From Global to Targeted Chromatin Proteomics: Mapping the Control Unit of Cellular Identity

Dissertation an der Fakultät für Biologie
der Ludwig-Maximilians-Universität München

Enes Ugur

München, 2023

Diese Dissertation wurde angefertigt
unter der Leitung von Prof. Dr. Heinrich Leonhardt
im Bereich Humanbiologie und BioImaging
an der Ludwig-Maximilians-Universität München

Erstgutachter:	Prof. Dr. Heinrich Leonhardt
Zweitgutachter:	Prof. Dr. Peter B. Becker
Dissertation eingereicht am:	09.08.2023
Mündliche Prüfung am:	07.12.2023

EIDESSTATTLICHE ERKLÄRUNG

Ich versichere hiermit an Eides statt, dass die vorgelegte Dissertation von mir selbstständig und ohne unerlaubte Hilfe angefertigt wurde.

München, den 09.08.2023 ....
Enes Ugur

ERKLÄRUNG

Hiermit erkläre ich, dass die vorliegende Dissertation weder ganz noch teilweise bei einer anderen Prüfungskommission vorgelegt wurde.
Ich habe mich nicht anderweitig einer Doktorprüfung ohne Erfolg unterzogen. Ich habe noch zu keinem früheren Zeitpunkt versucht, eine Dissertation einzureichen oder an einer Doktorprüfung teilzunehmen.

München, den 09.08.2023 ....
Enes Ugur

Meinen Eltern,
Meiner Familie

List of publications

1. **Ugur, E.**, de la Porte, A., Qin, W., Bultmann, S., Ivanova, A., Drukker, M., Mann, M., Wierer, M., & Leonhardt, H. (**2023**). Comprehensive chromatin proteomics resolves functional phases of pluripotency and identifies changes in regulatory components. *Nucleic Acids Research*, *51*(6), 2671-2690. https://doi.org/10.1093/nar/gkad058
2. Stolz, P., Mantero, A. S., Tvardovskiy, A., **Ugur, E.**, Wange, L. E., Mulholland, C. B., Cheng, Y., Wierer, M., Enard, W., Schneider, R., Bartke, T., Leonhardt, H., Elsässer, S. J., & Bultmann, S. (**2022**). TET1 regulates gene expression and repression of endogenous retroviruses independent of DNA demethylation. *Nucleic Acids Research*, *50*(15), 8491-8511. https://doi.org/10.1093/nar/gkac642
3. Wang, Z., Fan, R., Russo, A., Cernilogar, F. M., Nuber, A., Schirge, S., Shcherbakova, I., Dzhilyanova, I., **Ugur, E.**, Anton, T., Richter, L., Leonhardt, H., Lickert, H., & Schotta, G. (**2022**). Dominant role of DNA methylation over H3K9me3 for IAP silencing in endoderm. *Nature Communications*, *13*(1), 5447. https://doi.org/10.1038/s41467-022-32978-7
4. Qin, W., **Ugur, E.**, Mulholland, C. B., Bultmann, S., Solovei, I., Modic, M., Smets, M., Wierer, M., Forné, I., Imhof, A., Cardoso, M. C., & Leonhardt, H. (**2021**). Phosphorylation of the HP1β hinge region sequesters KAP1 in heterochromatin and promotes the exit from naïve pluripotency. *Nucleic Acids Research*, *49*(13), 7406–7423. https://doi.org/10.1093/nar/gkab548
5. Kempf, J. M., Weser, S., Bartoschek, M. D., Metzeler, K. H., Vick, B., Herold, T., Völse, K., Mattes, R., Scholz, M., Wange, L. E., Festini, M., **Ugur, E.**, Roas, M., Weigert, O., Bultmann, S., Leonhardt, H., Schotta, G., Hiddemann, W., Jeremias, I., & Spiekermann, K. (**2021**). Loss-of-function mutations in the histone methyltransferase EZH2 promote chemotherapy resistance in AML. *Scientific Reports*, *11*(1), 5838. https://doi.org/10.1038/s41598-021-84708-6
6. Bartoschek, M. D., **Ugur, E.**, Nguyen, T.-A., Rodschinka, G., Wierer, M., Lang, K., & Bultmann, S. (**2021**). Identification of permissive amber suppression sites

for efficient non-canonical amino acid incorporation in mammalian cells. *Nucleic Acids Research*, *49*(11), e62-e62.
https://doi.org/10.1093/nar/gkab132

7. Qin, W., Stengl, A., **Ugur, E.**, Leidescher, S., Ryan, J., Cardoso, M. C., & Leonhardt, H. (**2021**). HP1β carries an acidic linker domain and requires H3K9me3 for phase separation. *Nucleus (Austin, Tex.)*, *12*(1), 44–57.
https://doi.org/10.1080/19491034.2021.1889858
8. Mulholland, C. B., Traube, F. R., **Ugur, E.**, Parsa, E., Eckl, E.-M., Schönung, M., Modic, M., Bartoschek, M. D., Stolz, P., Ryan, J., Carell, T., Leonhardt, H., & Bultmann, S. (**2020**). Distinct and stage-specific contributions of TET1 and TET2 to stepwise cytosine oxidation in the transition from naive to primed pluripotency. *Scientific Reports*, *10*(1), 12066.
https://doi.org/10.1038/s41598-020-68600-3
9. **Ugur, E.**, Bartoschek, M. D., & Leonhardt, H. (**2020**). Locus-Specific Chromatin Proteome Revealed by Mass Spectrometry-Based CasID. *Methods in Molecular Biology* , *2175*, 109–121.
https://doi.org/10.1007/978-1-0716-0763-3_9
10. Mulholland, C. B., Nishiyama, A., Ryan, J., Nakamura, R., Yiğit, M., Glück, I. M., Trummer, C., Qin, W., Bartoschek, M. D., Traube, F. R., Parsa, E., **Ugur, E.**, Modic, M., Acharya, A., Stolz, P., Ziegenhain, C., Wierer, M., Enard, W., Carell, T., Bultmann, S., Leonhardt, H. (**2020**). Recent evolution of a TET-controlled and DPPA3/STELLA-driven pathway of passive DNA demethylation in mammals. *Nature Communications*, *11*(1), 5972.
https://doi.org/10.1038/s41467-020-19603-1

Declaration of contributions

1. **Comprehensive chromatin proteomics resolves functional phases of pluripotency and identifies changes in regulatory components**
   Enes Ugur and Michael Wierer conceived and designed the study. Heinrich Leonhardt, Michael Wierer and Matthias Mann supervised the study. Enes Ugur performed all experiments and MS data analysis. Enes Ugur, Heinrich Leonhardt and Michael Wierer interpreted the data. Alexandra de la Porte established the same culture conditions for human and mouse PSCs and conducted the cell culture under the supervision of Micha Drukker. Weihua Qin and Alina Ivanova performed validation experiments. Enes Ugur and Sebastian Bultmann programmed the web application. Enes Ugur, Heinrich Leonhardt, Michael Wierer and Matthias Mann wrote the manuscript and prepared the figures.

2. **TET1 regulates gene expression and repression of endogenous retroviruses independent of DNA demethylation**
   Paul Stolz and Sebastian Bultmann conceived and designed the study. Sebastian Bultmann supervised the study. Enes Ugur conducted TET1 ChIP and mass spectrometry analyses and interpreted the ChIP-MS data.

3. **Dominant role of DNA methylation over H3K9me3 for IAP silencing in endoderm**
   Gunnar Schotta, Zeyang Wang and Rui Fan conceived and designed the study. Gunnar Schotta supervised the study. Enes Ugur performed definitive and visceral endoderm differentiation and assessed H3K9me3 by western blot with and without SUV39H1/2 inhibition.

4. **Phosphorylation of the HP1β hinge region sequesters KAP1 in heterochromatin and promotes the exit from naïve pluripotency**
   Weihua Qin conceived and designed the study. Heinrich Leonhardt supervised the study. Enes Ugur performed ChIP and diGly enrichments, performed mass spectrometry analyses and subsequently interpreted the data.

5. **Loss-of-function mutations in the histone methyltransferase EZH2 promote chemotherapy resistance in AML**

Karsten Spiekermann, Julia M. Kempf and Sabrina Weser conceived and designed the study. Karsten Spiekermann supervised the study. Enes Ugur performed full proteome analysis and subsequently interpreted the data.

6. **Identification of permissive amber suppression sites for efficient non-canonical amino acid incorporation in mammalian cells**
   Michael Bartoschek and Sebastian Bultmann conceived and designed the study. Sebastian Bultmann and Kathrin Lang supervised the study. Michael Bartoschek, Enes Ugur, Kathrin Lang and Sebastian Bultmann interpreted the data. Michael Bartoschek and Enes Ugur conducted proteomic experiments. Enes Ugur performed mass spectrometry and analysis.

7. **HP1β carries an acidic linker domain and requires H3K9me3 for phase separation**
   Weihua Qin conceived and designed the study. Heinrich Leonhardt supervised the study. Enes Ugur performed proteome analysis of phase-separated droplets.

8. **Distinct and stage-specific contributions of TET1 and TET2 to stepwise cytosine oxidation in the transition from naive to primed pluripotency**
   Christopher B. Mulholland and Sebastian Bultmann conceived and designed the study. Sebastian Bultmann and Heinrich Leonhardt supervised the study. Enes Ugur conducted DPPA3 ChIP-MS and subsequently interpreted the data.

9. **Locus-Specific Chromatin Proteome Revealed by Mass Spectrometry-Based CasID**
   Enes Ugur conceived and designed the study. Heinrich Leonhardt supervised the study. Enes Ugur, Michael D. Bartoschek and Heinrich Leonhardt prepared, discussed and approved the manuscript.

10. **Recent evolution of a TET-controlled and DPPA3/STELLA-driven pathway of passive DNA demethylation in mammals**
    Christopher B. Mulholland and Sebastian Bultmann conceived and designed the study. Sebastian Bultmann and Heinrich Leonhardt supervised the study. Enes Ugur conducted full proteome analysis and subsequent data interpretation.

For each publication all authors read, discussed and approved the manuscript.

München, den 09.08.2023 ....

Enes Ugur Heinrich Leonhardt

CONTENTS

1 SUMMARY 1

2 INTRODUCTION 3

2.1 PROTEOMICS: FROM THE "GRUNDSTOFF" OF ANIMAL MATTER TO "NATURE'S ROBOTS" 3

2.1.1 Sample preparation in bottom-up Proteomics 6

2.1.2 Liquid chromatography coupled with tandem mass spectrometry 9

2.1.3 Data acquisition strategies in MS-based proteomics 12

2.1.4 Computational analysis of MS-based proteomics data 15

2.2 TO DIFFERENTIATE OR NOT TO DIFFERENTIATE: HALLMARKS OF PLURIPOTENCY 18

2.2.1 Origins of pluripotency 18

2.2.2 Phased progression of pluripotency 20

2.2.3 Transcriptional hallmarks of pluripotency phase transitions 22

2.2.4 Epigenetic hallmarks of pluripotency phase transitions 24

2.3 THE CHROMATOME: THE CONTROL UNIT OF CELLULAR IDENTITY 29

2.3.1 Global chromatin proteomics 32

2.3.2 Interactome-resolving chromatin proteomics 37

2.3.3 Gene locus-resolving chromatin proteomics 42

2.3.4 Histone PTM-resolving chromatin proteomics 45

3 DISCUSSION 49

3.1 ACHIEVING COMPREHENSIVE AND ACCURATE GLOBAL CHROMATIN PROTEOMICS 52

3.1.1 Improving the chromatin selectivity of ChAC-DIA 52

3.1.2 Expanding the scope of global chromatin proteomics 54

3.1.3 Towards high-throughput and single-cell ChAC-DIA 55

3.1.4 High-resolution DIA-MS for chromatin proteomics 57

3.1.5 Chromatome analysis and multi-modal data integration 59

3.2 CHROMATOME REORGANIZATION DURING PLURIPOTENCY 61

3.2.1 Establishment of a repressive chromatin state towards pluripotency exit 61

3.2.2 The epigenetic paradox of pluripotency 64

3.2.3 Co-existence of different cell-identity governing transcription factors 66

3.2.4 Conserved pluripotency features between mouse and human PSCs 66

4 BIBLIOGRAPHY 69

5 ANNEX ____ 105

5.1 ORIGINAL PUBLICATIONS ____ 105
5.2 LIST OF ABBREVIATIONS ____ 301
5.3 CURRICULUM VITAE ____ 304
5.4 ACKNOWLEDGEMENTS ____ 305

1 SUMMARY

Cellular identity is established and maintained by the chromatome, which consists of transcriptional, epigenetic and structural regulators of the chromatin proteome. Serving as a control hub, the chromatome processes incoming signaling cues and modifies the transcriptional program, resulting in a specific cellular phenotype. To fully understand cell-type specific gene regulation, multi-level chromatome analysis is necessary. Chromatin-associated proteins can be explored using mass spectrometry (MS)-based proteomic methods to assess (i) total protein abundances, (ii) chromatin-associated individual complexes, (iii) global or (iv) locus-specific chromatin compositions, and (v) nucleotide and histone (post-translational) modifications.

Global chromatin proteomics trails behind other areas of proteomics in terms of data comprehensiveness, accuracy, and throughput. The main aim of this work was the development of an MS-based proteomic method, Chromatin Aggregation Capture followed by Data Independent Acquisition (ChAC-DIA), which enables the comprehensive identification and accurate quantification of chromatin regulators, including those present in low quantities, across different pluripotency stages. ChAC-DIA identified 2-3 times more chromatin-associated proteins with enhanced accuracy and efficiency, required 100-fold less sample material, and halved the MS data acquisition time compared to prior methods. By applying ChAC-DIA an extensive atlas was constructed that encompasses proteomes, chromatomes, and chromatin affinities across the three key phases of pluripotency. The data served not only to verify *bona fide* pluripotency regulators such as REX1, OCT6 and SOX1, but also to identify new phase-specific factors like JADE1/2/3, QSER1, SUV39H1/2 and FLYWCH1. Moreover, this study offers a straightforward strategy for distinguishing between translation-driven changes in chromatin binding and alterations in nuclear localization or chromatin affinity. Using this approach, we observed that certain heterochromatic proteins, such as HP1β, KAP1, and SUV39H1, exhibited enhanced chromatin affinities towards the exit from pluripotency,

which we could demonstrate to be a conserved feature in both mouse and human.

In subsequent collaborative endeavors, chromatin proteomics was applied to study epigenetic regulations in several biological contexts. In three distinct projects, chromatin immunoprecipitation combined with MS was employed to analyze the interaction networks of the naive pluripotency marker DPPA3, the histone H3 lysine 9 trimethyl (H3K9me3) reader HP1β and the methylcytosine dioxygenase TET1. Moreover, the KAP1-dependent ubiquitinome was investigated, the composition of HP1β-driven phase-separated droplets was studied, and a proteomic workflow was developed to screen for the efficient incorporation of non-canonical amino acids into target proteins. This work also provided a detailed protocol for probing locus-specific chromatin composition as well as full proteome analyses upon genetic perturbations targeting epigenetic modifiers in an acute myeloid leukemia cell culture model and embryonic stem cells at various pluripotency stages. In a last collaboration, it was tested whether the histone methyltransferases SUV39H1/2 primarily contribute to H3K9me3 formation in visceral endoderm descendants.

In summary, this work provides a powerful method to study the global chromatome of any model in development and disease, sheds new light on dynamic rearrangements of pluripotency governing regulatory complexes and contributes to a broad range of epigenetic research by harnessing multi-level chromatome analyses.

2 INTRODUCTION

2.1 Proteomics: from the “Grundstoff” of animal matter to “nature’s robots”

In 1789, the world was roused by the French Revolution, as the people began to question the aristocracy as a ruling system. While the streets of Paris were filled with cries for liberty, equality, and fraternity, a quieter revolution was taking place in the laboratories of Antoine Fourcroy and of other contemporary chemists. Fourcroy was one of the major contributors to the chemical revolution of the 17th and 18th centuries (Bensaude-Vincent, 1990). With the advent of precision instruments and mathematizing chemistry, alchemy transitioned to modern chemistry by the introduction of the concept of elements. Focused on studying animal tissues to unlock the secrets of life itself, Fourcroy discovered that the substances he had previously thought to be one and the same - albumin, fibrin, and gelatin - were in fact three distinct variances of a novel class of substances, termed “Eiweisskörper” (Tanford and Reynolds, 2001). While the world outside was consumed by upheaval and turmoil, Fourcroy's discovery quietly paved the way for a new era of scientific inquiry, that would eventually lead to our modern understanding of proteins.

Half a century later, when the concept of molecules had not yet been established, Dutch chemist Gerrit Mulder analyzed “Eiweisskörper” by measuring the elemental compositions of serum or egg albumin and fibrin. Mulder’s meticulous work revealed the first quantitative evidence for Fourcroy’s earlier observation, namely that all proteins shared the same elemental composition, suggesting that they were part of the same chemical group (Tanford and Reynolds, 2001). It was this groundbreaking discovery that Mulder shared with the Swedish chemist Jacob Berzelius in a letter in 1838 stating that he has found the “Grundstoff” of animal matter. Intrigued, Berzelius postulated that this might be the “principal substance of animal nutrition” produced by plants, spread by herbivores and consumed by carnivores. Instead of "Grundstoff", Berzelius proposed the term "proteins" derived from the Greek word proteios (πρώτειος), meaning "standing in front" or "of prime importance" (Vickery, 1950).

During the subsequent thirty years, the concept of atomic organization within molecules significantly evolved, spearheaded by August Kekulé's discovery of the tetravalence of carbon (Benfey, 1958). Concurrently, the first amino acids were discovered and until the end of the 19th century almost all amino acids had been described and were known to be the primary components of proteins. Still, it was ambiguous how these amino acids were held together to build a protein. This mystery was unraveled independently in 1902 by Franz Hofmeister (Hofmeister, 1902) and Emil Fischer (Fischer, 1906), who presented their findings at the annual meeting of the "Gesellschaft deutscher Naturforscher und Ärzte" (Tanford and Reynolds, 2001).

By the beginning of the 20th century, proteins were known to be the building blocks of life and consist of amino acids which are linked by peptide bonds. Yet, the sequential arrangement of amino acids within a protein, and the relevance of this sequence, was still unclear (Crick and Anderson, 1989). After a decade-long endeavor from 1945 to 1955, English biochemist Frederik Sanger succeeded in sequencing insulin, utilizing a combination of chemical hydrolysis of peptides and subsequent electrophoresis and chromatography (Sanger, Thompson and Kitai, 1955; Tanford and Reynolds, 2001). From the mid-20th century onwards, protein sequencing was then dominated by Pehr Edman's method, "Edman sequencing", which involved the N-terminal labeling of a polypeptide chain, followed by sequential cleavage and identification of individual amino acids through chromatography (Edman *et al.*, 1950; M. Mann, 2016). Edman sequencing along other methods to study proteins, primed our understanding of proteins as the actual functional entities inside cells, leading to the concept of proteins as "nature's robots" (Tanford and Reynolds, 2001). From the late 1980s on then, mass spectrometry gained momentum as a new technology in proteomics research.

But how did mass spectrometry evolve to become the method of choice for identifying and quantifying proteins? The fundamental principle underlying mass spectrometry is the measurement of the mass-to-charge ratio (m/z) of ions. One could therefore say that the 1897 discovery of the electron by Joseph John Thomson set the stage for the eventual emergence of mass spectrometry (Figure 1) (Thomson, 1897; Yates, 2011). In subsequent years, Thomson and his assistant, Francis William Aston, experimented with positive rays and discovered non-radioactive elemental isotopes, such as neon isotopes, thereby pioneering the study of atomic and molecular masses (Thomson, 1914). Thomson underscored the significance of these discoveries with his statement, "*The positive rays thus seem to promise to furnish a method of investigating the structure of the molecule, a subject certainly of no less importance than that of the structure of the atom*" (Thomson, 1914). After returning from World War I, Aston made another significant contribution by engineering the first mass spectrometer (Aston, 1919). Subsequent advancements, including the development of time-of-flight (TOF) and quadrupole mass analyzers, facilitated the analysis of small molecules (Stephens, 1946; Paul and Steinwedel, 1953). Nevertheless, it was not until the development of electrospray ionization (ESI) and matrix-assisted laser desorption/ionization (MALDI) in the 1980s that mass spectrometry became a powerful tool for the analysis of proteins (Karas and Hillenkamp, 1988; Fenn *et al.*, 1989; Mann, Meng and Fenn, 1989).

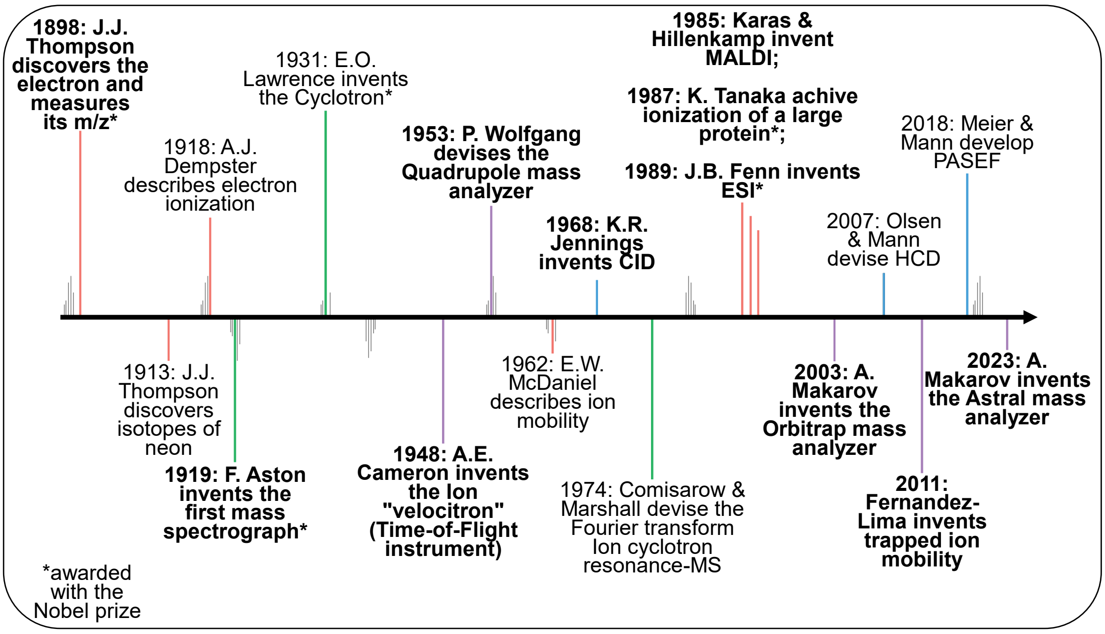

1898: J.J. Thompson discovers the electron and measures its m/z*
1913: J.J. Thompson discovers isotopes of neon
1918: A.J. Dempster describes electron ionization
1919: F. Aston invents the first mass spectrograph*
1931: E.O. Lawrence invents the Cyclotron*
1948: A.E. Cameron invents the Ion "velocitron" (Time-of-Flight instrument)
1953: P. Wolfgang devises the Quadrupole mass analyzer
1962: E.W. McDaniel describes ion mobility
1968: K.R. Jennings invents CID
1974: Comisarow & Marshall devise the Fourier transform Ion cyclotron resonance-MS
1985: Karas & Hillenkamp invent MALDI;
1987: K. Tanaka achive ionization of a large protein*;
1989: J.B. Fenn invents ESI*
2003: A. Makarov invents the Orbitrap mass analyzer
2007: Olsen & Mann devise HCD
2011: Fernandez-Lima invents trapped ion mobility
2018: Meier & Mann develop PASEF
2023: A. Makarov invents the Astral mass analyzer
*awarded with the Nobel prize

Figure 1: Overview of milestones in mass spectrometry-based proteomics. This timeline illustrates important scientific discoveries for MS-based proteomics, indicated by red lines, alongside the inventions of MS instruments, denoted by green lines, and mass analyzers, denoted by violet lines. Further advancements in MS instruments or MS acquisition strategies are highlighted by blue lines.

The introduction of a tandem MS strategy (MS/MS) and the integration of mass spectrometry with liquid chromatography (LC-MS/MS) by Donald Hunt and his team improved MS sensitivity and selectivity, allowing for the analysis of complex peptide and protein mixtures (Hunt *et al.*, 1981, 1992; Mann, 2016). Another milestone achievement was the development of the Orbitrap mass analyzer by Alexander Makarov and colleagues, which combines the high resolution and mass accuracy of magnetic sector instruments with the high sensitivity of quadrupole and ion trap devices (Hardman and Makarov, 2003). Since then, Orbitrap-based mass spectrometry has emerged as an efficient tool for analyzing complex proteomes. Recently, TOF instruments that leverage ion mobility-based peptide ion separation have gained popularity due to their high scan speeds and ion capacity (Fernandez-Lima, Kaplan and Park, 2011; Ridgeway *et al.*, 2018). Hence, the most recent integration of both Orbitrap and a TOF-like mass analyzer, termed Astral (Asymmetric Track Lossless), has attracted substantial interest (Heil *et al.*, 2023). Further technological advancements in mass spectrometry, liquid chromatography and sample preparation, coupled with sophisticated downstream computational analysis, have driven unprecedented biological insights. These insights range from the characterization of tens of thousands of canonical or perturbed proteomes and post-translational modifications (PTMs) (Perez-Riverol *et al.*, 2022), to the comprehensive mapping of thousands of protein interactomes (Hein *et al.*, 2015; Cho *et al.*, 2022) and to the identification of numerous novel clinical disease markers (Crutchfield *et al.*, 2016). At present, MS-based proteomics, alongside other *in vitro* protein characterization methods and imaging technologies, continues to enrich our understanding of proteins, increasingly substantiating their role as "nature's robots".

2.1.1 Sample preparation in bottom-up Proteomics

The identification and quantification of proteins in mass spectrometry-based proteomics depend on one of these three principal approaches: bottom-up, top-down or middle-down (Figure 2) (Chait, 2006). Bottom-up proteomics is particularly efficient in the identification and quantification of thousands of proteins as it involves the fragmentation of proteins into smaller peptides consisting of 7-30 amino acids (aa), thereby facilitating their analysis. Conversely, the top-down approach involves analyzing intact proteins. While this is effective in discerning combinatorial PTMs and proteoforms (*i.e.* all possible forms a protein product of a single gene can have (Smith, Kelleher and Consortium for Top Down Proteomics, 2013)), it necessitates sample fractionation and yields complex mass spectra, reducing throughput and overall proteome coverage (Catherman, Skinner and Kelleher, 2014). The middle-down approach involves the sequence-specific digestion of proteins into 30-100 aa long peptides, enabling the detection of combinatorial PTMs and proteoforms with fewer experimental challenges than top-down proteomics (Taverna *et al.*, 2007). Currently, most mass spectrometry-based proteomics experiments follow a bottom-up strategy to achieve a deep analysis of full proteomes in cell culture or tissue-derived samples (Gillet, Leitner and Aebersold, 2016).

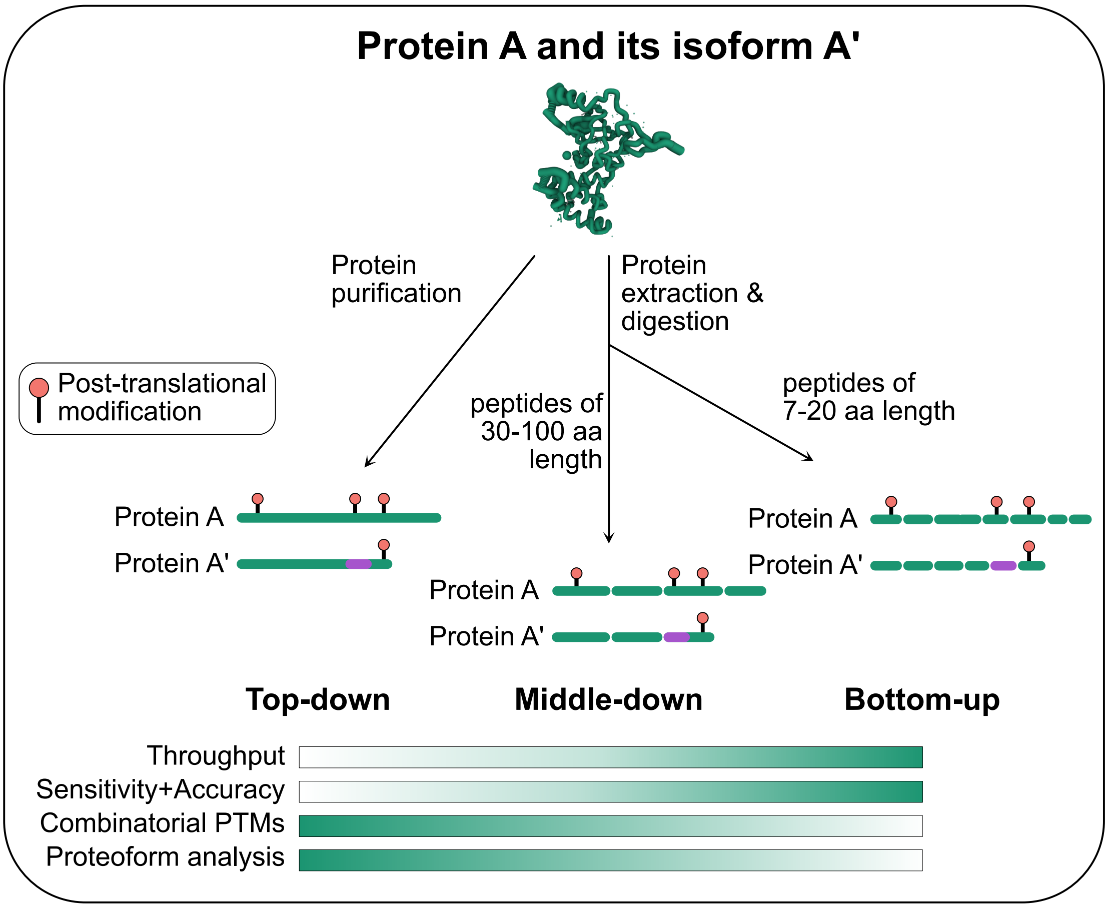


Figure 2: Overview of MS sample preparation strategies and respective outcomes. The throughput and sensitivity decrease progressively from bottom-up to top-down approaches, while the number of measurable combinatorial PTMs increases. In addition, top-down proteomics facilitates the identification of proteoforms.

In any mass spectrometry experiment, sample preparation is essential for achieving high reproducibility and sensitivity. The core experimental steps in a standard bottom-up proteomics method include: (i) cell lysis preferably using mass spectrometry-compatible detergents such as sodium deoxycholate (SDC) or guanidinium hydrochloride (GdCl); (ii) sample boiling and sonication, or treatment with micrococcal nuclease (MNase) to maximize protein solubilization and minimize viscosity caused by chromatin; (iii) in-solution protein digestion with trypsin; and (iv) peptide desalting prior to LC-MS/MS.

However, investigations of PTMs, protein-protein interactions, subcellular structures, or other specific proteomics applications often require more sophisticated sample preparation techniques. This is, for instance, the case when detergents incompatible with mass spectrometry are used, as they might interfere with peptide ionization or compromise mass spectrometry sensitivity. Hence, protein cleanup is required which is achieved by techniques such as peptide digestion in filter centrifuge tubes (FASP, filter-aided sample preparation) or sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS-PAGE) followed by in-gel digestion (Shevchenko *et al.*, 1996; Wilm *et al.*, 1996; Wiśniewski *et al.*, 2009). Other alternatives include acetone precipitation followed by in-solution digestion (Görg *et al.*, 1998) or protein aggregation capture (PAC) (Batth *et al.*, 2019). PAC involves unbiased precipitation of proteins on sub-micron beads, enables single-pot purification, stringent washing conditions, and on-bead digestion of proteins.

Sample preparation strategies frequently involve the reduction of disulfide bonds and the alkylation of cysteines prior to protein digestion to enhance digestion efficiency. This is generally followed by trypsin-based protein digestion, given its defined and frequent cleavage pattern at the C-terminal sides of arginine and lysine (Olsen, Ong and Mann, 2004). Following protein digestion, peptide mixtures are desalted by, for example, solid-phase extraction with the STop And Go Extraction tips (StageTip) (Rappsilber, Ishihama and Mann, 2003; Rappsilber, Mann and Ishihama, 2007). In this method, peptides are captured within a hydrophobic C_{18} (Octadecyl) disk embedded into a pipet tip, while salts are largely not retained. After several washing steps, peptides are eluted into an MS-compatible buffer.

To enhance the proteome coverage, it is beneficial to perform peptide fractionation prior to LC-MS/MS (also known as offline fractionation). An orthogonal fractionation method, complementing the MS-coupled online LC is favorable. Well-established offline fractionation strategies include continuous strong anion/cation exchange (SAX/SCX, respectively) and high pH reversed-phase fractionation (Ducret *et al.*, 1998; Issaq *et al.*, 2002). The combination of bottom-up proteomics with offline fractionation has enabled the identification of over 12,000 proteins, yielding nearly similar identification rates as those found in transcriptomics (Kulak, Geyer and Mann, 2017). With the continual advancement of MS technology and the introduction of novel MS-coupled chromatography devices, it is expected that proteomic analyses will routinely achieve the same depth as transcriptomic analyses. The upcoming chapter will explain the LC-MS/MS setup employed throughout this work and also explore recent technological developments.

2.1.2 Liquid chromatography coupled with tandem mass spectrometry

The combination of liquid chromatography with tandem mass spectrometry allows for harnessing three different dimensions for peptide identification and quantification: retention time, m/z and peptide intensity. The incorporation of an ion mobility analyzer into the mass spectrometry device introduces an additional dimension, known as the ion mobility space (Fernandez-Lima, Kaplan and Park, 2011). Retention time is defined as the duration a peptide requires to pass through the LC-column connected to the mass spectrometer before it undergoes ionization. This process usually involves a reversed-phase high-performance liquid chromatography (HPLC) setup operating at low pH and under high-pressure conditions. This setup incorporates a 15-50 cm long analytical column packed with hydrophobic C_{18} coated silica beads. A slow flow rate of 100-400 nL/min is applied to promote highly efficient and consistent peptide ionization (Davis *et al.*, 1995; Figeys and Aebersold, 1998).

Traditional nanoflow HPLC instruments have limitations in robustness and speed, leading to the development of alternative setups, such as the Evosep One LC system (Falkenby *et al.*, 2014; Bache *et al.*, 2018). This system facilitates peptide elution directly from a StageTip at low pressure, which allows for the preparation of a peptide gradient before the application of high pressure in the analytical column. The Evosep One exhibits substantially improved robustness and throughput, enabling, for example, full proteome measurements within 5 minutes per sample, remarkably resulting in the identification of up to 5,000 proteins (Z. Wang *et al.*, 2022). LC methods consistently separate peptides based primarily on hydrophobicity and other physicochemical attributes in a linear acetonitrile gradient. Consequently, peptides of the same kind are co-eluted, thus reducing the overall sample complexity analyzed by the mass spectrometer at any given time. Peptide elutions typically follow a bell-shaped abundance distribution, referred to as the chromatographic peak. At each point of this peak, peptides can be selected for tandem mass spectrometric analysis.

For analysis, mass spectrometers require gaseous and ionized peptides, which can be obtained by soft or hard ionization techniques. While hard ionization can lead to random peptide fragmentation, complicating the identification process, soft ionization mitigates this fragmentation. The most frequently used soft ionization technique is currently ESI, which has negligible peptide fragmentation characteristics, can generate multiple charged ions and can be easily coupled to an HPLC (Fenn *et al.*, 1989; El-Aneed, Cohen and Banoub, 2009). Generating multiple charges comes with the advantage that higher masses can be resolved as each charge reduces the m/z. Alternatively, peptides are ionized by MALDI (Hillenkamp *et al.*, 1991). Unlike ESI, which works with peptides in solution, MALDI requires samples to be dried and embedded in a matrix. The sample matrix mixture is irradiated by UV pulses which induces gaseous ion formation of matrix components as well as peptides. In ESI, charged droplets containing peptides in a solvent are continuously emitted from the LC column, forming a Taylor cone (Wilm and Mann, 1994). These droplets either progressively evaporate, leading to the formation of gaseous peptide ions (charged residue mechanism) (Fernandez de la Mora, 2000) or ions are expelled from droplets through electrostatic repulsion (ion evaporation mechanism) (Kebarle, 2000; Kebarle and Verkerk, 2009). The exact principle of gaseous ion formation in ESI is not fully elucidated and may involve both described mechanisms, depending on the respective droplet size and the analyte's properties (Aliyari and Konermann, 2022).

But what happens in the mass spectrometer, the subsequent "black box", after peptide separation and ionization? In the quadrupole-Orbitrap hybrid mass spectrometer used in this work (Figure 3), charge-neutral peptides are first filtered out and ionized peptides are focused and directed into a quadrupole mass analyzer (Woodward and Crawford, 1963). The quadrupole consists of four cylindrical rods which generate an electrodynamic field. Each metal rod is charged inversely to its opposite rod, and the charges are swiftly switched as ions traverse the quadrupole. Compared to other mass analyzers, the quadrupole has limitations in resolution and accuracy and, hence, is not used as a mass analyzer when coupled to an Orbitrap (Hardman and Makarov, 2003) or a TOF (Glenn, 1952) mass analyzer. Instead, the quadrupole is often employed as a mass filter due to its speed and efficiency, selectively guiding ions based on their m/z to the mass analyzer. The chosen m/z range typically spans 400-1600 Thomson (Th, unit of m/z), which the Orbitrap mass analyzer can cover. Filtered peptide ions enter the C-trap, essentially a bent quadrupole that traps ions by absorbing their kinetic energy using an electromagnetic field and nitrogen (Olsen *et al.*, 2005). Ions are trapped for a defined amount of time, which is in the range of milliseconds, while the C-trap in parallel directs ions to the Orbitrap for a full mass scan of all incoming ions ("MS1 scan"). The Orbitrap mass analyzer, introduced by Mark Hardman and Alexander Makarov in 2003, consists of a spindle-shaped central electrode surrounded by two hemispherical electrodes, creating a cylindrical structure (Hardman and Makarov, 2003). Ions entering the Orbitrap oscillate around the central electrode without physical contact as the trap potential swiftly changes. Within the Orbitrap, the oscillation frequency depends on a peptide ion's m/z, and the MS1 scan is achieved using an induced image current based on the axial movement of ions. This current, which is recorded for tens to hundreds of milliseconds, generates a potential difference between the two halves of the outer electrode, which is then used to determine the m/z of the ions. MS1 scans can take hundreds of milliseconds to several seconds and determine the m/z and intensities of peptide ions. Complex proteomes require an additional MS scan of fragmented peptide ions ("MS2 scan") for peptide sequence identification. This is achieved by guiding peptide ions at selected m/z from the quadrupole to the C-trap, which feeds these ion packages into the neighboring collision chamber for fragmentation. Techniques such as collision-induced dissociation (CID) and higher energy collisional dissociation (HCD) are commonly used for fragmentation, resulting in the formation of b-ions (containing the C-terminus of the peptide) and y-ions (containing the N-terminus of the peptide) (Steen and Mann, 2004). These fragmented peptide ions are then returned to the Orbitrap for mass analysis in the MS2 scan, ultimately enabling peptide sequence determination (Shuken, 2023).

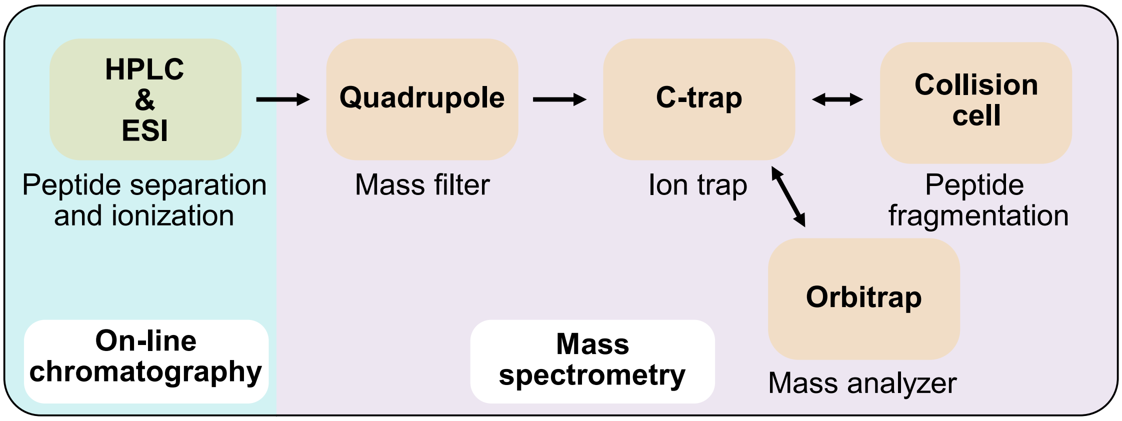


Figure 3: Schematic of key components of the Orbitrap Exploris 480 MS instrument. Peptides are separated by HPLC and ionized by ESI. These gaseous peptide ions are directed to the quadrupole, which filters peptide ions based on a predefined m/z range, charge state and other criteria. While the quadrupole can function as a mass analyzer, it primarily serves as a mass filter in the Orbitrap Exploris 480. Filtered peptide ions are trapped within the C-trap and a subset of these ions is sent to the Orbitrap for the

initial full MS scan (MS1). This is followed by a selective injection of peptide ions from the C-trap to the collision cell, where peptide ions are fragmented and then fed back into the C-trap. The C-trap then sends fragmented peptide ions to the Orbitrap for the final MS scan (MS2).

The results presented in this work were acquired, in part, using the Q Exactive HF-X, the predecessor of Thermo Fisher Scientific's quadrupole-Orbitrap hybrid mass spectrometer. This instrument exhibited an enhanced ion source, more efficient isotope detection capabilities and faster tandem MS acquisition speeds than its predecessor, the Q Exactive HF. As a result, the Q Exactive HF-X could attain the same proteome coverage as the Q Exactive HF in half the gradient time or with 10-fold lower sample loads (Kelstrup *et al.*, 2018). The next quadrupole-Orbitrap mass spectrometer, the Exploris 480 (Bekker-Jensen, Martínez-Val, *et al.*, 2020), was designed with a focus on ease of maintenance and is capable of identifying over 1,000 proteins from a 5 ng full proteome sample following a 5 min LC gradient. Despite these technological advancements, the C-trap, which filters and prepares ions for the Orbitrap, remains a limitation for the overall capacity of hybrid quadrupole-Orbitrap instruments. TIMS-TOF instruments, in contrast, leverage the high ion capacity of one or multiple TIMS devices paired with a fast TOF mass analyzer. When combined with scan modes that synchronize quadrupole ion selection with the ion mobility, these instruments harbor significant potential for future high-sensitivity and high-throughput proteomics research (Meier *et al.*, 2015, 2020; Meier, Park and Mann, 2021; Skowronek *et al.*, 2022). Recent publications highlighted the advancements achieved by using an Orbitrap for MS1 scans and the TOF-like mass analyzer Astral for MS2 scans, enabling the identification of 10,000 proteins in 48 minutes (Stewart *et al.*, 2023). It is very likely that future technologies will continue to push the boundaries, providing deeper biological insights through mass spectrometry.

2.1.3 Data acquisition strategies in MS-based proteomics

Bottom-up MS-based proteomics experiments primarily utilize data-dependent acquisition (DDA) for peptide selection following the initial MS1 scan (Figure 4). In DDA, the MS1 acquisition is succeeded by the selection of the 10-15 most abundant peptide ions for MS2 acquisition (Stahl *et al.*, 1996; Sinitcyn, Rudolph and Cox, 2018). While this technique simplifies computational data analysis by yielding MS2-level spectra that can be matched to each selected MS1-level peak, it significantly reduces the protein identification rate in a single MS run. DDA-based experiments often provide redundant information on highly abundant proteins, while the overlap between technical replicates falls below 75% due to missed identifications (also known as missing values), which renders the quantification of low-abundant proteins challenging (Tabb *et al.*, 2010). The disadvantages of DDA become clearer when analyzing samples covering a broad range of protein abundances since highly abundant peptide ions can "suppress" the signal of less abundant ones. Furthermore, the semi-stochastic nature of DDA complicates the analysis of enriched cellular components consisting of highly repetitive structural elements such as nucleosomes within the chromatin proteome (also known as the chromatome) (Imhof and Bonaldi, 2005). For instance, DNA-binding proteins span a wide dynamic range within the chromatome, making the identification of low-abundance peptide ions more challenging. Previous studies of the chromatome attempted to circumvent this issue by employing extensive peptide-level fractionation prior to LC-MS/MS (Ginno *et al.*, 2018) or by employing prolonged MS acquisition times (van Mierlo, Wester and Marks, 2019). Despite these labor-intensive and costly solutions, the coverage of chromatin-associated proteins, remains below expectations based on transcriptome analyses.

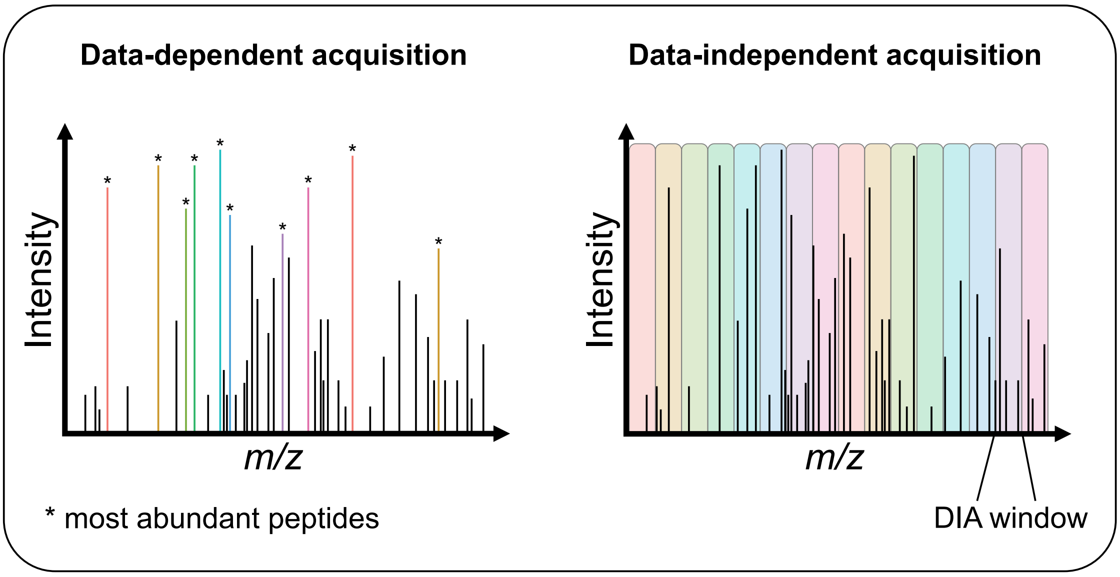


Figure 4: Schematic representation of a full MS scan (MS1) addressed either to DDA-based (left) or DIA-based (right) MS2 acquisition. In DDA, the most abundant peptides from the preceding MS1 scan at any given time are selected for further fragmentation and MS2 acquisition. This often includes the 10-15 most abundant peptides. Conversely, in DIA, the m/z range is subdivided into isolation windows of 25-35 Thomson (Thomson or Th, unit of the m/z). A greater number of these DIA isolation windows will yield a higher resolution of a given MS1 scan than fewer

windows, but will also increase the time required for MS2 acquisition. This increase in acquisition time consequently results in fewer total acquired MS1 scans. Therefore, the number of DIA windows is frequently empirically optimized according to the specific kind of sample under analysis.

One potential solution to this issue is the implementation of an alternative acquisition strategy known as data-independent acquisition (DIA). Compared with DDA, this strategy results in a higher identification rate of peptides with fewer missing values across replicates, while it offers more precise protein quantifications (Venable *et al.*, 2004; Gillet *et al.*, 2012; Bruderer *et al.*, 2015, 2017; Ludwig *et al.*, 2018). In DIA, all precursor ions falling within a predefined m/z window undergo fragmentation and are acquired at the MS2 level, resulting in highly complex spectra. This computational challenge was initially addressed by harnessing sample-specific DDA-based spectral libraries to deconvolute the data (Gillet *et al.*, 2012). To build a comprehensive spectral library, sample-specific peptides can be offline fractionated prior to acquisition (Schubert *et al.*, 2015). Alternatively, a spectral library can be generated directly from DIA measurements, and a hybrid spectral library can be created by combining DDA and DIA libraries (Bader *et al.*, 2020). A second approach, known as library-free or direct DIA, involves a spectrum-centric strategy where DIA-MS2 spectra are deconvoluted into pseudo-MS/MS spectra, followed by conventional database searches (Bruderer *et al.*, 2015; Tsou *et al.*, 2015). To combine the advantages of spectral libraries and direct DIA, a third approach utilizes computational tools that predict retention times and MS2 spectra based on trained deep neural networks (DNNs) within the respective peptide search space (Gessulat *et al.*, 2019; Tiwary *et al.*, 2019; Demichev *et al.*, 2020). These peptide-centric approaches map DIA-MS2 spectra to clear MS2 spectra from the DNN-based library, offering considerable convenience by eliminating the need for a spectral library while achieving comparable protein identification rates and quantification accuracies in significantly shorter analysis times (Bekker-Jensen, Bernhardt, *et al.*, 2020; Demichev *et al.*, 2020; Pino *et al.*, 2020; Lou *et al.*, 2023). Importantly, this recent computational breakthrough has led to widespread acceptance of DIA in the field of proteomics. The next chapter describes the general principles of computational MS analysis and briefly discusses additional challenges posed by DIA measurements.

2.1.4 Computational analysis of MS-based proteomics data

> *"One day I commented to my then student, Matthias Mann, that each peak in one of these multiple peak spectra was really an independent measure of the parent ion mass. Therefore, there should be some way of averaging those independent values to get a more reliable and accurate measure of the parent molecule mass than any single peak spectrum could provide. Two days later he had worked out a computer algorithm that transformed the multiple peaks into a single peak that would be obtained if all the ions had a single massless charge. The m/z value for that peak is thus the M_r [i.e., relative molecular mass] value for that species."* (John B. Fenn, Nobel Lecture)

Soft ionization of biological compounds through ESI or MALDI enabled the mass spectrometry of peptides and proteins. The next logical step was to develop computational methodologies to link measured spectra with their parent ions, facilitating the analysis of complex peptide or protein mixtures. Modern mass spectrometers acquire data across multiple dimensions, rendering each MS feature a higher dimensional object. The most direct information retrieved through mass spectrometry is the m/z, corresponding to one isotopic variant of a peptide ion. Additional peaks representing other isotopic forms are also recorded, generating isotope patterns.

John Fenn's insight and Matthias Mann's algorithmic approach enabled for the first time the computational matching of peptides with corresponding isotope patterns and thus paved the way for seminal advancements in computational mass spectrometry data analysis (Mann, Meng and Fenn, 1989; Mann, Højrup and Roepstorff, 1993). For complex peptide mixtures covering a total proteome, these isotope patterns can be interpreted using models such as the Averagine model which employs an average representation of an amino acid, known as an Averagine, to calculate isotope pattern compositions. Computed isotope patterns are composed of isotope mass differences and relative heights of isotopic peaks for a given mass which are then matched with acquired mass spectra (Senko, Beu and McLaffertycor, 1995; Craig and Beavis, 2004; Nesvizhskii, 2007).

A conventional peptide identification process begins with the generation of a database from a user-defined FASTA file, which encompasses all proteins presumed to be present in a sample. These proteins are *in silico* digested, following the protease digestion pattern of choice. For instance, in the case of trypsin, peptide bonds are cleaved between the carboxyl group of arginine or lysine and the amino group of the adjacent amino acid. Frequently, 1-2 missed cleavage sites are allowed for this *in silico* digestion to account for often missed cleavage sites. Subsequently, theoretical MS2 spectra are generated based on the calculated peptide sequences and the corresponding fragmentation technique. Each acquired spectrum of unfragmented and fragmented peptide ions is then compared to all theoretical MS2 spectra within a predefined mass tolerance, typically within several parts per million. The peptide spectrum match (PSM) with the highest score is considered to represent the identity of the peptide most accurately. Various software solutions, such as SEQUEST (Eng, McCormack and Yates, 1994), Mascot (Perkins *et al.*, 1999) and Andromeda (Cox *et al.*, 2011) have been developed to generate PSMs (Verheggen *et al.*, 2020).

However, solely relying on PSM scores derived from the peptide database can lead to numerous false PSM matches. The likelihood of false identifications increases with the number of theoretical peptides being investigated (Colaert *et al.*, 2011). Consequently, MS raw data processing tools implement a false discovery rate (FDR) control by matching acquired fragmentation spectra against a decoy database, to define a PSM score threshold that maintains the FDR within an acceptable range, often around 1%. Certain tools, such as Andromeda/MaxQuant, incorporate additional peptide attributes, such as length, charge state, or number of missed cleavages, in addition to the search engine score.

For the analysis of highly complex LC-MS/MS data, especially in the case of DIA based measurements, this peptide-centric approach alone is inadequate for maintaining a low FDR. Therefore, additional filtering criteria are applied to PSMs. Recently, DNNs have been trained to discriminate between correct PSMs and false identifications. These DNNs have been seamlessly integrated into mass spectrometry analysis tools, enhancing the precision and reliability of PSM identification (Demichev *et al.*, 2020; Meyer, 2021; Sinitcyn *et al.*, 2021). In summary, the computational analysis of mass spectrometry data has evolved considerably since the pioneering work of John Fenn, Matthias Mann and other following scientists. Through the integration of advanced algorithms, statistical controls, and machine learning approaches, present-day tools are capable of processing highly complex datasets with improved comprehensiveness and accuracy. As research in this field continues to advance, it is anticipated that new methodologies and software tools will further streamline and enhance the MS acquisition workflow (Figure 5).

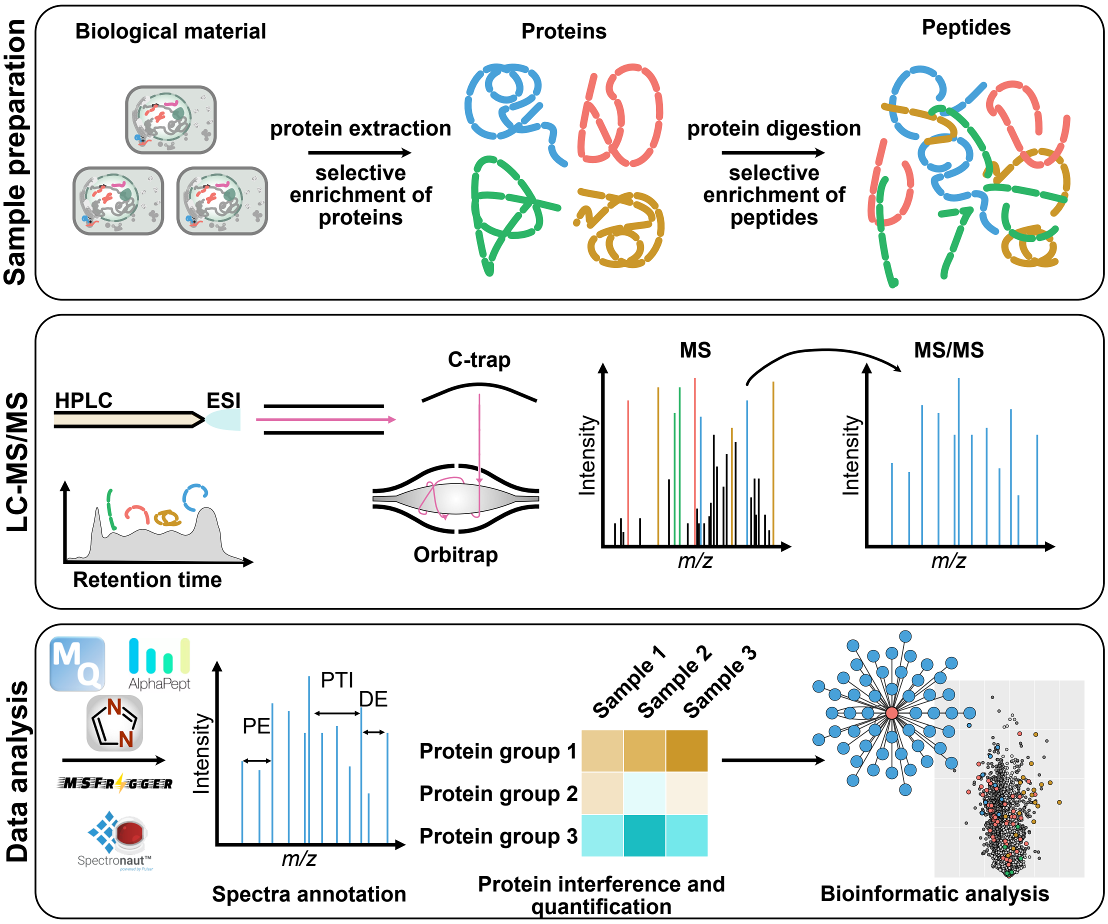


Figure 5: Graphical summary of a bottom-up proteomics workflow from sample preparation to tandem mass spectrometry and data analysis. The biological material for analysis may be sourced from cell cultures, tissues, or body fluids. The first two types of samples undergo cellular lysis for protein extraction, which may be followed by selective enrichment of specific proteins, such as a protein complex of interest. These proteins are then digested, and the resulting peptides can either be selectively enriched (for instance, for PTMs) or desalted in preparation for mass spectrometry. The LC-MS/MS step involves peptide separation by HPLC, subsequent peptide ionization, and the acquisition of MS1 and MS2 scans. Software tools, including AlphaPept, MaxQuant, DIA-NN, MSFragger, or Spectronaut, annotate the acquired MS1 and MS2 spectra based on a provided protein database and the raw MS files, using the retention time, m/z and abundance (or intensity) of each peptide. Subsequently, protein groups are assembled, encompassing all proteins that share the detected peptides and cannot be differentiated from one another (typically these are protein isoforms). The quantities for these protein groups are calculated so that researchers can use this information for downstream bioinformatic data analysis. Figure adapted from Hein *et al.*, 2013.

2.2 To differentiate or not to differentiate: Hallmarks of pluripotency

2.2.1 Origins of pluripotency

Cell fate decisions follow a hierarchical structure in which differentiation capacity diminishes from the totipotent zygote to unipotent stem cells (De Los Angeles *et al.*, 2015). The zygote forms post-fertilization and after fusion of paternal and maternal gametes via a process known as the maternal-to-zygotic transition (MZT). This process necessitates gametic chromatin decompaction, global DNA demethylation, genomic reorganization and chromatin remodeling (Ladstätter and Tachibana, 2019). At this stage, the zygote is transcriptionally silent, depending on mRNAs and proteins derived from the maternal gamete (Eckersley-Maslin, Alda-Catalinas and Reik, 2018). In mice, embryonic transcription begins with two transcription bursts post-zygotic genome activation (ZGA) following the first cell division (Jukam, Shariati and Skotheim, 2017; Zhang *et al.*, 2022). Subsequent cell divisions lead to embryo compaction and polarization at the 8-cell stage, followed by the morula formation at the 16-cell stage. During this preimplantation stage, inner cavitation and specification of the inner cell mass (ICM) to the primitive endoderm and pluripotent stem cells (PSCs) and outer cells to the trophectoderm occurs (Fiorentino, Torres-Padilla and Scialdone, 2020). PSCs form the epiblast within the ICM, from which point the fate of individual cells is determined based on their position within the embryo (Hillman, Sherman and Graham, 1972). PSCs, characterized by tightly regulated and conserved epigenomes and transcriptomes among mammals, can differentiate *in vivo* into every cell type except for the trophectoderm and extraembryonic cell types (Boroviak *et al.*, 2015; Smith, 2017; Takahashi, Kobayashi and Hiratani, 2018; Kinoshita *et al.*, 2021). The timing of PSC emergence and the duration of the pluripotent phase vary among mammals. In mice, PSCs emerge at embryonic day (E) 3.5 and the pluripotent phase ends at E6.5, while in humans, PSCs emerge at E5 and the pluripotent phase ends at E16. Since their first discovery in 1981 in mice, PSCs were established *in vitro* and found to self-renew, meaning that they can divide indefinitely (Evans and Kaufman, 1981; Martin, 1981). *In vitro*, PSCs can differentiate into all three germ layers upon exposure to various combinations of growth factors and other signaling molecules. Furthermore, *in vitro* cultured PSCs can produce germ cells in chimeras, which can then develop into fertile adults (Solter, 2006), a hallmark of pluripotency. Additional *in vitro* experiments have shown that PSCs can be "directed" to differentiate into trophoblasts through epigenetic modeling (Zijlmans *et al.*, 2022) or extraembryonic mesoderm by the addition of specific signaling molecules (Pham *et al.*, 2022).

Around three decades after the discovery of PSCs, it became widely recognized that the ICM is not their sole origin and that PSCs can be reprogrammed from other cell types. Before this finding, it was demonstrated that cell identities could be converted by nuclear transfer (Briggs and King, 1952; Gurdon, Elsdale and Fischberg, 1958) or the ectopic expression of certain master regulatory transcription factors (Davis, Weintraub and Lassar, 1987). Subsequent research revealed that nuclear transfer of a somatic cell into an oocyte (Wilmut *et al.*, 1997) or their hybridization with ES cells allows for reprogramming of somatic cell types to totipotent zygotes or PSCs (Tada *et al.*, 2001). This suggested that oocytes and PSCs contain components sufficient for reprogramming somatic cells. Around the same time, researchers identified transcription factors necessary for maintaining pluripotency, including OCT4, SOX2, and NANOG (Nichols *et al.*, 1998; Avilion *et al.*, 2003; Chambers *et al.*, 2003). Building on this foundational work, Kazutoshi Takahashi and Shinya Yamanaka screened for reprogramming factors through the ectopic expression of 24 candidate genes, finding that four of these were sufficient to reprogram mouse or human fibroblasts into induced pluripotent stem (iPS) cells (Takahashi and Yamanaka, 2006; Takahashi *et al.*, 2007). These factors, named "Yamanaka factors" or OSKM (OCT4, SOX2, KLF4, and c-MYC), function by modifying the epigenetic landscape, reorganizing the chromatin architecture, activating pluripotency genes, and silencing lineage-specific genes. The generated iPSCs exhibited the same morphology, transcriptome, cell cycle, and teratoma formation capabilities as PSCs. They were subsequently shown to generate adult chimeras with germline competence (Maherali *et al.*, 2007; Okita, Ichisaka and Yamanaka, 2007). *In vivo*, natural reprogramming can occur in somatic tissues under certain circumstances, such as skin renewal in the adult organism (Mosteiro *et al.*, 2016; Yadav, Quivy and Almouzni, 2018). Although all somatic tissues lack PSCs unless actively reprogrammed, there are two additional methods to obtain PSCs from (primordial) germ cells. First, primordial germ cells can generate PSCs by supplementing media with leukemia inhibitory factor (LIF) and basic fibroblast growth factor (bFGF, also known as FGF2) (Shamblott *et al.*, 1998). Second, unfertilized oocytes can be artificially activated to differentiate into parthenogenetic ES cells (Robertson, Evans and Kaufman, 1983). These cells possess the full differentiation potential of PSCs and can be diploid (Kim *et al.*, 2007) or haploid (Sagi *et al.*, 2016). Thus, by directed reprogramming, pluripotent cells can be derived from all stages of development and adulthood, which is indicative of a highly conserved transcriptional program regulated by the Yamanaka factors. Intriguingly, the very factors that are essential for establishing and maintaining pluripotency also play a significant role in guiding lineage selection from gastrulation onwards (Thomson *et al.*, 2011).

2.2.2 Phased progression of pluripotency

The emergence of PSCs aligns *in vivo* with the implantation of the blastocyst into the uterine wall where cells of the ICM pre- and post-implantation are considered as pluripotent (Figure 6). Consequently, pluripotency was suggested to exist in at least two stages (Nichols and Smith, 2009). However, it was quickly recognized that pluripotency must be a continuum with three primary stages: ground (also referred to as naive), followed by the post-implantation phases formative and primed as well as several transient states in-between (Nichols and Smith, 2009; Bedzhov and Zernicka-Goetz, 2014; Hackett and Surani, 2014; De Los Angeles *et al.*, 2015; Morgani, Nichols and Hadjantonakis, 2017; Rossant and Tam, 2017; Shahbazi *et al.*, 2017; Smith, 2017; Neagu *et al.*, 2020). Mouse ground PSCs (mouse embryonic stem cells, mESCs) are maintained in culture using LIF (Smith *et al.*, 1988; Williams *et al.*, 1988) along with inhibitors of MAPK-ERK (PD0325901) and GSK3β (CHIR99021) in a serum-free culture medium (Silva *et al.*, 2008; Ying *et al.*, 2008; Marks *et al.*, 2012). On the other hand, formative PSCs (epiblast-like cells, EpiLCs) are either transiently differentiated from mESCs by MAPK-ERK (FGF2) and SMAD (ActivinA) activation for two days or, alternatively, established as a stable cell line (formative stem cells, FSCs) by, for instance, SMAD activation along WNT (XAV939) and retinoic acid (BMS493) inhibition for several passages (Kinoshita *et al.*, 2021; Wang *et al.*, 2021). Primed PSCs (epiblast stem cells, EpiSCs) are differentiated from mESCs as are transient formative PSCs by MAPK-ERK and SMAD activation for at least seven days (Brons *et al.*, 2007; Tesar *et al.*, 2007; Hayashi *et al.*, 2011).

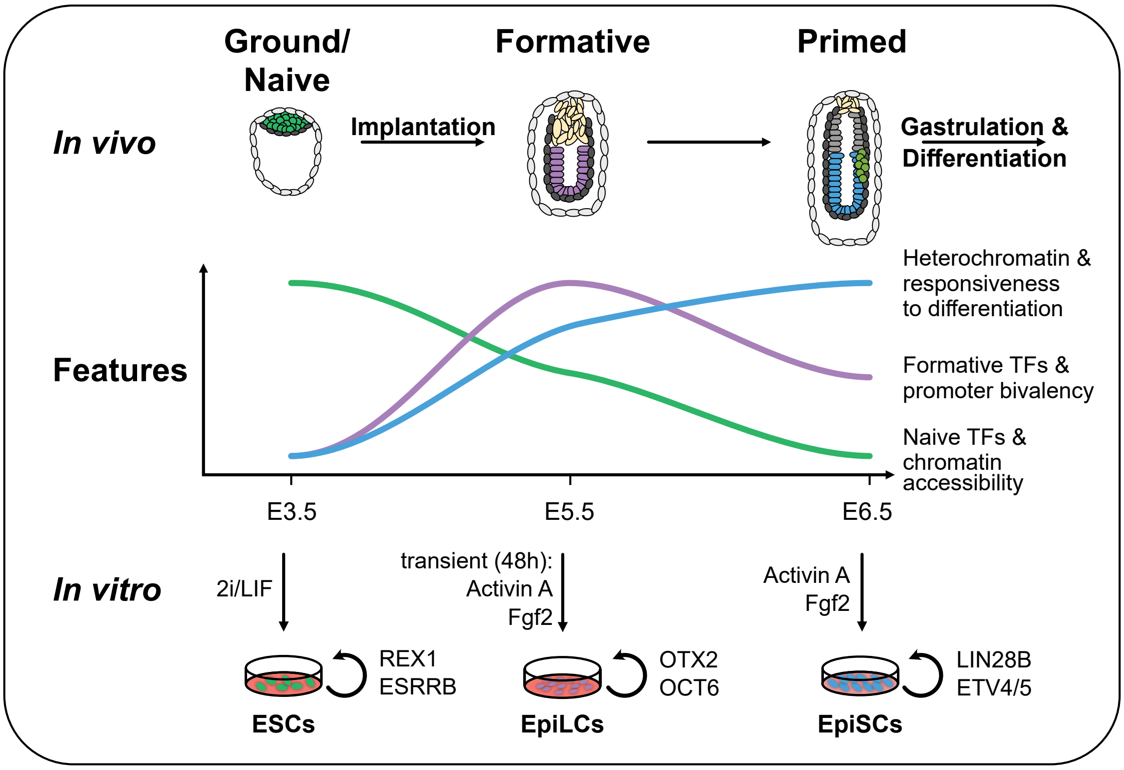


Figure 6: The phased progression of pluripotency. Pluripotent stem cells populate *in vivo* the inner cell mass of the blastocyst and can differentiate into nearly any cell type within an organism. Pluripotency is not a static state; instead, it spans a continuum comprising three primary phases: ground/naive, formative, and primed. These phases are crucial in preparing the cells for differentiation, while maintaining their stem cell properties, and each harbors distinct characteristics. Ground state PSCs are characterized by highly accessible chromatin, which is preserved by ground-specific transcription factors. Upon implantation into the uterus, formative PSCs establish bivalency at developmental genes, thus possessing a more repressive epigenome. Finally, before gastrulation, primed PSCs reach a nadir in chromatin accessibility and are readily responsive for differentiation. These three phases can be recapitulated *in vitro* by specific cell culture conditions which activate the same master transcription factors as *in vivo*.

But what is the relevance of these different phases? Pluripotency is often misunderstood as the immediate capability of PSCs to differentiate into virtually any cell type in an organism, while, in fact, the phased progression of pluripotency is essential for differentiation (Nichols and Smith, 2009). The following analogy should illustrate the current concept of pluripotency: The phased progression of pluripotency begins in an undifferentiated state like childhood, where the potential professions and required skills for a later profession are unknown. However, through the gradual acquisition of fundamental knowledge and capabilities, an individual becomes equipped to specialize in a particular career. In a parallel manner, naive PSCs can differentiate into any cell type, but lack the specific attributes, environment, and signaling pathways necessary for each lineage. As differentiation proceeds, the cell acquires the required characteristics and gradually restricts its developmental potential as it moves toward terminal differentiation. Bypassing the formative or primed phase therefore hinders the correct lineage specification and causes developmental anomalies. In the mentioned analogy, this would correspond to skipping essential parts of professional education and yet attempting to execute a specialized profession. To understand the characteristics of each pluripotency phase the following chapters will give an overview of transcriptomic and epigenetic hallmarks of the pluripotency transitions.

2.2.3 Transcriptional hallmarks of pluripotency phase transitions

The three primary stages of pluripotency all share the Yamanaka factors as a core circuit of pioneering transcription factors that establish and maintain pluripotency (Boiani and Schöler, 2005; Boroviak *et al.*, 2015; Kinoshita *et al.*, 2021). Pioneering transcription factors have the propensity to recognize and bind DNA binding motifs on nucleosomes, generally inaccessible to other transcription factors (Zaret, 2020). By recruiting coactivators or corepressors to the target locus, pioneering transcription factors alter the nucleosome arrangement and chromatin accessibility. SOX2, for example, partially unwraps DNA from nucleosomes and disrupts internucleosome interactions (Dodonova *et al.*, 2020), whereas OCT4 requires only parts of its DNA binding motif to bind a target site but then relies on chromatin remodelers such as BRG1 to rearrange nucleosomes (King and Klose, 2017; Huertas *et al.*, 2020). OCT4 also enhances the pioneering activity of SOX2 by retaining it on nucleosomes for longer periods (Li *et al.*, 2019). Both SOX2 and OCT4 can induce passive DNA demethylation, altering the epigenetic modifications at target regions (Vanzan *et al.*, 2021). The Yamanaka factors jointly occupy gene enhancers within a shared protein-protein interactome, facilitating long-range enhancer-promoter interactions in concert with CTCF, cohesin, and condensin (Zinzen *et al.*, 2009; Mullen *et al.*, 2011; Trompouki *et al.*, 2011; Huang and Wang, 2014; Emani *et al.*, 2015; Rafiee *et al.*, 2016; Schlesinger and Meshorer, 2019; Han *et al.*, 2022). OCT4, SOX2, and NANOG form clusters of enhancers (also known as super-enhancers) that regulate genes critical to pluripotency (Whyte *et al.*, 2013). Moreover, PSCs depend on approximately 1750 essential genes (essentialome) for pluripotency maintenance and self-renewal, with 80% of these genes shared with at least one cancer cell line (Yilmaz *et al.*, 2018). Ground state PSCs, which correspond to the early pre-implantation epiblast at E3.5-4.5, form homogenous and round colonies in cell culture. These cells require maturation to commit to lineage decisions due to their permissive chromatin, characterized by sparse epigenetic marks and high plasticity (Gaspar-Maia *et al.*, 2011; Marks *et al.*, 2012; Melcer *et al.*, 2012; Boroviak *et al.*, 2014, 2015; Lee, Hore and Reik, 2014; Zylicz *et al.*, 2015; Eckersley-Maslin, Alda-Catalinas and Reik, 2018). Key transcription factors, including ESRRB, REX1, KLF4, KLF2, TFCP2L1, TBX3, and PRDM14, collaborate with OCT4, SOX2, and NANOG to sustain the naive state. Notably, ESRRB binds to the Nanog promoter, positively regulating its expression in cooperation with OCT4, and likewise modulates Oct4 expression in a NANOG-dependent manner (van den Berg *et al.*, 2008; Zhang *et al.*, 2008).

The transition from the naive to formative state at E4.5-5.5 is driven by the MAPK-ERK and GSK3β signaling pathways, with MAPK-ERK signaling being active within the first six hours of exit from the naive state and becoming dispensable afterwards (Burdon *et al.*, 1999; Kunath *et al.*, 2007; Ying *et al.*, 2008; Wray *et al.*, 2011; Yang *et al.*, 2019). This transition is marked by changes in morphology, colony formation, and a metabolic shift from oxidative phosphorylation and glycolysis in the naive state to mainly glycolysis in the formative and later primed states (Zhou *et al.*, 2012; Kalkan *et al.*, 2017; Tsogtbaatar *et al.*, 2020; Dierolf *et al.*, 2022). In formative PSCs, OCT4 interacts with the formative-specific transcription factor OTX2 and relocates to developmental gene enhancers regulating the late epiblast (Buecker *et al.*, 2014; Kinoshita *et al.*, 2021; Wang *et al.*, 2021). Furthermore, the core pluripotency network collaborates with key transcription factors that govern formative pluripotency: OCT6, OTX2, SOX3, SALL2, and ZIC2 (Buecker *et al.*, 2014; Dunn *et al.*, 2014; Boroviak *et al.*, 2015; Kurimoto *et al.*, 2015; Shirane *et al.*, 2016; Weinberger *et al.*, 2016; Kalkan *et al.*, 2017).

Primed PSCs emerge between E5.5-6.5 and are partially fate-determined based on their position within the epiblast (Lawson, Meneses and Pedersen, 1991; Tam and Zhou, 1996), leading to substantial heterogeneity (Kojima *et al.*, 2014; Tsakiridis *et al.*, 2014). Although primed PSCs share a core pluripotency network with naive and formative PSCs, including OCT4, SOX2, and NANOG (Nichols and Smith, 2009; Hackett and Surani, 2014; Kalkan *et al.*, 2017; Smith, 2017), they exhibit differential enhancer utilization: for instance, in the case of Oct4, the proximal instead of distal enhancer is active (Chen *et al.*, 2008; Factor *et al.*, 2014). The deletion of the Oct4 proximal enhancer, in combination with

LIF supplementation, but not LIF supplementation alone, has been demonstrated to be sufficient for reprogramming primed PSCs back to the ground state (Bao *et al.*, 2009). In line with the shared pluripotency core, primed PSCs are unable to yield chimeras upon injection into pre-implantation epiblasts but can contribute to chimerism when injected into post-implantation epiblasts (Huang *et al.*, 2012; Masaki *et al.*, 2016; Weinberger *et al.*, 2016). Consequently, primed PSCs are considered pluripotent and undifferentiated despite the upregulation of differentiation markers such as NES or SALL3 (Buecker *et al.*, 2014).

2.2.4 Epigenetic hallmarks of pluripotency phase transitions

Epigenetic modifications play a pivotal role in PSCs transitioning from naive to primed states, preparing the cells for differentiation from gastrulation onwards. The previously discussed transformation in transcriptional networks coincides with a rigorously regulated reorganization of the epigenome, including DNA methylation and histone modifications (Takahashi, Kobayashi and Hiratani, 2018). Cytosine DNA methylation (5mC), a reversible epigenetic mark perpetuated by DNA methyltransferase 1 (DNMT1) throughout the cell cycle (Leonhardt *et al.*, 1992), is essential for comprehensive embryonic development, genome stability, transcriptional repression of developmental genes, transposable elements (TE), and X chromosome inactivation (Walsh, Chaillet and Bestor, 1998; Okano *et al.*, 1999; Rowe and Trono, 2011; Roulois *et al.*, 2015; Schübeler, 2015). While DNA methylation typically occurs in symmetrical CpG dinucleotides, it is also observable in promoter and enhancer regions, as well as gene bodies (Doskocil and Sorm, 1962; Reik, Dean and Walter, 2001; Zemach *et al.*, 2010). It serves dual interconnected roles: it primarily suppresses gene expression and prevents genome instability (Schübeler, 2015). However, DNA methylation at cytosine comes at the expense of spontaneous deamination which results in a C→T transition. In pluripotent stem cells, CpG methylation silences lineage-specific genes, facilitating stable expression of pluripotency-associated genes. Moreover, it represses repetitive DNA elements, such as retrotransposons, safeguarding genome stability (Greenberg and Bourc'his, 2019; Dahlet *et al.*, 2020; Petryk *et al.*, 2021). Global methylation levels and methylated regions differ across cell types and within cell populations and influence cell identity. For instance, methylation can fluctuate within the same pluripotent state, guiding early cell fate decisions (Bogdanović and Lister, 2017; Rulands *et al.*, 2018). Aberrant maintenance of DNA methylation frequently leads to diseases like cancer or neurological disorders (Petryk *et al.*, 2021).

Intriguingly, the ground state epiblast exhibits global DNA hypomethylation, despite the protective functions of this mark (Monk, Boubelik and Lehnert, 1987; Sanford *et al.*, 1987; Howlett and Reik, 1991; Hayashi *et al.*, 2008; Ficz *et al.*, 2013; Hackett *et al.*, 2013; Lee, Hore and Reik, 2014; Messerschmidt, Knowles and Solter, 2014). The low 5mC level in the ground state is primarily preserved by the impairment of maintenance DNA methylation and active demethylation initiated by the Ten-eleven Translocation (TET) family of dioxygenases TET1 and TET2 (Tahiliani *et al.*, 2009; von Meyenn *et al.*, 2016; Mulholland, Traube, *et al.*, 2020). Like TET1 and TET2, TET3 can also oxidize 5mC to its demethylation intermediates 5-hydroxymethylcytosine (5hmC), 5-formylcytosine (5fC), and 5-carboxylcytosine (5caC), but TET3 expression is notably low during pluripotency (Mulholland, Traube, *et al.*, 2020). Each demethylation intermediate can lead to loss of 5mC during DNA replication which is known as “passive” DNA demethylation. In “active” DNA demethylation, 5fC and 5caC are excised by the thymine DNA glycosylase (TDG) and subsequently replaced by cytosine through the base excision repair pathway (Cortellino *et al.*, 2011; He *et al.*, 2011; Greenberg and Bourc’his, 2019). In the ground state epiblast, hypomethylation is further facilitated by the low expression of the *de novo* DNA methyltransferases Dnmt3a and Dnmt3b (Carlson, Page and Bestor, 1992; Leitch *et al.*, 2013; Buecker *et al.*, 2014; Guo *et al.*, 2014). The global increase of DNA methylation in the formative phase strongly aligns with the upregulation of Uhrf1 along with Dnmt3a, Dnmt3b and their co-factor Dnmt3l (Okano *et al.*, 1999; Seisenberger *et al.*, 2012; Auclair *et al.*, 2014; von Meyenn *et al.*, 2016; Yang *et al.*, 2019; Wang *et al.*, 2021). This is followed by a raise in the DNA demethylation intermediates, which do not correspond to the overall lower level of TET2 and the stable level of TET1, suggesting an increased enzymatic activity (Ito *et al.*, 2011; Mulholland, Traube, *et al.*, 2020). DNA methylation levels between the formative and primed states are strikingly similar, even though DNMT3A/B/L levels diminish (Yang *et al.*, 2019; Wang *et al.*, 2021). The tight control of DNA hypomethylation during preimplantation development and subsequent global DNA hypermethylation is particularly conserved in placental mammals, but its exact function remains unknown. For instance, DNA methylation does not significantly affect the binding of pluripotency transcription factors (Dean *et al.*, 2001; Smith *et al.*, 2012; Ivanova *et al.*, 2020). However, there is increasing evidence that hypomethylation regulates the expression of pluripotency-related factors, including OCT4 and NANOG (Gao *et al.*, 2013; Olariu, Lövkvist and Sneppen, 2016; Shanak and Helms, 2020).

In addition to DNA methylation, pluripotency is dynamically modulated by histone modifications. One such modification is histone H3 lysine 9 trimethylation (H3K9me3), a prominent epigenetic mark associated with heterochromatin. This mark plays a crucial role in both differentiation and somatic reprogramming (Stancheva, 2005; Matoba *et al.*, 2014; Nicetto *et al.*, 2019). Out of the four known H3K9 trimethyltransferases (SUV39H1, SUV39H2, SETDB1, and SETDB2), only the knockout (KO) of SETDB1 results in lethality around implantation due to impaired ICM development (Dodge *et al.*, 2004), highlighting its importance in early embryogenesis. SETDB1 interacts with OCT4 to maintain pluripotency by preventing differentiation into the trophectoderm (Lohmann *et al.*, 2010). Moreover, the loss of SETDB1 induces a transcriptional program that resembles the totipotent two-cell (2C) stage by activation of the 2C stage master transcription factor Dux (Wu *et al.*, 2020). SETDB2 is present at low levels during pluripotency and remains poorly characterized in this context. The remaining two H3K9me3-specific methyltransferases SUV39H1 and SUV39H2 partner with CBX1/3/5 (also known as HP1β/γ/α) to promote heterochromatinization by depositing H3K9me3 in telomeric repeats, centromeric minor satellite repeats, and pericentromeric major satellite repeats, often in conjunction with DNA methylation (O'Carroll, Scherthan, *et al.*, 2001; Martens *et al.*, 2005; Bulut-Karslioglu *et al.*, 2014; Kinoshita *et al.*, 2021). During interphase, H3K9me3-containing repetitive regions within pericentromeres form clusters known as chromocenters, giving rise to constitutive heterochromatin (Peters *et al.*, 2003; Déjardin, 2015). However, in naive PSCs pericentromeres are characterized by high levels of H3K27me3, resulting in a moderate expression of pericentromeric satellites, while primed PSCs display replacement of H3K27me3 by DNA methylation and H3K9me3, leading to a lower expression of pericentromeric satellites (Tosolini *et al.*, 2018). Moreover, SUV39H1 and SUV39H2 are required for silencing most long interspersed nuclear elements (LINEs) and endogenous retroviruses (ERVs) in pluripotency. In committed cells, this silencing is later replaced by DNA methylation-driven mechanisms (Bulut-Karslioglu *et al.*, 2014).

H3K27me3, an alternative repressive histone PTM, is typically deposited in CpG-rich promoters by the Polycomb repressive complex 2 (PRC2). A loss of PRC2 subunits or its collaborating complex, PRC1, results in embryonic lethality from gastrulation onwards (Faust *et al.*, 1998; O'Carroll, Erhardt, *et al.*, 2001; Pasini *et al.*, 2004). Naive PSCs both *in vivo* and *in vitro* exhibit high levels of H3K27me3, coinciding with DNA hypomethylation. This correlation has been demonstrated to be crucial for early mammalian development due to the mutually exclusive nature of both epigenetic marks (Brinkman *et al.*, 2012; Saksouk *et al.*, 2014; Liu *et al.*, 2016). In fact, global increase of H3K27me3 turned out to be a result of DNA hypomethylation in PSCs (Reddington *et al.*, 2013; van Mierlo *et al.*, 2019) and might serve as a gatekeeper against early priming of PSCs (Zheng *et al.*, 2016; Santos-Barriopedro, van Mierlo and Vermeulen, 2021) and trophoblast differentiation (Zijlmans *et al.*, 2022). Moreover, at developmental gene promoters, H3K27me3 frequently coincides with H3K4me3, an activating mark deposited by the MLL family of methyltransferases (also known as bivalent promoters). These bivalent promoters, often spatially clustered (Joshi *et al.*, 2015; Dunican *et al.*, 2020), can swiftly turn on or off, yet still allow low expression of the corresponding genes. Hence, bivalent promoters are poised for activation (Bernstein *et al.*, 2006). Naive PSCs display fewer bivalent promoters (Marks *et al.*, 2012; Liu *et al.*, 2016; Zheng *et al.*, 2016), but during the formative phase, H3K4me3/H3K27me3 peak at promoters of developmental genes, although overall H3K27me3 levels diminish and are replaced by DNA methylation (Marks *et al.*, 2012; van Mierlo *et al.*, 2019; Wang *et al.*, 2021). The global DNA hypermethylation is excluded from bivalent promoters by active TET1-mediated DNA demethylation (Wu *et al.*, 2011; Neri *et al.*, 2013; Ross and Bogdanovic, 2019). Interestingly, bivalency is already more common in serum/LIF cultured PSCs, suggesting a potential role in the transition to formative pluripotency (Azuara et al. 2006; Bernstein et al. 2006). Loss of PRC2 in primed but not naive PSCs results in failures in pluripotency maintenance and spontaneous cell differentiation in primed PSCs (Moody *et al.*, 2017; Geng, Zhang and Jiang, 2019). Bivalency can vary depending on the distribution of H3K4me3 over the promoter and widespread distributions of H3K4me3 can be specifically found in 400 developmental gene promoters. Most of these robust bivalent promoters were associated with genes required for germ layer differentiation or morphological changes such as anterior/posterior pattern formation (Joshi *et al.*, 2015; Xiang *et al.*, 2020; Wang *et al.*, 2021), which provides additional evidence for the preparatory nature of the formative phase. Collectively, the definition of these transcriptional (Figure 6) and epigenetic (Figure 7) hallmarks of pluripotency has markedly advanced our understanding of the mechanisms that determine cell identity during early embryogenesis. It is important, however, to recognize that our current understanding of pluripotency largely rests on data derived from RNA or DNA-focused techniques, which do not provide a comprehensive picture of the chromatin composition – the actual control unit of cellular identity. Chromatin proteomics approaches that resolve the dynamic reorganization of the chromatin proteome could therefore aid in further deciphering early mammalian development and the reprogramming of iPSCs.

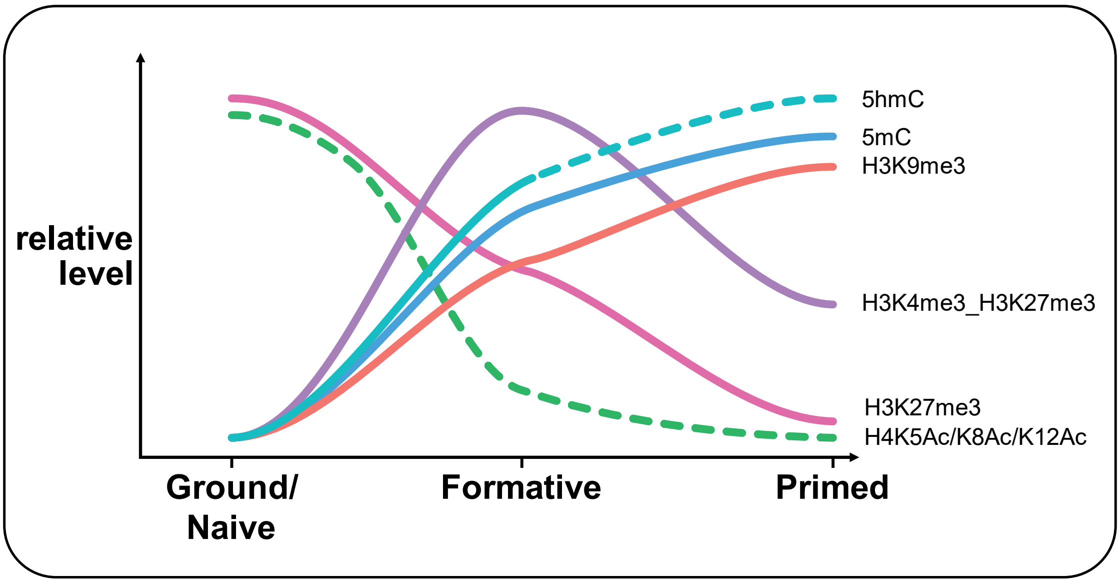


Figure 7: Important epigenetic hallmarks of pluripotency. While repressive epigenetic marks such as (hydroxy-)methylcytosine and H3K9me3 peak at the primed phase, activating marks like the acetylation of H4K5, H4K8 and H4K12 decrease. An exception is H3K27me3, a repressive epigenetic mark, which reaches a minimum on bulk level in the primed phase. However, bivalent sites with H3K4me3 and H3K27me3 peak at the formative phase. Dashed lines indicate that the relative level of the respective epigenetic mark is not fully resolved yet.

2.3 The Chromatome: The control unit of cellular identity

Cellular identity is defined by the phenotype, state and ontogeny of a cell (Morris, 2019). The phenotype includes “[...] all the manifold biological appearances, including chemical, structural and behavioral attributes, that we can observe about an organism but excludes its genetic constitution” (Churchill, 1974). Depending on environmental cues, a single cell type can exhibit several varied phenotypes, often referred to as cellular states (Morris, 2019). Information on the cellular lineage complements the phenotype and cellular state and helps to further distinguish similar cell types from each other. To give an example, macrophages exhibit a tissue-specific "macrophage" phenotype, encompass a spectrum of polarization states dictated by environmental inputs and originate from diverse lineages such as circulating monocytes or embryonic precursors (Gentek, Molawi and Sieweke, 2014; Boutilier and Elsawa, 2021; Ricketts *et al.*, 2021). Importantly, the term phenotype can be decomposed into quantifiable attributes such as morphology, spatial localization, transcriptome, epigenome, proteome and metabolome. These attributes are interconnected; for instance, the phosphorylation status of signaling proteins can drive downstream activation of chromatin-associated transcription factors, initiating a gene expression program that leads to a specific proteome. The proteome, which most directly reflects cellular function, shapes entities like the epigenetic landscape, thereby establishing and sustaining cell identity. This simplified description of cell identity regulation underscores its intricacy, requiring multi-omics approaches for its analysis (Ye and Sarkar, 2018).

Current omics methods cover a broad range of cellular phenotypes by providing comprehensive information on these cellular attributes (Figure 8): (i) genomics enables the identification of genetic variants which are implicated in diseases and influence treatment responses. With growing genomic information on a population scale, genetic markers can aid in predicting certain diseases and thus allow more personalized medical treatments (The 1000 Genomes Project Consortium, 2012). Genomics can also be harnessed to understand evolutionary relationships between different species. (ii) Epigenomics characterizes reversible and partially heritable DNA and histone modifications, binding sites of chromatin-associated proteins, chromatin accessibility and organization (Rottach, Leonhardt and Spada, 2009; Taudt, Colomé-Tatché and Johannes, 2016; Klemm, Shipony and Greenleaf, 2019; M. Nakamura *et al.*, 2021). These epigenomic layers encode diverse, context-dependent chromatin states such as silenced heterochromatic regions that reflect cellular identities and tissue-specific (disease) states (Baylin *et al.*, 2001). (iii) Transcriptomics identifies and quantifies various RNAs, including canonical and isoform messenger RNAs (mRNAs), long non-coding RNAs (lncRNAs) as well as short and circular RNAs (Hasin, Seldin and Lusis, 2017). Transcriptomic studies have revealed the genome's complexity, demonstrating that approximately 80% of the genome is transcribed and only 3% encodes proteins (The ENCODE Project Consortium, 2012). Single-cell mRNA sequencing applications are regularly utilized to characterize the cellular composition of organelles and tissues and can resolve the hierarchical differentiation path of a specific cell type (Treutlein *et al.*, 2014; Kanton *et al.*, 2019; Fleck *et al.*, 2022). (iv) Proteomics characterizes proteins, their isoforms, PTMs, protein-protein interactions, protein structure and protein localization. Proteomic methods provide information on protein or peptide abundance at the tissue, cell type, or single-cell level (Neagu *et al.*, 2022). (v) Metabolomics can resolve chemically heterogeneous small molecules which can be predictive of enzymatic activity; for instance, cellular methyltransferases have a single donor, S-adenosylmethionine (SAM). When SAM donates a methyl group, it is irreversibly converted to S-adenosylhomocysteine (SAH). Therefore, SAM turnover predicts global methyltransferase activity (Wooderchak, Zhou and Hevel, 2008; Wong, Qian and Yu, 2017; Gonzalez-Covarrubias, Martínez-Martínez and Del Bosque-Plata, 2022). Unlike the proteome, the metabolome lacks an analogous blueprint like mRNAs for the proteome. Therefore, multiple experiments are required to analyze the metabolome comprehensively (Sindelar and Patti, 2020). Moreover, metabolites have endogenous or exogenous origins, which further complicates investigations of the cellular metabolome (Lankadurai, Nagato and Simpson, 2013).

Information obtained from one omics approach can be further leveraged when data from several omics layers are integrated. These multi-omics analyses are powerful in identifying regulatory pathways that are not evident from any single omics technique alone. For example, analyzing the transcriptome and proteome of a differentiation model across a time course can reveal the initiation of transcriptional programs and corresponding dynamics in protein-level regulation that do not simply correlate with transcript-levels (Liu, Beyer and Aebersold, 2016). It can uncover how long a protein is maintained above a certain level while its corresponding gene is no longer transcribed. In addition, information can be obtained to complement transcriptome data, such as information on PTMs or protein-protein interactions of transcription factors that might modulate their activity. For instance, a recent multi-omics approach applied to study embryonic stem cell transitions has shown that phosphorylation signatures are rewired ahead of transcriptional and proteomic changes (Yang *et al.*, 2019). Lastly, these omics approaches can be refined by adding spatial information. When combined with artificial intelligence (AI)-based automated segmentation of cellular structures, these methodologies become highly efficient in revealing hidden patterns of cellular identity (Moffitt, Lundberg and Heyn, 2022).

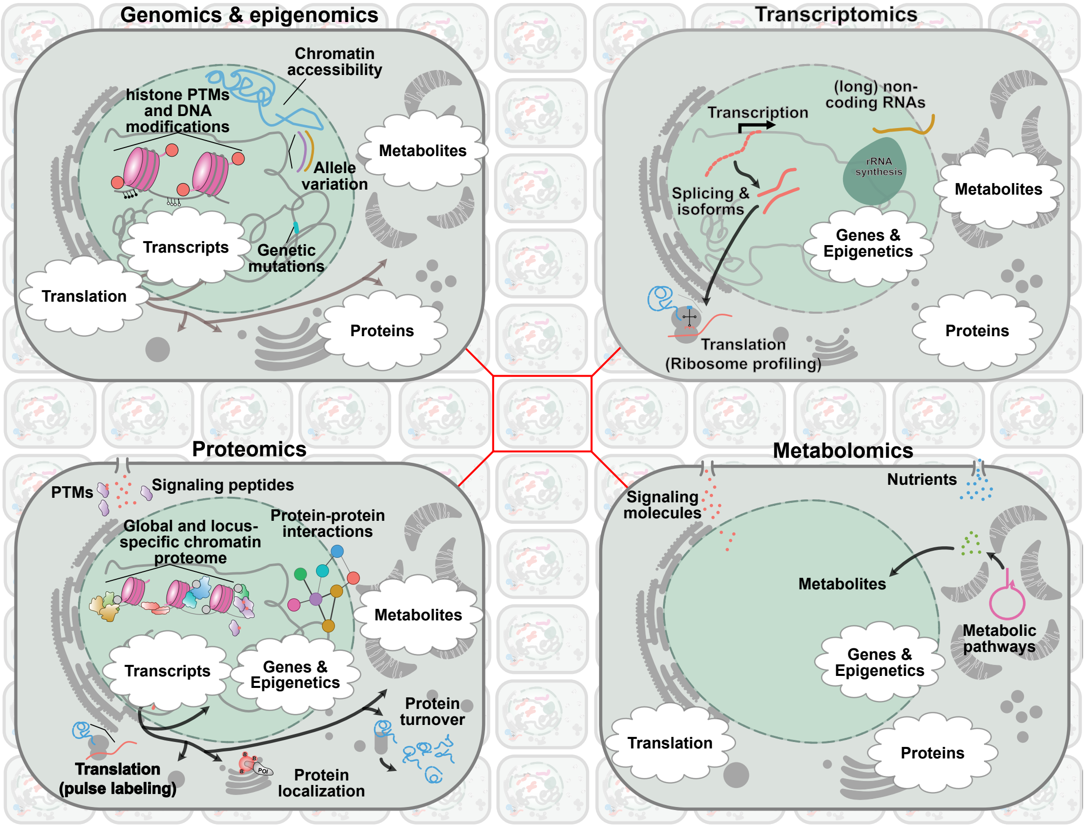


Figure 8: Schematic representation of omics techniques and the resolved cellular modalities. Genomics identifies allele variations and genetic mutations, aiding disease prediction, personalized treatment, and understanding evolutionary relationships. Epigenomics describes the heritable and reversible DNA and histone modifications which reflect cellular identities and disease states. Transcriptomics quantifies RNAs such as mRNAs or long non-coding RNAs and allows the detection of alternative splicing events. Through ribosome profiling, it also provides insight into translation. Proteomics quantifies proteins and their interactions, PTMs, turnover and localization. It further facilitates chromatin proteomics on a global or local scale and allows to investigate translational speed. Metabolomics quantifies small and chemically diverse molecules of both endogenous and exogenous origin, revealing intricate information despite its experimental complexity. Figure adapted from (Hein *et al.*, 2013).

Collectively, each omics approach provides unique insights into aspects of cellular regulation and function. Therefore, a comprehensive and integrated approach, combining multiple omics techniques, is required for understanding cellular function and identity. The analysis of the chromatin composition and underyling PTMs could further complement the existing omics approaches. The following chapters will highlight different areas of chromatin proteomics and focus on the strengths and weaknesses of current methods for investigating the global or targeted chromatin compositions.

2.3.1 Global chromatin proteomics

DNA- and chromatin-binding proteins orchestrate gene expression and thereby play an instrumental role in determining cellular identity. Transcription factors modulate regulatory regions such as promoters or enhancers and establish transcriptional programs by recruiting additional proteins that lead to either activation or repression of transcription. These genome-wide chromatin binding or dissociation events can be explored by analyzing the global chromatome. Global chromatin proteomics strategies offer unique advantages over methods like Chromatin Immunoprecipitation coupled to Mass Spectrometry (ChIP-MS) such as the quantification of numerous chromatin-associated proteins in an unbiased manner and within a single experiment. Several methods exist for studying global chromatin compositions. One straightforward approach involves studying the general protein abundance from whole-cell lysates via mass spectrometry. While this approach can potentially identify and quantify all proteins within a cell, it does not provide information on their spatial distribution, which is crucial to study the chromatome as protein localization can be altered in a cell type specific manner. Another challenge in using total protein measurements to study chromatin-associated proteins is that transcription factors can often be missed if not specifically enriched due to the limited sensitivity of mass spectrometry instrumentation, high complexity of cellular lysates, and the typically low abundance of transcription factors. As a result, prior studies have leveraged high-resolution mass spectrometry in combination with biochemical purification of (i) native (Shiio *et al.*, 2003; Torrente *et al.*, 2011; Räschle *et al.*, 2015; Kulej *et al.*, 2017; Federation *et al.*, 2020), (ii) formaldehyde (FA)-crosslinked (Kustatscher, Hégarat, *et al.*, 2014; Kustatscher, Wills, *et al.*, 2014; Ginno *et al.*, 2018; Aranda *et al.*, 2019) or (iii) biotinylated chromatin (Rafiee *et al.*, 2016; Aranda *et al.*, 2019) (Figure 9).

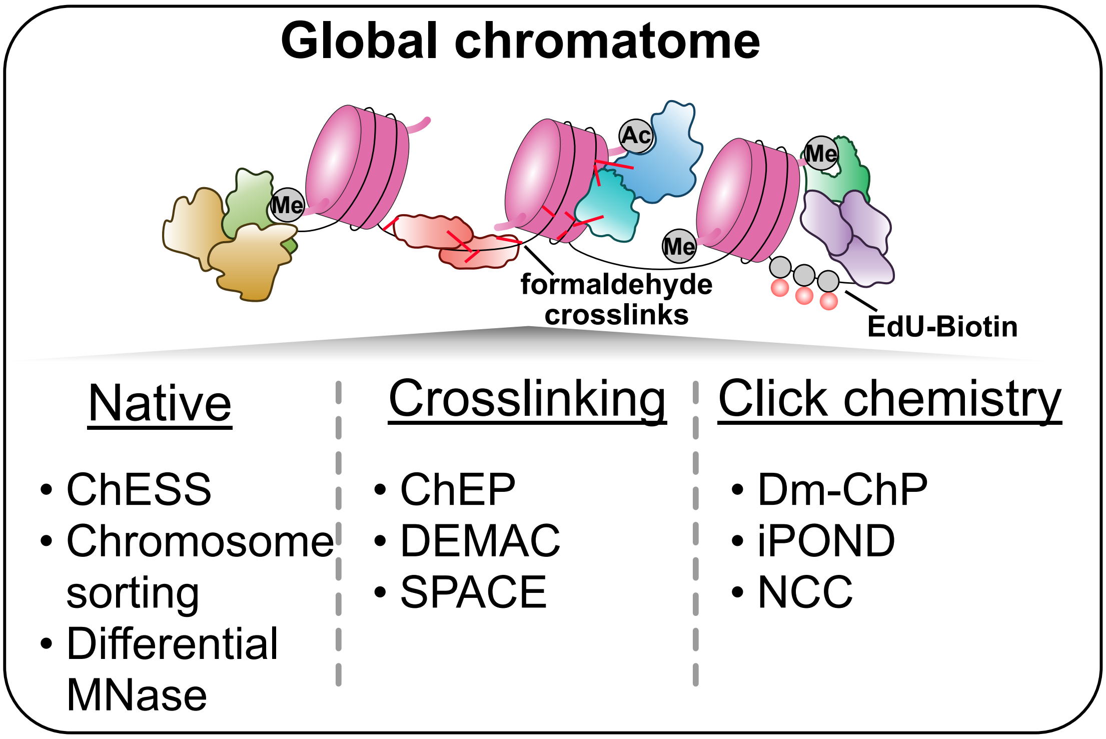


Figure 9: Overview of methods to study the global chromatin composition. The approaches for global chromatin proteomics research involve analyzing the biochemically purified chromatin composition within either native or formaldehyde-crosslinked chromatin. Alternatively, DNA can be labeled with biotin and subsequently subjected to streptavidin pulldown, with or without formaldehyde crosslinking.

Native global chromatin proteomics

Efficient chromatin purification strategies under native conditions involve isolating the nucleus, precipitating and fragmenting chromatin either enzymatically or mechanically, and then conducting an in-depth proteomic analysis of the chromatome (Sigismondo, Papageorgiou and Krijgsveld, 2022). A pioneering study employing native chromatin purification involved nuclear isolation followed by nuclear lysis and the pelleting of the insoluble chromatin fraction (Shiio *et al.*, 2003). This fraction was then resolubilized with detergents, and the resultant chromatome was subjected to proteomic analysis. Applied to human B lymphocytes, this study characterized in total 64 known nuclear proteins including the transcription factor MYC with a high background of 218 non-nuclear proteins. This study underscored the utility of cellular fractionation in identifying lower abundant nuclear proteins, thereby paving the way for future chromatin proteomics studies in different organisms such as *Saccharomyces cerevisiae* or different biological contexts such as mitosis

(Gassmann, Henzing and Earnshaw, 2005; Uchiyama *et al.*, 2005; Xie, Bandhakavi and Griffin, 2005; Chu *et al.*, 2006; Khoudoli *et al.*, 2008). Although these studies were mostly aimed at identifying as many chromatin-associated proteins as possible, they did not show much progress in the number of identified proteins. Still, these studies led to novel biological insights such as the identification of histone variants and their quantities in the rice plant *Oryza sativa* (Tan *et al.*, 2007). Another early chromatin proteomics study that went beyond merely cataloging the chromatome unraveled the chromatome reorganization upon UV laser-induced DNA lesions (Chou *et al.*, 2010). This study showed that subunits of PRC1, PRC2, and NuRD complexes are recruited to DNA damage sites in a PARP1- and PARP2-dependent manner. A more recent study has applied this kind of native chromatin purification on 2i/LIF and serum/LIF cultured mouse PSCs, identifying 1841 proteins in the chromatin fraction and revealing the enrichment of the LIF-downstream transcription factors KLF4 and TFCP2L1 in 2i/LIF cultured PSC chromatomes (van Mierlo, Wester and Marks, 2018).

Significant advancements in chromatin proteomics were made by a study that compared three distinct chromatin extraction methods: salt-based extraction, MNase digestion-based extraction, and partial MNase-based extraction into eu- and heterochromatin (Henikoff *et al.*, 2009; Torrente *et al.*, 2011). Each method individually identified approximately 1,000 potentially chromatin-associated proteins which resulted in a total of 1,900 unique proteins, a significant achievement given the technological status in the early 2010s. Interestingly, only 40% of the identified proteins were annotated as nuclear, suggesting a high background of non-nuclear proteins. A later study on rat liver extracts involving differential incubation of isolated nuclei with MNase or Deoxyribonuclease I (DNaseI) over a time course of 65 min with 5 min intervals was able to identify 694 proteins, including many novel chromatin-associated proteins (Dutta *et al.*, 2012). The study found that more euchromatic proteins were released after shorter periods of MNase/DNaseI digestion compared to heterochromatic proteins, a finding which can be attributed to the preference of MNase for nucleosome-free regions (Axel, 1975; Bloom and Anderson, 1978). A subsequent application of this method on PSCs and early differentiation models, led to the discovery of SMARCD1 as a novel regulator of ectodermal differentiation (Alajem *et al.*, 2015). Building on these insights, a subsequent study utilized differential MNase digestion to study changes to the chromatome mediated by Herpes Simplex Virus Type 1 (Kulej *et al.*, 2017). By analyzing histone PTM levels, the study was able to correlate histone PTM changes with respective epigenetic writers and erasers of histone PTMs. A follow-up study applied data-independent acquisition during mass spectrometry to this method, identifying 1,797 proteins (Federation *et al.*, 2020). Notably, this study demonstrated for the first time that global chromatin proteomics methods can discern subtle changes to the chromatome such as those caused by targeted degradation of histone modifiers, a finding with significant implications for drug discovery applications. Taking it a step further, Oliviero and colleagues employed hypotonic swelling and sucrose cushion centrifugation to isolate nuclei, which was then followed by high salt extraction and benzonase treatment (Oliviero *et al.*, 2022). By characterizing chromatin-associated proteins across various mouse organs, they observed age-related changes in the chromatome across thousands of proteins. Their findings underscored the significant influence of biological context on the dynamic composition of chromatin. While native chromatin purification holds theoretical potential to represent *in vivo* chromatin composition most accurately, the high heterogeneity of chromatin makes the comprehensive purification of chromatin-associated proteins a challenging task. Additionally, these methods have been found to yield high backgrounds of non-chromatin proteins.

Crosslinking-based global chromatin proteomics

The next generation in chromatin isolation protocols aimed to increase the purity of samples by implementing formaldehyde crosslinking, which primarily creates covalent bonds between DNA and lysines of proximal proteins (Hoffman *et al.*, 2015). This is followed by a selective purification of the crosslinked material. A popular crosslinking-based method is Chromatin Enrichment for Proteomics (ChEP) that enables chromatin isolation under denaturing conditions and upon centrifugation (Kustatscher, Hégarat, *et al.*, 2014; Kustatscher, Wills, *et al.*, 2014). ChEP further implemented an RNA digestion step via Ribonuclease A (RNaseA) prior to nuclear lysis and chromatin shearing in a glycerol-containing buffer. The first application of ChEP aimed to predict the chromatin association probability of proteins by implementing this method on four cell lines – HeLa, HepG2, MCF-7, and U2OS – and under different conditions such as ionizing radiation as well as α-Amanitin, DMSO, or doxycycline treatment (Kustatscher, Hégarat, *et al.*, 2014). A single ChEP application identified roughly 2,000 proteins, while all 63 ChEP conditions in total resulted in the identification of 7,635 human proteins, at least in one condition. Subsequent classification was performed using a random forest machine learning algorithm, which calculated an interphase chromatin probability, providing an estimate of a protein's likelihood of being chromatin-associated. This study set new standards for cataloguing chromatin-associated proteins, employing numerous test conditions, cell lines, machine learning algorithms, and subsequent estimation of chromatin association. It allowed for the creation of a quantitative protein ontology term, “interphase chromatin”. Moreover, this methodology was employed to analyze chromatin changes influenced by Cdk1- and Cdk2-mediated phosphorylation activities during the S-phase, which facilitated the identification of novel cell cycle-sensitive chromatin binders, likely involved in S-phase regulation.

The versatility of ChEP has been demonstrated in various biological systems. For instance, Samejima and colleagues used it to examine early events during the transition from interphase to mitosis (Samejima *et al.*, 2022). They discovered that initial prophase changes mainly occur at nuclear pores, on the nuclear envelope's inner surface, and within the nucleolus. Interestingly, most interphase chromatin proteins remain associated with chromatin until nuclear envelope breakdown (NEBD), after which their levels sharply decrease and cytoplasmic proteins accumulate on chromatin. This study offers critical insights into the successive waves of chromatin proteome remodeling that occur during nuclear disassembly and mitotic chromosome formation. ChEP's effectiveness has also been demonstrated in the study of the developmental plasticity of naive human pluripotent stem cells (hPSCs) (Zijlmans *et al.*, 2022). Using an integrated multi-omics approach, Zijlmans and colleagues discovered that PRC2 activity opposes trophoblast induction in naive hPSCs and blastoids, revealing that naive pluripotent cells are not epigenetically unrestricted, but rather face chromatin barriers that limit their differentiation potential. This work has substantial implications for understanding infertility and developmental disorders. Moreover, ChEP has been successfully used to study the human malaria parasite *Plasmodium falciparum,* which exhibits a highly organized chromatin structure underlying tight epigenetic regulation (Batugedara *et al.*, 2020). By analyzing 12 diverse eukaryotic genomes through comparative genomics, parasite-specific chromatin-associated domains were discovered. Subsequently, ChEP was employed to align the chromatin reorganization with chromatome changes, leading to the identification of parasite specific proteins that mediate chromatin reorganization.

An alternative strategy for crosslinked chromatome isolation is DEMAC, which uses the differential density of protein-DNA crosslinks in a caesium chloride gradient (Ginno *et al.*, 2018). While this method has enhanced our understanding of transcription factor retention at mitotic chromatin, it requires a labor-intensive process of 3 days of ultracentrifugation and sample fractionation, which still results in a relatively low yield of around 3000 proteins with low reproducibility between replicates and high background of non-nuclear proteins.

Click-chemistry-based global chromatin proteomics

Besides these methods, click chemistry-based techniques such as Isolation of Proteins on Nascent DNA (iPOND) and DNA-mediated Chromatin Pull-down (Dm-ChP) have been effectively employed for chromatin enrichment (Kliszczak *et al.*, 2011; Sirbu *et al.*, 2011). These methods incorporate the thymidine analog EdU (5-Ethynyl-2'-deoxyuridine) into new DNA strands, followed by crosslinking to azide biotin and enrichment via streptavidin. Chromatin-associated proteins can then be digested on the beads and identified through MS. iPOND, in particular, has been used to investigate protein dynamics at replication forks across multiple cell types (Lopez-Contreras *et al.*, 2013) and to study the mechanisms and resistance to anticancer drugs that interfere with DNA replication (Ribeyre *et al.*, 2016; K. Nakamura *et al.*, 2021).

In summary, global chromatin proteomics methods provided numerous insights into biological processes but suffer from a lack of chromatome coverage and data variability. This is primarily due to stochastic MS acquisition strategies and difficulties in sample preparation. Furthermore, these methods do not offer information on direct protein-histone interactions, such as through additional protein-protein crosslinking, which would permit the study of the chromatin landscape with actual evidence for direct interactions within native complexes. Therefore, the focus of the present work was to combine a robust and reproducible chromatin purification protocol that permits protein-protein crosslinking with a highly sensitive MS acquisition strategy. This integrated approach aims to analyze the chromatomes of naive, formative, and primed PSCs in a more holistic manner.

2.3.2 Interactome-resolving chromatin proteomics

Global chromatin proteomics provides valuable insights into the general association or dissociation of transcription factors, epigenetic regulators and other chromatin organizing proteins. However, these methods fail to reveal the precise chromatin environment surrounding a particular chromatin-associated protein. To address this issue, it is essential to utilize methods capable of dissecting the interactomes of chromatin-associated proteins. MS-based proteomics has emerged as a powerful tool for studying protein-protein interactions (Gingras *et al.*, 2007). There are three primary strategies for investigating chromatin-associated complexes: (i) affinity purification via antibodies; (ii) proximity biotinylation labeling followed by streptavidin-pulldown and (iii) chemical proteomics approaches (Figure 10).

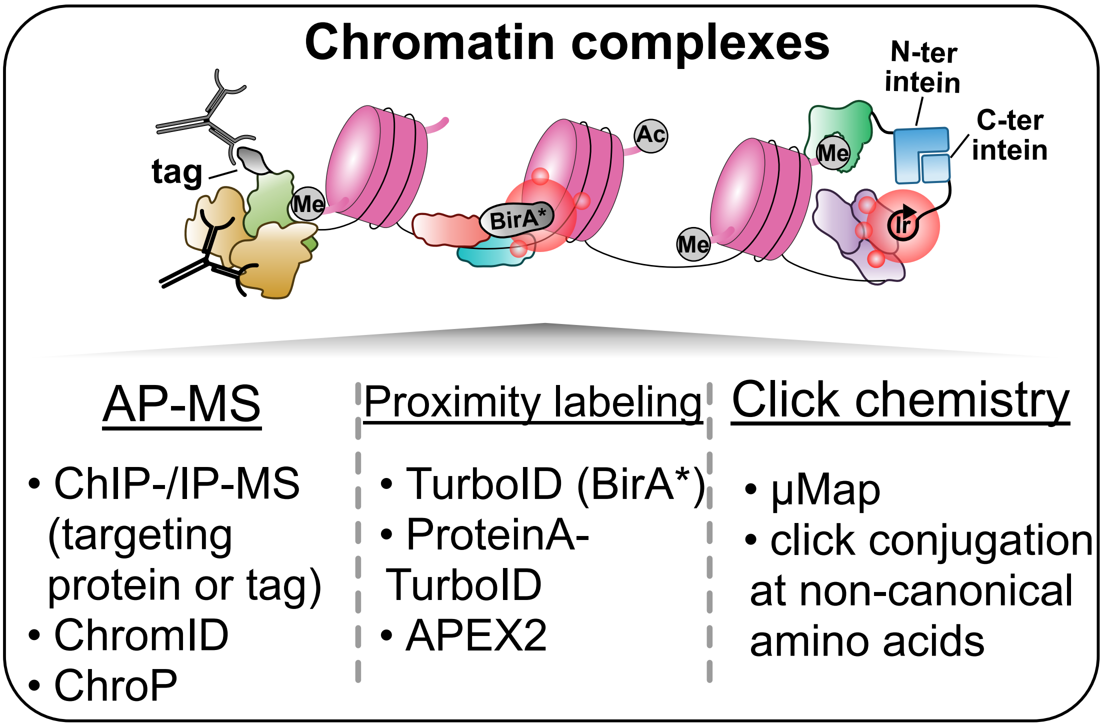


Figure 10: Overview of methods to study protein-protein interactions at chromatin. Chromatin complex compositions and more general protein-protein interactions can be investigated by (i) affinity purification using antibodies; (ii) proximity biotinylation labeling followed by streptavidin pulldown; and (iii) chemical proteomics approaches.

Affinity purification coupled with mass spectrometry (AP-MS) may involve incubating a cellular lysate with a labeled bait, such as a specifically modified DNA oligomer or histone (Bartke *et al.*, 2010; Spruijt *et al.*, 2013). For instance, Spruijt and colleagues employed MS-based proteomics to identify readers for 5mC and its oxidized derivatives in mESCs, neural progenitor cells (NPCs) and adult mouse brain. This method involved using modified DNA oligomers as bait to “fish” for interacting proteins. The DNA oligomers were coupled to a biotin moiety, enabling the selective capture of proteins binding specifically to the modified cytosines. By comparing the protein profiles across three cell types and tissues, Spruijt and colleagues identified distinct and overlapping sets of proteins interacting with each cytosine modification. Alternatively, AP-MS can involve immunoprecipitation (IP) of a protein with its interaction partners in either a native (Lambert *et al.*, 2009; Vermeulen *et al.*, 2010) or crosslinked state (Ji *et al.*, 2015; Wierer and Mann, 2016).

Native IP-MS experiments start with cell lysis and DNA shearing or digestion followed by protein quantification and normalization of protein inputs per replicate. A bait protein is then selectively enriched by using a specific antibody against the bait or an epitope tag that is fused to the bait. The antibody-protein complex is subsequently captured by protein A/G-coated beads and thoroughly washed to minimize non-specific interactors of the bait protein. The bait-associated proteins are then digested on the beads, and the resulting peptide mixture is subjected to LC-MS/MS for identification and quantification of the bait interactors (Gingras *et al.*, 2007).

While numerous research efforts deciphered the interactomes of single proteins using IP-MS, fewer studies have focused on a systems-wide approach to capture an organellar or cellular interactome. A groundbreaking study by Hein and colleagues is one such example in which the authors characterized over 1,100 different baits, covering a substantial part of the expressed proteome with over 28,000 interactions (Hein *et al.*, 2015). Notably, the authors introduced an innovative data analysis strategy in which significant interactors are assessed for their stoichiometry across different IPs to estimate stable and more transient protein complexes. The most recent endeavors to map interactomes comprehensively include aligning cellular localization with protein interactomes, which were subsequently integrated into a searchable web-based tool to aid future research (Y. Qin *et al.*, 2021; Cho *et al.*, 2022). Note that Cho and colleagues harnessed CRISPR-mediated genome editing to create a library of 1,310 fluorescently tagged cell lines. This resource allowed them to investigate the subcellular localization and physical interactions of the corresponding proteins. By employing unsupervised clustering and machine learning for image analysis, they identified colocalizing or interacting proteins. Considering the number of baits analyzed by both microscopy and mass spectrometry, this study demonstrated a strong correlation between the subcellular distribution of proteins and their cellular functions, implying that protein localization is inherently predictive of biological function.

In contrast to native IP-MS, experiments can integrate chromatin crosslinking before cell lysis and immunoprecipitation to preserve protein-DNA interactions (Wierer and Mann, 2016). Analogous to ChIP-Sequencing, these methods are generally referred to as ChIP-MS. Unlike IP-MS, ChIP-MS involves formaldehyde crosslinking and requires rigorous chromatin shearing or digestion, along with additional quality controls to ensure optimal DNA fragment sizes of below 1000 base pairs (bp). Notably, Rapid Immunoprecipitation Mass Spec of Endogenous proteins (RIME) is the first variation of ChIP-MS that introduced on-bead digestion after pulldown without requiring SDS-PAGE gel-based separation of the bead eluate and additional cleanup (Mohammed *et al.*, 2013). The Chromatin Proteomics (ChroP) method builds upon ChIP-MS by incorporating a PTM analysis of the enriched histones post ChIP (Soldi and Bonaldi, 2014). This is achieved by selectively enriching core histones through cutting the corresponding bands from an SDS-PAGE gel and subsequently analyzing each histone band by MS. This robust method permits the identification of the histone PTM context in which a chromatin binder is predominantly present, information that can be downstream validated by ChIP-Seq of the histone PTMs and the respective chromatin binders. ChroP has been recently optimized for IP-MS without crosslinking (Nicosia and Bonaldi, 2021).

Common ChIP-MS protocols start with whole cell lysates and can thus include cytoplasmic proteins that appear as interactors of a bait protein that is *in vivo* localized to chromatin. In Chromatin Immunoprecipitation - Selective Isolation of Chromatin-Associated Proteins (ChIP-SICAP), chromatin is selectively enriched first, and then ChIP-MS is performed (Rafiee *et al.*, 2016). Rafiee and colleagues demonstrated the effectiveness of their method by performing ChIP-SICAP on NANOG, finding that their method captured fewer proteins than ChIP-MS or RIME overall, but the captured proteins were more frequently *bona fide* NANOG-interactors, suggesting a reduction of background proteins. ChIP-MS and its derivations have been instrumental for many impactful biological insights (Wang *et al.*, 2013; Ji *et al.*, 2015; Kloet *et al.*, 2016; Sigismondo *et al.*, 2023), such as the characterization of dynamic reorganization of polycomb group complexes during embryonic stem cell differentiation (Kloet *et al.*, 2016).

IP experiments are constrained by the quality of the antibody, making it challenging to compare different IPs. One alternative involves proximity biotinylation methods, where the bait protein is fused to a biotin ligase derived from *Escherichia coli* (*e.g.* BirA, TurboID, miniTurboID) or the modified ascorbate peroxidase APEX2 from the soybean *Glycine max* (Roux *et al.*, 2012; Lam *et al.*, 2015; Mick *et al.*, 2015; Branon *et al.*, 2018). These methods capture proteins within a radius of about 10 nm. After biotinylation, cells can be lysed, and the lysate combined with streptavidin-coated beads. The robust interaction between biotin and streptavidin permits high salt and detergent concentrations during downstream washes, thereby reducing contaminants. The first generation of engineered promiscuous biotinylating enzymes required extensive biotinylation times, but newer versions require only 1-10 minutes (*e.g.* APEX2, TurboID), facilitating tightly controlled time course experiments. Furthermore, methods employing split variants of these enzymes exist, where each half is fused to a different protein. The enzyme is only active when both proteins are in close proximity, offering new experimental possibilities, such as exploring the common interactomes of two interacting proteins (Schopp *et al.*, 2017; Han *et al.*, 2019; Cho *et al.*, 2020). Based on these approaches, Rodrigo Villaseñor and colleagues devised ChromID, where instead of fusing the biotinylating enzyme to a full-length protein, it is fused to an epigenetic reader domain such as an H3K9me3-binding chromo domain (Villaseñor *et al.*, 2020). This method allows for capturing the local interactome of a specific epigenetic modification. Moreover, Irene Santos-Barriopedro and colleagues introduced ProtA-Turbo, a Protein A fusion with the proximity biotinylation enzyme TurboID. By permeabilizing target cells, purified ProtA-Turbo can be directed towards proteins or post-translational modifications using bait-specific antibodies, analogous to CUT&RUN (Skene and Henikoff, 2017; Santos-Barriopedro, van Mierlo and Vermeulen, 2021). Proteins that are in the vicinity of the bait are then biotinylated and can be specifically enriched for MS. Hence, the method eliminates the need for CRISPR-based introduction of a protein tag or cell transfection and can be applied to primary cells, while being limited by the antibody quality. The authors demonstrated this workflow by targeting Emerin, H3K9me3, and BRG1, revealing *e.g.* that FLYWCH1 is a novel interactor at H3K9me3-marked (peri)centromeres.

Another alternative to IP-MS approaches is based on chemical proteomics, which facilitates context-specific interactome analysis. One such method is the *in vivo* Crosslinking-Assisted and Stable Isotope Labeling by Amino acids in Cell culture (iCLASPI). This technique employs site-specific photo-crosslinking by harnessing diazirine-containing unnatural amino acids, stabilizing both strong and transient interactors for further co-purification (Kleiner *et al.*, 2018). A recent addition to the chemical proteomics-based methodologies is termed µMap, which tracelessly incorporates iridium-photosensitizers into the nuclear micro-environment using engineered split inteins (Seath *et al.*, 2023). This approach captures protein-protein interactions within an approximately 10 nm radius through photo-crosslinking. Remarkably, µMap was demonstrated to reveal altered interactomes caused by single aa mutations or drug treatments. For instance, µMap was applied to the histone mutation H2A E92K, which is implicated in various cancer types. The findings suggest that the local chromatin micro-environment is indeed sensitive to the H2A E92K mutation, impacting a specific subset of acidic patch-binding proteins. This insight could be crucial in identifying new therapeutic intervention opportunities.

In summary, the plethora of methods explored in this chapter provides complementary insights into protein interactomes, enabling the identification of stable and transient interactions, and extending our understanding of functional chromatin-complexes and histone PTM readers. However, each approach comes with unique challenges, such as the need for genetic engineering, reliance on high-quality antibodies, and the potential co-isolation of impurities leading to misclassification of interactors. Hence, researchers must wisely select and optimize the method best suited to their specific research objectives and experimental constraints.

2.3.3 Gene locus-resolving chromatin proteomics

The initial sections of this chromatin proteomics chapter discussed the analysis of proteins globally associated with chromatin and the interactome of an individual chromatin-associated protein. A more intricate understanding of genomic regulation can be achieved by deciphering the local proteome at a specific genetic locus. This has prompted the recent development of several techniques designed to capture the locus-specific chromatome, which presents unique biochemical challenges. The complexity arises due to ubiquitous structural chromatin binders, such as histones that hamper the identification of local proteins. Moreover, a short single locus accounts for a tiny fraction of the entire genome, making it difficult to distinguish and adequately enrich from a multitude of similar structures. Gauchier and colleagues estimated that most single-step affinity purification methods provide a maximum of 1,000-fold enrichment of a target locus, which results in a single 3 kb locus proteome consisting of 99.9% background proteins and only 0.1% locus-specific factors (Gauchier *et al.*, 2020). Comprehensive locus-specific chromatome analysis is further challenging due to low abundant chromatin binders as well as highly dynamic binding and dissociation events at a target region (Wierer and Mann, 2016; Gauchier *et al.*, 2020; Sigismondo, Papageorgiou and Krijgsveld, 2022). To specifically enrich the local chromatome, techniques have been developed that (i) employ direct hybridization of labeled DNA oligos, (ii) use a catalytically inactive Cas9 (dead Cas9, dCas9) fused to a biotinylating enzyme, or (iii) leverage genomic engineering strategies followed by next generation sequencing techniques (Figure 11).

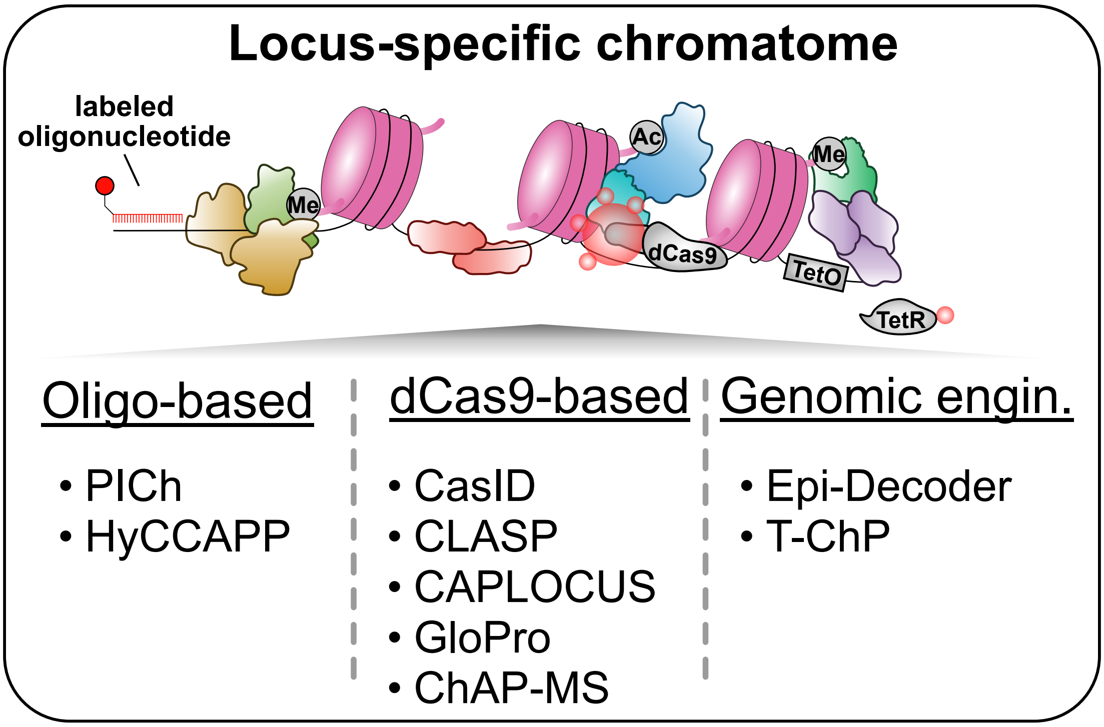


Figure 11: Methods to study the locus-specific chromatin composition. Chromatin proteomics techniques that resolve the local chromatome include (i) the use of direct hybridization of labeled DNA oligos; (ii) the application of a dCas9 fused to a biotinylating enzyme; and (iii) the employment of genomic engineering strategies followed by next-generation sequencing techniques.

Proteomics of isolated chromatin segments (PICh) pioneered locus-specific chromatin proteomics (Déjardin and Kingston, 2009). This method requires formaldehyde crosslinked and purified chromatin, incubated with locus-complementary and desthiobiotin-labeled DNA oligonucleotides. PICh was first tested on telomeric chromatin and has successfully identified the six shelterin proteins that protect the telomeres (de Lange, 2005). In addition, PICh uncovered hundreds of novel telomere-binding proteins and enabled the characterization of the alternative lengthening of telomeres pathway. While PICh can enrich repetitive genomic regions, it is not suitable for studying the composition of single-copy small loci in mammalian genomes (Gauchier *et al.*, 2020).

The advent of CRISPR-Cas9 technology enabled the characterization of more comprehensive locus-specific chromatomes. In an initial effort termed CRISPR-ChAP-MS (CRISPR-based Chromatin Affinity Purification with Mass Spectrometry) dCas9 was fused to protein A and captured by IgG-coated beads (Waldrip *et al.*, 2014). By using guide RNAs specific to the promoter region of the GAL1 gene in *Saccharomyces cerevisiae* and subsequent formaldehyde crosslinking, dCas9 can be enriched while being bound to the region of interest. dCas9 can be also fused to the promiscuous BirA* so that proteins in proximity to dCas9 are biotinylated and enriched (Schmidtmann *et al.*, 2016). Alternatively, dCas9 can be biotinylated by a co-expressed BirA, thus enabling direct dCas9 pulldown under more stringent washing conditions than in CRISPR-ChAP-MS (Liu *et al.*, 2017). All these methods have shown effectiveness in identifying locus-specific proteins, especially when combined with proximity-labeling instead of direct dCas9 pulldowns (Gauchier *et al.*, 2020). For instance, CasID has shown promise in identifying shelterin proteins at telomeres and thereby implemented an orthogonal control of dCas9 localization by an additional GFP-fusion. Other proximity-labeling techniques like CAPLOCUS utilize dCas9 fused with the promiscuous and rapid APEX2 to generate highly reactive biotin-phenoxyl radicals that covalently attach to tyrosine moieties in nearby proteins (Gao *et al.*, 2018; Myers *et al.*, 2018; Qiu *et al.*, 2019). CAPLOCUS achieved true single-locus chromatome resolution of a non-repetitive 233 bp long region by fusing dCas9 additionally to multiple MCP coat proteins that are bound by MS2 coat protein fused APEX2. As a result, two single copy loci on chromosome 11 could be characterized regarding their chromatome.

The genome engineering-based strategy Epi-Decoder harnesses DNA barcoding and sequencing to identify local proteins (Korthout *et al.*, 2018). Briefly, short DNA barcodes are integrated into the genome as molecular identifiers of chromatin states, and a library of yeast clones, each containing unique barcodes and a specific protein are generated. The method involves crosslinking proteins to DNA followed by ChIP, and then identifying and quantifying barcodes via sequencing. Applying this method, Korthout and colleagues showed the degree of protein binding to specific genomic loci, providing insights into chromatin state and binding of 469 proteins. While several of these were known DNA-binding proteins, a significant amount was not expected to be DNA-binding due to their known roles in RNA processing or metabolic functions. Nonetheless, Epi-Decoder in its current state does not provide a strong alternative to MS-based methods, which have achieved much higher sensitivity in recent years and do not require a library of tagged proteins for each species or the integration of DNA barcodes into the region of interest. In contrast, one limitation of the MS-based strategies is the high sample input requirement. Hence, gene locus-resolving techniques require further optimizations.

2.3.4 Histone PTM-resolving chromatin proteomics

Changes to the chromatome composition, such as the binding of transcription factors, chromatin remodelers or epigenetic regulators, can be efficiently resolved by the chromatin proteomics methods previously discussed. However, binding of for instance epigenetic regulators to chromatin is not directly indicative of a certain outcome. Therefore, MS-based methods have emerged that readily identify and quantify epigenetic modifications, such as histone PTMs (hPTMs) or DNA methylation levels (Huang *et al.*, 2015; Noberini, Robusti and Bonaldi, 2022; Sigismondo, Papageorgiou and Krijgsveld, 2022). Histones are typically enriched for mass spectrometry through acidic extraction and acetone-mediated precipitation (Shechter *et al.*, 2007). LC-MS/MS resolves the positions of one or multiple hPTMs within a peptide, which is derived from the mass difference between the observed and theoretical peptide. To date, around 30 unique hPTMs have been described, with frequent analysis performed for methylation (mono-, di-, or tri-), acetylation or ubiquitination on lysines, or phosphorylation on serines, tyrosines, or threonines (Millán-Zambrano *et al.*, 2022). High-resolution MS1 and MS2 scans are crucial for the precise identification of neighboring modifications on histones, distinguishing similarly modified peptides or resolving PTMs with nearly identical masses, such as lysine tri-methylation (42.047 Da) and acetylation (42.011 Da). Moreover, histone-derived peptides can share their mass but differ in their amino acid sequence (isobaric peptides) which can be resolved by MS instruments capable of performing a third MS scan (Huang *et al.*, 2015; Yu *et al.*, 2020).

Histones, rich in lysine and arginine, result in peptides that are too short for MS analysis upon trypsin-mediated digestion. Therefore, a single MS strategy cannot expose the combinatorial PTMs of one histone and accurately quantify them. Three primary MS strategies for hPTM analysis have been developed: (i) top-down, (ii) middle-down, and (iii) bottom-up proteomics (Figure 12). As introduced earlier in the chapter on sample preparation, these methods mainly differ in the size of peptides or intact proteins analyzed. Top-down proteomics, which analyses intact histones without enzymatic digestion, preserves the combinatorial information of hPTMs. However, the complexity of the full MS scan due to multiple charge states can be challenging and is incompatible with identifying and quantifying multiple histones from a complex sample. Alternatively, middle-down proteomics partially resolves combinatorial hPTMs also from more complex samples. However, longer peptides from middle-down proteomics are not well-retained on conventional or improved reversed-phase LC methods, necessitating specialized LC setups. In contrast, the standard bottom-up proteomics workflow focuses on 7-20 amino acid long peptides, which are achievable in the case of histones through a chemical modification of lysines by propionylation. As a result, trypsin does not digest after lysines, yielding an ArgC-like digestion pattern without requiring the more expensive and potentially error-prone ArgC enzyme (Golghalyani *et al.*, 2017). Bottom-up proteomics offers the clear advantage of applicability to complex samples and for *de novo* identification of hPTMs. However, it largely lacks the ability to resolve combinatorial hPTMs and is usually restricted to 1-3 lysines per peptide.

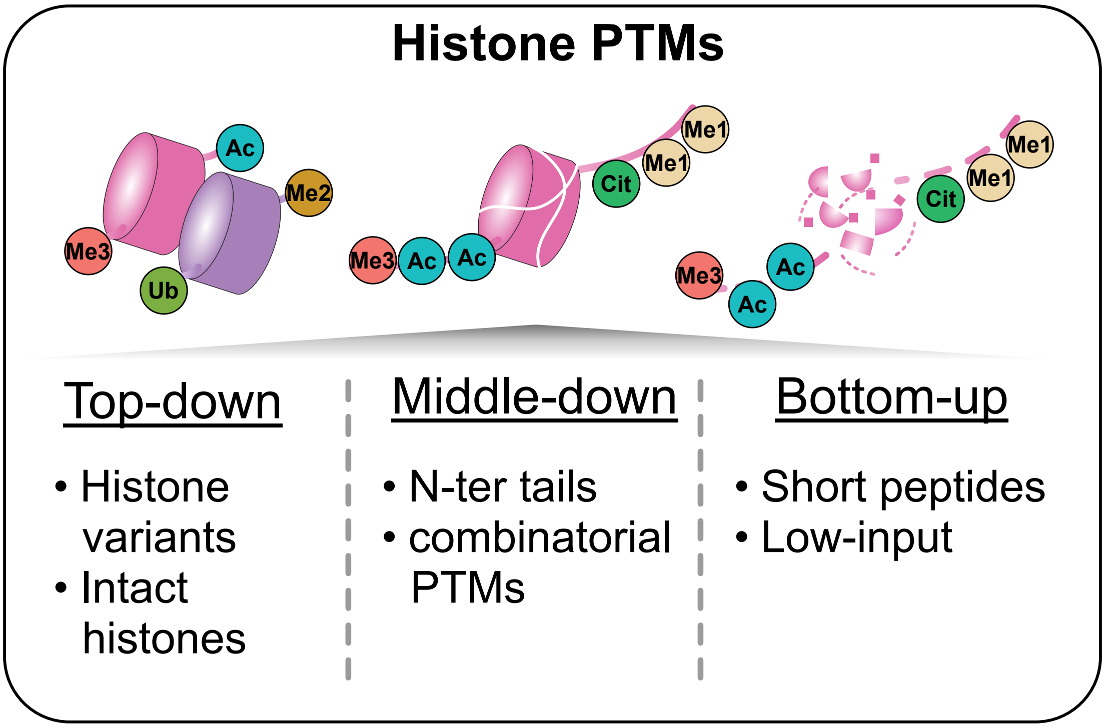


Figure 12: Methods to study hPTMs. The MS-based investigation of bulk hPTMs can be based on top-down, middle-down or bottom-up proteomics. These methods, which primarily differ in the size of the peptides or intact proteins analyzed, provide advantages when analyzing complex samples (bottom-up) or combinatorial hPTMs (middle-down and top-down) including histone variant-specific PTMs (mostly by top-down).

Combining hPTM analysis with other omics approaches can yield powerful insights. For example, Moussaieff and colleagues integrated metabolome data with global hPTM levels in differentiating pluripotent stem cells (Moussaieff *et al.*, 2015). By this, the authors discovered an intricate link between glycolysis-mediated reduction in Acetyl-CoA levels during differentiation and the downregulation of histone acetylation. They further demonstrated that PSCs direct pyruvate toward Acetyl-CoA, but not lactate, leading to increased histone acetylation. Furthermore, they showed that inhibition of Acetyl-CoA production causes a loss of pluripotency, while preventing its usage through small molecule inhibitors can significantly delay the differentiation of the cells.

In conclusion, mass spectrometry-based analysis of hPTMs is a complex, yet rewarding field. The ability to identify, locate, and quantify hPTMs has significantly broadened our understanding of the epigenetic landscape. Further advancements in proteomic methodologies and computational methods continue to expand the boundaries of this research area.

3 DISCUSSION

This work includes the development and application of MS-based chromatin proteomics methods across different biological fields (Figure 13):

Loss of epigenetic modifiers can perturb the proteome which can be effectively resolved by full proteome measurements. In this work, we conducted full proteome analyses in combinations of *Tet* KOs in naive and formative PSCs. Our results demonstrated that, unlike TET1, TET2 levels decrease during the naive to formative transition, while TET2 remains primarily responsible for most 5-formylcytosine formation (Mulholland, Traube, *et al.*, 2020). Moreover, we analyzed the full proteome of an acute myeloid leukemia (AML) cell culture model (K562 cells) with an *Ezh2* KO. This analysis uncovered differential expression of hundreds of direct EZH2-target genes, such as the upregulation of the drug efflux regulator FHL1. Intriguingly, FHL1 was recently described to promote resistance against cytarabine, a chemotherapy drug for AML patients (Luo *et al.*, 2016; Fu *et al.*, 2020). These results offer a potential explanation for the development of cytarabine resistance in AML patients with loss-of-function mutations in *Ezh2* (Kempf *et al.*, 2021). Epigenetic modifications not only regulate gene expression but also ERVs which, in turn, regulate crucial transcriptional programs during development (Wang *et al.*, 2014). ERVs are silenced via DNA methylation, H3K9me3 or a combination of both (Groh and Schotta, 2017). In this context, we sought to identify which heterochromatic pathways silence ERVs in visceral and definitive endoderm. Our results indicated that ERVs are silenced in the definitive endoderm via DNA methylation and in the visceral endoderm via SETDB1-mediated H3K9me3 formation (Wang *et al.*, 2022).

ChIP-MS-based interactome analysis was performed in three separate collaborations. The first interactome analysis involved the naive pluripotency marker DPPA3, which led to the identification of its primary interaction partner UHRF1, an essential co-factor of the DNA methyltransferase DNMT1, along with several nuclear import and export related proteins. The DPPA3-UHRF1 interaction was found to be crucial for passive DNA demethylation in ground phase PSCs, as DPPA3 evicts UHRF1 from chromatin and shuttles it out of the nucleus (Mulholland, Nishiyama, *et al.*, 2020). The second interactome analysis focused on the histone H3K9me3 reader HP1β and revealed its interaction not only with heterochromatic proteins like KAP1, but also with factors involved in pluripotency regulation. Analysis of the KAP1-dependent ubiquitinome uncovered that many of its ubiquitin targets are related to pluripotency, such as REX1 or NR0B1. These findings propose a novel function for HP1β and KAP1 in pluripotency regulation, likely associated with their ability to maintain a repressive chromatin environment (Qin, Ugur, *et al.*, 2021). In a related project dealing with the phase separation of heterochromatic compartments, we could demonstrate that HP1β-driven phase separated droplets contain numerous H3K9me3-associated proteins but also transcriptional activators (Qin, Stengl, *et al.*, 2021). The third interactome analysis focused on TET1 and uncovered its role as a central hub for chromatin associations of several epigenetic complexes, thereby regulating histone modifications independently of its catalytic activity (Stolz *et al.*, 2022).

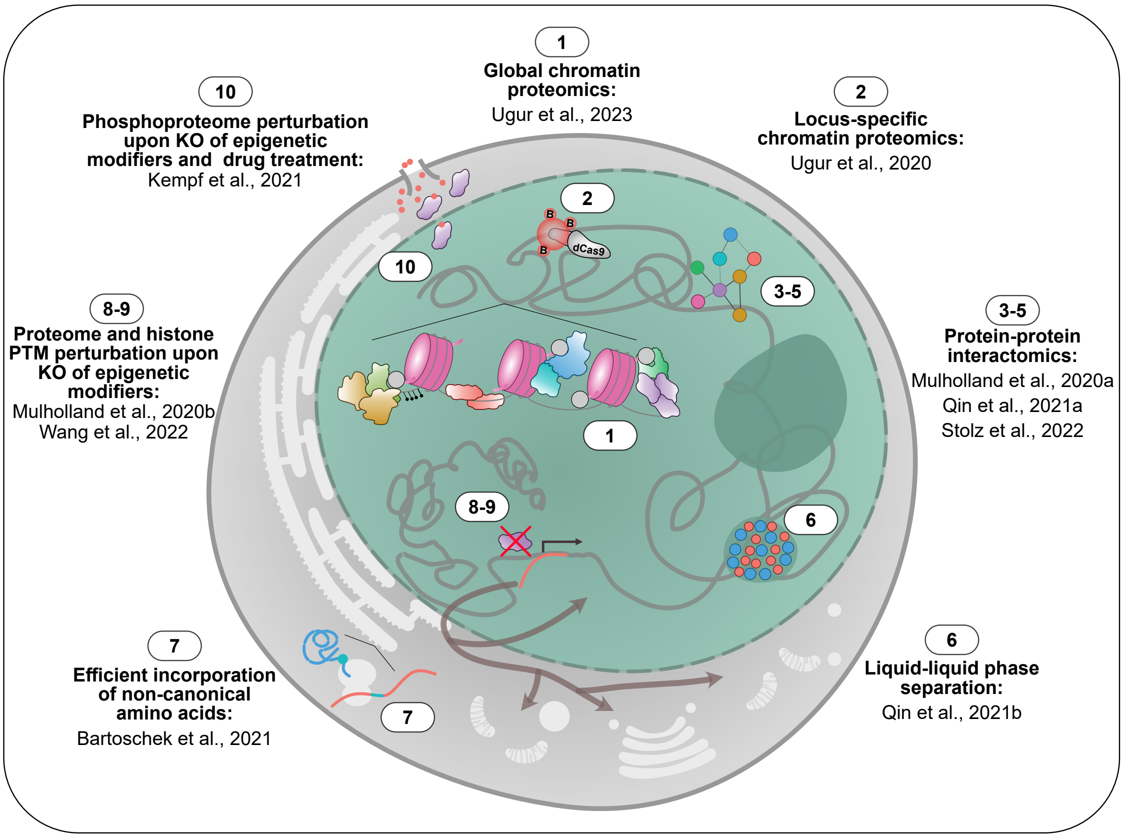

1
Global chromatin proteomics:
Ugur et al., 2023
10
Phosphoproteome perturbation upon KO of epigenetic modifiers and drug treatment:
Kempf et al., 2021
2
Locus-specific chromatin proteomics:
Ugur et al., 2020
8-9
Proteome and histone PTM perturbation upon KO of epigenetic modifiers:
Mulholland et al., 2020b
Wang et al., 2022
3-5
Protein-protein interactomics:
Mulholland et al., 2020a
Qin et al., 2021a
Stolz et al., 2022
7
Efficient incorporation of non-canonical amino acids:
Bartoschek et al., 2021
6
Liquid-liquid phase separation:
Qin et al., 2021b
2
B
B
B
dCas9
10
3-5
1
8-9
6
7

Figure 13: Schematic overview of MS-based chromatin proteomics applications covered by this work. Several biological fields were covered by the application of chromatin proteomics ranging from global chromatin compositions in the three major phases of pluripotency (Ugur *et al.*, 2023) over to locus-specific chromatin proteomes (Ugur, Bartoschek and Leonhardt, 2020) and protein-protein interactomes of different epigenetic regulators such as DPPA3 (Mulholland, Nishiyama, *et al.*, 2020), HP1β (Qin, Ugur, *et al.*, 2021) and TET1 (Stolz *et al.*, 2022). This work also covers the analysis of the proteomic composition of liquid-liquid phase separated droplets driven by HP1β (Qin, Stengl, *et al.*, 2021) and provides an approach for efficient incorporation of non-canonical amino acids (Bartoschek *et al.*, 2021). Lastly, alterations on proteome and histone PTM level are analyzed upon perturbations of several epigenetic modifiers such as TET1 and TET2 (Mulholland, Traube, *et al.*, 2020), EZH2 (Kempf *et al.*, 2021) and SETDB1 (Wang *et al.*, 2022).

On the methodological side, we established and applied a proteomics-based assay to assess the incorporation rates of non-canonical amino acids (ncAAs) at endogenous amber stop codons. The resulting data were integrated into a model that assigns scores to ncAA incorporation sites and proposes codon-optimized sequences without altering the amino acid sequence if necessary. This model enables efficient site-directed incorporation of an ncAA carrying, for example, a biotin moiety into a chromatin binding protein, which can be harnessed for interactome studies (Bartoschek *et al.*, 2021). Moreover, this work presents a step-by-step protocol for investigating locus-specific chromatin binders, which serves as a comprehensive guide for future research (Ugur, Bartoschek and Leonhardt, 2020).

The following discussion will center around the primary contribution of this work, which is the expansion of the chromatin proteomics toolkit through the development of a novel global chromatin proteomics method and the investigation of the chromatome reorganization during pluripotency phase transitions.

3.1 Achieving comprehensive and accurate global chromatin proteomics

3.1.1 Improving the chromatin selectivity of ChAC-DIA

Pioneering studies have demonstrated the feasibility of unbiased chromatin purification and proteomic analysis (Shiio *et al.*, 2003; Meshorer *et al.*, 2006; Torrente *et al.*, 2011; Kustatscher, Hégarat, *et al.*, 2014; Kustatscher, Wills, *et al.*, 2014; Kulej *et al.*, 2017; Ginno *et al.*, 2018; Aranda *et al.*, 2019). However, a comprehensive and reproducible analysis of the chromatome remained an unaddressed challenge. Here, we devised ChAC-DIA, a global chromatin purification and high-resolution DIA-MS method, which is conducted in a single tube and thus is compatible with minute sample amounts. This streamlined approach requires only three hours of hands-on experimentation time. It effectively minimizes the presence of non-chromatin proteins and enables the identification of over twice the number of DNA-binding proteins compared to alternative methodologies within only half of the MS acquisition time (van Mierlo, Wester and Marks, 2019; van Mierlo and Vermeulen, 2021).

At present, ChAC-DIA identifies over 5000 proteins in a single experiment, but only enriches around 2000 proteins compared to the full proteome. These 2000 enriched proteins predominantly include known nuclear and DNA-binding proteins, such as DNMT1 or ESRRB. However, around 20% of the identified and annotated DNA-binding proteins are not significantly enriched in the chromatome compared to the full proteome. Hence, further improvements are needed to enhance the chromatin selectivity of the ChAC-DIA method.

One potential improvement is to modify the initial chromatin purification step by exploring different detergents. The current ChAC-DIA protocol employs a ChEP-like strategy with the addition of SDS and Urea. While SDS and Urea are potent chaotropic detergents, alternative detergents with nonionic (Digitonin and Triton X-100), zwitterionic (CHAPS) or anionic (SDC) properties may enable efficient reduction of background proteins either globally or selectively (Linke, 2009). Importantly, detergents could be used in combination to exploit their individual selectivity or could be paired with salts to reduce non-formaldehyde crosslinked ionic interactions. Moreover, introducing an RNA digestion step before nuclear lysis could reduce RNA-binding proteins in the resulting chromatin pellet. Such an RNA digestion step was already implemented in previous global chromatin proteomics methods (Kustatscher, Wills, *et al.*, 2014), but introduces additional variability across replicates (Ugur *et al.*, 2023). The additional variability might be due to the digestion conditions for RNaseA (for 15 min at 37°C and 1100 rpm), and requires further optimization for increased reproducibility.

In addition to a modified chromatin purification step, the current ChAC-DIA protocol could be significantly enhanced by incorporating protein-protein crosslinking, a strategy that promises two main benefits: (i) improved selectivity for mitotically retained bookmarking transcription factors and (ii) amino acid resolution of direct protein-histone interactions. The first benefit is related to the inability of formaldehyde-crosslinking to conserve DNA interactions with transcription factors that are mitotically retained, such as SOX2 (Teves *et al.*, 2016). These bookmarking transcription factors can be preserved on DNA through the application of additional protein-protein crosslinking agents like DSG (disuccinimidyl glutarate) or EGS (ethylene glycol bis(succinimidyl succinate)) (Festuccia *et al.*, 2016, 2019). Furthermore, protein-protein crosslinkers enable MS-based analysis of interlinks and intralinks, *i.e.*, crosslinks between two proteins or within a protein, respectively, thereby providing evidence for protein-protein interactions at single amino acid resolution and the three-dimensional protein conformation (Chen and Rappsilber, 2023). It is very likely that the information derived from these crosslinks will give insights into uncharacterized proteins (Lenz *et al.*, 2021), especially if combined with AlphaLink, a deep learning-based structure modelling tool which enables the estimation of complex compositions based on crosslinking data (Stahl, Brock and Rappsilber, 2023). However, previous attempts to integrate protein-protein crosslinkers into a chromatin purification workflow resulted in insoluble pellets which rendered this approach impractical for chromatin proteomics (Kustatscher, Wills, *et al.*, 2014; van Mierlo and Vermeulen, 2021). In this work, we also optimized protein-protein crosslinking of global chromatomes (Ugur *et al.*, 2023). For this, we adjusted the crosslinker concentrations to avoid over-crosslinking of the pellet and reformulated the buffer compositions to enable chromatin pellet solubilization as well as crosslink preservation. These optimizations rendered a combination of ChAC-DIA with the protein-protein crosslinker DSSO (disuccinimidyl sulfoxide) feasible. Importantly, DSSO crosslinking neither compromises the number of identified proteins nor the reproducibility across replicates (Ugur *et al.*, 2023). Additional preliminary experiments led to the identification of approximately 1000 unique crosslinks in a 9-hour MS acquisition timeframe, with 94% of these crosslinks being related to nuclear proteins. The protein-protein interaction landscape derived from these 1000 crosslinks centered mostly around nucleosomes. Remarkably, performing protein aggregation capture (PAC) for chromatin samples significantly increased specificity for nuclear proteins and yielded five times more crosslinks compared to acetone precipitation. Further sample fractionation and the adoption of an LC-MS3 acquisition strategy could enhance the number of identified crosslinks. Additionally, the use of enrichable crosslinkers and switching to SCX-based peptide fractionation could further advance the crosslinking MS application of our protocol (Klykov *et al.*, 2018; Steigenberger *et al.*, 2019; Matzinger *et al.*, 2020). It is important to note that while the feasibility of crosslinking MS of chromatin samples has been demonstrated, the next challenge lies in leveraging this approach to gain novel insights into chromatin binding protein complexes and their conformation.

3.1.2 Expanding the scope of global chromatin proteomics

The current ChAC-DIA workflow yields information on the global chromatin abundance of proteins. The versatility of ChAC-DIA allows for its implementation in other workflows, thereby expanding the range of biological data captured. First, in ChIP-MS experiments, whole cell lysates or nuclear lysates are used, which both can contain numerous background proteins that do not interact with the chromatin-associated protein of interest. Strategies, such as using beads with reduced non-specific binding, have been employed to address this issue (Mali *et al.*, 2016; van Andel *et al.*, 2022). However, these strategies may not reduce the co-enrichment of non-chromatin-associated interactors of the bait protein. To prevent the co-enrichment of these background proteins, ChAC-DIA could be performed prior to the ChIP step, potentially enabling a more specific chromatin interactome of the target proteins. Alternatively, proximity labeling techniques like APEX2 or µmap could be utilized instead of ChIP (Lam *et al.*, 2015; Seath *et al.*, 2023). The µmap technique utilizes engineered split inteins to introduce iridium-photosensitizers into the nuclear environment, activating diazirine warheads that generate reactive carbenes. These carbenes crosslink with proximal proteins and may potentially enhance the specificity of the chromatin-specific interactome post ChAC-DIA.

Second, ChAC-DIA yields a high amount of histone proteins and can be therefore used to investigate hPTMs along with chromatome composition. However, the current protocol follows a bottom-up proteomics strategy, rendering the identification of lysine and arginine-rich histone N-termini challenging. This challenge could be overcome by implementing a middle-down ChAC-DIA approach using ArgC or GluC instead of trypsin and LysC for protein digestion. This middle-down approach would likely enable a more comprehensive analysis of hPTMs and their combinations (Sidoli and Garcia, 2017; Janssen *et al.*, 2019). Moreover, a middle-down ChAC-DIA approach could facilitate the identification of additional chromatin binding proteins that are frequently missed in MS when using trypsin digestion, as reported in prior studies (Giansanti *et al.*, 2016). Given that the middle-down approach shows a good chromatome coverage, a single workflow could then potentially address both chromatome and hPTMs. Alternatively, the same chromatin sample could be split into two after the PAC step, allowing for two complementary protein digestion protocols to obtain a comprehensive chromatome analysis combined with information on combinatorial histone PTMs.

Lastly, ChAC-DIA could be performed following pulse-chase labeling with metabolites of epigenetic writers, such as the ^{13}C containing isotope of S-Adenosyl methionine. This strategy could align the chromatome levels of DNA and histone methyltransferases with their respective enzymatic activities during differentiation or genetic perturbations. It would provide a temporal dimension to the study of chromatin, which is crucial for understanding the kinetics of chromatin alterations.

3.1.3 Towards high-throughput and single-cell global chromatome analysis

The interpretation of bulk chromatomes is challenging due to the cell cycle-dependent fluctuations in some protein levels (Kasvandik *et al.*, 2019) and chromatin interactions (Bérubé, Smeenk and Picketts, 2000). To address this issue, sorting cells based on their cell cycle phase, potentially determined by DNA content, followed by ChAC-DIA analysis could be a potential solution. Preliminary experiments employing this strategy yielded promising results, as several marker proteins of the G1 (ORC1), S (ESCO2), and G2 (AURKA) phases exhibited the expected chromatin enrichment pattern (Figure 14): ORC1 accumulates at chromatin in G1 phase (Ohta *et al.*, 2003), while ESCO2 promoters sister chromatid cohesion during S phase (Vega *et al.*, 2005) and AURKA-levels peak at centrosomes in G2 phase (Marumoto *et al.*, 2002).

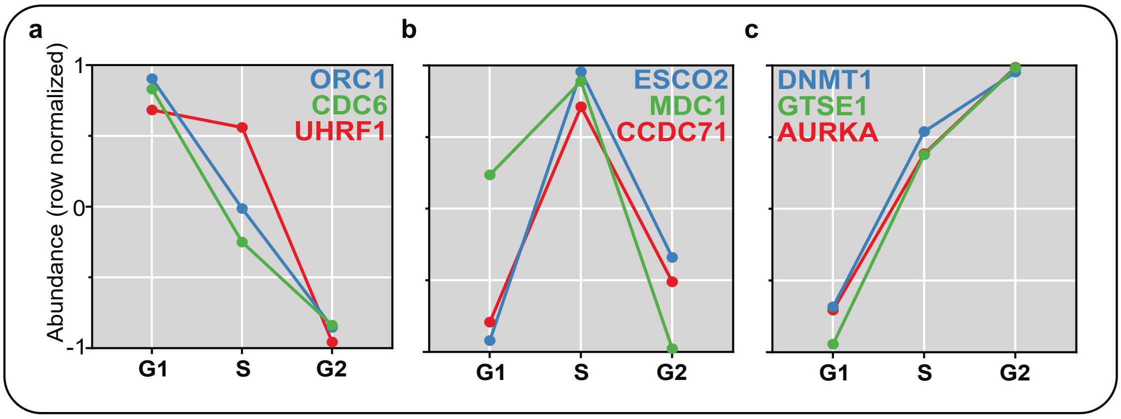


Figure 14: Profile plots illustrating the chromatin enrichment of selected proteins in G1 (a), S (b), and G2 (c) phases. Cells were sorted based on their DNA content, estimated by Hoechst staining. Each cell cycle phase was analyzed in triplicate, with each replicate containing 200,000-250,000 cells. Besides marker proteins of the respective cell cycle phases, UHRF1 and DNMT1 are highlighted due to their differential chromatin enrichment across the cell cycle.

Furthermore, we observed an enrichment of UHRF1 during the G1/S phase and DNMT1 during the S/G2 phase at chromatin. Interestingly, the targeting of DNMT1 to hemi-methylated sites during DNA replication is guided by UHRF1-dependent PAF15 di-ubiquitination in early S phase and Histone H3 di-ubiquitination in late S phase (Nishiyama *et al.*, 2020), which may explain the opposite pattern in DNMT1 and UHRF1 chromatin enrichment.

Implementing cell cycle-specific chromatome measurements, however, would inflate the number of samples per condition, leading to extended MS measurement times. Moreover, large-scale chromatin proteomics experiments are currently challenging with ChAC-DIA as the experimental hands-on time increases with each additional sample and the MS workflow is limited to nine samples per day. To enable multiplexed and high throughput chromatome analysis, three improvements can be incorporated into the current workflow. First, chemical labeling of peptides, such as with TMT (tandem mass tag), would enable the measurement of up to 16 samples in a single MS run (Wang *et al.*, 2020). TMT labeling hence offers a potential solution if MS time is limiting, despite the potential reduction in overall chromatome coverage. Second, ChAC-DIA could be adapted to a 96-well format to enable high-throughput chromatome analysis. One challenge is that high-speed centrifugation, as performed during ChAC-DIA at 20,000 g, is not compatible with standard 96-well plates. However, preliminary results suggest that chromatin purification is effective even when using lower centrifugal forces. Third, the HPLC setup in the current ChAC-DIA workflow limits high-throughput chromatome analysis due to long gradient preparation times (around 20-30 minutes) and additional washing steps between each set of replicates (around 1 hour). Initial experiments indicate that integrating the EvoSep One, an alternative HPLC instrument (Bache *et al.*, 2018), could facilitate robust and rapid chromatome analysis. This is due to automatic peptide desalting and gradient preformation by the EvoSep One, which significantly reduce the overhead time. Preliminary results suggest that up to 30 samples per day can be analyzed instead of the current nine, albeit with less identified proteins in total, which may be acceptable for high-throughput experiments. High-throughput chromatome analysis becomes more relevant with increasing numbers of experimental conditions or when single cell resolution in chromatome analysis is achieved.

Moving towards single-cell chromatin proteomics represents a significant milestone in the field. Previous single-cell analyses across multiple modalities, including the transcriptome, proteome, and epigenome, have contributed to the construction of comprehensive cell atlases (Mereu *et al.*, 2020; Danese *et al.*, 2021). Incorporating single-cell chromatin proteomics would add another dimension to these modalities and provide valuable insights into cellular heterogeneity and chromatin dynamics. To achieve single-cell resolution in ChAC-DIA, several protocol optimizations are required. The current lowest input threshold for ChAC-DIA is around 10,000 cells. To mitigate protein loss, individual cells could be sorted into separate wells and directly crosslinked as whole cells without nuclei isolation. Volumes at each step could be reduced to a few microliters to accommodate the smaller sample size. Additionally, the pelleting of chromatin following the addition of SDS and Urea, coupled with the removal of the nucleoplasmic supernatant, must be performed with minimal sample loss. To facilitate chromatin pelleting, TMT-labeled carrier cells could be used in this crucial step.

3.1.4 High-resolution DIA-MS for chromatin proteomics

DIA-MS generates highly complex mass spectra that, until recently, posed significant challenges in data deconvolution. Two established approaches address this issue: The first involves generating a sample-specific peptide library based on DDA measurements. The second, more recent solution bypasses the need for a peptide library (direct DIA) by utilizing deep neural network-based *in silico* prediction of mass spectra. Both approaches have demonstrated superior accuracy and comprehensiveness compared to DDA-based measurements (Bruderer *et al.*, 2017; Bekker-Jensen, Bernhardt, *et al.*, 2020; Demichev *et al.*, 2020; Pino *et al.*, 2020). For instance, Bruderer and colleagues demonstrated double the number of protein identifications in full proteome measurements, while Bekker-Jensen and colleagues identified 50% more phosphorylation sites in library-based DIA mode compared to DDA (Bruderer *et al.*, 2017; Bekker-Jensen, Bernhardt, *et al.*, 2020). With our ChAC-DIA, we demonstrated twice the number of protein identifications in optimized direct DIA acquisitions. Consequently, the implementation of direct DIA measurements in our study further reduces instrument time and ensures a comprehensive coverage of the chromatome, including low-abundant transcriptional or epigenetic regulators.

ChAC-DIA identified 80% of the transcribed and annotated chromatin binding proteins. One can assume that the remaining transcripts are at least in part also translated and may harbor crucial functions despite their low abundance. Therefore, to enhance chromatome coverage, further optimizations in instrumentation, data acquisition, and analysis strategies are needed. First, recently an alternative solution to further deconvolute DIA mass spectra emerged, termed synchronized parallel accumulation - serial fragmentation (synchro-PASEF). This strategy harnesses ion mobility analyzing mass spectrometers like the timsTOF Ultra (Skowronek *et al.*, 2022) to closely monitor the injected ion cloud and to assign the relationship between peptide precursors and fragment ions which mitigates peptide interferences and ultimately simplifies DIA-derived spectra. Second, newer mass spectrometers, including Bruker's timsTOF Ultra, Thermo's Orbitrap Astral, and Sciex' ZenoTOF 7600 are more sensitive than the Orbitraps Q-Exactive HF-X and Exploris 480, which were utilized for this work (Demichev *et al.*, 2022; Z. Wang *et al.*, 2022; Heil *et al.*, 2023). These instruments identify more proteins within the same LC gradient length or achieve the same protein identification rate in shorter LC gradients. Third, the ChAC-DIA dataset presented in this work has been analyzed without a spectral library. Incorporating a hybrid DIA- and DDA-library strategy could enhance proteome coverage without significantly affecting replicate reproducibility (Lou *et al.*, 2020; Shahbazy *et al.*, 2023). Preliminary results revealed that this approach increases the protein identification rate by 15-20%.

Moreover, high-resolution DIA-MS on the newest generation of MS instruments could facilitate deep chromatome measurements of hundreds of cell lines and states to ultimately catalogue chromatin binding proteins. This is a necessity since current gene ontology-based annotations likely underestimate the number of chromatin or DNA binding proteins (641 and 2314, respectively). For instance, some centromeric proteins (CENPE, CENPF, CENPH, CENPI, CENPJ, CENPK, CENPL, CENPM, CENPN, CENPO, CENPP, CENPQ, CENPU, CENPV), chromatin remodelers (CHAF1A and CHAF1B) and histone modifiers (EHMT1, KAT14, KAT8) are annotated nuclear proteins with no chromatin or DNA binding function. Of note, these classifications, derived from the Gene Ontology knowledgebase (The Gene Ontology Consortium, 2019; Thomas *et al.*, 2022), are limited in the covered cell types and states as well as the underlying experimental data and, hence, cannot predict protein localization in every cell type. For instance, proteins categorized as cytoplasmic in one cell type may occasionally bind to chromatin in another cell type or state. To comprehensively capture the chromatin binding ability of proteins, it is necessary to create a chromatome atlas encompassing diverse cell types and states.

3.1.5 Chromatome analysis and multi-modal data integration

Our results indicate a moderate to weak correlation between the chromatome and either the transcriptome or proteome, respectively. This discrepancy appears to be driven by cell type-specific chromatin-associations of proteins such as those observed for transcription factors in the LIF, Activin A and WNT pathways. For example, the WNT-related transcription factors TCF7 and CTNNB1 (β-Catenin) exhibit higher abundance in the chromatin fraction compared to the nuclear or cytoplasmic fractions of naive PSCs. However, their chromatin-levels decline in formative PSCs, in line with the inhibition of upstream WNT components (Kinoshita *et al.*, 2021; Wang *et al.*, 2021). The experimental workflow presented in this work thus enabled multiple ways of data interpretation: (i) selective purification of chromatin-binding proteins allows for the construction of a global chromatome atlas specific to the cell type of interest; (ii) enrichment of proteins in the chromatome compared to the proteome enables the identification of high-confidence chromatin-associated proteins; (iii) normalization of the chromatome to protein levels (*i.e.*, relative chromatin binding) enables the differentiation between chromatin binding events mediated by changes in total protein levels and those mediated by subcellular relocalization or chromatin affinity.

These analyses could be further expanded by integrating the chromatome data with data from other cellular modalities. First, a near-to-complete chromatome comprises numerous epigenetic reader, writer and eraser proteins without direct information on hPTMs. This information can be inferred either directly from the ChAC-DIA measurements as described in the previous section or from dedicated hPTM measurements. The latter approach has previously provided detailed insights into the regulation of PRC sub-complexes during mouse (van Mierlo *et al.*, 2019) and human (Zijlmans *et al.*, 2022) pluripotency transitions. It is likely that an improved chromatome measurement, such as the one presented here, could yield deeper insights when combined with hPTM measurements. Second, subcellular relocalization events, particularly for signaling pathway components, often correlate with phosphorylation or dephosphorylation events (Nardozzi, Lott and Cingolani, 2010). Thus, integrating chromatome with phosphoproteome data could lead to more precise observations regarding the nuclear translocation and activity of signaling pathway components. Third, chromatome data could be analyzed alongside epigenomic data, such as that obtained by ATAC-Seq followed by transcription factor motif analyses. This would enable the comparison of altered transcription factor binding in the chromatome with changes in transcription factor motif accessibility. In summary, further optimization of ChAC-DIA and chromatin-related multi-modal data integration harbors a great potential for (i) providing a comprehensive view of chromatome reorganization during cell identity changes, (ii) enabling high-throughput screening experiments encompassing hundreds of test conditions within a reasonable MS measurement time and (iii) facilitating in-depth interpretation of chromatome reorganizations.

3.2 Chromatome reorganization during pluripotency

3.2.1 Establishment of a repressive chromatin state towards pluripotency exit

After establishing comprehensive and accurate chromatome measurements, we created a chromatome atlas for mouse ground, formative, and primed PSCs. The majority of chromatome alterations occurred between the ground and formative phases, whereas the proteome still showed considerable changes between formative and primed phases. These findings align with recent publications indicating that formative PSCs are transcriptionally and epigenetically distinct from ground state PSCs and to a lesser extent from primed PSCs (Smith, 2017; Kinoshita *et al.*, 2021; Wang *et al.*, 2021). Interestingly, we observed an increase in histone H1 and HMG variants in the chromatome of primed PSCs compared to formative PSCs which is indicative of a potentially more compact chromatin state. Our dataset facilitates the identification of novel phase-specific chromatin binders that, given their selective chromatin enrichment, likely play a significant role in the phased progression of pluripotency. In our study, we focused on epigenetic modifiers and identified three notable groups in our dataset associated with H3K4me3, H4 acetylation, or H3K9me3, respectively.

The identified H3K4me3-associated proteins in our study include QSER1, which displays increased chromatin enrichment during the formative phase of pluripotency. Recently, QSER1 together with TET1 has been demonstrated to protect bivalent promoters from *de novo* methylation in human ESCs (Dixon *et al.*, 2021). Given that *de novo* methylation peaks at the formative phase, QSER1 might play a similar protective role against DNA methylation in mouse formative PSCs. This aligns with prior studies and our recent work exploring the non-catalytic functions of TET1, demonstrating that TET1 interacts with PRC2 and enhances its activity at bivalent sites under serum/LIF conditions (Chrysanthou *et al.*, 2022; Stolz *et al.*, 2022). Moreover, our dataset reveals a prominent interaction between QSER1 and TET1, suggesting a conserved mechanism wherein QSER1 and TET1 collaborate to prevent *de novo* methylation of genes crucial to development. Another interesting observation was the formative-specific chromatin enrichment of EZHIP, a protein known to mimic H3K27me3 and inhibit PRC2 spreading (Ragazzini *et al.*, 2019; Jain *et* *al.*, 2020). Given that bulk H3K27me3 levels decrease during the transition from ground to formative PSCs, while gene promoter bivalency concurrently increases (Gonzales-Cope *et al.*, 2016; Wang *et al.*, 2021), EZHIP may work in conjunction with QSER1 and TET1 to target PRC2 to formative-specific sites.

Another notable case of phase-specific regulation of epigenetic modifications involves the HBO1 complex, which regulates replication licensing and MCM complex formation by Histone H4 and H3 acetylation (Miotto and Struhl, 2010; Wong *et al.*, 2010). The core HBO1 complex includes the acetyltransferase KAT7 along with ING4, ING5, and MEAF6. Additionally, there are mutually exclusive accessory subunits: JADE1, JADE2, JADE3 as well as BRPF1 and BRPF3 (Iizuka *et al.*, 2009; Xiao *et al.*, 2021). Which particular histone and lysine is targeted by this complex depends on the HBO1 complex composition, particularly the respective accessory subunit (Lalonde *et al.*, 2013). Our chromatome dataset indicates a stable chromatin association of the HBO1 core. However, JADE1, BRPF1 and BRPF3 demonstrated primary enrichment in the ground PSC chromatome. In contrast, JADE3 and JADE2 exhibited specificity for the chromatomes of formative and primed PSCs, respectively. Considering that JADE1 directs HBO1 to Histone H4 K5/8/12 sites, it is remarkable that these sites are downregulated in later phases of pluripotency and, hence, are in line with the reduction of JADE1 (Gonzales-Cope *et al.*, 2016). The global chromatome alterations of HBO1 subunits correlated well with stoichiometry estimations for each subunit derived from ChIP-MS experiments of KAT7, indicating a pluripotency phase-specific targeting of the HBO1 complex.

Moreover, our chromatome analysis revealed enrichment of the H3K9 trimethyltransferases SUV39H1 and SUV39H2 in chromatomes of post-implantation reflecting PSCs. This coincides with an increase in bulk H3K9me3 (Tosolini *et al.*, 2018) and also OCT4-mediated upregulation of *Suv39h1* within the same developmental time frame (Bernard *et al.*, 2022). It should be noted, however, that increased chromatin abundance does not necessarily equate to an increase in SUV39H1/2-mediated H3K9me3. Supporting this, we also detected an enrichment of SETDB1, along with its nuclear translocating co-factor ATF7IP in the chromatomes of formative and primed PSCs (Beyer *et al.*, 2016; Tsusaka, Shimura and Shinkai, 2019). Therefore, we specifically inhibited the catalytic activity of SUV39H1 and SUV39H2 in ground and formative PSCs using Chaetocin and compared bulk H3K9me3-levels. The reduction of H3K9me3 was more efficient in formative PSCs, suggesting a transition from a more SETDB1-driven H3K9me3 deposition in ground PSCs to a SUV39H1- and SUV39H2-driven H3K9me3 deposition in formative and primed PSCs.

The global decrease in the transcriptionally activating H4 acetylation and increase in repressive H3K9 trimethylation goes along with the increased relative chromatin binding of heterochromatic proteins in formative and primed PSCs, such as KAP1, CBX3, FLYWCH1 and histone H1. Similar to these "housekeeping" heterochromatic proteins, we observed higher relative chromatin binding of SUMO1-3 and several SUMO E3 ligases. Recently, SUMOylation of histone H1 was described to be essential for H1 accumulation at chromatin and driving H1-mediated heterochromatin formation in mouse PSCs (Sheban *et al.*, 2022). In line with this, our chromatome analysis demonstrates that both histone H1 and the three SUMO variants along with candidate SUMO E3 ligases are enriched in chromatin towards the exit from pluripotency. Moreover, a comparison with human primed PSCs revealed that some of these heterochromatic proteins have a conserved relative chromatin binding. This suggests that not only the overall abundance but also an increased chromatin affinity of heterochromatic proteins might be a common feature of late pluripotency. So which events contribute jointly to a more restrictive chromatin state of formative and primed PSCs (Figure 15)? First, H3K9me3-levels are more frequently mediated by SUV39H1/2 towards the exit from pluripotency. Second, these H3K9me3 sites along with SUMOylation of H1b likely enable, in turn, the higher chromatin affinity of repressive proteins such as the CBX proteins, KAP1 or FLYWCH1. Third, this is all accompanied by lower levels of the activating histone marks H4K5/8/12Ac due to HBO1 complex reorganization, ultimately resulting in further heterochromatinization and loss of pluripotency. In summary, the described chromatome alterations allow an unprecedented insight into cell identity regulation which, however, requires further characterization by orthogonal experiments and ideally translation to *in vivo* models.

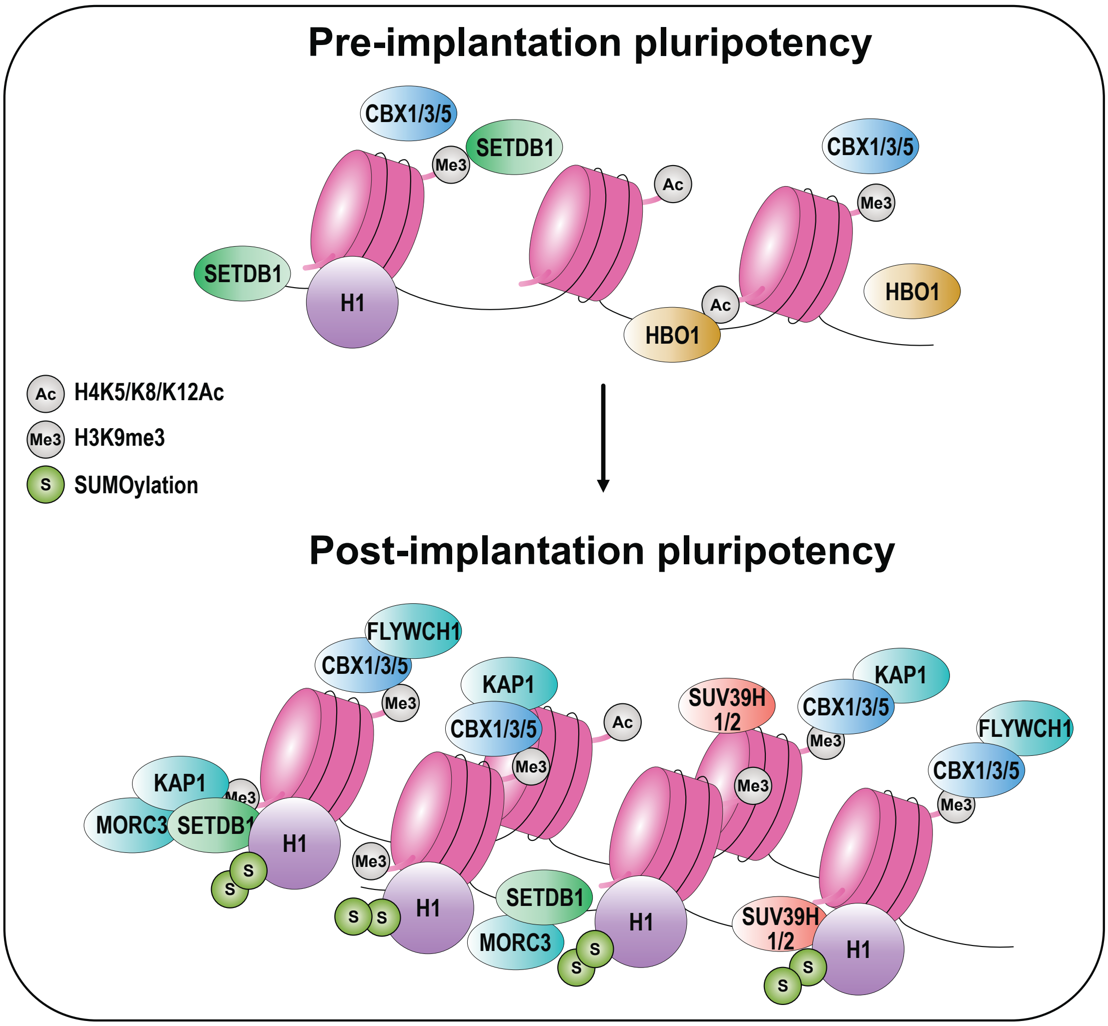


Figure 15: Schematic of heterochromatinization towards the exit from pluripotency. We suggest that from ground to formative and primed phases of pluripotency, H3K9me3-deposition is more frequently mediated by SUV39H1 and SUV39H2. These H3K9me3 sites, along with the SUMOylation of the linker Histone H1, likely contribute to the enhanced chromatin affinity of repressive proteins such as CBX1, CBX3, CBX5, KAP1 and FLYWCH1. Lastly, there is a downregulation of the activating histone marks H4K5/8/12Ac, likely due to the reorganization of the HBO1 complex. This progressive alteration leads to further heterochromatinization and eventual loss of pluripotency.

3.2.2 The epigenetic paradox of pluripotency

Remarkably, while changes in repressive epigenetic modifications are highly conserved in PSCs, they are not essential for maintaining pluripotency or cell viability (Surani, Hayashi and Hajkova, 2007; Meissner, 2010). In fact, PSCs have been observed to tolerate the simultaneous loss of multiple repressive epigenetic modifications (Walter *et al.*, 2016; Tosolini *et al.*, 2018; van Mierlo *et al.*, 2019), whereas the loss of a single epigenetic modifier is often lethal from gastrulation onwards (Faust *et al.*, 1998; O'Carroll, Erhardt, *et al.*, 2001; O'Carroll, Scherthan, *et al.*, 2001; Pasini *et al.*, 2004; Petryk *et al.*, 2021). The only exception to this observation is the loss of *Setdb1*, which is lethal even in ground state PSCs (Dodge *et al.*, 2004; Bilodeau *et al.*, 2009; Yuan *et al.*, 2009). However, it is not fully resolved whether this is related to the silencing function of SETDB1 (Lohmann *et al.*, 2010) or its role in bookmarking poised and cell type-specific enhancers (Barral *et al.*, 2022). Apart from this unresolved exception, the question remains why somatic cell types do need repressive epigenetic marks while PSCs do not? This "epigenetic paradox" of pluripotency points towards a dominance of pluripotency-specific transcription factors and upstream signaling pathways over epigenetic modifications during different phases of pluripotency (Ura *et al.*, 2008; Tsai *et al.*, 2012; Wu *et al.*, 2014; Festuccia, Gonzalez and Navarro, 2017; Mulholland, Nishiyama, *et al.*, 2020). Alternatively, one could ask whether the reduced levels of repressive epigenetic marks are an essential feature of PSCs. For example, pluripotency-specific transcription factors may bind target sites more efficiently due to the lower levels of repressive epigenetic modifications in PSCs compared to differentiated cell types (Festuccia, Gonzalez, and Navarro 2017; Meshorer et al. 2006). Here, previous studies demonstrated that neither overexpression of some heterochromatic proteins such as DNMT1 or SETDB1 nor the knockout of *e.g.* H3K27me3-specific demethylases altered the pluripotent state (Cho, Park and Kang, 2013; Meng *et al.*, 2021; Choudhury *et al.*, 2023). Hence, self-renewal and maintenance of PSCs heavily rely on the activity of pluripotency-associated transcription factors, likely due to their pioneering binding activities. But is there no function at all for repressive epigenetic modifications in pluripotency? While repressive epigenetic modifications may not have a major impact on pluripotency maintenance and cell viability, they do modulate the differentiation capacity of PSCs. For instance, primed PSCs with higher levels of repressive epigenetic modifications are highly inefficient in contributing to blastocyst chimeras (James *et al.*, 2006). Additionally, by selectively reducing the expression of master transcription factors from previous pluripotency phases, epigenetic modifications prevent the reversal of primed PSCs to naive PSCs unless active reprogramming by the Yamanaka factors is induced (Bao *et al.*, 2009; Bultmann *et al.*, 2012; Takahashi, Kobayashi and Hiratani, 2018). It is worth noting that even the pioneering activity of OCT4 and its cooperativity with SOX2, which is crucial for pluripotency transitions (Boyer *et al.*, 2005; Chen *et al.*, 2008; Jerabek *et al.*, 2014; Merino *et al.*, 2014), can be influenced by repressive histone modifications such as H3K27me3 (Sinha *et al.*, 2023).

3.2.3 Co-existence of different cell-identity governing transcription factors

The extent to which master transcription factors, which are essential for establishing specific cell identities, coexist and shape the identity in a concentration-dependent manner (Loh and Lim, 2011; Sarkar and Hochedlinger, 2013) or alternatively undergo rapid switches in response to signaling cues (Boroviak *et al.*, 2015), remains a complex question. Our work, however, has shed light on several such coexisting master transcription factors. For example, we found that GATA4, which plays a pivotal role in endodermal differentiation (Matsuda *et al.*, 1994; Watt *et al.*, 2007), is present in ground PSCs. Similarly, OTX2, a master regulator of formative pluripotency (Buecker *et al.*, 2014; Yang *et al.*, 2014), was detected in ground PSCs and further persisted in primed PSCs. These findings suggest that phase-specific transcription factors coexist and potentially drive the transition to a new cellular identity when they reach a certain concentration threshold. The chromatome offers a unique advantage in addressing this question, as it offers direct evidence of protein presence and chromatin association, unlike the transcriptome or the full proteome.

To gain a deeper insight into the impact of master transcription factors on the chromatin composition, the preblastoderm embryo chromatin assembly extract (DREX), a cell-free *in vitro* reconstitution method of the *Drosophila* chromatome, presents an intriguing model system (Eggers and Becker, 2021). DREX provides a highly adaptable and controlled chromatin environment that could be seamlessly integrated with ChAC-DIA. By selectively titrating the levels of transcription factors in this reconstituted Drosophila chromatome, we could uncover the degree to which these proteins impact the chromatin composition in a concentration-dependent way. This approach would provide further mechanistic insights into the distinction of master transcription factors, thereby enhancing our understanding of their role in cellular state transitions.

3.2.4 Conserved pluripotency features between mouse and human PSCs

Our work primarily focuses on mouse PSCs but also includes the chromatome analysis of conventionally cultured hESCs. These cells are broadly considered the equivalent of mouse primed PSCs and, unlike their mouse counterparts, are the major state acquired upon reprogramming (Davidson, Mason and Pera, 2015; Weinberger *et al.*, 2016). Given the clinical interest in hESCs, their comprehensive characterization is highly relevant. Despite the shared, conserved molecular mechanisms between mouse and human development, differences exist in the required *in vitro* growth conditions and speed of pluripotency transitions. Mouse PSCs transition through pluripotency phases approximately three times faster than human PSCs (Xue *et al.*, 2013; Nakamura *et al.*, 2016; Guo *et al.*, 2021). This discrepancy in timing extends beyond pluripotency and was further observed for instance during body segmentation events (Lázaro *et al.*, 2023). Here, additional cellular modalities such as the chromatome can aid in precisely placing hESCs within the mouse pluripotency spectrum. Therefore, we performed chromatome analysis of hESCs after establishing the same culture conditions for hESCs and mouse formative and primed PSCs to rule out chromatome differences mediated by differential growth conditions. Our data offered five critical observations. First, approximately 75% of the high-confidence chromatome, comprising 2628 proteins, is shared between hESCs and all three tested mouse PSCs. This overlap supports the notion of conserved molecular mechanisms governing pluripotency. Second, the chromatome levels of these shared proteins strongly correlate between hESCs and mouse primed PSCs, yet still exhibit significant correlation between hESCs and mouse formative PSCs. Third, lower chromatome levels of ground phase pluripotency factors constitute the major difference between hESCs and mouse formative PSCs. Fourth, hESCs differ from mouse primed PSCs in the chromatome-association of germ cell differentiation markers, which is consistent with the capability of hESCs (similar to mouse formative PSCs) to differentiate into primordial germ cells. Lastly, we also found evidence of distinct hESC features, such as overall lower activity in the HIPPO signaling pathway compared to all three mouse PSCs. These observations collectively indicate that hESCs are developmentally at a mouse primed PSC-like phase but also distinctly share features with mouse formative PSCs.

Future research could more systematically approach inter-organismal comparisons to answer several outstanding questions. For instance, what are the underlying reasons for the variability in the rate of progression through pluripotency and other stages of embryogenesis among different mammals? Could these disparities in developmental speed influence specific capacities, such as the ability of hESCs to differentiate into germ cells, even while they are primarily considered primed? Another question is whether extended developmental time frames could lead to the emergence of a greater number of intermediate cell identities. While it is challenging to collect embryonic samples from many different mammals, novel reprogramming techniques enable the non-invasive generation of iPSCs (Geuder *et al.*, 2021). In this context, it is also worth noting that early embryonic DNA methylation dynamics are strikingly similar among placental mammals (*Boroeutheria*), characterized by a wave of global DNA demethylation (Monk, Boubelik and Lehnert, 1987; Dean *et al.*, 2001). Yet, it remains unclear whether similar patterns apply to non-placental mammals (Mulholland, Nishiyama, *et al.*, 2020). To pinpoint the specific proteins controlling developmental capacities and speed and to refine the nomenclature of pluripotency phases across species, it would be beneficial to expand these investigations to a broader range of *Boroeutheria* and, ideally, mammals from additional orders.

In conclusion, our work offers several applications of chromatin proteomics, leading to novel biological insights that range from basic epigenetic mechanisms (Mulholland, Nishiyama, *et al.*, 2020; Mulholland, Traube, *et al.*, 2020; Qin, Stengl, *et al.*, 2021; Qin, Ugur, *et al.*, 2021; Stolz *et al.*, 2022; Wang *et al.*, 2022) to disease models (Kempf *et al.*, 2021). We have also expanded the proteomic toolkit for future chromatin-related applications, introduced new data analysis methods, and provided insights into the intricate regulation of pluripotency progression (Ugur, Bartoschek and Leonhardt, 2020; Bartoschek *et al.*, 2021; Ugur *et al.*, 2023). We anticipate that further optimization and applications of chromatin proteomics, combined with the analysis of other cellular modalities, will help to further unravel regulatory events in cell identity establishment and perturbation.

4 BIBLIOGRAPHY

Alajem, A. *et al.* (2015) "Differential Association of Chromatin Proteins Identifies BAF60a/SMARCD1 as a Regulator of Embryonic Stem Cell Differentiation," *Cell Reports*, pp. 2019–2031. doi:10.1016/j.celrep.2015.02.064.

Aliyari, E. and Konermann, L. (2022) "Formation of gaseous peptide ions from electrospray droplets: Competition between the ion evaporation mechanism and charged residue mechanism," *Analytical chemistry*, 94(21), pp. 7713–7721. doi:10.1021/acs.analchem.2c01355.

van Andel, E. *et al.* (2022) "Highly specific protein identification by immunoprecipitation-mass spectrometry using antifouling microbeads," *ACS applied materials & interfaces*, 14(20), pp. 23102–23116. doi:10.1021/acsami.1c22734.

Aranda, S. *et al.* (2019) "Chromatin capture links the metabolic enzyme AHCY to stem cell proliferation," *Science advances*, 5(3), p. eaav2448. doi:10.1126/sciadv.aav2448.

Aston, F.W. (1919) "LXXIV. A positive ray spectrograph," *The London Edinburgh and Dublin Philosophical Magazine and Journal of Science*, 38(228), pp. 707–714. doi:10.1080/14786441208636004.

Auclair, G. *et al.* (2014) "Ontogeny of CpG island methylation and specificity of DNMT3 methyltransferases during embryonic development in the mouse," *Genome biology*, 15(12), p. 545. doi:10.1186/s13059-014-0545-5.

Avilion, A.A. *et al.* (2003) "Multipotent cell lineages in early mouse development depend on SOX2 function," *Genes & development*, 17(1), pp. 126–140. doi:10.1101/gad.224503.

Axel, R. (1975) "Cleavage of DNA in nuclei and chromatin with staphylococcal nuclease," *Biochemistry*, 14(13), pp. 2921–2925. doi:10.1021/bi00684a020.

Bache, N. *et al.* (2018) "A Novel LC System Embeds Analytes in Pre-formed Gradients for Rapid, Ultra-robust Proteomics," *Molecular & cellular proteomics: MCP*, 17(11), pp. 2284–2296. doi:10.1074/mcp.TIR118.000853.

Bader, J.M. *et al.* (2020) "Proteome profiling in cerebrospinal fluid reveals novel biomarkers of Alzheimer's disease," *Molecular systems biology*, 16(6), p. e9356. doi:10.15252/msb.20199356.

Bao, S. *et al.* (2009) "Epigenetic reversion of post-implantation epiblast to pluripotent embryonic stem cells," *Nature*, 461(7268), pp. 1292–1295. doi:10.1038/nature08534.

Barral, A. *et al.* (2022) "SETDB1/NSD-dependent H3K9me3/H3K36me3 dual heterochromatin maintains gene expression profiles by bookmarking poised enhancers," *Molecular cell*, 82(4), pp. 816-832.e12. doi:10.1016/j.molcel.2021.12.037.

Bartke, T. *et al.* (2010) "Nucleosome-interacting proteins regulated by DNA and histone methylation," *Cell*, 143(3), pp. 470–484. doi:10.1016/j.cell.2010.10.012.

Bartoschek, M.D. *et al.* (2021) "Identification of permissive amber suppression sites for efficient non-canonical amino acid incorporation in mammalian cells," *Nucleic acids research*, 49(11), p. e62. doi:10.1093/nar/gkab132.

Batth, T.S. *et al.* (2019) "Protein Aggregation Capture on Microparticles Enables Multipurpose Proteomics Sample Preparation," *Molecular & cellular proteomics: MCP*, 18(5), pp. 1027–1035. doi:10.1074/mcp.TIR118.001270.

Batugedara, G. *et al.* (2020) "The chromatin bound proteome of the human malaria parasite," *Microbial genomics*, 6(2). doi:10.1099/mgen.0.000327.

Baylin, S.B. *et al.* (2001) "Aberrant patterns of DNA methylation, chromatin formation and gene expression in cancer," *Human molecular genetics*, 10(7), pp. 687–692. doi:10.1093/hmg/10.7.687.

Bedzhov, I. and Zernicka-Goetz, M. (2014) "Self-organizing properties of mouse pluripotent cells initiate morphogenesis upon implantation," *Cell*, 156(5), pp. 1032–1044. doi:10.1016/j.cell.2014.01.023.

Bekker-Jensen, D.B., Martínez-Val, A., *et al.* (2020) "A compact quadrupole-orbitrap mass spectrometer with FAIMS Interface improves proteome coverage in short LC gradients," *Molecular & cellular proteomics: MCP*, 19(4), pp. 716–729. doi:10.1074/mcp.TIR119.001906.

Bekker-Jensen, D.B., Bernhardt, O.M., *et al.* (2020) "Rapid and site-specific deep phosphoproteome profiling by data-independent acquisition without the need for spectral libraries," *Nature communications*, 11(1), p. 787. doi:10.1038/s41467-020-14609-1.

Benfey, O.T. (1958) "August Kekule and the birth of the structural theory of organic chemistry in 1858," *Journal of chemical education*, 35(1), p. 21. doi:10.1021/ed035p21.

Bensaude-Vincent, B. (1990) "A view of the chemical revolution through contemporary textbooks: Lavoisier, Fourcroy and Chaptal," *British journal for the history of science*, 23(4), pp. 435–460. doi:10.1017/s0007087400028089.

van den Berg, D.L.C. *et al.* (2008) "Estrogen-related receptor beta interacts with Oct4 to positively regulate Nanog gene expression," *Molecular and cellular biology*, 28(19), pp. 5986–5995. doi:10.1128/MCB.00301-08.

Bernard, L.D. *et al.* (2022) "OCT4 activates a Suv39h1-repressive antisense lncRNA to couple histone H3 Lysine 9 methylation to pluripotency," *Nucleic acids research*, 50(13), pp. 7367–7379. doi:10.1093/nar/gkac550.

Bernstein, B.E. *et al.* (2006) "A bivalent chromatin structure marks key developmental genes in embryonic stem cells," *Cell*, 125(2), pp. 315–326. doi:10.1016/j.cell.2006.02.041.

Bérubé, N.G., Smeenk, C.A. and Picketts, D.J. (2000) "Cell cycle-dependent phosphorylation of the ATRX protein correlates with changes in nuclear matrix and chromatin association," *Human molecular genetics*, 9(4), pp. 539–547. doi:10.1093/hmg/9.4.539.

Beyer, S. *et al.* (2016) "Canonical Wnt signalling regulates nuclear export of Setdb1 during skeletal muscle terminal differentiation," *Cell discovery*, 2(1), p. 16037. doi:10.1038/celldisc.2016.37.

Bilodeau, S. *et al.* (2009) "SetDB1 contributes to repression of genes encoding developmental regulators and maintenance of ES cell state," *Genes & development*, 23(21), pp. 2484–2489. doi:10.1101/gad.1837309.

Bloom, K.S. and Anderson, J.N. (1978) "Fractionation of hen oviduct chromatin into transcriptionally active and inactive regions after selective micrococcal nuclease digestion," *Cell*, 15(1), pp. 141–150. doi:10.1016/0092-8674(78)90090-9.

Bogdanović, O. and Lister, R. (2017) "DNA methylation and the preservation of cell identity," *Current opinion in genetics & development*, 46, pp. 9–14. doi:10.1016/j.gde.2017.06.007.

Boiani, M. and Schöler, H.R. (2005) "Regulatory networks in embryo-derived pluripotent stem cells," *Nature reviews. Molecular cell biology*, 6(11), pp. 872–884. doi:10.1038/nrm1744.

Boroviak, T. *et al.* (2014) "The ability of inner-cell-mass cells to self-renew as embryonic stem cells is acquired following epiblast specification," *Nature cell biology*, 16(6), pp. 513–525. doi:10.1038/ncb2965.

Boroviak, T. *et al.* (2015) "Lineage-specific profiling delineates the emergence and progression of naive pluripotency in mammalian embryogenesis," *Developmental cell*, 35(3), pp. 366–382. doi:10.1016/j.devcel.2015.10.011.

Boutilier, A.J. and Elsawa, S.F. (2021) "Macrophage polarization states in the tumor microenvironment," *International journal of molecular sciences*, 22(13), p. 6995. doi:10.3390/ijms22136995.

Boyer, L.A. *et al.* (2005) "Core transcriptional regulatory circuitry in human embryonic stem cells," *Cell*, 122(6), pp. 947–956. doi:10.1016/j.cell.2005.08.020.

Branon, T.C. *et al.* (2018) “Efficient proximity labeling in living cells and organisms with TurboID,” *Nature biotechnology*, 36(9), pp. 880–887. doi:10.1038/nbt.4201.

Briggs, R. and King, T.J. (1952) “Transplantation of living nuclei from blastula cells into enucleated frogs’ eggs,” *Proceedings of the National Academy of Sciences of the United States of America*, 38(5), pp. 455–463. doi:10.1073/pnas.38.5.455.

Brinkman, A.B. *et al.* (2012) “Sequential ChIP-bisulfite sequencing enables direct genome-scale investigation of chromatin and DNA methylation cross-talk,” *Genome research*, 22(6), pp. 1128–1138. doi:10.1101/gr.133728.111.

Brons, I.G.M. *et al.* (2007) “Derivation of pluripotent epiblast stem cells from mammalian embryos,” *Nature*, 448(7150), pp. 191–195. doi:10.1038/nature05950.

Bruderer, R. *et al.* (2015) “Extending the limits of quantitative proteome profiling with data-independent acquisition and application to acetaminophen-treated three-dimensional liver microtissues,” *Molecular & cellular proteomics: MCP*, 14(5), pp. 1400–1410. doi:10.1074/mcp.M114.044305.

Bruderer, R. *et al.* (2017) “Optimization of Experimental Parameters in Data-Independent Mass Spectrometry Significantly Increases Depth and Reproducibility of Results,” *Molecular & cellular proteomics: MCP*, 16(12), pp. 2296–2309. doi:10.1074/mcp.RA117.000314.

Buecker, C. *et al.* (2014) “Reorganization of enhancer patterns in transition from naive to primed pluripotency,” *Cell stem cell*, 14(6), pp. 838–853. doi:10.1016/j.stem.2014.04.003.

Bultmann, S. *et al.* (2012) “Targeted transcriptional activation of silent oct4 pluripotency gene by combining designer TALEs and inhibition of epigenetic modifiers,” *Nucleic acids research*, 40(12), pp. 5368–5377. doi:10.1093/nar/gks199.

Bulut-Karslioglu, A. *et al.* (2014) “Suv39h-dependent H3K9me3 marks intact retrotransposons and silences LINE elements in mouse embryonic stem cells,” *Molecular cell*, 55(2), pp. 277–290. doi:10.1016/j.molcel.2014.05.029.

Burdon, T. *et al.* (1999) “Suppression of SHP-2 and ERK signalling promotes self-renewal of mouse embryonic stem cells,” *Developmental biology*, 210(1), pp. 30–43. doi:10.1006/dbio.1999.9265.

Carlson, L.L., Page, A.W. and Bestor, T.H. (1992) “Properties and localization of DNA methyltransferase in preimplantation mouse embryos: implications for genomic imprinting,” *Genes & development*, 6(12B), pp. 2536–2541. doi:10.1101/gad.6.12b.2536.

Catherman, A.D., Skinner, O.S. and Kelleher, N.L. (2014) “Top Down proteomics: facts and perspectives,” *Biochemical and biophysical research communications*, 445(4), pp. 683–693. doi:10.1016/j.bbrc.2014.02.041.

Chait, B.T. (2006) “Chemistry. Mass spectrometry: bottom-up or top-down?,” *Science*, pp. 65–66. doi:10.1126/science.1133987.

Chambers, I. *et al.* (2003) “Functional expression cloning of Nanog, a pluripotency sustaining factor in embryonic stem cells,” *Cell*, 113(5), pp. 643–655. doi:10.1016/s0092-8674(03)00392-1.

Chen, X. *et al.* (2008) “Integration of external signaling pathways with the core transcriptional network in embryonic stem cells,” *Cell*, 133(6), pp. 1106–1117. doi:10.1016/j.cell.2008.04.043.

Chen, Z.A. and Rappsilber, J. (2023) “Protein structure dynamics by crosslinking mass spectrometry,” *Current opinion in structural biology*, 80(102599), p. 102599. doi:10.1016/j.sbi.2023.102599.

Cho, K.F. *et al.* (2020) “Split-TurboID enables contact-dependent proximity labeling in cells,” *Proceedings of the National Academy of Sciences of the United States of America*, 117(22), pp. 12143–12154. doi:10.1073/pnas.1919528117.

Cho, N.H. *et al.* (2022) “OpenCell: Endogenous tagging for the cartography of human cellular organization,” *Science*, 375(6585), p. eabi6983. doi:10.1126/science.abi6983.

Cho, S., Park, J.S. and Kang, Y.-K. (2013) “Regulated nuclear entry of over-expressed Setdb1,” *Genes to cells: devoted to molecular & cellular mechanisms*, 18(8), pp. 694–703. doi:10.1111/gtc.12068.

Chou, D.M. *et al.* (2010) “A chromatin localization screen reveals poly (ADP ribose)-regulated recruitment of the repressive polycomb and NuRD complexes to sites of DNA damage,” *Proceedings of the National Academy of Sciences of the United States of America*, 107(43), pp. 18475–18480. doi:10.1073/pnas.1012946107.

Choudhury, S. *et al.* (2023) “Generation of a transgenic mouse embryonic stem cell line overexpressing DNMT1,” *Stem cell research*, 71(103141), p. 103141. doi:10.1016/j.scr.2023.103141.

Chrysanthou, S. *et al.* (2022) “The DNA dioxygenase Tet1 regulates H3K27 modification and embryonic stem cell biology independent of its catalytic activity,” *Nucleic acids research*, 50(6), pp. 3169–3189. doi:10.1093/nar/gkac089.

Chu, D.S. *et al.* (2006) “Sperm chromatin proteomics identifies evolutionarily conserved fertility factors,” *Nature*, 443(7107), pp. 101–105. doi:10.1038/nature05050.

Churchill, F.B. (1974) “William Johannsen and the genotype concept,” *Journal of the history of biology*, 7(1), pp. 5–30. doi:10.1007/bf00179291.

Colaert, N. *et al.* (2011) “Analysis of the resolution limitations of peptide identification algorithms,” *Journal of proteome research*, 10(12), pp. 5555–5561. doi:10.1021/pr200913a.

Cortellino, S. *et al.* (2011) "Thymine DNA glycosylase is essential for active DNA demethylation by linked deamination-base excision repair," *Cell*, 146(1), pp. 67–79. doi:10.1016/j.cell.2011.06.020.

Cox, J. *et al.* (2011) "Andromeda: a peptide search engine integrated into the MaxQuant environment," *Journal of proteome research*, 10(4), pp. 1794–1805. doi:10.1021/pr101065j.

Craig, R. and Beavis, R.C. (2004) "TANDEM: matching proteins with tandem mass spectra," *Bioinformatics (Oxford, England)*, 20(9), pp. 1466–1467. doi:10.1093/bioinformatics/bth092.

Crick, F. and Anderson, P.W. (1989) "What mad pursuit: A personal view of scientific discovery," *Physics today*, 42(7), pp. 68–70. doi:10.1063/1.2811088.

Crutchfield, C.A. *et al.* (2016) "Advances in mass spectrometry-based clinical biomarker discovery," *Clinical proteomics*, 13(1), p. 1. doi:10.1186/s12014-015-9102-9.

Dahlet, T. *et al.* (2020) "Genome-wide analysis in the mouse embryo reveals the importance of DNA methylation for transcription integrity," *Nature communications*, 11(1), p. 3153. doi:10.1038/s41467-020-16919-w.

Danese, A. *et al.* (2021) "EpiScanpy: integrated single-cell epigenomic analysis," *Nature communications*, 12(1), p. 5228. doi:10.1038/s41467-021-25131-3.

Davidson, K.C., Mason, E.A. and Pera, M.F. (2015) "The pluripotent state in mouse and human," *Development*, 142(18), pp. 3090–3099. doi:10.1242/dev.116061.

Davis, M.T. *et al.* (1995) "A microscale electrospray interface for online, capillary liquid chromatography/tandem mass spectrometry of complex peptide mixtures," *Analytical chemistry*, 67(24), pp. 4549–4556. doi:10.1021/ac00120a019.

Davis, R.L., Weintraub, H. and Lassar, A.B. (1987) "Expression of a single transfected cDNA converts fibroblasts to myoblasts," *Cell*, 51(6), pp. 987–1000. doi:10.1016/0092-8674(87)90585-x.

De Los Angeles, A. *et al.* (2015) "Hallmarks of pluripotency," *Nature*, 525(7570), pp. 469–478. doi:10.1038/nature15515.

Dean, W. *et al.* (2001) "Conservation of methylation reprogramming in mammalian development: aberrant reprogramming in cloned embryos," *Proceedings of the National Academy of Sciences of the United States of America*, 98(24), pp. 13734–13738. doi:10.1073/pnas.241522698.

Déjardin, J. (2015) "Switching between epigenetic states at pericentromeric heterochromatin," *Trends in genetics: TIG*, 31(11), pp. 661–672. doi:10.1016/j.tig.2015.09.003.

Déjardin, J. and Kingston, R.E. (2009) "Purification of proteins associated with specific genomic Loci," *Cell*, 136(1), pp. 175–186. doi:10.1016/j.cell.2008.11.045.

Demichev, V. *et al.* (2020) "DIA-NN: neural networks and interference correction enable deep proteome coverage in high throughput," *Nature methods*, 17(1), pp. 41–44. doi:10.1038/s41592-019-0638-x.

Demichev, V. *et al.* (2022) "dia-PASEF data analysis using FragPipe and DIA-NN for deep proteomics of low sample amounts," *Nature communications*, 13(1), p. 3944. doi:10.1038/s41467-022-31492-0.

Dierolf, J.G. *et al.* (2022) "Modulation of PKM1/2 levels by steric blocking morpholinos alters the metabolic and pluripotent state of Murine pluripotent stem cells," *Stem cells and development*, 31(11–12), pp. 278–295. doi:10.1089/scd.2021.0347.

Dixon, G. *et al.* (2021) "QSER1 protects DNA methylation valleys from de novo methylation," *Science*, 372(6538), p. eabd0875. doi:10.1126/science.abd0875.

Dodge, J.E. *et al.* (2004) "Histone H3-K9 methyltransferase ESET is essential for early development," *Molecular and cellular biology*, 24(6), pp. 2478–2486. doi:10.1128/MCB.24.6.2478-2486.2004.

Dodonova, S.O. *et al.* (2020) "Nucleosome-bound SOX2 and SOX11 structures elucidate pioneer factor function," *Nature*, 580(7805), pp. 669–672. doi:10.1038/s41586-020-2195-y.

Doskocil, J. and Sorm, F. (1962) "Distribution of 5-methylcytosine in pyrimidine sequences of deoxyribonucleic acids," *Biochimica et biophysica acta*, 55(6), pp. 953–959. doi:10.1016/0006-3002(62)90909-5.

Ducret, A. *et al.* (1998) "High throughput protein characterization by automated reverse-phase chromatography/electrospray tandem mass spectrometry," *Protein science: a publication of the Protein Society*, 7(3), pp. 706–719. doi:10.1002/pro.5560070320.

Dunican, D.S. *et al.* (2020) "Bivalent promoter hypermethylation in cancer is linked to the H327me3/H3K4me3 ratio in embryonic stem cells," *BMC biology*, 18(1), p. 25. doi:10.1186/s12915-020-0752-3.

Dunn, S.-J. *et al.* (2014) "Defining an essential transcription factor program for naïve pluripotency," *Science*, 344(6188), pp. 1156–1160. doi:10.1126/science.1248882.

Dutta, B. *et al.* (2012) "Elucidating the temporal dynamics of chromatin-associated protein release upon DNA digestion by quantitative proteomic approach," *Journal of proteomics*, 75(17), pp. 5493–5506. doi:10.1016/j.jprot.2012.06.030.

Eckersley-Maslin, M.A., Alda-Catalinas, C. and Reik, W. (2018) "Dynamics of the epigenetic landscape during the maternal-to-zygotic transition," *Nature reviews. Molecular cell biology*, 19(7), pp. 436–450. doi:10.1038/s41580-018-0008-z.

Edman, P. *et al.* (1950) "Method for determination of the amino acid sequence in peptides," *Acta chemica Scandinavica (Copenhagen, Denmark: 1989)*, 4, pp. 283–293. doi:10.3891/acta.chem.scand.04-0283.

Eggers, N. and Becker, P.B. (2021) "Cell-free genomics reveal intrinsic, cooperative and competitive determinants of chromatin interactions," *Nucleic acids research*, 49(13), pp. 7602–7617. doi:10.1093/nar/gkab558.

El-Aneed, A., Cohen, A. and Banoub, J. (2009) "Mass spectrometry, review of the basics: Electrospray, MALDI, and commonly used mass analyzers," *Applied spectroscopy reviews*, 44(3), pp. 210–230. doi:10.1080/05704920902717872.

Emani, M.R. *et al.* (2015) "The L1TD1 protein interactome reveals the importance of post-transcriptional regulation in human pluripotency," *Stem cell reports*, 4(3), pp. 519–528. doi:10.1016/j.stemcr.2015.01.014.

Eng, J.K., McCormack, A.L. and Yates, J.R. (1994) "An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database," *Journal of the American Society for Mass Spectrometry*, 5(11), pp. 976–989. doi:10.1016/1044-0305(94)80016-2.

Evans, M.J. and Kaufman, M.H. (1981) "Establishment in culture of pluripotential cells from mouse embryos," *Nature*, 292(5819), pp. 154–156. doi:10.1038/292154a0.

Factor, D.C. *et al.* (2014) "Epigenomic comparison reveals activation of 'seed' enhancers during transition from naive to primed pluripotency," *Cell stem cell*, 14(6), pp. 854–863. doi:10.1016/j.stem.2014.05.005.

Falkenby, L.G. *et al.* (2014) "Integrated solid-phase extraction-capillary liquid chromatography (speLC) interfaced to ESI-MS/MS for fast characterization and quantification of protein and proteomes," *Journal of proteome research*, 13(12), pp. 6169–6175. doi:10.1021/pr5008575.

Faust, C. *et al.* (1998) "The Polycomb-group gene eed is required for normal morphogenetic movements during gastrulation in the mouse embryo," *Development (Cambridge, England)*, 125(22), pp. 4495–4506. doi:10.1242/dev.125.22.4495.

Federation, A.J. *et al.* (2020) "Highly parallel quantification and compartment localization of transcription factors and nuclear proteins," *Cell reports*, 30(8), pp. 2463-2471.e5. doi:10.1016/j.celrep.2020.01.096.

Fenn, J.B. *et al.* (1989) "Electrospray ionization for mass spectrometry of large biomolecules," *Science*, 246(4926), pp. 64–71. doi:10.1126/science.2675315.

Fernandez de la Mora, J. (2000) "Electrospray ionization of large multiply charged species proceeds via Dole's charged residue mechanism," *Analytica chimica acta*, 406(1), pp. 93–104. doi:10.1016/s0003-2670(99)00601-7.

Fernandez-Lima, F.A., Kaplan, D.A. and Park, M.A. (2011) "Note: Integration of trapped ion mobility spectrometry with mass spectrometry," *The Review of scientific instruments*, 82(12), p. 126106. doi:10.1063/1.3665933.

Festuccia, N. *et al.* (2016) “Mitotic binding of Esrrb marks key regulatory regions of the pluripotency network,” *Nature cell biology*, 18(11), pp. 1139–1148. doi:10.1038/ncb3418.

Festuccia, N. *et al.* (2019) “Transcription factor activity and nucleosome organization in mitosis,” *Genome research*, 29(2), pp. 250–260. doi:10.1101/gr.243048.118.

Festuccia, N., Gonzalez, I. and Navarro, P. (2017) “The epigenetic paradox of pluripotent ES cells,” *Journal of molecular biology*, 429(10), pp. 1476–1503. doi:10.1016/j.jmb.2016.12.009.

Ficz, G. *et al.* (2013) “FGF signaling inhibition in ESCs drives rapid genome-wide demethylation to the epigenetic ground state of pluripotency,” *Cell stem cell*, 13(3), pp. 351–359. doi:10.1016/j.stem.2013.06.004.

Figeys, D. and Aebersold, R. (1998) “Nanoflow solvent gradient delivery from a microfabricated device for protein identifications by electrospray ionization mass spectrometry,” *Analytical chemistry*, 70(18), pp. 3721–3727. doi:10.1021/ac980502j.

Fiorentino, J., Torres-Padilla, M.-E. and Scialdone, A. (2020) “Measuring and modeling single-cell heterogeneity and fate decision in mouse embryos,” *Annual review of genetics*, 54(1), pp. 167–187. doi:10.1146/annurev-genet-021920-110200.

Fischer, E. (1906) “Untersuchungen über Aminosäuren, Polypeptide und Proteïne,” *Berichte der Deutschen Chemischen Gesellschaft*, 39(1), pp. 530–610. doi:10.1002/cber.19060390190.

Fleck, J.S. *et al.* (2022) “Inferring and perturbing cell fate regulomes in human brain organoids,” *Nature* [Preprint]. doi:10.1038/s41586-022-05279-8.

Fu, Y. *et al.* (2020) “Genome-wide identification of FHL1 as a powerful prognostic candidate and potential therapeutic target in acute myeloid leukaemia,” *EBioMedicine*, 52(102664), p. 102664. doi:10.1016/j.ebiom.2020.102664.

Gao, X.D. *et al.* (2018) “C-BERST: defining subnuclear proteomic landscapes at genomic elements with dCas9-APEX2,” *Nature methods*, 15(6), pp. 433–436. doi:10.1038/s41592-018-0006-2.

Gao, Y. *et al.* (2013) “Replacement of Oct4 by Tet1 during iPSC induction reveals an important role of DNA methylation and hydroxymethylation in reprogramming,” *Cell stem cell*, 12(4), pp. 453–469. doi:10.1016/j.stem.2013.02.005.

Gaspar-Maia, A. *et al.* (2011) “Open chromatin in pluripotency and reprogramming,” *Nature reviews. Molecular cell biology*, 12(1), pp. 36–47. doi:10.1038/nrm3036.

Gassmann, R., Henzing, A.J. and Earnshaw, W.C. (2005) “Novel components of human mitotic chromosomes identified by proteomic analysis of the chromosome scaffold fraction," *Chromosoma*, 113(7), pp. 385–397. doi:10.1007/s00412-004-0326-0.

Gauchier, M. *et al.* (2020) "Purification and enrichment of specific chromatin loci," *Nature methods*, 17(4), pp. 380–389. doi:10.1038/s41592-020-0765-4.

The Gene Ontology Consortium (2019) "The Gene Ontology Resource: 20 years and still GOing strong," *Nucleic acids research*, 47(D1), pp. D330–D338. doi:10.1093/nar/gky1055.

Geng, T., Zhang, D. and Jiang, W. (2019) "Epigenetic regulation of transition among different pluripotent states: Concise review," *Stem cells (Dayton, Ohio)*, 37(11), pp. 1372–1380. doi:10.1002/stem.3064.

Gentek, R., Molawi, K. and Sieweke, M.H. (2014) "Tissue macrophage identity and self-renewal," *Immunological reviews*, 262(1), pp. 56–73. doi:10.1111/imr.12224.

Gessulat, S. *et al.* (2019) "Prosit: proteome-wide prediction of peptide tandem mass spectra by deep learning," *Nature methods*, 16(6), pp. 509–518. doi:10.1038/s41592-019-0426-7.

Geuder, J. *et al.* (2021) "A non-invasive method to generate induced pluripotent stem cells from primate urine," *Scientific reports*, 11(1), p. 3516. doi:10.1038/s41598-021-82883-0.

Giansanti, P. *et al.* (2016) "Six alternative proteases for mass spectrometry-based proteomics beyond trypsin," *Nature protocols*, 11(5), pp. 993–1006. doi:10.1038/nprot.2016.057.

Gillet, L.C. *et al.* (2012) "Targeted data extraction of the MS/MS spectra generated by data-independent acquisition: a new concept for consistent and accurate proteome analysis," *Molecular & cellular proteomics: MCP*, 11(6), p. O111.016717. doi:10.1074/mcp.O111.016717.

Gillet, L.C., Leitner, A. and Aebersold, R. (2016) "Mass Spectrometry Applied to Bottom-Up Proteomics: Entering the High-Throughput Era for Hypothesis Testing," *Annual review of analytical chemistry* , 9(1), pp. 449–472. doi:10.1146/annurev-anchem-071015-041535.

Gingras, A.-C. *et al.* (2007) "Analysis of protein complexes using mass spectrometry," *Nature reviews. Molecular cell biology*, 8(8), pp. 645–654. doi:10.1038/nrm2208.

Ginno, P.A. *et al.* (2018) "Cell cycle-resolved chromatin proteomics reveals the extent of mitotic preservation of the genomic regulatory landscape," *Nature communications*, 9(1), p. 4048. doi:10.1038/s41467-018-06007-5.

Glenn, W.E. (1952) *A Time of Flight Mass Spectrograph (thesis)*. U.S. Atomic Energy Commission, Technical Information Service. Available at: https://play.google.com/store/books/details?id=9ynecHgVjAsC.

Golghalyani, V. *et al.* (2017) "ArgC-like digestion: Complementary or alternative to tryptic digestion?," *Journal of proteome research*, 16(2), pp. 978–987. doi:10.1021/acs.jproteome.6b00921.

Gonzales-Cope, M. *et al.* (2016) “Histone H4 acetylation and the epigenetic reader Brd4 are critical regulators of pluripotency in embryonic stem cells,” *BMC genomics*, 17, p. 95. doi:10.1186/s12864-016-2414-y.

Gonzalez-Covarrubias, V., Martínez-Martínez, E. and Del Bosque-Plata, L. (2022) “The potential of metabolomics in biomedical applications,” *Metabolites*, 12(2), p. 194. doi:10.3390/metabo12020194.

Görg, A. *et al.* (1998) “Two-dimensional electrophoresis of proteins in an immobilized pH 4-12 gradient,” *Electrophoresis*, 19(8–9), pp. 1516–1519. doi:10.1002/elps.1150190850.

Greenberg, M.V.C. and Bourc’his, D. (2019) “The diverse roles of DNA methylation in mammalian development and disease,” *Nature reviews. Molecular cell biology*, 20(10), pp. 590–607. doi:10.1038/s41580-019-0159-6.

Groh, S. and Schotta, G. (2017) “Silencing of endogenous retroviruses by heterochromatin,” *Cellular and molecular life sciences: CMLS*, 74(11), pp. 2055–2065. doi:10.1007/s00018-017-2454-8.

Guo, F. *et al.* (2014) “Active and passive demethylation of male and female pronuclear DNA in the mammalian zygote,” *Cell stem cell*, 15(4), pp. 447–459. doi:10.1016/j.stem.2014.08.003.

Guo, G. *et al.* (2021) “Human naive epiblast cells possess unrestricted lineage potential,” *Cell Stem Cell*, pp. 1040-1056.e6. doi:10.1016/j.stem.2021.02.025.

Gurdon, J.B., Elsdale, T.R. and Fischberg, M. (1958) “Sexually mature individuals of Xenopus laevis from the transplantation of single somatic nuclei,” *Nature*, 182(4627), pp. 64–65. doi:10.1038/182064a0.

Hackett, J.A. *et al.* (2013) “Synergistic mechanisms of DNA demethylation during transition to ground-state pluripotency,” *Stem cell reports*, 1(6), pp. 518–531. doi:10.1016/j.stemcr.2013.11.010.

Hackett, J.A. and Surani, M.A. (2014) “Regulatory principles of pluripotency: from the ground state up,” *Cell stem cell*, 15(4), pp. 416–430. doi:10.1016/j.stem.2014.09.015.

Han, D. *et al.* (2022) “A balanced Oct4 interactome is crucial for maintaining pluripotency,” *Science advances*, 8(7), p. eabe4375. doi:10.1126/sciadv.abe4375.

Han, Y. *et al.* (2019) “Directed evolution of split APEX2 peroxidase,” *ACS chemical biology*, 14(4), pp. 619–635. doi:10.1021/acschembio.8b00919.

Hardman, M. and Makarov, A.A. (2003) “Interfacing the orbitrap mass analyzer to an electrospray ion source,” *Analytical chemistry*, 75(7), pp. 1699–1705. doi:10.1021/ac0258047.

Hasin, Y., Seldin, M. and Lusis, A. (2017) “Multi-omics approaches to disease,” *Genome biology*, 18(1), pp. 1–15. doi:10.1186/s13059-017-1215-1.

Hayashi, K. *et al.* (2008) “Dynamic equilibrium and heterogeneity of mouse pluripotent stem cells with distinct functional and epigenetic states,” *Cell stem cell*, 3(4), pp. 391–401. doi:10.1016/j.stem.2008.07.027.

Hayashi, K. *et al.* (2011) “Reconstitution of the mouse germ cell specification pathway in culture by pluripotent stem cells,” *Cell*, 146(4), pp. 519–532. doi:10.1016/j.cell.2011.06.052.

He, Y.-F. *et al.* (2011) “Tet-mediated formation of 5-carboxylcytosine and its excision by TDG in mammalian DNA,” *Science*, 333(6047), pp. 1303–1307. doi:10.1126/science.1210944.

Heil, L.R. *et al.* (2023) “Evaluating the performance of the Astral mass analyzer for quantitative proteomics using data independent acquisition,” *bioRxiv.* (2023-06). doi:10.1101/2023.06.03.543570.

Hein, M.Y. *et al.* (2013) “Proteomic Analysis of Cellular Systems,” in *Handbook of Systems Biology.* Elsevier, pp. 3–25. doi:10.1016/b978-0-12-385944-0.00001-0.

Hein, M.Y. *et al.* (2015) “A human interactome in three quantitative dimensions organized by stoichiometries and abundances,” *Cell*, 163(3), pp. 712–723. doi:10.1016/j.cell.2015.09.053.

Henikoff, S. *et al.* (2009) “Genome-wide profiling of salt fractions maps physical properties of chromatin,” *Genome research*, 19(3), pp. 460–469. doi:10.1101/gr.087619.108.

Hillenkamp, F. *et al.* (1991) “Matrix-assisted laser desorption/ionization mass spectrometry of biopolymers,” *Analytical chemistry*, 63(24), pp. 1193A-1203A. doi:10.1021/ac00024a716.

Hillman, N., Sherman, M.I. and Graham, C. (1972) “The effect of spatial arrangement on cell determination during mouse development,” *Journal of embryology and experimental morphology*, 28(2), pp. 263–278. Available at: https://www.ncbi.nlm.nih.gov/pubmed/4674567.

Hoffman, E.A. *et al.* (2015) “Formaldehyde crosslinking: a tool for the study of chromatin complexes,” *The Journal of biological chemistry*, 290(44), pp. 26404–26411. doi:10.1074/jbc.R115.651679.

Hofmeister, F. (1902) “Über Bau und Gruppierung der Eiweisskörper,” *Ergebnisse der Physiologie*, 1(1), pp. 759–802. doi:10.1007/bf02188398.

Howlett, S.K. and Reik, W. (1991) “Methylation levels of maternal and paternal genomes during preimplantation development,” *Development (Cambridge, England)*, 113(1), pp. 119–127. doi:10.1242/dev.113.1.119.

Huang, H. *et al.* (2015) “Quantitative proteomic analysis of histone modifications,” *Chemical reviews*, 115(6), pp. 2376–2418. doi:10.1021/cr500491u.

Huang, X. and Wang, J. (2014) “The extended pluripotency protein interactome and its links to reprogramming,” *Current opinion in genetics & development*, 28, pp. 16–24. doi:10.1016/j.gde.2014.08.003.

Huang, Y. *et al.* (2012) “In Vivo differentiation potential of epiblast stem cells revealed by chimeric embryo formation,” *Cell reports*, 2(6), pp. 1571–1578. doi:10.1016/j.celrep.2012.10.022.

Huertas, J. *et al.* (2020) "Nucleosomal DNA dynamics mediate Oct4 pioneer factor binding," *Biophysical journal*, 118(9), pp. 2280–2296. doi:10.1016/j.bpj.2019.12.038.

Hunt, D.F. *et al.* (1981) "Sequence analysis of polypeptides by collision activated dissociation on a triple quadrupole mass spectrometer," *Biomedical mass spectrometry*, 8(9), pp. 397–408. doi:10.1002/bms.1200080909.

Hunt, D.F. *et al.* (1992) "Characterization of peptides bound to the class I MHC molecule HLA-A2.1 by mass spectrometry," *Science*, pp. 1261–1263. doi:10.1126/science.1546328.

Iizuka, M. *et al.* (2009) "Histone acetyltransferase Hbo1: catalytic activity, cellular abundance, and links to primary cancers," *Gene*, 436(1–2), pp. 108–114. doi:10.1016/j.gene.2009.01.020.

Imhof, A. and Bonaldi, T. (2005) "'Chromatomics' the analysis of the chromatome," *Molecular bioSystems*, 1(2), pp. 112–116. doi:10.1039/B502845K.

Issaq, H.J. *et al.* (2002) "Methods for fractionation, separation and profiling of proteins and peptides," *Electrophoresis*, 23(17), pp. 3048–3061. doi:10.1002/1522-2683(200209)23:17<3048::AID-ELPS3048>3.0.CO;2-L.

Ito, S. *et al.* (2011) "Tet proteins can convert 5-methylcytosine to 5-formylcytosine and 5-carboxylcytosine," *Science*, 333(6047), pp. 1300–1303. doi:10.1126/science.1210597.

Ivanova, E. *et al.* (2020) "DNA methylation changes during preimplantation development reveal inter-species differences and reprogramming events at imprinted genes," *Clinical epigenetics*, 12(1), p. 64. doi:10.1186/s13148-020-00857-x.

Jain, S.U. *et al.* (2020) "H3 K27M and EZHIP Impede H3K27-Methylation Spreading by Inhibiting Allosterically Stimulated PRC2," *Molecular cell*, 80(4), pp. 726-735.e7. doi:10.1016/j.molcel.2020.09.028.

James, D. *et al.* (2006) "Contribution of human embryonic stem cells to mouse blastocysts," *Developmental biology*, 295(1), pp. 90–102. doi:10.1016/j.ydbio.2006.03.026.

Janssen, K.A. *et al.* (2019) "Quantitation of single and combinatorial histone modifications by integrated chromatography of bottom-up peptides and middle-down polypeptide tails," *Journal of the American Society for Mass Spectrometry*, 30(12), pp. 2449–2459. doi:10.1007/s13361-019-02303-6.

Jerabek, S. *et al.* (2014) "OCT4: dynamic DNA binding pioneers stem cell pluripotency," *Biochimica et biophysica acta*, 1839(3), pp. 138–154. doi:10.1016/j.bbagrm.2013.10.001.

Ji, X. *et al.* (2015) "Chromatin proteomic profiling reveals novel proteins associated with histone-marked genomic regions," *Proceedings of the*

National Academy of Sciences of the United States of America, 112(12), pp. 3841–3846. doi:10.1073/pnas.1502971112.

Joshi, O. *et al.* (2015) "Dynamic Reorganization of Extremely Long-Range Promoter-Promoter Interactions between Two States of Pluripotency," *Cell stem cell*, 17(6), pp. 748–757. doi:10.1016/j.stem.2015.11.010.

Jukam, D., Shariati, S.A.M. and Skotheim, J.M. (2017) "Zygotic genome activation in vertebrates," *Developmental cell*, 42(4), pp. 316–332. doi:10.1016/j.devcel.2017.07.026.

Kalkan, T. *et al.* (2017) "Tracking the embryonic stem cell transition from ground state pluripotency," *Development* , 144(7), pp. 1221–1234. doi:10.1242/dev.142711.

Kanton, S. *et al.* (2019) "Organoid single-cell genomic atlas uncovers human-specific features of brain development," *Nature*, 574(7778), pp. 418–422. doi:10.1038/s41586-019-1654-9.

Karas, M. and Hillenkamp, F. (1988) "Laser desorption ionization of proteins with molecular masses exceeding 10,000 daltons," *Analytical chemistry*, 60(20), pp. 2299–2301. doi:10.1021/ac00171a028.

Kasvandik, S. *et al.* (2019) "Cell cycle-balanced expression of pluripotency regulators via cyclin-dependent kinase 1," *bioRxiv*. bioRxiv (764639). doi:10.1101/764639.

Kebarle, P. (2000) "A brief overview of the present status of the mechanisms involved in electrospray mass spectrometry," *Journal of mass spectrometry*, 35(7), pp. 804–817. doi:10.1002/1096-9888(200007)35:7<804::AID-JMS22>3.0.CO;2-Q.

Kebarle, P. and Verkerk, U.H. (2009) "Electrospray: from ions in solution to ions in the gas phase, what we know now," *Mass spectrometry reviews*, 28(6), pp. 898–917. doi:10.1002/mas.20247.

Kelstrup, C.D. *et al.* (2018) "Performance evaluation of the Q Exactive HF-X for shotgun proteomics," *Journal of proteome research*, 17(1), pp. 727–738. doi:10.1021/acs.jproteome.7b00602.

Kempf, J.M. *et al.* (2021) "Loss-of-function mutations in the histone methyltransferase EZH2 promote chemotherapy resistance in AML," *Scientific reports*, 11(1), p. 5838. doi:10.1038/s41598-021-84708-6.

Khoudoli, G.A. *et al.* (2008) "Temporal profiling of the chromatin proteome reveals system-wide responses to replication inhibition," *Current biology: CB*, 18(11), pp. 838–843. doi:10.1016/j.cub.2008.04.075.

Kim, K. *et al.* (2007) "Recombination signatures distinguish embryonic stem cells derived by parthenogenesis and somatic cell nuclear transfer," *Cell stem cell*, 1(3), pp. 346–352. doi:10.1016/j.stem.2007.07.001.

King, H.W. and Klose, R.J. (2017) "The pioneer factor OCT4 requires the chromatin remodeller BRG1 to support gene regulatory element function in mouse embryonic stem cells," *eLife*, 6, p. e22631. doi:10.7554/elife.22631.

Kinoshita, M. *et al.* (2021) “Capture of mouse and human stem cells with features of formative pluripotency,” *Cell stem cell*, 28(3), pp. 453-471.e8. doi:10.1016/j.stem.2020.11.005.

Kleiner, R.E. *et al.* (2018) “A chemical proteomics approach to reveal direct protein-protein interactions in living cells,” *Cell chemical biology*, 25(1), pp. 110-120.e3. doi:10.1016/j.chembiol.2017.10.001.

Klemm, S.L., Shipony, Z. and Greenleaf, W.J. (2019) “Chromatin accessibility and the regulatory epigenome,” *Nature reviews. Genetics*, 20(4), pp. 207–220. doi:10.1038/s41576-018-0089-8.

Kliszczak, A.E. *et al.* (2011) “DNA mediated chromatin pull-down for the study of chromatin replication,” *Scientific reports*, 1, p. 95. doi:10.1038/srep00095.

Kloet, S.L. *et al.* (2016) “The dynamic interactome and genomic targets of Polycomb complexes during stem-cell differentiation,” *Nature structural & molecular biology*, 23(7), pp. 682–690. doi:10.1038/nsmb.3248.

Klykov, O. *et al.* (2018) “Efficient and robust proteome-wide approaches for cross-linking mass spectrometry,” *Nature protocols*, 13(12), pp. 2964–2990. doi:10.1038/s41596-018-0074-x.

Kojima, Y. *et al.* (2014) “The transcriptional and functional properties of mouse epiblast stem cells resemble the anterior primitive streak,” *Cell stem cell*, 14(1), pp. 107–120. doi:10.1016/j.stem.2013.09.014.

Korthout, T. *et al.* (2018) “Decoding the chromatin proteome of a single genomic locus by DNA sequencing,” *PLoS biology*, 16(7), p. e2005542. doi:10.1371/journal.pbio.2005542.

Kulak, N.A., Geyer, P.E. and Mann, M. (2017) “Loss-less Nano-fractionator for High Sensitivity, High Coverage Proteomics,” *Molecular & cellular proteomics: MCP*, 16(4), pp. 694–705. doi:10.1074/mcp.O116.065136.

Kulej, K. *et al.* (2017) “Time-resolved Global and Chromatin Proteomics during Herpes Simplex Virus Type 1 (HSV-1) Infection,” *Molecular & cellular proteomics: MCP*, 16(4 suppl 1), pp. S92–S107. doi:10.1074/mcp.M116.065987.

Kunath, T. *et al.* (2007) “FGF stimulation of the Erk1/2 signalling cascade triggers transition of pluripotent embryonic stem cells from self-renewal to lineage commitment,” *Development (Cambridge, England)*, 134(16), pp. 2895–2902. doi:10.1242/dev.02880.

Kurimoto, K. *et al.* (2015) “Quantitative Dynamics of Chromatin Remodeling during Germ Cell Specification from Mouse Embryonic Stem Cells,” *Cell stem cell*, 16(5), pp. 517–532. doi:10.1016/j.stem.2015.03.002.

Kustatscher, G., Wills, K.L.H., *et al.* (2014) “Chromatin enrichment for proteomics,” *Nature protocols*, 9(9), pp. 2090–2099. doi:10.1038/nprot.2014.142.

Kustatscher, G., Hégarat, N., *et al.* (2014) “Proteomics of a fuzzy organelle: interphase chromatin,” *The EMBO journal*, 33(6), pp. 648–664. doi:10.1002/embj.201387614.

Ladstätter, S. and Tachibana, K. (2019) “Genomic insights into chromatin reprogramming to totipotency in embryos,” *The journal of cell biology*, 218(1), pp. 70–82. doi:10.1083/jcb.201807044.

Lalonde, M.-E. *et al.* (2013) “Exchange of associated factors directs a switch in HBO1 acetyltransferase histone tail specificity,” *Genes & development*, 27(18), pp. 2009–2024. doi:10.1101/gad.223396.113.

Lam, S.S. *et al.* (2015) “Directed evolution of APEX2 for electron microscopy and proximity labeling,” *Nature methods*, 12(1), pp. 51–54. doi:10.1038/nmeth.3179.

Lambert, J.-P. *et al.* (2009) “A Novel Proteomics Approach for the Discovery of Chromatin-associated Protein Networks,” *Molecular & cellular proteomics: MCP*, 8(4), pp. 870–882. doi:10.1074/mcp.M800447-MCP200.

de Lange, T. (2005) “Shelterin: the protein complex that shapes and safeguards human telomeres,” *Genes & development*, 19(18), pp. 2100–2110. doi:10.1101/gad.1346005.

Lankadurai, B.P., Nagato, E.G. and Simpson, M.J. (2013) “Environmental metabolomics: an emerging approach to study organism responses to environmental stressors,” *Environmental reviews*, 21(3), pp. 180–205. doi:10.1139/er-2013-0011.

Lawson, K.A., Meneses, J.J. and Pedersen, R.A. (1991) “Clonal analysis of epiblast fate during germ layer formation in the mouse embryo,” *Development (Cambridge, England)*, 113(3), pp. 891–911. doi:10.1242/dev.113.3.891.

Lázaro, J. *et al.* (2023) “A stem cell zoo uncovers intracellular scaling of developmental tempo across mammals,” *Cell stem cell*, 30, pp. 938–949. doi:10.1016/j.stem.2023.05.014.

Lee, H.J., Hore, T.A. and Reik, W. (2014) “Reprogramming the methylome: Erasing memory and creating diversity,” *Cell stem cell*, 14(6), pp. 710–719. doi:10.1016/j.stem.2014.05.008.

Leitch, H.G. *et al.* (2013) “Naive pluripotency is associated with global DNA hypomethylation,” *Nature structural & molecular biology*, 20(3), pp. 311–316. doi:10.1038/nsmb.2510.

Lenz, S. *et al.* (2021) “Reliable identification of protein-protein interactions by crosslinking mass spectrometry,” *Nature communications*, 12(1), p. 3564. doi:10.1038/s41467-021-23666-z.

Leonhardt, H. *et al.* (1992) “A targeting sequence directs DNA methyltransferase to sites of DNA replication in mammalian nuclei,” *Cell*, 71(5), pp. 865–873. doi:10.1016/0092-8674(92)90561-p.

Li, S. *et al.* (2019) “Nonreciprocal and conditional cooperativity directs the pioneer activity of pluripotency transcription factors,” *Cell reports*, 28(10), pp. 2689-2703.e4. doi:10.1016/j.celrep.2019.07.103.

Linke, D. (2009) “Detergents: an overview,” *Methods in enzymology*, 463, pp. 603–617. doi:10.1016/S0076-6879(09)63034-2.

Liu, X. *et al.* (2016) “Distinct features of H3K4me3 and H3K27me3 chromatin domains in pre-implantation embryos,” *Nature*, 537(7621), pp. 558–562. doi:10.1038/nature19362.

Liu, X. *et al.* (2017) “In situ capture of chromatin interactions by biotinylated dCas9,” *Cell*, 170(5), pp. 1028-1043.e19. doi:10.1016/j.cell.2017.08.003.

Liu, Y., Beyer, A. and Aebersold, R. (2016) “On the dependency of cellular protein levels on mRNA abundance,” *Cell*, 165(3), pp. 535–550. doi:10.1016/j.cell.2016.03.014.

Loh, K.M. and Lim, B. (2011) “A precarious balance: pluripotency factors as lineage specifiers,” *Cell stem cell*, 8(4), pp. 363–369. doi:10.1016/j.stem.2011.03.013.

Lohmann, F. *et al.* (2010) “KMT1E mediated H3K9 methylation is required for the maintenance of embryonic stem cells by repressing trophectoderm differentiation,” *Stem cells (Dayton, Ohio)*, 28(2), pp. 201–212. doi:10.1002/stem.278.

Lopez-Contreras, A.J. *et al.* (2013) “A proteomic characterization of factors enriched at nascent DNA molecules,” *Cell reports*, 3(4), pp. 1105–1116. doi:10.1016/j.celrep.2013.03.009.

Lou, R. *et al.* (2020) “Hybrid spectral library combining DIA-MS data and a targeted virtual library substantially deepens the proteome coverage,” *iScience*, 23(3), p. 100903. doi:10.1016/j.isci.2020.100903.

Lou, R. *et al.* (2023) “Benchmarking commonly used software suites and analysis workflows for DIA proteomics and phosphoproteomics,” *Nature communications*, 14(1), p. 94. doi:10.1038/s41467-022-35740-1.

Ludwig, C. *et al.* (2018) “Data-independent acquisition-based SWATH-MS for quantitative proteomics: a tutorial,” *Molecular systems biology*, 14(8), p. e8126. doi:10.15252/msb.20178126.

Luo, H. *et al.* (2016) “Microarray-based analysis and clinical validation identify ubiquitin-conjugating enzyme E2E1 (UBE2E1) as a prognostic factor in acute myeloid leukemia,” *Journal of hematology & oncology*, 9(1), p. 125. doi:10.1186/s13045-016-0356-0.

Maherali, N. *et al.* (2007) “Directly reprogrammed fibroblasts show global epigenetic remodeling and widespread tissue contribution,” *Cell stem cell*, 1(1), pp. 55–70. doi:10.1016/j.stem.2007.05.014.

Mali, S. *et al.* (2016) “Observations on different resin strategies for affinity purification mass spectrometry of a tagged protein,” *Analytical biochemistry*, 515, pp. 26–32. doi:10.1016/j.ab.2016.09.022.

Mann (2016) “Origins of mass spectrometry-based proteomics,” *Nature reviews. Molecular cell biology*. Springer Science and Business Media LLC, p. 678. doi:10.1038/nrm.2016.135.

Mann, M. (2016) “The rise of mass spectrometry and the fall of Edman degradation,” *Clinical chemistry*, 62(1), pp. 293–294. doi:10.1373/clinchem.2014.237271.

Mann, M., Højrup, P. and Roepstorff, P. (1993) “Use of mass spectrometric molecular weight information to identify proteins in sequence databases,” *Biological mass spectrometry*, 22(6), pp. 338–345. doi:10.1002/bms.1200220605.

Mann, Meng and Fenn (1989) “Interpreting mass spectra of multiply charged ions,” *Analytical chemistry*, 61(15), pp. 1702–1708. doi:10.1021/ac00190a023.

Marks, H. *et al.* (2012) “The transcriptional and epigenomic foundations of ground state pluripotency,” *Cell*, 149(3), pp. 590–604. doi:10.1016/j.cell.2012.03.026.

Martens, J.H.A. *et al.* (2005) “The profile of repeat-associated histone lysine methylation states in the mouse epigenome,” *The EMBO journal*, 24(4), pp. 800–812. doi:10.1038/sj.emboj.7600545.

Martin, G.R. (1981) “Isolation of a pluripotent cell line from early mouse embryos cultured in medium conditioned by teratocarcinoma stem cells,” *Proceedings of the National Academy of Sciences of the United States of America*, 78(12), pp. 7634–7638. doi:10.1073/pnas.78.12.7634.

Marumoto, T. *et al.* (2002) “Roles of aurora-A kinase in mitotic entry and G2 checkpoint in mammalian cells,” *Genes to cells: devoted to molecular & cellular mechanisms*, 7(11), pp. 1173–1182. doi:10.1046/j.1365-2443.2002.00592.x.

Masaki, H. *et al.* (2016) “Inhibition of apoptosis overcomes stage-related compatibility barriers to chimera formation in mouse embryos,” *Cell stem cell*, 19(5), pp. 587–592. doi:10.1016/j.stem.2016.10.013.

Matoba, S. *et al.* (2014) “Embryonic development following somatic cell nuclear transfer impeded by persisting histone methylation,” *Cell*, 159(4), pp. 884–895. doi:10.1016/j.cell.2014.09.055.

Matsuda, K. *et al.* (1994) “Expression of GATA-binding transcription factors in rat hepatocytes,” *FEBS letters*, 353(3), pp. 269–272. doi:10.1016/0014-5793(94)01062-5.

Matzinger, M. *et al.* (2020) “Fast and Highly Efficient Affinity Enrichment of Azide-A-DSBSO Cross-Linked Peptides,” *Journal of proteome research*, 19(5), pp. 2071–2079. doi:10.1021/acs.jproteome.0c00003.

Meier, F. *et al.* (2015) “Parallel accumulation-serial fragmentation (PASEF): Multiplying sequencing speed and sensitivity by synchronized scans in a trapped ion mobility device," *Journal of proteome research*, 14(12), pp. 5378–5387. doi:10.1021/acs.jproteome.5b00932.

Meier, F. *et al.* (2020) "diaPASEF: parallel accumulation–serial fragmentation combined with data-independent acquisition," *Nature methods*, 17(12), pp. 1229–1236. doi:10.1038/s41592-020-00998-0.

Meier, F., Park, M.A. and Mann, M. (2021) "Trapped ion mobility spectrometry and parallel accumulation-serial fragmentation in proteomics," *Molecular & cellular proteomics: MCP*, 20(100138), p. 100138. doi:10.1016/j.mcpro.2021.100138.

Meissner, A. (2010) "Epigenetic modifications in pluripotent and differentiated cells," *Nature biotechnology*, 28(10), pp. 1079–1088. doi:10.1038/nbt.1684.

Melcer, S. *et al.* (2012) "Histone modifications and lamin A regulate chromatin protein dynamics in early embryonic stem cell differentiation," *Nature communications*, 3(1), p. 910. doi:10.1038/ncomms1915.

Meng, Y. *et al.* (2021) "Depletion of demethylase KDM6 enhances early neuroectoderm commitment of human PSCs," *Frontiers in cell and developmental biology*, 9, p. 702462. doi:10.3389/fcell.2021.702462.

Mereu, E. *et al.* (2020) "Benchmarking single-cell RNA-sequencing protocols for cell atlas projects," *Nature biotechnology*, 38(6), pp. 747–755. doi:10.1038/s41587-020-0469-4.

Merino, F. *et al.* (2014) "Structural basis for the SOX-dependent genomic redistribution of OCT4 in stem cell differentiation," *Structure (London, England: 1993)*, 22(9), pp. 1274–1286. doi:10.1016/j.str.2014.06.014.

Meshorer, E. *et al.* (2006) "Hyperdynamic plasticity of chromatin proteins in pluripotent embryonic stem cells," *Developmental cell*, 10(1), pp. 105–116. doi:10.1016/j.devcel.2005.10.017.

Messerschmidt, D.M., Knowles, B.B. and Solter, D. (2014) "DNA methylation dynamics during epigenetic reprogramming in the germline and preimplantation embryos," *Genes & development*, 28(8), pp. 812–828. doi:10.1101/gad.234294.113.

von Meyenn, F. *et al.* (2016) "Impairment of DNA Methylation Maintenance Is the Main Cause of Global Demethylation in Naive Embryonic Stem Cells," *Molecular cell*, 62(6), pp. 848–861. doi:10.1016/j.molcel.2016.04.025.

Meyer, J.G. (2021) "Deep learning neural network tools for proteomics," *Cell reports methods*, 1(2), p. 100003. doi:10.1016/j.crmeth.2021.100003.

Mick, D.U. *et al.* (2015) "Proteomics of primary cilia by proximity labeling," *Developmental cell*, 35(4), pp. 497–512. doi:10.1016/j.devcel.2015.10.015.

van Mierlo, G. *et al.* (2019) "Integrative proteomic profiling reveals PRC2-dependent epigenetic crosstalk maintains ground-state pluripotency," *Cell stem cell*, 24(1), pp. 123-137.e8. doi:10.1016/j.stem.2018.10.017.

van Mierlo, G. and Vermeulen, M. (2021) “Chromatin proteomics to study epigenetics - challenges and opportunities,” *Molecular & cellular proteomics: MCP*, 20, p. 100056. doi:10.1074/mcp.R120.002208.

van Mierlo, G., Wester, R.A. and Marks, H. (2018) “Quantitative subcellular proteomics using SILAC reveals enhanced metabolic buffering in the pluripotent ground state,” *Stem cell research*, 33, pp. 135–145. doi:10.1016/j.scr.2018.09.017.

van Mierlo, G., Wester, R.A. and Marks, H. (2019) “A mass spectrometry survey of chromatin-associated proteins in pluripotency and early lineage commitment,” *Proteomics*, 19(14), p. e1900047. doi:10.1002/pmic.201900047.

Millán-Zambrano, G. *et al.* (2022) “Histone post-translational modifications - cause and consequence of genome function,” *Nature reviews. Genetics*, 23(9), pp. 563–580. doi:10.1038/s41576-022-00468-7.

Miotto, B. and Struhl, K. (2010) “HBO1 histone acetylase activity is essential for DNA replication licensing and inhibited by Geminin,” *Molecular cell*, 37(1), pp. 57–66. doi:10.1016/j.molcel.2009.12.012.

Moffitt, J.R., Lundberg, E. and Heyn, H. (2022) “The emerging landscape of spatial profiling technologies,” *Nature reviews. Genetics*, 23(12), pp. 741–759. doi:10.1038/s41576-022-00515-3.

Mohammed, H. *et al.* (2013) “Endogenous purification reveals GREB1 as a key estrogen receptor regulatory factor,” *Cell reports*, 3(2), pp. 342–349. doi:10.1016/j.celrep.2013.01.010.

Monk, M., Boubelik, M. and Lehnert, S. (1987) “Temporal and regional changes in DNA methylation in the embryonic, extraembryonic and germ cell lineages during mouse embryo development,” *Development (Cambridge, England)*, 99(3), pp. 371–382. doi:10.1242/dev.99.3.371.

Moody, J.D. *et al.* (2017) “First critical repressive H3K27me3 marks in embryonic stem cells identified using designed protein inhibitor,” *Proceedings of the National Academy of Sciences of the United States of America*, 114(38), pp. 10125–10130. doi:10.1073/pnas.1706907114.

Morgani, S., Nichols, J. and Hadjantonakis, A.-K. (2017) “The many faces of Pluripotency: in vitro adaptations of a continuum of in vivo states,” *BMC developmental biology*, 17(1), p. 7. doi:10.1186/s12861-017-0150-4.

Morris, S.A. (2019) “The evolving concept of cell identity in the single cell era,” *Development*, 146(12), p. dev169748. doi:10.1242/dev.169748.

Mosteiro, L. *et al.* (2016) “Tissue damage and senescence provide critical signals for cellular reprogramming in vivo,” *Science*, 354(6315), p. aaf4445. doi:10.1126/science.aaf4445.

Moussaieff, A. *et al.* (2015) “Glycolysis-mediated changes in acetyl-CoA and histone acetylation control the early differentiation of embryonic stem cells,” *Cell metabolism*, 21(3), pp. 392–402. doi:10.1016/j.cmet.2015.02.002.

Mulholland, C.B., Traube, F.R., *et al.* (2020) "Distinct and stage-specific contributions of TET1 and TET2 to stepwise cytosine oxidation in the transition from naive to primed pluripotency," *Scientific reports*, 10(1), p. 12066. doi:10.1038/s41598-020-68600-3.

Mulholland, C.B., Nishiyama, A., *et al.* (2020) "Recent evolution of a TET-controlled and DPPA3/STELLA-driven pathway of passive DNA demethylation in mammals," *Nature communications*, 11(1), p. 5972. doi:10.1038/s41467-020-19603-1.

Mullen, A.C. *et al.* (2011) "Master transcription factors determine cell-type-specific responses to TGF-β signaling," *Cell*, 147(3), pp. 565–576. doi:10.1016/j.cell.2011.08.050.

Myers, S.A. *et al.* (2018) "Discovery of proteins associated with a predefined genomic locus via dCas9-APEX-mediated proximity labeling," *Nature methods*, 15(6), pp. 437–439. doi:10.1038/s41592-018-0007-1.

Nakamura, K. *et al.* (2021) "Proteome dynamics at broken replication forks reveal a distinct ATM-directed repair response suppressing DNA double-strand break ubiquitination," *Molecular cell*, 81(5), pp. 1084-1099.e6. doi:10.1016/j.molcel.2020.12.025.

Nakamura, M. *et al.* (2021) "CRISPR technologies for precise epigenome editing," *Nature cell biology*, 23(1), pp. 11–22. doi:10.1038/s41556-020-00620-7.

Nakamura, T. *et al.* (2016) "A developmental coordinate of pluripotency among mice, monkeys and humans," *Nature*, 537(7618), pp. 57–62. doi:10.1038/nature19096.

Nardozzi, J.D., Lott, K. and Cingolani, G. (2010) "Phosphorylation meets nuclear import: a review," *Cell communication and signaling: CCS*, 8(1), p. 32. doi:10.1186/1478-811X-8-32.

Neagu, A. *et al.* (2020) "In vitro capture and characterization of embryonic rosette-stage pluripotency between naive and primed states," *Nature cell biology*, 22(5), pp. 534–545. doi:10.1038/s41556-020-0508-x.

Neagu, A.-N. *et al.* (2022) "Applications of tandem Mass Spectrometry (MS/MS) in protein analysis for biomedical research," *Molecules (Basel, Switzerland)*, 27(8), p. 2411. doi:10.3390/molecules27082411.

Neri, F. *et al.* (2013) "Genome-wide analysis identifies a functional association of Tet1 and Polycomb repressive complex 2 in mouse embryonic stem cells," *Genome biology*, 14(8), p. R91. doi:10.1186/gb-2013-14-8-r91.

Nesvizhskii, A.I. (2007) "Protein identification by tandem mass spectrometry and sequence database searching," *Methods in molecular biology (Clifton, N.J.)*, 367, pp. 87–119. doi:10.1385/1-59745-275-0:87.

Nicetto, D. *et al.* (2019) "H3K9me3-heterochromatin loss at protein-coding genes enables developmental lineage specification," *Science*, 363(6424), pp. 294–297. doi:10.1126/science.aau0583.

Nichols, J. *et al.* (1998) "Formation of pluripotent stem cells in the mammalian embryo depends on the POU transcription factor Oct4," *Cell*, 95(3), pp. 379–391. doi:10.1016/s0092-8674(00)81769-9.

Nichols, J. and Smith, A. (2009) "Naive and primed pluripotent states," *Cell stem cell*, 4(6), pp. 487–492. doi:10.1016/j.stem.2009.05.015.

Nicosia, L. and Bonaldi, T. (2021) "Native Chromatin Proteomics (N-ChroP) to characterize histone post-translational modification (PTM) combinatorics at distinct genomic regions," *Methods in molecular biology (Clifton, N.J.)*, 2351, pp. 251–274. doi:10.1007/978-1-0716-1597-3_14.

Nishiyama, A. *et al.* (2020) "Two distinct modes of DNMT1 recruitment ensure stable maintenance DNA methylation," *Nature communications*, 11(1), p. 1222. doi:10.1038/s41467-020-15006-4.

Noberini, R., Robusti, G. and Bonaldi, T. (2022) "Mass spectrometry-based characterization of histones in clinical samples: applications, progress, and challenges," *The FEBS journal*, 289(5), pp. 1191–1213. doi:10.1111/febs.15707.

O'Carroll, D., Scherthan, H., *et al.* (2001) "Loss of the Suv39h Histone Methyltransferases Impairs Mammalian Heterochromatin and Genome Stability," *Cell*, 107, pp. 323–337.

O'Carroll, D., Erhardt, S., *et al.* (2001) "The polycomb-group gene Ezh2 is required for early mouse development," *Molecular and cellular biology*, 21(13), pp. 4330–4336. doi:10.1128/MCB.21.13.4330-4336.2001.

Ohta, S. *et al.* (2003) "The ORC1 cycle in human cells: II. Dynamic changes in the human ORC complex during the cell cycle," *The journal of biological chemistry*, 278(42), pp. 41535–41540. doi:10.1074/jbc.M307535200.

Okano, M. *et al.* (1999) "DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development," *Cell*, 99(3), pp. 247–257. doi:10.1016/s0092-8674(00)81656-6.

Okita, K., Ichisaka, T. and Yamanaka, S. (2007) "Generation of germline-competent induced pluripotent stem cells," *Nature*, 448(7151), pp. 313–317. doi:10.1038/nature05934.

Olariu, V., Lövkvist, C. and Sneppen, K. (2016) "Nanog, Oct4 and Tet1 interplay in establishing pluripotency," *Scientific reports*, 6(1), p. 25438. doi:10.1038/srep25438.

Oliviero, G. *et al.* (2022) "Distinct and diverse chromatin proteomes of ageing mouse organs reveal protein signatures that correlate with physiological functions," *eLife*, 11, p. e73524. doi:10.7554/elife.73524.

Olsen, J.V. *et al.* (2005) "Parts per million mass accuracy on an Orbitrap mass spectrometer via lock mass injection into a C-trap," *Molecular & cellular proteomics: MCP*, 4(12), pp. 2010–2021. doi:10.1074/mcp.T500030-MCP200.

Olsen, J.V., Ong, S.-E. and Mann, M. (2004) “Trypsin cleaves exclusively C-terminal to arginine and lysine residues,” *Molecular & cellular proteomics: MCP*, 3(6), pp. 608–614. doi:10.1074/mcp.T400003-MCP200.

Pasini, D. *et al.* (2004) “Suz12 is essential for mouse development and for EZH2 histone methyltransferase activity,” *The EMBO journal*, 23(20), pp. 4061–4071. doi:10.1038/sj.emboj.7600402.

Paul, W. and Steinwedel, H. (1953) “Notizen: Ein neues Massenspektrometer ohne Magnetfeld,” *Zeitschrift für Naturforschung A*, 8(7), pp. 448–450. doi:10.1515/zna-1953-0710.

Perez-Riverol, Y. *et al.* (2022) “The PRIDE database resources in 2022: a hub for mass spectrometry-based proteomics evidences,” *Nucleic acids research*, 50(D1), pp. D543–D552. doi:10.1093/nar/gkab1038.

Perkins, D.N. *et al.* (1999) “Probability-based protein identification by searching sequence databases using mass spectrometry data,” *Electrophoresis*, 20(18), pp. 3551–3567. doi:10.1002/(SICI)1522-2683(19991201)20:18<3551::AID-ELPS3551>3.0.CO;2-2.

Peters, A.H.F.M. *et al.* (2003) “Partitioning and plasticity of repressive histone methylation states in mammalian chromatin,” *Molecular cell*, 12(6), pp. 1577–1589. doi:10.1016/s1097-2765(03)00477-5.

Petryk, N. *et al.* (2021) “Staying true to yourself: mechanisms of DNA methylation maintenance in mammals,” *Nucleic acids research*, 49(6), pp. 3020–3032. doi:10.1093/nar/gkaa1154.

Pham, T.X.A. *et al.* (2022) “Modeling human extraembryonic mesoderm cells using naive pluripotent stem cells,” *Cell stem cell*, 29(9), pp. 1346-1365.e10. doi:10.1016/j.stem.2022.08.001.

Pino, L.K. *et al.* (2020) “Acquiring and analyzing data independent acquisition proteomics experiments without spectrum libraries,” *Molecular & cellular proteomics: MCP*, 19(7), pp. 1088–1103. doi:10.1074/mcp.P119.001913.

Qin, Stengl, A., *et al.* (2021) “HP1β carries an acidic linker domain and requires H3K9me3 for phase separation,” *Nucleus (Austin, Tex.)*, 12(1), pp. 44–57. doi:10.1080/19491034.2021.1889858.

Qin, Ugur, E., *et al.* (2021) “Phosphorylation of the HP1β hinge region sequesters KAP1 in heterochromatin and promotes the exit from naïve pluripotency,” *Nucleic acids research*, 49(13), pp. 7406–7423. doi:10.1093/nar/gkab548.

Qin, Y. *et al.* (2021) “A multi-scale map of cell structure fusing protein images and interactions,” *Nature*, 600(7889), pp. 536–542. doi:10.1038/s41586-021-04115-9.

Qiu, W. *et al.* (2019) “Determination of local chromatin interactions using a combined CRISPR and peroxidase APEX2 system,” *Nucleic acids research*, 47(9), p. e52. doi:10.1093/nar/gkz134.

Rafiee, M.-R. *et al.* (2016) "Expanding the Circuitry of Pluripotency by Selective Isolation of Chromatin-Associated Proteins," *Molecular cell*, 64(3), pp. 624–635. doi:10.1016/j.molcel.2016.09.019.

Ragazzini, R. *et al.* (2019) "EZHIP constrains Polycomb Repressive Complex 2 activity in germ cells," *Nature communications*, 10(1), p. 3858. doi:10.1038/s41467-019-11800-x.

Rappsilber, J., Ishihama, Y. and Mann, M. (2003) "Stop and go extraction tips for matrix-assisted laser desorption/ionization, nanoelectrospray, and LC/MS sample pretreatment in proteomics," *Analytical chemistry*, 75(3), pp. 663–670. doi:10.1021/ac026117i.

Rappsilber, J., Mann, M. and Ishihama, Y. (2007) "Protocol for micro-purification, enrichment, pre-fractionation and storage of peptides for proteomics using StageTips," *Nature protocols*, 2(8), pp. 1896–1906. doi:10.1038/nprot.2007.261.

Räschle, M. *et al.* (2015) "DNA repair. Proteomics reveals dynamic assembly of repair complexes during bypass of DNA cross-links," *Science*, 348(6234), p. 1253671. doi:10.1126/science.1253671.

Reddington, J.P. *et al.* (2013) "Redistribution of H3K27me3 upon DNA hypomethylation results in de-repression of Polycomb target genes," *Genome biology*, 14(3), p. R25. doi:10.1186/gb-2013-14-3-r25.

Reik, W., Dean, W. and Walter, J. (2001) "Epigenetic reprogramming in mammalian development," *Science*, 293(5532), pp. 1089–1093. doi:10.1126/science.1063443.

Ribeyre, C. *et al.* (2016) "Nascent DNA proteomics reveals a chromatin remodeler required for topoisomerase I loading at replication forks," *Cell reports*, 15(2), pp. 300–309. doi:10.1016/j.celrep.2016.03.027.

Ricketts, T.D. *et al.* (2021) "Mechanisms of macrophage plasticity in the tumor environment: Manipulating activation state to improve outcomes," *Frontiers in immunology*, 12, p. 642285. doi:10.3389/fimmu.2021.642285.

Ridgeway, M.E. *et al.* (2018) "Trapped ion mobility spectrometry: A short review," *International journal of mass spectrometry*, 425, pp. 22–35. doi:10.1016/j.ijms.2018.01.006.

Robertson, E.J., Evans, M.J. and Kaufman, M.H. (1983) "X-chromosome instability in pluripotential stem cell lines derived from parthenogenetic embryos," *Journal of embryology and experimental morphology*, 74, pp. 297–309. Available at: https://www.ncbi.nlm.nih.gov/pubmed/6886600.

Ross, S.E. and Bogdanovic, O. (2019) "TET enzymes, DNA demethylation and pluripotency," *Biochemical Society transactions*, 47(3), pp. 875–885. doi:10.1042/BST20180606.

Rossant, J. and Tam, P.P.L. (2017) "New Insights into Early Human Development: Lessons for Stem Cell Derivation and Differentiation," *Cell stem cell*, 20(1), pp. 18–28. doi:10.1016/j.stem.2016.12.004.

Rottach, A., Leonhardt, H. and Spada, F. (2009) "DNA methylation-mediated epigenetic control," *Journal of cellular biochemistry*, 108(1), pp. 43–51. doi:10.1002/jcb.22253.

Roulois, D. *et al.* (2015) "DNA-demethylating agents target colorectal cancer cells by inducing viral mimicry by endogenous transcripts," *Cell*, 162(5), pp. 961–973. doi:10.1016/j.cell.2015.07.056.

Roux, K.J. *et al.* (2012) "A promiscuous biotin ligase fusion protein identifies proximal and interacting proteins in mammalian cells," *The Journal of cell biology*, 196(6), pp. 801–810. doi:10.1083/jcb.201112098.

Rowe, H.M. and Trono, D. (2011) "Dynamic control of endogenous retroviruses during development," *Virology*, 411(2), pp. 273–287. doi:10.1016/j.virol.2010.12.007.

Rulands, S. *et al.* (2018) "Genome-Scale Oscillations in DNA Methylation during Exit from Pluripotency," *Cell systems*, 7(1), pp. 63-76.e12. doi:10.1016/j.cels.2018.06.012.

Sagi, I. *et al.* (2016) "Derivation and differentiation of haploid human embryonic stem cells," *Nature*, 532(7597), pp. 107–111. doi:10.1038/nature17408.

Saksouk, N. *et al.* (2014) "Redundant mechanisms to form silent chromatin at pericentromeric regions rely on BEND3 and DNA methylation," *Molecular cell*, 56(4), pp. 580–594. doi:10.1016/j.molcel.2014.10.001.

Samejima, I. *et al.* (2022) "Mapping the invisible chromatin transactions of prophase chromosome remodeling," *Molecular cell*, 82(3), pp. 696-708.e4. doi:10.1016/j.molcel.2021.12.039.

Sanford, J.P. *et al.* (1987) "Differences in DNA methylation during oogenesis and spermatogenesis and their persistence during early embryogenesis in the mouse," *Genes & development*, 1(10), pp. 1039–1046. doi:10.1101/gad.1.10.1039.

Sanger, F., Thompson, E.O. and Kitai, R. (1955) "The amide groups of insulin," *The Biochemical journal*, 59(3), pp. 509–518. doi:10.1042/bj0590509.

Santos-Barriopedro, I., van Mierlo, G. and Vermeulen, M. (2021) "Off-the-shelf proximity biotinylation for interaction proteomics," *Nature communications*, 12(1), p. 5015. doi:10.1038/s41467-021-25338-4.

Sarkar, A. and Hochedlinger, K. (2013) "The sox family of transcription factors: versatile regulators of stem and progenitor cell fate," *Cell stem cell*, 12(1), pp. 15–30. doi:10.1016/j.stem.2012.12.007.

Schlesinger, S. and Meshorer, E. (2019) "Open chromatin, epigenetic plasticity, and nuclear organization in pluripotency," *Developmental cell*, 48(2), pp. 135–150. doi:10.1016/j.devcel.2019.01.003.

Schmidtmann, E. *et al.* (2016) "Determination of local chromatin composition by CasID," *Nucleus*, 7(5), pp. 476–484. doi:10.1080/19491034.2016.1239000.

Schopp, I.M. *et al.* (2017) “Split-BioID a conditional proteomics approach to monitor the composition of spatiotemporally defined protein complexes,” *Nature communications*, 8(1), p. 15690. doi:10.1038/ncomms15690.

Schübeler, D. (2015) “Function and information content of DNA methylation,” *Nature*, 517(7534), pp. 321–326. doi:10.1038/nature14192.

Schubert, O.T. *et al.* (2015) “Building high-quality assay libraries for targeted analysis of SWATH MS data,” *Nature protocols*, 10(3), pp. 426–441. doi:10.1038/nprot.2015.015.

Seath, C.P. *et al.* (2023) “Tracking chromatin state changes using nanoscale photo-proximity labelling,” *Nature*, 616(7957), pp. 574–580. doi:10.1038/s41586-023-05914-y.

Seisenberger, S. *et al.* (2012) “The dynamics of genome-wide DNA methylation reprogramming in mouse primordial germ cells,” *Molecular cell*, 48(6), pp. 849–862. doi:10.1016/j.molcel.2012.11.001.

Senko, M.W., Beu, S.C. and McLaffertycor, F.W. (1995) “Determination of monoisotopic masses and ion populations for large biomolecules from resolved isotopic distributions,” *Journal of the American Society for Mass Spectrometry*, 6(4), pp. 229–233. doi:10.1016/1044-0305(95)00017-8.

Shahbazi, M.N. *et al.* (2017) “Pluripotent state transitions coordinate morphogenesis in mouse and human embryos,” *Nature*, 552(7684), pp. 239–243. doi:10.1038/nature24675.

Shahbazy, M. *et al.* (2023) “Benchmarking bioinformatics pipelines in data-independent acquisition mass spectrometry for immunopeptidomics,” *Molecular & cellular proteomics: MCP*, 22(4), p. 100515. doi:10.1016/j.mcpro.2023.100515.

Shamblott, M.J. *et al.* (1998) “Derivation of pluripotent stem cells from cultured human primordial germ cells,” *Proceedings of the National Academy of Sciences of the United States of America*, 95(23), pp. 13726–13731. doi:10.1073/pnas.95.23.13726.

Shanak, S. and Helms, V. (2020) “DNA methylation and the core pluripotency network,” *Developmental biology*, 464(2), pp. 145–160. doi:10.1016/j.ydbio.2020.06.001.

Sheban, D. *et al.* (2022) “SUMOylation of linker histone H1 drives chromatin condensation and restriction of embryonic cell fate identity,” *Molecular cell*, 82(1), pp. 106-122.e9. doi:10.1016/j.molcel.2021.11.011.

Shechter, D. *et al.* (2007) “Extraction, purification and analysis of histones,” *Nature protocols*, 2(6), pp. 1445–1457. doi:10.1038/nprot.2007.202.

Shevchenko, A. *et al.* (1996) “Mass spectrometric sequencing of proteins from silver-stained polyacrylamide gels,” *Analytical chemistry*, 68(5), pp. 850–858. doi:10.1021/ac950914h.

Shiio, Y. *et al.* (2003) “Quantitative proteomic analysis of chromatin-associated factors,” *Journal of the American Society for Mass Spectrometry*, 14(7), pp. 696–703. doi:10.1016/S1044-0305(03)00204-6.

Shirane, K. *et al.* (2016) “Global landscape and regulatory principles of DNA methylation reprogramming for germ cell specification by mouse pluripotent stem cells,” *Developmental cell*, 39(1), pp. 87–103. doi:10.1016/j.devcel.2016.08.008.

Shuken, S.R. (2023) “An introduction to mass spectrometry-based proteomics,” *Journal of proteome research*, 22(7), pp. 2151–2171. doi:10.1021/acs.jproteome.2c00838.

Sidoli, S. and Garcia, B.A. (2017) “Middle-down proteomics: a still unexploited resource for chromatin biology,” *Expert review of proteomics*, 14(7), pp. 617–626. doi:10.1080/14789450.2017.1345632.

Sigismondo, G. *et al.* (2023) “Multi-layered chromatin proteomics identifies cell vulnerabilities in DNA repair,” *Nucleic acids research*, 51(2), pp. 687–711. doi:10.1093/nar/gkac1264.

Sigismondo, G., Papageorgiou, D.N. and Krijgsveld, J. (2022) “Cracking chromatin with proteomics: From chromatome to histone modifications,” *Proteomics*, 22(15–16), p. e2100206. doi:10.1002/pmic.202100206.

Silva, J. *et al.* (2008) “Promotion of reprogramming to ground state pluripotency by signal inhibition,” *PLoS biology*, 6(10), p. e253. doi:10.1371/journal.pbio.0060253.

Sindelar, M. and Patti, G.J. (2020) “Chemical discovery in the era of metabolomics,” *Journal of the American Chemical Society*, 142(20), pp. 9097–9105. doi:10.1021/jacs.9b13198.

Sinha, K.K. *et al.* (2023) “Histone modifications regulate pioneer transcription factor cooperativity,” *Nature*, 619, pp. 378–384. doi:10.1038/s41586-023-06112-6.

Sinitcyn, P. *et al.* (2021) “MaxDIA enables library-based and library-free data-independent acquisition proteomics,” *Nature biotechnology*, 39(12), pp. 1563–1573. doi:10.1038/s41587-021-00968-7.

Sinitcyn, P., Rudolph, J.D. and Cox, J. (2018) “Computational Methods for Understanding Mass Spectrometry–Based Shotgun Proteomics Data,” *Annual Review of Biomedical Data Science*, 1(1), pp. 207–234. doi:10.1146/annurev-biodatasci-080917-013516.

Sirbu, B.M. *et al.* (2011) “Analysis of protein dynamics at active, stalled, and collapsed replication forks,” *Genes & development*, 25(12), pp. 1320–1327. doi:10.1101/gad.2053211.

Skene, P.J. and Henikoff, S. (2017) “An efficient targeted nuclease strategy for high-resolution mapping of DNA binding sites,” *eLife*, 6, p. e21856. doi:10.7554/eLife.21856.

Skowronek, P. *et al.* (2022) "Synchro-PASEF allows precursor-specific fragment ion extraction and interference removal in data-independent acquisition," *Molecular & cellular proteomics: MCP*, 22(2), p. 100489. doi:10.1016/j.mcpro.2022.100489.

Smith, A. (2017) "Formative pluripotency: the executive phase in a developmental continuum," *Development (Cambridge, England)*, 144(3), pp. 365–373. doi:10.1242/dev.142679.

Smith, A.G. *et al.* (1988) "Inhibition of pluripotential embryonic stem cell differentiation by purified polypeptides," *Nature*, 336(6200), pp. 688–690. doi:10.1038/336688a0.

Smith, L.M., Kelleher, N.L. and Consortium for Top Down Proteomics (2013) "Proteoform: a single term describing protein complexity," *Nature methods*, 10(3), pp. 186–187. doi:10.1038/nmeth.2369.

Smith, Z.D. *et al.* (2012) "A unique regulatory phase of DNA methylation in the early mammalian embryo," *Nature*, 484(7394), pp. 339–344. doi:10.1038/nature10960.

Soldi, M. and Bonaldi, T. (2014) "The ChroP approach combines ChIP and mass spectrometry to dissect locus-specific proteomic landscapes of chromatin," *Journal of visualized experiments: JoVE* [Preprint], (86). doi:10.3791/51220.

Solter, D. (2006) "From teratocarcinomas to embryonic stem cells and beyond: a history of embryonic stem cell research," *Nature reviews. Genetics*, 7(4), pp. 319–327. doi:10.1038/nrg1827.

Spruijt, C.G. *et al.* (2013) "Dynamic readers for 5-(hydroxy)methylcytosine and its oxidized derivatives," *Cell*, 152(5), pp. 1146–1159. doi:10.1016/j.cell.2013.02.004.

Stahl, D.C. *et al.* (1996) "Data-controlled automation of liquid chromatography/tandem mass spectrometry analysis of peptide mixtures," *Journal of the American Society for Mass Spectrometry*, 7(6), pp. 532–540. doi:10.1016/1044-0305(96)00057-8.

Stahl, K., Brock, O. and Rappsilber, J. (2023) "Modelling protein complexes with crosslinking mass spectrometry and deep learning," *bioRxiv.* (2023-06). doi:10.1101/2023.06.07.544059.

Stancheva, I. (2005) "Caught in conspiracy: cooperation between DNA methylation and histone H3K9 methylation in the establishment and maintenance of heterochromatin," *Biochemistry and cell biology*, 83(3), pp. 385–395. doi:10.1139/o05-043.

Steen, H. and Mann, M. (2004) "The ABC's (and XYZ's) of peptide sequencing," *Nature reviews. Molecular cell biology*, 5(9), pp. 699–711. doi:10.1038/nrm1468.

Steigenberger, B. *et al.* (2019) "PhoX: An IMAC-Enrichable Cross-Linking Reagent," *ACS central science*, 5(9), pp. 1514–1522. doi:10.1021/acscentsci.9b00416.

Stephens, W.E. (1946) "A pulsed mass spectrometer with time disaersion," *Physics Review*, 69, p. 691. Available at: https://cir.nii.ac.jp/crid/1573387449129435264.

Stewart, H. *et al.* (2023) "Parallelized acquisition of Orbitrap and Astral analyzers enables high-throughput quantitative analysis." (2023-06). doi:10.1101/2023.06.02.543408.

Stolz, P. *et al.* (2022) "TET1 regulates gene expression and repression of endogenous retroviruses independent of DNA demethylation," *Nucleic acids research*, 50(15), pp. 8491–8511. doi:10.1093/nar/gkac642.

Surani, M.A., Hayashi, K. and Hajkova, P. (2007) "Genetic and epigenetic regulators of pluripotency," *Cell*, 128(4), pp. 747–762. doi:10.1016/j.cell.2007.02.010.

Tabb, D.L. *et al.* (2010) "Repeatability and reproducibility in proteomic identifications by liquid chromatography-tandem mass spectrometry," *Journal of proteome research*, 9(2), pp. 761–776. doi:10.1021/pr9006365.

Tada, M. *et al.* (2001) "Nuclear reprogramming of somatic cells by in vitro hybridization with ES cells," *Current biology: CB*, 11(19), pp. 1553–1558. doi:10.1016/s0960-9822(01)00459-6.

Tahiliani, M. *et al.* (2009) "Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1," *Science*, 324(5929), pp. 930–935. doi:10.1126/science.1170116.

Takahashi, K. *et al.* (2007) "Induction of pluripotent stem cells from adult human fibroblasts by defined factors," *Cell*, 131(5), pp. 861–872. doi:10.1016/j.cell.2007.11.019.

Takahashi, K. and Yamanaka, S. (2006) "Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors," *Cell*, 126(4), pp. 663–676. doi:10.1016/j.cell.2006.07.024.

Takahashi, S., Kobayashi, S. and Hiratani, I. (2018) "Epigenetic differences between naïve and primed pluripotent stem cells," *Cellular and molecular life sciences: CMLS*, 75(7), pp. 1191–1203. doi:10.1007/s00018-017-2703-x.

Tam, P.P. and Zhou, S.X. (1996) "The allocation of epiblast cells to ectodermal and germ-line lineages is influenced by the position of the cells in the gastrulating mouse embryo," *Developmental biology*, 178(1), pp. 124–132. doi:10.1006/dbio.1996.0203.

Tan, F. *et al.* (2007) "Proteome and phosphoproteome analysis of chromatin associated proteins in rice (Oryza sativa)," *Proteomics*, 7(24), pp. 4511–4527. doi:10.1002/pmic.200700580.

Tanford, C. and Reynolds, J. (2001) *Nature's robots: A history of proteins.* Oxford, England: Oxford University Press, pp. 1–41.

Taudt, A., Colomé-Tatché, M. and Johannes, F. (2016) “Genetic sources of population epigenomic variation,” *Nature reviews. Genetics*, 17(6), pp. 319–332. doi:10.1038/nrg.2016.45.

Taverna, S.D. *et al.* (2007) “Long-distance combinatorial linkage between methylation and acetylation on histone H3 N termini,” *Proceedings of the National Academy of Sciences of the United States of America*, 104(7), pp. 2086–2091. doi:10.1073/pnas.0610993104.

Tesar, P.J. *et al.* (2007) “New cell lines from mouse epiblast share defining features with human embryonic stem cells,” *Nature*, 448(7150), pp. 196–199. doi:10.1038/nature05972.

Teves, S.S. *et al.* (2016) “A dynamic mode of mitotic bookmarking by transcription factors,” *eLife*, 5, p. e22280. doi:10.7554/elife.22280.

The 1000 Genomes Project Consortium (2012) “An integrated map of genetic variation from 1,092 human genomes,” *Nature*, 491(7422), pp. 56–65. doi:10.1038/nature11632.

The ENCODE Project Consortium (2012) “An integrated encyclopedia of DNA elements in the human genome,” *Nature*, 489(7414), pp. 57–74. doi:10.1038/nature11247.

Thomas, P.D. *et al.* (2022) “PANTHER: Making genome-scale phylogenetics accessible to all,” *Protein science: a publication of the Protein Society*, 31(1), pp. 8–22. doi:10.1002/pro.4218.

Thomson (1897) “XL. Cathode Rays,” *The London Edinburgh and Dublin Philosophical Magazine and Journal of Science*, 44(269), pp. 293–316. doi:10.1080/14786449708621070.

Thomson, J.J. (1914) “Rays of positive electricity and their application to chemical analysis,” *The journal of the Röntgen Society*, 10(39), pp. 41–42. doi:10.1259/jrs.1914.0021.

Thomson, M. *et al.* (2011) “Pluripotency factors in embryonic stem cells regulate differentiation into germ layers,” *Cell*, 145(6), pp. 875–889. doi:10.1016/j.cell.2011.05.017.

Tiwary, S. *et al.* (2019) “High-quality MS/MS spectrum prediction for data-dependent and data-independent acquisition data analysis,” *Nature methods*, 16(6), pp. 519–525. doi:10.1038/s41592-019-0427-6.

Torrente, M.P. *et al.* (2011) “Proteomic interrogation of human chromatin,” *PloS one*, 6(9), p. e24747. doi:10.1371/journal.pone.0024747.

Tosolini, M. *et al.* (2018) “Contrasting epigenetic states of heterochromatin in the different types of mouse pluripotent stem cells,” *Scientific reports*, 8(1), p. 5776. doi:10.1038/s41598-018-23822-4.

Treutlein, B. *et al.* (2014) “Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-seq,” *Nature*, 509(7500), pp. 371–375. doi:10.1038/nature13173.

Trompouki, E. *et al.* (2011) “Lineage regulators direct BMP and wnt pathways to cell-specific programs during differentiation and regeneration,” *Cell*, 147(3), pp. 577–589. doi:10.1016/j.cell.2011.09.044.

Tsai, C.-C. *et al.* (2012) “Oct4 and Nanog directly regulate Dnmt1 to maintain self-renewal and undifferentiated state in mesenchymal stem cells,” *Molecular cell*, 47(2), pp. 169–182. doi:10.1016/j.molcel.2012.06.020.

Tsakiridis, A. *et al.* (2014) “Distinct Wnt-driven primitive streak-like populations reflect in vivo lineage precursors,” *Development (Cambridge, England)*, 141(6), pp. 1209–1221. doi:10.1242/dev.101014.

Tsogtbaatar, E. *et al.* (2020) “Energy metabolism regulates stem cell pluripotency,” *Frontiers in cell and developmental biology*, 8, p. 87. doi:10.3389/fcell.2020.00087.

Tsou, C.-C. *et al.* (2015) “DIA-Umpire: comprehensive computational framework for data-independent acquisition proteomics,” *Nature methods*, 12(3), pp. 258–264. doi:10.1038/nmeth.3255.

Tsusaka, T., Shimura, C. and Shinkai, Y. (2019) “ATF7IP regulates SETDB1 nuclear localization and increases its ubiquitination,” *EMBO reports*, 20(12), p. e48297. doi:10.15252/embr.201948297.

Uchiyama, S. *et al.* (2005) “Proteome analysis of human metaphase chromosomes,” *The journal of biological chemistry*, 280(17), pp. 16994–17004. doi:10.1074/jbc.M412774200.

Ugur, E. *et al.* (2023) “Comprehensive chromatin proteomics resolves functional phases of pluripotency and identifies changes in regulatory components,” *Nucleic acids research*, 51(6), pp. 2671–2690. doi:10.1093/nar/gkad058.

Ugur, E., Bartoschek, M.D. and Leonhardt, H. (2020) “Locus-Specific Chromatin Proteome Revealed by Mass Spectrometry-Based CasID,” *Methods in molecular biology*, 2175, pp. 109–121. doi:10.1007/978-1-0716-0763-3_9.

Ura, H. *et al.* (2008) “STAT3 and Oct-3/4 control histone modification through induction of Eed in embryonic stem cells,” *The journal of biological chemistry*, 283(15), pp. 9713–9723. doi:10.1074/jbc.M707275200.

Vanzan, L. *et al.* (2021) “High throughput screening identifies SOX2 as a super pioneer factor that inhibits DNA methylation maintenance at its binding sites,” *Nature communications*, 12(1), p. 3337. doi:10.1038/s41467-021-23630-x.

Vega, H. *et al.* (2005) “Roberts syndrome is caused by mutations in ESCO2, a human homolog of yeast ECO1 that is essential for the establishment of sister chromatid cohesion,” *Nature genetics*, 37(5), pp. 468–470. doi:10.1038/ng1548.

Venable, J.D. *et al.* (2004) “Automated approach for quantitative analysis of complex peptide mixtures from tandem mass spectra,” *Nature methods*, 1(1), pp. 39–45. doi:10.1038/nmeth705.

Verheggen, K. *et al.* (2020) “Anatomy and evolution of database search engines-a central component of mass spectrometry based proteomic workflows,” *Mass spectrometry reviews*, 39(3), pp. 292–306. doi:10.1002/mas.21543.

Vermeulen, M. *et al.* (2010) “Quantitative interaction proteomics and genome-wide profiling of epigenetic histone marks and their readers,” *Cell*, 142(6), pp. 967–980. doi:10.1016/j.cell.2010.08.020.

Vickery, H.B. (1950) “The origin of the word protein,” *The Yale journal of biology and medicine*, 22(5), pp. 387–393. Available at: https://www.ncbi.nlm.nih.gov/pubmed/15413335.

Villaseñor, R. *et al.* (2020) “ChromID identifies the protein interactome at chromatin marks,” *Nature biotechnology*, 38(6), pp. 728–736. doi:10.1038/s41587-020-0434-2.

Waldrip, Z.J. *et al.* (2014) “A CRISPR-based approach for proteomic analysis of a single genomic locus,” *Epigenetics: official journal of the DNA Methylation Society*, 9(9), pp. 1207–1211. doi:10.4161/epi.29919.

Walsh, C.P., Chaillet, J.R. and Bestor, T.H. (1998) “Transcription of IAP endogenous retroviruses is constrained by cytosine methylation,” *Nature genetics*, 20(2), pp. 116–117. doi:10.1038/2413.

Walter, M. *et al.* (2016) “An epigenetic switch ensures transposon repression upon dynamic loss of DNA methylation in embryonic stem cells,” *eLife*, 5, p. e11418. doi:10.7554/eLife.11418.

Wang *et al.* (2020) “High-throughput and deep-proteome profiling by 16-plex tandem mass tag labeling coupled with two-dimensional chromatography and mass spectrometry,” *Journal of visualized experiments: JoVE*, (162), p. e61684. doi:10.3791/61684.

Wang *et al.* (2022) “Dominant role of DNA methylation over H3K9me3 for IAP silencing in endoderm,” *Nature communications*, 13(1), p. 5447. doi:10.1038/s41467-022-32978-7.

Wang, C.I. *et al.* (2013) “Chromatin proteins captured by ChIP-mass spectrometry are linked to dosage compensation in Drosophila,” *Nature structural & molecular biology*, 20(2), pp. 202–209. doi:10.1038/nsmb.2477.

Wang, J. *et al.* (2014) “Primate-specific endogenous retrovirus-driven transcription defines naive-like stem cells,” *Nature*, 516(7531), pp. 405–409. doi:10.1038/nature13804.

Wang, Xiaoxiao *et al.* (2021) “Formative pluripotent stem cells show features of epiblast cells poised for gastrulation,” *Cell research*, 31(5), pp. 526–541. doi:10.1038/s41422-021-00477-x.

Wang, Z. *et al.* (2022) “High-throughput proteomics of nanogram-scale samples with Zeno SWATH MS,” *eLife*, 11. doi:10.7554/eLife.83947.

Watt, A.J. *et al.* (2007) “Development of the mammalian liver and ventral pancreas is dependent on GATA4,” *BMC developmental biology*, 7(1), p. 37. doi:10.1186/1471-213X-7-37.

Weinberger, L. *et al.* (2016) "Dynamic stem cell states: naive to primed pluripotency in rodents and humans," *Nature reviews. Molecular cell biology*, 17(3), pp. 155–169. doi:10.1038/nrm.2015.28.

Whyte, W.A. *et al.* (2013) "Master transcription factors and mediator establish super-enhancers at key cell identity genes," *Cell*, 153(2), pp. 307–319. doi:10.1016/j.cell.2013.03.035.

Wierer, M. and Mann, M. (2016) "Proteomics to study DNA-bound and chromatin-associated gene regulatory complexes," *Human molecular genetics*, 25(R2), pp. R106–R114. doi:10.1093/hmg/ddw208.

Williams, R.L. *et al.* (1988) "Myeloid leukaemia inhibitory factor maintains the developmental potential of embryonic stem cells," *Nature*, 336(6200), pp. 684–687. doi:10.1038/336684a0.

Wilm *et al.* (1996) "Femtomole sequencing of proteins from polyacrylamide gels by nano-electrospray mass spectrometry," *Nature*, 379(6564), pp. 466–469. doi:10.1038/379466a0.

Wilm, M.S. and Mann, M. (1994) "Electrospray and Taylor-Cone theory, Dole's beam of macromolecules at last?," *International journal of mass spectrometry and ion processes*, 136(2–3), pp. 167–180. doi:10.1016/0168-1176(94)04024-9.

Wilmut, I. *et al.* (1997) "Viable offspring derived from fetal and adult mammalian cells," *Nature*, 385(6619), pp. 810–813. doi:10.1038/385810a0.

Wiśniewski, J.R. *et al.* (2009) "Universal sample preparation method for proteome analysis," *Nature methods*, 6(5), pp. 359–362. doi:10.1038/nmeth.1322.

Wong, C.C., Qian, Y. and Yu, J. (2017) "Interplay between epigenetics and metabolism in oncogenesis: mechanisms and therapeutic approaches," *Oncogene*, 36(24), pp. 3359–3374. doi:10.1038/onc.2016.485.

Wong, P.G. *et al.* (2010) "Chromatin unfolding by Cdt1 regulates MCM loading via opposing functions of HBO1 and HDAC11-geminin," *Cell cycle* , 9(21), pp. 4351–4363. doi:10.4161/cc.9.21.13596.

Wooderchak, W.L., Zhou, Z.S. and Hevel, J. (2008) "Assays for S-adenosylmethionine (AdoMet/SAM)-dependent methyltransferases," *Current Protocols in Toxicology*, 38(1), pp. 4–26. doi:10.1002/0471140856.tx0426s38.

Woodward, C.E. and Crawford, C.K. (1963) "DESIGN OF A QUADRUPOLE MASS SPECTROMETER," *Technical Report 176* [Preprint]. doi:https://www.osti.gov/biblio/4757193.

Wray, J. *et al.* (2011) "Inhibition of glycogen synthase kinase-3 alleviates Tcf3 repression of the pluripotency network and increases embryonic stem cell resistance to differentiation," *Nature cell biology*, 13(7), pp. 838–845. doi:10.1038/ncb2267.

Wu, F.-R. *et al.* (2014) "H3K27me3 may be associated with Oct4 and Sox2 in mouse preimplantation embryos," *Genetics and molecular research: GMR*, 13(4), pp. 10121–10129. doi:10.4238/2014.December.4.6.

Wu, H. *et al.* (2011) "Dual functions of Tet1 in transcriptional regulation in mouse embryonic stem cells," *Nature*, 473(7347), pp. 389–393. doi:10.1038/nature09934.

Wu, K. *et al.* (2020) "SETDB1-mediated cell fate transition between 2C-like and pluripotent states," *Cell reports*, 30(1), pp. 25-36.e6. doi:10.1016/j.celrep.2019.12.010.

Xiang, Y. *et al.* (2020) "Epigenomic analysis of gastrulation identifies a unique chromatin state for primed pluripotency," *Nature genetics*, 52(1), pp. 95–105. doi:10.1038/s41588-019-0545-1.

Xiao, Y. *et al.* (2021) "HBO1 is a versatile histone acyltransferase critical for promoter histone acylations," *Nucleic acids research*, 49(14), pp. 8037–8059. doi:10.1093/nar/gkab607.

Xie, H., Bandhakavi, S. and Griffin, T.J. (2005) "Evaluating preparative isoelectric focusing of complex peptide mixtures for tandem mass spectrometry-based proteomics: a case study in profiling chromatin-enriched subcellular fractions in Saccharomyces cerevisiae," *Analytical chemistry*, 77(10), pp. 3198–3207. doi:10.1021/ac0482256.

Xue, L. *et al.* (2013) "Global expression profiling reveals genetic programs underlying the developmental divergence between mouse and human embryogenesis," *BMC genomics*, 14(1), p. 568. doi:10.1186/1471-2164-14-568.

Yadav, T., Quivy, J.-P. and Almouzni, G. (2018) "Chromatin plasticity: A versatile landscape that underlies cell fate and identity," *Science*, 361(6409), pp. 1332–1336. doi:10.1126/science.aat8950.

Yang, P. *et al.* (2019) "Multi-omic profiling reveals dynamics of the phased progression of pluripotency," *Cell systems*, 8(5), pp. 427-445.e10. doi:10.1016/j.cels.2019.03.012.

Yang, S.-H. *et al.* (2014) "Otx2 and Oct4 drive early enhancer activation during embryonic stem cell transition from naive pluripotency," *Cell reports*, 7(6), pp. 1968–1981. doi:10.1016/j.celrep.2014.05.037.

Yates, J.R., III (2011) "A century of mass spectrometry: from atoms to proteomes," *Nature methods*, 8(8), pp. 633–637. doi:10.1038/nmeth.1659.

Ye, Z. and Sarkar, C.A. (2018) "Towards a quantitative understanding of cell identity," *Trends in cell biology*, 28(12), pp. 1030–1048. doi:10.1016/j.tcb.2018.09.002.

Yilmaz, A. *et al.* (2018) "Defining essential genes for human pluripotent stem cells by CRISPR–Cas9 screening in haploid cells," *Nature cell biology*, 20(5), pp. 610–619. doi:10.1038/s41556-018-0088-1.

Ying, Q.-L. *et al.* (2008) “The ground state of embryonic stem cell self-renewal,” *Nature*, 453(7194), pp. 519–523. doi:10.1038/nature06968.

Yu, Q. *et al.* (2020) “Benchmarking the orbitrap tribrid eclipse for next generation multiplexed proteomics,” *Analytical chemistry*, 92(9), pp. 6478–6485. doi:10.1021/acs.analchem.9b05685.

Yuan, P. *et al.* (2009) “Eset partners with Oct4 to restrict extraembryonic trophoblast lineage potential in embryonic stem cells,” *Genes & development*, 23(21), pp. 2507–2520. doi:10.1101/gad.1831909.

Zaret, K.S. (2020) “Pioneer transcription factors initiating gene network changes,” *Annual review of genetics*, 54(1), pp. 367–385. doi:10.1146/annurev-genet-030220-015007.

Zemach, A. *et al.* (2010) “Genome-wide evolutionary analysis of eukaryotic DNA methylation,” *Science*, 328(5980), pp. 916–919. doi:10.1126/science.1186366.

Zhang, C. *et al.* (2022) “Profiling and functional characterization of maternal mRNA translation during mouse maternal-to-zygotic transition,” *Science advances*, 8(5), p. eabj3967. doi:10.1126/sciadv.abj3967.

Zhang, X. *et al.* (2008) “Esrrb activates Oct4 transcription and sustains self-renewal and pluripotency in embryonic stem cells,” *The journal of biological chemistry*, 283(51), pp. 35825–35833. doi:10.1074/jbc.M803481200.

Zheng, H. *et al.* (2016) “Resetting Epigenetic Memory by Reprogramming of Histone Modifications in Mammals,” *Molecular cell*, 63(6), pp. 1066–1079. doi:10.1016/j.molcel.2016.08.032.

Zhou, W. *et al.* (2012) “HIF1α induced switch from bivalent to exclusively glycolytic metabolism during ESC-to-EpiSC/hESC transition,” *The EMBO journal*, 31(9), pp. 2103–2116. doi:10.1038/emboj.2012.71.

Zijlmans, D.W. *et al.* (2022) “Integrated multi-omics reveal polycomb repressive complex 2 restricts human trophoblast induction,” *Nature cell biology*, 24(6), pp. 858–871. doi:10.1038/s41556-022-00932-w.

Zinzen, R.P. *et al.* (2009) “Combinatorial binding predicts spatio-temporal cis-regulatory activity,” *Nature*, 462(7269), pp. 65–70. doi:10.1038/nature08531.

Zylicz, J.J. *et al.* (2015) “Chromatin dynamics and the role of G9a in gene regulation and enhancer silencing during early mouse development,” *eLife*, 4, p. e09571. doi:10.7554/eLife.09571.

5 ANNEX

5.1 Original Publications

Ugur, E., de la Porte, A., Qin, W., Bultmann, S., Ivanova, A., Drukker, M., Mann, M., Wierer, M., & Leonhardt, H. (**2023**). **Comprehensive chromatin proteomics resolves functional phases of pluripotency and identifies changes in regulatory components**. Nucleic Acids Research, 51(6), 2671-2690.

License: CC-BY-4.0

https://doi.org/10.1093/nar/gkad058

Published online 20 February 2023 *Nucleic Acids Research, 2023, Vol. 51, No. 6* ***2671–2690***
https://doi.org/10.1093/nar/gkad058

Comprehensive chromatin proteomics resolves functional phases of pluripotency and identifies changes in regulatory components

Enes Ugur [1,2], Alexandra de la Porte[3], Weihua Qin [1], Sebastian Bultmann [1], Alina Ivanova[1,7], Micha Drukker [3,4], Matthias Mann [2,6,*], Michael Wierer [2,5,*] and Heinrich Leonhardt [1,*]

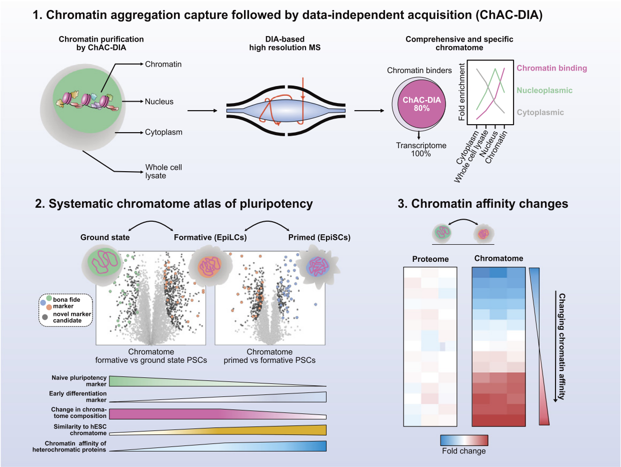
[1]Faculty of Biology and Center for Molecular Biosystems (BioSysM), Human Biology and BioImaging, Ludwig-Maximilians-Universität München, Munich 81377, Germany, [2]Department of Proteomics and Signal Transduction, Max-Planck Institute of Biochemistry, Martinsried 82152, Germany, [3]Institute of Stem Cell Research, Helmholtz Center Munich, Neuherberg 85764, Germany, [4]Division of Drug Discovery and Safety, Leiden Academic Centre for Drug Research (LACDR), Leiden University, Gorlaeus Building, 2333 CC RA Leiden, The Netherlands, [5]Proteomics Research Infrastructure, University of Copenhagen, DK-2200 Copenhagen, Denmark, [6]Novo Nordisk Foundation Center for Protein Research, Faculty of Health Sciences, University of Copenhagen, DK-2200 Copenhagen, Denmark and [7]Present address: Department of Totipotency, Max-Planck Institute of Biochemistry, Martinsried 82152, Germany

Received August 27, 2022; Revised January 11, 2023; Editorial Decision January 16, 2023; Accepted January 19, 2023

ABSTRACT

The establishment of cellular identity is driven by transcriptional and epigenetic regulators of the chromatin proteome - the chromatome. Comprehensive analyses of the chromatome composition and dynamics can therefore greatly improve our understanding of gene regulatory mechanisms. Here, we developed an accurate mass spectrometry (MS)-based proteomic method called Chromatin Aggregation Capture (ChAC) followed by Data-Independent Acquisition (DIA) and analyzed chromatome reorganizations during major phases of pluripotency. This enabled us to generate a comprehensive atlas of proteomes, chromatomes, and chromatin affinities for the ground, formative and primed pluripotency states, and to pinpoint the specific binding and rearrangement of regulatory components. These comprehensive datasets combined with extensive analyses identified phase-specific factors like QSER1 and JADE1/2/3 and provide a detailed foundation for an in-depth understanding of mechanisms that govern the phased progression of pluripotency. The technical advances reported here can be readily applied to other models in development and disease.

GRAPHICAL ABSTRACT


1. Chromatin aggregation capture followed by data-independent acquisition (ChAC-DIA)
Chromatin purification by ChAC-DIA
Chromatin
Nucleus
Cytoplasm
Whole cell lysate
DIA-based high resolution MS
Comprehensive and specific chromatome
Chromatin binders
ChAC-DIA 80%
Transcriptome 100%
Fold enrichment
Chromatin binding
Nucleoplasmic
Cytoplasmic
Cytoplasm
Whole cell lysate
Nucleus
Chromatin
2. Systematic chromatome atlas of pluripotency
Ground state
Formative (EpiLCs)
Primed (EpiSCs)
bona fide marker
novel marker candidate
Chromatome formative vs ground state PSCs
Chromatome primed vs formative PSCs
Naive pluripotency marker
Early differentiation marker
Change in chromatome composition
Similarity to hESC chromatome
Chromatin affinity of heterochromatic proteins
3. Chromatin affinity changes
Proteome
Chromatome
Changing chromatin affinity
Fold change

INTRODUCTION

DNA- and chromatin-binding proteins regulate gene expression and thereby govern cellular identity. During early embryonic development, the chromatin of pluripotent stem cells (PSCs) undergoes dynamic changes that are conserved among mammals (1–5). Pluripotency progresses in separate phases controlled by distinct signaling pathways and downstream transcription factors (3,6,7). Three major interme-

---

[*] To whom correspondence should be addressed. Tel: +49 89 2180 74232; Email: h.leonhardt@lmu.de
Correspondence may also be addressed to Michael Wierer. Tel: +45 35 3369 53; Email: michael.wierer@sund.ku.dk
Correspondence may also be addressed to Matthias Mann. Tel: +49 89 8578-2557; Email: mmann@biochem.mpg.de

© The Author(s) 2023. Published by Oxford University Press on behalf of Nucleic Acids Research.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

diate phases of pluripotency have been described: naive (also referred to as ground state), formative, and primed (3). Ground state PSCs harbor a homogeneously organized and transcriptionally permissive chromatin with high plasticity and low levels of repressive epigenetic marks (8,9). In transition to the formative phase, PSCs gain trimethylation of lysines 4 and 27 of histone H3 at promoters, and the exclusive ability to differentiate into primordial germ cells, while losing the expression of certain naive genes (1,10). Finally, at the primed phase, PSCs are partially fate determined, yet still share a core regulatory circuitry with earlier pluripotency phases (3,11–14).

Current systems-wide knowledge of pluripotency is primarily based on transcriptome and epigenome analyses, and chromatin accessibility data (1,10,14–16). For instance, previous studies revealed that major chromatin reorganization and compaction occur at the formative phase (10). However, how this chromatin reorganization affects chromatin proteome composition, the chromatome (17), remains unknown. Moreover, although the expression of chromatin binders, such as transcription factors, has been extensively studied in PSCs (18–21), changes in expression do not inevitably entail changes in chromatin association. The latter has not been studied comprehensively on a global scale and instead mostly has been studied by focusing on specific transcription factors or histone PTM-associated proteins (22–26). Therefore, the complete picture of the chromatome structure and dynamics in functional phases of pluripotency is still largely missing.

Previous attempts to quantify global chromatomes combined high-resolution mass spectrometry (MS) with the biochemical purification of native (27,28) or formaldehyde (FA) crosslinked chromatin (29–32). Although these methods greatly contributed to the understanding of the chromatome, they offer limited insights as they cannot detect low-abundant DNA-binding factors that are known to play key regulatory roles despite low abundance. Furthermore, current sample preparation strategies require millions of cells (15–50 mio.) and multiple purification steps, which impairs overall protein recovery and quantification (30,31). Therefore, the current view of the chromatome remains incomplete.

To overcome these difficulties, we developed a method that combines a new streamlined chromatin purification strategy, Chromatin Aggregation Capture (ChAC), with Data-Independent Acquisition (DIA) MS-based proteomics, a powerful strategy for rapid, accurate, and reproducible proteomics analysis with a broad dynamic range that allows identification of low-abundant proteins starting with 100–250k cells. Using this method, we generated accurate and comprehensive chromatome maps of mouse naive, formative and primed PSCs that cover 80% of transcribed chromatin binders in single MS runs. Our analysis of these datasets revealed striking chromatome changes between different functional phases of pluripotency and provided evidence for novel, low-abundant chromatin binders that are dynamically regulated in pluripotency transitions. Additionally, by comparing the abundance of proteins in chromatomes and proteomes, we were able to infer chromatin reorganizations mediated by differential affinities or subcellular localizations. Finally, we applied this approach to chromatomes of human PSCs to provide a mouse-to-human comparison of the pluripotency chromatome. Collectively, we present a comprehensive atlas of proteomes and chromatomes for the three pluripotency phases, thus revealing previously unknown details about how cell identity governing proteins are recruited to or evicted from chromatin in the process of pluripotency transitions. We have made the datasets available and searchable on an interactive web application, accessible on: https://pluripotency.shinyapps.io/Chromatome_Atlas/.

MATERIALS AND METHODS

Cell culture

Naive J1 mESCs were cultured in serum-free media consisting of: N2B27 (50% neurobasal medium (Life Technologies), 50% DMEM/F12 (Life Technologies)), 2i (1 μM PD032591 and 3 μM CHIR99021 (Axon Medchem, Netherlands)), 1000 U/ml recombinant leukemia inhibitory factor (LIF, Millipore), and 0.3% BSA (Gibco), 2 mM L-glutamine (Life Technologies), 0.1 mM β-mercaptoethanol (Life Technologies), N2 supplement (Life Technologies), B27 serum-free supplement (Life Technologies), and 100 U/ml penicillin, 100 μg/ml streptomycin (Sigma). Formative EpiLCs were derived by differentiating naive mESCs (33) for 48 h using the same serum-free media for naive mESCs devoid of 2i, LIF, and BSA and supplemented with 10 ng/ml Fgf2 (R&D Systems), 20 ng/ml Activin A (R&D Systems) and 0.1× Knockout Serum Replacement (KSR) (Life Technologies). Both, naive mESCs and EpiLCs, were cultured on 0.2% gelatin-treated flasks. The media of EpiLCs was changed once after 24 h and all cells were harvested after 48 h. Cells were tested negative for Mycoplasma contamination by PCR.

Identical culture conditions for mouse formative and primed as well as human ESCs

129S2C1a mouse EpiSCs (34) and J1 EpiLCs that were compared directly to human ESCs H9 were cultured in UPPS medium consisting of StemMACS iPS Brew XF (Miltenyi Biotec) supplemented with 1 μM IWR-1 (Sigma) and 0.5 μM CHIR (Tocris) (35). ESCs, EpiSCs and compared EpiLCs were cultured on plates coated with Matrigel (Corning) diluted 1:100 in DMEM/F-12 (Thermo Fisher Scientific).

For all experiments, cells were differentiated/cultured in three independent flasks and are therefore considered to be three biological replicates. Cells were split upon harvesting for total proteome (5×10^6 cells per replicate) and chromatome (15×10^6 cells per replicate) analyses and flash-frozen. The following descriptions are based on the above-mentioned amounts. Systematic downscaling showed that as few as 1×10^4 to 1×10^5 cells per replicate may suffice (see also Materials and Methods details).

Total proteome sample preparation

Previously flash-frozen samples were quickly placed on ice and pellets were solubilized in 200 μl lysis buffer (6 M guanidinium Chloride, 100 mM Tris–HCl pH 8.5, 2 mM

DTT) and heated for 10 min at 99°C under constant shaking at 1400 rpm. Subsequently, samples were sonicated at 4°C in 30 s on/off intervals for 15 cycles using a Bioruptor® Plus sonication instrument (Diagenode) at high-intensity settings. If the viscosity of the samples was sufficiently reduced, protein concentrations were estimated, otherwise, sonication was repeated. For concentration measurements, the Pierce™ BCA Protein Assay Kit (23225, Thermo Fisher Scientific) was employed following the manufacturer's instructions. After at least 20 min of incubation with 40 mM chloroacetamide, 30 μg of each proteome sample was diluted in a 30 μl lysis buffer supplemented with CAA and DTT. Samples were diluted in 270 μl digestion buffer (10% acetonitrile, 25 mM Tris–HCl pH 8.5, 0.6 μg Trypsin/sample (Pierce™ Trypsin Protease, 90058, Thermo Fisher Scientific) and 0.6 μg/sample LysC (Pierce™ LysC Protease, 90051, Thermo Fisher Scientific) and proteins digested for 16 h at 37°C with constant shaking at 1100 rpm.

To stop protease activity 1% (v/v) trifluoroacetic acid (TFA) was added the next day and samples were loaded on self-made StageTips consisting of three layers of SDB-RPS matrix (Empore) (36) that were previously equilibrated by 0.1% (v/v) TFA. After loading, two washing steps with 0.1% (v/v) TFA were scheduled, and peptides were eluted by 80% acetonitrile and 2% ammonium hydroxide. Upon evaporation of the eluates in a SpeedVac centrifuge, samples were resuspended in 20 μl 0.1% TFA and 2% acetonitrile. After complete solubilization of peptides by constant shaking for 10 min at 2,000 rpm, peptide concentrations were estimated on a Nanodrop™ 2000 spectrophotometer (Thermo Fisher Scientific) at 280 nm.

Chromatin aggregation capture

Previously flash-frozen samples were quickly placed on ice and pellets were solubilized in 1 ml cellular lysis buffer (20 mM HEPES pH 7.4, 10 mM NaCl, 3 mM $MgCl_2$, 0.1% NP40, freshly added 1× cOmplete™ EDTA-free Protease Inhibitor Cocktail (04693132001, Roche)) and incubated for 10 min on ice. Nuclei were pelleted by centrifugation (2300 g, 5 min, 4°C) and the supernatant was discarded. In the differential fraction analysis (Figure 2A), the supernatant was saved as the cytosolic fraction. Upon a second wash of the nuclei pellet with the cellular lysis buffer, the nuclei were taken into 3 ml crosslinking buffer (PBS pH 7.4 (806552, Sigma), 1× cOmplete™ EDTA-free Protease Inhibitor Cocktail). Formaldehyde (28906, Thermo Fisher Scientific) was added to a final concentration of 1% and samples were incubated for 10 min on an orbital shaker at room temperature. Excess formaldehyde was then quenched by 125 mM Glycine for 5 min and crosslinked cells were washed twice with ice-cold PBS. Nuclei were lysed in 300 μl SDS buffer (50 mM HEPES pH 7.4, 10 mM EDTA pH 8.0, 4% UltraPure™ SDS Solution (24730020, Invitrogen), freshly added 1× cOmplete™ EDTA-free Protease Inhibitor Cocktail) by gentle pipetting. After 10 min incubation at room temperature, 900 μl freshly prepared Urea buffer (10 mM HEPES pH 7.4, 1 mM EDTA pH 8.0, 8 M urea (U4883, Sigma)) was added. Tubes were carefully inverted 7 times and centrifuged at 20 000 g and room temperature for 30 min. The supernatant was discarded without perturbing the pellet. The pellet was resuspended in 300 μl Sonication buffer (10 mM HEPES pH 7.4, 2 mM $MgCl_2$, freshly added 1× cOmplete™ EDTA-free Protease Inhibitor Cocktail). Before sonication, two additional wash steps can be scheduled (one SDS and urea wash and one SDS only wash) (30), but to our hands, this did not notably improve the chromatin enrichment efficiency. The chromatin samples were sonicated using a Bioruptor® Plus at 4°C for 15 cycles (30 s on, 60 s off). The protein concentration was estimated by the Pierce™ BCA Protein Assay Kit.

Next, protein aggregation capture (PAC) was performed. Here 1000 μg of undiluted Sera-Mag™ beads (1 mg, GE24152105050250, Sigma) per 100 μg chromatin solution were washed three times by 70% acetonitrile. 300 μl of the chromatin solution corresponding to 100 μg was added after the last wash to the beads and 700 μl 100% acetonitrile was added to each sample. Chromatome-bead mixtures were vortexed. After 10 min incubation on the bench, the samples were again vortexed and rested on the bench. Samples were then placed into a magnetic rack. A first wash followed this with 700 μl 100% acetonitrile, a second wash with 1 ml 95% acetonitrile, and a third wash with 1 ml 70% ethanol. The remaining ethanol was allowed to evaporate and beads were resuspended in 400 μl 50 mM HEPES pH 8.5 supplemented with fresh 5 mM TCEP and 5.5 mM CAA. Samples were incubated for 30 min at room temperature upon which LysC (1:200) and Trypsin (1:100) were added. Proteins were digested overnight at 37°C. From this step on, samples were treated exactly like the total proteome samples.

Chromatin aggregation capture of <1 million cells

Chromatin aggregation capture for sub-million amounts of cells was performed with some additional modifications to the standard protocol. Here, cells were directly harvested into a DNAse-/RNase-free 1.5 ml tube (0030108051, Eppendorf). Nuclei were then isolated by 0.5 ml of cellular lysis buffer and the nuclei pellet was resuspended in 666 μl crosslinking buffer. After crosslinking with 1% formaldehyde and subsequent formaldehyde quenching with 125 mM Glycine, the chromatin extraction was performed again by SDS and Urea washes with careful pipetting so that nothing would stick to the pipette tip. Of note, with <100 000 cells the chromatin is not visually pelleted but rather a smear that spreads at the wall of the tube. For 10 000 cells even this smear is not visible anymore and it is advised to use a thermal shaker at 1,500 rpm instead of pipetting. For 10 000–250 000 cells the protein yield after sonication was between 10–16 μg. Here, we used 10 μg as input for the PAC purification and 1500 μg magnetic beads per replicate since smaller amounts require a higher bead-to-protein ratio (37). After the peptide cleanup, these samples were resuspended in 8 μl of 0.1% TFA and 2% acetonitrile.

Chromatin immunoprecipitation for MS analysis

Chromatin immunoprecipitation for subsequent MS analysis (ChIP-MS) using a KAT7 (Abcam, ab70183), H3K4me3 (Abcam, ab8580), H3K9me3 (Abcam, ab8898) or normal rabbit IgG (Cell Signaling Technology, #2729) antibody was performed in triplicates in naive, formative and primed PSCs. ChIP-MS was performed like previously described (38–40), but without nuclei isolation and MNase digestion. Briefly, for each replicate, independently grown 10 × 10^6 cells were harvested and crosslinked in 1% paraformaldehyde. Lysis of cells was performed in IP buffer (1.7% Triton X-100, 100 mM NaCl, 50 mM Tris–HCl pH 8.0, 5 mM EDTA pH 8.0, 0.3% SDS, and freshly added 1× protease inhibitor cocktail). After 10 min incubation on ice, samples were sonicated for 15 min in a Bioruptor Plus (30 s on/off cycles, Diagenode). Shearing efficiency was checked after overnight reverse crosslinking and proteinase K digestion of samples on a 1% agarose gel. Shearing had to be repeated twice to reach an average DNA length of ∼150–1000 bp. Protein concentrations were estimated by BCA assay (Thermo). Samples were subsequently diluted to 1 mg/ml in 1 ml. 2 μg of the antibody was added to each replicate and samples were incubated O/N at 4°C under constant rotation. 80 μl of protein A sepharose bead slurry volume was added to each sample. After two hours of incubation at 4°C and under constant rotation, beads were washed three times by a low salt buffer (50 mM HEPES pH 7.5, 140 mM NaCl, 1% Triton X-100) and once by a high salt buffer (50 mM HEPES pH 7.5, 500 mM NaCl, 1% Triton X-100). In case of histone pulldowns, a third wash buffer was used (50 mM HEPES pH 7.5, 250 mM LiCl, 1% Triton X-100) after the high salt wash. Samples were then washed three times by TBS. Supernatants were discarded and beads were resuspended in 50 μl 2 mM DTT for 30 min at 37°C and subsequently 40 mM CAA for 5 min at 37°C (both diluted in 2 M Urea and 50 mM Tris–HCl pH 7.5). Then proteins were on-bead digested by Trypsin (20 μg/ml) O/N at 25°C. The next day, protease activity was stopped by 1% TFA and peptides were cleaned up on StageTips consisting of three layers of C18 material (Empore) (36). After elution from StageTips peptides were speedvac dried and resuspended in 20 μl of A* buffer (0.1% TFA and 2% acetonitrile). Peptide concentrations were estimated on a Nanodrop™ 2000 spectrophotometer (Thermo Fisher Scientific) at 280 nm.

Acid histone extraction

5 Mio. cells were harvested and nuclei were isolated by cellular lysis buffer (20 mM HEPES pH 7.4, 10 mM NaCl, 3 mM $MgCl_2$, 0.1% NP40, freshly added 1× cOmplete™ EDTA-free Protease Inhibitor Cocktail (04693132001, Roche)) and histones were extracted by 0.2 N HCl at a density of 5 Mio. nuclei/500 μl. Samples were incubated O/N at 4°C under constant rotation. After spinning at 16 000 g for 10 min at 4°C, the histone containing supernatant was acetone precipitated (5 volumes acetone: 1 volume histones). Histones were solubilized in DNase- and RNase-free water (Thermo Fisher Scientific, 10977035).

SDS-PAGE and western blot

8 μg of the chromatome and full proteome extracts and 1 μg of acid histone extracts were separated on SDS-PAGE. Proteins were transferred onto a nitrocellulose membrane and incubated with an antibody against QSER1 (Abcam, ab86072, 1:1000) or H3K9me3 (Abcam, ab8898, 1:1000). The secondary antibody of goat-anti-rabbit IgG (H + L)–HRP conjugate was used with a dilution of 1:5000. Blots were developed with the Pierce ECL western blotting substrate (Thermo Scientific, 32109) and scanned by the Amersham™ Imager 600 system.

Nanoflow LC–MS/MS measurements for proteomes and chromatomes

Peptides were separated prior to MS by liquid chromatography on an Easy-nLC 1200 (Thermo Fisher Scientific) on in-house packed 50 cm columns of ReproSilPur C18-AQ 1.9-μm resin (Dr Maisch GmbH). By employing a binary buffer system (buffer A: 0.1% formic acid and buffer B: 0.1% formic acid and 80% acetonitrile) with successively increasing buffer B percentage (from 5% in the beginning to 95% at the end) peptides were eluted for 120 min under a constant flow rate of 300 nl/min. Via a nanoelectrospray source, peptides were then injected into an Orbitrap Exploris™ 480 mass spectrometer (Thermo Fisher Scientific). Samples were scheduled in triplicates and a subsequent washing step while the column temperature was constantly at 60°C. Thereby the operational parameters were monitored in real-time by SprayQc.

DDA-based runs consisted of a top12 shotgun proteomics method within a range of 300–1650 m/z, a default charge state of 2, and a maximum injection time of 25 ms. The resolution of full scans was set to 60 000 and the normalized AGC target was set to 300%. For MS2 scans the orbitrap resolution was set to 15 000 and the normalized AGC target to 100%. The maximum injection time was 28 ms.

DIA-based runs employed an orbitrap resolution of 120 000 for full scans in a scan range of 350–1400 m/z. The maximum injection time was set to 45 ms. For MS2 acquisitions the mass range was set to 361–1033 with isolation windows of 22.4 m/z. A window overlap of 1 m/z was set as default. The orbitrap resolution for MS2 scans was at 30 000, the normalized AGC target was at 1000%, and the maximum injection time was at 54 ms. The tested DIA methods varied within the range of the isolation windows which were 37.3 m/z for in total of 18 windows and 16.8 m/z for in total of 40 windows.

MS data quantification

DIA-NN-based analysis of raw MS data acquired in DIA mode was performed by using version 1.7.17 beta 12 in ‘high accuracy’ mode. Instead of a previously measured precursor library, spectra and RTs were predicted by a deep learning-based algorithm and spectral libraries were generated from FASTA files. Cross-run normalization was established in an RT-dependent manner. Missed cleavages were set to 1. N-terminal methionine excision was activated and cysteine carbamidomethylation was set as a fixed modification. Proteins were grouped with the additional command ‘–relaxed-prot-inf’. Match-between runs was enabled and the precursor FDR was set to 1%.

The DIA raw files were analyzed with the Spectronaut Pulsar X software package (Biognosys, version 14.10.201222.47784) (41) applying the default Biognosys factory settings for DIA analysis (*Q*-value cutoff at precursor and protein level was set to 0.01). Imputation of missing values was disabled.

The DDA raw files were analyzed with MaxQuant 1.6.11.0 (42). 'Match between runs' was enabled and the FDR was adjusted to 1%, including proteins and peptides. The MaxLFQ algorithm was enabled for the relative quantification of proteins (43). Contaminants were defined by using the Andromeda search engine (44).

Statistical analyses

Downstream analysis of raw data output was performed with Perseus (version 1.6.0.9) (45). For the calculation of CVs, proteins or precursors with <2 out of 3 valid values were filtered out. For GO term counts the filtering was more strict and 3 out of 3 valid values were required. GO enrichment analyses of differentially enriched proteins (Figure 2A) were performed against the background of total identified proteins by employing a Benjamini-Hochberg FDR-corrected Fisher's Exact test. The analysis was thereby performed individually for each cluster. The functional enrichment analysis of proteins enriched by ChAC-DIA versus total proteome was performed by ranking proteins according to their enrichment in the ChAC-DIA fraction. The functional enrichment analysis was thereby based on STRING (46).

Student's t-tests were performed after imputation of missing values. The latter was always performed based on a Gaussian distribution relative to the standard deviations of measured values (width of 0.2 and a downshift of 1.8 standard deviations). Both, one- and two-sided *t*-tests were calculated with a permutation-based FDR of 0.05 and an $s0 = 1$ if not otherwise declared. For the multiple sample test based on an ANOVA (Figure 2A) we chose a minimal 1.5-fold change. We performed imputation for missing values, except for supplementary heatmaps that represent the data without imputation (Supplementary Figures S4–S9). Student's *t*-tests of normalized chromatomes were performed after calculating pairwise differences of ChAC-DIA and total proteome values. The complete catalog of proteins found in the naive, formative, and primed states can be found in Supplementary Table 3.

Correlations between samples in the differential fraction analysis experiment were calculated with Perseus, and the correlations between transcriptomes, proteomes, and chromatomes were calculated with GraphPad Prism (version 9.1.0).

Analysis of ChIP-MS experiments was performed by first filtering out proteins that were identified less than twice in a set of triplicates. A two-sided Student's *t*-test of the $\log_2$ transformed LFQ intensities (specific pulldown vs normal IgG pulldown) was performed to obtain significantly enriched proteins. By definition, a permutation-based false discovery rate of 5% and a fold change cut-off of $\log_2 = 1$ were applied. For stoichiometry calculations of the HBO1 complex, iBAQ values were $\log_2$ transformed and normalized to KAT7.

Web application development

Row-normalized *z*-scores for each significantly changing protein across the ChAC-DIA purification steps were generated for an interactive profile plot representation of the data. Significant chromatome and proteome changes during pluripotency were represented in an interactive heatmap as mean row differences of $\log_2$ intensities.

The web application was programmed using R Shiny with the following libraries besides base R packages for data processing and visualization: shiny (1.7.1), shinydashboard (0.7.2), shinyHeatmaply (0.2.0), plotly (4.10.0), heatmaply (1.3.0) and png (0.1–7). From the tidyverse (1.3.1) family we further utilized tidyr (1.2.0), dplyr (1.0.9), and ggplot2 (3.3.6).

RESULTS

Chromatin aggregation capture (ChAC) followed by data-independent MS acquisition (DIA) enables near-complete chromatome identification and high-precision quantification

We hypothesized that accurate and comprehensive chromatin proteomics could be accomplished by combining Chromatin Aggregation Capture (ChAC) with Data Independent Acquisition (DIA). The method comprises nuclei isolation and formaldehyde crosslinking followed by an initial chromatin enrichment under denaturing conditions similar to the Chromatin enrichment for proteomics (ChEP) protocol (30). This is followed by an additional purification based on the protein aggregation capture (PAC) technique (37) to generate specific and pure chromatin fractions, and achieve highly accurate quantification by DIA-based MS using the DIA-NN software package (47). Briefly, in DIA, all peptide precursors that fall into a predefined mass-to-charge (m/z) window are fragmented and acquired on the MS2-level compared to selecting the top N most abundant peptide ions in a typical Data-Dependent MS Acquisition experiment (DDA) (41,48–51). The application of DIA is especially relevant for the analysis of enriched cellular structures that consist of highly repetitive structural elements such as nucleosomes. Here, DIA is much more sensitive and accurate for lower abundant proteins than the more semi-stochastic DDA-based approach (52,53). To improve chromatome quantification accuracy and comprehensiveness, we optimized the protocol, MS acquisition strategy (Supplementary Figure S1A–C), and raw data analysis (Supplementary Figure S1D–H) (Supplementary Table 1).

To benchmark the chromatome protocol, we performed ChAC-DIA in naive mouse embryonic stem cells (mESCs) and compared it to a recent ChEP-based chromatome data set of mESCs (PRIDE: PXD011782) (54). ChAC-DIA identified over 2.5 times more proteins in half of the MS acquisition time (Figure 1B). In addition, ChAC-DIA quantified proteins more reproducibly with median coefficients of variation (CVs) of 4% compared to 16% in the previous study (Figure 1B and Supplementary Table 1). The CV differences were even more pronounced at the peptide ion level (Supplementary Figure S1E).

Next, we classified nuclear, DNA-binding, RNA-binding, or chromatin-binding proteins based on their

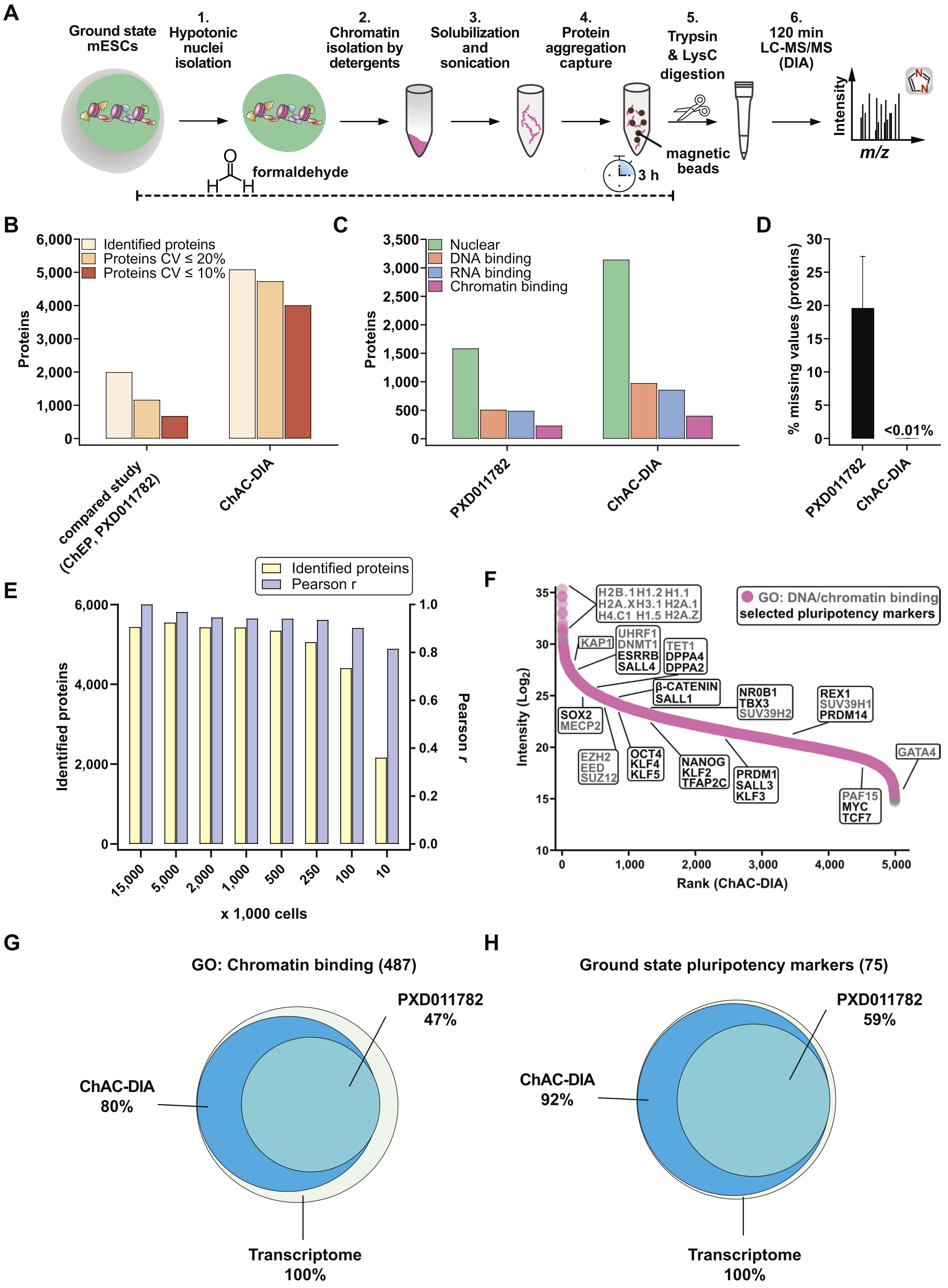


Figure 1. Chromatin aggregation capture (ChAC) followed by data-independent MS acquisition (DIA) enables near-complete chromatome identification and high-precision quantification. (**A**) Schematic workflow of ChAC-DIA. (**B**) Total numbers of identified proteins with representations of the coefficient of variation (CV) below 20% and 10%. ChAC-DIA results obtained in library-free mode by DIA-NN were benchmarked against a previous study based on the ChEP protocol (PRIDE: PXD011782). In both cases, mouse naive PSCs were used. (**C**) Total numbers of proteins falling into a gene ontology (GO) category. (**D**) Percentage of missing intensity values on protein level across replicates. (**E**) Total numbers of identified proteins and Pearson correlation coefficients of ChAC-DIA applied on different cell amounts. Pearson r reflects the correlation with the standard protocol comprising 15 Mio cells. (**F**) Protein abundance rank based on the ChAC-DIA-derived naive PSC chromatome. Chromatin binding proteins are highlighted in pink. Protein names in black indicate examples of *bona fide* pluripotency factors. Protein names in gray indicate other chromatin binders and the highest ranked nine proteins. (**G**) Venn diagram of proteins annotated as chromatin binding in ChAC-DIA, the compared study, and a transcriptome data set of naive PSCs (ArrayExpress: E-MTAB-6797). (**H**) Venn diagram of literature derived *bona fide* naive pluripotency factors identified by ChAC-DIA, the compared study, and a transcriptome data set of naive PSCs (ArrayExpress: E-MTAB-6797). See also Supplementary Figures S1 and S2.

Gene Ontology (GO) annotations (55). ChAC-DIA identified more than twice the number of nuclear and DNA-binding proteins, and three times more unique peptides of DNA-binding proteins as the previous ChEP method despite half of the required MS time (Figure 1C and Supplementary Figure S1E). Furthermore, annotated chromatin proteins had significantly fewer missing values across replicates (Figure 1D) and smaller CVs (Supplementary Figure 1H).

To make the method applicable to rare stem cell populations, we examined how input amounts affect the performance of our method. Cell numbers between 100K to 5 Mio. correlated well with the original protocol comprising 15 Mio. cells (Pearson correlation > 0.9) and 250k to 5 Mio. cells were sufficient for stable identification rates of over 5000 proteins (Figure 1E). Notably, ChAC-DIA with as few as 10k cells still resulted in over 2000 protein identifications. Ranking proteins quantified by ChAC-DIA according to their abundance revealed specific enrichment of histones and *bona fide* naive pluripotency factors as compared to a full proteome (Figure 1F, Supplementary Figure S2A, and Supplementary Table 1).

To further assess the comprehensiveness of ChAC-DIA, we compared the results to naive mESC transcriptome data. Among approximately 13000 expressed transcripts, 487 encode proteins annotated as chromatin binders, of which 80% were identified by ChAC-DIA (Figure 1G). Among *bona fide* naive pluripotency factors, 92% were identified by ChAC-DIA. Given that not all transcripts are translated into proteins with the same efficiency, we also compared the results obtained by ChAC-DIA to a full proteome analysis covering around 7000 proteins and observed that ChAC-DIA identified the same number of known chromatin binders that were also present in the full proteome data (Supplementary Figure S2B–D). We speculated that these annotated chromatin binding proteins might be missed due to overall low expression levels. However, we found that only some of these transcripts are lowly expressed (Supplementary Figure S2E). We, therefore, checked whether these missing proteins harbor additional cellular localizations and thus might not be frequently nuclear in naive mESCs. Indeed, these missing proteins are more often annotated cytoplasmic or membrane-associated proteins (Supplementary Figure S2F). Half of the missed proteins were identified and enriched in purified cytoplasmic fractions of naive mESCs (Supplementary Figure S2G).

Taken together, our results validated ChAC-DIA as a rapid and highly accurate method for analyzing the chromatome that uses only 100–250K cells and achieves unprecedented, almost complete chromatome coverage, including low-abundant proteins.

Chromatome mapping reveals a specific enrichment of chromatin-associated proteins in ground state PSCs

To define high-confidence chromatomes of ground state PSCs and thereby assess the specificity of chromatin enrichment by ChAC-DIA, we analyzed all fractions obtained during the chromatin purification in triplicates (i.e. whole cell lysate, cytoplasmic and nuclei fractions, ChAC-DIA after 1–3 washes). In total, we identified 8567 proteins, and the triplicates correlated well with each other ($R^2 > 0.95$). We observed that the correlation between the chromatin and nuclei fractions was weak ($R^2 = 0.66$) (Supplementary Figure S3A–D). Filtering for proteins with significantly different quantities between the fractions (ANOVA FDR < 0.05, fold change difference ≥ 1.5), resulted in 5464 proteins which explains the low correlation between the fractions. Unsupervised hierarchical cluster analysis of these proteins revealed nine distinct clusters (Figure 2A and Supplementary Table 2).

Two clusters (II and III), harboring 1141 proteins, were significantly enriched in the chromatomes (ChAC-DIA after 1–3 washes), but not in the nuclei or any other fraction. Therefore, proteins in clusters II and III comprise high-confidence chromatin binders. Importantly, well-known pluripotency proteins such as DNMT1, ESRRB, SALL4 or SOX2 are most abundant within these two clusters. Cluster II contained the highest enrichment of general chromatin-specific GO categories such as 'nucleosome' or 'nucleosomal DNA binding' (Supplementary Figure S3E and Supplementary Table 2). Euchromatic and heterochromatic proteins were equally enriched within this cluster. In cluster III, mitotic chromatin binders were overrepresented, resulting in GO categories such as 'mitotic prometaphase'. Clusters I and IV revealed significant enrichment of proteins in the nuclei fraction and a strong depletion in the chromatomes indicating that these two clusters captured nucleoplasmic proteins (Figure 2B). In line with this, well-characterized nucleoplasmic proteins such as RANGAP1 or CDK11B were categorized within these two clusters. In contrast, proteins in clusters V-IX were enriched for cytoplasm-specific GO categories (e.g. 'Golgi membrane', 'structural constituent of the ribosome' or 'Mitochondrion') (Supplementary Figure S3F). PCA analysis of the six different fractions confirmed that the three chromatin fractions are distinct from the nuclei fraction (Figure 2C).

Pluripotency phases are guided by distinct signaling pathways that lead to the translocation of otherwise cytoplasmic transcription factors into the nucleus (56–59). For example, naive pluripotent stem cells harbor active WNT and LIF pathways, while the GSK, FGF2 and Activin A pathways are inactive. Our data captured these features accurately, as we observed the chromatin-association of transcription factors linked to the WNT and LIF pathways, while those related to GSK, FGF2 and Activin A were mostly cytoplasmic (Figure 2D–H). For instance, β-CATENIN, the effector of WNT signaling, was equally distributed between the cytoplasmic and chromatin fractions, while being less abundant in the nuclear fraction (Figure 2D). We also observed chromatin enrichment of the LIF pathway transcription factors like KLF4 and KLF5, as well as STAT1 and STAT3, which, although being less abundant at chromatin than in the cytoplasm, still showed chromatin enrichment over the nuclear fraction (Figure 2E). In contrast, GSK, FGF2 and Activin A-related transcription factors were depleted from the chromatin fractions (Figure 2F–H). Taken together, we confirmed that ChAC-DIA selectively enriched components of the chromatome by reducing background proteins, even hard to separate mitochondrial or ribosomal proteins. This enabled the identification of not

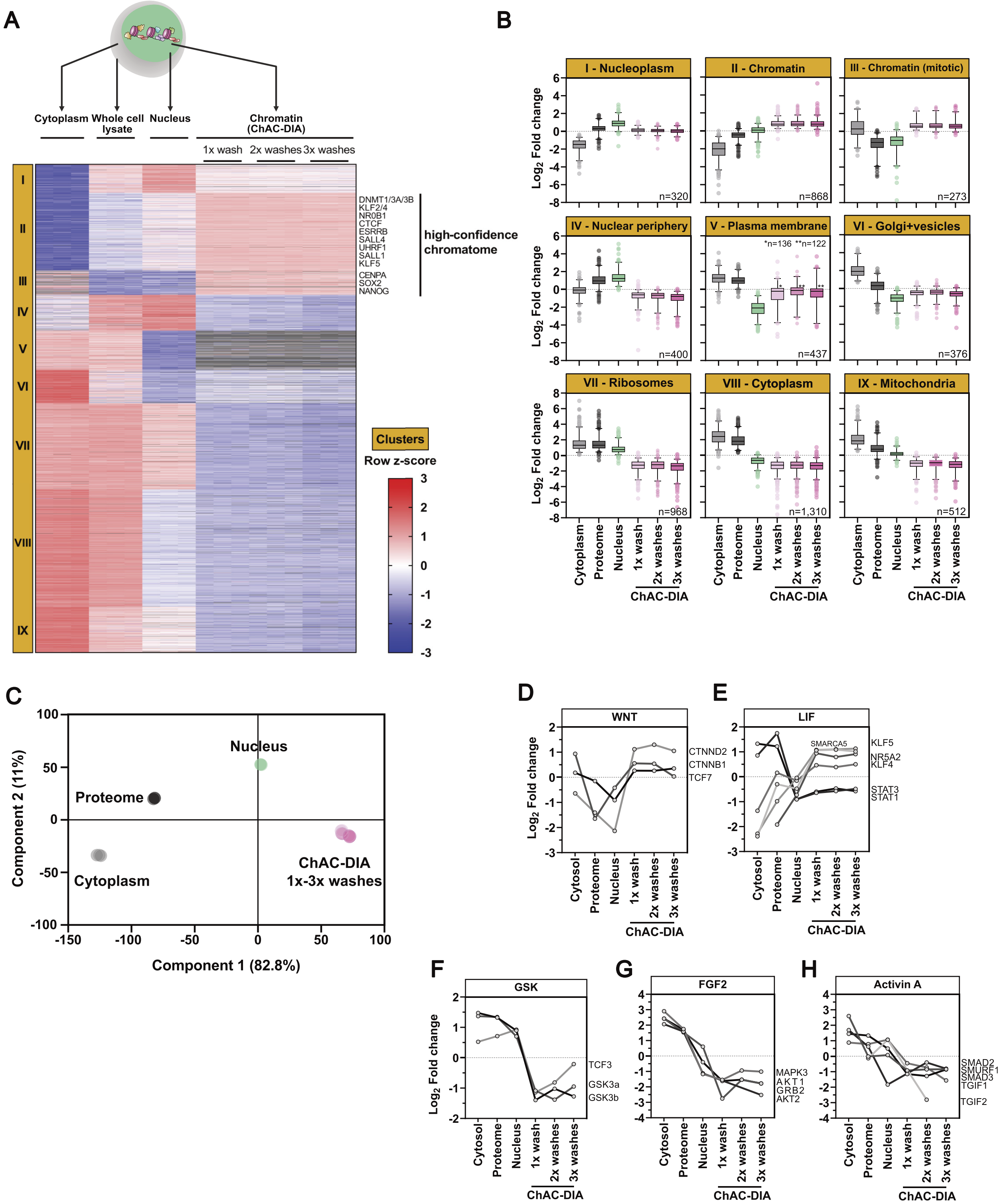


Figure 2. Chromatome mapping reveals a specific enrichment of chromatin-associated proteins in ground state PSCs. (**A**) Different fractions along the ChAC-DIA protocol were processed and measured. After ANOVA testing (FDR < 0.05, fold change difference ≥ 2) results were visualized in a heatmap generated by unsupervised hierarchical k-means clustering of z-scored intensities. In total nine clusters were identified. Proteins that are enriched only in the chromatome fractions are highlighted as the high-confidence chromatome. (**B**) Boxplot representation of row-scaled fold changes within each cluster. Cluster names are based on the most prominent GO-enriched terms (see Supplementary Figure 3E). (**C**) Principal component analysis (PCA) of the six different fractions. (**D–H**) Individual intensity profile plots of several proteins that are components of the WNT, LIF, GSK, FGF2 or Activin A pathways. See also Supplementary Figure S3.

only the majority of the annotated chromatome, but the expansion of the existent GO annotations. Thus, ChAC-DIA provides a high-confidence global map of the chromatome. Furthermore, analyzing chromatome data in combination with the overall proteome, and proteomes derived from different cellular fractions, allowed us to dissect events such as nuclear translocation and chromatin binding of proteins related to pluripotency-regulating pathways.

Chromatome atlas of mouse naive, formative and primed pluripotent stem cells identifies groups of chromatin proteins with distinct binding patterns

Two recent studies provided evidence that the formative phase is a discrete pluripotent state during embryonic development that is transcriptionally distinct from naive and primed pluripotency phases (1,10). To examine this further, we analyzed chromatomes of naive, formative, and primed PSCs (Figure 3A). We observed that 1403 proteins significantly changed in the chromatome during the differentiation of naive to formative PSCs, while the proteome revealed 1683 significantly regulated proteins (P value < 0.05, FC ≥ 2) (Figure 3A). In contrast, between formative and primed PSCs, only 859 proteins were significantly regulated on chromatome level and 1451 on proteome level. This suggests a more drastic reorganization of the chromatome during the transition from naive to formative pluripotency.

Next, we analyzed the chromatome changes based on a list of PSC phase-specific factors that we derived from the literature (Supplementary Table 3) (1,4,6,7,10,13,15,33,54,60–67). ChAC-DIA data confirmed that the abundance of the core pluripotency circuitry (OCT4, MYC, SOX2 and SALL4) is maintained throughout pluripotency; whereas state-specific markers displayed phase-dependent selective enrichment in the chromatome (Figure 3B–D and Supplementary Table 3). The naive chromatome was characterized by high levels of REX1, ESRRB, KLF4 and TET2 while the *de novo* methyltransferases DNMT3A and DNMT3B, OTX2, and OCT6 (or POU3F1), were highly enriched in the formative chromatome (Figure 3C). We observed a slight enrichment of lineage-specific transcription factors such as NES as early as the formative state.

In contrast to the formative chromatome, the primed chromatome was characterized by lower levels of early post-implantation-specific proteins like DPPA4 (15) and OCT6 (7) and higher levels of *bona fide* primed-specific transcription factors such as SOX1 (10) and SALL3 (60). Similarly, naive factors like ESRRB, HMCES and TET2 were further decreased in the primed chromatome while lineage-specific factors such as RAI1 and SIX6 (Figure 3D) were significantly enriched, which fits the partially fate-determined identity of primed PSCs. Among the primed-specific chromatin constituents, several histone H1 variants and high mobility group (HMG) proteins were also observed. The enrichment of these proteins governing chromatin structure and compaction could in part account for the previously described reduced chromatin plasticity and accessibility at the primed phase (1,5,10). Although major chromatome changes were already established at the formative state, these results demonstrate that formative and primed pluripotency are characterized by distinct chromatin landscapes.

These findings point to gradual chromatin recruitment or eviction of pluripotency governing factors during naive to primed transition. Interestingly, we observed similar chromatin-enrichment patterns for proteins related to epigenetic regulation, transcriptional regulation, and chromatin remodeling, as well as hundreds of zinc finger proteins with mostly unknown functions in pluripotency regulation (Supplementary Figures S4–S9). Approximately 70% of proteins harboring a zinc finger domain significantly change between naive and primed pluripotency, which fits well with the recently reported zinc finger protein-driven regulation of transposable elements during early embryonic development (68,69).

In summary, we provide the first systematic and near-comprehensive chromatome atlas of naive, formative, and primed PSCs (Supplementary Figures S4–S9, Supplementary Table 3) and provide an interactive web application for easy access to the data set (Supplementary Figure S10). We show that the chromatome reflects distinct features of pluripotency phases and a tightly regulated pluripotency phase transition process.

Identification of novel pluripotency phase-specific proteins through chromatome analysis

Using the comprehensive chromatome dataset we next sought to pinpoint novel pluripotency phase-specific proteins that bind chromatin in a similar manner to *bona fide* phase-specific proteins such as TBX3, OCT6 or SOX1 (Figure 4A–C). To achieve this, we ranked proteins according to their fold change between each pluripotency phase and observed differential enrichments of proteins associated with H3K4me3 or H3K9me3. For instance, we found that QSER1 increases at chromatin from naive to formative and decreases from formative to primed (Figure 4D, E). Previous studies have shown that QSER1, along with TET1, protects bivalent promoters from de novo methylation in human ESCs. (70). Our chromatome data shows that QSER1 and the *de novo* methyltransferases peak at the formative phase, potentially indicating a conserved role of QSER1 in mouse PSCs. Other H3K4me3-related proteins are preferentially enriched in the naive chromatome (e.g. KAT6B) or the primed chromatome (e.g. KAT6A, ZNF800).

Among the H3K9me3-associated proteins, we observed that two trimethyltransferases of H3K9, SUV39H1 and SUV39H2, increase at chromatin from naive to formative, while SUV39H1 decreases from formative to primed. To test whether SUV39H1/2 inhibition by their specific inhibitor Chaetocin could provide evidence for increased catalytic activity of these enzymes in formative vs naive pluripotent stem cells, we treated wild-type PSCs with or without Chaetocin and compared to Suv39h double knockout mESCs in both naive and formative states. We then quantified H3K9me3 abundance by western blot, which revealed lower levels of H3K9me3 in formative PSCs upon 0.1 μM Chaetocin treatment than in naive PSCs (Figure 4F). Our results suggest increased catalytic activities of SUV39H1/2 in formative PSCs, consistent with the increased chromatin binding of both enzymes revealed by

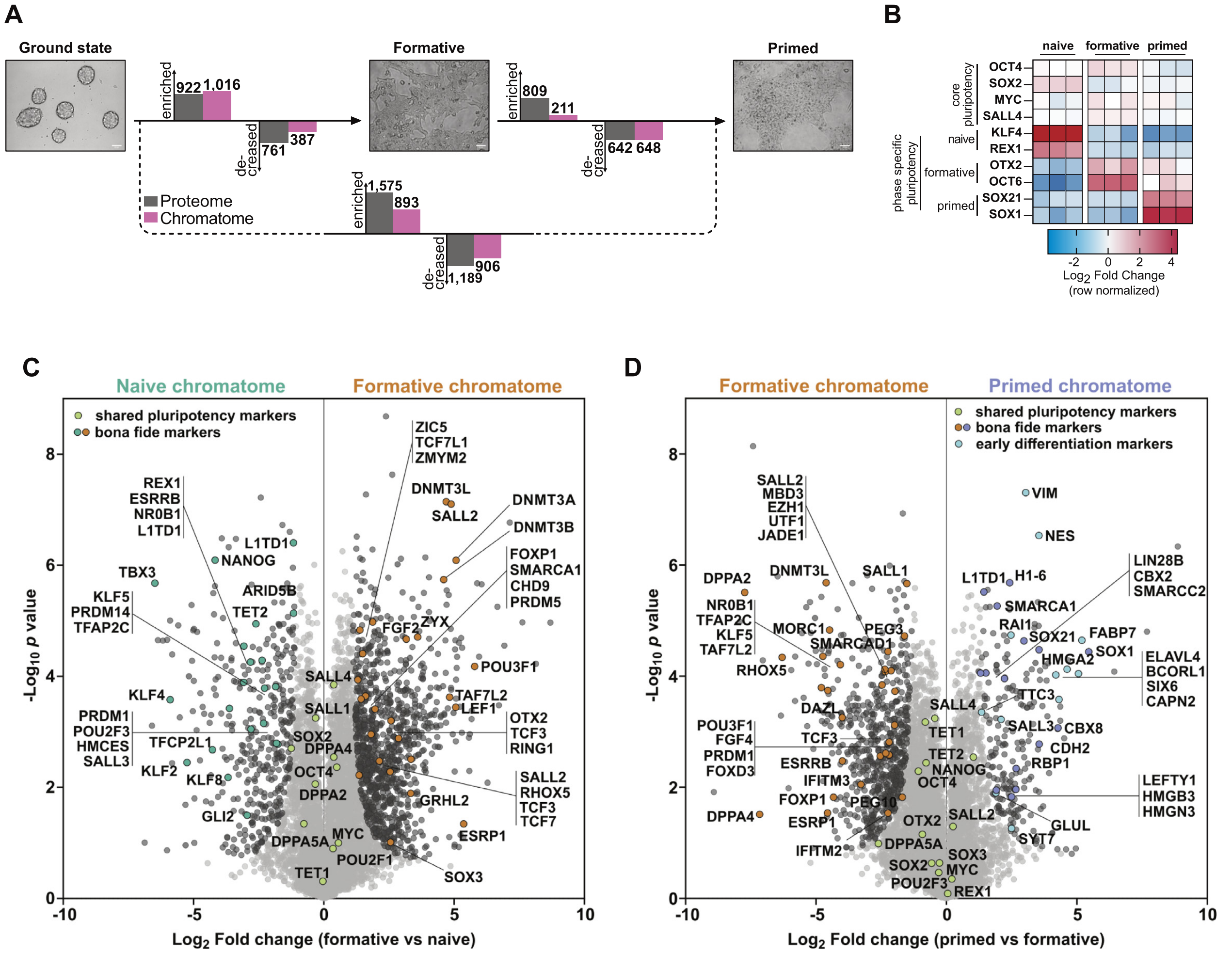


Figure 3. Chromatome atlas of mouse naive, formative, and primed pluripotent stem cells identifies groups of chromatin proteins with distinct binding patterns. (**A**) Schematic representation of compared cell lines and total significant changes between respective proteomes and chromatomes (Student's *t*-test, *P* value < 0.05, FC ≥ 2). (**B**) Heatmap representation of *bona fide* pluripotency factors. Fold changes are row-normalized by subtracting the mean $\log_2$ fold-change from each value. (**C, D**) Volcano plots of chromatomes based on Student's *t*-test displayed in (A). Light grey dots: not significantly enriched proteins. Black dots: significantly enriched proteins. Green dots: shared pluripotency factors. Blue dots: early differentiation markers. $n = 3$ biological replicates, meaning independently cultured/differentiated PSCs of the same genetic background. See also Supplementary Figures S4–S10.

ChAC-DIA. We further observed a SUV39H1-like pattern for DNMT3L and ZNF462. Proteins that continuously decreased in their chromatin association from naive to primed included LIRE1 and PHF11, while FLYWCH1, SUV39H2, UHRF2, CBX3, CBX5 and MKI67 increased from naive to primed.

To validate the global chromatome change of the described H3K4me3- and H3K9me3-associated proteins, we performed ChIP-MS of both histone PTMs and compared the ChAC-DIA results to the ChIP-MS data (Figure 4G–J). We observed a high level of similarity between the two datasets for well-described H3K4me3- or H3K9me3-associated proteins. However, some proteins showed slightly different levels in the global chromatome compared to specific regions with H3K9me3. A good example is FLYWCH1, a low-abundant chromatin binder at H3K9me3-rich regions which has not been detected in previous chromatome or proteome studies of PSCs (60,71). FLYWCH1 chromatin binding increases along with H3K9me3 from naive to primed PSCs (Supplementary Figure S6C) but is most abundant at H3K9me3 sites in formative PSCs, suggesting alternative mechanisms of chromatin association beyond H3K9me3 binding.

We further observed several chromatin-associated complexes among these phase-specific proteins (Supplementary Figure S9). One interesting example is the HBO1 complex, which acetylates several lysines at histones H3 and H4 and by this co-regulates the origin of replication licensing and MCM complex formation (72,73). The specificity of the complex is determined by the association of the mutually exclusive accessory subunits JADE1/2/3 and BRPF1/3 (74). Our chromatome data suggests that the core HBO1 complex (KAT7, ING4/5, MEAF6) remains at a constant level from naive to primed, while the accessory subunits are dynamically regulated. JADE1, BRPF1 and BRPF3 were mostly enriched in the naive chromatome, while JADE3 peaked at the formative phase and JADE2 peaked in the primed phase (Figure 4A–C, K). Since global

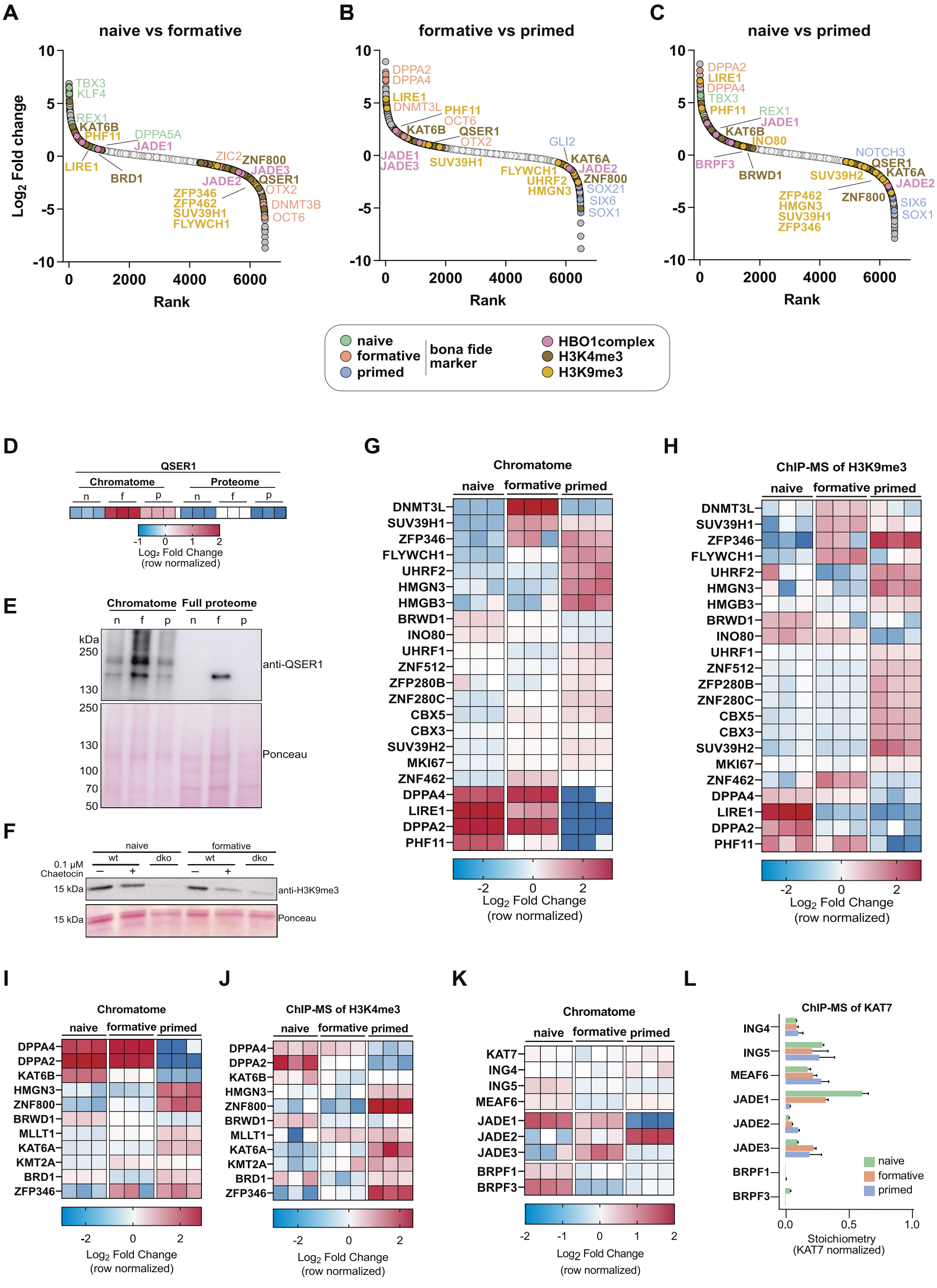


Figure 4. Identification of novel pluripotency phase-specific proteins through chromatome analysis. (**A–C**) Protein rank based on the $\log_2$ fold change between naive versus formative (A), formative versus primed (B), or naive versus primed (C) PSC chromatomes. *Bona fide* pluripotency phase-specific proteins are highlighted alongside H3K4me3-, H3K9me3- or HBO1 complex-associated proteins. Light grey dots are not significantly changing proteins while dark grey dots are significantly changing. (**D**) Heatmap representation of QSER1 abundance in chromatomes and proteomes of naive, formative and primed PSCs. Each replicate value was normalized to the mean of the row. (**E**) Western Blot of QSER1 in the chromatome and the whole cell lysate and ponceau staining of the respective western blot membrane. (**F**) Western blot of H3K9me3 upon chaetocin treatment (0.1 μM) in WT and Suv39h1/2 double knockout (dko) mESCs at the naive and formative phase and ponceau staining of the respective western blot membrane. (D–J) Heatmap representation of H3K9me3- (**G, H**) and H3K4me3-associated (**I, J**) proteins and their abundance in chromatomes (G, I) and respective ChIP-MS experiments (H, J) of naive, formative and primed PSCs. Each replicate value was normalized to the mean of the row. (**K**) Heatmap representation of HBO1 complex proteins and their abundance in chromatomes of naive, formative, and primed PSCs. Each replicate value was normalized to the mean of the row. (**L**) Bar diagram of KAT7-normalized protein stoichiometries after KAT7 ChIP-MS in naive, formative, and primed PSCs. Error bars represent the standard deviation of independent triplicates. See also Supplementary Figure S11.

chromatome changes might not reflect the actual changes within the HBO1 complex, we calculated the complex stoichiometries after performing ChIP-MS on the HBO1 catalytic subunit KAT7 (Figure 4L and Supplementary Figure S11). The ChIP-MS data revealed that KAT7 indeed interacts in a stable ratio with ING4/5 and MEAF6, but selectively interacts with JADE1/2/3 and hardly with BRPF1. This latter finding might hint towards a cell-type dependent BRPF1/3 interaction with KAT7 or more frequent interactions of BRPF1/3 with other complexes (e.g. MOZ/MORF complex, Supplementary Figure S9). The switch between JADE1/2/3 across pluripotency implies that the complex might target different lysines in a pluripotency phase-specific manner.

Collectively, we used the comprehensive chromatome dataset to identify novel pluripotency phase-specific proteins that bind chromatin in a manner similar to known phase-specific proteins. We found that especially proteins associated with H3K4me3 and H3K9me3 show phase-specific enrichment patterns and that these patterns can be confirmed by ChIP-MS.

Determination of relative chromatin binding reveals regulatory changes along pluripotency phases

Next, we correlated the transcriptome changes during the naive to formative transition (75) with the respective proteome and chromatome changes. As expected and previously reported (60,76,77) the proteome showed a moderately positive correlation with the transcriptome (Figure 5A), due to mechanisms regulating translation and protein stability. Consequently, transcriptome and chromatome showed the lowest correlation (Figure 5B) indicating that transcriptional data can only provide limited coverage of regulatory chromatin changes. Interestingly, the comparison of proteome and chromatome changes revealed also a moderate positive correlation (Figure 5C), pointing to mechanisms controlling chromatin binding and dissociation. In line with these observations, proteins related to active signaling pathways in postimplantation pluripotency like the FGF2, Activin A, and Notch pathways were differentially enriched in the chromatome, while they changed neither on transcriptome nor on proteome level.

Proteome-independent changes in the chromatome contain valuable information and point to either altered chromatin affinity or subcellular localization and availability of individual proteins (Figure 5D). We, therefore, computed proteome normalized chromatome changes to estimate the relative changes in chromatin binding. We subtracted the Log_2 chromatome-intensity of a protein from its mean Log_2 proteome intensity across triplicates and subsequently filtered for significant proteins by ANOVA testing ($\mathrm{FDR} < 0.05$ and $\mathrm{FC} > 2$) (Figure 5E and Supplementary Table 3). Based on our differential chromatin fraction analysis, we defined high-confidence chromatin binders as proteins that are significantly enriched in the chromatome over the proteome.

We observed that 1518 proteins significantly changed in relative chromatin binding from naive to primed pluripotency. Hierarchical clustering yielded five distinct clusters harboring proteins with different trends in relative chromatin binding across pluripotency phases. GO analysis of these five clusters against the background of total identified proteins revealed distinct functional categories (Benjamini-Hochberg $\mathrm{FDR} < 0.05$) (Figure 5F and Supplementary Table 3). In the cluster of proteins with a peak in relative chromatin binding at the formative phase (cluster II) categories related to signaling pathways like 'β-catenin degradation' or 'RAF activation' were enriched (Figure 5F). Importantly, cluster III showed an increased relative chromatin binding at the formative and primed phases and was enriched for categories associated with a repressive chromatin state like 'heterochromatin' or 'transcription corepressor activity'. More specifically, this cluster harbored essential heterochromatic proteins such as SETDB1, SETDB2, KAP1, CBX3 and CBX5 suggesting a functional relation of their formative and primed specific enrichment to the incremental heterochromatinization towards the exit from pluripotency. Interestingly, this cluster III was also enriched for GO categories related to 'SUMOylation of transcription factors', 'SUMOylation of chromatin organization proteins', and SUMOylation-dependent 'PML bodies'. In line with this observation, SUMOylation was reported to regulate heterochromatinization in naive mouse PSCs (78). Notably, histone H1.0, whose function in chromatin compaction depends also on its SUMOylation (79), peaked in its relative chromatin binding at the primed phase. These results suggest that besides the binding of classical heterochromatin factors, SUMOylation also contributes to heterochromatin formation at the formative and primed phases. Among the proteins with decreasing relative chromatin binding (clusters IV and V) are enzymes involved in DNA and histone demethylation or DNA repair like TDG, APOBEC3, NTHL1, KDM4C and KDM6A. Thus, lower levels of these proteins would translate into an increase of repressive epigenetic marks, which is expected to promote repressive chromatin states and reduce chromatin plasticity.

These findings are indicative of an increased chromatin affinity of heterochromatic proteins at the formative and primed phases which may enhance in turn further heterochromatinization and prepare pluripotent stem cells for differentiation.

The chromatome of conventionally cultured human ESCs is most similar to the mouse primed state

Previous reports compared the epigenome, transcriptome, and proteome of conventional human ESCs (hESCs) with mouse PSCs and have shown that hESCs are more similar to post-implantation mouse PSCs (34,60,80,81). Here, we used our method to examine the correspondence between different pluripotency states of hESCs and mouse PSCs. A Venn diagram representation of the high-confidence chromatomes for all three mouse PSCs and hESCs revealed an overlap of approximately 75% (Figure 6A and Supplementary Table 4). The strongest overlap was between proteins related to chromatin remodeling, histone modifications, and developmental processes (Supplementary Figure S12A). A PCA of the high-confidence chromatomes resulted in a clear separation of all three mouse PSCs from hESCs on PC1. PC2 in turn separates hESCs and

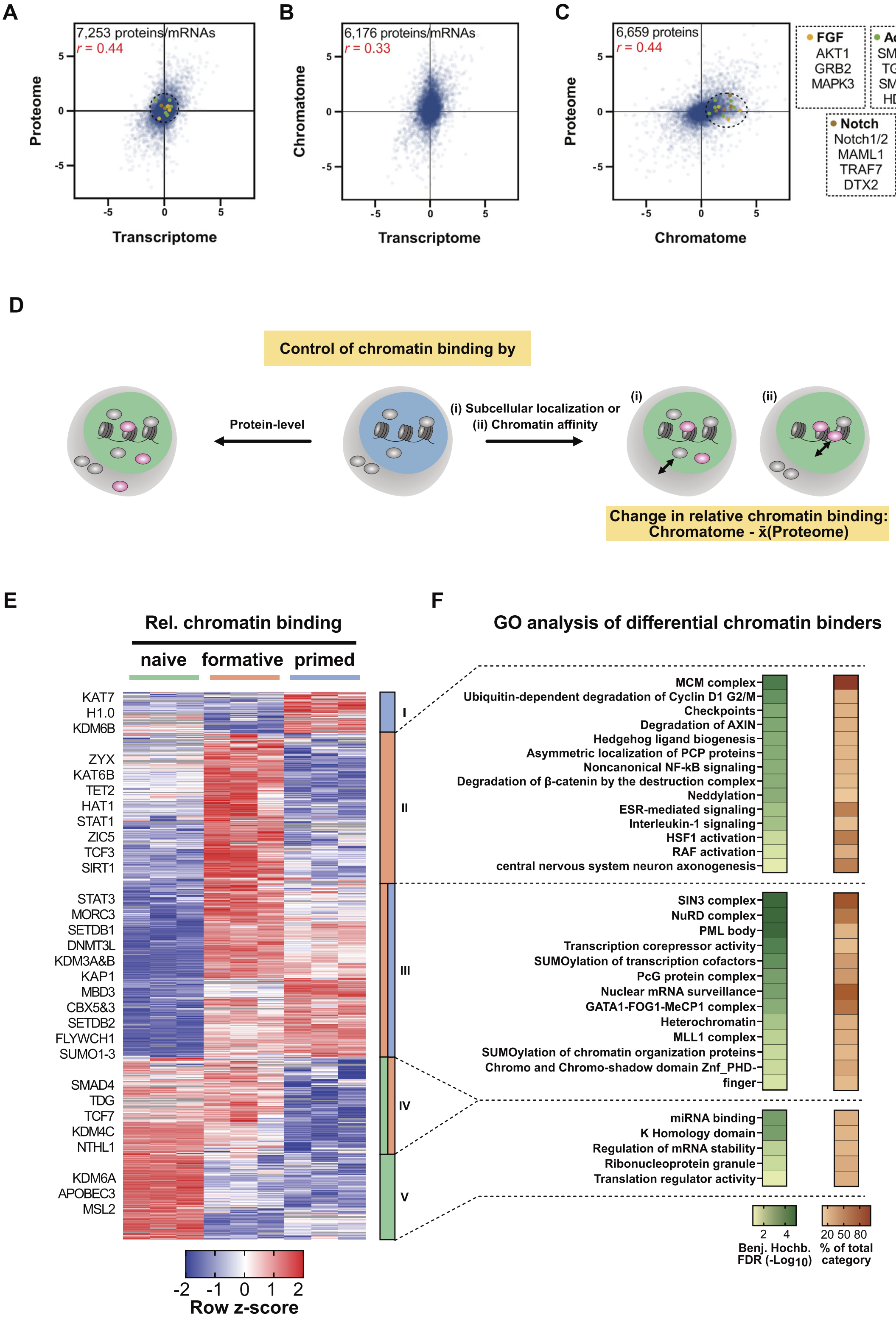


Figure 5. Relative chromatin binding reveals higher chromatin affinity of heterochromatic proteins in formative and primed PSCs. (**A**–**C**) Correlations of transcriptomes (ArrayExpress: E-MTAB-6797), proteomes, and chromatomes of formative vs naive PSCs of isogenic background (J1). Only proteins/mRNAs that were identified in both compared data sets are displayed. Pearson correlation coefficients are indicated in red. (**D**) Schematic representation of the relative chromatin binding concept. (**E**) Row *z*-scored relative chromatin binding changes between naive, formative, and primed PSCs filtered for ANOVA significant changes (FDR < 0.05, FC ≥ 2) and high-confidence chromatin binders. The relative chromatin binding was computed by subtracting the FC on chromatome level by the mean proteome FC of either formative versus naive or primed versus formative PSCs, respectively. (**F**) GO analyses of proteins enriched in clusters II, III or V of the hierarchical clustering from (E). As a comparison, the whole set of identified proteins was utilized.

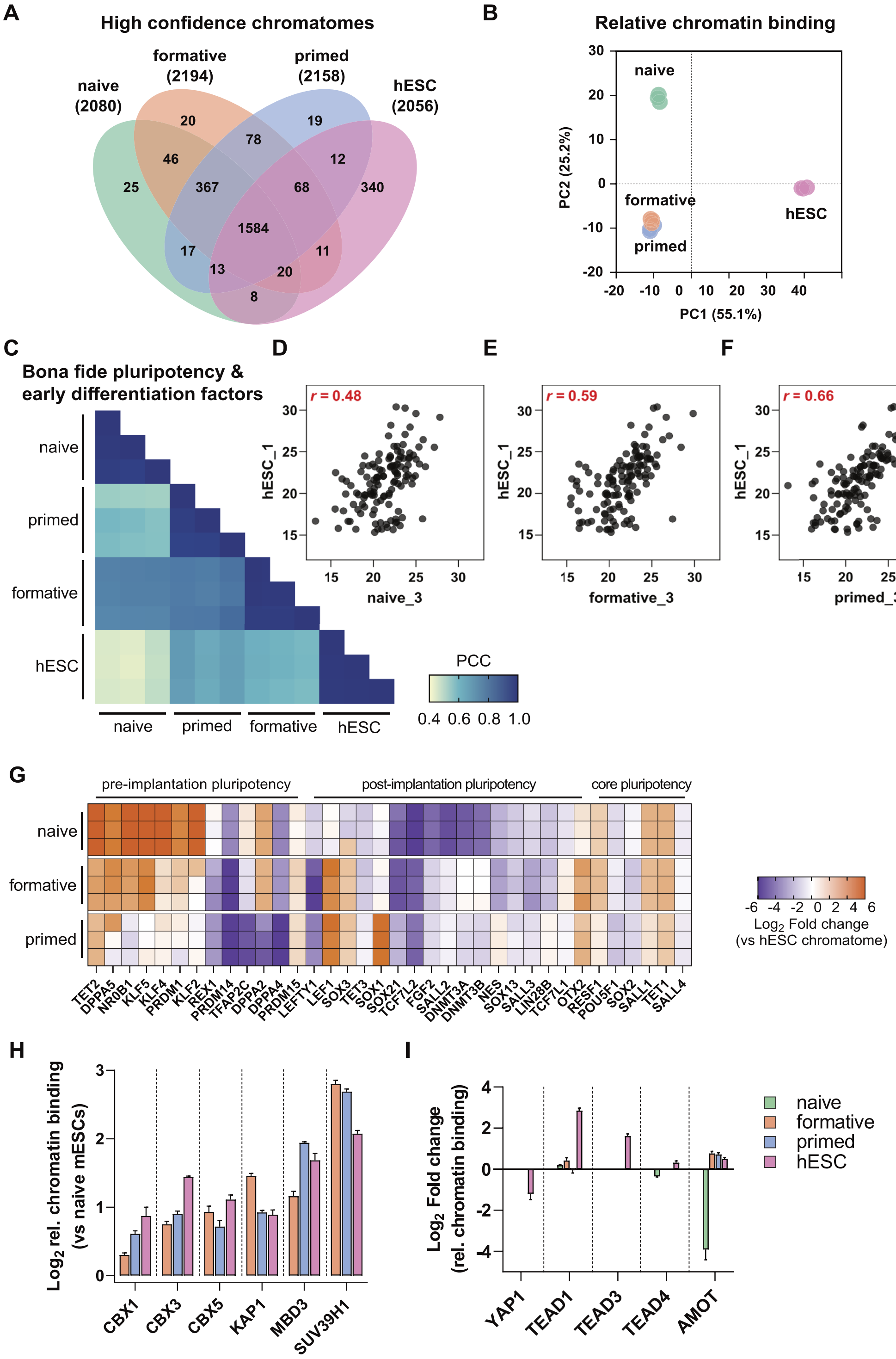


Figure 6. The chromatome of conventionally cultured human ESCs is most similar to the mouse primed state. (**A**) Venn diagram of high-confidence chromatomes in all tested cell lines. The high-confidence chromatome was defined by a Student's T-test between each cell line's chromatome vs proteome (Student's t-test, P value < 0.05 and FC ≥ 1.5). (**B**) PCA of the high-confidence chromatomes of the tested four cell lines on relative chromatin binding level. (**C**) Pearson correlations of chromatomes filtered for literature-derived *bona fide* pluripotency and differentiation factors. (D–F) Scatter plots of one replicate of hESCs versus naive (**D**), formative (**E**) or primed PSCs (**F**) and Pearson correlation coefficient from (C) are displayed in red. (**G**) hESC-normalized chromatomes from each mouse PSC to hESCs in $\log_2$. The selection comprises *bona fide* pluripotency factors. (**H**) Relative chromatin bindings of a selection of heterochromatic proteins after normalization to their respective relative chromatin bindings in naive mESCs. The bars represent mean values and the error bar is based on the standard error of the mean. SUV39H1 was not identified in the full proteome of naive mESCs which is why the relative chromatin binding was imputed by a fixed value: 0. (**I**) Relative chromatin bindings of proteins related to the HIPPO signaling pathway in all analyzed cell lines. Bars represent mean values and the error bar is based on the standard error of the mean. See also Supplementary Figure S12.

mouse formative and primed PSCs from mouse naive PSCs (Figure 6B).

To further dissect whether hESCs correspond more to the early or late mouse post-implantation stage, we computed correlations between the chromatomes of all four cell lines selected for *bona fide* pluripotency and early differentiation factors (Figure 6C). We noted an incremental increase in the correlation of hESCs with naive, formative, and primed PSCs (Pearson, $r = 0.48$ for naive, 0.59 for formative, and 0.66 for primed PSCs) (Figure 6D–F), while chromatomes of formative and primed PSCs correlated better to each other (Pearson, $r = 0.78$) than to naive PSCs (Pearson, $r = 0.74$ and $r = 0.57$, respectively). We observed similar differences on the relative chromatin binding and total proteome levels (Supplementary Figure S12B, C).

For an in-depth view of pluripotency factors and their contribution to cell identity, we computed the chromatome difference between a given mouse PSC-line and hESCs for each *bona fide* pluripotency factor (Figure 6G and Supplementary Table 4). A step-wise loss of pre-implantation pluripotency markers was observed from naive to primed PSCs with some remarkable exceptions; TFAP2C, DPPA2, DPPA4 and PRDM14 were more similar in their chromatin abundance between both naive and formative PSCs and hESCs. These proteins are indicative of germline competence, a capability that mouse formative PSCs and conventional hESCs harbor, while mouse naive PSCs first require differentiation to the formative state (33,82–85). Moreover, REX1, a well-characterized naive pluripotency and germline marker, was more strongly associated with the hESC chromatome than mouse formative and primed PSCs, likely reflecting the more heterogeneous nature of hESCs or species-specific differences (86). In a PCA based on these *bona fide* pluripotency factors only, mouse formative PSCs were even further separated from primed PSCs but not from naive PSCs (Supplementary Figure S12D, F). A scatter plot of the protein loading values uncovered that the main causes of this separation were naive pluripotency factors such as NR0B1, KLF2 and KLF4 (Supplementary Figure S12E). Thus, these naive factors were less associated with chromatin in hESCs and mouse primed PSCs than formative or naive PSCs. Conversely, post-implantation pluripotency factors contributed to the higher similarity between hESCs and primed PSCs. Of note, we did not observe differences in the chromatin association of the core pluripotency circuitry such as OCT4 or SALL4 (Supplementary Figure S12F, G).

The relative chromatin binding of well-known heterochromatic proteins (CBX1, CBX3, CBX5, KAP1, MBD3 and SUV39H1) revealed similar high levels in hESCs as in formative and primed PSCs (Figure 6H, see also Figure 5). An increased relative chromatin binding of heterochromatic proteins seems thus to be a common hallmark of post-implantation PSCs, indicating that higher chromatin compaction involves enhanced chromatin association of heterochromatic proteins. However, we also observed notable differences between hESCs and mouse post-implantation PSCs, like for the HIPPO signaling pathway (Figure 6I). This pathway is highly active in pluripotent epiblast cells and upon its activation the downstream proteins YAP1 and TAZ are kept cytoplasmic (56,87). Interestingly, we observed YAP1 and TAZ only in the full proteome fractions, except for hESCs where YAP1 was also present in the chromatin fraction. This was in agreement with a higher relative chromatin binding of the YAP1 cofactors TEAD1/3/4 in hESCs, likely suggesting a more inactive state of the HIPPO pathway in hESCs than in closely related mouse pluripotency phases.

In summary, the conventional hESC chromatome is similar to mouse PSC chromatomes reflecting post-implantation, particularly the mouse primed stage. This is largely due to lower levels of naive-specific transcription factors in these chromatomes. However, hESCs differ from mouse primed PSCs in the chromatin association of e.g. essential germline factors and the HIPPO pathway, indicating that hESCs have some similarities to mouse formative-like chromatomes and that the HIPPO pathway is regulated differently between mouse and human PSCs.

DISCUSSION

Previous studies have established methods for chromatin purification and measurement (29–32,88,89). These techniques, however, require large numbers of cells and have limited accuracy and comprehensiveness, often failing to detect low-abundant proteins such as regulatory factors. In this study, we combined a stringent and simple chromatin preparation strategy of crosslinked nuclei with an additional purification step by protein aggregation capture (PAC) and optimized DIA-based MS. Our method only requires three hours of experimental hands-on time and confidently reduces non-chromatin proteins while identifying more than twice the number of DNA-binding proteins compared to other methods in half of the MS acquisition time (54,90). In addition, recent deep neural network-based computational processing of DIA measurements without a peptide library (direct DIA) can now outperform DDA in accuracy and comprehensiveness (47,50,91,92). Thus, our direct DIA measurements additionally decreased instrument time, while providing a near-complete chromatome coverage. However, it is possible that a library-based analysis would increase the current chromatome depth further, and may represent a potential future opportunity.

The datasets generated here allowed us to perform several different types of analysis. Given that ChAC-DIA selectively enriched components of the chromatome, we were able to assemble a high-confidence global map of the chromatome. By comparing chromatome and proteome data, including proteomic data derived from different cellular fractions, for different pluripotency phases, we identified proteins affected by nuclear translocation or chromatin binding. For example, we observed chromatin enrichment of cytoplasmic transcription factors such as those involved in WNT and LIF pathways, and not GSK, FGF2 and Activin A pathways in naive PSCs, which has implications for their role in pluripotency regulation. Furthermore, normalizing the chromatome to protein levels enabled a global assessment of changes in relative chromatin binding which may be caused by either altered chromatin affinity and accessibility or differential subcellular localization and availability. Our method thus enables accurate and comprehensive chromatome and relative chromatin binding measurements despite limited cell numbers, making it ideally suited for analyzing minute tissue samples or rare subpopulations of cells.

Additionally, ChAC-DIA enables the quantification of low-abundant transcriptional or epigenetic regulators, and we identified several low-abundant chromatin binders that are pluripotency phase-specific. Besides well-described factors, we find many phase-specific proteins with still unknown functions in pluripotency regulation. Given their phase-specific chromatin association, many of them are likely to contribute to the regulation of cellular identity. One such example is EZHIP which was only identified in the formative phase. EZHIP was recently described to inhibit H3K27me3 by mimicking the H3K27M oncohistone and thus preventing the PRC2 complex from spreading along chromatin (93,94). Bulk levels of H3K27me3 are known to be downregulated from naive to primed pluripotency while bivalent sites harboring H3K4me3 and H3K27me3 are enriched (10,95). In our chromatome data set, we observed that EZH1 increases at chromatin between the naive and formative PSCs which does not fit a global downregulation of H3K27me3. Interestingly, this goes along with an increase in EZHIP in the formative chromatome implying a possible role of PRC2 inhibition or redirection to other regions by EZHIP in formative PSCs. Moreover, low-abundant epigenetic writers such as SUV39H1/2, SUV420H1/2, SETDB2 or TET1–TET3 featured phase-specific enrichment at chromatin. Remarkably, all three TET proteins showed a distinct redistribution along the exit from pluripotency, starting with TET1 and TET2 being most abundant in the naive state and TET3 being mostly chromatin-associated in the primed state. This was also observed in conventional hESCs where TET2 and TET1 are even less associated with chromatin than in mouse primed PSCs.

The chromatome correlates weakly with the transcriptome and proteome and is, therefore, an important complement to previous studies of pluripotency. Our results provide a system-wide view of pluripotency by offering a chromatome atlas with specifically enriched proteins for each analyzed pluripotency phase. Our observations are in line with the recent finding that formative pluripotency is an essential state which is transcriptionally and epigenetically distinct from naive pluripotency and to a smaller degree also from primed pluripotency (1,3,10,13,62,96). The underlying chromatome changes fit in with the phased progression model of pluripotency (3). Moreover, formative and primed PSCs share the majority of open chromatin sites while there is little overlap between formative and naive PSCs (1). Our data support this observation by showing that the chromatome undergoes larger changes from naive to formative, than from formative to primed pluripotency. The chromatin composition is further reorganized between formative and primed PSCs, mainly driven by transcription factors triggering early differentiation as well as histone H1 and HMG variants guiding chromatin compaction. The histone H1 chromatin enrichment is in agreement with an increased relative chromatin binding of SUMO1–3 and SUMOylating enzymes of chromatin organizing proteins. SUMOylation of histone H1 was recently described as a mechanism for heterochromatinization in ESCs (79), thus suggesting a role for SUMOylation in further chromatin compaction from formative to primed pluripotency. An increased relative chromatin binding was observed for additional heterochromatic proteins, such as KAP1 and CBX3, at the formative and primed phases. Surprisingly, this increased relative chromatin binding of heterochromatic proteins was conserved in conventional hESCs. We conclude that heterochromatic proteins not only become more abundant towards the exit from pluripotency, but also have a stronger affinity for chromatin. One potential explanation for this enhanced affinity is that the increase of repressive epigenetic marks during the transition from naive to primed pluripotency provides additional binding sites for heterochromatic proteins, thereby giving rise to a more repressive chromatome signature.

Conventionally cultured hESCs are reminiscent of mouse primed PSCs regarding their epigenome, transcriptome and underlying signaling cues (56,80). Still, human embryonic development comprises pluripotent phases that differ in length and growth conditions when compared to mouse (1,3,4,97–99). It remains unclear whether hESCs are the direct counterpart of mouse primed PSCs and to what extent they share unique features with mouse formative PSCs. A quantitative comparison of the high-confidence chromatomes revealed that mouse primed PSCs correlated best with hESCs. Of note, a comparable correlation range was previously described on transcriptome and full proteome levels (33,60). In our hands, the correlation between hESCs and mouse primed PSCs increased even further when only *bona fide* pluripotency and early differentiation factors were considered. Here, chromatome-levels of naive pluripotency factors were the main difference between mouse primed PSCs and hESCs on the one side and mouse formative and naive PSCs on the other side. One major distinction between hESCs and mouse primed PSCs was the high chromatin association of essential germline factors like DPPA2, PRDM14 and TFAP2C in hESCs which resembles formative pluripotency in the mouse. This finding may explain the differential developmental capacities of hESCs and mouse primed PSCs. In addition, the hESC chromatome provided evidence for a less active HIPPO pathway compared to all three mouse PSCs, likely reflecting more species-specific signaling mechanisms.

Our study sheds light on the important question of whether cell identity-defining transcription factors coexist, suggesting an ongoing competition with each other (100,101), or abruptly change across pluripotency phases (4). For all three phases and especially for the formative phase we observed that transcription factors were gradually recruited or evicted from chromatin. For instance, OTX2, a key transcription factor of formative pluripotency (15,102), peaks in abundance at the formative state, but is still associated with chromatin in naive and primed PSCs. Thus, our findings support the model of coexisting phase-specific transcription factors that ultimately define cellular identity if a certain critical threshold is exceeded.

In conclusion, we present a robust chromatin proteomics method to detect changes in the abundance and affinity of even low-abundant proteins. We offer a rich resource for the proteomes, chromatomes and relative chromatin bindings in mouse naive, formative and primed PSCs, as well as hESCs that are a basis for identifying and investigating novel regulatory mechanisms of pluripotency. Further investigations of candidate phase-specific proteins highlighted herein may help detangle the connection between pluripotency and lineage priming and support clinical applications of iPSCs. The dramatically improved sensitivity now makes it possible to also study rare subpopulations of cells. The comprehensive capture of chromatomes and chromatin affinities provides a deep and unbiased view of regulatory events underlying the establishment, maintenance, and change of cellular identity.

DATA AVAILABILITY

The mass spectrometry proteomics data has been deposited to the ProteomeXchange Consortium via the PRIDE (103) partner repository with the dataset identifiers PXD034448 for chromatomes and proteomes and PXD039556 for ChIP-MS. To make the proteome and chromatome files better comprehensible, they have been assigned to experiments (Raw data list, see PXD034448). Source data are provided in this paper.

The used RNA-Seq dataset is derived from the ArrayExpress with the following accession code: E-MTAB-6797.

CODE AVAILABILITY

The underlying custom code for the provided web application, accessible on https://pluripotency.shinyapps.io/Chromatome_Atlas/, can be found at the Github repository https://github.com/ugur-enes/pluripotency_chromatome_shinyapp.git.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank Dr Igor Paron and Tim Heymann for excellent MS technical assistance, Dr Johannes Bruno Müller-Reif for columns, and Jeannette Koch and Ruzica Barisic for technical assistance. We thank Dr Christopher B. Mulholland, Dr Michael D. Bartoschek as well as Dr Carsten Marr for valuable and insightful scientific discussions. We thank Dr Vadim Demichev for generously providing access to the most recent version of DIA-NN and for answering questions on setting up DIA-NN. We thank Dr Ludovic Vallier and Dr Derk ten Berge for kindly providing mEpiSCs. E.U. gratefully acknowledges the International Max Planck Research School for Molecular Life Sciences (IMPRS-LS) and the research training group 1721 (RTG 1721) for training and support.

Author contributions: E.U. and M.W. designed the study; H.L., M.W. and M.M. supervised the study. E.U. performed all experiments and MS data analysis, and E.U., H.L. and M.W. interpreted the data. E.U., H.L., M.W. and M.M. wrote the manuscript and prepared the figures. A.d.l.P. established the same culture conditions for human and mouse ESCs and conducted the cell culture under the supervision of M.D. W.Q. performed western blots and A.I. performed KAT7-ChIP-MS. E.U. and S.B. programmed the web application. All authors discussed the results, then read and approved the manuscript.

FUNDING

Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) [213249687 – SFB1064 to H.L. (A17) and S.B. (A22)]. Funding for open access charge: Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) [213249687 – SFB1064 to H.L. (A17)].

Conflict of interest statement. None declared.

REFERENCES

1. Kinoshita,M., Barber,M., Mansfield,W., Cui,Y., Spindlow,D., Stirparo,G.G., Dietmann,S., Nichols,J. and Smith,A. (2021) Capture of mouse and human stem cells with features of formative pluripotency. *Cell Stem Cell*, **28**, 453–471.
2. Takahashi,S., Kobayashi,S. and Hiratani,I. (2018) Epigenetic differences between naïve and primed pluripotent stem cells. *Cell. Mol. Life Sci.*, **75**, 1191–1203.
3. Smith,A. (2017) Formative pluripotency: the executive phase in a developmental continuum. *Development*, **144**, 365–373.
4. Boroviak,T., Loos,R., Lombard,P., Okahara,J., Behr,R., Sasaki,E., Nichols,J., Smith,A. and Bertone,P. (2015) Lineage-specific profiling delineates the emergence and progression of naive pluripotency in mammalian embryogenesis. *Dev. Cell*, **35**, 366–382.
5. Meshorer,E., Yellajoshula,D., George,E., Scambler,P.J., Brown,D.T. and Misteli,T. (2006) Hyperdynamic plasticity of chromatin proteins in pluripotent embryonic stem cells. *Dev. Cell*, **10**, 105–116.
6. Neagu,A., van Genderen,E., Escudero,I., Verwegen,L., Kurek,D., Lehmann,J., Stel,J., Dirks,R.A.M., van Mierlo,G., Maas,A. *et al.* (2020) In vitro capture and characterization of embryonic rosette-stage pluripotency between naive and primed states. *Nat. Cell Biol.*, **22**, 534–545.
7. Morgani,S., Nichols,J. and Hadjantonakis,A.-K. (2017) The many faces of pluripotency: in vitro adaptations of a continuum of in vivo states. *BMC Dev. Biol.*, **17**, 7.
8. Melcer,S., Hezroni,H., Rand,E., Nissim-Rafinia,M., Skoultchi,A., Stewart,C.L., Bustin,M. and Meshorer,E. (2012) Histone modifications and lamin A regulate chromatin protein dynamics in early embryonic stem cell differentiation. *Nat. Commun.*, **3**, 910.
9. Zylicz,J.J., Dietmann,S., Günesdogan,U., Hackett,J.A., Cougot,D., Lee,C. and Surani,M.A. (2015) Chromatin dynamics and the role of G9a in gene regulation and enhancer silencing during early mouse development. *Elife*, **4**, e09571.
10. Wang,X., Xiang,Y., Yu,Y., Wang,R., Zhang,Y., Xu,Q., Sun,H., Zhao,Z.-A., Jiang,X., Wang,X. *et al.* (2021) Formative pluripotent stem cells show features of epiblast cells poised for gastrulation. *Cell Res.*, **31**, 526–541.
11. Hackett,J.A. and Surani,M.A. (2014) Regulatory principles of pluripotency: from the ground state up. *Cell Stem Cell*, **15**, 416–430.
12. Nichols,J. and Smith,A. (2009) Naive and primed pluripotent states. *Cell Stem Cell*, **4**, 487–492.
13. Kalkan,T., Olova,N., Roode,M., Mulas,C., Lee,H.J., Nett,I., Marks,H., Walker,R., Stunnenberg,H.G., Lilley,K.S. *et al.* (2017) Tracking the embryonic stem cell transition from ground state pluripotency. *Development*, **144**, 1221–1234.
14. Xiang,Y., Zhang,Y., Xu,Q., Zhou,C., Liu,B., Du,Z., Zhang,K., Zhang,B., Wang,X., Gayen,S. *et al.* (2020) Epigenomic analysis of gastrulation identifies a unique chromatin state for primed pluripotency. *Nat. Genet.*, **52**, 95–105.
15. Buecker,C., Srinivasan,R., Wu,Z., Calo,E., Acampora,D., Faial,T., Simeone,A., Tan,M., Swigut,T. and Wysocka,J. (2014) Reorganization of enhancer patterns in transition from naive to primed pluripotency. *Cell Stem Cell*, **14**, 838–853.
16. Chovanec,P., Collier,A.J., Krueger,C., Várnai,C., Semprich,C.I., Schoenfelder,S., Corcoran,A.E. and Rugg-Gunn,P.J. (2021) Widespread reorganisation of pluripotent factor binding and gene regulatory interactions between human pluripotent states. *Nat. Commun.*, **12**, 2098.

17. Imhof,A. and Bonaldi,T. (2005) ‘Chromatomics’ the analysis of the chromatome. *Mol. Biosyst.*, **1**, 112–116.
18. Nichols,J., Zevnik,B., Anastassiadis,K., Niwa,H., Klewe-Nebenius,D., Chambers,I., Schöler,H. and Smith,A. (1998) Formation of pluripotent stem cells in the mammalian embryo depends on the POU transcription factor Oct4. *Cell*, **95**, 379–391.
19. Loh,Y.-H., Wu,Q., Chew,J.-L., Vega,V.B., Zhang,W., Chen,X., Bourque,G., George,J., Leong,B., Liu,J. *et al.* (2006) The Oct4 and Nanog transcription network regulates pluripotency in mouse embryonic stem cells. *Nat. Genet.*, **38**, 431–440.
20. Mitsui,K., Tokuzawa,Y., Itoh,H., Segawa,K., Murakami,M., Takahashi,K., Maruyama,M., Maeda,M. and Yamanaka,S. (2003) The homeoprotein Nanog is required for maintenance of pluripotency in mouse epiblast and ES cells. *Cell*, **113**, 631–642.
21. Masui,S., Nakatake,Y., Toyooka,Y., Shimosato,D., Yagi,R., Takahashi,K., Okochi,H., Okuda,A., Matoba,R., Sharov,A.A. *et al.* (2007) Pluripotency governed by Sox2 via regulation of Oct3/4 expression in mouse embryonic stem cells. *Nat. Cell Biol.*, **9**, 625–635.
22. Ji,X., Dadon,D.B., Abraham,B.J., Lee,T.I., Jaenisch,R., Bradner,J.E. and Young,R.A. (2015) Chromatin proteomic profiling reveals novel proteins associated with histone-marked genomic regions. *Proc. Natl. Acad. Sci. U.S.A.*, **112**, 3841–3846.
23. Alajem,A., Biran,A., Harikumar,A., Sailaja,B.S., Aaronson,Y., Livyatan,I., Nissim-Rafinia,M., Sommer,A.G., Mostoslavsky,G., Gerbasi,V.R. *et al.* (2015) Differential Association of Chromatin Proteins identifies BAF60a/SMARCD1 as a regulator of embryonic stem cell differentiation. *Cell Rep.*, **10**, 2019–2031.
24. Rafiee,M.-R., Girardot,C., Sigismondo,G. and Krijgsveld,J. (2016) Expanding the circuitry of pluripotency by selective isolation of chromatin-associated proteins. *Mol. Cell*, **64**, 624–635.
25. Villaseñor,R., Pfaendler,R., Ambrosi,C., Butz,S., Giuliani,S., Bryan,E., Sheahan,T.W., Gable,A.L., Schmolka,N., Manzo,M. *et al.* (2020) ChromID identifies the protein interactome at chromatin marks. *Nat. Biotechnol.*, **38**, 728–736.
26. Kloet,S.L., Makowski,M.M., Baymaz,H.I., van Voorthuijsen,L., Karemaker,I.D., Santanach,A., Jansen,P.W.T.C., Di Croce,L. and Vermeulen,M. (2016) The dynamic interactome and genomic targets of Polycomb complexes during stem-cell differentiation. *Nat. Struct. Mol. Biol.*, **23**, 682–690.
27. Shiio,Y., Eisenman,R.N., Yi,E.C., Donohoe,S., Goodlett,D.R. and Aebersold,R. (2003) Quantitative proteomic analysis of chromatin-associated factors. *J. Am. Soc. Mass Spectrom.*, **14**, 696–703.
28. Räschle,M., Smeenk,G., Hansen,R.K., Temu,T., Oka,Y., Hein,M.Y., Nagaraj,N., Long,D.T., Walter,J.C., Hofmann,K. *et al.* (2015) DNA repair. Proteomics reveals dynamic assembly of repair complexes during bypass of DNA cross-links. *Science*, **348**, 1253671.
29. Kustatscher,G., Hégarat,N., Wills,K.L.H., Furlan,C., Bukowski-Wills,J.-C., Hochegger,H. and Rappsilber,J. (2014) Proteomics of a fuzzy organelle: interphase chromatin. *EMBO J.*, **33**, 648–664.
30. Kustatscher,G., Wills,K.L.H., Furlan,C. and Rappsilber,J. (2014) Chromatin enrichment for proteomics. *Nat. Protoc.*, **9**, 2090–2099.
31. Ginno,P.A., Burger,L., Seebacher,J., Iesmantavicius,V. and Schübeler,D. (2018) Cell cycle-resolved chromatin proteomics reveals the extent of mitotic preservation of the genomic regulatory landscape. *Nat. Commun.*, **9**, 4048.
32. Aranda,S., Alcaine-Colet,A., Blanco,E., Borràs,E., Caillot,C., Sabidó,E. and Di Croce,L. (2019) Chromatin capture links the metabolic enzyme AHCY to stem cell proliferation. *Sci. Adv.*, **5**, eaav2448.
33. Hayashi,K., Ohta,H., Kurimoto,K., Aramaki,S. and Saitou,M. (2011) Reconstitution of the mouse germ cell specification pathway in culture by pluripotent stem cells. *Cell*, **146**, 519–532.
34. Brons,I.G.M., Smithers,L.E., Trotter,M.W.B., Rugg-Gunn,P., Sun,B., Chuva de Sousa Lopes,S.M., Howlett,S.K., Clarkson,A., Ahrlund-Richter,L., Pedersen,R.A. *et al.* (2007) Derivation of pluripotent epiblast stem cells from mammalian embryos. *Nature*, **448**, 191–195.
35. Stauske,M., Rodriguez Polo,I., Haas,W., Knorr,D.Y., Borchert,T., Streckfuss-Bömeke,K., Dressel,R., Bartels,I., Tiburcy,M., Zimmermann,W.-H. *et al.* (2020) Non-Human primate iPSC generation, cultivation, and cardiac differentiation under chemically defined conditions. *Cells*, **9**, 1349.
36. Rappsilber,J., Mann,M. and Ishihama,Y. (2007) Protocol for micro-purification, enrichment, pre-fractionation and storage of peptides for proteomics using StageTips. *Nat. Protoc.*, **2**, 1896–1906.
37. Batth,T.S., Tollenaere,M.A.X., Rüther,P., Gonzalez-Franquesa,A., Prabhakar,B.S., Bekker-Jensen,S., Deshmukh,A.S. and Olsen,J.V. (2019) Protein aggregation capture on microparticles enables multipurpose proteomics sample preparation. *Mol. Cell. Proteomics*, **18**, 1027–1035.
38. Mulholland,C.B., Nishiyama,A., Ryan,J., Nakamura,R., Yiğit,M., Glück,I.M., Trummer,C., Qin,W., Bartoschek,M.D., Traube,F.R. *et al.* (2020) Recent evolution of a TET-controlled and DPPA3/STELLA-driven pathway of passive DNA demethylation in mammals. *Nat. Commun.*, **11**, 5972.
39. Stolz,P., Mantero,A.S., Tvardovskiy,A., Ugur,E., Wange,L.E., Mulholland,C.B., Cheng,Y., Wierer,M., Enard,W., Schneider,R. *et al.* (2022) TET1 regulates gene expression and repression of endogenous retroviruses independent of DNA demethylation. *Nucleic Acids Res.*, **50**, 8491–8511.
40. Qin,W., Ugur,E., Mulholland,C.B., Bultmann,S., Solovei,I., Modic,M., Smets,M., Wierer,M., Forné,I., Imhof,A. *et al.* (2021) Phosphorylation of the HP1β hinge region sequesters KAP1 in heterochromatin and promotes the exit from naïve pluripotency. *Nucleic Acids Res.*, **49**, 7406–7423.
41. Bruderer,R., Bernhardt,O.M., Gandhi,T., Miladinović,S.M., Cheng,L.-Y., Messner,S., Ehrenberger,T., Zanotelli,V., Butscheid,Y., Escher,C. *et al.* (2015) Extending the limits of quantitative proteome profiling with data-independent acquisition and application to acetaminophen-treated three-dimensional liver microtissues. *Mol. Cell. Proteomics*, **14**, 1400–1410.
42. Cox,J. and Mann,M. (2008) MaxQuant enables high peptide identification rates, individualized ppb-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.*, **26**, 1367.
43. Cox,J., Hein,M.Y., Luber,C.A., Paron,I., Nagaraj,N. and Mann,M. (2014) Accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction, termed MaxLFQ. *Mol. Cell. Proteomics*, **13**, 2513–2526.
44. Cox,J., Neuhauser,N., Michalski,A., Scheltema,R.A., Olsen,J.V. and Mann,M. (2011) Andromeda: a peptide search engine integrated into the MaxQuant environment. *J. Proteome Res.*, **10**, 1794–1805.
45. Tyanova,S., Temu,T., Sinitcyn,P., Carlson,A., Hein,M.Y., Geiger,T., Mann,M. and Cox,J. (2016) The Perseus computational platform for comprehensive analysis of (prote) omics data. *Nat. Methods*, **13**, 731.
46. Snel,B., Lehmann,G., Bork,P. and Huynen,M.A. (2000) STRING: a web-server to retrieve and display the repeatedly occurring neighbourhood of a gene. *Nucleic Acids Res.*, **28**, 3442–3444.
47. Demichev,V., Messner,C.B., Vernardis,S.I., Lilley,K.S. and Ralser,M. (2020) DIA-NN: neural networks and interference correction enable deep proteome coverage in high throughput. *Nat. Methods*, **17**, 41–44.
48. Venable,J.D., Dong,M.-Q., Wohlschlegel,J., Dillin,A. and Yates,J.R. (2004) Automated approach for quantitative analysis of complex peptide mixtures from tandem mass spectra. *Nat. Methods*, **1**, 39–45.
49. Gillet,L.C., Navarro,P., Tate,S., Röst,H., Selevsek,N., Reiter,L., Bonner,R. and Aebersold,R. (2012) Targeted data extraction of the MS/MS spectra generated by data-independent acquisition: a new concept for consistent and accurate proteome analysis. *Mol. Cell. Proteomics*, **11**, O111.016717.
50. Bruderer,R., Bernhardt,O.M., Gandhi,T., Xuan,Y., Sondermann,J., Schmidt,M., Gomez-Varela,D. and Reiter,L. (2017) Optimization of experimental parameters in data-independent mass spectrometry significantly increases depth and reproducibility of results. *Mol. Cell. Proteomics*, **16**, 2296–2309.
51. Ludwig,C., Gillet,L., Rosenberger,G., Amon,S., Collins,B.C. and Aebersold,R. (2018) Data-independent acquisition-based SWATH-MS for quantitative proteomics: a tutorial. *Mol. Syst. Biol.*, **14**, e8126.
52. Hansen,F.M., Tanzer,M.C., Brüning,F., Bludau,I., Stafford,C., Schulman,B.A., Robles,M.S., Karayel,O. and Mann,M. (2021) Data-independent acquisition method for ubiquitinome analysis reveals regulation of circadian biology. *Nat. Commun.*, **12**, 254.
53. Steger,M., Demichev,V., Backman,M., Ohmayer,U., Ihmor,P., Müller,S., Ralser,M. and Daub,H. (2021) Time-resolved in vivo

ubiquitinome profiling by DIA-MS reveals USP7 targets on a proteome-wide scale. *Nat. Commun.*, **12**, 5399.

54. van Mierlo,G., Wester,R.A. and Marks,H. (2019) A mass spectrometry survey of chromatin-associated proteins in pluripotency and early lineage commitment. *Proteomics*, **19**, e1900047.
55. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Michael Cherry,J., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T. *et al.* (2000) Gene ontology: tool for the unification of biology. *Nat. Genet.*, **25**, 25–29.
56. Weinberger,L., Ayyash,M., Novershtern,N. and Hanna,J.H. (2016) Dynamic stem cell states: naive to primed pluripotency in rodents and humans. *Nat. Rev. Mol. Cell Biol.*, **17**, 155–169.
57. Fuerer,C., Nusse,R. and Ten Berge,D. (2008) Wnt signalling in development and disease. Max Delbrück Center for Molecular Medicine meeting on Wnt signaling in development and disease. *EMBO Rep.*, **9**, 134–138.
58. ten Berge,D., Kurek,D., Blauwkamp,T., Koole,W., Maas,A., Eroglu,E., Siu,R.K. and Nusse,R. (2011) Embryonic stem cells require Wnt proteins to prevent differentiation to epiblast stem cells. *Nat. Cell Biol.*, **13**, 1070–1075.
59. Vallier,L., Alexander,M. and Pedersen,R.A. (2005) Activin/nodal and FGF pathways cooperate to maintain pluripotency of human embryonic stem cells. *J. Cell Sci.*, **118**, 4495–4509.
60. Yang,P., Humphrey,S.J., Cinghu,S., Pathania,R., Oldfield,A.J., Kumar,D., Perera,D., Yang,J.Y.H., James,D.E., Mann,M. *et al.* (2019) Multi-omic profiling reveals dynamics of the phased progression of pluripotency. *Cell Syst.*, **8**, 427–445.
61. van Mierlo,G., Dirks,R.A.M., De Clerck,L., Brinkman,A.B., Huth,M., Kloet,S.L., Saksouk,N., Kroeze,L.I., Willems,S., Farlik,M. *et al.* (2019) Integrative proteomic profiling reveals PRC2-dependent epigenetic crosstalk maintains ground-state pluripotency. *Cell Stem Cell*, **24**, 123–137.
62. Kalkan,T., Bornelöv,S., Mulas,C., Diamanti,E., Lohoff,T., Ralser,M., Middelkamp,S., Lombard,P., Nichols,J. and Smith,A. (2019) Complementary activity of ETV5, RBPJ, and TCF3 drives formative transition from naive pluripotency. *Cell Stem Cell*, **24**, 785–801.
63. Gretarsson,K.H. and Hackett,J.A. (2020) Dppa2 and Dppa4 counteract de novo methylation to establish a permissive epigenome for development. *Nat. Struct. Mol. Biol.*, **27**, 706–716.
64. Hackett,J.A., Huang,Y., Günesdogan,U., Gretarsson,K.A., Kobayashi,T. and Surani,M.A. (2018) Tracing the transitions from pluripotency to germ cell fate with CRISPR screening. *Nat. Commun.*, **9**, 4292.
65. Pastor,W.A., Chen,D., Liu,W., Kim,R., Sahakyan,A., Lukianchikov,A., Plath,K., Jacobsen,S.E. and Clark,A.T. (2016) Naive human pluripotent cells feature a methylation landscape devoid of blastocyst or germline memory. *Cell Stem Cell*, **18**, 323–329.
66. Tesar,P.J., Chenoweth,J.G., Brook,F.A., Davies,T.J., Evans,E.P., Mack,D.L., Gardner,R.L. and McKay,R.D.G. (2007) New cell lines from mouse epiblast share defining features with human embryonic stem cells. *Nature*, **448**, 196–199.
67. Takashima,Y., Guo,G., Loos,R., Nichols,J., Ficz,G., Krueger,F., Oxley,D., Santos,F., Clarke,J., Mansfield,W. *et al.* (2014) Resetting transcription factor control circuitry toward ground-state pluripotency in human. *Cell*, **158**, 1254–1269.
68. Ecco,G., Imbeault,M. and Trono,D. (2017) KRAB zinc finger proteins. *Development*, **144**, 2719–2729.
69. Macfarlan,T.S., Gifford,W.D., Driscoll,S., Lettieri,K., Rowe,H.M., Bonanomi,D., Firth,A., Singer,O., Trono,D. and Pfaff,S.L. (2012) Embryonic stem cell potency fluctuates with endogenous retrovirus activity. *Nature*, **487**, 57–63.
70. Dixon,G., Pan,H., Yang,D., Rosen,B.P., Jashari,T., Verma,N., Pulecio,J., Caspi,I., Lee,K., Stransky,S. *et al.* (2021) QSER1 protects DNA methylation valleys from de novo methylation. *Science*, **372**, eabd0875.
71. Santos-Barriopedro,I., van Mierlo,G. and Vermeulen,M. (2021) Off-the-shelf proximity biotinylation for interaction proteomics. *Nat. Commun.*, **12**, 5015.
72. Miotto,B. and Struhl,K. (2010) HBO1 histone acetylase activity is essential for DNA replication licensing and inhibited by Geminin. *Mol. Cell*, **37**, 57–66.
73. Wong,P.G., Glozak,M.A., Cao,T.V., Vaziri,C., Seto,E. and Alexandrow,M. (2010) Chromatin unfolding by Cdt1 regulates MCM loading via opposing functions of HBO1 and HDAC11-geminin. *Cell Cycle*, **9**, 4351–4363.
74. Lalonde,M.-E., Avvakumov,N., Glass,K.C., Joncas,F.-H., Saksouk,N., Holliday,M., Paquet,E., Yan,K., Tong,Q., Klein,B.J. *et al.* (2013) Exchange of associated factors directs a switch in HBO1 acetyltransferase histone tail specificity. *Genes Dev.*, **27**, 2009–2024.
75. Mulholland,C.B., Traube,F.R., Ugur,E., Parsa,E., Eckl,E.M., Schönung,M., Modic,M., Bartoschek,M.D., Stolz,P., Ryan,J. *et al.* (2020) Distinct and stage-specific contributions of TET1 and TET2 to stepwise cytosine oxidation in the transition from naive to primed pluripotency. *Sci. Rep.*, **10**, 12066.
76. Kempf,J.M., Weser,S., Bartoschek,M.D., Metzeler,K.H., Vick,B., Herold,T., Völse,K., Mattes,R., Scholz,M., Wange,L.E. *et al.* (2021) Loss-of-function mutations in the histone methyltransferase EZH2 promote chemotherapy resistance in AML. *Sci. Rep.*, **11**, 5838.
77. Griffin,T.J., Gygi,S.P., Ideker,T., Rist,B., Eng,J., Hood,L. and Aebersold,R. (2002) Complementary profiling of gene expression at the transcriptome and proteome levels in saccharomyces cerevisiae* S. *Mol. Cell. Proteomics*, **1**, 323–333.
78. Theurillat,I., Hendriks,I.A., Cossec,J.-C., Andrieux,A., Nielsen,M.L. and Dejean,A. (2020) Extensive SUMO modification of repressive chromatin factors distinguishes pluripotent from somatic cells. *Cell Rep.*, **33**, 108251.
79. Sheban,D., Shani,T., Maor,R., Aguilera-Castrejon,A., Mor,N., Oldak,B., Shmueli,M.D., Eisenberg-Lerner,A., Bayerl,J., Hebert,J. *et al.* (2022) SUMOylation of linker histone H1 drives chromatin condensation and restriction of embryonic cell fate identity. *Mol. Cell*, **82**, 106–122.
80. Davidson,K.C., Mason,E.A. and Pera,M.F. (2015) The pluripotent state in mouse and human. *Development*, **142**, 3090–3099.
81. Gafni,O., Weinberger,L., Mansour,A.A., Manor,Y.S., Chomsky,E., Ben-Yosef,D., Kalma,Y., Viukov,S., Maza,I., Zviran,A. *et al.* (2013) Derivation of novel human ground state naive pluripotent stem cells. *Nature*, **504**, 282–286.
82. Grabole,N., Tischler,J., Hackett,J.A., Kim,S., Tang,F., Leitch,H.G., Magnúsdóttir,E. and Surani,M.A. (2013) Prdm14 promotes germline fate and naive pluripotency by repressing FGF signalling and DNA methylation. *EMBO Rep.*, **14**, 629–637.
83. Maldonado-Saldivia,J., van den Bergen,J., Krouskos,M., Gilchrist,M., Lee,C., Li,R., Sinclair,A.H., Surani,M.A. and Western,P.S. (2007) Dppa2 and Dppa4 are closely linked SAP motif genes restricted to pluripotent cells and the germ line. *Stem Cells*, **25**, 19–28.
84. Schemmer,J., Araúzo-Bravo,M.J., Haas,N., Schäfer,S., Weber,S.N., Becker,A., Eckert,D., Zimmer,A., Nettersheim,D. and Schorle,H. (2013) Transcription factor TFAP2C regulates major programs required for murine fetal germ cell maintenance and haploinsufficiency predisposes to teratomas in male mice. *PLoS One*, **8**, e71113.
85. Lim,J.J., Shim,M.S., Lee,J.E. and Lee,D.R. (2014) Three-step method for proliferation and differentiation of human embryonic stem cell (hESC)-derived male germ cells. *PLoS One*, **9**, e90454.
86. Graf,T. and Stadtfeld,M. (2008) Heterogeneity of embryonic and adult stem cells. *Cell Stem Cell*, **3**, 480–483.
87. Kagiwada,S., Aramaki,S., Wu,G., Shin,B., Kutejova,E., Obridge,D., Adachi,K., Wrana,J.L., Hübner,K. and Schöler,H.R. (2021) YAP establishes epiblast responsiveness to inductive signals for germ cell fate. *Development*, **148**, dev199732.
88. Torrente,M.P., Zee,B.M., Young,N.L., Baliban,R.C., LeRoy,G., Floudas,C.A., Hake,S.B. and Garcia,B.A. (2011) Proteomic interrogation of human chromatin. *PLoS One*, **6**, e24747.
89. Kulej,K., Avgousti,D.C., Sidoli,S., Herrmann,C., Della Fera,A.N., Kim,E.T., Garcia,B.A. and Weitzman,M.D. (2017) Time-resolved global and chromatin proteomics during Herpes Simplex virus type 1 (HSV-1) infection. *Mol. Cell. Proteomics*, **16**, S92–S107.
90. van Mierlo,G. and Vermeulen,M. (2021) Chromatin proteomics to study epigenetics - challenges and opportunities. *Mol. Cell. Proteomics*, **20**, 100056.
91. Pino,L.K., Just,S.C., MacCoss,M.J. and Searle,B.C. (2020) Acquiring and analyzing data independent acquisition proteomics experiments without spectrum libraries. *Mol. Cell. Proteomics*, **19**, 1088–1103.

92. Bekker-Jensen,D.B., Bernhardt,O.M., Hogrebe,A., Martinez-Val,A., Verbeke,L., Gandhi,T., Kelstrup,C.D., Reiter,L. and Olsen,J.V. (2020) Rapid and site-specific deep phosphoproteome profiling by data-independent acquisition without the need for spectral libraries. *Nat. Commun.*, **11**, 787.
93. Ragazzini,R., Pérez-Palacios,R., Baymaz,I.H., Diop,S., Ancelin,K., Zielinski,D., Michaud,A., Givelet,M., Borsos,M., Aflaki,S. *et al.* (2019) EZHIP constrains Polycomb Repressive Complex 2 activity in germ cells. *Nat. Commun.*, **10**, 3858.
94. Jain,S.U., Rashoff,A.Q., Krabbenhoft,S.D., Hoelper,D., Do,T.J., Gibson,T.J., Lundgren,S.M., Bondra,E.R., Deshmukh,S., Harutyunyan,A.S. *et al.* (2020) H3 K27M and EZHIP impede H3K27-methylation spreading by inhibiting allosterically stimulated PRC2. *Mol. Cell*, **80**, 726–735.
95. Gonzales-Cope,M., Sidoli,S., Bhanu,N.V., Won,K.-J. and Garcia,B.A. (2016) Histone H4 acetylation and the epigenetic reader Brd4 are critical regulators of pluripotency in embryonic stem cells. *Bmc Genomics (Electronic Resource)*, **17**, 95.
96. Kinoshita,M. and Smith,A. (2018) Pluripotency deconstructed. *Dev. Growth Differ.*, **60**, 44–52.
97. Rossant,J. and Tam,P.P.L. (2017) New insights into early Human development: lessons for stem cell derivation and differentiation. *Cell Stem Cell*, **20**, 18–28.
98. Nakamura,T., Okamoto,I., Sasaki,K., Yabuta,Y., Iwatani,C., Tsuchiya,H., Seita,Y., Nakamura,S., Yamamoto,T. and Saitou,M. (2016) A developmental coordinate of pluripotency among mice, monkeys and humans. *Nature*, **537**, 57–62.
99. Guo,G., Stirparo,G.G., Strawbridge,S.E., Spindlow,D., Yang,J., Clarke,J., Dattani,A., Yanagida,A., Li,M.A., Myers,S. *et al.* (2021) Human naive epiblast cells possess unrestricted lineage potential. *Cell Stem Cell*, **28**, 1040–1056.
100. Loh,K.M. and Lim,B. (2011) A precarious balance: pluripotency factors as lineage specifiers. *Cell Stem Cell*, **8**, 363–369.
101. Sarkar,A. and Hochedlinger,K. (2013) The sox family of transcription factors: versatile regulators of stem and progenitor cell fate. *Cell Stem Cell*, **12**, 15–30.
102. Yang,S.-H., Kalkan,T., Morissroe,C., Marks,H., Stunnenberg,H., Smith,A. and Sharrocks,A.D. (2014) Otx2 and Oct4 drive early enhancer activation during embryonic stem cell transition from naive pluripotency. *Cell Rep.*, **7**, 1968–1981.
103. Perez-Riverol,Y., Bai,J., Bandla,C., García-Seisdedos,D., Hewapathirana,S., Kamatchinathan,S., Kundu,D.J., Prakash,A., Frericks-Zipper,A., Eisenacher,M. *et al.* (2022) The PRIDE database resources in 2022: a hub for mass spectrometry-based proteomics evidences. *Nucleic Acids Res.*, **50**, D543–D552.

SUPPLEMENTARY FIGURES for:

Comprehensive chromatin proteomics resolves functional phases of pluripotency and identifies changes in regulatory components

Authors:
Enes Ugur, Alexandra de la Porte, Weihua Qin, Sebastian Bultmann, Alina Ivanova, Micha Drukker, Matthias Mann*, Michael Wierer*, Heinrich Leonhardt*

Correspondence:
mmann@biochem.mpg.de, michael.wierer@sund.ku.dk and h.leonhardt@lmu.de

This PDF includes:

Figure S1 (Related to Figure 1)
Figure S2 (Related to Figure 1)
Figure S3 (Related to Figure 2)
Figure S4 (Related to Figure 3)
Figure S5 (Related to Figure 3)
Figure S6 (Related to Figure 3)
Figure S7 (Related to Figure 3)
Figure S8 (Related to Figure 3)
Figure S9 (Related to Figure 3)
Figure S10 (Related to Figure 3)
Figure S11 (Related to Figure 4)
Figure S12 (Related to Figure 6)

Figure S1 (Related to Figure 1)

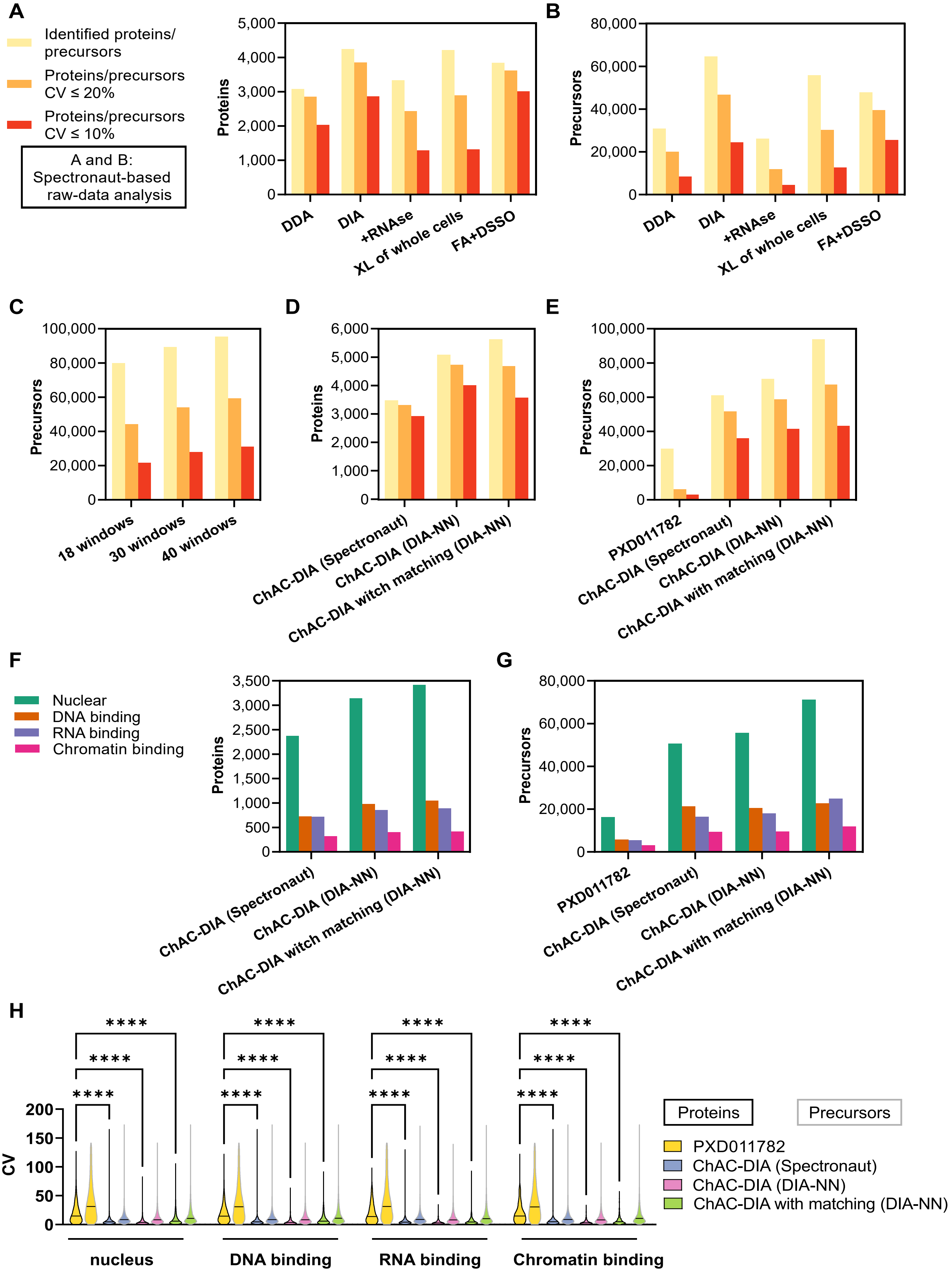


Figure S2 (Related to Figure 1)

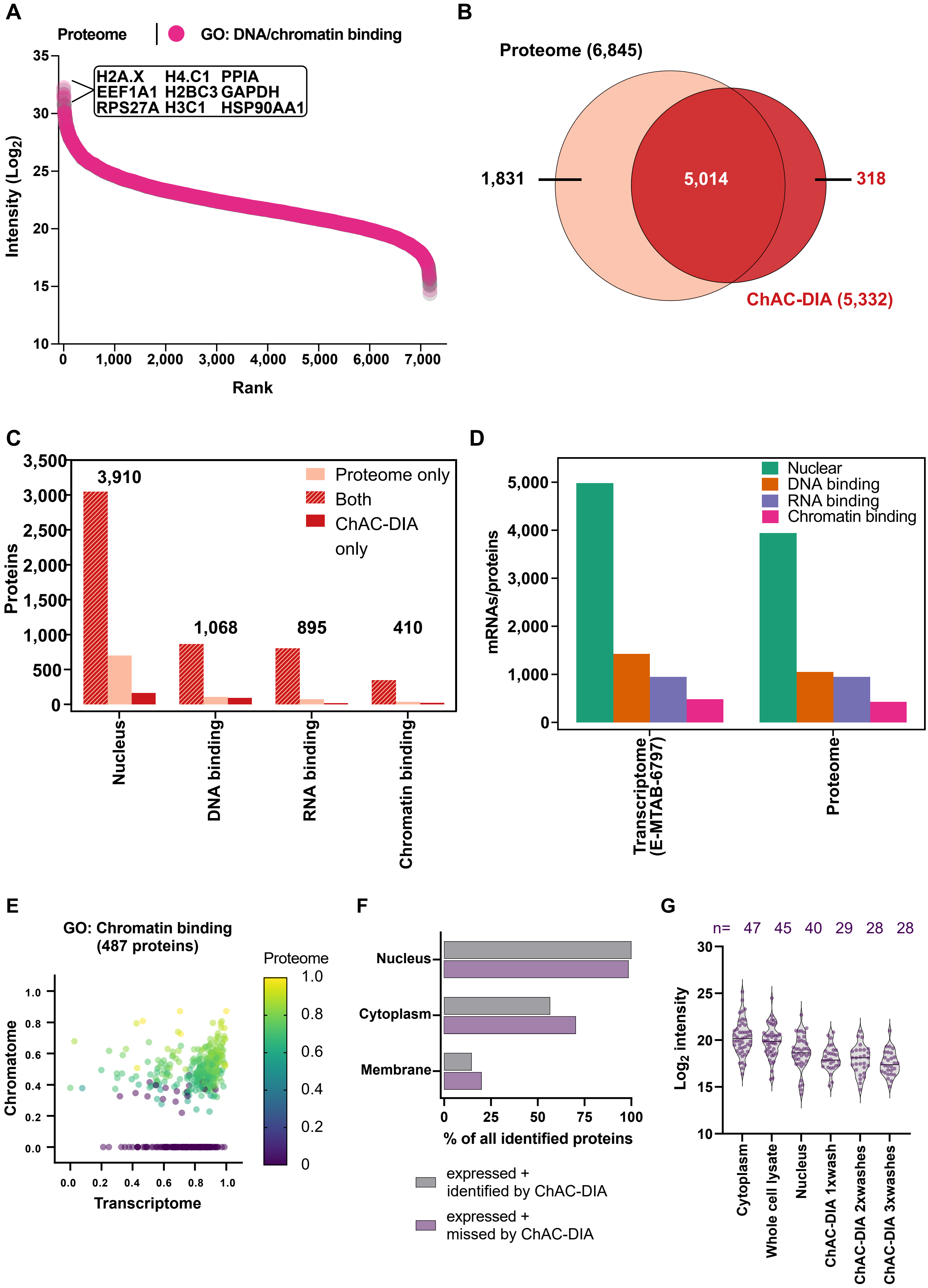


Figure S3 (Related to Figure 2)

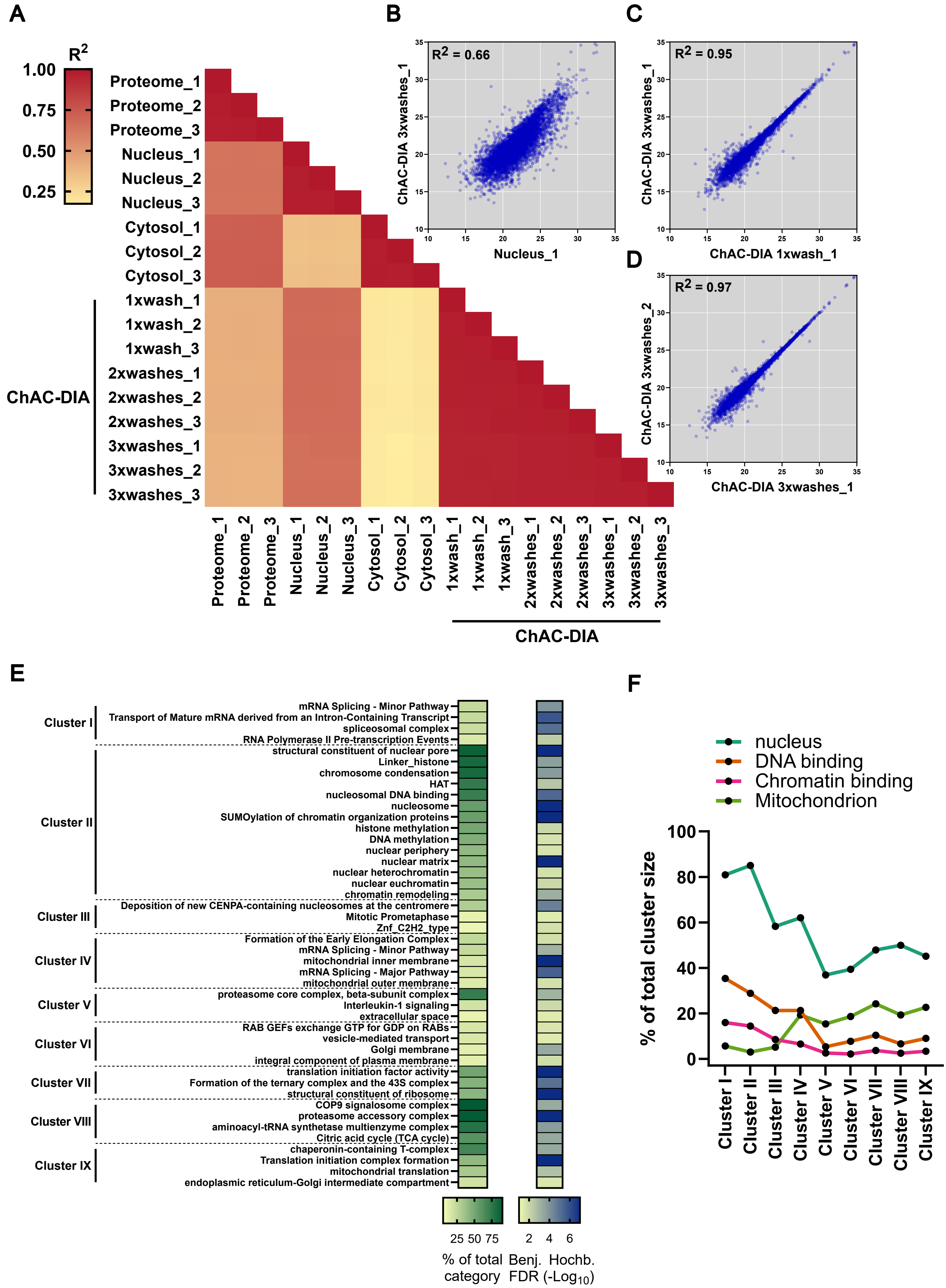


Figure S4 (Related to Figure 3)

Pluripotency/differentiation marker proteins

A) ANOVA significant (FDR < 0.05, fold change difference ≥ 2)

B) Absent in at least one pluripotency phase

C) Not ANOVA significant between any pluripotency phase

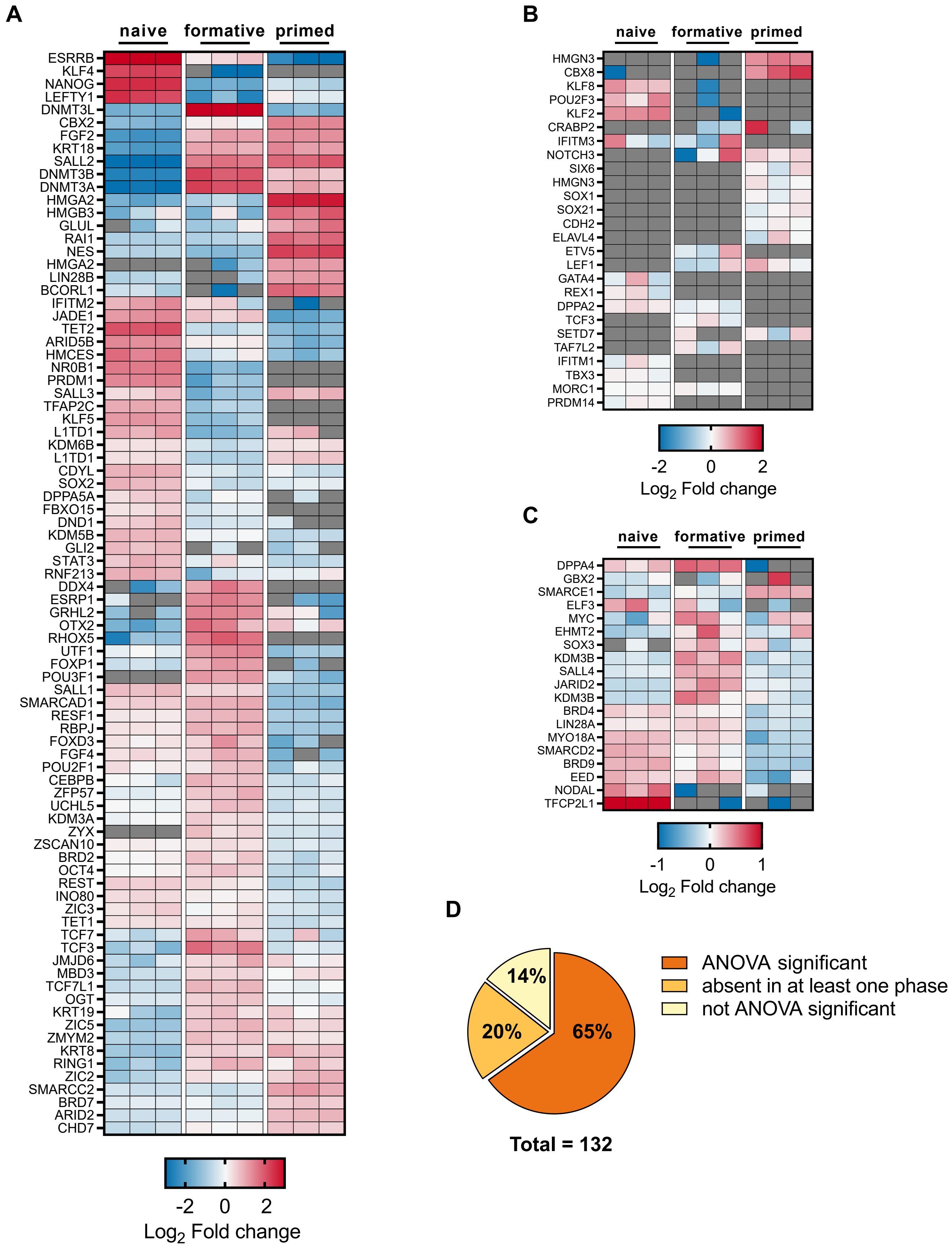


Figure S5 (Related to Figure 3)

GO "Transcription factor activity"

A) ANOVA significant (FDR < 0.05, fold change difference ≥ 2)

B) Absent in at least one pluripotency phase

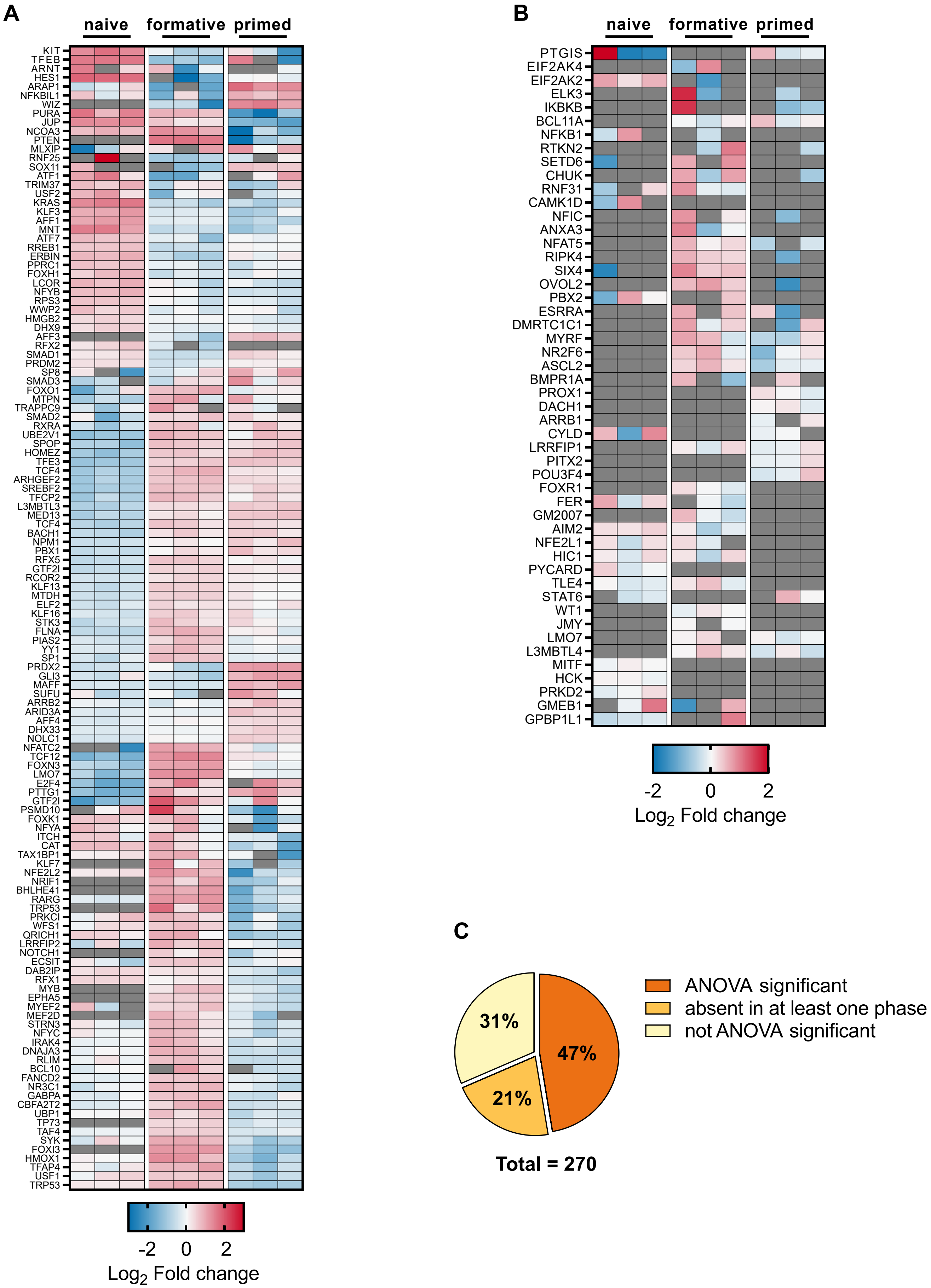


Figure S6 (Related to Figure 3)

Proteins related in epigenetic regulation

A) ANOVA significant (FDR < 0.05, fold change difference ≥ 2)

B) Absent in at least one pluripotency phase

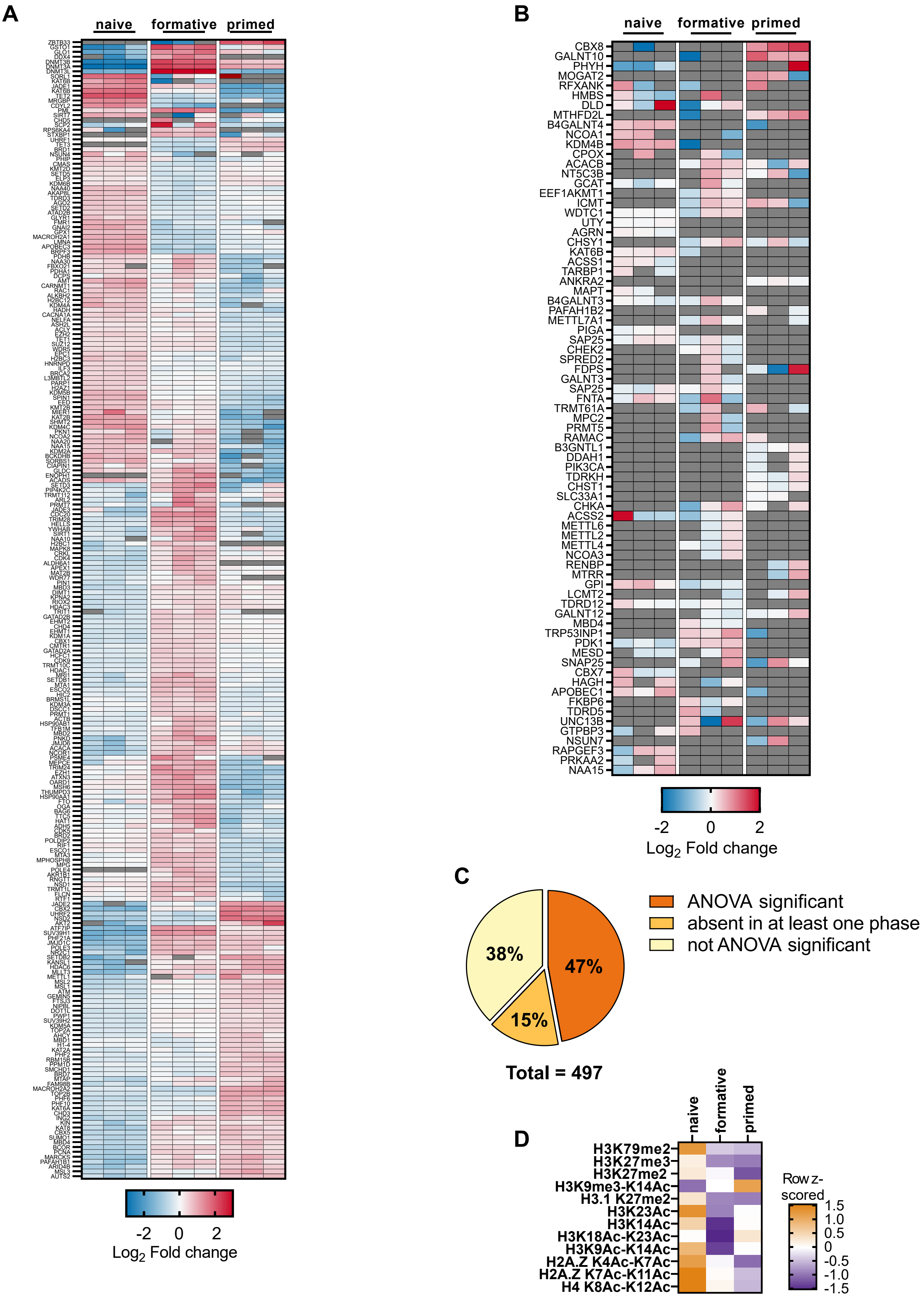


Figure S7 (Related to Figure 3)

Zinc finger (domain containing) proteins

A) ANOVA significant (FDR < 0.05, fold change difference ≥ 2)

B) Absent in at least one pluripotency phase

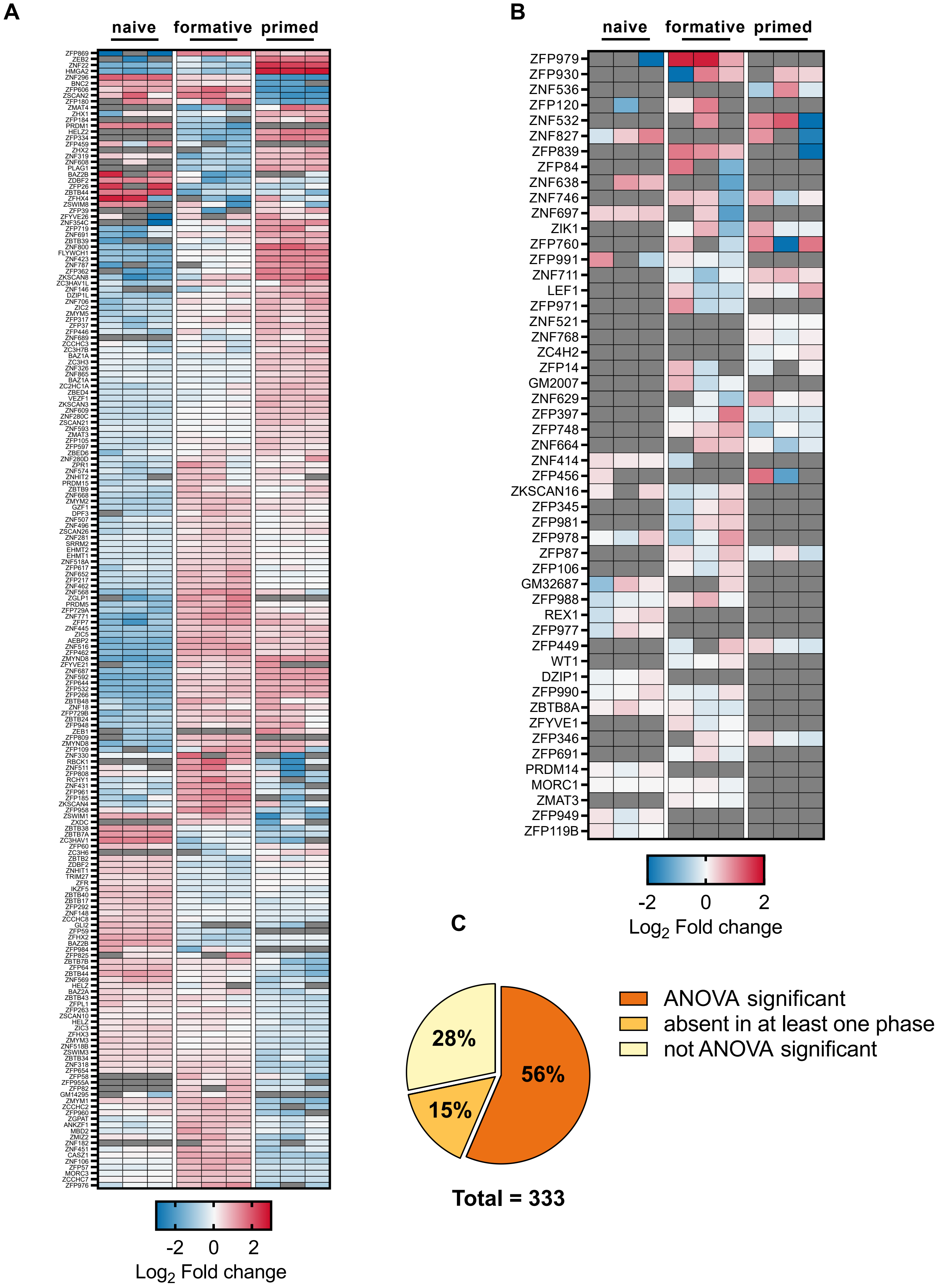


Figure S8 (Related to Figure 3)

Chromatin remodeler

A) ANOVA significant (FDR< 0.05, fold change difference ≥ 2)

B) Absent in at least one pluripotency phase

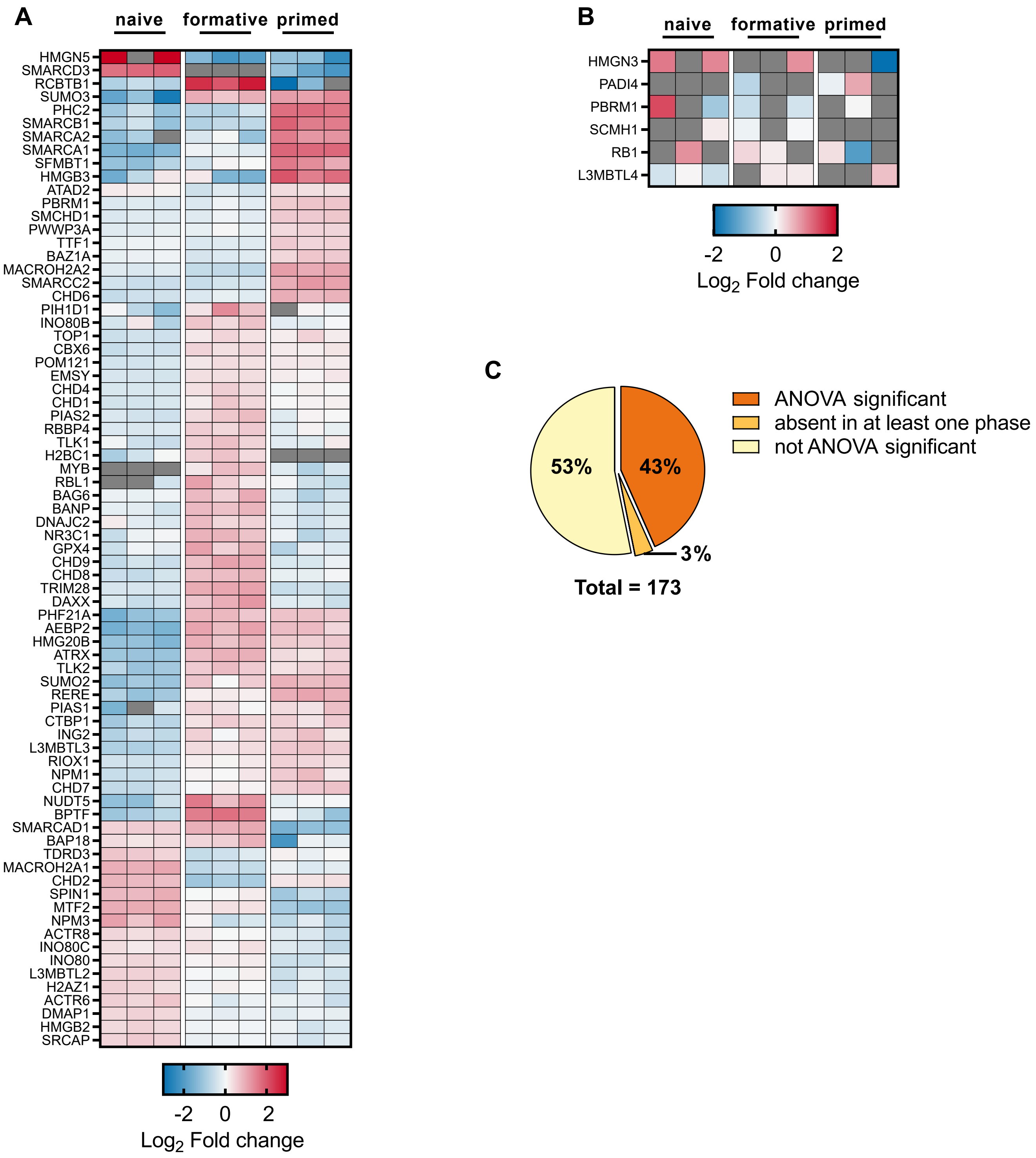


Figure S9 (Related to Figure 3)

Chromatin binding complexes

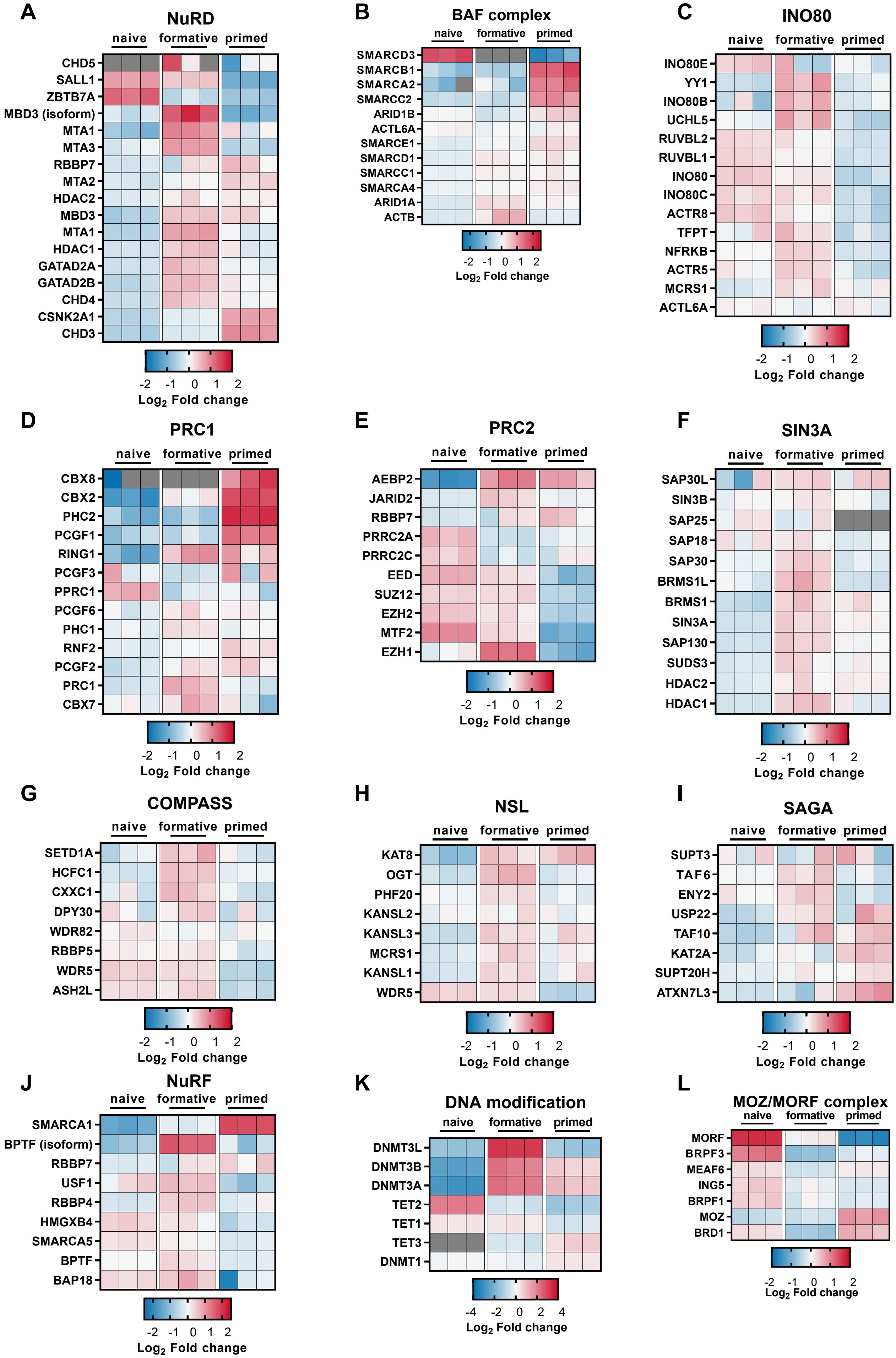
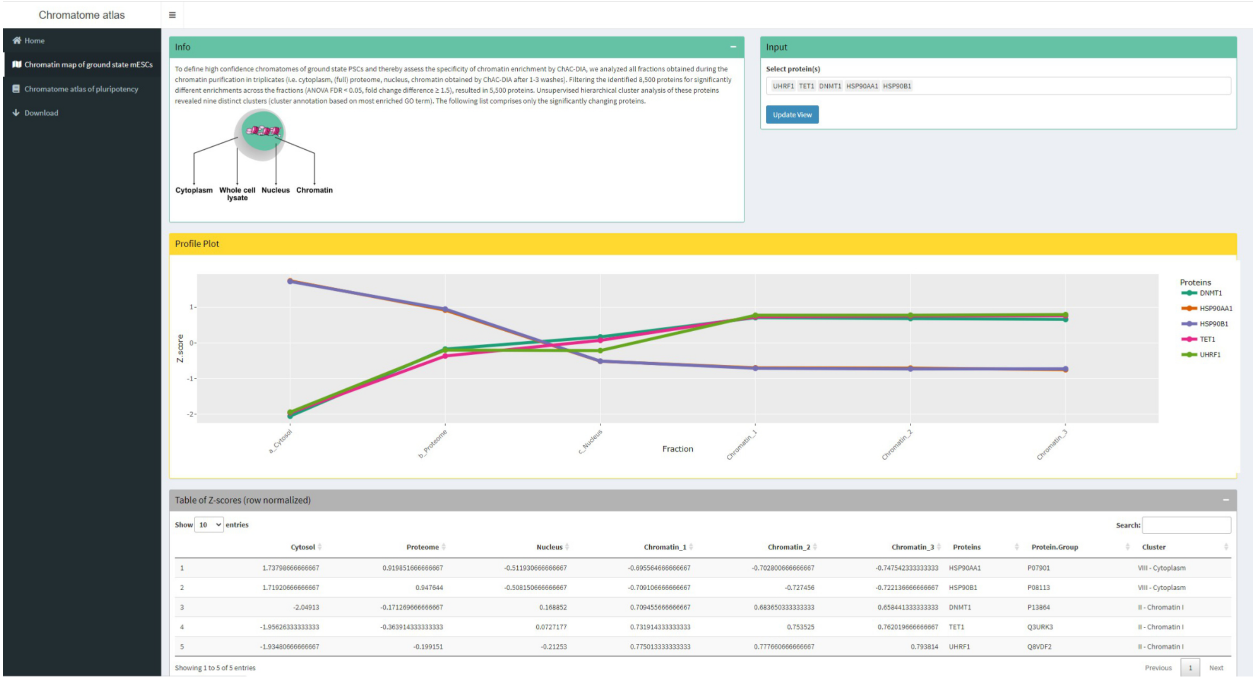


Figure S10 (Related to Figure 3)

Interactive Web Application:

https://pluripotency.shinyapps.io/Chromatome_Atlas/

A



To define high confidence chromatomes of ground state PSCs and thereby assess the specificity of chromatin enrichment by ChAC-DIA, we analyzed all fractions obtained during the chromatin purification in triplicates (i.e. cytoplasm, (full) proteome, nucleus, chromatin obtained by ChAC-DIA after 1-3 washes). Filtering the identified 8,500 proteins for significantly different enrichments across the fractions (ANOVA FDR < 0.05, fold change difference ≥ 1.5), resulted in 5,500 proteins. Unsupervised hierarchical cluster analysis of these proteins revealed nine distinct clusters (cluster annotation based on most enriched GO term). The following list comprises only the significantly changing proteins.

Select protein(s)
UHRF1 TET1 DNMT1 HSP90AA1 HSP90B1

Table of Z-scores (row normalized)

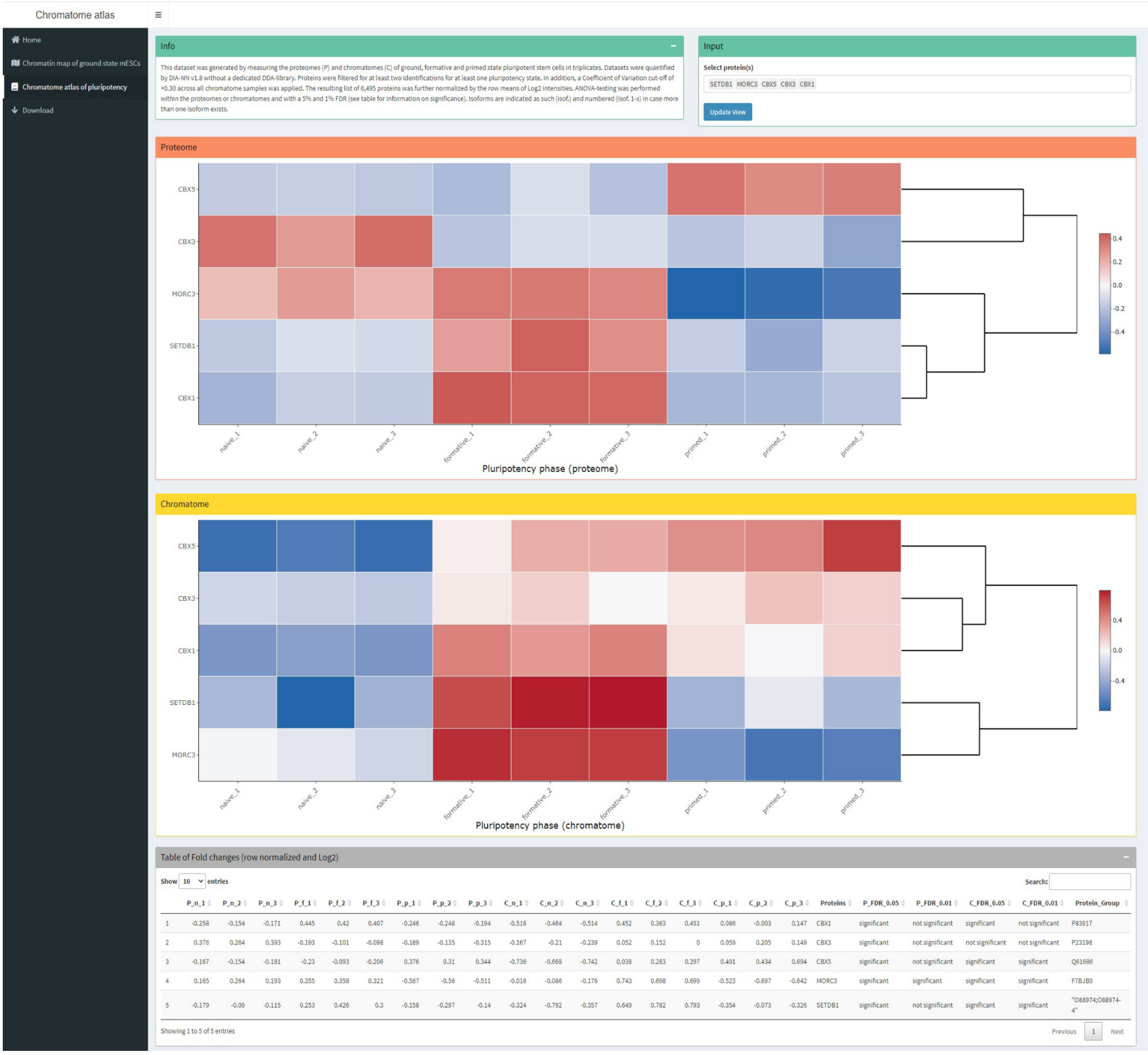
Show 10 entries
Search:

	Cytosol	Proteome	Nucleus	Chromatin_1	Chromatin_2	Chromatin_3	Proteins	Protein.Group	Cluster
1	1.73798666666667	0.919851666666667	-0.511930666666667	-0.695564666666667	-0.702800666666667	-0.747542333333333	HSP90AA1	P07901	VIII - Cytoplasm
2	1.71920666666667	0.947644	-0.508150666666667	-0.709106666666667	-0.727456	-0.722136666666667	HSP90B1	P08113	VIII - Cytoplasm
3	-2.04913	-0.171269666666667	0.168852	0.709455666666667	0.683650333333333	0.658441333333333	DNMT1	P13864	II - Chromatin I
4	-1.95626333333333	-0.363914333333333	0.0727177	0.731914333333333	0.753525	0.762019666666667	TET1	Q3URK3	II - Chromatin I
5	-1.93480666666667	-0.199151	-0.21253	0.775013333333333	0.777660666666667	0.793814	UHRF1	Q8VDF2	II - Chromatin I

Showing 1 to 5 of 5 entries

Previous 1 Next

B



This dataset was generated by measuring the proteomes (P) and chromatomes (C) of ground, formative and primed state pluripotent stem cells in triplicates. Datasets were quantified by DIA-NN v1.8 without a dedicated DDA-library. Proteins were filtered for at least two identifications for at least one pluripotency state. In addition, a Coefficient of Variation cut-off of >0.30 across all chromatome samples was applied. The resulting list of 6,495 proteins was further normalized by the row means of Log2 intensities. ANOVA-testing was performed within the proteomes or chromatomes and with a 5% and 1% FDR (see table for information on significance). Isoforms are indicated as such (isof.) and numbered (isof. 1-x) in case more than one isoform exists.

Select protein(s)
SETDB1 MORC3 CBX5 CBX3 CBX1

Table of Fold changes (row normalized and Log2)

Show 10 entries
Search:

	P_n_1	P_n_2	P_n_3	P_f_1	P_f_2	P_f_3	P_p_1	P_p_2	P_p_3	C_n_1	C_n_2	C_n_3	C_f_1	C_f_2	C_f_3	C_p_1	C_p_2	C_p_3	Proteins	P_FDR_0.05	P_FDR_0.01	C_FDR_0.05	C_FDR_0.01	Protein_Group
1	-0.258	-0.154	-0.171	0.445	0.42	0.407	-0.246	-0.248	-0.194	-0.516	-0.464	-0.514	0.452	0.363	0.451	0.086	-0.003	0.147	CBX1	significant	not significant	significant	not significant	P83917
2	0.376	0.264	0.393	-0.193	-0.101	-0.098	-0.189	-0.135	-0.315	-0.167	-0.21	-0.239	0.052	0.152	0	0.059	0.205	0.149	CBX3	significant	not significant	not significant	not significant	P23198
3	-0.167	-0.154	-0.181	-0.23	-0.093	-0.206	0.376	0.31	0.344	-0.736	-0.668	-0.742	0.038	0.283	0.297	0.401	0.434	0.694	CBX5	significant	not significant	significant	significant	Q61686
4	0.165	0.264	0.193	0.355	0.358	0.321	-0.587	-0.56	-0.511	-0.016	-0.086	-0.176	0.743	0.698	0.699	-0.523	-0.697	-0.642	MORC3	significant	significant	significant	significant	F7BJB9
5	-0.179	-0.09	-0.115	0.253	0.426	0.3	-0.158	-0.297	-0.14	-0.324	-0.792	-0.357	0.649	0.782	0.793	-0.354	-0.073	-0.326	SETDB1	significant	not significant	significant	significant	"O88974;O88974-4"

Showing 1 to 5 of 5 entries

Previous 1 Next

Figure S11 (Related to Figure 4)

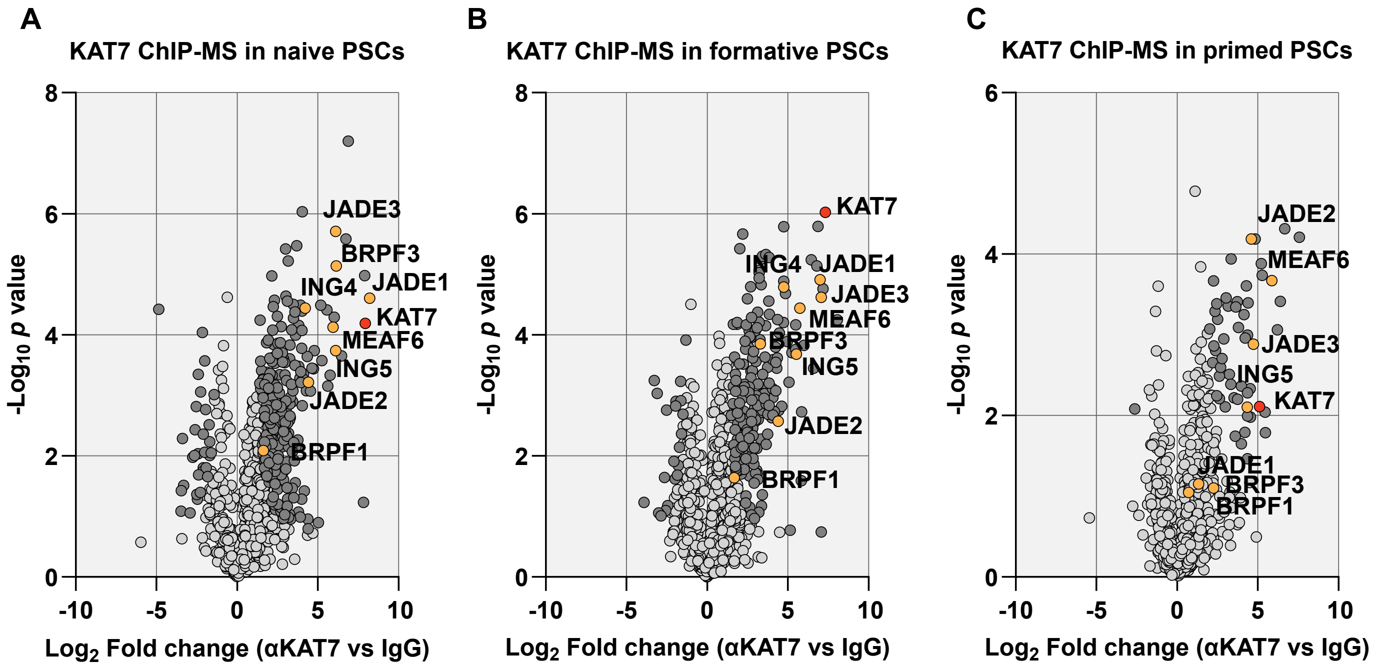
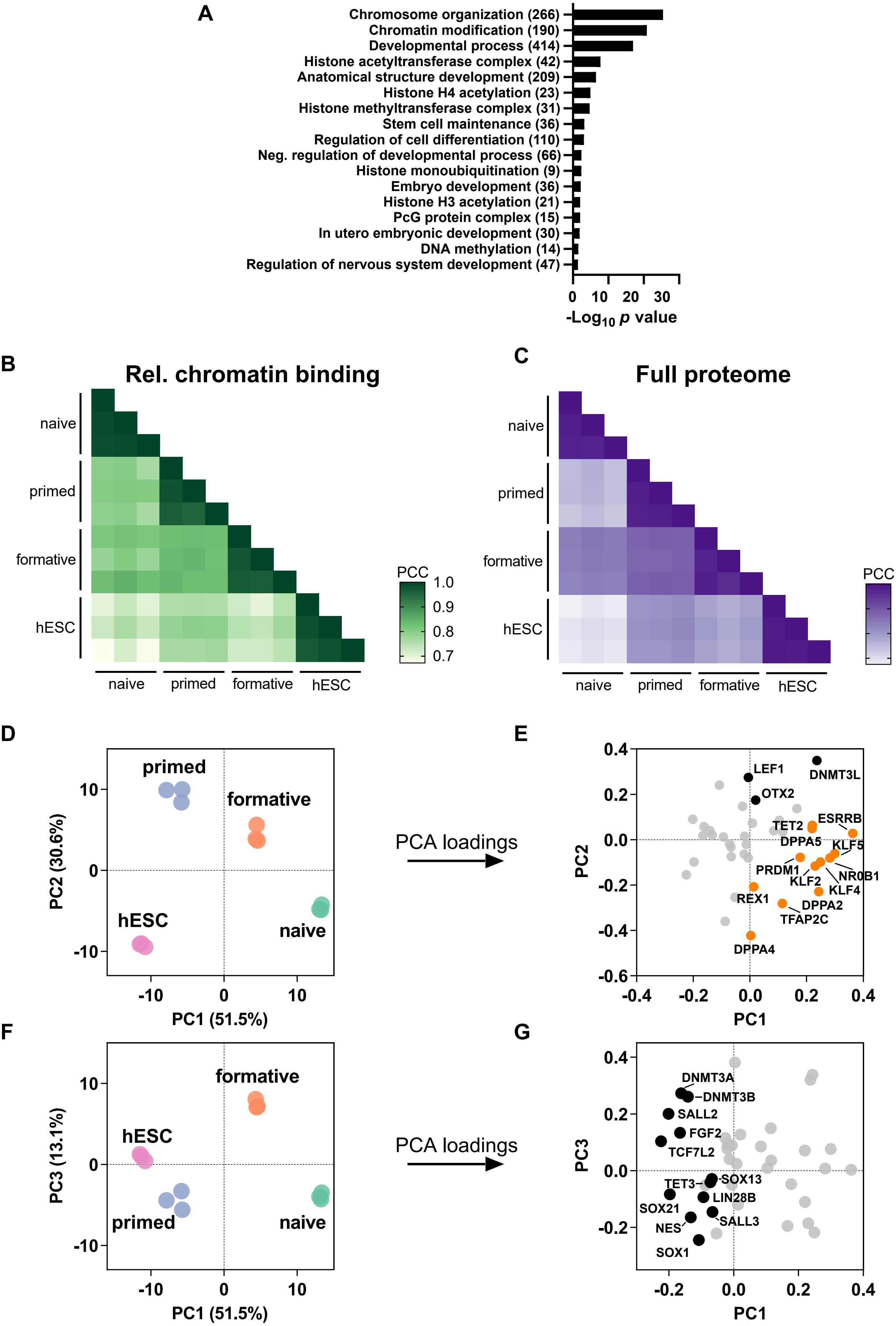


Figure S12 (Related to Figure 5)



Supplemental Figure legends

Figure S1. Evaluation of ChAC-DIA improvements, Related to Figure 1

A, B Total numbers of identified proteins (**A**) or precursors (**B**) with representation of the percentage coefficient of variation (CV) below 20% and 10%. We performed our protocol with changing only one parameter at once. DDA: same ChAC workflow, but without DIA. DIA: PAC step is omitted (instead, acetone precipitation is performed). +RNase: additional incubation of nuclei with RNaseA for 15 min at 37°C. XL of whole cells: FA crosslinking is performed before nuclei isolation. FA+DSSO: double crosslinking with FA and DSSO. The data was analyzed with Spectronaut. Experimental conditions were kept comparable by using the same cell pool. Briefly, DIA improved the protein identification rate by 37.9% with constant CVs around 7.3% compared to just DDA. Despite more than doubling precursor numbers, CVs on peptide-level were even reduced from 15.6% for DDA to 12.7% for DIA. Strikingly, CVs were further improved by the additional PAC step (median CVs for proteins 4.1% and for precursors 9.6%). Moreover, additional RNAse addition prior to nuclei lysis and the formaldehyde crosslinking of whole cells instead of nuclei impaired CVs especially on precursor level (22.2% and 18.4%, respectively). A combination of DIA and PAC after nuclei isolation without RNase addition therefore gave the best results in terms of sensitivity and reproducibility. **C** Effect of precursor isolation window numbers in DIA on total precursor identifications and CV. The data was analyzed with DIA-NN. MS2 resolution was constant at 30,000. Here, we found that 30 or 40 isolation windows outperform 15 windows regarding total precursor identifications by 11.8% and 19.3%, respectively, while keeping the CVs constant. **D, E** Total numbers of identified proteins (**D**) or precursors (**E**) with representation of the percentage CVs below 20% and 10% (corresponding experiment to **Figure 1B**). We compare a previous Chromatin Enrichment for Proteomics (ChEP)-based study (PRIDE: PXD011782) to ChAC-DIA quantified in directDIA mode by either Spectronaut or DIA-NN with or without matching across different purification fractions which are mentioned in **Figure 2A**. **F, G** Total numbers of proteins (**F**) or precursors (**G**) falling into a gene ontology (GO) category (corresponding experiment to **Figure 1C**). **H** Distribution of CVs of proteins selected by GO category. Asterisks represent digits after the decimal point of 0.05. For better readability only comparisons on protein-level are shown. However, each comparison between PXD011782 and any given ChAC-DIA analysis method was to the same extent significant on precursor-level.

Figure S2. Comparison of chromatome to proteome and transcriptome, Related to Figure 1

A Venn diagram of proteins identified in proteome and chromatome of naive PSCs. Biological replicates: n = 3. **B** Numbers of proteins falling into a GO category and that are either identified in the proteome and chromatome or in one of the experiments exclusively. **C** Numbers of proteins falling into a GO category in the proteome or transcriptome of naïve PSCs. **D** Protein abundance rank based on the naive PSC proteome. Chromatin binding and DNA-binding proteins are highlighted in magenta. Displayed protein names indicate the highest ranked 9 proteins.

E Scatter plot of all mESC expressed mRNAs corresponding to annotated "chromatin-binding" binding proteins. Abundances of mRNAs/proteins are scaled from 0 to 1. Total proteome abundances are used to color each individual mRNA/protein. Proteome and chromatome raw files were only matched within technical triplicates using DIA-NN v1.8. **F** Analysis of expressed mRNAs corresponding to "chromatin binding" proteins which have additional "nuclear", "cytoplasmic" or "membrane" localization. The bar diagram represents their % among all proteins missed (violet) or identified (grey) by ChAC-DIA. **G** Abundances of all expressed mRNAs corresponding to "chromatin binding" proteins across different ChAC-DIA purification steps. MS raw files were analyzed together with matching between runs by DIA-NN v1.8.

Figure S3. Reproducibility of differential fraction analysis during ChAC-DIA, Related to Figure 2

A-D Unsuprevised hierarchical clustering of R^2 values (**A**) and scatter plots of nucleus vs ChAC-DIA (3x washes) (**B**), ChAC-DIA 1x wash vs 3x washes (**C**) and two replicates of ChAC-DIA (3x washes) (**D**). **E** Fisher’s exact test to assess enriched GO terms of significantly enriched proteins in each cluster based on unsupervised hierarchical clustering in **Figure 2A**. P values are colour coded ($-\text{Log}_{10}$) and dot diameters correspond to group sizes (Log_2). **F** Percentage of a given GO category from the total cluster size of each cluster.

Figure S4. Complete chromatome list of pluripotency or differentiation related proteins, Related to Figure 3

A-C Heatmap representation of Log_2 FCs for significant differences between pluripotency phases (**A**), proteins absent in at least one pluripotency phase (**B**) or no ANOVA-based significant differences (**C**). **D** Total group size and percentage of (non-)significant differences.

Figure S5. Complete chromatome list of proteins annotated with „Transcription factor activity“, Related to Figure 3

A, B Heatmap representation of Log_2 FCs for significant differences between pluripotency phases (**A**) or proteins absent in at least one pluripotency phase (**B**). **C** Total group size and percentage of (non-)significant differences.

Figure S6. Complete chromatome list of proteins related to epigenetic regulation, Related to Figure 3

A, B Heatmap representation of Log_2 FCs for significant differences between pluripotency phases (**A**) or proteins absent in at least one pluripotency phase (**B**). **C** Total group size and percentage of (non-)significant differences. **D** Selection of identified and ANOVA-significant histone posttranslational modifications (PTMs) identified by ChAC-DIA. Data was analyzed by Spectronaut, column-wise normalized to median column intensity and subsequently row z-scored and averaged across triplicates.

Figure S7. Complete chromatome list harboring a Zinc finger domain, Related to Figure 3

A, B Heatmap representation of Log_2 FCs for significant differences between pluripotency phases (**A**) or proteins absent in at least one pluripotency phase (**B**). **C** Total group size and percentage of (non-)significant differences.

Figure S8. Complete chromatome list of proteins annotated with „chromatin remodeler“ or „chromatin organization“, Related to Figure 3

A, B Heatmap representation of Log$_2$ FCs for significant differences between pluripotency phases (**A**) or proteins absent in at least one pluripotency phase (**B**). **C** Total group size and percentage of (non-)significant differences.

Figure S9. Examples of chromatin-associated complexes, Related to Figure 3.

A-L Heatmap representation of chromatin-associated complexes and their Log$_2$ FCs between pluripotency phases.

Figure S10. Interactive web application with example input, Related to Figure 3.

A Chromatin map of ground state mESCs. Profile plot shows relative abundance of proteins of interest in the cytoplasmic, full proteome, nuclear and chromatome (after 1-3 washes) fractions **B** Chromatome atlas of pluripotency. Users can search for proteins of interest and their respective full proteome and chromatome levels. Unlike for heatmaps shown in Supplementary Figures S4-S9, missing values were imputed.

Figure S11. ChIP-MS of KAT7 in naive, formative and primed PSCs, Related to Figure 4.

A-C Volcano plots of KAT7 (Abcam, ab70183) vs Normal Rabbit IgG control (Cell Signaling, 2729) ChIP-MS in WT mouse PSCs at naive, formative and primed pluripotency states (n = 3 independent replicates, except for primed IgG control, which had 2 independent replicates). Dark grey dots: significantly enriched proteins after KAT7 pulldown. Red dot: KAT7. Orange dots: Proteins associated with HBO1 complex. Light grey dots: not significantly enriched proteins. Statistical significance is based on a Student's t-test with a permutation-based FDR of 0.05 and an s_0-cutoff of >1 (based on Log$_2$ FCs). Represented Log$_2$ FCs are computed by using LFQ values and are not normalized for KAT7-levels, unlike **Figure 4L**, which shows KAT7-normalized iBAQ values.

Figure S12. Comparison of proteomes and chromatomes between mouse naive, formative and primed PSCs as well as hESCs, Related to Figure 6

A Fisher’s exact results obtained from comparing the shared high confidence chromatome across all four tested cell lines against the total set of high confidence chromatome binders. Numbers in brackets represent groups size. **B, C** Pearson correlations of relative chromatin binding (**B**) or full proteomes (**C**) filtered for pluripotency or differentiation markers as in **Figure 6C**. Underlying data was filtered for only valid values (**B**) or at least 6 valid values in total and missing values were imputed based on a gaussian distribution relative to the standard deviations of measured values (width of 0.2 and a downshift of 1.8 standard deviations) (**C**). **D-G** PCA representations of projections (**D, F**) and individual loadings (**E, G**) based on chromatome values of bona fide pluripotency and differentiation markers as represented in **Figure 6G** and from each mouse PSC and hESCs. **D, F** are based on PC1 and PC2 whereas **E, G** on PC1 and PC3. Orange dots represent pre- implantation markers and black dots post-implantation markers.

Stolz, P., Mantero, A. S., Tvardovskiy, A., **Ugur, E.**, Wange, L. E., Mulholland, C. B., Cheng, Y., Wierer, M., Enard, W., Schneider, R., Bartke, T., Leonhardt, H., Elsässer, S. J., & Bultmann, S. (**2022**). **TET1 regulates gene expression and repression of endogenous retroviruses independent of DNA demethylation**. Nucleic Acids Research, 50(15), 8491-8511.

License: CC-BY-4.0

https://doi.org/10.1093/nar/gkac642

Published online 29 July 2022

Nucleic Acids Research, 2022, Vol. 50, No. 15 **8491–8511**
https://doi.org/10.1093/nar/gkac642

TET1 regulates gene expression and repression of endogenous retroviruses independent of DNA demethylation

Paul Stolz [1], Angelo Salazar Mantero[2], Andrey Tvardovskiy[3], Enes Ugur [1,5], Lucas E. Wange[4], Christopher B. Mulholland[1], Yuying Cheng[2], Michael Wierer [5], Wolfgang Enard[4], Robert Schneider[3], Till Bartke [3], Heinrich Leonhardt [1], Simon J. Elsässer [2] and Sebastian Bultmann [1,*]

[1]Faculty of Biology and Center for Molecular Biosystems (BioSysM), Human Biology and BioImaging, Ludwig-Maximilians-Universität München, Munich 81377, Germany, [2]Science for Life Laboratory, Department of Medical Biochemistry and Biophysics, Karolinska Institutet 17165 Stockholm, Sweden, Ming Wai Lau Centre for Reparative Medicine, Stockholm Node, Karolinska Institutet 17177 Stockholm, Sweden, [3]Institute of Functional Epigenetics (IFE), Helmholtz Zentrum München, 85764 Neuherberg, Germany, [4]Faculty of Biology, Anthropology and Human Genomics, Ludwig-Maximilians-Universität München 82152, Planegg-Martinsried, Germany and [5]Department of Proteomics and Signal Transduction, Max-Planck Institute of Biochemistry, Martinsried 82152, Germany

Received October 11, 2021; Revised April 25, 2022; Editorial Decision July 07, 2022; Accepted July 20, 2022

ABSTRACT

DNA methylation (5-methylcytosine (5mC)) is critical for genome stability and transcriptional regulation in mammals. The discovery that ten-eleven translocation (TET) proteins catalyze the oxidation of 5mC to 5-hydroxymethylcytosine (5hmC), 5-formylcytosine (5fC), and 5-carboxylcytosine (5caC) revolutionized our perspective on the complexity and regulation of DNA modifications. However, to what extent the regulatory functions of TET1 can be attributed to its catalytic activity remains unclear. Here, we use genome engineering and quantitative multi-omics approaches to dissect the precise catalytic vs. non-catalytic functions of TET1 in murine embryonic stem cells (mESCs). Our study identifies TET1 as an essential interaction hub for multiple chromatin modifying complexes and a global regulator of histone modifications. Strikingly, we find that the majority of transcriptional regulation depends on non-catalytic functions of TET1. In particular, we show that TET1 is critical for the establishment of H3K9me3 and H4K20me3 at endogenous retroviral elements (ERVs) and their silencing that is independent of its canonical role in DNA demethylation. Furthermore, we provide evidence that this repression of ERVs depends on the interaction between TET1 and SIN3A. In summary, we demonstrate that the non-catalytic functions of TET1 are critical for regulation of gene expression and the silencing of endogenous retroviruses in mESCs.

INTRODUCTION

DNA methylation is essential for the regulation of gene expression and genome stability in mammals (1). During development, methylated cytosine (5-methylcytosine (5mC)) serves as an epigenetic modification that prevents illegitimate cell fate decisions and contributes to coordination of the step-wise exit of pluripotency (2). The genome-wide landscape of 5mC is established during development by the de novo DNA methyltransferases DNMT3A and DNMT3B and maintained through subsequent cell divisions by the DNA methyltransferase DNMT1. The global 5mC patterns can be altered by the inhibition of maintenance DNA methylation and/or via the action of the Ten-eleven Translocation (TET) family of dioxygenases (3). The three mammalian homologs, TET1, TET2, and TET3 share a conserved C-terminal dioxygenase domain, which can catalyze the stepwise oxidation from 5mC to 5-hydroxymethylcytosine (5hmC), 5-formylcytosine (5fC), and 5-carboxylcytosine (5caC) (4–7). These oxidized cytosine derivatives have been described as intermediates of passive and active DNA demethylation (6,8–10), yet may also represent stable epigenetic marks on their own (11,12).

TET1 and TET3 possess a CXXC-type zinc finger domain that promotes their targeting to CpG-rich sequences,

---

[*]To whom correspondence should be addressed. Tel: +49 89 / 2180 71 026; Fax: +49 89 / 2180 74 236; Email: bultmann@bio.lmu.de

© The Author(s) 2022. Published by Oxford University Press on behalf of Nucleic Acids Research.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

whereas TET2 associates with IDAX, an independent CXXC domain-containing protein (13). The expression of TET proteins is highly dynamic during mouse preimplantation development. TET3 is strongly expressed in oocytes and zygotes followed by rapid depletion over the following cleavage stages, while TET1 and TET2 expression increase up to the blastocyst stage (14–16). In murine embryonic stem cells (mESCs), TET1 and TET2 are the main TET proteins expressed, whereas TET3 is present at very low to undetectable levels (17). Loss of all TET proteins is incompatible with normal mammalian development (18–21), as evidenced by the failure of TET-deficient mice to develop beyond gastrulation (20,21). In comparison, single TET mutants exhibit less severe yet distinct phenotypes, suggesting that each enzyme can partially compensate for loss of the other (22–24).

TET proteins demethylate regulatory regions including promoters, enhancers and distal regulatory elements (25). For instance, R-loop-dependent demethylation by TET1 is critical for transcriptional activation of the *Tcf21* promoter (26) and active DNA demethylation mediated by TET1 and TET2 has been demonstrated to facilitate somatic cell reprogramming (27). Furthermore, TET-catalytic activity restricts Polycomb domain boundaries to the promoters of developmentally regulated genes (28). In general, active DNA demethylation by TET1 as well as TET2 is responsible for maintaining the distinctive global DNA hypomethylation signature of naive mESCs, albeit indirectly via the locus-specific demethylation and transcriptional activation of *Dppa3* (29). Beyond this, it has become increasingly clear that TET proteins also regulate transcription independently of their catalytic activity. For example, the phenotype of full-length TET1 knockout (KO) mice differs from that of mice lacking the TET1 catalytic domain (23). Furthermore, TET1 mainly suppresses gene expression independent of its DNA demethylase activity in adipocytes (Villivalam *et al.*, 2020). Similarly, TET2 can activate gene expression independent of its catalytic activity via the direct interaction with the O-linked *N*-acetylglucosamine (O-GlcNAc) transferase (OGT) (30).

TET1 binds through its CXXC domain, both active and bivalent promoters and can act as either a transcriptional repressor or activator depending on the associated chromatin modifying complexes (13). At this, TET1 interacts with several protein complexes including Polycomb Repressive Complex 2 (PRC2) and the SIN3A histone deacetylase (SIN3A/HDAC) complex to regulate transcription (31–33). Several early studies demonstrated that TET1 accumulates at PRC2 targets and promotes the recruitment of the histone 3 lysine 27 trimethylation (H3K27me3)-depositing enzyme EZH2 to these sites (31–34). In addition, TET1 is also described to associate with SIN3A/HDAC, OGT, the histone acetyltransferase MOF, and chromatin remodeler MBD3/NuRD (32,35–37). These findings suggest that TET1 can regulate gene expression by coordinating chromatin modifying complexes.

In addition to gene regulation, TET1 has also been implicated in the repression of transposable elements (TEs) (38). In vertebrates, TEs are highly decorated by DNA methylation, which is essential for genomic stability (39–43). Counterintuitively, in mESCs young non-long terminal repeat (non-LTR) LINE-1 (L1) elements are highly decorated with 5hmC and maintained in a hypomethylated state by TET1, while their repression is mediated by SIN3A in a TET1-dependent manner (38). Furthermore, LTR-containing endogenous retroviruses (ERVs) were described to be specifically upregulated in TET triple KO (TKO) mESCs potentially due to loss of TRIM28 (also known as KAP1) binding (25). Besides DNA methylation, retrotransposons are repressed by the establishment of histone 3 lysine 9 trimethylation (H3K9me3) and histone 4 lysine 20 trimethylation (H4K20me3) (44–46). However, it is unclear how TET1-SIN3A is involved in the silencing machinery, repressing L1 elements. Furthermore, it is an open question if TET1-SIN3A might also regulate the activity of LTR retrotransposons, such as ERVs.

Taken together, these findings suggest that TET1 can mediate transcriptional regulation in a catalytically independent manner. However, the underlying molecular mechanisms as well as the extent of TET1's non-catalytic functions remain poorly understood.

Here, we systematically dissected the non-catalytic role of TET1 in mESCs. We used genome engineering and a quantitative multi-omics approach to compare a TET1 KO with a catalytically inactive TET1 mESC line. In particular, we find that (i) a large proportion of transcriptional changes are independent of TET1-mediated DNA demethylation; (ii) TET1 associates with different chromatin modifiers and is important for the establishment of specific histone modifications, namely H3K27me3, pan histone 4 lysine 5 + 8 + 12 + 16 acetylation (pH4Kac) and H4K20me3 and (iii) that loss of the TET1 protein but not its catalytic activity causes a specific loss of H3K9me3 at ERV1, ERVK and ERVL elements. Finally, we highlight that the interplay between TET1 and SIN3A is a main driver of ERV repression. Our results demonstrate that TET1 has a pivotal non-catalytic role in regulating gene expression and ERV silencing in mESCs.

MATERIALS AND METHODS

Cell culture

The generation of Tet1 KO (clone H9) and Tet1 CM (clone D7) mESC lines was described previously (17,29).

Mouse ESCs were cultured in 'Serum LIF' conditions and as independent replicates for 6 days prior to experiments. Here the cells were maintained on 0.2% gelatin-coated dishes in Dulbecco's modified Eagle's medium (Sigma) supplemented with 16% fetal bovine serum (FBS, Sigma), 0.1 mM ß-mercaptoethanol (Invitrogen), 2 mM L-glutamine (Sigma), 1× MEM Non-essential amino acids (Sigma), 100 U/ml penicillin, 100 μg/ml streptomycin (Sigma), homemade recombinant LIF tested for efficient self-renewal maintenance.

For the generation of piggybac doxycycline inducible cell lines, mESCs were cultured in 'Serum LIF 2i media'. Those were the same conditions as described above, but supplemented with 2i (1 μM PD032591 and 3 μM CHIR99021 (Axon Medchem, Netherlands)).

All cell lines were regularly tested for Mycoplasma contamination by PCR.

Piggybac constructs and cell line generation

The piggybac dox inducible TET1 (#102421) and TET1 CM (#102422) vector constructs were obtained from addgene (23). To generate the TET1 piggybac donor vector carrying a mutation at the Sin3a interaction domain (SID) (47), two overlapping PCR fragments were amplified.

Primers:

Sin3a_PsyI_FWD: 5' gtccatggactgcagtagacgtggtcatgggg aagaagagc 3'

Sin3a_NheI_REV: 5' ttactatactctatagctagctgctcttgcttcttc tgatc 3'

Sin3a_SID_FWD: 5' caagtggtagccatagaagccGCCactcag GCCtcagaag 3'

Sin3a_SID_REV: 5' cttctgaGGCctgagtGGCggcttctatgg ctaccacttg 3'

The resulting DNA fragments were cloned into the TET1 or TET1 CM piggybac vector digested with PsyI and NheI (Thermo Fisher Scientific) using a Gibson Cloning Kit (NEB).

To generate stable mESC lines carrying doxycycline-inducible forms of *Tet1, Tet1CM* or *Tet1 Sin3a mut.,* Tet1 KO mES cells were seeded at 0.5 mio mESCs in a 6-well plate and transfected with 1.5 μg of the pPB-tetO(hCMV1)-HA-Tet1mHxD(201R2)-IV (#102422, addgene) or pPB-tetO(hCMV1)-HA-Tet1(201R2)-IV (#102421, addgene) or pPB-tetO(hCMV1)-HA-Tet1Sin3a(201R2)-IV plasmid, 0.5 μg of the PiggyBac transposase vector (#PB200PA-1System, Biosciences) and 0.5 μg of the pPB-CAG-rtTA-IRES-Hygro (#102423, addgene) plasmid using Lipofectamine 3000 (Thermo Fisher Scientific) according to manufacturer's instructions. Two days after transfection, cells were plated at 10% confluency into a p100 plate and selected with Hygromycin (125 μg/ml) for 5–6 days. To enrich positive clones, cells were induced with doxycycline (1 μg/ml) for 24 h then sorted with flow cytometry on thresholded levels of mVenus expression. The mVenus fluorophore is under the control of the same promoter as Tet1 via an IRES sequence and therefore a fluorescent readout of successful induction. To ensure a stable pool the cell lines were sorted twice for mVenus expression. Post sorting, cells were plated back into media without doxycycline for 7 days before commencing experiments.

Western blot

Western blots for TET1 rescue and HP1β were performed as described previously (48) using monoclonal antibody rat anti-TET1 5D6 (1:10) (49), rabbit anti-HP1β (1:1000, 10478, abcam), rabbit anti-HP1β (1:1000, 8676, Cell Signaling) and polyclonal mouse anti-Tubulin (1:2500; T9026, Sigma-Aldrich) as loading control. Briefly, 1 million cells were collected and washed with ice-cold PBS (D8537, Sigma-Aldrich). The cells were lysed in 75 μl ice-cold RIPA buffer (50 mM TRIS/HCl pH 8.0, 150 mM NaCl, 0.1% UltraPure™ SDS Solution (24730020, Invitrogen), 0.5% sodium deoxycholate detergent, 1% Triton X-100; freshly add 1× cOmplete™ EDTA-free Protease Inhibitor Cocktail (04693132001, Roche), 2 mM PMSF, 0.1 U/μl Benzonase), mixed with 25 μl 4× Laemmli and boiled for 10 min at 95°C. Samples were separated by 8% (TET1) and 10% (HP1β) SDS-Page Mini-Protean system (Bio-Rad) and transferred to a nitrocellulose membrane (0,2 μM) using wet transfer (Bio-Rad). After blocking (1h, 5% milk in PBS-Tween), the blots were probed with the before mentioned primary antibodies and the corresponding secondary antibodies goat anti-rat (1:5000; 112-035-068, Jackson ImmunoResearch), goat anti-rabbit (1:5000, 170-6515, Bio-Rad) and goat anti-mouse (1:5000; A9044, Sigma-Aldrich) conjugated to horseradish peroxidase (HRP) and visualized using an ECL detection kit (Thermo Scientific Pierce).

MINUTE-ChIP

The quantitative multiplexed ChIP experiments were conducted as previously described (50). In short, three cell lines (WT J1, Tet1 KO H9 and Tet1 CM D7) were cultured as independent quadruplicates, cell pellets of 2 mio cells were lysed in Lysis Buffer and digested with 6 U/μl Micrococcal nuclease for 10 min at 37ºC. T7-adapters with 6 bp unique molecular identifying (UMI) sequences and 8bp sample barcodes were ligated to the chromatin fragments for 2 h at 23ºC and subsequently for 16 h at 16ºC. The twelve samples were thereafter pooled together and 2 mio cell equivalents of digested and barcoded chromatin was used for immunoprecipitation using antibodies for the histone marks H3K4me3 (04-745, Millipore), H3K27me3 (07-449, Millipore), H3K27me1 (61015, Active Motif), H4K20me3 (07-463, Millipore), H4K20me1 (ab9051, Abcam), pH4Kac (06-598, Sigma) and H3K9me3 (39161, Active Motif). The antibodies were coupled to SureBeads Protein A (1614013, Bio-Rad) and Protein G (1614023, Bio-Rad) magnetic beads and the immunoprecipitation was conducted for 4 h at 4ºC with rotation, followed by quick washes using RIPA and LiCl buffers. The immunoprecipitated chromatin was eluted from the beads and subjected to Proteinase K for 1 h at 63ºC. A sample consisting of 0.2 mio cell equivalent from the pooled lysates was also subjected to Proteinase K digestion as input for later normalization purposes. The digested DNA was cleaned up using AMPureXP SPRI beads (A63881, Beckman Coulter). The barcoded DNA fragments were in vitro transcribed for 16 h at 37ºC followed by DNase digestion for 15 min at 37ºC and purified using Silane beads (37002D, Thermo Fisher Scientific). RA3 RNA adapters were ligated to the transcripts for 2 h at 25ºC followed by reverse transcription to cDNA using a paired end primer. The cDNA was cleaned up using AMPureXP SPRI beads. 150 ng of cDNA was used for library PCR using a different barcoded primer for each sample. Finally, the libraries were diluted to 4 nM and combined for sequencing using Illumina sequencing.

MINUTE-ChIP analysis

We conducted the MINUTE-ChIP data analysis as previously described (51). The bioinformatic pipeline for MINUTE-ChIP data analysis is available at github (https://github.com/NBISweden/minute).

Preparation of FASTQ files. Sequencing was performed using 50:8:34 cycles (Read1:Index1:Read2) Illumina bcl2fastq was used to demultiplex paired-end sequencing reads by 8nt index1 read (PCR barcode). NextSeq lanes were merged into single fastq files, creating the primary fastq files. Read1 starts with 6nt UMI and 8nt barcode in the format NNNNNNABCDEFGH.

Primary analysis. MINUTE-ChIP multiplexed FASTQ files were processed using minute, a data processing pipeline implemented in Snakemake (52). In order to ensure reproducibility, a conda environment was set. Source code and configuration are available on GitHub: https://github.com/NBISweden/minute. Main steps performed are described below.

Adaptor removal. Read pairs matching parts of the adaptor sequence (SBS3 or T7 promoter) in either read1 or read2 were removed using cutadapt v3.2 (53).

Demultiplexing and deduplication. Reads were demultiplexed using cutadapt v3.2 allowing only one mismatch per barcode. Demultiplexed reads were written into sample-specific fastq files used for subsequent mapping and GEO submission.

Mapping. Sample-specific paired fastq files were mapped to the mouse genome (mm10) using bowtie2 (v2.3.5.1) with –fast parameter. Alignments were processed into sorted BAM files with samtools (v1.10). Pooled BAM files were generated from replicates using samtools.

Deduplication. Duplicate reads are marked using UMI-sensitive deduplication tool je-suite (v2.0.RC) (https://github.com/gbcs-embl/Je/). Read pairs are marked as duplicates if their read1 (first-in-pair) sequences have the same UMI (allowing for 1 mismatch) and map to the same location in the genome. Blacklisted regions were then removed from BAM files using BEDTools (v2.29.2).

Generation of coverage tracks and quantitative scaling. Input coverage tracks with 1bp resolution in BigWig format were generated from BAM files using deepTools (v3.5.0) bamCoverage and scaled to a reads-per-genome- coverage of one (1xRPGC, also referred to as '1× normalization'). ChIP coverage tracks were generated from BAM files using deepTools (v3.5.0) bamCoverage. Quantitative scaling of the ChIP-Seq tracks amongst conditions within each pool was based on their Input-Normalized Mapped Read Count (INRC). INRC was calculated by dividing the number of unique mm10-mapped reads by the respective number of Input reads: #mapped[ChIP]/#mapped[Input]. This essentially corrected for an uneven representation of barcodes in the Input and we previously demonstrated that the INRC is proportional to the amount of epitope present in each condition (50). Wildtype mESC (replicates combined) were chosen as the reference condition, which was scaled to 1x coverage (also termed Reads per Genome Coverage, RPGC). All other conditions were scaled relative to the reference using the ratio of INRCs multiplied by the scaling factor determined for 1x normalization of the reference: (#mapped[ChIP]/#mapped[Input])/(#mapped[ChIP_Reference]/#mapped[Input_Reference]) × scaling factor.

Quality control. FastQC was run on all FASTQ files to assess general sequencing quality.

Picard (v2.24.1) was used to determine insert size distribution, duplication rate, estimated library size. Mapping stats were generated from BAM files using samtools (v1.10) idxstats and flagstat commands. Final reports with all the statistics generated throughout the pipeline execution are gathered with MultiQC (54).

ChIP analysis of published data sets

We analysed published ChIP-seq reads of TET1 (34), SIN3A (55), SETDB1 (56) and H3K9ac (57) of WT mESC cultured in SL medium. Reads were aligned to the mouse genome (mm10) with Bowtie (v.1.2.2) with parameters '-a -m 3 -n 3 –best –strata'. Subsequent ChIP–seq analysis was carried out on data of merged replicates. Peak calling and signal pile up was performed using MACS2 callpeak (58) with the parameters '–extsize 150' for ChIP, '–extsize 220–nomodel -B –nolambda' for all samples. Reads mapping to Repeats (defined by RepeatMasker mm10) were extracted using custom R scripts.

Enzymatic methylome sequencing (EM-seq)

Three cell lines (WT J1, Tet1 KO H9 and Tet1 CM D7) were cultured as independent triplicates. The genomic DNA was isolated using the QIAamp DNA Mini Kit (QIAGEN). DNA concentration was measured using Nanodrop (NanoPhotometer NP80, Implen). The gDNA was then diluted to 10 ng/μl in 200 μl TE buffer. To control the conversion efficiency 0.01 ng pUC19 methylated DNA and 0.2 ng unmethylated lambda DNA were added. The DNA was sheared into 350–400 bp fragments using the Bioruptor Plus sonication device (Diagenode) (30 s on/off, 20 cycles). Bioanalyzer (Agilent) was used to control for the shearing efficiency. For library preparation 200 ng of the sheared DNA were used. The final EM-seq library preparation was performed according to the manufacturer's instructions (New England Biolabs).

EM-seq processing and analysis

The EM-seq library was a paired end sequencing run, 2 × 150 bp (Novogene). Raw reads were first trimmed using Trim Galore (v.0.3.1). Alignments were carried out to the mouse genome (mm10) using bsmap (v.2.90) using the parameters '-s 12 -v 10 -r 2 -I 1'. CpG-methylation calls were extracted from the mapping output using bsmaps methratio.py. Analysis was restricted to CpG with a coverage >10. methylKit (59) was used to identify differentially methylated regions between the respective contrasts for the following genomic features: (i) all 1-kb tiles (containing a minimum of three CpGs) detected by EM-seq; (ii) repeats (defined by RepeatMasker mm10); (iii) gene promoters (defined as gene start sites −2 kb/+2 kb) and (iv) gene bodies (defined as longest isoform per gene) and CpG islands (as defined by (60)). Differentially methylated regions were identified as regions with $P < 0.05$ and a difference in methylation means between two groups >20%. DNA methylation browser track figures were created using IGV (v2.9.2).

Relative quantification of histone post translational modification abundances using LC-MS/MS

Histones were acid extracted as described previously (61). In brief, mESCs were lysed in 10× cell pellet volume of ice-cold hypotonic lysis buffer (15 mM Tris|HCl (pH 7.5), 60 mM KCl, 11 mM CaCl2, 5 mM NaCl, 5 mM MgCl2, 250 mM sucrose, 1 mM dithiothreitol, 10 mM sodium butyrate) supplemented with 0.1% NP-40 on ice for 5 min. Nuclei were pelleted by centrifugation (1000g, 2 min, 4○C) and washed twice in ice-cold hypotonic lysis buffer w/o NP-40. Nuclei were resuspended in 5× nuclei pellet volumes of ice-cold 0.2 M sulfuric acid and mixed on a rotation wheel for 120 min at 4°C. Insolubilized nuclear debris was pelleted by centrifugation (16 000g, 10 min, 4°C). Supernatant was transferred to a fresh low-protein binding Eppendorf tube and histone proteins were precipitated by adding ice-cold trichloroacetic acid (TCA) to the final concentration of 20% (v/v) followed by 60 min incubation on ice. Precipitated histone proteins were pelleted by centrifugation (16 000g, 10 min, 4°C), washed 3 times with acetone (–20°C) and resuspended in MS grade water.

Extracted histones were prepared for LC–MS/MS analysis using hybrid chemical derivatization method as described previously (62). In brief, 4 μg aliquots of purified histones were diluted with MS grade water to a total volume of 18 μl and buffered to pH 8.5 by addition of 2 μl of 1 M triethylammonium bicarbonate buffer (TEAB). Propionic anhydride was mixed with MS grade water in a ratio of 1:100 and 2 μl of the anhydride-mixture was added immediately to the histone sample, with vortexing, and the resulting mixture was incubated for 5 min at room temperature. The reaction was quenched by adding 2 μl of 80 mm hydroxylamine followed by 20 min incubation at room temperature. Tryptic digestion was performed overnight with 0.5 μg trypsin per sample at 37○C. A 1% v/v solution of phenyl isocyanate (PIC) in acetonitrile was freshly prepared and 6 μl added to each sample and incubated for 60 min at 37°C. Samples were acidified by adding trifluoroacetic acid (TFA) to the final concentration of 1%. Peptides were de-salted with C18 spin columns (Pierce™) following the manufacture protocol. Peptides were eluted from C18 spin columns with 70% acetonitrile, partially dried in a speedvac and resuspended in 30 μl 0.1% TFA.

The resulting peptide mixtures were analyzed using nanoflow liquid chromatography–tandem mass spectrometry (LC–MS/MS) on a Q-Exactive HF mass spectrometer coupled to an Ultimate 3000 nano-UPLC (Ultimate 3000, Dionex, Sunnyvale, CA) in data-dependant acquisition (DDA) mode. ~300 ng peptide aliquot was used per one sample per one injection. Peptides were loaded automatically on a trap column (300 μm inner diameter × 5 mm, Acclaim PepMap100 C18, 5 μm, 100 Å; LC Packings, Sunnyvale, USA) prior to C18 reversed phase chromatography on the analytical column (nanoEase MZ HSS T3 Column, 100 Å, 1.8 μm, 75 μm × 250 mm; Waters, Milford, USA). Peptides were separated at flow rate of 0.250 μl per minute by a linear gradient from 1% buffer B (0.1% (v/v) formic acid, 98% (v/v) acetonitrile) to 25% buffer B over 40 min followed by a linear gradient to 40% B in 20 min, then to 85% B in 5 min. After 5 min at 85% buffer B, the gradient was reduced to 1% buffer B over 2 min and then allowed to equilibrate for 8 min. Full mass range spectra were at 60 000 resolution (at m/z 400), and product ions spectra were collected in a 'top 15' data-dependent scan cycle at 15 000 resolution.

RAW MS data were analyzed using EpiProfile 2.0 software (63). The reported relative abundances of histone modifications were validated by manual re-quantification using an open-source Skyline software.

Cell growth and morphology analysis

The time evolution of cell growth and cell morphology was determined using the PHIO Cellwatcher (www.phio.de). WT J1, Tet1 KO and Tet1 CM mESCs lines were cultured in Serum LIF media as described. The Cellwatcher was placed inside the incubator and images with a large field of view of 10 mm^2 were automatically recorded every 30 min. The cell proliferation and morphology data were gained with PHIO's automatic AI-based analysis platform and were accessed through PHIO's data dashboard www.phio-cells.com.

For cell counting, WT J1, Tet1 KO and Tet1 CM mESCs lines were seeded in 6-well plates at densities of 0.35 mio mESCs/well in five replicates. The cells were collected and counted after 24 and 48 h using an automated cell counter (Countstar BioTech).

RNA-seq library

For RNA-seq, three different cell lines (WT J1, Tet1 KO H9, Tet1 CM D7) were cultured as independent quadruplicates. RNA was isolated using the NucleoSpin Triprep Kit (Machery-Nagel) according to the manufacturer's instructions. Isolated total RNA was normalised and subjected to RNA sequencing using a version of the prime-seq method (64). This method is based on the single cell RNA-seq method mcSCRB-seq (65) and is a three prime counting method that includes a sample specific barcode sequence and unique molecular identifiers (UMI) for accurate quantification of gene expression. Here we used the Nextera XT Kit (Illumina) for sequencing library preparation as described in the mcSCRB-seq protocol (65). Illumina paired end sequencing was performed on an HiSeq 1500 instrument for the first two experiments and on a NextSeq 1000 instrument for the third experiment. The first read was 16–28 bases long and covered the sample barcode and UMI, the second read was 50–109 bases long and read the cDNA fragment. Raw data was demultiplexed using deML (66), adapters and poly A tails were trimmed using cutadapt (53) and further preprocessed using the zUMIs pipeline (67) with STAR (68). Reads were mapped to the mouse genome (mm10) with either Ensembl annotation for the first experiment (GRCm38 release 102) or Gencode annotation (v M25) for the later experiments.

RNA-seq processing and analysis

RNA-seq libraries were processed and mapped to the mouse genome (mm10) using the zUMIs pipeline (67). UMI count tables were filtered for low counts using HTSFilter

(69). Differential expression analysis was performed in R using DESeq2 (70) and genes with an adjusted P <0.05 and an LFC >abs(1) were considered to be differentially expressed. Differential expression analysis over transposable elements was performed using TEtranscript (71).

Immunofluorescence staining

For immunostaining, mESCs were grown on coverslips coated with Geltrex (Life Technologies), thereby allowing better visualization during microscopic analysis. All steps during immunostaining were performed at room temperature. Coverslips were rinsed two times with PBS (pH 7.4; 140 mM NaCl, 2.7 mM KCl, 6.5 mM Na_2HPO_4, 1.5 mM KH_2PO_4) prewarmed to 37°C, cells fixed for 10 min with 4% paraformaldehyde (pH 7.0; prepared from paraformaldehyde powder (Merck) by heating in PBS up to 60°C; stored at –20°C), washed three times by dipping in PBST (PBS, 0.01% Tween20), permeabilized for 5 min in PBS supplemented with 0.5% Triton X-100, and washed two times by dipping in PBS. Primary and secondary antibodies were diluted in blocking solution (PBST, 4% BSA). Coverslips were incubated with primary and secondary antibody solutions (PBST, 4% BSA) in dark humid chambers for 1 h and washed three times by dipping in PBST after primary and secondary antibodies. For DNA counterstaining, coverslips were incubated 6 min in PBST containing a final concentration of 2 μg/ml DAPI (Sigma-Aldrich) and washed three times for 10 min with PBST. Coverslips were mounted in antifade medium (Vectashield, Vector Laboratories) and sealed with colorless nail polish.

Following primary antibodies were used: polyclonal rabbit anti-HP1β (1:300; 10478, abcam), monoclonal mouse anti-HP1α (1:100, 05-689, Sigma-Aldrich), monoclonal mouse anti-HP1γ (1:100, MA3-054, Invitrogen) and monoclonal rat anti-TET1 (1:10; 5D6). Following secondary antibodies were used: polyclonal donkey anti-rabbit Alexa 488 (1:500; 711-547-003, Dianova), polyclonal donkey anti-rat 488 (1:500, A-21208, Life technologies), polyclonal donkey anti-rabbit Alexa 647 (1:500, A-21244, ThermoFisher Scientific), polyclonal donkey anti-mouse Alexa 647 (1:500, A-31571, Invitrogen).

Immunofluorescence imaging and analysis

Images were acquired on the Leica TCS SP8 X using 63× glycerol immersion objective and high-content screening Operetta microscope using a 20× objective. DAPI or fluorophores were excited with 405, 488 or 594 nm laser lines. Within each experiment, cells were imaged using the same settings on the microscope (camera exposure time, laser power and gain) to compare signal intensities between cell lines.

Images were analyzed using Fiji software (ImageJ 1.51j) for SP8 images and Harmony software package for Operetta images.

The coefficient of variance (CV) of the respective fluorescent signal was calculated as follows: (standard deviation/mean) × 100. The mean fluorescence and standard deviation of the fluorescence signal was acquired and calculated with the Operetta microscope and Harmony software package. To calculate the CV of the KO + TET1 and KO + TET1 SIN3A mut. rescue experiments, we used a TET1 antibody staining to identify cells with TET1 expression. The cells were separated into TET1 positive (488 nm mean intensity > 1500) and TET1 negative (488 nm mean intensity < 1500) and the CV calculated of the respective population.

Mass spectrometry-based proteomic analysis of chromatin immunoprecipitated samples

Chromatin immunoprecipitation coupled to Mass Spectrometry (ChIP-MS) of TET1 was performed in triplicates for WT and TET1 KO mESCs under Serum LIF condition. For the pulldown a direct TET1 antibody (09-872-I, Sigma-Aldrich) was employed. ChIP-MS was performed as described previously, but without MNase digestion (72). Briefly, for each replicate a 15 cm cell culture dish was cultured for 2 days and 15 mio cells were crosslinked by 1% paraformaldehyde. Cells were lysed by the IP buffer (1.7% Triton X-100, 100 mM NaCl, 50 mM Tris–HCl pH 8.0, 5 mM EDTA pH 8.0, 0.3% SDS and freshly added 1x protease inhibitor cocktail) by pipetting and resting for 10 min on ice. Chromatin was sheared by sonication for 15 min in a Bioruptor Plus (30 s on/off cycles, Diagenode). Shearing efficiency was checked after overnight reverse crosslinking and proteinase K digestion of samples on a 1% agarose gel. Protein concentrations were estimated by BCA assay (Thermo) and samples were diluted to 1.3 mg/ml in 1 ml. 1.7 μg of the antibody was added to each replicate and samples were incubated O/N at 4°C under constant rotation. The next day magnetic protein A/G beads (20 μl slurry volume/sample, Sigma) were added to each sample to wash out unspecific interactors. After two low salt (50 mM HEPES pH 7.5, 140 mM NaCl, 1% Triton X-100), one high salt (50 mM HEPES pH 7.5, 500 mM NaCl, 1% Triton X-100) and two TBS washes, proteins were incubated in 2 mM DTT and subsequently 40 mM CAA (both diluted in 2 M Urea and 50 mM Tris–HCl pH 7.5). Then proteins were on-bead digested by Trypsin (20 μg/ml) O/N at 25°C. The next day, protease activity was stopped by 1% TFA and peptides were cleaned-up on Stage Tips consisting of three layers of C18 material (Empore) (73). After elution from Stage Tips peptides were speedvac dried and resuspended in 20 μl of A^* buffer (0.1% TFA and 2% acetonitrile). Peptide concentrations were estimated by nanodrop measurements at 280 nm.

300 ng of each peptide solution was analyzed on a quadrupole Orbitrap mass spectrometer (Orbitrap Exploris™ 480, Thermo Fisher Scientific) after nanoflow liquid chromatography on an in-house packed 50 cm column (ReproSil-Pur C18-AQ 1.9 μM resin, Dr Maisch GmbH) coupled to an Easy-nLC 1200 (Thermo Fisher Scientific) over a linear acetonitrile gradient for 120 min. Data-dependent acquisition was employed and thereby the most abundant 12 peptides were selected for MS/MS scans. The target value for full scan MS spectra was set to 3×10^6 and the resolution was at 60 000. The m/z range was adjusted to 400–1650 m/z and the maximum injection time was limited to 20 ms.

Subsequent data analysis of raw MS files was first accomplished by the MaxQuant software package (version 1.6.0.7) (74). Protein sequences were acquired over the Uniprot database (reviewed and unreviewed, version 2020) as a FASTA file. The MaxQuant analysis comprised the 'Match between runs' option, a false discovery rate for both peptides (minimum length of 7 amino acids) and proteins of 1% and determination of proteins amounts by the MaxLFQ algorithm (75). Downstream analysis was then performed with the Perseus software package (version 1.6.0.9). A two-sided Student's *t*-test of the $\log_2$ transformed LFQ intensities was performed to obtain significantly enriched proteins. By definition, a permutation-based false discovery rate of 5% and a fold change cut-off of $\log_2 = 1$ was applied.

RESULTS

TET1 regulates gene expression mainly independent of its catalytic activity in mESCs

To dissect the catalytic and non-catalytic contributions of TET1, we used our previously described *Tet1* knockout (Tet1 KO) and *Tet1* catalytic mutant (Tet1 CM) mESCs (17,29). All cell lines were cultured in standard mESC media containing serum and leukemia inhibitory factor LIF (SL). We observed a striking difference in growth and morphology among wildtype (WT), Tet1 KO and Tet1 CM cells. Compared with WT and Tet1 CM cells, Tet1 KO mESC colonies exhibited a much flatter and less rounded morphology, a classical morphological hallmark of reduced pluripotency and spontaneous differentiation (Supplementary Figure 1A, B). While both Tet1 KO and Tet1 CM showed impaired cell growth, only Tet1 KO cells were altered in shape and size (Supplementary Figure 1A, B). To determine the transcriptional consequences of TET1 inactivation compared with total loss of TET1 proteins, we performed bulk RNA-seq (prime-seq (64)) on Tet1 KO, Tet1 CM, and WT mESCs. Differential gene expression analysis between WT and each of the TET1 mutant cell lines revealed that loss and catalytic inactivation of TET1 resulted in transcriptional activation as well as repression (Figure 1A), in line with TET1's dual role in transcriptional regulation (33). Strikingly, however, we found in Tet1 KO mESCs ∼5 times more genes (2020) to be differentially expressed than in Tet1 CM mESCs (459). This small subset of genes deregulated in Tet1 CM mESCs was almost entirely composed of genes also deregulated in Tet1 KO mESCs (Supplementary Figure 2A), strongly suggesting that these are catalytically-dependent TET1 targets. While these catalytically-dependent genes exhibited the same directionality of expression changes (up- or downregulation) in both Tet1 KOs and Tet1 CMs, the extent of deregulation in terms of fold-change was more severe in Tet1 KO mESCs (Supplementary Figure 2B). This discrepancy in comparison to Tet1 CM mESCs implies that these genes are subject to synergistic catalytic and non-catalytic regulation by TET1. Next, we performed a Gene Set enrichment analysis to investigate whether genes controlled by TET1 cluster into functional groups. We detected several significantly deregulated gene sets with enriched Gene Ontology (GO) terms in the Tet1 KO mESCs, yet no encriched gene sets in the Tet1 CM mESCs (Supplementary Table 1). In line with our observation of a differentiated cell morphology upon TET1 loss, we found several developmental GO terms such as 'gastrulation', 'embryonic organ development', and 'cell differentiation' enriched among significantly upregulated genes in Tet1 KO mESCs. In contrast, significantly downregulated genes in Tet1 KOs were associated with naive pluripotency GO terms such as 'germ cell development', 'response to leukemia inhibitory factor', and 'spermatogenesis' (Supplementary Table 1). These findings indicate that TET1 is important for maintaining the balance between pluripotency and lineage commitment.

To further investigate whether these changes in gene expression are dependent or independent of TET1's catalytic activity, we performed two rescue experiments. In particular, we used PiggyBac-mediated transposition to stably express TET1 or TET1 CM in Tet1 KO mESCs upon induction with doxycycline (Supplementary Figure 2C) (23). We then performed bulk RNA-seq (prime-seq (64)) to study the global effect on the transcriptome upon re-expression of TET1 or TET1 CM. In contrast to reintroducing TET1, TET1 CM cannot stimulate active DNA demethylation and hence cannot rescue genome-wide DNA modification levels (23). However re-expression of both TET1 or TET1 CM resulted in the repression of developmental markers upregulated in Tet KO cells such as genes involved in gastrulation (e.g. *Ets2*, *Mbp* and *Nog*, Figure 1B, Supplementary Figure 2D). Similarly, genes downregulated in Tet1 KO cells such as those involved in germ cell development (e.g. *Zfp42* and *Prdm14*) were upregulated after re-expression of either TET1 or TET1 CM (Figure 1C, Supplementary Figure 2D). Taken together, these results are consistent with previous findings (32,33,76) and reveal that loss of TET1 results in the upregulation of developmental genes as well as the downregulation of naive pluripotency markers. Remarkably, we find that TET1 controls these genes largely independently of its catalytic activity.

Finally, we asked whether the transcriptional dysregulation in TET1 mutant ESCs might be attributable to changes in DNA methylation. To address this question we performed enzymatic methylome sequencing (EM-seq, Supplementary Table 2). Strikingly, the loss of TET1 resulted in widespread promoter hypermethylation (Supplementary Figure 2E). However, we found that, in the majority of cases, increased promoter methylation was not accompanied by changes in gene expression, in line with previous studies (23,77,78) (Figure 1D, Supplementary Figure 2E). Only a small cluster of genes were found to be both downregulated and exhibit promoter hypermethylation in Tet1 KO as well as Tet1 CM mESCs, suggesting that there are relatively few bona fide catalytic targets of TET1 (Supplementary Figure 2E and F). The majority of studies have reported hypermethylation (17,23,32,79,80) while some have shown hypomethylation in Tet1 KO mESCs (81,82). Overall, we observed genome-wide hypermethylation in Tet1 KO mESCs, which was less pronounced in Tet1 CM mESCs (Figure 1E). We detected an increase in DNA methylation at promoter, enhancer, gene bodies and TEs in Tet1 KO and Tet1 CM compared to WT mESCs. DNA methylation gains were broadly correlated between Tet1 KO and Tet1 CM, with Tet1 KO showing a larger effect size (Figure 1E, Supplementary Figure 2G). Collectively, we found that TET1

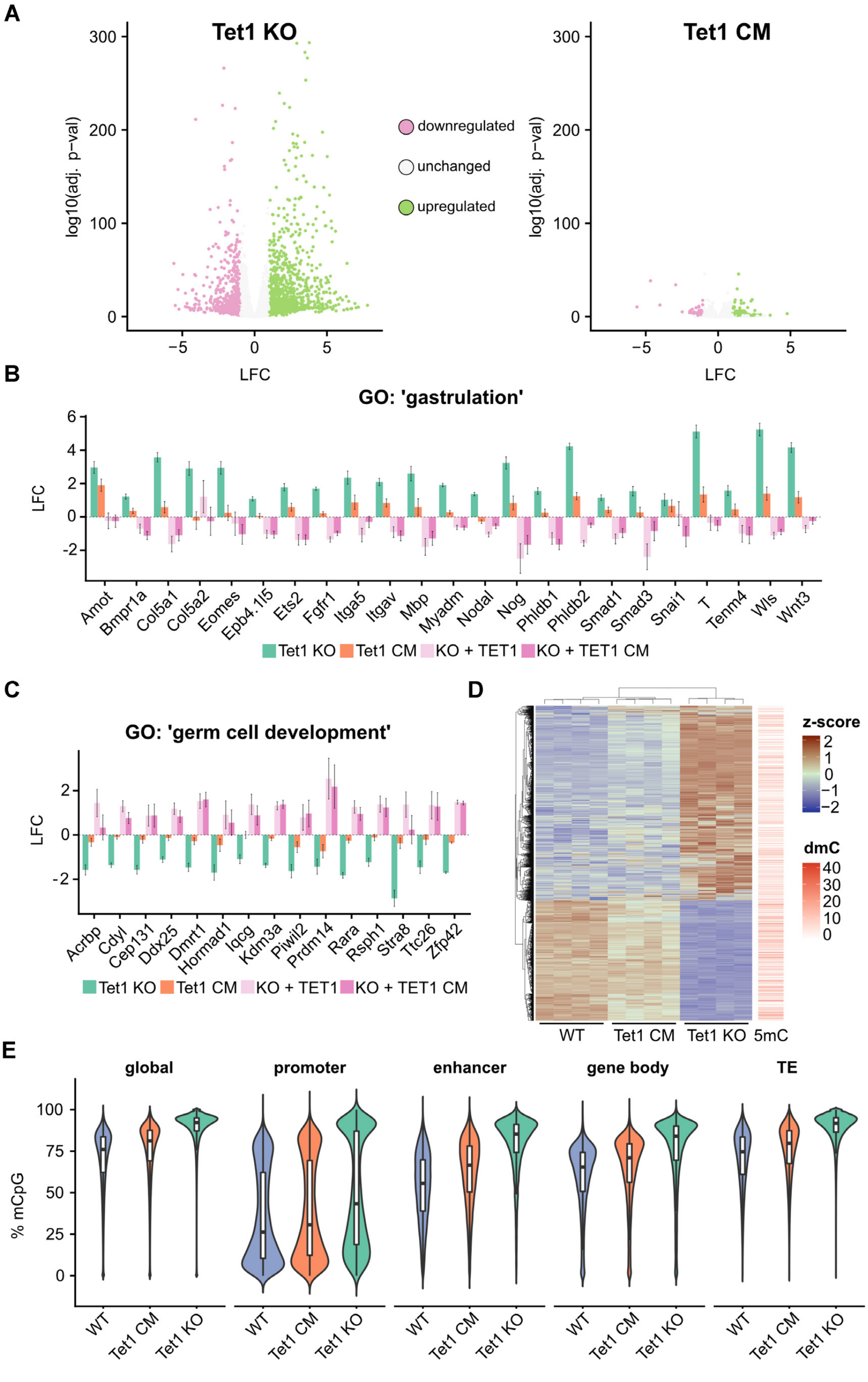


Figure 1. TET1 regulates gene expression mainly independent of its catalytic activity. (**A**) Volcano plots illustrating the transcriptional changes ($\log_2$-fold change, LFC) of Tet1 knockout (Tet1 KO) and catalytic mutant (Tet1 CM) mESCs relative to WT mESCs as assessed by RNA-seq. Green dots: Upregulated genes (Tet1 KO = 1250; Tet1 CM = 91). Violet dots: Downregulated genes (Tet1 KO = 770; Tet1 CM = 39). Grey dots: Unchanged expression. The threshold for significant changes was applied for an adjusted P-value <0.05 and LFC <–1 or >1 ($n = 4$ independent replicates). (**B**) Expression of selected genes from the GO cluster 'gastrulation', depicting the LFC of Tet1 KO and Tet1 CM relative to WT mESCs and Tet1 KO mESCs re-expressing TET1 or TET1 CM relative to Tet1 KO mESCs. (**C**) same analysis as in (B) depicted for genes in the GO term 'germ cell development' ($n = 3$ independent replicates). (**D**) Heat map of the hierarchical clustering of the RNA-seq expression z-scores and promoter DNA methylation in Tet1 KO mESCs significantly up- or downregulated genes. Promoter DNA methylation was assessed by enzymatic methylome sequencing (EM-seq, $n = 3$ independent replicates). Red bars indicate the delta DNA methylation (dmC, Tet1 KO – WT)) at the corresponding promoter. (**E**) Violin plots showing the percentage of methylated CpG dinucleotides globally, at promoters, enhancers, gene bodies and transposable elements (TE) in WT, Tet1 KO and Tet1 CM mESCs determined by EM-seq.

predominantly regulates gene expression independently of its catalytic activity with only a small subset of genes depending on promoter demethylation by TET1.

Loss of TET1 alters the chromatin modification landscape

To gain further insights into possible mechanisms by which TET1 regulates transcription independent of DNA demethylation, we asked if the loss of TET1 is accompanied by changes in the chromatin landscape. To this end, we compared the relative abundances of core histone modifications among Tet1 KO, Tet1 CM and WT mESCs using quantitative LC–MS/MS analysis. We observed a profound global reduction of H3K27me3, pH4Kac as well as H4K20me3 in Tet1 KO mESCs (Figure 2A, Supplementary Figure 3A). Conversely, the corresponding monomethylation states H3K27me1 and H4K20me1 were significantly, but to a lower extent increased in Tet1 KO mESCs (Figure 2A). We also detected significant, albeit less pronounced changes of several other histone modifications such as H3K18me1, H3K23me1, H3K9ac and H3K14ac in Tet1 KO mESCs (Figure 2A, Supplementary Figure 3A). Similar to the transcriptomics data, these profound changes in histone modification levels were only observed in Tet1 KO cells with the exception of H4K20me3, which exhibited a modest downregulation in Tet1 CM cells (KO = 49% and CM = 18% reduction compared to WT) (Figure 2A, Supplementary Figure 3A). Notably, we observed a significant downregulation of the *EZH2* transcript level. However, in total these global reductions in histone modification levels in Tet1 KO mESCs cannot be explained by transcriptional deregulation of the responsible histone modifying enzyme complexes (Supplementary Figure 3B). Taken together, these results demonstrate that TET1 predominantly regulates global H3K27me3, pH4Kac and H4K20me3 histone modification states via catalytic-independent mechanisms.

To investigate how loss of TET1 affects the genomic distributions of H3K27me3, pH4Kac and H4K20me3, we acquired genome-wide histone modification profiles using the quantitative ChIP-Seq method MINUTE-ChIP (50). MINUTE-ChIP uses a barcoding and pooling approach to enable quantitative comparisons between samples. This allowed us to profile quadruplicates of WT, Tet1 CM and KO mESCs in the same pool. The global readcount analysis from these MINUTE-ChIP experiments confirmed the global trends observed by mass spectrometry, with Tet1 KO mESCs exhibiting significantly reduced levels of H3K27me3, pH4Kac and H4K20me3 (Supplementary Figure 4). Of note, in contrast to the LC-MS/MS data, global H4K20me3 levels were unchanged in the MINUTE-ChIP data from Tet1 CM mESCs.

Next, we focused our analysis on the distribution of H3K27me3, H3K4me3, pH4Kac and H4K20me3 across selected genomic elements including active promoters, inactive promoters, enhancers, gene bodies of active and inactive genes, and TEs (Figure 2B). For H3K27me3, pH4Kac and H4K20me3, we detected a strong reduction over all analyzed genomic elements in Tet1 KO mESCs, but only minimal reductions in H3K4me3. In general, most histone modifications such as H3K4me3, H3K27me3 and H4K20me3 exhibit well-defined patterns of enrichment over distinct genomic elements in WT mESCs (44). In line with prior reports, H3K4me3 was found at enhancers, active genes, and mainly at active promoters and, as in histone LC–MS/MS measurements, changed only subtly in Tet1 KO and Tet1 CM mESCs (Figure 2A and B). H3K27me3 was mainly enriched at inactive promoters and within inactive gene bodies, but significantly reduced upon TET1 loss (Figure 2B). Furthermore, pH4Kac was enriched at enhancers, active promoters, and within active gene bodies. At all three elements we observed a significant reduction in Tet1 KO mESCs (Figure 2B). We also found H4K20me3 to be enriched over TEs, but significantly reduced in Tet1 KO mESCs (Figure 2B). Additionally, we performed a chromatin-state discovery and genome annotation analysis with ChromHMM to investigate the enrichment of H3K4me3, H3K27me3, H3K27me1, pH4Kac, H4K20me3 and H4K20me1 at defined chromatin states. Amongst many smaller alterations, we detected a pronounced loss of H3K27me3 at poised promoters and a strong reduction of H4K20me3 at H3K9-marked heterochromatin (Supplementary Figure 5).

Next, we wondered whether the reduction of histone marks at promoters correlates with changes in gene expression and DNA methylation observed in Tet1 KO mESCs. We compared H3K4me3, H3K27me3, pH4Kac and DNA methylation levels over genes down- or upregulated in Tet1 KO mESCs. To narrow our focus on direct targets of TET1, we used published ChIP-seq data to preselect for genes bound by TET1 (34). We observed in Tet1 KO mESCs a reduction of H3K27me3 at upregulated genes, whereas at downregulated genes changes in H3K27me3 were less prominent (Figure 2C). H3K4me3 levels were unchanged at upregulated genes, but were slightly decreased at downregulated genes (Figure 2C). Furthermore, we detected a strong loss of pH4Kac at downregulated genes in Tet1 KO mESCs and almost no change at upregulated genes (Figure 2C). We asked if the changes in histone modification levels at up- and downregulated genes correspond to DNA hyper- or hypomethylation. We observed DNA hypermethylation at up- and downregulated genes in Tet1 KO mESCs and a similar but smaller increase in DNA methylation in Tet1 CM mESCs (Figure 2C). At multiple gastrulation and germ cell development markers, the loss of specific histone modifications correlated with expression changes observed in Tet1 KO mESCs. For instance, we detected a pronounced loss of H3K27me3 but only minor changes in H3K4me3 at the genomic locus of the upregulated gastrulation marker *Wnt3* in Tet1 KO mESCs. In contrast, the downregulated germ cell development marker *Zfp42* exhibited a clear loss of pH4Kac only in the Tet1 KO mESCs (Supplementary Figure 6A). In both cases we observed an increase of DNA methylation in Tet1 KO mESC at the promoter region and gene body (Supplementary Figure 6A). In summary, our data shows that the transcriptional deregulation observed in Tet1 KO mESCs cannot be attributed to changes in DNA methylation but rather global perturbation of histone modifications.

Since H4K20me3 was mainly enriched over TEs, we next analyzed whether H4K20me3 was specifically lost at distinct TE families in Tet1 KO mESCs. We detected a major

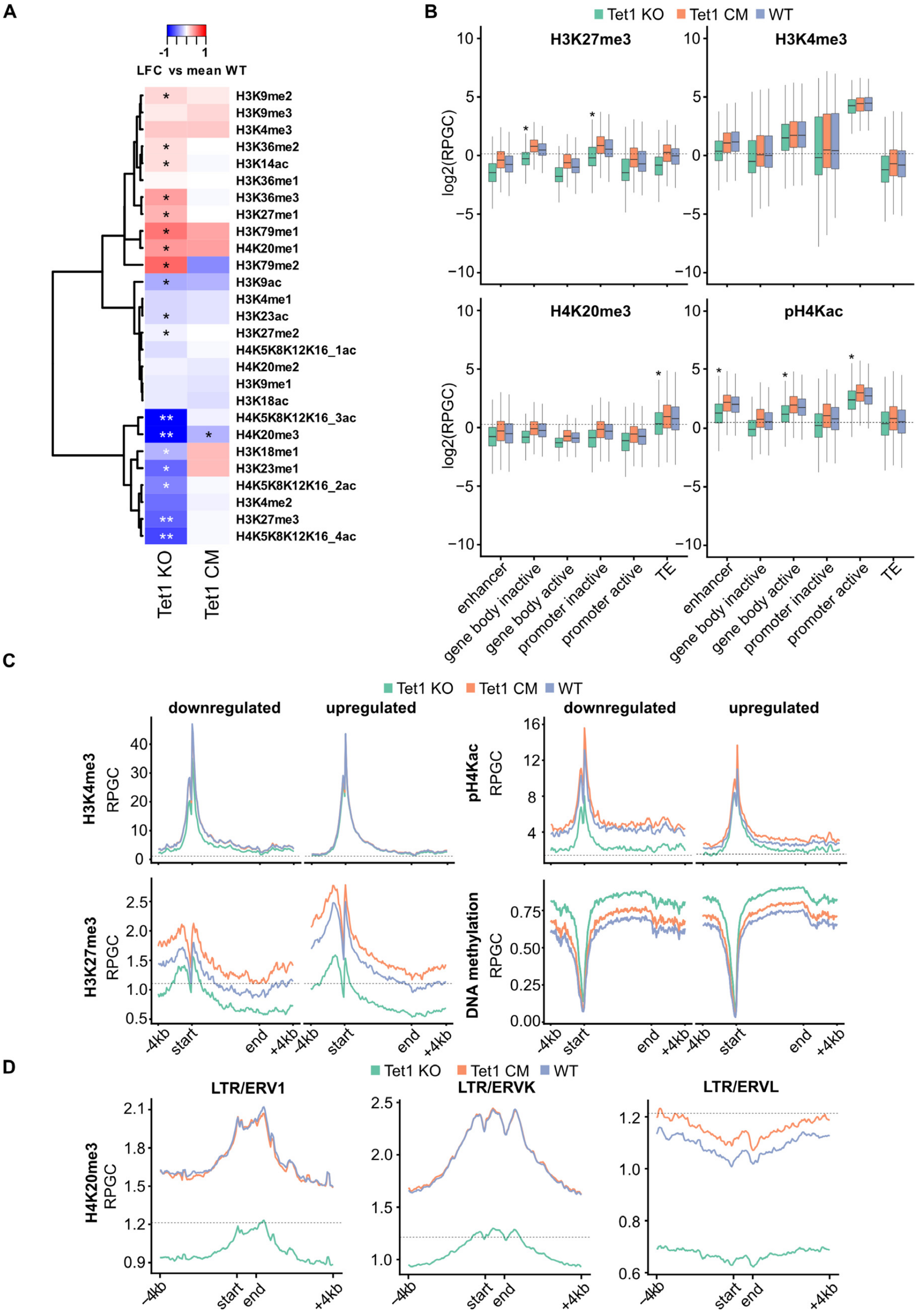


Figure 2. Tet1 KO mESCs display a reduction in histone marks. (**A**) Heatmap depicting hierarchical clustering of individual histone post-translational modification abundances. Calculated is the $\log_2$-fold change (LFC) relative to the mean abundances in WT mESCs. LC-MS/MS quantification of Tet1 KO, Tet1 CM, and WT mESCs ($n = 3$ independent replicates). Each row represents distinct histone modification states and the color gradient indicates the LFC. Significant changes (adjusted P-value < 0.05 and < 0.01) in the Tet1 KO and CM relative to WT mESCs are marked with * and **, respectively. (B–D) The y-axis indicates reads per genomic content (RPGC). The dotted line indicates the genome average RPGC of the respective histone signal ($n = 4$ independent replicates). (**B**) Average quantitative MINUTE-ChIP signal displayed as boxplots of H3K27me3, H3K4me3, H4K20me3 and pH4Kac comparing Tet1 KO, Tet1 CM and WT at enhancer, gene body inactive/active, promoter inactive/active and transposable element (TE). Significant changes were marked with * (one-sided *t*-test, adjusted P-value <0.05, Tet1 KO relative to WT mESCs), see Supplementary Table 3 for a full list. Horizontal black lines within boxes represent median values, boxes indicate the lower and upper quartiles, and whiskers indicate the 1.5 interquartile range. (**C**) Average quantitative MINUTE-ChIP profiles of H3K4me3, H3K27me3, and pH4Kac and DNA methylation levels using EM-seq data in Tet1 KO, Tet1 CM and WT mESCs across gene bodies significantly down- or upregulated in Tet1 KO mESCs. Up- and downregulated genes were preselected for TET1 binding in WT mESCs. TET1 binding sites were identified using published ChIP-seq data of wild-type mESC cultured with the same medium conditions (34). (**D**) Average quantitative MINUTE-ChIP profiles of H4K20me3 across ERV1, ERVK and ERVL elements ±4 kb in Tet1 KO, Tet1 CM and WT mESCs.

loss of H4K20me3 at ERV1 and ERVK elements (Figure 2D, Supplementary Figure 6B). Additionally, we detected a less pronounced loss at L1 and ERVL elements (Figure 2D, Supplementary Figure 6B). TET1 seems to mainly regulate H4K20me3 levels at ERV1 and ERVK elements, raising the intriguing question how TET1 is involved in heterochromatin formation at these genetic elements. Collectively, we identify the non-catalytic role of TET1 to be a global regulator of H3K27me3, pH4Kac and H4K20me3 levels.

TET1 associates with different chromatin modifiers and regulates ERV expression

We next asked if the dramatic drop in H4K20me3 at ERVs also correlates with changes in TE expression. In contrast to Tet1 CM mESCs, we identified in our RNA-seq data of Tet1 KO mESCs multiple TEs that were significantly upregulated (Figure 3A). In line with their loss of H4K20me3, we observed the strongest upregulation at ERV1 and ERVK elements (Figure 3A and B, Supplementary Figure 6B, 7). However, ERVL elements exhibited the greatest number of significantly upregulated ERVs in Tet1 KO mESCs ($n = 695$), compared to ERV1 ($n = 522$) and ERVK ($n = 50$) (Figure 3A). Furthermore, expression of exogenous TET1 or TET1 CM was able to reverse the ERV upregulation in Tet1 KO mESCs (Figure 3B), suggesting that ERVs are regulated independently of TET1's catalytic activity.

Previously, TET1 binding was shown to strongly correlate with CpG density (32–34). In line with this observation, we found that ERV1 and ERVK elements in particular displayed a higher CpG density than expected by their GC content (Supplementary Figure 6C) and ERV elements with a higher observed over expected (O/E) CpG ratio were also more likely to be upregulated in Tet1 KO mESCs (Supplementary Figure 6C and D). Furthermore, using published TET1 ChIP-seq data (34) we found that TET1 was enriched at ERV1, ERVK and ERVL elements (Figure 3C). At the same time, all three ERV classes were hypermethylated in Tet1 KO mESCs. In the Tet1 CM mESC the increase in DNA methylation was significant, but less pronounced compared to Tet1 KO mESCs (Figure 3C). Our finding that DNA methylation is not sufficient to silence ERV elements in mESCs is in line with previous studies (83,84). Taken together, these findings suggest that TET1 binds ERV1, ERVK and ERVL elements due to their high CpG density and facilitates a repressive mechanism which is independent of DNA methylation and involves H4K20me3.

The repression of TEs, especially of ERVs, relies on the cooperation of several epigenetic pathways. In particular, the establishment and maintenance of H3K9me3 is crucial for ERV1 and ERVK silencing (83). However, we did not detect a global loss of H3K9me3 in our histone LC–MS/MS measurements in Tet1 KO mESCs (Figure 2A). To investigate if H3K9me3 is specifically lost at ERVs in Tet1 KO mESCs, we exploited our quantitative MINUTE-ChIP approach. In accordance with the LC–MS/MS data, we did not observe a global reduction of H3K9me3 in Tet1 KO mESCs using quantitative ChIP-seq (Supplementary Figure 4 and 6E). Moreover, when all TEs were assessed as a single group, H3K9me3 levels appeared to be essentially unchanged in Tet1 KO mESCs. However, a more detailed analysis of individual TE families revealed a significant drop of H3K9me3 at ERV1, ERVK and ERVL in Tet1 KO mESCs (Figure 3C). We found that in Tet1 KO mESCs at specific ERV elements the loss of H3K9me3 and H4K20me3 cooccurs with an increase in DNA methylation and an upregulation of ERV elements (Figure 4). ERVL transcriptional activation correlates with the expression of the 2C marker *Zscan4* (85). In Tet1 KO mESCs, we detected a significant upregulation of the ERVL elements *MERVL-int* and *MT2_Mm* and the *Zscan4* cluster (Figure 3B, Supplementary Figure 6F). Interestingly, the activation of ERVL and *Zscan4* was significantly stronger in Tet1 KO compared to Tet1 CM mESCs and we detected a significant loss of both H3K9me3 and H4K20me3 at *MERVL-in*t and *MT2_Mm* (Supplementary Figure 7 and 8). In addition, we could rescue the *MERVL-int*, *MT2_Mm*, and *Zscan4* expression by reintroducing TET1 and TET1CM (Figure 3B and Supplementary Figure 6F). Previously, TET-mediated DNA demethylation was reported to regulate ERVL and *Zscan4* expression (23,25). In contrast, our data indicates a predominant non-catalytic role of TET1. Collectively, these findings describe a novel role of TET1 in ERV silencing independent of DNA demethylation. We demonstrate for the first time that TET1 is critical for H3K9me3 and H4K20me3 deposition and silencing of ERV1 and ERVK.

The interplay between TET1 and SIN3A is crucial for ERV repression

Next, we aimed to investigate the underlying mechanism that regulates TET1-dependent silencing of ERV1, ERVK, and ERVL elements. Since we found that deposition of H3K9me3 and H4K20me3 is dependent on non-catalytic activities of TET1, we performed ChIP-MS on TET1 to identify interaction partners potentially involved in this process. Using this strategy, we identified a large number of different chromatin modifiers associating with TET1 (Figure 5A). In line with previous studies, we detected the core PRC2 complex (EED, SUZ12, EZH2) and many subunits of the SIN3A/HDAC complex (31,32). Strikingly, we also identified heterochromatin protein 1 (HP1) beta (HP1β, also known as CBX1), MORC3 and SMARCAD1 to be significantly enriched, and TRIM28 as well as HP1 gamma (HP1γ also known as CBX3) just below significance threshold (Figure 5A). Interestingly, these proteins were found to be associated with the formation of H3K9me3-marked heterochromatin in particular at ERVs (83,86–88).

A well-established pathway in ERV silencing is the binding of HP1 proteins to H3K9me3, recruiting SUV39H and SUV4-20H, and the subsequent spreading of H3K9me3 and H4K20me3 (89). To investigate if the ERV-specific loss of H3K9me3 might impact HP1β localization, we used immunofluorescence to examine the distribution of HP1 in Tet1 KO, Tet1 CM and WT mESCs. Intriguingly, HP1β became depleted from heterochromatic foci, i.e. chromocenters and exhibited an overall more homogenous distribution in the nucleus upon loss of TET1 protein but not upon loss of TET1 catalytic activity (Figure 5B). At the same time we observed only a minor reduction of HP1β at the transcript level and no obvious change on the protein level in Tet1 KO mESCs (Supplementary Figure 9A). To quantify our obser-

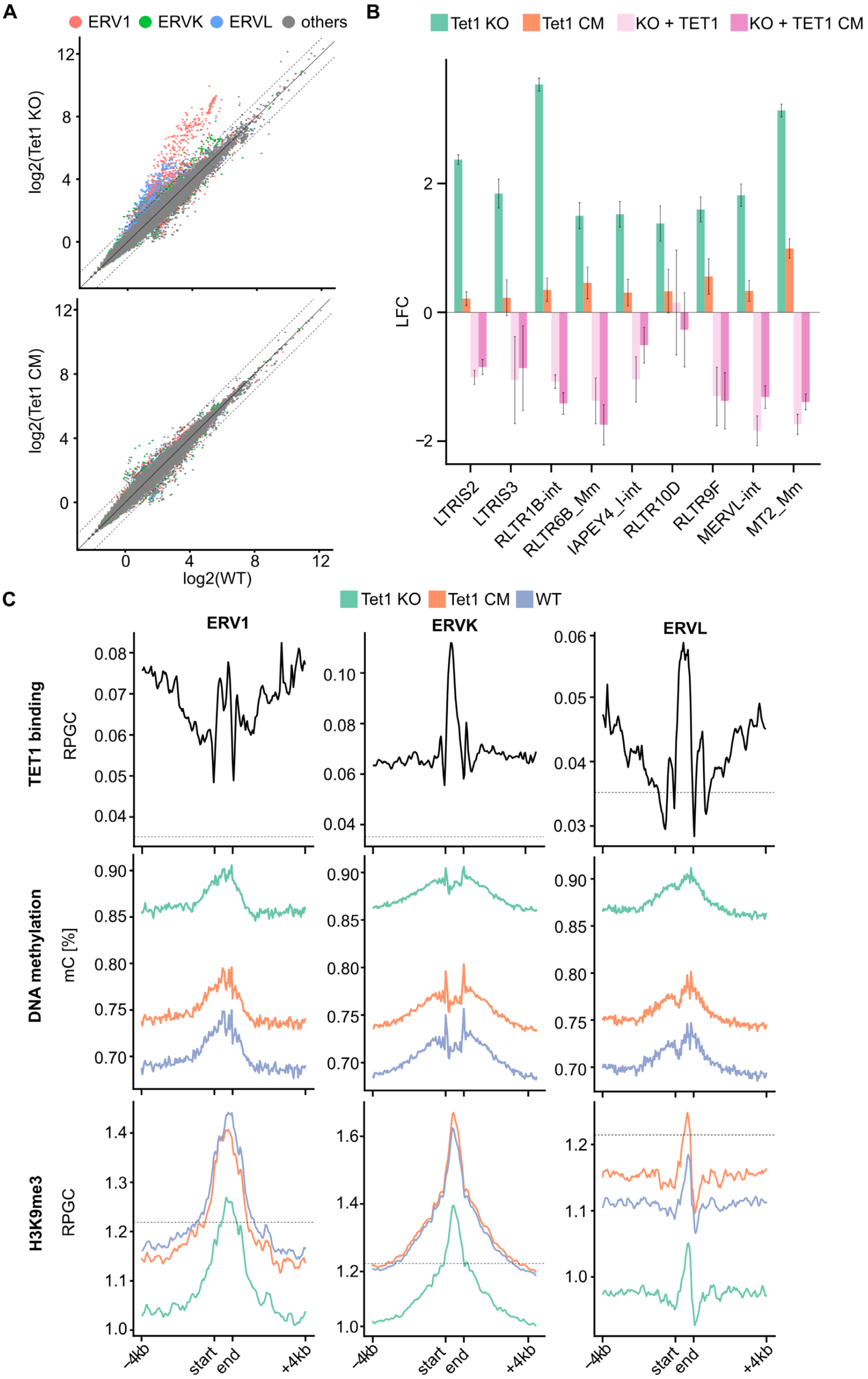


Figure 3. TET1 regulates H3K9me3 deposition and ERV silencing. (**A**) Scatter plot depicting log2 transformed counts of single TEs (transposable elements) comparing Tet1 KO versus WT and Tet1 CM versus WT. Red dots: ERV1, green dots: ERVK, blue dots: ERVL and grey dots: other TEs. Significantly upregulated ERV elements in Tet1 KO mESCs: ERV1 (*n* = 522), ERVK (*n* = 50), ERVL (*n* = 695). (**B**) LFC of differentially expressed ERVs in Tet1 KO relative to WT mESCs. Comparing ERV expression in Tet1 KO and Tet1 CM relative to WT mESCs and ERV expression when re-expressing TET1 or TET1 CM in Tet1 KO relative to Tet1 KO mESCs. LFC = $\log_2$ fold change (*n* = 3 independent replicates). (**C**) Average quantitative MINUTE-ChIP profiles of H3K9me3, ChIP profile of TET1 binding using published ChIP-seq data of mESC cultured under the same medium conditions (34) and percentage of DNA methylation using EM-seq data across ERV1, ERVK and ERVL elements ±4 kb (kilo base) comparing Tet1 KO, Tet1 CM and WT. For ChIP the y-axis shows reads per genomic content (RPGC). The dotted line indicates the genome average RPGC of the respective histone signal (*n* = 4 independent replicates).

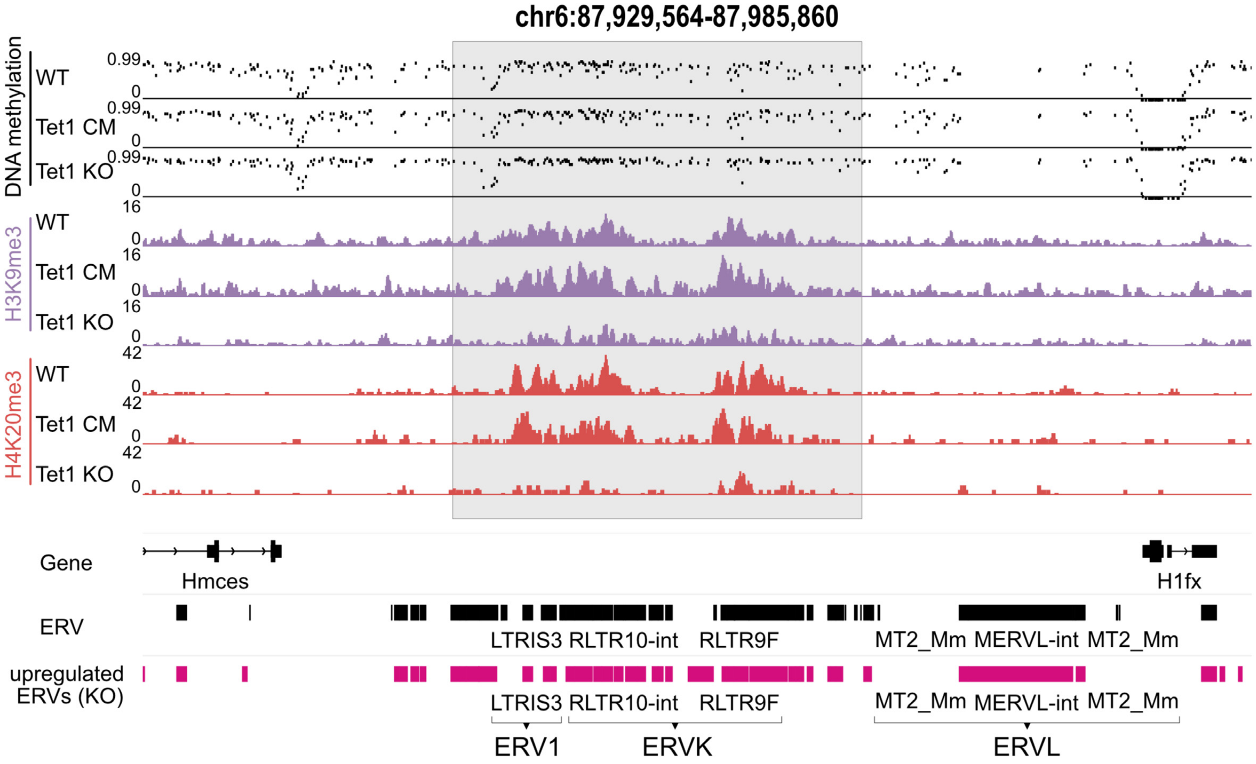


Figure 4. DNA methylation, H3K9me3 and H4K20me3 at upregulated ERVs in Tet1 KO mESCs. Representative genome browser tracks of EM-seq data, H3K9me3 and H4K20me3 ChIP in WT, Tet1 KO and Tet1 CM. Pink bars indicated the log fold change of ERVs upregulated in Tet1 KO cells. Individual upregulated ERVs are named and classified in ERV1, ERVK and ERVL. Regions with a gain of DNA methylation and loss of H3K9me3 and H4K20me3 in Tet1 KO cells are marked in grey.

vation, we performed high-throughput microscopy and calculated the coefficient of variation (CV) of the HP1β signal, commonly used as a benchmark for fluorescence signal distribution (90,91). High CV values correspond to a heterogenous and lower CV values to a more homogenous signal distribution. While the HP1β signal in WT and Tet1 CM mESCs displayed similar CV values, we observed significantly lower CV values for HP1β in Tet1 KO mESCs (Figure 5B). In addition to HP1β, mammals possess two other paralogs of HP1, namely, HP1α and HP1γ. All three have overlapping, but distinct functions in heterochromatin formation (92,93). Therefore, we also analyzed the CV values of HP1α and HP1γ under the same conditions as for HP1β in WT, Tet1 KO and Tet1 CM mESCs. Compared with HP1β, the distribution of HP1α exhibited a more limited but still significant reduction in focal heterochromatin accumulation in Tet1 KO mESCs (Supplementary Figure 9B). In the case of HP1γ, the extent of this reduction in heterogeneity was even more severe in Tet1 KO mESCs (Supplementary Figure 9C). Although not as dramatic as Tet1 KO mESCs, we also observed significant decreases in the focal patterning of both HP1α and HP1γ in Tet1 CM mESCs (Supplementary Figure 9B, C). In summary, our data indicates that TET1 associates with heterochromatin proteins and might be a regulator of HP1 formation at heterochromatic regions.

It is well accepted that the turnover of histone acetylation is crucial for heterochromatin formation (94–97). Since we and others have found the SIN3A/HDAC complex to be among the most abundant interactors of TET1 (32,98,99), we investigated whether the TET1-SIN3A/HDAC interaction is involved in TET-mediated regulation of ERVs. To this end, we first assessed whether SIN3A occupies the same ERVs as TET1. We identified a considerable overlap between TET1 and SIN3A bound ERVs, many of which were also found to be upregulated in Tet1 KO mESCs (Figure 5C, Supplementary Figure 9D). Next, we asked whether the TET1 and SIN3A interaction is critical for the transcriptional regulation of these ERVs. To answer this question, we expressed a version of TET1 harbouring a mutation described to disrupt the interaction with SIN3A (TET1 SIN3A mut.) in Tet1 KO mESCs (Supplementary Figure 9E) (47). The two amino acids (L897 and L900) critical for the SIN3A interaction are not part of the catalytic domain of TET1. Intriguingly, ERV repression was restored by WT TET1, but not the TET1 SIN3A mut. (Figure 5D). Of note, we also identified a subset of genes where the TET1 SIN3A mut. rescues gene expression (e.g. *Esrrb*, *Lefty* and *Pvalb*), suggesting additional pathways independent of SIN3A (Supplementary Figure 9F).

Finally, we asked whether reexpresing TET1 can restore HP1β localization. After re-expressing TET1, TET1CM and the TET1 SIN3A mut., we selected TET1 positive mESCs using a TET1 antibody staining and calculated the CV of the HP1β signal for WT, Tet1 KO and the three rescue cell lines. The re-expression of TET1, TET1CM and

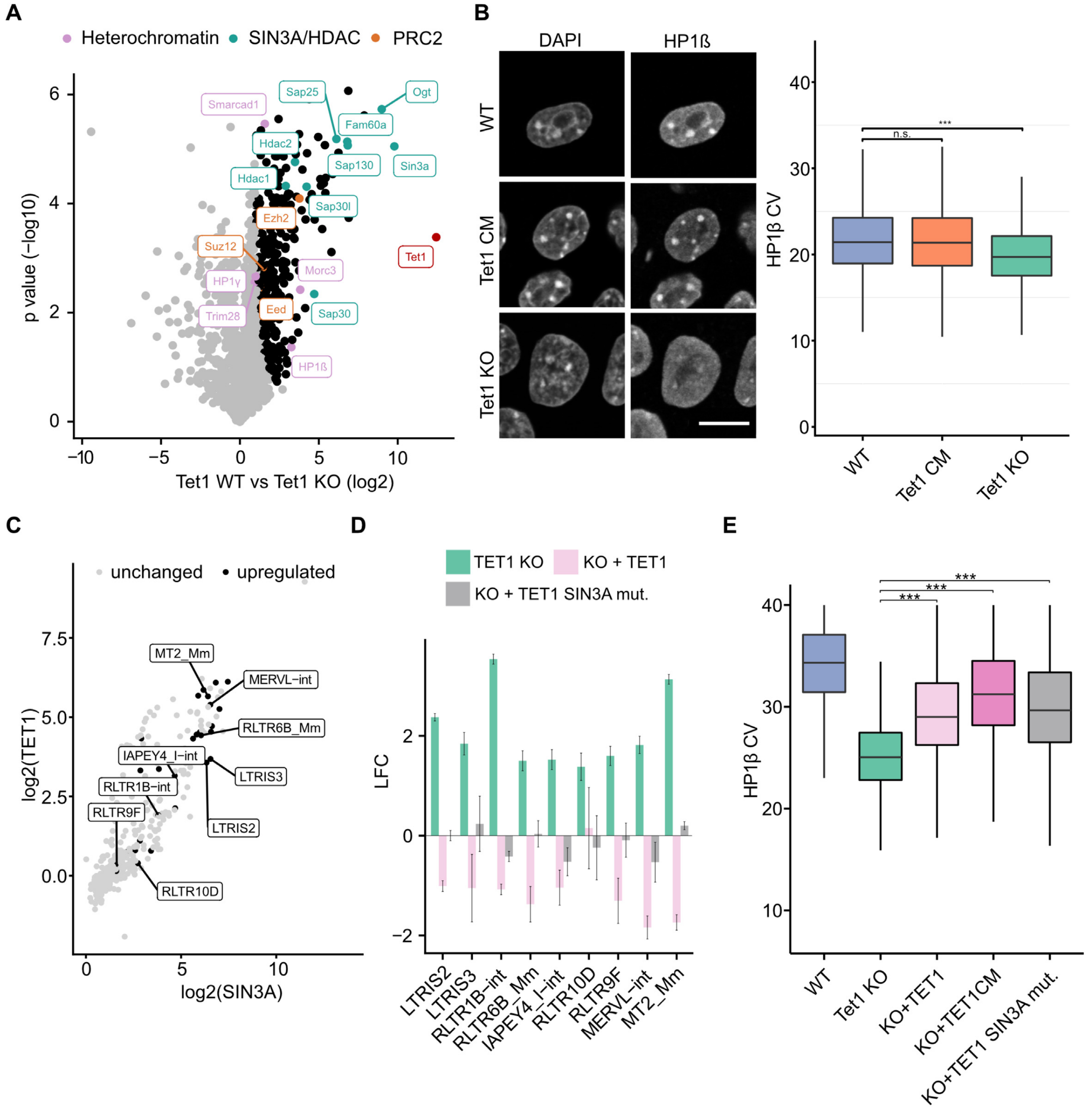


Figure 5. The TET1-SIN3A interaction is crucial for ERV regulation. (**A**) Volcano plot of TET1 ChIP-MS experiment in WT and Tet1 KO mESCs ($n = 3$ independent replicates). Black dots: significantly enriched after TET1 pulldown. Purple dots: Proteins associated with heterochromatin formation. Turquoise dots: Members of the SIN3A/HDAC complex. Orange dots: Core complex members of PRC2. Statistical significance is determined by performing a Student's t-test with a permutation-based false discovery rate of 0.05 and a cutoff of >1 of $\log_2$ transformed fold change. (**B**) Left: Immunofluorescence images of WT, Tet1 CM and Tet1 KO mESC stained for DAPI and HP1β. Scale bar = 10 μm. Images were taken using a confocal microscope. Right: Boxplots showing the coefficient of variation (CV) calculated from HP1β signal intensities, comparing WT ($n = 27\,588$), Tet1 CM ($n = 40\,160$), and Tet1 KO ($n = 25\,882$). Images were taken using an Operetta microscope. ANOVA + Tukey's honestly significant difference post-hoc test: $^{****}P < 0.0001$. (**C**) Scatter plot comparing log2 transformed fold change enrichment of TET1 and SIN3A at transposable elements (TE) using published ChIP-seq data from mESCs cultured under the same conditions (34,55). Gray dots: unchanged expression of TE in Tet1 KO relative to WT mESCs. Black dots: upregulated TE in Tet1 KO relative to WT mESCs. (**D**) Expression of differentially expressed ERVs in Tet1 KO relative to WT mESCs as $\log_2$ transformed fold changes. Comparing ERV expression of Tet1 KO relative to WT mESCs and re-expressing TET1 or TET1 SIN3A mut. in Tet1 KO mESCs relative to Tet1 KO mESCs. (**E**) Boxplots depicting the coefficient of variation (CV) calculated from HP1β signal intensities comparing WT ($n = 4617$), Tet1 KO ($n = 9334$), KO + TET1 ($n = 3757$), KO + TET1CM ($n = 1136$) and KO + TET1 SIN3A mut. ($n = 1885$) TET1 and TET1 SIN3A mut. negative and positive cells. For the TET1 rescue cell lines, TET1 staining was used to select for TET1 positive (signal intensity > 1000) mESCs before the CV was calculated. ANOVA + Tukey's honestly significant difference post-hoc test: $^{****}P < 0.0001$. Horizontal black lines within boxes represent median values, boxes indicate the lower and upper quartiles, and whiskers indicate the 1.5 interquartile range. Representative confocal images of HP1β and TET1 stainings (Supplementary Figure 10).

TET1 SIN3A mut. restored the HP1β localization to heterochromatic regions (Figure 5E, Supplementary Figure 10). Interestingly, the TET1 SIN3A mut. efficiently restored HP1β localization, but in contrast to WT TET1 did not silence ERV expression (Figure 5D). These findings are in line with the observation that HP1 proteins alone are not sufficient to silence ERVs in mESCs (100) and might suggest that the TET1-SIN3A mut. can still directly recruit HP1β to heterochromatin, but not silence ERV expression without SIN3A deacetylation activity. Deacetylation of the H3 tail is crucial for H3K9 methylation efficiency by SETDB1 (101). To investigate if H3K9ac, SETDB1, TET1 and SIN3A correlate at ERV1, ERVK and ERVL elements we used published ChIP-seq data (34,56,57) and our MINUTE-ChIP data of H3K9me3, H4K20me3, pH4Kac, H3K4me3 and H3K27me3. Interestingly, we found that H3K9ac, SETDB1, SIN3A and TET1 occupancy were highly correlated at ERV elements (Figure 6A). On the contrary, at all other TEs excluding ERVs, TET1 and SIN3A binding were not associated with SETDB1 and H3K9ac occupancy (Figure 6A). This might suggest that TET1-SIN3A are involved in deacetylation and the subsequent methylation of H3K9 via SETDB1 to control repression specifically of ERV elements. In summary, we identified TET1 as a key regulator of ERV expression in mESCs. Furthermore, our findings suggest that SIN3A is important for DNA demethylation independent regulation of ERVs by TET1.

DISCUSSION

Whereas the role of TET1 in active DNA demethylation is well described (102), the non-catalytic functions of TET1 remain unclear. In contrast to earlier studies suggesting that TET1 KO mice are viable (18,22), a recent study reported that TET1 KO mice display severe gastrulation defects and are not viable after E9.5 (23). These discrepancies can be assigned to differences in the *Tet1* knockout targeting strategy. The viability of some *Tet1* KO strains seems to be the consequence of a hypomorphic deletion, which allows an N-terminal fragment of TET1 to be expressed. Importantly, this fragment does not contain the catalytic domain of TET1, suggesting TET1 to have key non-catalytic functions (23). Here, we aimed to systematically decipher those DNA demethylation independent functions of TET1 in mESCs.

In agreement with the current literature, our transcriptomics analysis revealed a deregulation of pluripotency and gastrulation markers in Tet1 KO mESCs (22,23,103,104). Interestingly, our rescue experiments, DNA methylation analysis and systematic comparison of Tet1 KO and Tet1 CM mESCs showed that the transcriptional changes can mainly be attributed to the non-catalytic functions of TET1. These findings are supported by a number of previous studies, suggesting a non-catalytic role of TET1 in mESCs, reprogramming or thermogenesis (23,27,32,105). While this manuscript was in the review process, another study demonstrated that TET1 regulates H3K27me3 in mESCs independent of its catalytic activity (106). In line with our observations, Chrysanthou *et al.* showed that TET1 regulates developmental genes together with PRC2 and SIN3A independent of its DNA demethylation activity. Further, non-catalytic functions of TET1 are critical for early development, while the catalytic functions gain importance in late gestation and postnatal development (106). To note, our RNA-seq and rescue data also shows some minor transcriptional effects in Tet1 CM mESCs at genes significantly deregulated in Tet1 KO mESCs. Therefore, in some cases the catalytic and non-catalytic functions of TET1 might cooperate to regulate transcription. These findings and several other studies highlight the relevance of TET1-dependent active DNA demethylation in different biological systems (102,107–110). Together, suggesting that TET1 catalytic functions are highly context-dependent. TET1 is also important for recruiting TET2 to chromatin (37). In mESCs and different biological settings, TET2 might be partially compensating for the catalytically inactive TET1. Those compensatory effects of TET2 or the blocking of CpG sites by the presence of catalytic inactive TET1 could explain the less pronounced hypermethylation in Tet1 CM mESCs observed in our EM-seq data. In line with this hypothesis, we and others recently proposed that TET1 and TET2 have coordinated roles in DNA demethylation (111). While active DNA demethylation in mESCs seems to have few transcriptional effects, TET catalytic functions in DNA demethylation or the oxidative derivatives 5hmC, 5fC and 5caC themselves are important during differentiation, gastrulation and in somatic cells. This hypothesis is supported by previous findings, demonstrating that TET-dependent active DNA demethylation at promoters of lineage factors is critical for their activation during lineage commitment, gastrulation and reprogramming (20,77,112).

The predominant non-catalytic role of TET1 in mESCs prompted us to further study the underlying mechanisms of TET1 regulating transcription. TET1 was previously shown to associate with the chromatin modifying complexes PRC2, SIN3A/HDAC, OGT, MBD3/NURD and MOF (31,32,35–37,98,99). Here, we used a LC–MS/MS approach to identify the global interplay of TET1 with different histone modifications. Whereas loss of TET1 was reported to result in a reduction of H3K27me3 at promoters (34,36), our data reveals a genome-wide reduction of this mark independent of TET1 catalytic activity. Additionally, we identified a global reduction of H4K20me3 as well as pH4Kac only in Tet1 KO and not in Tet1 CM mESCs. It has been suggested that TET1-dependent DNA demethylation facilitates other chromatin modifiers to bind and restructure chromatin in order to activate or repress transcription. In contrast, our data suggest that in mESCs active DNA demethylation by TET1 is not required for the proper regulation of chromatin states, as we did not detect global alterations of histone modifications in Tet1 CM. Further, changes in DNA methylation did in most cases not correlate with the observed gene expression and histone modifications changes, suggesting a DNA methylation independent mechanism in mESCs. Alternatively, TET1 might act as an interaction hub for chromatin modifiers and/or is important for the composition of different regulatory chromatin complexes.

Our data identifies TET1 as a novel interactor of the heterochromatin machinery and a regulator of ERV elements. We show that ERV1, ERVK and ERVL lose H3K9me3 and H4K20me3 in Tet1 KO mESCs. Furthermore, we find that

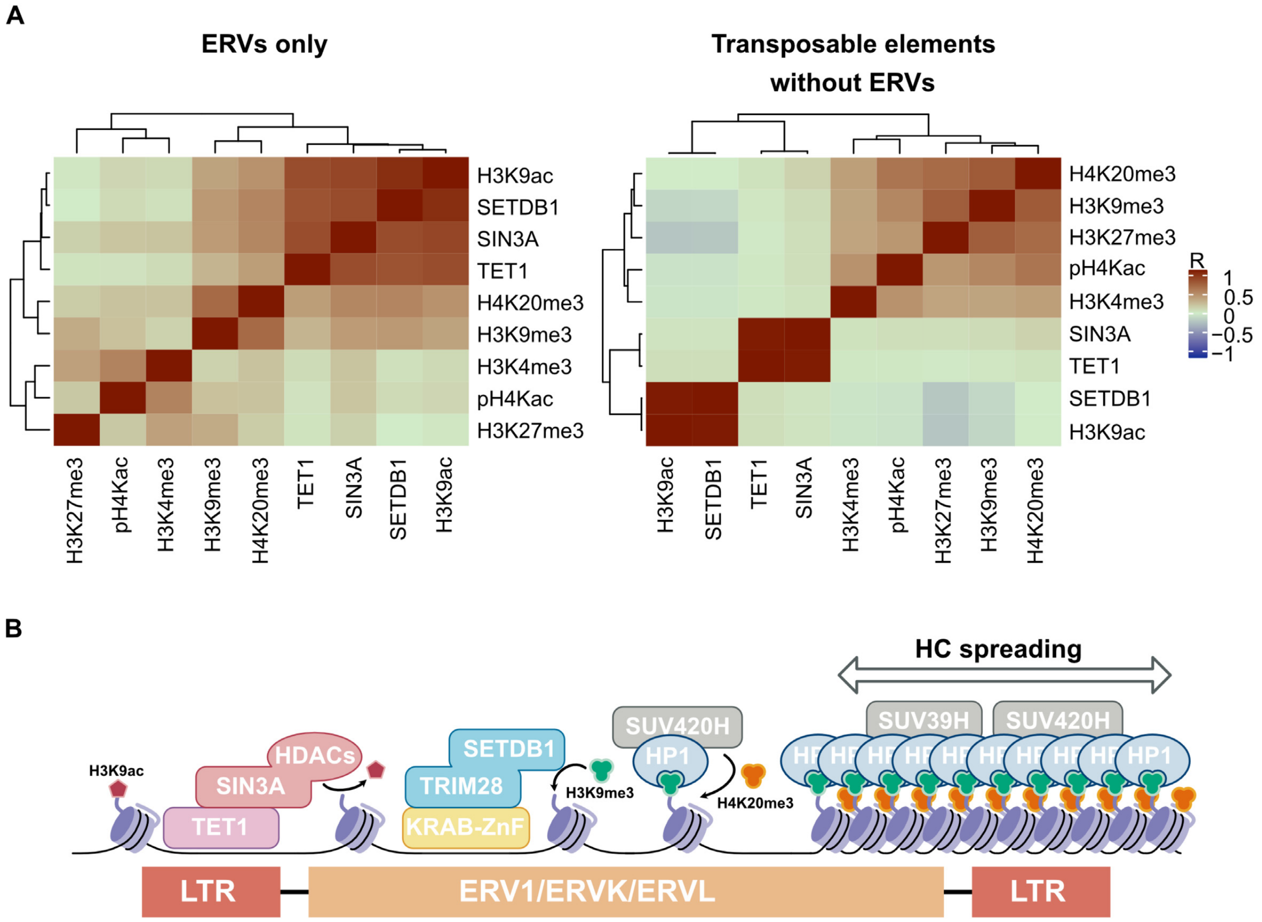


Figure 6. TET1-SIN3A/HDAC-mediated acetylation turnover might regulate H3K9me3/H4K20me3-mediated silencing of ERVs in mESCs. (**A**) Correlation matrix of ChIP-seq data of H3K9ac, SETDB1, SIN3A, TET1, H3K9me3, H4K20me3, pH4Kac, H3K4me3, and H3K27me3 at individual copies of only ERV elements ($n = 258\ 668$) or at individual copies of transposable elements (TEs) excluding ERVs (n = 757,079). The correlation coefficient (R) is indicated by a color gradient. (**B**) Model figure illustrating the proposed TET1-SIN3A/HDAC-mediated ERV1, ERVK and ERVL silencing mechanism. TET1 recruits the SIN3A/HDAC complex to ERV1, ERVK and ERVL elements. SIN3A/HDAC-mediated deacetylation of H3K9ac facilitates the recruitment of the KRAB-ZnF/TRIM28/SETDB1 silencing complex and the subsequent installation of the heterochromatin mark H3K9me3. HP1 proteins bind H3K9me3, recruit SUV39H and SUV4-20H for the establishment of H3K9me3 and H4K20me3 domains, ultimately causing heterochromatin (HC) spreading.

TET1 associates with different proteins involved in heterochromatin formation. SMARCAD1 is a chromatin remodeler and was recently shown to regulate IAP elements (86), however we only observed a minor upregulation of most IAPs in Tet1 KO mESCs. Only recently, MORC3 was identified as a regulator of ERV elements and H3K9me3 (87). Among others, MORC3 regulates the LTRIS family, which we found significantly upregulated in Tet1 KO mESCs. To this end, future studies will be important to dissect a potential TET1-MORC3 interplay in ERV silencing.

In general, only little is known about the role of TET1 in ERV silencing. Previously, TET enzymes were proposed to regulate ERVL LTRs (25). ERVL expression is related to *Zscan4* expression and other markers of the 2 cell (2C) state (85). Interestingly, the *Zscan4* cluster was reported to be regulated by DNA demethylation (23). In contrast, our data indicates a more prominent upregulation in Tet1 KO mESC than in Tet1 CM mESCs. Furthermore, we could rescue the 2C markers when reexpressing TET1 CM in Tet1 KO mESCs. These findings indicate that *Zscan4* and MERVL regulation depend on both DNA demethylation and non-catalytic functions of TET1. In general, the finding that 2C markers are upregulated is contradictory to the concurrent upregulation of differentiation markers in Tet1 KO mESC. Serum LIF cultured mESCs exhibit a heterogeneous cell population and are known to include 2C-like cells (113). One explanation could be that the loss of TET1, besides mainly priming cells for differentiation, also promotes the expansion of the 2C-like cell subpopulation in Serum LIF mESC cultures.

Despite hypermethylation at ERVs in Tet1 KO mESCs, we could rescue normal ERV repression when reintroducing either TET1 or TET1 CM. Our finding that TET1 regulates ERV expression independently of its DNA demethylation function is in line with the observation that ERV silencing mediated by TRIM28 and SETDB1 is DNA methylation independent (83,84,114,115). In addition, non-LTR containing LINE1 elements are repressed independent of DNA methylation turnover, but by SIN3A in a TET1-dependent manner (38). To note, TRIM28/SETDB1 can also act synergistically with DNA methylation to silence IAP elements (88). One possible explanation for the simultaneous hypermethylation and activation of ERVs in Tet1 KO mESCs could be that 5mC-insensitive transcription factors are able to engage ERVs in the absence of TET1 (116).

Using immunofluorescence, we demonstrate for the first time that loss of TET1 leads to a displacement of HP1β, HP1γ, and HP1α from heterochromatin foci. Our data does not show that HP1 proteins are lost at ERVs in Tet1 KO mESCs. However, the loss of H3K9me3 and H4K20me3 at ERVs could explain the displacement of HP1 proteins from heterochromatic regions, prompting the question how TET1 influences the maintenance of heterochromatin in mESCs. The current model of heterochromatin formation proposes that site specific KRAB-Znf transcription factors recruit TRIM28 and its interaction partner SETDB1 to DNA. The latter installs H3K9me3, which is bound by HP1 and subsequently recruits SUV39H and SUV4-20H for spreading of H3K9me3 and H4K20me3 (89,117). We cannot completely rule out an indirect effect causing the loss of H3K9me3 and H4K20me3 in Tet1 KO mESC. However, our rescue experiments and quantitative ChIP data showing ERV silencing upon TET1 expression together with a specific loss of H3K9me3 at ERVs suggest that TET1 acts upstream of SETDB1. The loss of H3K9me3 in Tet1 KO mESCs could explain the delocalization of HP1β. Our interaction data and HP1β rescue experiments suggest that TET1 might also directly interact with HP1β independently of SIN3A and recruit HP1β directly to specific ERVs without inducing repression. This hypothesis would be in line with the finding that the deletion of HP1α, β, or γ alone does not lead to deregulation of ERV1 and ERVK, showing that TRIM28/SETDB1-mediated H3K9me3 deposition is sufficient for ERV silencing (100).

It is important to note that deacetylation and heterochromatin establishment are tightly connected (94,101,118–122). Furthermore, deacetylation of the H3 tail by SIN3A/HDAC is necessary for transcriptional repression and the loss of SIN3A causes a delocalization of HP1α (123) (101,121,124). Intriguingly, our TET1 ChIP-MS data identified a large number of the SIN3A/HDAC complex members as interactors. TET1 might be important for SIN3A/HDAC recruitment or complex composition, as SIN3A lacks any DNA-binding activity (125). Additionally, our rescue data strongly suggests that TET1 regulates ERVs in a SIN3A-dependent manner. Correlating binding of SETDB1, SIN3A, and TET1 and levels of H3K9ac, H3K9me3, and H4K20me3 revealed an overlap at ERV1, ERVK and ERVL elements, but not at other groups of TEs (Figure 6A). Therefore, we propose that the TET1-SIN3A/HDAC axis is crucial to control the constant acetylation turnover at ERV1, ERVK and ERVL, enabling the repression and installation of H3K9me3/H4K20me3 by TRIM28-SETDB1 (Figure 6B). We suggest that in mESC loss of TET1 interferes with correct placement and function of the SIN3A/HDAC complex at ERV elements. Subsequent accumulation of H3K9ac could interfere with TRIM28 or SETDB1 recruitment, resulting in a reduction of H3K9me3 at ERV1, ERVK and ERVL elements, displacement of HP1, and following loss of H4K20me3 (Figure 6B). It will be intriguing to further decipher the details of the underlying mechanism in the future.

Collectively, our results demonstrate that TET1 regulates gene expression independently of active DNA demethylation in mESCs. We provide novel insights into the mechanisms underlying TET1's non-catalytic functions in transcriptional regulation, including identifying TET1 as a global regulator of histone modifications. Moreover, we show that TET1 associates with different proteins involved in heterochromatin formation to suppress the expression of ERV1, ERVK and ERVL elements. Finally, we provide evidence that the mechanism of TET1-mediated silencing of ERV1, ERVK and ERVL elements critically depends on the interaction between TET1 and SIN3A but not the catalytic activity of TET1. Our study reveals the importance of disentangling the non-catalytic and catalytic roles of TET enzymes in different biological contexts. This will be of particular relevance for furthering our understanding of *Tet* mutations and their molecular consequences in cancer and disease.

DATA AVAILABILITY

EM-seq and RNA-seq data generated in this study is available at https://www.ebi.ac.uk/arrayexpress/ via the accession numbers E-MTAB-10933 and E-MTAB-10937, respectively. ChIP-seq data is available under the accession number GSE183465 at https://www.ncbi.nlm.nih.gov/geo/. The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE (126) partner repository with the dataset identifier PXD028566 and PXD028850.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank Klára Nováková, Paulina Spurk, Deis Haxholli and Fotios Gkountromichos for help with the experiments and Jeannette Koch and Ruzica Barisic for technical support. We thank Michael Bartoschek for critically reading the manuscript. We thank Weihua Qin for discussion and exchange of ideas. We thank Joshua Kie (Gene Center Flow Cytometry Core Facility, LMU Munich, Germany) for technical support with flow-cytometry. We thank Dr Igor Paron and Dr Christian Deiml for MS technical assistance. We thank Dr Philipp Paulitschke and Konstantin Schaffer (PHIO Scientific GmbH) for their support. We thank Dr H. Blum, Dr S. Krebs (Laboratory for Functional Genome Analysis, LMU Munich) for next-generation sequencing services. P.S. gratefully acknowledges the IRTG 1064 graduate school and E.U. the International Max Planck Research School for Molecular Life Sciences (IMPRS-LS) for training and support.
Author contributions: P.S. and S.B. designed and conceived the study. S.B. supervised the study. P.S., C.B.M., H.L. and S.B. interpreted the data. P.S. performed molecular, cellular and microscopy experiments and generated the EM-seq

library. P.S. generated stable cell lines with the help from C.B.M.. P.S. generated the RNA-seq library with help and supervision from L.E.W. and W.E. S.B. analyzed the data. A.S.M. performed quantitative ChIP-seq experiments with the supervision from S.E.. S.E. and Y.C. analysed the quantitative ChIP-seq data. A.T. and E.U. conducted proteomic experiments under the supervision of R.S., T.B., M.W. and H.L. A.T. and E.U. performed mass spectrometry analyses. P.S. and S.B. wrote the manuscript and prepared figures. All authors read, discussed, and approved the manuscript.

FUNDING

Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) [283300243 to S.B., 213249687/SFB 1064 to R.S., T.B., H.L., S.B., 325871075/SFB 1309 to R.S., SFB1243 (A14) to W.E., L.E.W.]; W.E. was supported by the Cyliax Foundation; R.S. and T.B. were supported by the Helmholtz Gesellschaft; S.E. acknowledges funding from Vetenskapsrådet [2015-04815, 2020-04313]; H2020 ERC-2016-StG [715024 RAPID]; Ming Wai Lau Center for Reparative Medicine, Cancerfonden [2015/430]. Funding for open access charge: German Research Foundation [213249687/SFB 1064].

Conflict of interest statement. None declared.

REFERENCES

1. Smith,Z.D. and Meissner,A. (2013) DNA methylation: roles in mammalian development. *Nat. Rev. Genet.*, **14**, 204–220.
2. Seisenberger,S., Peat,J.R. and Reik,W. (2013) Conceptual links between DNA methylation reprogramming in the early embryo and primordial germ cells. *Curr. Opin. Cell Biol.*, **25**, 281–288.
3. Tahiliani,M., Koh,K.P., Shen,Y., Pastor,W.A., Bandukwala,H., Brudno,Y., Agarwal,S., Iyer,L.M., Liu,D.R., Aravind,L. *et al.* (2009) Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1. *Science*, **324**, 930–935.
4. Iyer,L.M., Tahiliani,M., Rao,A. and Aravind,L. (2009) Prediction of novel families of enzymes involved in oxidative and other complex modifications of bases in nucleic acids. *Cell Cycle*, **8**, 1698–1710.
5. Ito,S., Shen,L., Dai,Q., Wu,S.C., Collins,L.B., Swenberg,J.A., He,C. and Zhang,Y. (2011) Tet proteins can convert 5-methylcytosine to 5-formylcytosine and 5-carboxylcytosine. *Science*, **333**, 1300–1303.
6. He,Y.-F., Li,B.-Z., Li,Z., Liu,P., Wang,Y., Tang,Q., Ding,J., Jia,Y., Chen,Z., Li,L. *et al.* (2011) Tet-mediated formation of 5-carboxylcytosine and its excision by TDG in mammalian DNA. *Science*, **333**, 1303–1307.
7. Pfaffeneder,T., Hackner,B., Truss,M., Münzel,M., Müller,M., Deiml,C.A., Hagemeier,C. and Carell,T. (2011) The discovery of 5-formylcytosine in embryonic stem cell DNA. *Angew. Chem. Int. Ed Engl.*, **50**, 7008–7012.
8. Maiti,A. and Drohat,A.C. (2011) Thymine DNA glycosylase can rapidly excise 5-formylcytosine and 5-carboxylcytosine: potential implications for active demethylation of CpG sites. *J. Biol. Chem.*, **286**, 35334–35338.
9. Hashimoto,H., Liu,Y., Upadhyay,A.K., Chang,Y., Howerton,S.B., Vertino,P.M., Zhang,X. and Cheng,X. (2012) Recognition and potential mechanisms for replication and erasure of cytosine hydroxymethylation. *Nucleic Acids Res.*, **40**, 4841–4849.
10. Otani,J., Kimura,H., Sharif,J., Endo,T.A., Mishima,Y., Kawakami,T., Koseki,H., Shirakawa,M., Suetake,I. and Tajima,S. (2013) Cell cycle-dependent turnover of 5-hydroxymethyl cytosine in mouse embryonic stem cells. *PLoS One*, **8**, e82961.
11. Bachman,M., Uribe-Lewis,S., Yang,X., Williams,M., Murrell,A. and Balasubramanian,S. (2014) 5-Hydroxymethylcytosine is a predominantly stable DNA modification. *Nat. Chem.*, **6**, 1049–1055.
12. Bachman,M., Uribe-Lewis,S., Yang,X., Burgess,H.E., Iurlaro,M., Reik,W., Murrell,A. and Balasubramanian,S. (2015) 5-Formylcytosine can be a stable DNA modification in mammals. *Nat. Chem. Biol.*, **11**, 555–557.
13. Yang,J., Bashkenova,N., Zang,R., Huang,X. and Wang,J. (2020) The roles of TET family proteins in development and stem cells. *Development*, **2**, 147.
14. Gu,T.-P., Guo,F., Yang,H., Wu,H.-P., Xu,G.-F., Liu,W., Xie,Z.-G., Shi,L., He,X., Jin,S.-G. *et al.* (2011) The role of tet3 DNA dioxygenase in epigenetic reprogramming by oocytes. *Nature*, **477**, 606–610.
15. Wossidlo,M., Nakamura,T., Lepikhov,K., Marques,C.J., Zakhartchenko,V., Boiani,M., Arand,J., Nakano,T., Reik,W. and Walter,J. (2011) 5-Hydroxymethylcytosine in the mammalian zygote is linked with epigenetic reprogramming. *Nat. Commun.*, **2**, 241.
16. Iqbal,K., Jin,S.-G., Pfeifer,G.P. and Szabó,P.E. (2011) Reprogramming of the paternal genome upon fertilization involves genome-wide oxidation of 5-methylcytosine. *Proc. Natl. Acad. Sci. U.S.A.*, **108**, 3642–3647.
17. Mulholland,C.B., Traube,F.R., Ugur,E., Parsa,E., Eckl,E.-M., Schönung,M., Modic,M., Bartoschek,M.D., Stolz,P., Ryan,J. *et al.* (2020) Distinct and stage-specific contributions of TET1 and TET2 to stepwise cytosine oxidation in the transition from naive to primed pluripotency. *Sci. Rep.*, **10**, 12066.
18. Dawlaty,M.M., Breiling,A., Le,T., Barrasa,M.I., Raddatz,G., Gao,Q., Powell,B.E., Cheng,A.W., Faull,K.F., Lyko,F. *et al.* (2014) Loss of tet enzymes compromises proper differentiation of embryonic stem cells. *Dev. Cell*, **29**, 102–111.
19. Kang,J., Lienhard,M., Pastor,W.A., Chawla,A., Novotny,M., Tsagaratou,A., Lasken,R.S., Thompson,E.C., Surani,M.A., Koralov,S.B. *et al.* (2015) Simultaneous deletion of the methylcytosine oxidases tet1 and tet3 increases transcriptome variability in early embryogenesis. *Proc. Natl. Acad. Sci. U.S.A.*, **112**, E4236–E4245.
20. Dai,H.-Q., Wang,B.-A., Yang,L., Chen,J.-J., Zhu,G.-C., Sun,M.-L., Ge,H., Wang,R., Chapman,D.L., Tang,F. *et al.* (2016) TET-mediated DNA demethylation controls gastrulation by regulating lefty-nodal signalling. *Nature*, **538**, 528–532.
21. Li,X., Yue,X., Pastor,W.A., Lin,L., Georges,R., Chavez,L., Evans,S.M. and Rao,A. (2016) Tet proteins influence the balance between neuroectodermal and mesodermal fate choice by inhibiting wnt signaling. *Proc. Natl. Acad. Sci. U.S.A.*, **113**, E8267–E8276.
22. Dawlaty,M.M., Ganz,K., Powell,B.E., Hu,Y.-C., Markoulaki,S., Cheng,A.W., Gao,Q., Kim,J., Choi,S.-W., Page,D.C. *et al.* (2011) Tet1 is dispensable for maintaining pluripotency and its loss is compatible with embryonic and postnatal development. *Cell Stem Cell*, **9**, 166–175.
23. Khoueiry,R., Sohni,A., Thienpont,B., Luo,X., Velde,J.V., Bartoccetti,M., Boeckx,B., Zwijsen,A., Rao,A., Lambrechts,D. *et al.* (2017) Lineage-specific functions of TET1 in the postimplantation mouse embryo. *Nat. Genet.*, **49**, 1061–1072.
24. Moran-Crusio,K., Reavie,L., Shih,A., Abdel-Wahab,O., Ndiaye-Lobry,D., Lobry,C., Figueroa,M.E., Vasanthakumar,A., Patel,J., Zhao,X. *et al.* (2011) Tet2 loss leads to increased hematopoietic stem cell self-renewal and myeloid transformation. *Cancer Cell*, **20**, 11–24.
25. Lu,F., Liu,Y., Jiang,L., Yamaguchi,S. and Zhang,Y. (2014) Role of TET proteins in enhancer activity and telomere elongation. *Genes Dev.*, **28**, 2103–2119.
26. Arab,K., Karaulanov,E., Musheev,M., Trnka,P., Schäfer,A., Grummt,I. and Niehrs,C. (2019) GADD45A binds R-loops and recruits TET1 to CpG island promoters. *Nat. Genet.*, **51**, 217–223.
27. Costa,Y., Ding,J., Theunissen,T.W., Faiola,F., Hore,T.A., Shliaha,P.V., Fidalgo,M., Saunders,A., Lawrence,M., Dietmann,S. *et al.* (2013) NANOG-dependent function of TET1 and TET2 in establishment of pluripotency. *Nature*, **495**, 370–374.
28. Manzo,M., Wirz,J., Ambrosi,C., Villaseñor,R., Roschitzki,B. and Baubec,T. (2017) Isoform-specific localization of DNMT3A regulates DNA methylation fidelity at bivalent CpG islands. *EMBO J.*, **36**, 3421–3434.
29. Mulholland,C.B., Nishiyama,A., Ryan,J., Nakamura,R., Yiğit,M., Glück,I.M., Trummer,C., Qin,W., Bartoschek,M.D., Traube,F.R. *et al.* (2020) Recent evolution of a TET-controlled and DPPA3/STELLA-driven pathway of passive DNA demethylation in mammals. *Nat. Commun.*, **11**, 5972.

30. Chen,Q., Chen,Y., Bian,C., Fujiki,R. and Yu,X. (2013) TET2 promotes histone O-GlcNAcylation during gene transcription. *Nature*, **493**, 561–564.
31. Neri,F., Incarnato,D., Krepelova,A., Rapelli,S., Pagnani,A., Zecchina,R., Parlato,C. and Oliviero,S. (2013) Genome-wide analysis identifies a functional association of TET1 and polycomb repressive complex 2 in mouse embryonic stem cells. *Genome Biol.*, **14**, R91.
32. Williams,K., Christensen,J., Pedersen,M.T., Johansen,J.V., Cloos,P.A.C., Rappsilber,J. and Helin,K. (2011) TET1 and hydroxymethylcytosine in transcription and DNA methylation fidelity. *Nature*, **473**, 343–348.
33. Wu,H., D'Alessio,A.C., Ito,S., Xia,K., Wang,Z., Cui,K., Zhao,K., Sun,Y.E. and Zhang,Y. (2011) Dual functions of TET1 in transcriptional regulation in mouse embryonic stem cells. *Nature*, **473**, 389–393.
34. Gu,T., Lin,X., Cullen,S.M., Luo,M., Jeong,M., Estecio,M., Shen,J., Hardikar,S., Sun,D., Su,J. *et al.* (2018) DNMT3A and TET1 cooperate to regulate promoter epigenetic landscapes in mouse embryonic stem cells. *Genome Biol.*, **19**, 88.
35. Zhang,P., Rausch,C., Hastert,F.D., Boneva,B., Filatova,A., Patil,S.J., Nuber,U.A., Gao,Y., Zhao,X. and Cardoso,M.C. (2017) Methyl-CpG binding domain protein 1 regulates localization and activity of tet1 in a CXXC3 domain-dependent manner. *Nucleic Acids Res.*, **45**, 7118–7136.
36. Zhong,J., Li,X., Cai,W., Wang,Y., Dong,S., Yang,J., Zhang,J., 'an,Wu, N., Li,Y., Mao,F. *et al.* (2017) TET1 modulates H4K16 acetylation by controlling auto-acetylation of hMOF to affect gene regulation and DNA repair function. *Nucleic Acids Res.*, **45**, 672–684.
37. Vella,P., Scelfo,A., Jammula,S., Chiacchiera,F., Williams,K., Cuomo,A., Roberto,A., Christensen,J., Bonaldi,T., Helin,K. *et al.* (2013) Tet proteins connect the O-linked N-acetylglucosamine transferase ogt to chromatin in embryonic stem cells. *Mol. Cell*, **49**, 645–656.
38. de la Rica,L., Deniz,Ö., Cheng,K.C.L., Todd,C.D., Cruz,C., Houseley,J. and Branco,M.R. (2016) TET-dependent regulation of retrotransposable elements in mouse embryonic stem cells. *Genome Biol.*, **17**, 234.
39. Walsh,C.P., Chaillet,J.R. and Bestor,T.H. (1998) Transcription of IAP endogenous retroviruses is constrained by cytosine methylation. *Nat. Genet.*, **20**, 116–117.
40. Jackson-Grusby,L., Beard,C., Possemato,R., Tudor,M., Fambrough,D., Csankovszki,G., Dausman,J., Lee,P., Wilson,C., Lander,E. *et al.* (2001) Loss of genomic methylation causes p53-dependent apoptosis and epigenetic deregulation. *Nat. Genet.*, **27**, 31–39.
41. Chiappinelli,K.B., Strissel,P.L., Desrichard,A., Li,H., Henke,C., Akman,B., Hein,A., Rote,N.S., Cope,L.M., Snyder,A. *et al.* (2017) Inhibiting DNA methylation causes an interferon response in cancer via dsRNA including endogenous retroviruses. *Cell*, **169**, 361.
42. Rowe,H.M. and Trono,D. (2011) Dynamic control of endogenous retroviruses during development. *Virology*, **411**, 273–287.
43. Howard,G., Eiges,R., Gaudet,F., Jaenisch,R. and Eden,A. (2008) Activation and transposition of endogenous retroviral elements in hypomethylation induced tumors in mice. *Oncogene*, **27**, 404–408.
44. Mikkelsen,T.S., Ku,M., Jaffe,D.B., Issac,B., Lieberman,E., Giannoukos,G., Alvarez,P., Brockman,W., Kim,T.-K., Koche,R.P. *et al.* (2007) Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature*, **448**, 553–560.
45. Schotta,G. (2004) A silencing pathway to induce H3-K9 and H4-K20 trimethylation at constitutive heterochromatin. *Genes Dev.*, **18**, 1251–1262.
46. Aravin,A.A., Hannon,G.J. and Brennecke,J. (2007) The Piwi-piRNA pathway provides an adaptive defense in the transposon arms race. *Science*, **318**, 761–764.
47. Chandru,A., Bate,N., Vuister,G.W. and Cowley,S.M. (2018) Sin3A recruits tet1 to the PAH1 domain via a highly conserved sin3-interaction domain. *Sci. Rep.*, **8**, 14689.
48. Mulholland,C.B., Smets,M., Schmidtmann,E., Leidescher,S., Markaki,Y., Hofweber,M., Qin,W., Manzo,M., Kremmer,E., Thanisch,K. *et al.* (2015) A modular open platform for systematic functional studies under physiological conditions. *Nucleic Acids Res.*, **43**, e112.
49. Bauer,C., Göbel,K., Nagaraj,N., Colantuoni,C., Wang,M., Müller,U., Kremmer,E., Rottach,A. and Leonhardt,H. (2015) Phosphorylation of TET proteins is regulated via O-GlcNAcylation by the O-linked N-acetylglucosamine transferase (OGT). *J. Biol. Chem.*, **290**, 4801–4812.
50. Kumar,B. and Elsässer,S.J. (2019) Quantitative multiplexed ChIP reveals global alterations that shape promoter bivalency in ground state embryonic stem cells. *Cell Reports*, **28**, 3274–3284.
51. Navarro,C., Martin,M. and Elsässer,S. (2022) minute: a MINUTE-ChIP data analysis workflow. bioRxiv doi: https://doi.org/10.1101/2022.03.14.484318, 17 March 2022, preprint: not peer reviewed.
52. Mölder,F., Jablonski,K.P., Letcher,B., Hall,M.B., Tomkins-Tinch,C.H., Sochat,V., Forster,J., Lee,S., Twardziok,S.O., Kanitz,A. *et al.* (2021) Sustainable data analysis with snakemake. *F1000Res.*, **10**, 33.
53. Martin,M. (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*, **17**, 10.
54. Ewels,P., Magnusson,M., Lundin,S. and Käller,M. (2016) MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics*, **32**, 3047–3048.
55. Kloet,S.L., Karemaker,I.D., van Voorthuijsen,L., Lindeboom,R.G.H., Baltissen,M.P., Edupuganti,R.R., Poramba-Liyanage,D.W., Jansen,P.W.T.C. and Vermeulen,M. (2018) NuRD-interacting protein ZFP296 regulates genome-wide NuRD localization and differentiation of mouse embryonic stem cells. *Nat. Commun.*, **9**, 4588.
56. Bilodeau,S., Kagey,M.H., Frampton,G.M., Rahl,P.B. and Young,R.A. (2009) SetDB1 contributes to repression of genes encoding developmental regulators and maintenance of ES cell state. *Genes Dev.*, **23**, 2484–2489.
57. Karmodiya,K., Krebs,A.R., Oulad-Abdelghani,M., Kimura,H. and Tora,L. (2012) H3K9 and H3K14 acetylation co-occur at many gene regulatory elements, while H3K14ac marks a subset of inactive inducible promoters in mouse embryonic stem cells. *BMC Genomics*, **13**, 424.
58. Zhang,Y., Liu,T., Meyer,C.A., Eeckhoute,J., Johnson,D.S., Bernstein,B.E., Nusbaum,C., Myers,R.M., Brown,M., Li,W. *et al.* (2008) Model-based analysis of chip-Seq (MACS). *Genome Biol.*, **9**, R137.
59. Akalin,A., Kormaksson,M., Li,S., Garrett-Bakelman,F.E., Figueroa,M.E., Melnick,A. and Mason,C.E. (2012) methylKit: a comprehensive r package for the analysis of genome-wide DNA methylation profiles. *Genome Biol.*, **13**, R87.
60. Illingworth,R.S., Gruenewald-Schneider,U., Webb,S., Kerr,A.R.W., James,K.D., Turner,D.J., Smith,C., Harrison,D.J., Andrews,R. and Bird,A.P. (2010) Orphan CpG islands identify numerous conserved promoters in the mammalian genome. *PLoS Genet.*, **6**, e1001134.
61. Shechter,D., Dormann,H.L., Allis,C.D. and Hake,S.B. (2007) Extraction, purification and analysis of histones. *Nat. Protoc.*, **2**, 1445–1457.
62. Maile,T.M., Izrael-Tomasevic,A., Cheung,T., Guler,G.D., Tindell,C., Masselot,A., Liang,J., Zhao,F., Trojer,P., Classon,M. *et al.* (2015) Mass spectrometric quantification of histone post-translational modifications by a hybrid chemical labeling method. *Mol. Cell. Proteomics*, **14**, 1148–1158.
63. Yuan,Z.-F., Sidoli,S., Marchione,D.M., Simithy,J., Janssen,K.A., Szurgot,M.R. and Garcia,B.A. (2018) EpiProfile 2.0: a computational platform for processing epi-proteomics mass spectrometry data. *J. Proteome Res.*, **17**, 2533–2541.
64. Janjic,A., Wange,L.E., Bagnoli,J.W., Geuder,J., Nguyen,P., Richter,D., Vieth,B., Vick,B., Jeremias,I., Ziegenhain,C. *et al.* (2022) Prime-seq, efficient and powerful bulk RNA sequencing. *Genome Biol.*, **23**, 88.
65. Bagnoli,J.W., Ziegenhain,C., Janjic,A., Wange,L.E., Vieth,B., Parekh,S., Geuder,J., Hellmann,I. and Enard,W. (2018) Sensitive and powerful single-cell RNA sequencing using mcSCRB-seq. *Nat. Commun.*, **9**, 2937.
66. Renaud,G., Stenzel,U., Maricic,T., Wiebe,V. and Kelso,J. (2015) deML: robust demultiplexing of illumina sequences using a likelihood-based approach. *Bioinformatics*, **31**, 770–772.
67. Parekh,S., Ziegenhain,C., Vieth,B., Enard,W. and Hellmann,I. (2018) zUMIs - A fast and flexible pipeline to process RNA sequencing data with UMIs. *GigaScience*, **7**, giy059.

68. Dobin,A., Davis,C.A., Schlesinger,F., Drenkow,J., Zaleski,C., Jha,S., Batut,P., Chaisson,M. and Gingeras,T.R. (2013) STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, **29**, 15–21.
69. Rau,A., Gallopin,M., Celeux,G. and Jaffrézic,F. (2013) Data-based filtering for replicated high-throughput transcriptome sequencing experiments. *Bioinformatics*, **29**, 2146–2152.
70. Love,M.I., Huber,W. and Anders,S. (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.*, **15**, 550.
71. Jin,Y., Tam,O.H., Paniagua,E. and Hammell,M. (2015) TEtranscripts: a package for including transposable elements in differential expression analysis of RNA-seq datasets. *Bioinformatics*, **31**, 3593–3599.
72. Qin,W., Ugur,E., Mulholland,C.B., Bultmann,S., Solovei,I., Modic,M., Smets,M., Wierer,M., Forné,I., Imhof,A. *et al.* (2021) Phosphorylation of the HP1β hinge region sequesters KAP1 in heterochromatin and promotes the exit from naïve pluripotency. *Nucleic Acids Res.*, **49**, 7406–7423.
73. Rappsilber,J., Mann,M. and Ishihama,Y. (2007) Protocol for micro-purification, enrichment, pre-fractionation and storage of peptides for proteomics using stagetips. *Nat. Protoc.*, **2**, 1896–1906.
74. Cox,J. and Mann,M. (2008) MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nature Biotechnology*, **26**, 1367–1372.
75. Cox,J., Hein,M.Y., Luber,C.A., Paron,I., Nagaraj,N. and Mann,M. (2014) Accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction, termed MaxLFQ. *Mol. Cell. Proteomics*, **13**, 2513–2526.
76. Fidalgo,M., Huang,X., Guallar,D., Sanchez-Priego,C., Valdes,V.J., Saunders,A., Ding,J., Wu,W.-S., Clavel,C. and Wang,J. (2016) Zfp281 coordinates opposing functions of tet1 and tet2 in pluripotent states. *Cell Stem Cell*, **19**, 355–369.
77. Verma,N., Pan,H., Doré,L.C., Shukla,A., Li,Q.V., Pelham-Webb,B., Teijeiro,V., González,F., Krivtsov,A., Chang,C.-J. *et al.* (2018) TET proteins safeguard bivalent promoters from de novo methylation in human embryonic stem cells. *Nat. Genet.*, **50**, 83–95.
78. Zhang,X., Su,J., Jeong,M., Ko,M., Huang,Y., Park,H.J., Guzman,A., Lei,Y., Huang,Y.-H., Rao,A. *et al.* (2016) DNMT3A and TET2 compete and cooperate to repress lineage-specific transcription factors in hematopoietic stem cells. *Nat. Genet.*, **48**, 1014–1023.
79. Zhang,R.-R., Cui,Q.-Y., Murai,K., Lim,Y.C., Smith,Z.D., Jin,S., Ye,P., Rosa,L., Lee,Y.K., Wu,H.-P. *et al.* (2013) Tet1 regulates adult hippocampal neurogenesis and cognition. *Cell Stem Cell*, **13**, 237–245.
80. Zhang,W., Xia,W., Wang,Q., Towers,A.J., Chen,J., Gao,R., Zhang,Y., Yen,C.-A., Lee,A.Y., Li,Y. *et al.* (2016) Isoform switch of TET1 regulates DNA demethylation and mouse development. *Mol. Cell*, **64**, 1062–1073.
81. Hon,G.C., Song,C.-X., Du,T., Jin,F., Selvaraj,S., Lee,A.Y., Yen,C.-A., Ye,Z., Mao,S.-Q., Wang,B.-A. *et al.* (2014) 5mC oxidation by Tet2 modulates enhancer activity and timing of transcriptome reprogramming during differentiation. *Mol. Cell*, **56**, 286–297.
82. López-Moyado,I.F., Tsagaratou,A., Yuita,H., Seo,H., Delatte,B., Heinz,S., Benner,C. and Rao,A. (2019) Paradoxical association of TET loss of function with genome-wide DNA hypomethylation. *Proc. Natl. Acad. Sci. U.S.A.*, **116**, 16933–16942.
83. Matsui,T., Leung,D., Miyashita,H., Maksakova,I.A., Miyachi,H., Kimura,H., Tachibana,M., Lorincz,M.C. and Shinkai,Y. (2010) Proviral silencing in embryonic stem cells requires the histone methyltransferase ESET. *Nature*, **464**, 927–931.
84. Karimi,M.M., Goyal,P., Maksakova,I.A., Bilenky,M., Leung,D., Tang,J.X., Shinkai,Y., Mager,D.L., Jones,S., Hirst,M. *et al.* (2011) DNA methylation and SETDB1/H3K9me3 regulate predominantly distinct sets of genes, retroelements, and chimeric transcripts in mESCs. *Cell Stem Cell*, **8**, 676–687.
85. Macfarlan,T.S., Gifford,W.D., Driscoll,S., Lettieri,K., Rowe,H.M., Bonanomi,D., Firth,A., Singer,O., Trono,D. and Pfaff,S.L. (2012) Embryonic stem cell potency fluctuates with endogenous retrovirus activity. *Nature*, **487**, 57–63.
86. Sachs,P., Ding,D., Bergmaier,P., Lamp,B., Schlagheck,C., Finkernagel,F., Nist,A., Stiewe,T. and Mermoud,J.E. (2019) SMARCAD1 ATPase activity is required to silence endogenous retroviruses in embryonic stem cells. *Nat. Commun.*, **10**, 1335.
87. Groh,S., Milton,A.V., Marinelli,L., Sickinger,C.V., Bollig,H., de Almeida,G.P., Forné,I., Schmidt,A., Imhof,A. and Schotta,G. (2021)Morc3 silences endogenous retroviruses by enabling Daxx-mediated H3.3 incorporation. *Nat. Commun.*, **12**, 5996.
88. Rowe,H.M., Jakobsson,J., Mesnard,D., Rougemont,J., Reynard,S., Aktas,T., Maillard,P.V., Layard-Liesching,H., Verp,S., Marquis,J. *et al.* (2010) KAP1 controls endogenous retroviruses in embryonic stem cells. *Nature*, **463**, 237–240.
89. Groh,S. and Schotta,G. (2017) Silencing of endogenous retroviruses by heterochromatin. *Cell. Mol. Life Sci.*, **74**, 2055–2065.
90. Weihs,F., Wacnik,K., Turner,R.D., Culley,S., Henriques,R. and Foster,S.J. (2018) Heterogeneous localisation of membrane proteins in staphylococcus aureus. *Sci. Rep.*, **8**, 3657.
91. Osswald,M., Santos,A.F. and Morais-de-Sá,E. (2019) Light-Induced protein clustering for optogenetic interference and protein interaction analysis in S2 cells. *Biomolecules*, **2**, 9.
92. Larson,A.G., Elnatan,D., Keenen,M.M., Trnka,M.J., Johnston,J.B., Burlingame,A.L., Agard,D.A., Redding,S. and Narlikar,G.J. (2017) Liquid droplet formation by HP1α suggests a role for phase separation in heterochromatin. *Nature*, **547**, 236–240.
93. Qin,W., Ugur,E., Mulholland,C.B., Bultmann,S., Solovei,I., Modic,M., Smets,M., Wierer,M., Forné,I., Imhof,A. *et al.* (2021) Phosphorylation of the HP1β hinge region sequesters KAP1 in heterochromatin and promotes the exit from naïve pluripotency. *Nucleic Acids Res.*, **49**, 7406–7423.
94. Taddei,A., Maison,C., Roche,D. and Almouzni,G. (2001) Reversible disruption of pericentric heterochromatin and centromere function by inhibiting deacetylases. *Nat. Cell Biol.*, **3**, 114–120.
95. Braunstein,M., Sobel,R.E., Allis,C.D., Turner,B.M. and Broach,J.R. (1996) Efficient transcriptional silencing in saccharomyces cerevisiae requires a heterochromatin histone acetylation pattern. *Mol. Cell. Biol.*, **16**, 4349–4356.
96. Jeppesen,P. and Turner,B.M. (1993) The inactive x chromosome in female mammals is distinguished by a lack of histone H4 acetylation, a cytogenetic marker for gene expression. *Cell*, **74**, 281–289.
97. O'Neill,L.P. and Turner,B.M. (1995) Histone H4 acetylation distinguishes coding regions of the human genome from heterochromatin in a differentiation-dependent but transcription-independent manner. *EMBO J.*, **14**, 3946–3957.
98. Streubel,G., Fitzpatrick,D.J., Oliviero,G., Scelfo,A., Moran,B., Das,S., Munawar,N., Watson,A., Wynne,K., Negri,G.L. *et al.* (2017) Fam60a defines a variant sin3a-hdac complex in embryonic stem cells required for self-renewal. *EMBO J.*, **36**, 2216–2232.
99. Zhu,F., Zhu,Q., Ye,D., Zhang,Q., Yang,Y., Guo,X., Liu,Z., Jiapaer,Z., Wan,X., Wang,G. *et al.* (2018) Sin3a-Tet1 interaction activates gene transcription and is required for embryonic stem cell pluripotency. *Nucleic Acids Res.*, **46**, 6026–6040.
100. Maksakova,I.A., Goyal,P., Bullwinkel,J., Brown,J.P., Bilenky,M., Mager,D.L., Singh,P.B. and Lorincz,M.C. (2011) H3K9me3-binding proteins are dispensable for SETDB1/H3K9me3-dependent retroviral silencing. *Epigenetics Chromatin*, **4**, 12.
101. Schultz,D.C., Ayyanathan,K., Negorev,D., Maul,G.G. and Rauscher,F.J. 3rd (2002) SETDB1: a novel KAP-1-associated histone H3, lysine 9-specific methyltransferase that contributes to HP1-mediated silencing of euchromatic genes by KRAB zinc-finger proteins. *Genes Dev.*, **16**, 919–932.
102. Wu,X. and Zhang,Y. (2017) TET-mediated active DNA demethylation: mechanism, function and beyond. *Nat. Rev. Genet.*, **18**, 517–534.
103. Koh,K.P., Yabuuchi,A., Rao,S., Huang,Y., Cunniff,K., Nardone,J., Laiho,A., Tahiliani,M., Sommer,C.A., Mostoslavsky,G. *et al.* (2011) Tet1 and tet2 regulate 5-hydroxymethylcytosine production and cell lineage specification in mouse embryonic stem cells. *Cell Stem Cell*, **8**, 200–213.
104. Luo,X., van der Veer,B.K., Sun,L., Bartoccetti,M., Boretto,M., Vankelecom,H., Khoueiry,R. and Koh,K.P. (2020) Coordination of germ layer lineage choice by TET1 during primed pluripotency. *Genes Dev.*, **34**, 598–618.
105. Villivalam,S.D., You,D., Kim,J., Lim,H.W., Xiao,H., Zushin,P.-J.H., Oguri,Y., Amin,P. and Kang,S. (2020) TET1 is a beige

adipocyte-selective epigenetic suppressor of thermogenesis. *Nat. Commun.*, **11**, 4313.
106. Chrysanthou,S., Tang,Q., Lee,J., Taylor,S.J., Zhao,Y., Steidl,U., Zheng,D. and Dawlaty,M.M. (2022) The DNA dioxygenase tet1 regulates H3K27 modification and embryonic stem cell biology independent of its catalytic activity. *Nucleic Acids Res.*, **50**, 3169–3189.
107. Yamaguchi,S., Shen,L., Liu,Y., Sendler,D. and Zhang,Y. (2013) Role of tet1 in erasure of genomic imprinting. *Nature*, **504**, 460–464.
108. Kaas,G.A., Zhong,C., Eason,D.E., Ross,D.L., Vachhani,R.V., Ming,G.-L., King,J.R., Song,H. and Sweatt,J.D. (2013) TET1 controls CNS 5-methylcytosine hydroxylation, active DNA demethylation, gene transcription, and memory formation. *Neuron*, **79**, 1086–1093.
109. Tsagaratou,A. and Rao,A. (2013) TET proteins and 5-methylcytosine oxidation in the immune system. *Cold Spring Harb. Symp. Quant. Biol.*, **78**, 1–10.
110. Hu,X., Zhang,L., Mao,S.-Q., Li,Z., Chen,J., Zhang,R.-R., Wu,H.-P., Gao,J., Guo,F., Liu,W. *et al.* (2014) Tet and TDG mediate DNA demethylation essential for mesenchymal-to-epithelial transition in somatic cell reprogramming. *Cell Stem Cell*, **14**, 512–522.
111. Mulholland,C.B., Traube,F.R., Ugur,E., Parsa,E., Eckl,E.-M., Schönung,M., Modic,M., Bartoschek,M.D., Stolz,P., Ryan,J. *et al.* (2020) Distinct and stage-specific contributions of TET1 and TET2 to stepwise cytosine oxidation in the transition from naive to primed pluripotency. *Sci. Rep.*, **10**, 12066.
112. Baumann,V., Wiesbeck,M., Breunig,C.T., Braun,J.M., Köferle,A., Ninkovic,J., Götz,M. and Stricker,S.H. (2019) Targeted removal of epigenetic barriers during transcriptional reprogramming. *Nat. Commun.*, **10**, 2119.
113. Rodriguez-Terrones,D., Hartleben,G., Gaume,X., Eid,A., Guthmann,M., Iturbide,A. and Torres-Padilla,M. (2020) A distinct metabolic state arises during the emergence of 2-cell-like cells. *EMBO Reports*, **21**, e48354.
114. Ramírez,M.A., Pericuesta,E., Fernandez-Gonzalez,R., Moreira,P., Pintado,B. and Gutierrez-Adan,A. (2006) Transcriptional and post-transcriptional regulation of retrotransposons IAP and MuERV-L affect pluripotency of mice ES cells. *Reprod. Biol. Endocrinol.*, **4**, 55.
115. Dong,K.B., Maksakova,I.A., Mohn,F., Leung,D., Appanah,R., Lee,S., Yang,H.W., Lam,L.L., Mager,D.L., Schübeler,D. *et al.* (2008) DNA methylation in ES cells requires the lysine methyltransferase G9a but not its catalytic activity. *EMBO J.*, **27**, 2691–2701.
116. Yin,Y., Morgunova,E., Jolma,A., Kaasinen,E., Sahu,B., Khund-Sayeed,S., Das,P.K., Kivioja,T., Dave,K., Zhong,F. *et al.* (2017) Impact of cytosine methylation on DNA binding specificities of human transcription factors. *Science*, **356**, 6337.
117. Geis,F.K. and Goff,S.P. (2020) Silencing and transcriptional regulation of endogenous retroviruses: an overview. *Viruses*, **12**, 884.
118. Taddei,A., Roche,D., Sibarita,J.B., Turner,B.M. and Almouzni,G. (1999) Duplication and maintenance of heterochromatin domains. *J. Cell Biol.*, **147**, 1153–1166.
119. Nielsen,A.L., Ortiz,J.A., You,J., Oulad-Abdelghani,M., Khechumian,R., Gansmuller,A., Chambon,P. and Losson,R. (1999) Interaction with members of the heterochromatin protein 1 (HP1) family and histone deacetylation are differentially involved in transcriptional silencing by members of the TIF1 family. *EMBO J.*, **18**, 6385–6395.
120. Schultz,D.C., Friedman,J.R. and Rauscher,F.J. 3rd (2001) Targeting histone deacetylase complexes via KRAB-zinc finger proteins: the PHD and bromodomains of KAP-1 form a cooperative unit that recruits a novel isoform of the Mi-2alpha subunit of NuRD. *Genes Dev.*, **15**, 428–443.
121. Maison,C., Bailly,D., Peters,A.H.F.M., Quivy,J.-P., Roche,D., Taddei,A., Lachner,M., Jenuwein,T. and Almouzni,G. (2002) Higher-order structure in pericentric heterochromatin involves a distinct pattern of histone modification and an RNA component. *Nat. Genet.*, **30**, 329–334.
122. Rice,J.C. and Allis,C.D. (2001) Histone methylation versus histone acetylation: new insights into epigenetic regulation. *Curr. Opin. Cell Biol.*, **13**, 263–273.
123. Vermeulen,M., Walter,W., Le Guezennec,X., Kim,J., Edayathumangalam,R.S., Lasonder,E., Luger,K., Roeder,R.G., Logie,C., Berger,S.L. *et al.* (2006) A feed-forward repression mechanism anchors the Sin3/histone deacetylase and N-CoR/SMRT corepressors on chromatin. *Mol. Cell. Biol.*, **26**, 5226–5236.
124. Dannenberg,J.-H., David,G., Zhong,S., van der Torre,J., Wong,W.H. and Depinho,R.A. (2005) mSin3A corepressor regulates diverse transcriptional networks governing normal and neoplastic growth and survival. *Genes Dev.*, **19**, 1581–1595.
125. Silverstein,R.A. and Ekwall,K. (2005) Sin3: a flexible regulator of global gene expression and genome stability. *Curr. Genet.*, **47**, 1–17.
126. Perez-Riverol,Y., Csordas,A., Bai,J., Bernal-Llinares,M., Hewapathirana,S., Kundu,D.J., Inuganti,A., Griss,J., Mayer,G., Eisenacher,M. *et al.* (2019) The PRIDE database and related tools and resources in 2019: improving support for quantification data. *Nucleic Acids Res.*, **47**, D442–D450.

Wang, Z., Fan, R., Russo, A., Cernilogar, F. M., Nuber, A., Schirge, S., Shcherbakova, I., Dzhilyanova, I., **Ugur, E.**, Anton, T., Richter, L., Leonhardt, H., Lickert, H., & Schotta, G. (**2022**). **Dominant role of DNA methylation over H3K9me3 for IAP silencing in endoderm**. Nature Communications, 13(1), 5447.

License: CC-BY-4.0

https://doi.org/10.1038/s41467-022-32978-7

nature communications

Article

https://doi.org/10.1038/s41467-022-32978-7

Dominant role of DNA methylation over H3K9me3 for IAP silencing in endoderm

Received: 21 January 2020

Accepted: 25 August 2022

Published online: 19 September 2022

Check for updates

Zeyang Wang[1,9], Rui Fan[1,8,9], Angela Russo[1], Filippo M. Cernilogar[1], Alexander Nuber[1], Silvia Schirge[2,3], Irina Shcherbakova[1], Iva Dzhilyanova[1], Enes Ugur[4], Tobias Anton[4], Lisa Richter[5], Heinrich Leonhardt[4], Heiko Lickert[2,3,6,7] & Gunnar Schotta[1] ✉

Silencing of endogenous retroviruses (ERVs) is largely mediated by repressive chromatin modifications H3K9me3 and DNA methylation. On ERVs, these modifications are mainly deposited by the histone methyltransferase *Setdb1* and by the maintenance DNA methyltransferase *Dnmt1*. Knock-out of either *Setdb1* or *Dnmt1* leads to ERV de-repression in various cell types. However, it is currently not known if H3K9me3 and DNA methylation depend on each other for ERV silencing. Here we show that conditional knock-out of *Setdb1* in mouse embryonic endoderm results in ERV de-repression in visceral endoderm (VE) descendants and does not occur in definitive endoderm (DE). Deletion of *Setdb1* in VE progenitors results in loss of H3K9me3 and reduced DNA methylation of *Intracisternal A-particle* (*IAP*) elements, consistent with up-regulation of this ERV family. In DE, loss of *Setdb1* does not affect H3K9me3 nor DNA methylation, suggesting *Setdb1*-independent pathways for maintaining these modifications. Importantly, *Dnmt1* knock-out results in *IAP* de-repression in both visceral and definitive endoderm cells, while H3K9me3 is unaltered. Thus, our data suggest a dominant role of DNA methylation over H3K9me3 for *IAP* silencing in endoderm cells. Our findings suggest that Setdb1-meditated H3K9me3 is not sufficient for *IAP* silencing, but rather critical for maintaining high DNA methylation.

ERVs are remnants of retroviral germline integrations during evolution. In mammalian genomes, a large proportion of ERV Long Terminal Repeats (LTR) contribute to the physiological regulation of gene expression during development. In this sense, ERV initiated transcripts contribute to pluripotency regulation in both mice and human embryonic stem cells[1,2]. In contrast, aberrantly high activity of ERVs is associated with diseases and abnormal development. Dys-regulation of ERV LTRs can drive expression of oncogenes in human tumor cells[3–5] and, overexpression of ERVs is a feature of autoimmune diseases[6]. Thus, silencing mechanisms, which restrict ERV activity are important to ensure proper development and, misregulation may lead to disease.

Silencing of endogenous retroviruses is mediated by heterochromatin and, in particular, by establishment of H3K9me3 and DNA methylation[7]. The major H3K9me3 specific histone methyltransferase (HMTase) for ERVs is SETDB1[8], but additional HMTases, such as SUV39H, SETDB2 and PRDM enzymes, can contribute to establishing

[1]Division of Molecular Biology, Biomedical Center, Faculty of Medicine, LMU Munich, Munich, Germany. [2]Helmholtz Zentrum München, Institute of Stem Cell Research, Neuherberg, Germany. [3]Helmholtz Zentrum München, Institute of Diabetes and Regeneration Research, Neuherberg, Germany. [4]Biozentrum, LMU Munich, Munich, Germany. [5]Biomedical Center (BMC), Core Facility Flow Cytometry, Faculty of Medicine, LMU Munich, Munich, Germany. [6]German Center for Diabetes Research (DZD), Neuherberg, Germany. [7]Technische Universität München, Munich, Germany. [8]Present address: Embryonic Self-Organization Research Group, Max Planck Institute for Molecular Biomedicine, Münster, Germany. [9]These authors contributed equally: Zeyang Wang, Rui Fan.
✉e-mail: gunnar.schotta@bmc.med.lmu.de

this modification[9–12]. Establishment of H3K9me3 on ERVs depends on the sequence-specific recognition by KRAB-ZFP proteins[13]. The KRAB domain of these proteins is bound by the corepressor TRIM28, which then recruits SETDB1[14]. DNA methylation on ERVs is deposited during preimplantation development and then maintained by DNMT1. Establishment and maintenance of DNA methylation on ERVs relates to the H3K9me3 pathway. This is shown by impaired DNA methylation in *Trim28* ko embryos[15] and upon deletion of *Setdb1* in ESCs and other cell types[16,17]. Mechanistically, the connection between H3K9me3 and maintenance of DNA methylation is not fully understood. UHRF1 can target DNMT1 to ERVs through binding of H3K9me3 and hemi-methylated DNA[18,19], however, UHRF1 mutant proteins with impaired H3K9me3 binding can still maintain substantial levels of DNA methylation[20].

Although H3K9me3 and DNA methylation are both enriched on a subset of ERVs, e.g. *IAP* elements, the role of these modifications for silencing seems to differ in various cell types. Deletion of *Setdb1* in embryonic stem cells (ESC) leads to strong *IAP* de-repression, whereas constitutive *Dnmt1* ko ESCs did not show clear transcriptional changes of *IAP* elements[21]. However, acute deletion of *Dnmt1* in ESCs did result in transient *IAP* de-repression, demonstrating that DNA methylation has an important function for *IAP* silencing in ESCs[22]. In differentiated cells, DNA methylation was initially found to play a crucial role for *IAP* silencing, as *Dnmt1* ko embryos display strong *IAP* expression in various cell types[23]. Specific impairment of DNA methylation in neuronal cells could recapitulate these findings[24,25]. Interestingly, deletion of *Setdb1* in neuronal cells also results in *IAP* de-repression[26], although a broader investigation of the roles of *Setdb1* in differentiated cells revealed that *Setdb1* is dispensable for *IAP* silencing in some differentiated cell types[27]. These findings demonstrate that both modification pathways have important roles for *IAP* repression in distinct differentiated cell types. However, the interplay between these pathways has not been investigated in the same cell type.

Here, we investigate the role of *Setdb1* for ERV silencing upon germ layer differentiation in the endoderm lineage. We find that *Setdb1*-mediated ERV repression is restricted to extra-embryonic VE cells and does not occur in embryonic DE, suggesting ontogenesis-dependent regulatory mechanisms for ERV silencing. In both endoderm cell types, DNA methylation plays a dominant role for ERV and, in particular, *IAP* repression. Interestingly, H3K9me3 is maintained on de-repressed *IAP* elements in *Dnmt1* ko VE and DE cells, suggesting that H3K9me3 in absence of DNA methylation is not sufficient to establish transcriptional repression.

Results

Loss of *Setdb1* in embryonic endoderm cells leads to developmental defects

Setdb1 is highly expressed during mouse embryonic development (Supplementary Fig. 1a). To study the role of *Setdb1* specifically in endoderm lineage development we combined a conditional *Setdb1*^{flox} allele with the *Sox17-2A-iCre* knock-in allele which expresses Cre recombinase in *Sox17* expressing endoderm cells[28]. *Sox17* is expressed in both embryonic and extraembryonic endoderm cells, therefore deletion of *Setdb1* is expected to occur in both lineages (Supplementary Fig. 1b). As the conditional deletion of one allele of *Setdb1* in the endoderm was phenotypically normal, we assign *Setdb1*$^{flox/+}$*; Sox17-2A-iCre* or *Setdb1*$^{flox/+}$ mice as control and mutant *Setdb1*$^{flox/flox}$*; Sox17-2A-iCre* as *Setdb1*^{END}.

Setdb1^{END} embryos do not display notable differences in development until embryonic day 8.5, where mutant embryos cannot complete turning (Fig. 1a). At later developmental stages (E9.0) the posterior part of *Setdb1*^{END} embryos deteriorates (Fig. 1a), followed by death of mutant embryos during later stages of pregnancy. The expression of key endoderm transcription factors *Foxa2* and *Sox17* was unaltered in E7.5 embryos (Fig. 1b), and no obvious structural aberrations could be detected in E8.0 embryos (Supplementary Fig. 1c), demonstrating proper initiation of endoderm differentiation. Later developmental stages clearly show endoderm-derived gut tube structures (Fig. 1c, red arrows). In *Setdb1*^{END} embryos gut tube structures were not fully connected to neighboring tissue (Fig. 1c, red arrows), which could explain the turning defect. We did not observe notable differences in proliferation or apoptosis in *Setdb1*^{END} embryos (Supplementary Fig. 1d, e). Together our data demonstrate a crucial role for *Setdb1* in embryonic endoderm differentiation. In contrast to ESCs, where deletion of *Setdb1* results in cell death[8,29,30], *Setdb1*-deficient endoderm cells (between E7.5 - E8.5) do not display obvious viability or proliferation problems, but rather have functional defects which result in altered tissue integrity.

Setdb1^{END} embryos display ERV de-repression specifically in visceral endoderm cells

To study transcriptional changes upon *Setdb1* deletion in endoderm cells, we combined *Setdb1*^{flox}*; Sox17-2A-iCre* alleles with an *EGFP-Cre* reporter allele[31]. The resulting *Setdb1*^{flox}*; Sox17-2A-iCre; EGFP-reporter* embryos displayed EGFP signals in *Sox17* expressing cells, whereas no *EGFP* expression was detected in *Setdb1*^{flox}*; EGFP-reporter* embryos which lack *Cre* activity at E8.5 (Supplementary Fig. 2a, b). The specific *Cre* reporter activity allowed us to FACS-isolate *EGFP* positive endoderm cells from control and *Setdb1*^{END} embryos (Supplementary Fig. 2c). Expression of *Setdb1* was strongly reduced in *Setdb1*^{END} endoderm cells (Supplementary Fig. 2d), demonstrating efficient deletion of *Setdb1*. Transcriptional profiling of control vs. *Setdb1*^{END} endoderm cells revealed significant up-regulation of 166 genes and down-regulation of only four genes (Fig. 2a, Supplementary Data 1). Top upregulated are germline-specific genes, known targets of *Setdb1* also in other cell types and consistent with the role of *Setdb1* in gene repression[17,21]. The top downregulated gene, *Nepn*, is an important marker gene for endoderm development[32,33]. Reduced *Nepn* expression domain could be confirmed by in situ hybridization (Supplementary Fig. 2e) and further indicates defective endoderm development observed in *Setdb1*^{END} embryos.

To investigate the role of *Setdb1* in ERV silencing we analyzed the expression of ERV families in control vs. *Setdb1*^{END} ex vivo endoderm cells. We observed strong de-repression of several ERV families (Fig. 2b, Supplementary Data 2). Expression of *LINE* elements was largely unchanged with only *L1Md_T* upregulated (Supplementary Fig. 2f). A prominent ERV family which is targeted by *Setdb1* in different cell types is *IAPEz*. Interestingly, *IAPEz* transcripts retained coding potential and, therefore, expression of *IAPEz* can be detected by presence of capsid protein GAG[34]. To investigate if ERV de-repression occurs in all embryonic endoderm lineage cells in *Setdb*^{END} embryos, we performed immunofluorescence analysis for IAPEz GAG. In E8.0 embryos we observe strong labeling of embryonic endoderm cells with the *EGFP* reporter, however, only a subset of reporter positive cells displayed GAG labeling (Fig. 2c). This suggests that *Setdb1* knock-out results in IAPEz de-repression in only a subset of embryonic endoderm cells. Embryonic endoderm is ontogenetically derived from both definitive endoderm which emerges around E6.5/E7.0 during gastrulation and visceral endoderm, which forms before gastrulation (Supplementary Fig. 1b). Since visceral endoderm cells assume morphology and function of definitive endoderm cells during gastrulation, they can only be distinguished using specific markers at the beginning of gastrulation. Therefore, to test if *IAPEz* de-repression may be specific to definitive or visceral endoderm in vivo, we stained E7.5 embryos for IAPEz GAG together with AFP, a specific marker of visceral endoderm[35]. AFP is strongly expressed in the visceral endoderm of the proximal extra-embryonic part of the embryo at E7.5 (Fig. 2d, region above white dashed line). Visceral endoderm descendants in the embryonic part start losing AFP expression during gastrulation, but we could still detect several AFP positive cells in the embryonic endoderm region at

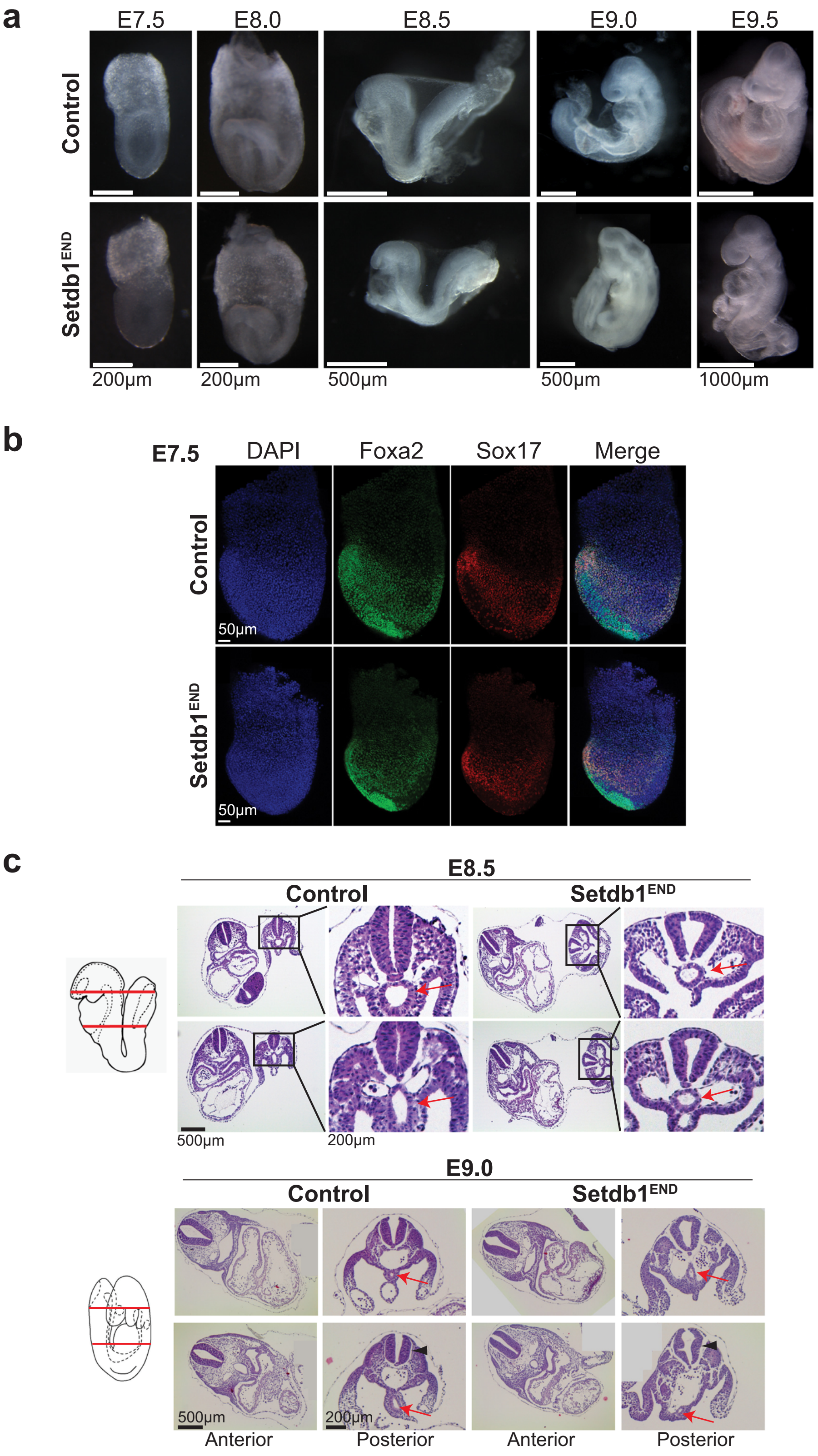


Fig. 1 | Loss of Setdb1 in endoderm leads to strong developmental defects during embryogenesis. **a** Lateral view of control and Setdb1^{END} embryos at E7.5 (late bud stage), E7.75 (head fold stage), E8.5 (~6 somite) and E9. No visible developmental defects can be detected in Setdb1^{END} embryos from E7.5 to E8.0. Setdb1^{END} embryos show an axis turning defect which manifests from E8.5 and leads to strong posterior truncation at E9.0. Representative images from $n = 3$ per genotype and stage. **b** Lateral view of E7.5 control and Setdb1^{END} embryos stained with Foxa2 and Sox17 antibodies (anterior to the left). The presence of both markers indicates that endoderm cells could be formed in Setdb1^{END} embryos. Representative images from $n = 3$ per genotype and stage. **c** Hematoxylin/Eosin staining of transverse sections of E8.5 and E9.0 control and Setdb1^{END} embryos. The approximate positions of the sections are indicated in the schematic. The black rectangle marks the region used for magnification. Red arrows indicate the hindgut region. The black arrowheads mark the neural tube. Representative images from $n = 3$ per genotype and stage.

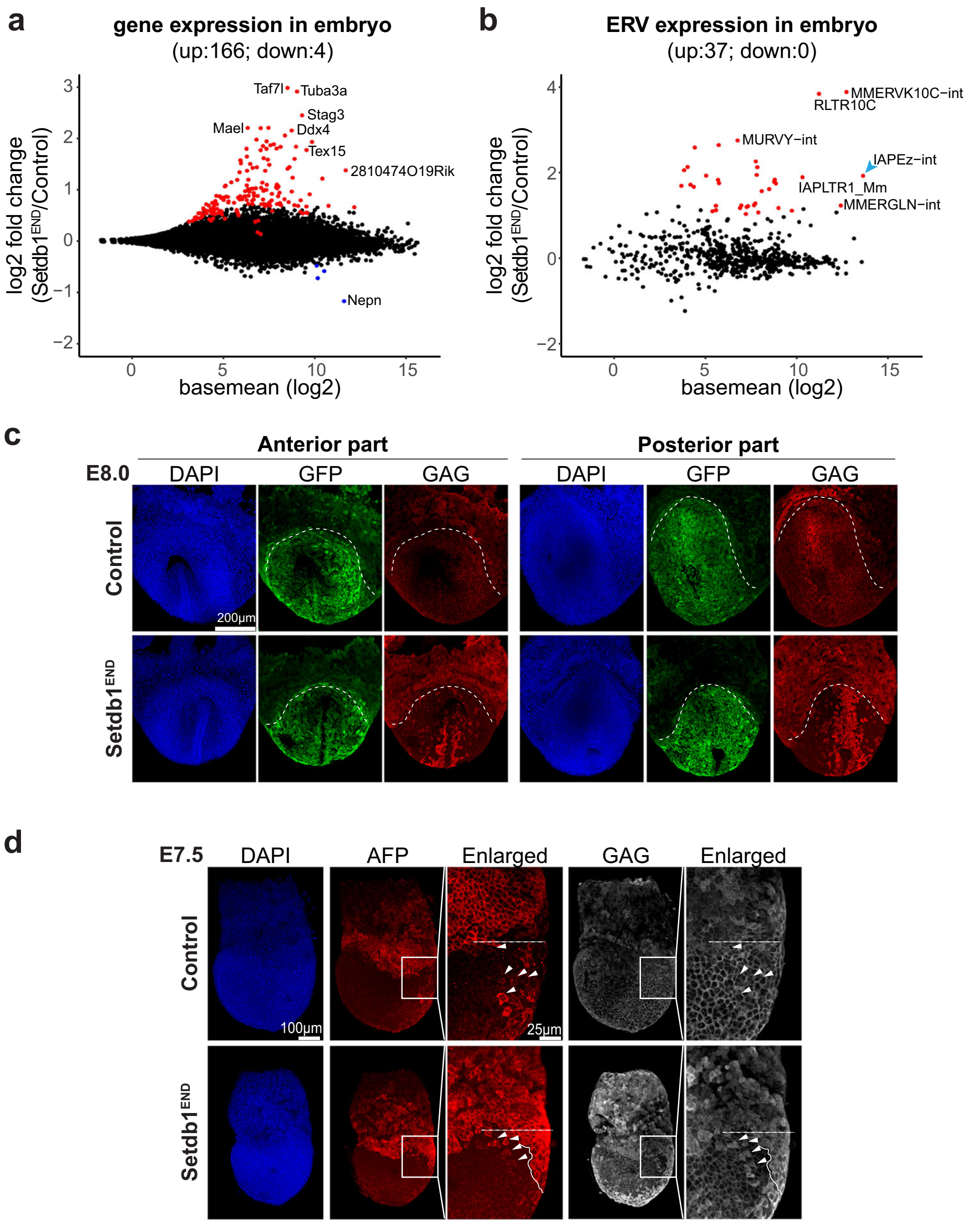


Fig. 2 | Loss of Setdb1 leads to selective de-repression of IAP elements in visceral endoderm cells. a Dot plot showing basemean expression vs. log2-fold change of protein coding genes in embryonic endoderm cells. Genes with significantly changed expression (adjusted p value <0.01; $n = 3$ for each condition) are colored (red = increased expression in Setdb1^{END} cells, blue = reduced expression in Setdb1^{END} cells). Selected genes are labeled. **b** Dot plot showing basemean expression vs. log2-fold change of ERV families in embryonic endoderm cells. ERV families with significantly changed expression (Wald test with Benjamini–Hochberg correction, adjusted p value <0.01, fold change >2; $n = 3$ for each condition) are colored (red = increased expression in Setdb1^{END} cells, blue = reduced expression in Setdb1^{END} cells). Selected ERV families are labeled. **c** Whole mount immunostaining of control and Setdb1^{END} embryos using GFP (to detect Cre reporter activity) and IAP-GAG antibodies. Strong expression of IAP-GAG can only be detected in a subpopulation of endoderm cells. Dashed lines indicate the border between extra-embryonic and embryonic part. Representative images from $n = 3$ per genotype and stage. **d** Lateral view of E7.5 embryos stained with AFP (to mark visceral endoderm cells) and IAP-GAG antibodies (anterior to the left). The boxed regions indicate the positions of the enlargements. Dashed lines indicate the border between extraembryonic and embryonic part. White arrowheads indicate AFP expressing cells which are integrated in the embryonic endoderm region. In Setdb1^{END} embryos, these cells display clear IAP-GAG staining. Representative images from $n = 3$ per genotype and stage.

E7.5 (Fig. 2d, region below white dashed line, cells marked by arrowheads). Notably, we detected clear IAPEz GAG expression in these AFP positive cells, whereas most AFP negative DE cells did not display GAG expression. These data suggest that *Setdb1* knockout results in specific *IAPEz* de-repression in the visceral, but not definitive endoderm.

ERV de-repression in *Setdb1*-deficient visceral endoderm progenitors in vitro

To investigate the molecular mechanisms underlying *Setdb1*-dependent ERV de-repression in visceral vs. definitive endoderm cells, we employed an in vitro differentiation system (Fig. 3a). ESCs from control and *Setdb1^{END}* mice were stimulated with Wnt3a and Activin to induce definitive endoderm cells[36]. *Gata6* overexpression in these ESCs triggered differentiation to extraembryonic endoderm (XEN), a progenitor stage of visceral endoderm cells[37]. Both, DE and XEN cells show expression of the *EGFP Cre* reporter (Supplementary Fig. 3a) and efficiently delete *Setdb1* (Supplementary Fig. 3b, c). In support of our hypothesis, that *Setdb1*-dependent de-repression of ERVs mainly occurs in the extra-embryonic endoderm lineage, we observed strong expression of IAP GAG in *Setdb1^{END}* XEN cells, but not in *Setdb1^{END}* DE

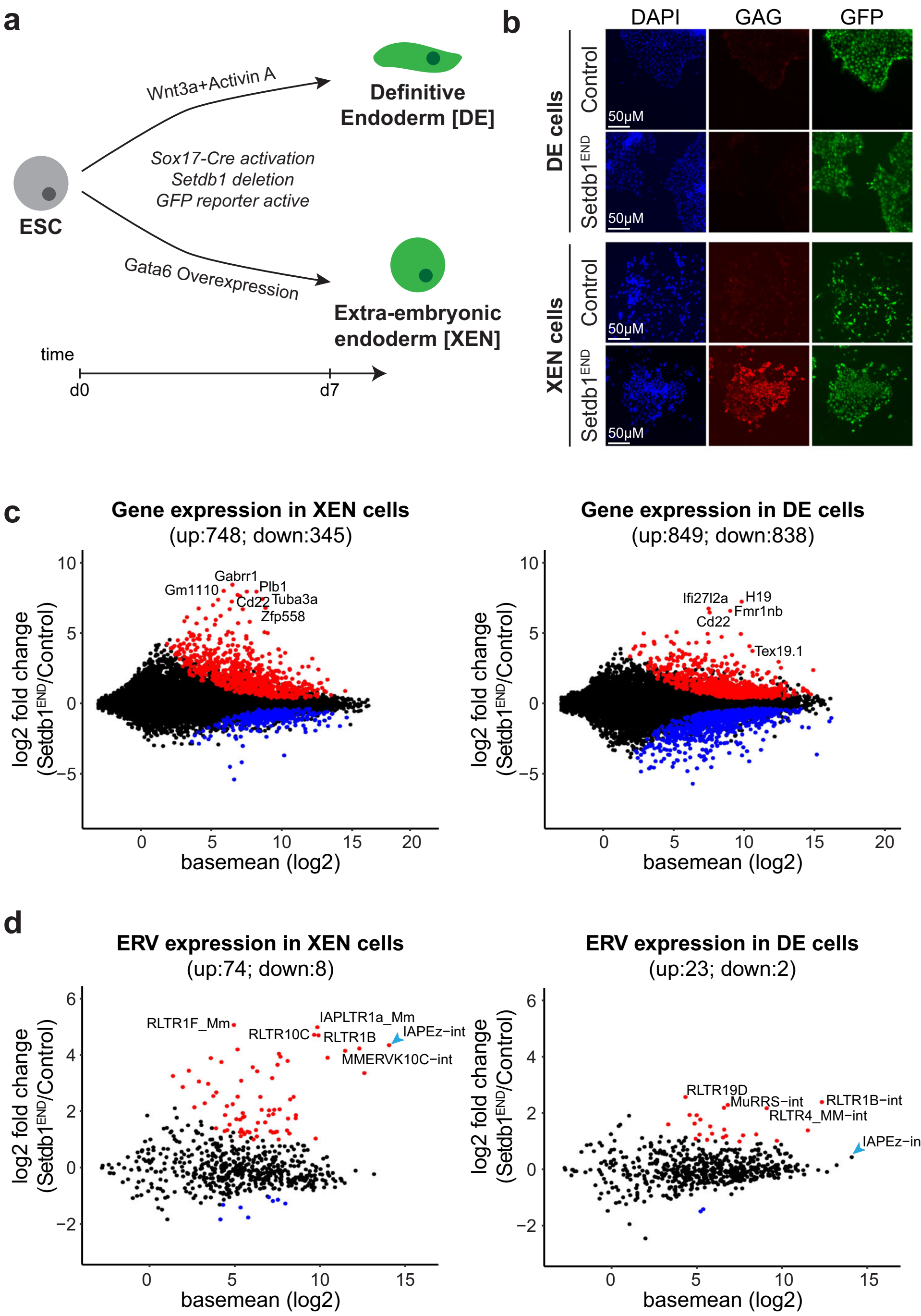


Fig. 3 | Setdb1 is critical for IAP silencing in visceral endoderm progenitors, but not in definitive endoderm cells. a Schematic of in vitro endoderm differentiation of control and Setdb1^{END} ESCs. Definitive endoderm cells were generated by Wnt3a/Activin stimulation. Visceral endoderm progenitor (XEN) cells were generated by overexpression of Gata6. **b** Immunofluorescence staining of in vitro differentiated control and Setdb1^{END} endoderm cells. Cells were stained after 7 days of in vitro differentiation using IAP-GAG (marks IAP de-repression) and GFP (marks Cre reporter activity) antibodies. Strong IAP de-repression is only observed in Setdb1^{END} XEN cells. Representative images from $n = 3$ experiments. **c** Dot plot showing basemean expression vs. log2-fold change of protein coding genes in in vitro differentiated control vs. Setdb1^{END} XEN and DE cells. Genes with significantly changed expression (Wald test with Benjamini–Hochberg correction, adjusted p value < 0.01; $n = 3$ for each condition) are colored (red = increased expression in Setdb1^{END} cells, blue = reduced expression in Setdb1^{END} cells). Selected genes are labeled. **d** Dot plot showing basemean expression vs. log2-fold change of ERV families in in vitro differentiated control vs. Setdb1^{END} XEN and DE cells. ERV families with significantly changed expression (Wald test with Benjamini–Hochberg correction, adjusted p value < 0.01; $n = 3$ for each condition) are colored (red = increased expression in Setdb1^{END} cells, blue = reduced expression in Setdb1^{END} cells). Selected ERV families are labeled.

cells (Fig. 3b). To characterize *Setdb1* dependent transcriptional changes, we isolated differentiated control and *Setdb1* deficient DE and XEN cells for RNA-seq. PCA analysis revealed clear clustering of ESCs, DE and XEN cells (Supplementary Fig. 3d). Examination of control genes for ESCs and endoderm markers confirmed efficient differentiation of both control and *Setdb1^{END}* cells (Supplementary Fig. 3e). We detected many up- and down-regulated genes in both *Setdb1^{END}* XEN and DE cells (Fig. 3c, Supplementary Data 3, 4).

When we analyzed ERV expression changes, we could detect strong de-repression of ERV classes in XEN cells, but only minor expression changes in DE cells (Fig. 3d, Supplementary Data 5, 6). *LINE* elements did not display strong expression changes (Supplementary Fig. 3f). *IAPEz* elements which we found de-repressed in visceral endoderm cells in *Setdb1*^{END} embryos did show strong de-repression in *Setdb1*^{END} XEN cells, but no change was observed in *Setdb1*^{END} DE cells (Fig. 3d, Supplementary Fig. 3g), even at a longer time period post *Setdb1* deletion (Supplementary Fig. 3h). Thus, in vitro differentiated XEN and DE cells reproduce the differential requirements for *Setdb1*-dependent ERV silencing, as observed in *Setdb1*^{END} embryos.

Impaired H3K9me3 and DNA methylation on IAPEz elements specifically in *Setdb1*^{END} XEN cells

Next, we aimed to investigate whether selective changes in repressive chromatin modifications might explain the differential response of XEN or DE cells to *Setdb1* loss. To investigate this question, we generated H3K9me3 ChIP-seq data from control and *Setdb1*^{END} XEN and DE cells. The appearance of H3K9me3 distribution was strikingly different in DE vs XEN cells. In particular, we detected a number of large megabase-size H3K9me3 domains in DE cells, which were not present in XEN cells (Supplementary Fig. 4). In *Setdb1*^{END} DE cells, most of these regions were not compromised, suggesting that other H3K9me3-specific HMTases maintain these regions. Interestingly, in XEN cells, deletion of *Setdb1* resulted in appearance of large H3K9me3 domains, which were not present in control DE or XEN cells (Supplementary Fig. 4). Together these data suggest that alterations in the balance of H3K9me3 specific HMTases can lead to large-scale changes in the genome-wide distribution of this modification. To investigate to which extent changes in H3K9me3 would relate to gene expression changes, we identified peaks which lose H3K9me3 in *Setdb1*^{END} cells. In *Setdb1*^{END} DE cells, we detected 722 peaks with lost H3K9me3 signal, 84 of which occurred in the vicinity of regulated genes (Supplementary Fig. 5a). Only a small set of genes was upregulated and would suggest a repressive role for H3K9me3. For example, *Triml2* is marked by *Setdb1*-dependent H3K9me3 in DE cells and loss of H3K9me3 coincided with de-repression of *Triml2*. In XEN cells, *Triml2* was not modified by H3K9me3 and expression did not change between control and *Setdb1*^{END} cells (Supplementary Fig. 5b). In XEN cells, we detected 17438 peaks with *Setdb1*-dependent H3K9me3, of which 614 were in the vicinity of regulated genes (Supplementary Fig. 5a). The majority of H3K9me3 marked genes was upregulated in *Setdb1*^{END} cells, suggesting a role for *Setdb1* in gene repression. For example, *Gabrr1* is H3K9me3 modified in both DE and XEN cells, but only in XEN cells, we detected loss of H3K9me3 and de-repression of *Gabrr1* (Supplementary Fig. 5c).

Next, we investigated H3K9me3 changes specifically on ERV families. In *Setdb1*^{END} XEN cells, we observed reduced H3K9me3 levels in several ERV families, including many upregulated ERVs (Fig. 4a, left panel). In contrast, H3K9me3 was reduced on very few ERV families in *Setdb1*^{END} DE cells (Fig. 4a, right panel). H3K9me3 was unaltered on *LINE* elements in both *Setdb1*^{END} XEN and DE cells (Supplementary Fig. 6a). Cumulative coverage analysis on *IAPEz* elements revealed that H3K9me3 was completely lost in *Setdb1*^{END} XEN cells, whereas no difference could be observed in *Setdb1*^{END} DE cells (Fig. 4b). These data suggest that *Setdb1* is the major H3K9me3 HMTase for *IAPEz* elements in XEN cells and, that other HMTases compensate for the loss of *Setdb1* in DE cells to maintain H3K9me3 on *IAPEz* elements and other ERVs.

Reduced H3K9me3 often correlates with reduced maintenance of DNA methylation[8,17]. To determine if reduced H3K9me3 in *Setdb1*^{END} XEN cells would compromise DNA methylation, we measured DNA methylation levels specifically on *IAPEz* and *LINE1* elements using locus-specific bisulfite sequencing. Control ESCs as well as XEN and DE cells display high DNA methylation levels across *IAP-LTR* and *IAP-GAG* regions (Fig. 4c). *LINE1* elements only showed moderate DNA methylation in ESCs and XEN cells, but full methylation in DE cells (Supplementary Fig. 6b). Upon loss of *Setdb1*, DNA methylation is only affected in XEN cells, where we detected reduced levels across the *IAP-GAG* region (Fig. 4c). These data agree with current models that maintenance of DNA methylation on repressive chromatin regions is coupled with the presence of H3K9me3[15–17]. Further, our data suggest that reduced H3K9me3 and DNA methylation allow higher transcriptional activity of ERVs in *Setdb1*^{END} XEN cells. It is interesting to note that DNA methylation reduction on *LINE1* elements does not coincide with strongly reduced H3K9me3 (Supplementary Fig. 6). Perhaps, *Setdb1* is required for recruitment of de novo methylation by *Dnmt3a/b* during differentiation and other H3K9me3 HMTases could deposit H3K9me3 in absence of *Setdb1* on these elements.

In DE cells, SETDB1 localizes to *IAP* elements (Supplementary Fig. 7a), indicating that SETDB1 could mediate H3K9me3 in this cell type. However, loss of *Setdb1* in DE cells even for extended time periods did not result in noticeable *IAPEz* de-repression (Supplementary Fig. 7b). In vitro differentiation allows DE cells to proliferate until around day 14 (Supplementary Fig. 7c), which would allow for passive loss of H3K9me3 upon *Setdb1* deletion, but we could not detect reduced H3K9me3 on *IAPEz* elements at day 12 or day 14 (Supplementary Fig. 7d). These data suggest that H3K9me3 is maintained by other histone methyltransferases. We assessed whether *Suv39h* enzymes would be responsible for H3K9me3 deposition in DE cells. In *Suv39h* dko DE cells, we detected largely unaltered H3K9me3, DNA methylation and ERV transcription (Supplementary Fig. 8), suggesting a minor role for IAP regulation. However, it is still possible that *Suv39h* enzymes could compensate for the loss of *Setdb1*, or that other H3K9me3 HMTases could mediate H3K9me3 in DE.

Loss of *Dnmt1* leads to ERV de-repression in both DE and XEN cells in presence of H3K9me3

To investigate if maintenance of repressive chromatin in *Setdb1*^{END} DE cells prevents ERV de-repression we used *Dnmt1* knock-out ESCs to study the effect of impaired DNA methylation on ERV activity. *Dnmt1* ko ESCs and genetic background matched wildtype ESCs were in vitro differentiated to XEN and DE cells, respectively. We then performed RNA-seq analysis to determine transcriptional changes in *Dnmt1* ko XEN and DE cells. PCA analysis showed clear clustering of ESCs, DE and XEN cells (Supplementary Fig. 9a). Expression of specific marker genes revealed that *Dnmt1* ko ESCs efficiently differentiate to XEN and DE cells (Supplementary Fig. 9b). Transcriptional changes of coding genes between *Dnmt1* ko XEN and DE cells were observed (Supplementary Fig. 9c, Supplementary Data 7–9), with little overlap to transcriptional changes observed in *Setdb1*^{END} cells. We then investigated transcriptional changes of ERV and LINE families in response to *Dnmt1* ko (Supplementary Fig. 9d, e, Supplementary Data 10–12). Undifferentiated *Dnmt1* ko ESCs did not show elevated *IAPEz* expression (Supplementary Fig. 9d), in agreement with previous studies[21]. Importantly, loss of *Dnmt1* resulted in strongly reduced DNA methylation (Supplementary Fig. 10) and *IAPEz* de-repression in both XEN and DE cells (Supplementary Fig. 9d). These data demonstrate that DNA methylation is critical for ERV silencing in both endoderm lineages. To test if upregulated *IAPEz* expression in *Dnmt1* ko endoderm cells would be due to impaired H3K9me3, we performed ChIP-seq analyses for this modification in control and *Dnmt1* ko ESCs, XEN and DE cells. In ESCs, H3K9me3 was not affected on IAP regions (Fig. 5a, b), suggesting that maintained H3K9me3 could support *IAP* silencing in ESCs. However, in both *Dnmt1* ko XEN and DE cells H3K9me3 was also maintained on *IAP* sequences, although *IAPEz* elements were strongly de-repressed (Fig. 5a, b). These data were supported by ChIP-qPCR analyses for H3K9me3 in wild type and *Dnmt1* ko XEN and DE cells, where we failed to detect striking changes in H3K9me3 on *IAPEz* regions, although other control regions, such as *H19* and *Polrmt* could display reduced signals (Supplementary Fig. 9f). Since *IAPEz* elements show little polymorphisms, small read mapping to unique elements is challenging and we cannot be sure which *IAPEz* insertions become transcriptionally active while maintaining H3K9me3 in *Dnmt1* ko DE

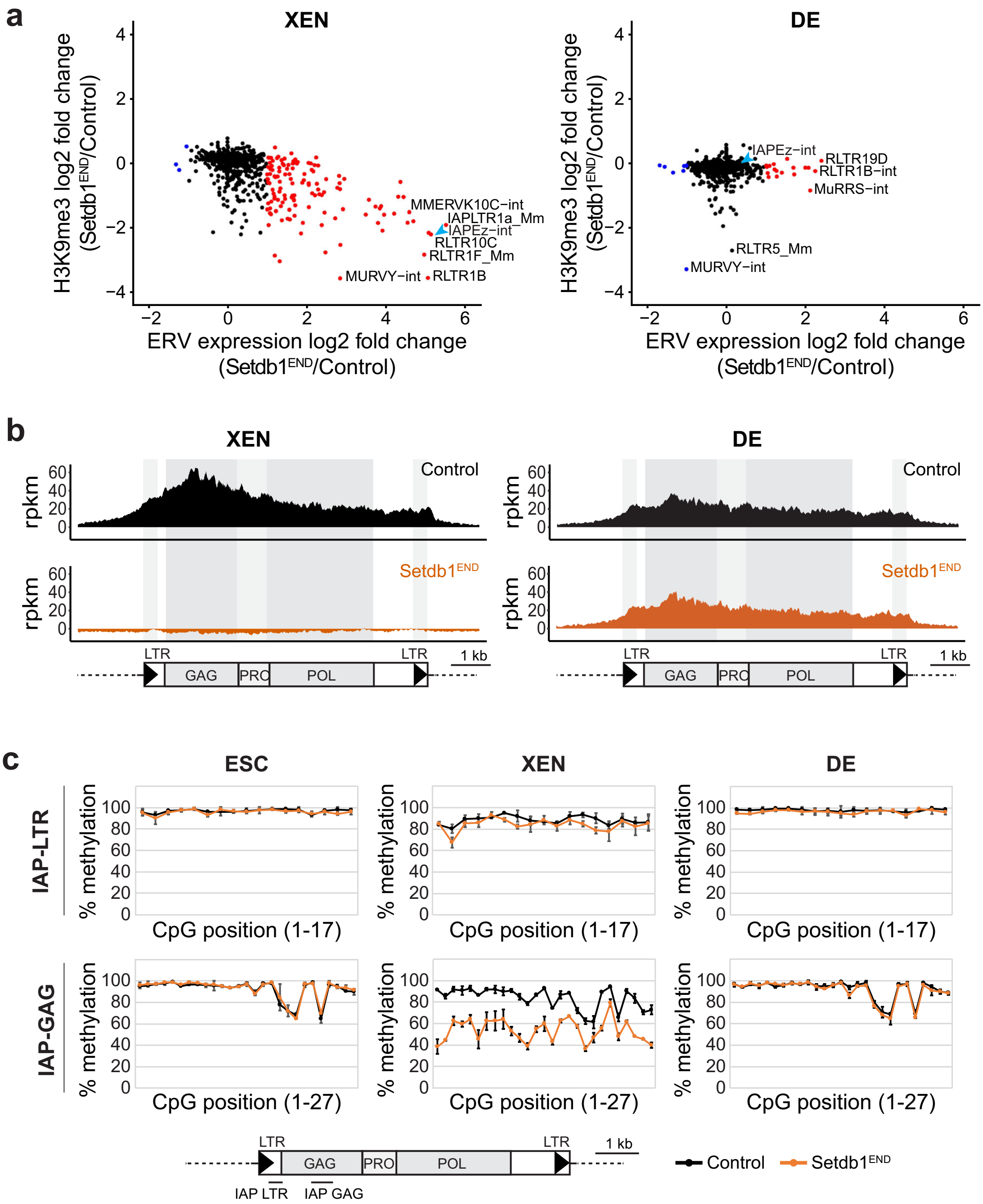


Fig. 4 | Setdb1 is critical to maintain H3K9me3 and DNA methylation on ERVs in XEN, but not DE cells. a Dot plot showing expression vs. H3K9me3 changes on ERV families between control and Setdb1^{END} XEN and DE cells. ERV families with significantly changed expression (fold change > 2; $n = 3$ for each condition) are colored (red = increased expression in Setdb1^{END} cells). Selected ERV families are labeled. **b** Cumulative H3K9me3 ChIP-seq coverage across IAP elements in control and Setdb1^{END} XEN and DE cells. The structure of IAP elements is shown schematically. (rpkm = reads per kilobase per million of reads). **c** Bisulfite-PCR analysis for DNA methylation in IAP-LTR and IAP-GAG regions. Positions of the PCR products are indicated in the schematic. Plots display the percentages of DNA methylation in individual CpG positions of IAP-LTR and IAP-GAG PCR fragments. DNA methylation analysis was performed in control and Setdb1^{END} ESCs, XEN and DE cells. Error bars depict standard deviation ($n = 2$; 500 sequences each). Source data are provided as a Source Data file.

and XEN cells. However, some *IAPEz* elements lack proper transcriptional termination at their 3′LTRs and, H3K9me3 presence can be detected in uniquely mapping regions neighboring the *IAPEz* insertion. Using this approach, we could identify examples of individual *IAPEz* elements that displayed significant transcriptional activity only in *Dnmt1* ko DE or XEN cells while maintaining H3K9me3 (Fig. 5c). Based on these data we conclude that DNA methylation is critical for IAP repression in endoderm lineages and, that the presence of H3K9me3 is not sufficient to establish a repressive chromatin environment across these elements.

Discussion

In this study, we have delineated the roles of H3K9me3 and DNA methylation for ERV regulation upon early embryonic vs. extra-embryonic endoderm development and differentiation. Our data demonstrate ontogenesis-dependent regulatory mechanism for IAP

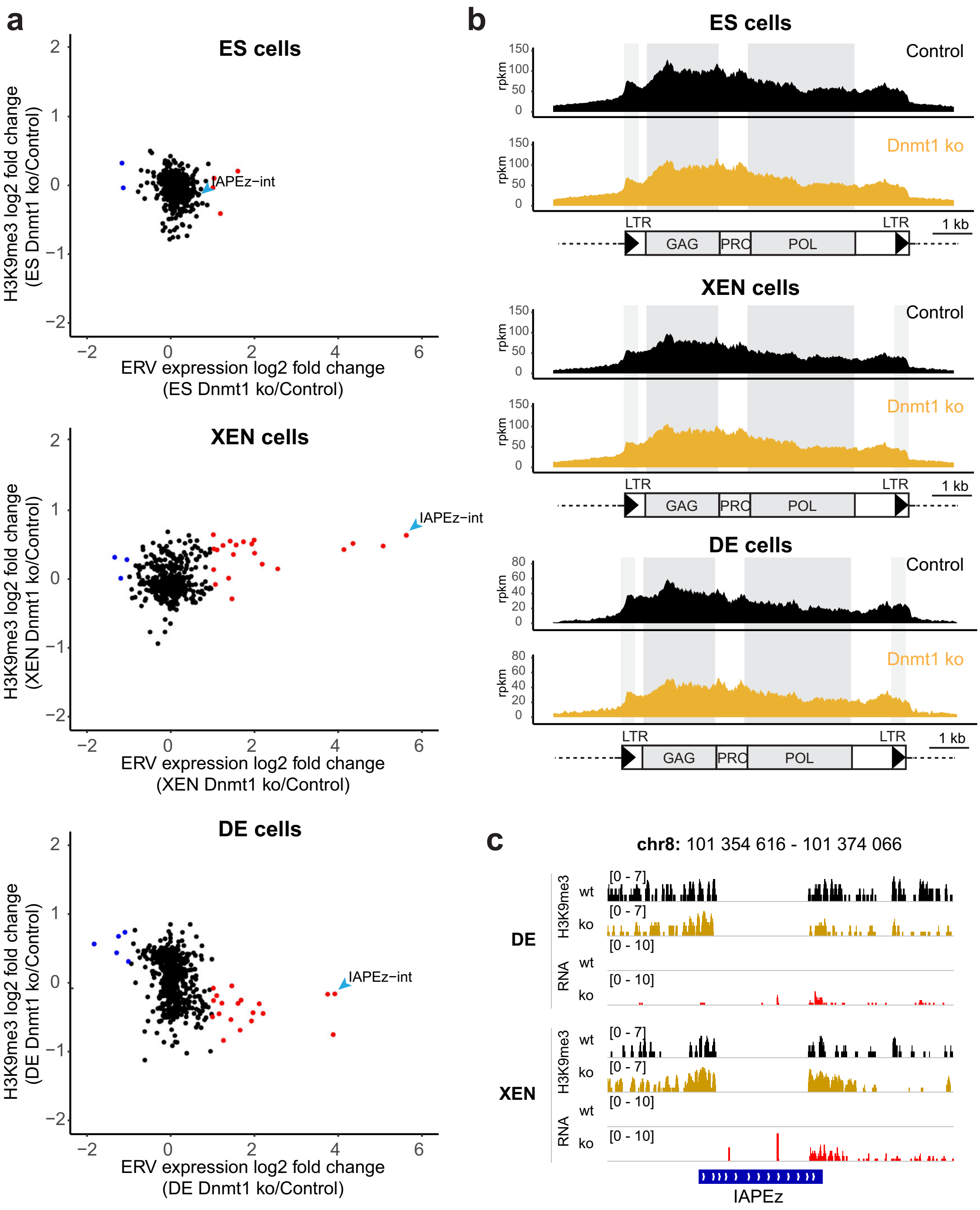


Fig. 5 | Loss of Dnmt1 leads to IAP de-repression in both XEN and DE cells, while H3K9me3 is maintained. a Dot plot showing basemean expression vs. H3K9me3 changes of ERV families in ESCs and in vitro differentiated wild type (J1) vs. Dnmt1 ko XEN and DE cells. ERV families with significantly changed expression (Wald test with Benjamini–Hochberg correction, adjusted p value < 0.01, fold change >2; $n = 2$ for each condition) are colored (red = increased expression in Dnmt1 ko cells, blue = reduced expression in Dnmt1 ko cells). Selected ERV families are labeled. **b** Cumulative H3K9me3 ChIP-seq coverage across IAP elements in control and Dnmt1 ko ES, XEN and DE cells. The structure of IAP elements is shown schematically. (rpkm = reads per kilobase per million of reads). **c** Improper transcriptional termination allows identification of individual IAPEz integrations with detectable expression in an H3K9me3 context. Genomic screenshot of an IAPEz integration with H3K9me3 coverage in uniquely mapping regions bordering the IAPEz sequence. In Dnmt1 ko DE and XEN cells, expression from this IAPEz element can be detected by presence of RNA in the uniquely mapping 3′ region of this element, which is likely due to improper transcriptional termination in the 3′LTR. H3K9me3 is unaltered in this region in Dnmt1 ko cells.

silencing in early endoderm development. In visceral endoderm cells, *Setdb1* is crucial to maintain H3K9me3 and DNA methylation on *IAP* elements. In definitive endoderm cells, H3K9me3 is not lost upon *Setdb1* deletion, suggesting compensation by other histone methyltransferases. This would be in agreement with a previous report that demonstrated reduced H3K9me3 only upon triple deletion of *Setdb1*, *Suv39h1* and *Suv39h2* in definitive endoderm-derived liver cells[38]. We could not detect reduced H3K9me3 and IAP-GAG expression in *Suv39h* dko definitive endoderm cells. This would suggest that *Suv39h* enzymes do not play a major role for mediating H3K9me3 on *IAP* in early definitive endoderm. However, because of the maintained H3K9me3 on *IAP* in *Setdb1*^{END} DE cells, *Suv39h* enzymes or other histone

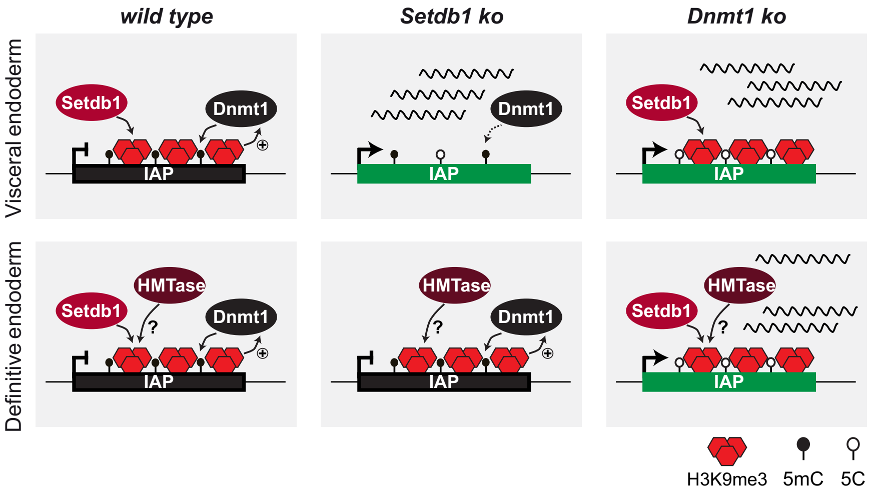


Fig. 6 | **Model.** Schematic depicts a dominant role of DNA methylation over H3K9me3 for ERV silencing in endoderm differentiation. Depletion of Setdb1 leads to augmented H3K9me3 and DNA methylation and impaired ERV silencing in visceral endoderm but not in definitive endoderm cells. In contrast, depletion of Dnmt1 impairs ERV repression in both visceral endoderm and definitive endoderm cells, although H3K9me3 is maintained. Together, our data suggest redundant cell type specific H3K9me3 maintenance pathways and a dominant role of DNA methylation for IAP silencing in endoderm cells.

methyltransferases (*Setdb2*, *Prdm* enzymes) may compensate the loss of *Setdb1*.

Our data suggest a dominant role of DNA methylation over H3K9me3 for *IAP* silencing (Fig. 6). Loss of DNA methylation resulted in *IAP* de-repression in both visceral and definitive endoderm. Surprisingly, H3K9me3 on *IAP* elements was maintained under these conditions. Thus, our data demonstrate that H3K9me3 is not sufficient for ERV silencing in endoderm-derived cells. Based on the reduced DNA methylation upon *Setdb1* deletion in XEN cells and other cell types[8,16,17] we speculate that one important role of *Setdb1*/H3K9me3 is to maintain high levels of DNA methylation. Currently, two mechanisms are being discussed: (1) H3K9me3 can be recognized by UHRF1 resulting in recruitment of DNMT1[18,19], although this view is challenged by largely maintained DNA methylation in *Uhrf1* mutants lacking the H3K9me3 binding domain[20]. (2) *Setdb1*/H3K9me3 may protect ERVs from access of TET enzymes which would remove DNA methylation[16]. The function of DNA methylation in ERV silencing is currently not fully understood on the mechanistic level. It is possible that transcription factors which could activate ERVs bind DNA in a methylation-sensitive manner[39]. DNA methylation may also help to facilitate establishment of other chromatin modifications, or recruitment of repressive chromatin binding factors. Future studies are required to better clarify the roles of H3K9me3 and DNA methylation in this context.

A surprising finding of our study was that *IAP* de-repression upon loss of *Setdb1* was limited to extraembryonic endoderm cells in vivo. We showed this by co-staining of IAP-GAG with AFP as marker for extraembryonic endoderm cells in E7.5 embryos. During this time extraembryonic endoderm-derived cells integrate into the embryonic part of definitive endoderm and, upon integration, assume very similar transcriptional and phenotypic properties[40]. As we did not use a lineage reporter to follow extraembryonic endoderm-derived cells in later embryonic stages we cannot definitely state that all IAP-GAG expressing cells are of extraembryonic origin. However, our in vitro differentiation experiments strongly support the differential mode of ERV regulation in definitive vs. extraembryonic endoderm cells. The mechanistic basis for this differential mode of regulation remains obscure. Differences in expression of chromatin regulators or different transcription factor networks are unlikely to contribute. Recently, single cell RNA-seq analyses of endoderm cells from different embryonic stages revealed an almost indistinguishable expression pattern of cells derived from embryonic or extraembryonic origin[41]. As no significant transcriptional differences appear to exist between embryonic vs extraembryonic endoderm-derived cells in vivo, we hypothesize that the chromatin composition of ERVs established during endoderm differentiation is epigenetically maintained. More experiments are needed to determine the exact chromatin composition of ERVs in different cell types and to understand the cell type-specific recruitment of chromatin modifying factors. The mixed ontogeny of endoderm cells provides a unique opportunity to study the role and the mechanisms of epigenetic inheritance in different developmental stages and during aging in an in vivo model.

Limitations of this study

Setdb1 depletion in in vitro DE differentiation has minor effects on transcription of *IAPEz* elements and H3K9me3 is largely maintained. The intricacies of the DE differentiation protocol do not allow monitoring the fate of individual differentiated cells. In particular, we cannot determine how many cell divisions after *Setdb1* depletion have passed. It is, therefore, possible that even after extensive DE differentiation (day 14), cells have not undergone enough replication rounds to passively lose all *Setdb1*-dependent H3K9me3. We may therefore underestimate the role of *Setdb1* in establishing H3K9me3 in DE cells.

Methods

Generation of conditional Setdb1 knock-out strains

Setdb1$^{ß\text{-}gal}$ mice (*Setdb1*$^{tm1a(EUCOMM)Wtsi}$) were crossed with actin-Flp recombinase mice[42] to remove the ß-gal cassette, resulting in the *Setdb1*^{flox} allele. To induce the specific deletion of *Setdb1* in endoderm cells, *Setdb1*^{flox} was combined with the *Sox17*$^{2A\text{-}iCre}$ mouse line[28]. To monitor *Sox17-2A-iCre* activity, the *CAG-CAT-EGFP* reporter allele was introduced[43].

Housing of mice was performed in the BMC animal core facility which is licensed by local authorities (Az. 5.1-5682/LMU/BMC/CAM, approved on 02-12-2017 by Landratsamt München) following the regulations of German Law (TierSchG, BGBl. I S. 1206, 1313).

Establishment of ESC lines and cell culture conditions

Blastocysts were isolated at 3.5 dpc and cultured on feeder coated 48 well plates with ESC medium containing MEK-1 inhibitor (PD098059; New England Biolabs). After 5-7 days, the inner cell mass outgrowth was trypsinized and transferred to larger feeder-coated plates. Established ESC lines were split every 2 days.

MEF and ESCs were cultured in medium based on DMEM (D6429, Sigma) containing 15% FCS (F7542 Sigma), non-essential amino acids

(M7145, Sigma), Penicillin/Streptomycin (P4333; Sigma), and 2-mercaptoethanol (Gibco, 31350-010). For ESC culture, the medium was supplemented with leukemia inhibitory factor (LIF).

Whole mount in situ hybridization

In situ hybridization of whole-mount embryos was performed as previously described[44].

Whole mount embryo immunostaining

Embryos were isolated in PBS$^+$ dissection medium [PBS containing Mg^{2+} and Ca^{2+}]. Isolated embryos were fixed for 20 min at RT in 2% PFA in PBS$^+$ followed by permeabilizing for 10-15 min in permeabilization solution [0.1 M glycine/0.1% Triton X-100]. Embryos were transferred into blocking solution [0.1% Tween-20; 10% FCS; 0.1% BSA; 3% Rabbit, Goat or Donkey serum]. Primary antibodies were added into the blocking solution and incubated o/n at 4 °C. The following antibodies were used: Foxa2 (Abcam, ab40874, 1:1000), Sox17 (Acris/Novus, GT15094,1:1000), GFP (Aves, GFP1020, 1:1000), AFP (R&D, AF5369, 1:1000) and IAP-GAG (Cullen lab, 1:1000). The next day, embryos were kept at RT for 2 hours. After 3 washes with PBST, embryos were incubated with secondary antibodies Donkey anti rabbit Alexa488 (Jackson Immuno Research, 711-545-152, 1:800), Donkey anti goat Alexa 555 (Invitrogen, A-21432, 1:1000), Donkey anti Chicken IgY Alexa488 (Jackson Immuno Research, 703-545-155, 1:800), Donkey anti mouse Cy3 (Jackson Immuno Research, 715-165-150, 1:800), Donkey anti rabbit Alexa647 (Jackson Immuno Research, 711-605-152, 1:500) for 3 hours at RT, followed by three washes. Embryos were then embedded in antifade medium (Invitrogen, P36930) for microscopy analysis.

β-galactosidase staining and histology

Embryos were fixed with 4% paraformaldehyde/PBS at 4 °C [E7.5/5 min; E8.5/10 min; E9.5/20 min]. After washes with LacZ rinse solution [2 mM MgCl2; 0.02% NP-40; 0.01% sodium deoxycholate in PBS], embryos were stained with X-gal staining solution [1 mg/ml dimethylformamide; 5 mM potassium ferricyanide; 5 mM potassium ferrocyanide in LacZ rinse solution] o/n at 37 °C.

For histological sections, embryos were fixed overnight in 4% formaldehyde and embedded in paraffin. The embedded embryos were sectioned and stained with Hematoxylin and Eosin.

In vitro endoderm differentiation

For in vitro differentiation towards definitive endoderm 0.1 Mio ESCs were seeded on FCS coated 6-well plates directly into endoderm differentiation medium (EDM) [500 ml Advanced DMEM/F-12 (1x) (Gibco/LifeTechnologies; 12634-10- 500 ml), 500 ml Advanced RPMI 1640 (1x) (Gibco/LifeTechnologies; 12633-012- 500 ml), 22 ml GlutaMAXTM – I CTSTM (Gibco/LifeTechnologies; 12860-01- 100 ml), 200 µl AlbuMAX 100 mg/ml (Gibco/LifeTechnologies; 11021-029 100 g, 22 ml HEPES 1 M (Gibco/LifeTechnologies; 15630-056- 100 ml), 70 µl Cytidine 150 mg/ml (SIGMA; C4654-5G), 0,9 ml ß-Mercaptoethanol 50 mM (Gibco/LifeTechnologies; 31350-10- 20 ml), 12 ml Pen/Strep (10000U/ml) (Gibco/LifeTechnologies; 10378016 – 100 ml), 1 ml Insulin-Transferin-Selenium Ethanolamine (Gibco/LifeTechnologies; 51500-056- 10 ml)], supplemented with 2 ng/ml of murine Wnt3a (1324 WN-CF, R&D systems) and 10 ng/ml of Activin A (338-AC, R&D systems). Cells were collected on day 7 for FACS isolation.

For in vitro differentiation towards extraembryonic endoderm, 0.2 million ESCs were seeded on gelatine coated 6-well plates directly in ESC medium and then were transduced with a lentiviral *Gata6* overexpression construct (#1582 pLenti6/EF1a-GATA6-IRES-Puro) on the next day. Two days after transduction, *Gata6* expressing cells were selected with 1 µg/ml puromycin. Five days after transduction, ESC medium was replaced with XEN medium [Advanced RPMI 1640 (1x) (Gibco/LifeTechnologies; 12633-012- 500 ml), supplemented with 15% FCS, 0.1 mM β-mercaptoethanol and 1% penicillin-streptomycin]. Cells were collected on day 7 for FACS isolation.

For the extended DE differentiation time-course cells were harvested at day 8 of differentiation. DE cells were detached with StemPro™ Accutase™ Cell Dissociation Reagent (Thermofisher) by incubation for 3 minutes at 37 °C. For later time-points cells were reseeded onto Biotechne Cultrex Stem Cell Qualified Reduced Growth Factor BME coated dishes and harvested at day 10, 12 and 14.

Cell cycle analysis

Cell cycle analysis of DE cells was performed using Hoechst 33342 Ready-Flow Reagent (Thermo Fisher Scientific) at 1 drop per 0.5 ml of cell suspension, incubated at 37 °C for 10 min. Cells were analyzed on FACSAriaFusion SORP or LSRFortessa SORP (both BD Biosciences) equipped with a UV laser (355 nm) for optimal excitation of Hoechst 33342. Dead cells were excluded using SYTOX™ Red Dead Cell Stain, for 633 or 635 nm excitation (ThermoFisher) in a 1:1000 dilution. Cell cycle profiles were analyzed with FlowJo v10.8.1 (BD). Doublets were excluded from analysis based on SSC-H versus -W and Hoechst-W versus -H plots.

Immunofluorescence microscopy

Cells were carefully washed once with PBS. Fixation was carried out with 3.7% formaldehyde (Carl Roth) in PBS for 10 min at RT. Cells were then permeabilized with 3 mM sodium citrate tribasic dehydrate (Merck), 0.1% v/v Triton X-100. Permeabilized cells were washed twice with PBS and twice in washing solution [PBS, 0.1% v/v Tween 20, 0.2% w/v BSA] for 5 min. Cells were blocked with blocking solution [PBS, 0.1 % v/v Tween 20, 2.5% w/v BSA] for 30 min and incubated overnight at 4 °C with primary antibodies in blocking solution. The following antibodies were used: GFP (Aves, GFP1020, 1:1000), IAP-GAG (Cullen lab, 1:1000). Cells were washed three times in washing solution for 10 min before incubation with secondary antibodies Donkey anti Chicken IgY Alexa488 (Jackson Immuno Research, 703-545-155, 1:800) and Donkey anti rabbit Alexa 555 (Invitrogen, A-31572, 1:1000) in blocking solution containing 10% normal goat serum (Dianova-Jackson Immuno Research) at RT for 1 h. After washing three times in PBS, 0.1% Tween 20 for 10 min, cells were embedded with Vectashield/DAPI (Vector Laboratories) on standard microscope slides (Carl Roth). The immunofluorescence staining was examined with Axiovert 200 M inverted microscope for transmitted light and epifluorescence (Carl Zeiss Microscopy) with the help of the AxioVision Special Edition Software (Carl Zeiss Microscopy).

Fluorescence activated cell sorting

Cells were resuspended in PBS/0.2% FCS before FACS collection. Cells from embryos were directly sorted into lysis buffer (Thermo Fisher, KIT0204) followed by RNA extraction. Cells from in vitro culture were sorted into PBS/0.2% FCS. FACS was performed using a FACS Aria instrument (BD Biosciences). Data were analysed using FlowJo software.

RT-qPCR analyses

Total RNA from three independent biological replicates of sorted cells was isolated using the RNA Clean & Concentrator kit (Zymo Research) including digestion of remaining genomic DNA according to producer ´s guidelines. qPCR was carried out with the Fast SYBR Green Master Mix (Applied Biosystems) in a LightCycler480 (Roche) according to the Fast SYBR Green Master Mix-protocol. Primers were evaluated for generating a single PCR product and for linear amplification in a wide range of DNA template dilutions. Every PCR-reaction was performed in a total volume of 10 µl in duplicates, triplicates or quadruplicates in a 384-well plate (Sarstedt). Two independent control genes (*Gapdh* and *HPRT*) were used as reference genes for qRT-PCR experiments and geometric mean of reference Ct values was used as normalization[45].

For qRT-PCR of repetitive regions like *IAP* elements, negative control samples that were not treated with reverse transcriptase were used to control for genomic DNA background. Ct-values were generated by the LightCycler480-Software (Roche) using the 2nd derivative max function and fold changes were calculated using the $2^{-\Delta\Delta Ct}$ method.

Western blot

Whole cell proteins extracts were prepared by resuspending 1 million cells in 40 μl of freshly prepared lysis buffer containing 50 mM Tris/HCl pH 7.5, 2% w/v SDS, 1% v/v Triton X-100, 1 mM PMSF, 0.5x Roche Complete Protease Inhibitor Cocktail. Samples were vortexed for 10 s at max speed and boiled for 10 min at 95 °C. After incubation with 1 μl of Benzonase/ 2.5 mM $MgCl_2$ at 37 °C for 15 min, protein extracts were mixed with 12 μl 4x sample buffer (Roth) and boiled again for 5 min at 95 °C. The boiled protein extracts were separated through SERVAGel TG PRiME 4-12 % precast SDS Page (SERVA Electrophoresis) in running buffer 25 mM Tris, 200 mM glycine, 1% (m/v) SDS at RT for 1 h and 25 mA per gel. Gels were blotted onto methanol activated PVDF membranes in a wet-blotting chamber (Bio-Rad Laboratories) containing blotting buffer 50 mM Tris, 40 mM glycine, 10% v/v methanol, 5 μM SDS for 1.5 h at 4 °C. Membranes were incubated in blocking buffer 1x PBS, 2.5% w/v BSA and 2.5 % w/v milk at RT for 1 h under mild agitation. Blocked membranes were incubated with primary Ab in blocking buffer at 4 °C for 16 h. The antibodies used were Setdb1 (Santa Cruz, sc66884-X, 1:250) and α-Tubulin (Sigma, T5168, 1:1000). Membranes were washed 3 times with PBST buffer 1x PBS, 0.1% v/v Tween 20 for 20 min. Incubation with secondary Ab 680RD Goat anti-Mouse (LI-COR, 926-68070, 1:3000) and 800CW Goat anti-Rabbit (LI-COR, 926-32211, 1:3000) was done in blocking buffer at RT for 1.5 h. The probed membranes were washed 3 times in PBST for 20 min. Based on the detection method, Immobilon Western Chemiluminescent HRP Substrate (Merck Millipore) was used for ECL method and IRDye 800CW Secondary Antibodies for LI-Cor method. Chemoluminescence was detected in by ChemiDoc MP Imaging System with the Image Lab Software using ECL Western blot detection reagent (Amersham Biosciences) or by Li-Cor Odyssey Imaging System with the Image studio software.

RNAseq analysis

The Agilent 2100 Bioanalyzer was used to assess RNA quality and only high-quality RNA samples (RIN > 8) were further processed for cDNA synthesis using SMART-Seq v4 Ultra Low Input RNA Kit (Clontech cat. 634888) according to the manufacturer´s instruction. cDNA was fragmented to an average size of 200-500 bp in a Covaris S220 device (5 min; 4 °C; PP 175; DF 10; CB 200). Fragmented cDNA was used as input for library preparation using MicroPlex Library Preparation Kit v2 (Diagenode, cat. C05010012) and processed according to the manufacturer´s instruction. RNA samples from Dnmt1 ko cells were Ribo-depleted using the NEBNext rRNA Depletion Kit (Human/Mouse/Rat) (NEB #E6310) and, RNAseq libraries were generated using the NEBNext Ultra II Directional RNA Library Prep Kit for Illumina (NEB #E7760) according to the manufacturer's instructions. All libraries were quality controlled by Qubit and Agilent DNA Bioanalyzer analysis. Deep sequencing was performed on HiSeq 1500 system according to the standard Illumina protocol for 50 bp single end reads.

ChIP-seq of histone modifications

0.5 million FACS-sorted cross-linked cells (1% formaldehyde, 10 min RT) were lysed in 100 μl Buffer-B-0.5 (50 mM Tris-HCl, pH 8.0, 10 mM EDTA, 0.5% SDS, 1x protease inhibitors -Roche) and sonicated in a microtube (Covaris; 520045) using a Covaris S220 device until most of the DNA fragments were between 200-500 base pairs long (settings: temperature 4 °C, duty cycle 2%, peak incident power 105 Watts, cycles per burst 200). After shearing, lysates were centrifuged for 10 min, 4 °C, 12000 g and supernatant diluted with 400 μl of Buffer-A (10 mM Tris-HCl pH 7.5, 1 mM EDTA, 0.5 mM EGTA, 1% Triton X-100, 0.1% SDS, 0.1% Na-deoxycholate, 140 mM NaCl, 1x protease inhibitors-Roche). 150 μl of sonicated chromatin was then incubated 4 h at 4 °C on a rotating wheel with 3 μg of H3K9me3 antibody (Active Motif; 39161) conjugated to 10 μl of magnetic beads. Beads were washed four times with Buffer-A (10 mM Tris-HCl pH 7.5, 1 mM EDTA, 0.5 mM EGTA, 1% Triton X-100, 0.1% SDS, 0.1% Na-deoxycholate, 140 mM NaCl, 1x protease inhibitors - Roche) and once with Buffer-C (10 mM Tris-HCl pH 8.0, 10 mM EDTA). Beads were re-suspended in 100 μl elution buffer (50 mM Tris-HCl, pH 8.0, 10 mM EDTA, 1% SDS) and incubated 20 min at 65 °C. Supernatant was transferred to a new tube. Crosslink reversal of immunoprecipitated DNA was carried out overnight at 65 °C. Then 100 μl TE (10 mM Tris-HCl pH 8.0, 1 mM EDTA) was added, RNA was degraded by 4 μl RNase A (10 mg/ml) for 1 hour at 37 °C and proteins were digested with 4 μl Proteinase K (10 mg/ml) at 55 °C for 2 hours. Finally, DNA was isolated by phenol:chloroform:isoamyl alcohol purification followed by ethanol precipitation. Purified DNA was used as input for library preparation using MicroPlex Library Preparation Kit v2 (Diagenode, cat. C05010012) and processed according to the manufacturer´s instruction. Libraries were quality controlled by Qubit and Agilent DNA Bioanalyzer analysis. Deep sequencing was performed on HiSeq 1500 system according to the standard Illumina protocol for 50 bp single-end reads.

For the extended DE differentiation time-course cells (day 12 and day 14) ChIPseq was was done using the following protocol. Briefly, 50.000 FACS-sorted cross-linked cells (1% formaldehyde, 10 min RT) were lysed in 100 ul Buffer-B-0.3 (50 mM Tris-HCl, pH 8.0, 10 mM EDTA, 0,3%SDS, 1x protease inhibitors -Roche) and and sonicated in a microtube (Covaris; 520045) using a Covaris S220 device until most of the DNA fragments were between 200-500 base pairs long (settings: temperature 4 °C, duty cycle 2%, peak incident power 105 Watts, cycles per burst 200). After shearing, lysates were diluted with 1 volume of Dilution Buffer (1 mM EGTA 300 mM NaCl, 2% Triton x-100, 0.2% sodium deoxycholate, 1x protease inhibitors-Roche). Sonicated chromatin) was then incubated 4 h at 4 °C on a rotating wheel with 1 ug of H3K9me3 (Diagenode C15410193) antibody conjugated to 10 μl of Protein-G Dynabeads (Thermofisher). Beads were washed four times with Buffer-A (10 mM Tris-HCl, pH 7.5, 1 mM EDTA, 0.5 mM EGTA,1% Triton X-100, 0.1% SDS, 0.1% Na-deoxycholate, 140 mM NaCl, 1x protease inhibitors) and once with Buffer-C (10 mM Tris-HCl, pH 8.0, 10 mM EDTA). Beads were then incubated with 70 μl elution buffer (0.5% SDS, 300 mM NaCl, 5 mM EDTA, 10 mM Tris HCl pH 8.0) containing 2 μl of Proteinase K (20 mg/ml) for 1 hour at 55 °C and 8 hours at 65 °C to revert formaldehyde crosslinking, and supernatant was transferred to a new tube. Another 30 μl of elution buffer was added to the beads for 1 minute and eluates were combined and incubated with another 1 μl of Proteinase K for 1 h at 55 °C. Finally, DNA was purified with SPRI AMPure XP beads (Beckman Coulter) (sample-to-beads ratio 1:2). Purified DNA was used as input for library preparation with Thruplex DNA-seq kit (Takara, cat. R400674) and processed according to the manufacturer´s instruction. Libraries were quality controlled by Qubit and Agilent DNA Bioanalyzer analysis. Deep sequencing was performed on Illumina NextSeq device.

Oxidative bisulfite analysis

Genomic DNA was prepared using the DNEasy Blood and Tissue Kit (Qiagen) and subjected to bisulfite conversion using the EpiTect Bisulfite Kit (Qiagen) according to the manufacturer's protocol. Jumpstart Taq polymerase (Sigma Aldrich) was used to amplify the IAP GAG region, IAP LTR region and a 200 bp region of LINE-1. PCR primers for bisulfite-converted DNA were modified by adding Illummina adaptors for library preparation based on previous studies[15,46,47]. The gel-purified amplicons were indexed with index primers/universal PCR primers and Illumina P5/P7 primers. Before amplification, the DNA was purified with SPRI AMPure XP beads (sample-to-beads ratio 1:0.8). Libraries were checked for quality control and correct fragment length on a Bioanalyzer 2100 (Agilent) and concentrations were determined with Qubit dsDNA HS Assay Kit (Life Technologies). Sequencing was carried out on a MiSeq sequencer (2 × 300 bp and 2 × 250 bp paired end) with v3 chemistry (Illumina).

Intracellular Staining for IAP-GAG

ESCs and endoderm differentiated cells were resuspended in 500 μl PBS containing 2 μl zombie aqua (Biolegend, cat no.423101). Fixation/Permeabilization was performed in 1 ml of Foxp3 fixation/permeabilization buffer. After 30 min incubation at 4 °C in the dark, samples were washed with 2 ml of 1x permeabilization buffer and centrifuged at 400 g at RT for 5 min. The pellet was resuspended in 100 μl of 1x permeabilization buffer after a second wash and incubated with primary antibodies IAP-GAG (Cullen lab, 1:1000) and Sox17 (Acris/Novus, GT15094, 1:1000) for at least 30 minutes at room temperature in the dark. After washes with 2 ml of 1x permeabilization buffer, the pellet was resuspended in 100 μl of 1x permeabilization buffer and incubated with secondary antibodies anti-rabbit (Jackson; 711605152, 1:800) and anti-goat (Invitrogen; A21432, 1:1000) at room temperature for 60 min in the dark. After two washes with 2 ml of 1x permeabilization buffer, cells were resuspended in 300 μl of FACS buffer. ESCs and differentiated cells were then analyzed by flow cytometry using FACS Aria instrument (BD Biosciences). Data were further processed using the FlowJo v10 Software. FITC-A channel (Sox17) was used to distinguish endoderm differentiated from undifferentiated cells.

Reagents

Cell lines, antibodies, primers and plasmids are listed in the Supplementary Information.

Bioinfomatic analysis

RNA-seq. Single end reads were aligned to the mouse genome version mm10 using STAR[48] with default options "--runThreadN 32 --quantMode TranscriptomeSAM GeneCounts --outSAMtype BAM SortedByCoordinate". Read counts for all genes and repeats were normalized using DESeq2[49]. Significantly changed genes were determined through pairwise comparisons using the DESeq2 results function (adjusted p value < 0.01). The expression levels of different repeat classes was assessed using Homer through analyzeRepeats.pl with the repeats function. The repeat definitions were loaded from UCSC. Significantly changed ERV families were determined through pairwise comparisons using the DESeq2 results function (log2 fold change threshold = 1, adjusted p value < 0.01). PCA analyses were done using the plotPCA function of the DESeq2 package. Bargraphs showing expression data for selected genes were plotted using ggplot2 with RSEM-normalized data (TPM = Transcript Per Million). Heatmap with differentially expressed ERV families was plotted with pheatmap using rlog-normalized expression values.

ChIP-seq. ChIP-seq single end reads were aligned to the mouse genome mm10 using Bowtie with options "-q -n 2 --best --chunkmbs 2000 -p 32 -S". The H3K9me3 enrichment of different repeat classes was assessed using Homer through analyzeRepeats.pl with the repeats function. The repeat definitions were loaded from UCSC. Correlation of expression and H3K9me3 changes for ERVs were plotted by log2-foldchange of H3K9me3 enrichment over input versus log2foldchange of expression. Cumulative read coverage across IAP elements was calculated using coverageBed and normalized to the library size. Coverage profiles were plotted using ggplot2.

H3K9me3 domains were identified using chromstaR[50] with options binsize = 5000, stepsize = 1000 in mode "separate". Differential H3K9me3 peaks were detected with chromstaR using options binsize = 1000, stepsize = 500 in mode "differential".

Bisulfite sequencing. Bisulfite sequencing paired end reads were aligned using CLC Genomics Workbench. Methylation analysis of sequencing data was performed using QUMA: quantification tool for methylation analysis[51].

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The data that support this study are available from the corresponding author upon reasonable request. NGS data (Supplementary Data 13) were deposited on NCBI GEO database with accession number GSE139128. Source data are provided with this paper.

References

1. Wang, J. et al. Primate-specific endogenous retrovirus-driven transcription defines naive-like stem cells. *Nature* **516**, 405–409 (2014).
2. Macfarlan, T. S. et al. Embryonic stem cell potency fluctuates with endogenous retrovirus activity. *Nature* **487**, 57–63 (2012).
3. Lamprecht, B. et al. Derepression of an endogenous long terminal repeat activates the CSF1R proto-oncogene in human lymphoma. *Nat. Med.* **16**, 571–579 (2010).
4. Wiesner, T. et al. Alternative transcription initiation leads to expression of a novel ALK isoform in cancer. *Nature* **526**, 453–457 (2015).
5. Lock, F. E. et al. Distinct isoform of FABP7 revealed by screening for retroelement-activated genes in diffuse large B-cell lymphoma. *Proc. Natl Acad. Sci. USA* **111**, E3534–E3543 (2014).
6. Volkman, H. E. & Stetson, D. B. The enemy within: endogenous retroelements and autoimmune disease. *Nat. Immunol.* **15**, 415–422 (2014).
7. Groh, S. & Schotta, G. Silencing of endogenous retroviruses by heterochromatin. *Cell Mol. Life Sci.* **74**, 2055–2065 (2017).
8. Matsui, T. et al. Proviral silencing in embryonic stem cells requires the histone methyltransferase ESET. *Nature* **464**, 927–931 (2010).
9. Bulut-Karslioglu, A. et al. Suv39h-dependent H3K9me3 marks intact retrotransposons and silences LINE elements in mouse embryonic stem cells. *Mol. Cell* **55**, 277–290 (2014).
10. Kim, K. C., Geng, L. & Huang, S. Inactivation of a histone methyltransferase by mutations in human cancers. *Cancer Res.* **63**, 7619–7623 (2003).
11. Pinheiro, I. et al. Prdm3 and Prdm16 are H3K9me1 methyltransferases required for mammalian heterochromatin integrity. *Cell* **150**, 948–960 (2012).
12. Falandry, C. et al. CLLD8/KMT1F is a lysine methyltransferase that is important for chromosome segregation. *J. Biol. Chem.* **285**, 20234–20241 (2010).
13. Ecco, G., Imbeault, M. & Trono, D. KRAB zinc finger proteins. *Development* **144**, 2719–2729 (2017).
14. Schultz, D. C., Ayyanathan, K., Negorev, D., Maul, G. G. & Rauscher, F. J. 3rd SETDB1: a novel KAP-1-associated histone H3, lysine 9-specific methyltransferase that contributes to HP1-mediated silencing of euchromatic genes by KRAB zinc-finger proteins. *Genes Dev.* **16**, 919–932 (2002).
15. Rowe, H. M. et al. De novo DNA methylation of endogenous retroviruses is shaped by KRAB-ZFPs/KAP1 and ESET. *Development* **140**, 519–529 (2013).
16. Leung, D. et al. Regulation of DNA methylation turnover at LTR retrotransposons and imprinted loci by the histone methyltransferase Setdb1. *Proc. Natl Acad. Sci. USA* **111**, 6690–6695 (2014).

17. Liu, S. et al. Setdb1 is required for germline development and silencing of H3K9me3-marked endogenous retroviruses in primordial germ cells. *Genes Dev*. **28**, 2041–2055 (2014).
18. Liu, X. et al. UHRF1 targets DNMT1 for DNA methylation through cooperative binding of hemi-methylated DNA and methylated H3K9. *Nat. Commun.* **4**, 1563 (2013).
19. Rothbart, S. B. et al. Association of UHRF1 with methylated H3K9 directs the maintenance of DNA methylation. *Nat. Struct. Mol. Biol.* **19**, 1155–1160 (2012).
20. Zhao, Q. et al. Dissecting the precise role of H3K9 methylation in crosstalk with DNA maintenance methylation in mammals. *Nat. Commun*. **7**, 12464 (2016).
21. Karimi, M. M. et al. DNA methylation and SETDB1/H3K9me3 regulate predominantly distinct sets of genes, retroelements, and chimeric transcripts in mESCs. *Cell Stem Cell* **8**, 676–687 (2011).
22. Sharif, J. et al. Activation of endogenous retroviruses in Dnmt1(-/-) ESCs involves disruption of SETDB1-mediated repression by NP95 binding to hemimethylated DNA. *Cell Stem Cell* **19**, 81–94 (2016).
23. Walsh, C. P., Chaillet, J. R. & Bestor, T. H. Transcription of IAP endogenous retroviruses is constrained by cytosine methylation. *Nat. Genet*. **20**, 116–117 (1998).
24. Hutnick, L. K. et al. DNA hypomethylation restricted to the murine forebrain induces cortical degeneration and impairs postnatal neuronal maturation. *Hum. Mol. Genet*. **18**, 2875–2888 (2009).
25. Ramesh, V. et al. Loss of Uhrf1 in neural stem cells leads to activation of retroviral elements and delayed neurodegeneration. *Genes Dev*. **30**, 2199–2212 (2016).
26. Tan, S. L. et al. Essential roles of the histone methyltransferase ESET in the epigenetic control of neural progenitor cells during development. *Development* **139**, 3806–3816 (2012).
27. Kato, M., Takemoto, K. & Shinkai, Y. A somatic role for the histone methyltransferase Setdb1 in endogenous retrovirus silencing. *Nat. Commun*. **9**, 1683 (2018).
28. Engert, S., Liao, W. P., Burtscher, I. & Lickert, H. Sox17-2A-iCre: a knock-in mouse line expressing Cre recombinase in endoderm and vascular endothelial cells. *Genesis* **47**, 603–610 (2009).
29. Lohmann, F. et al. KMT1E mediated H3K9 methylation is required for the maintenance of embryonic stem cells by repressing trophectoderm differentiation. *Stem Cells* **28**, 201–212 (2010).
30. Yuan, P. et al. Eset partners with Oct4 to restrict extraembryonic trophoblast lineage potential in embryonic stem cells. *Genes Dev.* **23**, 2507–2520 (2009).
31. Kawamoto, S. et al. A novel reporter mouse strain that expresses enhanced green fluorescent protein upon Cre-mediated recombination. *FEBS Lett.* **470**, 263–268 (2000).
32. Hou, J. et al. A regulatory network controls nephrocan expression and midgut patterning. *Development* **141**, 3772–3781 (2014).
33. Mochida, Y. et al. Nephrocan, a novel member of the small leucine-rich repeat protein family, is an inhibitor of transforming growth factor-beta signaling. *J. Biol. Chem.* **281**, 36044–36051 (2006).
34. Dewannieux, M., Dupressoir, A., Harper, F., Pierron, G. & Heidmann, T. Identification of autonomous IAP LTR retrotransposons mobile in mammalian cells. *Nat. Genet*. **36**, 534–539 (2004).
35. Kwon, G. S., Viotti, M. & Hadjantonakis, A. K. The endoderm of the mouse embryo arises by dynamic widespread intercalation of embryonic and extraembryonic lineages. *Dev. Cell* **15**, 509–520 (2008).
36. Cernilogar, F. M. et al. Pre-marked chromatin and transcription factor co-binding shape the pioneering activity of Foxa2. *Nucleic Acids Res*. https://doi.org/10.1093/nar/gkz627 (2019).
37. Wamaitha, S. E. et al. Gata6 potently initiates reprograming of pluripotent and differentiated cells to extraembryonic endoderm stem cells. *Genes Dev*. **29**, 1239–1255 (2015).
38. Nicetto, D. et al. H3K9me3-heterochromatin loss at protein-coding genes enables developmental lineage specification. *Science* **363**, 294–297 (2019).
39. Yin, Y. et al. Impact of cytosine methylation on DNA binding specificities of human transcription factors. *Science* **356**, https://doi.org/10.1126/science.aaj2239 (2017).
40. Viotti, M., Nowotschin, S. & Hadjantonakis, A. K. SOX17 links gut endoderm morphogenesis and germ layer segregation. *Nat. Cell Biol*. **16**, 1146–1156 (2014).
41. Nowotschin, S. et al. The emergent landscape of the mouse gut endoderm at single-cell resolution. *Nature* **569**, 361–367 (2019).
42. Rodriguez, C. I. et al. High-efficiency deleter mice show that FLPe is an alternative to Cre-loxP. *Nat. Genet*. **25**, 139–140 (2000).
43. Nakamura, T., Colbert, M. C. & Robbins, J. Neural crest cells retain multipotential characteristics in the developing valves and label the cardiac conduction system. *Circ. Res*. **98**, 1547–1554 (2006).
44. Engert, S. et al. Wnt/beta-catenin signalling regulates Sox17 expression and is essential for organizer and endoderm formation in the mouse. *Development* **140**, 3128–3138 (2013).
45. Vandesompele, J. et al. Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes. *Genome Biol*. **3**, RESEARCH0034 (2002).
46. Sadic, D. et al. Atrx promotes heterochromatin formation at retrotransposons. *EMBO Rep*. **16**, 836–850 (2015).
47. Tommasi, S. et al. Whole DNA methylome profiling in mice exposed to secondhand smoke. *Epigenetics* **7**, 1302–1314 (2012).
48. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
49. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol*. **15**, 550 (2014).
50. Taudt, A., Nguyen, M. A., Heinig, M., Johannes, F. & Colomé-Tatché, M. chromstaR: Tracking combinatorial chromatin state dynamics in space and time. *bioRxiv*, 038612, https://doi.org/10.1101/038612 (2016).
51. Kumaki, Y., Oda, M. & Okano, M. QUMA: quantification tool for methylation analysis. *Nucleic Acids Res*. **36**, W170–W175 (2008).

Acknowledgements

We deeply thank Brian Cullen for providing his last aliquots of the IAP-GAG antibody. We wish to thank the help of BMC core facilities for FACS and Bioinformatics for providing equipment, services, expertise, and assistance with data analysis. We are further thankful to the LAFUGA Sequencing facility and the Genomics Service Unit of the LMU. Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation), Project-ID 213249687 – SFB 1064 to G.S. and H.Le. and Projekt-ID 329628492 – SFB 1321 TP13 to G.S.

Author contributions

G.S., H.Li., H.Le., Z.W., R.F. contributed to concepts and approaches (designed the experimental approach) (conceived and designed the project); Z.W., R.F., A.R., F.M.C., A.N., S.S., I.S., I.D., E.U., T.A., L.R. performed the experiments; Z.W. performed the bioinformatic analysis with the help of G.S.; G.S. and Z.W. wrote the manuscript.

Funding

Open Access funding enabled and organized by Projekt DEAL.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-022-32978-7.

Correspondence and requests for materials should be addressed to Gunnar Schotta.

Peer review information *Nature Communications* thanks Maxim Greenberg and the other anonymous reviewer(s) for their contribution to the peer review of this work. Peer review reports are available.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2022

Qin, W., **Ugur, E.**, Mulholland, C. B., Bultmann, S., Solovei, I., Modic, M., Smets, M., Wierer, M., Forné, I., Imhof, A., Cardoso, M. C., & Leonhardt, H. (**2021**). **Phosphorylation of the HP1β hinge region sequesters KAP1 in heterochromatin and promotes the exit from naïve pluripotency**. Nucleic Acids Research, 49(13), 7406–7423.

License: CC-BY-4.0

https://doi.org/10.1093/nar/gkab548

7406–7423 Nucleic Acids Research, 2021, Vol. 49, No. 13
https://doi.org/10.1093/nar/gkab548

Published online 2 July 2021

Phosphorylation of the HP1β hinge region sequesters KAP1 in heterochromatin and promotes the exit from naïve pluripotency

Weihua Qin[1,*], Enes Ugur[1,2], Christopher B. Mulholland[1], Sebastian Bultmann[1], Irina Solovei[1], Miha Modic[3], Martha Smets[1], Michael Wierer[2], Ignasi Forné[4], Axel Imhof[4], M. Cristina Cardoso[5] and Heinrich Leonhardt[1,*]

[1]Faculty of Biology, Ludwig-Maximilians-Universität München, Butenandtstraße 1, D-81377 Munich, Germany, [2]Department of Proteomics and Signal Transduction, Max Planck Institute for Biochemistry, Am Klopferspitz 18, 82152 Martinsried, Germany, [3]The Francis Crick Institute and UCL Queen Square Institute of Neurology, London NW1 1AT, United Kingdom, [4]Biomedical Center Munich, Faculty of Medicine, Ludwig-Maximilians-Universität München, Großhaderner Str. 9, 82152 Planegg-Martinsried, Germany and [5]Cell Biology and Epigenetics, Department of Biology, Technical University of Darmstadt, 64287 Darmstadt, Germany

Received March 29, 2021; Revised May 31, 2021; Editorial Decision June 09, 2021; Accepted June 11, 2021

ABSTRACT

Heterochromatin binding protein HP1β plays an important role in chromatin organization and cell differentiation, however the underlying mechanisms remain unclear. Here, we generated *HP1β$^{-/-}$* embryonic stem cells and observed reduced heterochromatin clustering and impaired differentiation. We found that during stem cell differentiation, HP1β is phosphorylated at serine 89 by CK2, which creates a binding site for the pluripotency regulator KAP1. This phosphorylation dependent sequestration of KAP1 in heterochromatin compartments causes a downregulation of pluripotency factors and triggers pluripotency exit. Accordingly, *HP1β$^{-/-}$* and phospho-mutant cells exhibited impaired differentiation, while ubiquitination-deficient *KAP1$^{-/-}$* cells had the opposite phenotype with enhanced differentiation. These results suggest that KAP1 regulates pluripotency via its ubiquitination activity. We propose that the formation of subnuclear membraneless heterochromatin compartments may serve as a dynamic reservoir to trap or release cellular factors. The sequestration of essential regulators defines a novel and active role of heterochromatin in gene regulation and represents a dynamic mode of remote control to regulate cellular processes like cell fate decisions.

INTRODUCTION

Heterochromatin binding protein HP1 is a non-histone chromosomal protein and has a function in the establishment and maintenance of higher-order chromatin structures and gene silencing (1,2). In mammals, there are three homologues of HP1, termed HP1α, HP1β and HP1γ, encoded by *Cbx5*, *Cbx1* and *Cbx3* genes, respectively. HP1 homologues contain two conserved functional domains, an N-terminal chromodomain (CD) and a C-terminal chromoshadow domain (CSD), linked by a hinge region. The CD domain is responsible for recognition of di- and trimethylated K9 of histone H3 (H3K9me2 and H3K9me3) (3–5), while the CSD domain mediates interactions with other proteins (6,7). The intrinsically disordered regions (IDRs) and posttranslational modifications are likely responsible for the unique functions of HP1 homologues.

Recent studies testing the capacity of HP1 to induce phase separation revealed that only HP1α formed phase-separated droplets (8,9). This phase separation correlates with the formation of heterochromatin compartments (chromocenters) in the nucleus. Recently, we found that the charge of the hinge IDR (IDR-H) is a distinctive feature of HP1 homologues and plays a decisive role in liquid-liquid phase separation (LLPS) (10,11) and that HP1β also undergoes phase separation in a histone H3K9me3 dependent manner (11). HP1α/β together with other chromatin binding proteins, such as SUV39H1 and KAP1, coalesce around the solid chromatin scaffold (12–15).

In cells, HP1α and HP1γ locate at condensed heterochromatin and euchromatin, respectively, while HP1β accumulates mostly at condensed heterochromatin and to less ex-

---

[*]To whom correspondence should be addressed. Tel: +49 89 2180 74232; Fax: +49 89 2180 74236; Email: h.leonhardt@lmu.de
Correspondence may also be addressed to Weihua Qin. Tel: +49 89 2180 71132; Email: weihua@zi.biologie.uni-muenchen.de

© The Author(s) 2021. Published by Oxford University Press on behalf of Nucleic Acids Research.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

tent at euchromatin (16). The specific functions of HP1 proteins in chromatin organization correlate with their unique cellular roles during cell differentiation. HP1β knockout mice die perinatally and show impaired development of the cerebral neocortex and neuromuscular junctions (17). In mouse embryonic stem cells (mESCs), depletion of HP1β affects differentiation (18). However, how HP1β regulates cell differentiation is unclear.

To address this question, we generated $HP1\beta^{-/-}$ mESCs. These cells showed impaired naïve pluripotency exit and are defective in neural progenitor cell differentiation. We found that HP1β is phosphorylated at the serine 89 residue of the hinge region by casein kinase 2 (CK2). This phosphorylation creates a specific binding site for KAP1, which leads to sequestration of this pluripotency factor and downregulation of pluripotency genes. While phase separation and the formation of membraneless compartments has been implicated in the local enrichment of factors involved in the same cellular process, the sequestration of KAP1 represents a novel mechanism of transcriptional regulation and cell fate decision by remote controlled functional depletion.

MATERIALS AND METHODS

Cell culture, transfection and inhibitor treatment

Human embryonic kidney (HEK) 293T cells and baby hamster kidney (BHK) cells were cultured in DMEM supplemented with 10% fetal calf serum (FCS) and 50 μg/ml gentamicin (PAA).

Naive E14 mESCs (19) were cultured as described previously (20). In brief, cells were kept under naive conditions in N2B27 medium consisting of 50% DMEM/F12 (Life Technologies) supplemented with N2 (Life Technologies) and 50% neurobasal medium (Life Technologies) supplemented with serum-free B27 (Life Technologies), 2 mM L-glutamine (Life Technologies), 100 U/ml penicillin, 100 μg/ml streptomycin (PAA Laboratories GmbH) and 0.1 mM β-mercaptoethanol (Life Technologies). Naive mESCs were maintained on flasks treated with Geltrex (Life Technologies) diluted 1:100 in DMEM/F12 (Life Technologies) in N2B27 media containing 2i (1 μM PD032591 and 3 μM CHIR99021 (Axon Medchem, Netherlands)), 1000 U/ml recombinant leukemia inhibitory factor (LIF, Millipore) and 0.3% BSA (Gibco).

For the metastable state culture of mESCs, cells were cultured in gelatinized flasks in DMEM supplemented with 16% FCS, 0.1 mM β-mercaptoethanol (Invitrogen), 2 mM L-glutamine, 1× MEM non-essential amino acids, 100 U/ml penicillin, 100 μg/ml streptomycin (PAA) and 1000 U/ml recombinant leukemia inhibitory factor LIF (Millipore). For CRISPR-assisted cell line generation, the culture medium was supplemented with 2i.

To differentiate ESCs from naive to epiblast state, cells were plated on flasks treated with Geltrex (Life Technologies) in defined medium containing 20 ng/ml Activin A (R&D Systems), 10 ng/ml FGF2 (R&D Systems) and 0.1× Knockout Serum Replacement (Life Technologies). Media was changed after 24 h and epiblast cells were imaged at 48 h.

Mouse ESCs were transfected with Lipofectamine 3000 Reagent (Invitrogen) according to the manufacturer's instructions. HEK 293T and BHK cells were transfected using polyethylenimine (PEI) as transfection reagent (Sigma-Aldrich) according to the manufacturer's instructions. Cell fixation and microscopy were carried out as described (21).

To inhibit CK2 activity, 50 μM of the specific inhibitor 4,5,6,7-tetrabromobenzotriazole (TBB) was added directly after transfection. To check HP1β-pS89 levels in wt and $CK2a1^{as}$ cells, 10 μM of the adenine analog 1-NA-PP1 was supplemented to the medium overnight.

CRISPR/Cas-mediated gene editing and generation of stable cell lines

For generation of $HP1\beta^{-/-}$ mESCs, the MINtag strategy was used as described previously (11,22). After generation of the MIN-tagged line, the attB-RFP-Stop-PolyA was inserted into the N-terminus of the endogenous $HP1\beta^{attP/attP}$ locus by Bxb1 mediated recombination.

For generation of the $CK2a1^{as}$ cell line, genome editing was performed with slight modifications compared to a previous publication (23). Briefly, the two gRNAs for editing CK2a1 were designed using the CRISPR design tool from the Zhang Lab (MIT, www.genome-engineering.org), and got incorporated into the pSpCas9 (BB)-2A-GFP (px458) vector by BpiI restriction sites (23). To mutate CK2a1 at aa position 113 from phenylalanine to alanine, a 200 nt ssDNA donor oligo was synthesized by Integrated DNA Technologies (IDT). A HpyCH4V cutting site was incorporated into the donor oligo for screening. Mouse ESCs were transfected with the Cas9-gRNA vector and a donor oligo. 48 h after transfection, GFP positive cells were sorted using FACS and plated at clonal density. After one-week, individual clones were picked and expanded for genomic DNA isolation. The mutant clones were validated by PCR using respective primers, HpyCH4V digestion and DNA sequencing (Supplementary Figure S3E).

For generation of HP1β S89A and HP1β S89E cell lines, the specific gRNA was cloned into a vector expressing GFP and SpCas9 (px458: F. Zhang Lab). To mutate HP1β from serine (S) to alanine (A) or glutamic acid (E) in aa position 89, 200 nt ssDNA donor oligos were synthesized by Integrated DNA Technologies (IDT). For screening, the HypCH4V or HypCH4IV cutting site was incorporated into the donor oligo of S89A and S89E, respectively. The mutant lines were validated by PCR using the respective primers followed by HypCH4V or HypCH4IV (New England Biolabs) digestion and DNA sequencing. The expression of HP1β S89A and HP1β S89E was analyzed by western blot and immunostaining.

For generation of the $KAP1^{-/-}$ cell line, KAP1-specific gRNA was cloned into a puromycin-selectable vector expressing SpCas9 (px459: F. Zhang Lab). Mouse ESCs were transfected with the Cas9-gRNA vector and two days after transfection E14 mESCs were plated at clonal density in ESC media supplemented with 1 μg/ml puromycin (Gibco). Selection media was replaced by normal ESC media after 48 h and colonies were allowed to grow for a week. Single ESC colonies were transferred into 2 × 96-well plates. Selection of $KAP1^{-/-}$ clones was accomplished by amplifying the CRISPR/Cas targeted region via PCR followed by PstI digestion (FastDigest; Thermo Scientific). Positive clones were assessed by sanger sequencing and western blots by using antibodies against both N- and C-terminus of KAP1 (Figure 6B and Supplementary Figure S8).

To generate KAP1-GFP mESCs, gRNA specific to C-terminus of KAP1 locus was cloned into a puromycin-selectable vector expressing both SpCas9 (px459: F. Zhang Lab). mESCs were transfected with the Cas9-gRNA vector and a 719 bp donor synthesized by Integrated DNA Technologies (IDT). Two days after transfection, cells were subjected to puromycin (1 μg/ml) for two days. A week later, GFP positive cells were sorted using FACS (Supplementary Figure S10B).

For generation of stable mESC and HEK293T lines, 48 h after expression of GFP-tagged constructs (GFP-HP1β wt, GFP-HP1β S89/91A, GFP-HP1β S89/91D, GFP-KAP1 wt, GFP-KAP1 RH, GFP-KAP1 PVL, GFP-KAP1 RH/PVL), cells were plated at clonal density and subjected to blasticidin selection (10 μg/ml) for a week. Then, GFP positive cells were separated using a fluorescence activated cell sorting (FACS) Aria II instrument (Becton Dickinson).

The cell lines and expression constructs are listed in Supplementary Table S1.

Immunofluorescence staining

For immunostaining, mESCs were grown on coverslips coated with Geltrex (Life Technologies). After rinsing coverslips 2× with PBS (pH 7.4; 140 mM NaCl, 2.7 mM KCl, 6.5 mM Na2HPO4, 1.5 mM KH_2PO_4), cells were fixed for 10 min with 3.7% formaldehyde (Sigma), washed 3× for 10 min with PBST (PBS, 0.01% Tween20), permeabilized for 5 min in PBS supplemented with 0.5% Triton X-100 and washed 2× for 10 min with PBS. Primary and secondary antibodies (see Supplementary Table S1) were diluted in blocking solution (PBST, 3% BSA). After the incubation steps with the respective antibody in a humidified dark chamber for 1 h, coverslips were washed 3× for 10 min with PBST. For DNA counterstaining, coverslips were incubated in a solution of DAPI (400 ng/ml) in PBS-T for 5 min, before washing 3× for 10 min with PBST. Coverslips were mounted in antifade containing medium (Vectashield, Vector Laboratories) and sealed with colorless nail polish. Images were collected using a Leica TCS SP5 confocal microscope equipped with Plan Apo 63×/1.4 NA oil immersion objective and lasers with excitation lines 405, 488, 594 and 633 nm.

Co-immunoprecipitation and western blotting

For co-immunoprecipitation, 1×10^7 mESCs were lysed in lysis buffer (10 mM Tris/Cl pH7.5, 150 mM NaCl, 0.5 mM EDTA, 0.5% NP40, 1.5 mM $MgCl_2$, 0.5 U/ml Benzonase (Sigma-Aldrich), 1 mM PMSF, 1× mammalian Protease Inhibitor Cocktail (Serva®) at 4°C for 30 min. Lysate was cleared up by centrifugation at 20 000 × g at 4°C for 15 min and protein concentration was measured with Pierce™ 660 nm Protein Assay Reagent according to the manufacturer's instructions. Equal amounts of protein extracts were incubated with 80 μl of anti- HP1β-pS89 antibody for 2 h at 4°C under constant rotation. Then, 20 μl of protein G beads (GE) were added to the protein extracts and incubated overnight under constant rotation at 4°C. Beads were washed 3× with washing buffer (10 mM Tris/Cl pH7.5, 300 mM NaCl, 0.5 mM EDTA) and boiled in Laemmli buffer at 95°C for 10 min. Bound fractions were separated and visualized by western blotting.

For immunoprecipitation, HEK 293T cells stably expression GFP-HP1β wt and its mutants were treated with hypotonic buffer (10 mM Tris–HCl pH 8, 10 mM KCl, 1.5 mM $MgCl_2$, 1 mM DTT and 1× Protease Inhibitor, 2 mM PMSF) for 30 min and centrifuged at 1000 × g at 4°C to get the intact nuclei. Nuclei were lysed in a lysis buffer at 4°C for 30 min. Lysates were first cleared by centrifugation at 20 000 × g for 15 min at 4°C and then incubated with a GFP-Trap (Chromotek). Bound fractions were visualized by a coomassie stained polyacrylamide gel.

For the general detection of HP1β on western blots, a rabbit anti-HP1β antibody (Abcam and Cell Signaling Technology, see Supplementary Table S1) was used. For the specific detection of HP1β-pS89, antibodies against the peptide GKRKADpSDSEDKG were raised in mice and rats. RFP or Cherry fusion proteins were detected by the rat-anti-red antibody 5F8 (24). KAP1 was visualized by rabbit anti-KAP1 antibodies (Abcam and Proteintech, see Supplementary Table S1). Equal loading of cell lysates was assessed by a mouse anti-β-actin antibody (Sigma-Aldrich), a mouse anti-tubulin antibody (Sigma-Aldrich) and a polyclonal H3 antibody (Abcam, see Supplementary Table S1). Secondary antibodies, anti-rabbit (Biorad), anti-rat and anti-mouse (Dianova), were conjugated to horseradish peroxidase and visualized with ECL Plus reagent (GE Healthcare, Thermo Scientific). Signals were acquired on an Amersham Imager 600 (GE).

Antibodies used in this study are listed in Supplementary Table S1.

F3H assay

The F3H assay was performed as described previously (25). In brief, BHK cells containing lac operator arrays were transiently transfected on coverslips using PEI and fixed with 3.7% formaldehyde 16 h after transfection. For DNA counterstaining, coverslips were incubated in a solution of DAPI (400 ng/ml) in PBST and mounted in Vectashield. Images were taken using a SP5 Leica confocal microscope equipped with Plan Apo 63×/1.4 NA oil immersion objective and lasers with excitation lines: 405 nm for DAPI, 488 nm for GFP fusions, 561 nm for Cherry fusions and 633 nm for HP1β-pS89.

Flow cytometry analysis

For flow cytometry, plates were washed once with PBS, dissociated to single cells by trypsin-EDTA treatment, resuspended in PBS buffer supplemented with 2% FBS and 1 mM EDTA, and incubated with DyLight-650-conjugated anti-SSEA-1 (clone MC-480, MA1-022-D650, Life Technologies) antibody for 30–60 min on ice. Cells were spun down, resuspended in a buffer containing DAPI for live-dead cell staining and analyzed using a FACS Aria II (BD Biosciences). Cell debris was excluded by forward and side scatter gating. FlowJo was used for data analysis.

Mass spectrometry of in-gel digests

In-gel digests were performed according to standard protocols. Briefly, after washing the excised gel slices proteins were reduced by DTT, alkylated with iodoacetamide and digested with trypsin (Sequencing Grade Modified, Promega) overnight at 37°C. For protein identification the resulting peptides were purified on-line with C18 reversed cartridge (Dionex) and separated in an Ultimate 3000 RSLCnano system (Thermo Fisher Scientific), using in a 15-cm analytical column (75 μm ID home-packed with ReproSil-Pur C18-AQ 2.4 μm from Dr Maisch) with a 50-min gradient from 5 to 60% acetonitrile in 0.1% formic acid. The effluent from the HPLC was directly electrosprayed into Orbitrap-LTQ XL (Thermo Fisher Scientific) operated in data dependent mode to automatically switch between full scan MS and MS/MS acquisition. Survey full scan MS spectra (from m/z 300 –2000) were acquired in the Orbitrap with resolution $R = 60\,000$ at m/z 400 (after accumulation to a 'target value' of 500 000 in the linear ion trap). The six most intense peptide ions with charge states between 2 and 4 were sequentially isolated to a target value of 10 000 and fragmented in the linear ion trap by collision induced dissociation (CID). All fragment ion spectra were recorded in the LTQ part of the instrument. For all measurements with the Orbitrap detector, 3 lock-mass ions from ambient air were used for internal calibration. Typical MS conditions were spray voltage, 1.5 kV; no sheath and auxiliary gas flow; heated capillary temperature, 200°C; normalized CID energy 35%; activation $q = 0.25$; activation time = 30 ms. Proteins were identified using Mascot (Matrix Science, London, UK; version Mascot) against SwissProt_2011.02 database for human proteins (Fragment Tolerance: 0.80 Da, Fixed Modification for carbamidomethyl cysteine, Variable Modification for methionine oxidation, Max Missed Cleavage: 2).

Protein purification and histone isolation

HP1 cDNA was cloned into a pET28 expression vector, mutants were made using overlap extension PCR and proteins were subsequently expressed in *Escherichia coli*. Purifications of HP1β proteins were described previously (11).

KAP1 cDNA was cloned into a pCAG-GFP expression vector and respective mutants were made using overlap extension PCR. HEK293T were transfected with the plasmid coding for GFP-KAP1, harvested 48 h after transfection and lysed in lysis buffer (50 mM NaH_2PO_4 pH 8.0, 300 mM NaCl, 10 mM imidazole, 0,5% Tween-20, 2 mM MgCl2, 0.5 U/ml Benzonase, 1 mM PMSF, 1× mammalian protease inhibitor cocktail.) at 4°C for 30 min. Cell debris were removed by centrifugation at 20 000 × g for 15 min at 4°C. Cleared cell lysate was incubated with Ni-NTA-GBP beads for 1.5 h under constant rotation at 4°C. The beads were washed 3× with washing buffer (50 mM NaH_2PO_4 pH 7.5, 300 mM NaCl, 20 mM imidazole, 0.05% Tween-20) before eluting the protein with elution buffer (10 mM Tris pH 7.5, 100 mM KCl, 1 mM EDTA, 1 mM DTT and 250 mM imidazole). Protein concentration was assessed by measuring its GFP emission signal on a plate reader (TECAN) with purified GFP as standard reference.

Histone isolation was conducted as previously described with minor changes of the protocol (26). In brief, 15 × p100 HEK293T cells were harvested, and cell pellets were resuspended in a hypotonic buffer. To obtain pure nuclei, cells were disrupted using a homogenizer and nuclei were subsequently incubated in a chromatin dissociation buffer (10 Tris–HCl pH 8.0, 20 mM EDTA and 400 mM NaCl) for 30 min on ice. This chromatin dissociation step was repeated 4×. Afterwards, nuclei were resuspended in 0.4 N H2SO4 and incubated on a rotator at 4°C overnight. After centrifugation, histones in the supernatant were transferred into a fresh reaction tube and precipitated using 33% Trichloroacetic acid (TCA). After washing 3x with cold acetone, histones were dissolved in H2O and centrifuged at 2000 rpm for 5 min to remove precipitations. Histone concentrations were measured using the PierceTM 660 nm protein assay kit.

In vitro droplet assays

For the droplet assay, proteins were concentrated to ∼10 μg/μl using Amicon concentrators. After the concentration step, buffer was exchanged to 20 mM HEPES pH 7.2, 75 mM KCl, 1 mM DTT with Zeba™ Spin Desalting Columns. For the spin down assay, 30 μl of turbid solution was spun down at 2000 rpm for 5 min and 29 μl of supernatant was transferred into a Protein LoBind Tube (Eppendorf). The supernatant and droplets were boiled in 120 μl laemmli buffer at 95°C for 10 min. 10 μl of supernatant and droplets were loaded into a SDS-PAGE gel for following detection via coomassie stain.

For visualization of His-tagged-HP1β within the droplets, 500 ng of protein was labeled according to the Monolith NT™ Protein Labeling Kit RED-NHS from Nano Temper. After buffer exchange, 50 ng of the labeled protein was added into the droplet solution. For visualization of GFP-KAP1 within the droplets, 100 ng of protein was incubated with HP1β at 4°C for 3 min before adding histones.

Neuronal progenitor cell (NPC) differentiation

The differentiation of pluripotent ESCs into NPCs was based on a protocol described before (27). Simply, ESCs maintained with naïve medium (2i/LIF) were switched to the metastable culturing medium (serum/LIF) one week before the NPC differentiation. At the D0, 4×10^6 cells were plated onto bacteriological Petri dishes (Greiner) in 15 ml cellular aggregates (CA) medium (DMEM supplemented with 10% FCS, 2 mM L-glutamine, 1 × non-essential amino acids and 0.1 mM β-mercaptoethanol). At the D4, 5 μM of the retinoic acid (RA) was added into the CA medium. At the D8, the CAs were dissociated with freshly prepared trypsin and were plated onto PORN/laminin-coated plated with N2 medium (125 ml DMEM, 125 ml F-12, 1.25 ml insulin (25 μg/ml), 6.25 ml transferrin (50 μg/ml), 0.25 ml progesterone (20 nM), 0.25 ml putrescine (100 nM), 25 μl sodium selenite (30 nM), 0.5× L-glutamine, 1× Pen/Strep and 50 μg/ml BSA). Samples from different time points of differentiation were collected for analyses.

Alkaline phosphatase (AP) staining

One thousand mESCs were seeded into one well of a six-well plate and cultured for 6 days prior to the AP staining. The AP staining was performed as published previously (28) using the Alkaline Phosphatase Detection Kit (Sigma-Aldrich) according to the manufacturer's instructions.

RNA isolation and RNA sequencing and transcriptome analysis

For RNA-seq, RNA was isolated using the NucleoSpin Triprep Kit (Machery-Nagel) according to the manufacturer's instructions. Digital gene expression libraries for RNA-seq were produced using a modified version of single-cell RNA barcoding sequencing (SCRB-seq) optimized to accommodate bulk cells (29) in which a total of 70 ng of input RNA was used for the reverse-transcription of individual samples. RNA-seq libraries were sequenced on an Illumina HiSeq 1500. The libraries were sequenced paired end with 15–20 cycles to decode sample barcodes and UMI from read 1 and 45 cycles into the cDNA fragment. Similar sequencing qualities were confirmed by FastQC v0.10.1.

To generate principal component analysis (PCA) plot, SCRB-seq pools (i7) were demultiplexed from the Illumina barcode reads using deML (30). All reads were trimmed to the same length of 45 bp by cutadapt (31) (v1.8.3) and mapped using Spliced Transcripts Alignment to a Reference (STAR) (32) and mapped to the mouse genome (mm10). Gene-wise count/UMI tables were generated using the published Drop-seq pipeline (v1.0) (33). PCA was performed on the 1000 most variable genes to display the major variance between the genotype and differentiation state.

To check gene expression during NPC differentiation, RNA-seq libraries were processed and mapped to the mouse genome (mm10) using the zUMIs pipeline (34). UMI count tables were filtered for low, plasmids, counts using HTSFilter (35). Differential expression analysis was performed in R using DESeq2 (36) and genes with an adjusted $P<0.05$ were considered to be differentially expressed.

For GO analysis of biological processes the online tool (http://cbl-gorilla.cs.technion.ac.il/) was used (37,38). For the analysis of $HP1\beta^{-/-}$ and HP1β S89A, genes showing >1.5-fold changes (Supplementary Table S2) were considered. The upregulated and downregulated genes in $KAP1^{-/-}$ (Supplementary Table S4) were separately analyzed. The GO analyses were done by two unranked lists of genes with p-value thresholds of 1.0E−03 and 1.0E−05.

Chromatin immunoprecipitation and sample preparation for mass spectrometry

Chromatin immunoprecipitation coupled to mass spectrometry (ChIP-MS) of HP1β was performed in two technical replicates for WT and HP1β-KO mESCs and EpiLCs by using a direct HP1β antibody (Abcam). For each replicate, independently grown 15×10^6 cells were harvested and crosslinked as described previously (39). Next, nuclei were isolated with a mild lysis buffer (20 mM Tris–HCl pH 8.0, 85 mM KCl, 0.5% NP40, 1× PIC) and briefly pelleted for 5 min at 2000 × g and 4°C. To digest DNA, nuclei were resuspended in an MNase digestion buffer (1 M sorbitol, 50 mM Tris–HCl pH 8.0, 5 mM $CaCl_2$, 1× PIC). Subsequently, 2 μl MNase (NEB, 6000 gel units) was added and, after 1 min prewarming at 37°C, samples were incubated for 12.5 min at 37°C and at 1000 rpm in a thermal shaker. The reaction was quenched by the addition of EGTA to a final concentration of 50 mM. Nuclei were then spun down and resuspended in the IP-Buffer (50 mM Tris–HCl pH 8.0, 100 mM NaCl, 5 mM EDTA, 0.3% SDS, 1.7% Triton X-100, 1× PIC). Samples were addressed to brief sonication (3 × 30 s) at low setting in a Bioruptor Plus (Diagenode). Lysates were then centrifuged for 20 min at maximum speed and 4°C. To check the DNA digestion efficiency 20 μl of each sample was diluted to 5% in TBS and 10 μl proteinase K (Invitrogen) was added. These quality check samples were incubated O/N at 65°C under constant shaking to reverse FA-crosslinks. The next day, samples were incubated with 5 μl RNaseA (10 mg/ml) and incubated for 30 min at 37°C. The DNA was purified (Quagen Quaquick PCR purification kit) and DNA sizes were checked on an 1% agarose gel.

Meanwhile samples for ChIP-MS were kept on ice. If the shearing efficiency was in the range of 150–500 bp the protein concentration of the ChIP-MS samples was estimated by a BCA assay (Thermo). Each replicate was diluted to 1 mg/ml in 1 ml total volume and 1 μg of antibody was added. The samples were incubated O/N at 4°C under constant rotation.

The next day, for each sample 20 μl (slurry volume) of magnetic protein A/G beads (Sigma) were washed 3x in the IP buffer and subsequently aliquoted to the samples. The samples were incubated at 4°C under constant rotation for 2 h. To enrich for direct HP1β interactors, samples were washed three times with a low-salt buffer (50 mM HEPES (pH 7.5), 140 mM NaCl, 1% Triton X-100) and once with a high-salt buffer (50 mM HEPES (pH 7.5), 500 mM NaCl, 1% Triton X-100). To reduce the detergent for subsequent protein digestion and proteomic analysis, samples were washed twice with TBS. After the last wash the supernatant was discarded carefully and the beads were resuspended in the elution buffer I (2 M Urea, 50 mM Tris–HCl (pH 7.5), 2 mM DTT and 20 μg/ml Trypsin) and incubated for 30 min at 37°C in a thermal shaker at 1100 rpm. Next, the supernatants were saved, and the beads were resuspended in 50 μl of elution buffer II (2 M urea, 50 mM Tris–HCl (pH 7.5), 40 mM CAA). After 5 min of incubation at 37°C, both supernatants were combined, and digestion was continued O/N at 25°C. The next day, 1% TFA was added to stop the digestion and peptides were cleaned-up on Stage Tipps consisting of three layers of C18 material (Empore) (40). Eluted and speedvac dried peptides were resuspended in 8 μl of A* buffer (0.1% TFA and 2% acetonitrile) and peptide concentrations were estimated by nanodrop at 280 nm.

Full proteome sample preparation

For full proteome measurements cells were lysed in 6 M Guanidinium Chloride, 100 mM Tris–HCl pH 8.5 and freshly added 2 mM DTT by constant pipetting and subsequent boiling for 10 min at 99°C and 1700 rpm. Next, samples were quickly spun down and sonicated for in a

Bioruptor Plus (30 s on/off interval, high setting). Protein concentrations were estimated by a BCA assay and meanwhile CAA was added to a final concentration of 40 mM. After a minimum incubation time of 20 min, 30 μg of each lysate was diluted in 30 μl of the lysis buffer and diluted 1:10 in the digestion buffer (25 mM Tris–HCl pH 8.5 and 10% acetonitrile). Next, trypsin and LysC were added in a 1:100 protease to protein ratio. Digestion was carried out O/N at 37°C and 1000 rpm. The next day, the samples were acidified with 1% TFA and cleaned-up on three layers of SDB-RPS material (Empore). After elution and vacuum drying, the samples were resuspended in 20 μl A* buffer and peptide concentrations were estimated by nanodrop at 280 nm.

Enrichment of K-Gly-Gly peptides

The K-Gly-Gly enrichment was performed by using the PTMScan Ubiquitin Remnant Motif Kit (Cell Signaling Technology) according to the manufacturer's protocol.

Briefly, 1×10^8 cells were lysed in the Urea lysis buffer (20 mM HEPES pH 8.0, 9 M urea, 1 mM sodium orthovanadate, 2.5 mM sodium pyrophosphate, 1 mM β-glycerophosphate.) and digested by Trypsin/LysC in an enzyme to protein ratio of 1:50 and 1:250, respectively. This step was carried out in duplicates for the $KAP1^{-/-}$ and in triplicates for wt mESCs. Next, peptides were desalted using 200-mg tC18 Sep Pak Cartridges (Waters). After vacuum drying of the samples, peptides were resuspended in the IAP buffer (50 mM Mops (pH 7.2), 10 mM sodium phosphate, 50 mM NaCl) and addressed to K-Gly-Gly pulldown. Then, eluted peptides were desalted once more with C18 Stage Tips, dried with a speedvac and resuspended in 20 μl of A* buffer (0.1% TFA and 2% acetonitrile). Peptide concentrations were estimated by nanodrop at 280 nm.

Mass spectrometry of ChIP-MS, full proteomes and K-Gly-Gly peptides

For mass spectrometry on a quadrupole Orbitrap mass spectrometer (Q Exactive HF-X, ThermoFisher Scientific), 300 ng of peptide solution per replicate was separated by nanoflow liquid chromatography on an Easy-nLC 1200 (ThermoFisher Scientific) during an increasing acetonitrile gradient for 120 min. As a column an in-house packed 50 cm column of ReproSil-Pur C18-AQ 1.9 μM resin (Dr Maisch GmbH) was used. The flow rate was constantly monitored and kept at 300 nl/min and the column oven temperature was fixed at 60°C. The injection was performed through a nanoelectrospray source. After each set of replicates, an additional washing step was scheduled. Data acquisition was performed in a data-dependent mode by selecting for the most abundant 12 peptides for MS/MS scans. The m/z range was set to 400–1650 m/z. The max. injection time was at 20 ms. The target value for the full scan MS spectra was 3×10^6 and the resolution at 60 000.

MS data analysis

Raw MS files were first analyzed with the MaxQuant software package (version 1.6.0.7) (41). The FASTA files (reviewed and unreviewed) were obtained from Uniprot (version 2020). Contaminants were identified by the Andromeda search engine (42) with 245 entries. 'Match between runs' option was enabled and the false discovery rate for both peptides (minimum length of 7 amino acids) and proteins was set to 1%. Determination of the relative protein amounts followed the MaxLFQ algorithm (43), with a minimum ratio count of two peptides.

For the downstream analysis of the MaxQuant output, Perseus was used. Contaminants were filtered out, intensities were transformed to $\log_2$ and a two-sided Student's t-test with a permutation-based FDR of 0.05 and a fold change cut-off of $\log_2 = 1$ was applied.

RESULTS

HP1β plays a role in mESC differentiation

Recently we found that HP1β shows phase separation properties in the presence of H3K9me3 histones *in vitro* (11). We next investigated its function in heterochromatin organization in cells. To this end, we first inserted a multifunctional integrase (MIN) tag directly after the ATG start codon of *HP1β* (Supplementary Figure S1A) for subsequent systematic studies applying our previously described genome engineering strategy (22) (Supplementary Table S1). In a second step we used Bxb1 mediated recombination to insert a transcription termination sequence into the MIN, i.e. directly after the ATG, to generate $HP1\beta^{-/-}$ mESCs (Supplementary Figure S1B). Immunostaining and reverse transcription quantitative PCR (RT-qPCR) showed that HP1β was completely depleted from the cells (Supplementary Figure S1C and D). DAPI staining of DNA showed alteration in chromocenter number and size in the $HP1\beta^{-/-}$ compared to the wt mESCs (Figure 1A–C). $HP1\beta^{-/-}$ cells exhibit an increased number of chromocenters which were on average smaller in size, indicating a reduced chromocenter clustering.

To profile the HP1β interactome, we performed chromatin immunoprecipitation coupled to mass spectrometry (ChIP-MS) of HP1β in wt and $HP1\beta^{-/-}$ mESCs. Among the HP1β interaction partners, we found SUV39H1/2, HP1α and KAP1 (Figure 1D), all proteins involved in regulation of chromatin compartmentalization (15). We also detected several zinc finger proteins (ZFPs) and transcriptional factors involved in pluripotency regulation (highlighted in cyan and green, respectively, Figure 1D).

As chromatin reorganization is a shared feature of multiple differentiation pathways (44–46), we investigated the role of HP1β in this process. To this aim, we differentiated *wt* and $HP1\beta^{-/-}$ mESCs to neural progenitor cells (NPCs, Figure 1E) (27). To monitor differentiation, we analyzed the expression of stage-specific embryonic antigen-1 (SSEA-1), a marker of pluripotent cells, at the distinct stages of NPC differentiation. Before LIF removal (stage D0) both wt and $HP1\beta^{-/-}$ mESCs were pluripotent as evidenced by high SSEA-1 expression (Figure 1E). At D4 of the differentiation protocol less than 5% of the wt cells were SSEA-1 positive, while more than 90% of the $HP1\beta^{-/-}$ mESCs were still SSEA-1 positive and even at NPC commitment (D9) 32.2% of the cells still expressed the SSEA-1 marker (Figure 1E). These results suggest that depletion of HP1β impairs the exit from the pluripotent state.

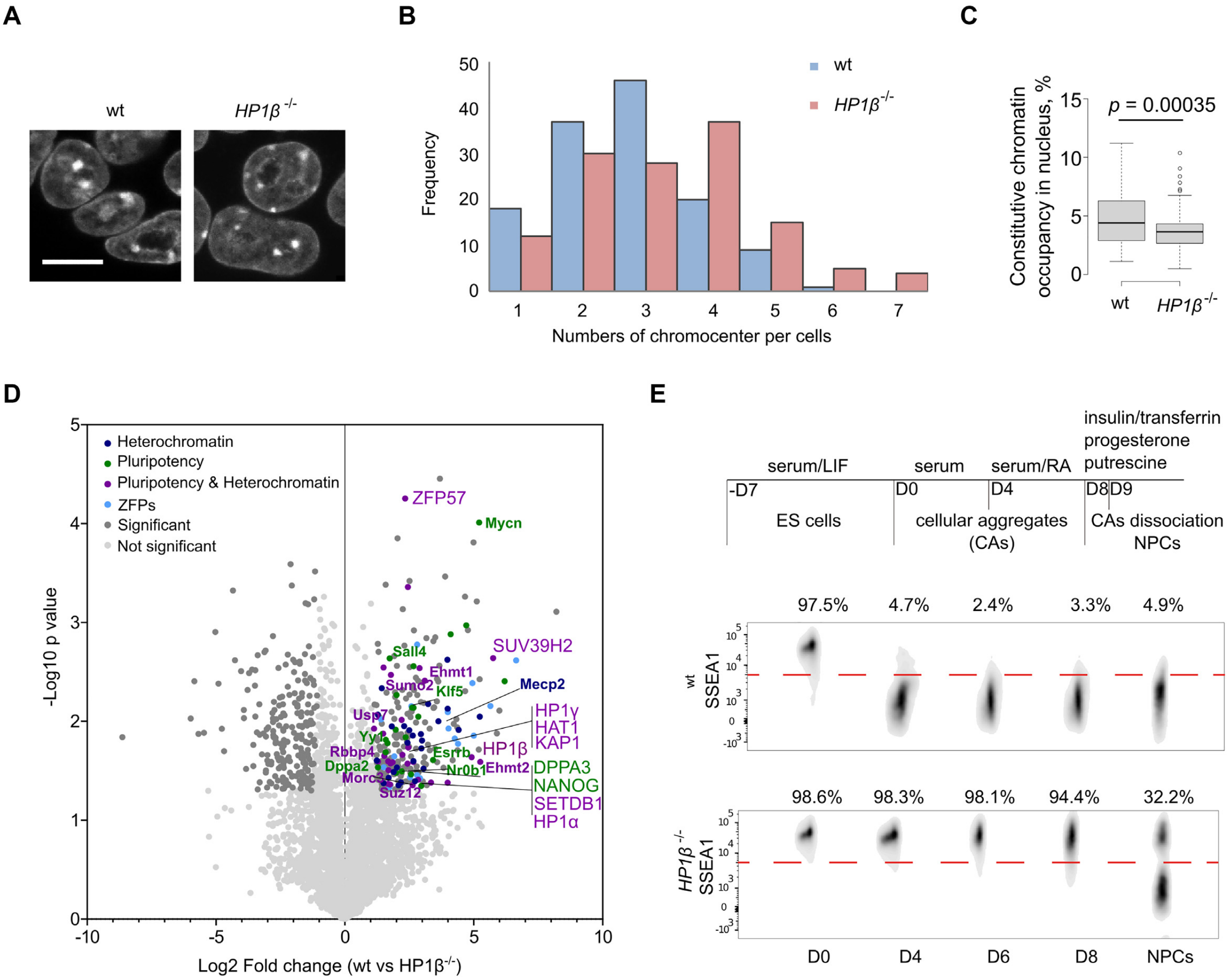


Figure 1. HP1β is required for neural progenitor cell (NPC) differentiation. (**A**–**C**) Depletion of HP1β leads to alterations in number and size of chromocenters. Images of mESCs stained with DAPI (A), scale bar: 10 μm. 131 nuclei for wt and *HP1β*$^{-/-}$ cells respectively were counted and frequency (y-axis) relative to the number of chromocenters per cell (x-axis) was plotted (B). The area of chromocenters and nucleus was measured with ImageJ to calculate the relative space occupied by chromocenters within the nucleus for wt and *HP1β*$^{-/-}$ mESCs as depicted in the box plot. Center lines show the medians; box limits indicate the 25th and 75th percentiles as determined by R software; whiskers extend 1.5× the interquartile range from the 25th and 75th percentiles, outliers are represented by dots. 109 and 102 individual cells were measured for wt and *HP1β*$^{-/-}$, respectively. Two-sided Student's t-test was done, **** $P < 0.001$ (C). (**D**) Volcano plot from HP1β ChIP-MS in wt and *HP1β*$^{-/-}$ mESCs ($n = 2$ biological replicates). Dark gray dots: significantly enriched proteins. Blue dots: proteins involved in heterochromatin regulation. Green dots: proteins involved in pluripotency. Purple dots: proteins involved in both heterochromatin and pluripotency. Cyan dots: zinc finger proteins (ZFPs). Statistical significance determined by performing a Student's *t* test with a permutation-based FDR of 0.05 and a cutoff of <2-fold enriched proteins. (**E**) Schematic representation of the NPC differentiation strategy and more details in Materials and Methods. Cells from distinct stages of differentiation were stained with a DyLight 650-conjugated anti-SSEA-1 antibody and analyzed by FACS.

HP1β is phosphorylated at serine 89 residue (HP1β-pS89) by casein kinase 2 (CK2)

To dissect the role of HP1β in pluripotency exit, we cultured mESCs with 2i/LIF (naïve) and serum/LIF (metastable state) media. In contrast to the naïve mESCs cultured with 2i/LIF, most cells in metastable state exhibit an altered transcriptional and epigenetic profile related to preimplantation epiblast cells (primed) (47,48). At the transcriptional level HP1β showed the lowest expression of all HP1 genes, with no significant changes at naïve and metastable state culturing conditions (Figure 2A and Supplementary Figure S2). However, we detected ∼2–3 fold increase of HP1β protein abundance by western blot analysis in the metastable state condition (Figure 2B).

Investigating possible posttranslational modifications of HP1β we noticed that GFP-HP1β purified from HEK293T cells migrates in coomassie stained protein gels as a double band of which the upper one disappeared upon incubation with antarctic phosphatase (Figure 2C). With mutational analyses, we mapped a phosphorylation at the serine 89 residue (Figures 2D, Supplementary Figure S3A and B). The phosphorylation of GFP-HP1β at the serine 89 residue was also detected with mESCs (Supplementary Figure S3C). To characterize the function of HP1β phosphorylation, we generated a monoclonal antibody against HP1β-

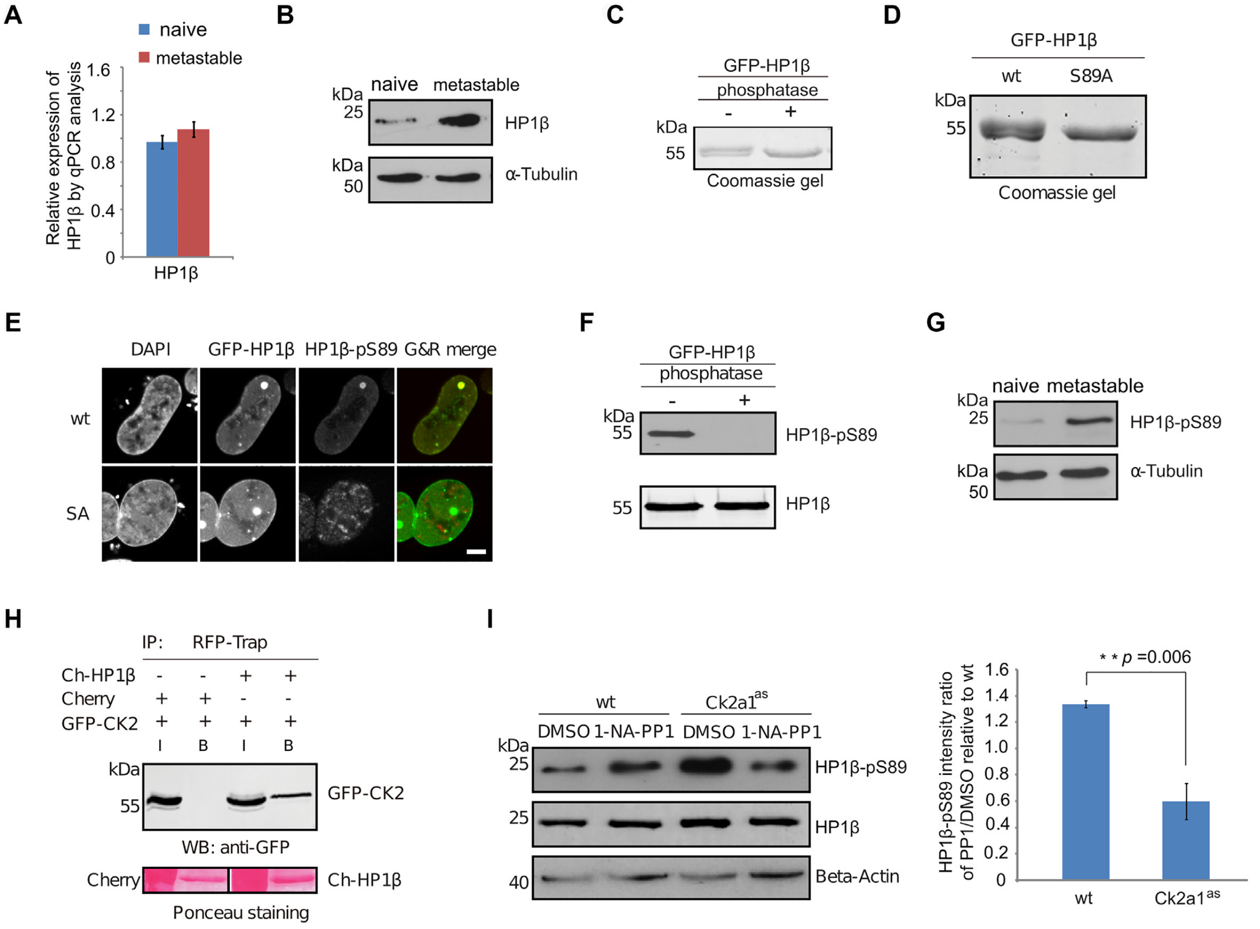


Figure 2. HP1β is phosphorylated at serine 89 residue. (**A**) Relative expression of HP1β in 2i/LIF (naïve) and metastable state conditions by RT-qPCR analysis. Values represent mean ± SEM from four biological replicates. (**B**) HP1β is upregulated in the metastable state condition. Total cell lysates of mESCs from naive and metastable culturing conditions were separated and visualized by anti-HP1β antibody. The anti-Tubulin blot was used as a loading control. (**C**, **D**) HP1β is highly phosphorylated on the serine 89 residue. GFP-HP1β purified from HEK293T cells was incubated with alkaline phosphatase and visualized in a coomassie stained gel (C). GFP-HP1β, wt and mutant, purified from HEK293T cells are visualized in a coomassie stained gel (D). (**E**) Characterization of a HP1β-pS89 monoclonal antibody by immunostaining. GFP-HP1β wt and mutant GFP-HP1β S89A fusion proteins were transiently expressed in BHK cells. HP1β proteins were anchored at a lac operator (*lacO*) array inserted in the genome and visible as a spot of enriched GFP fluorescence in the nucleus. Cell nuclei were stained with DAPI and HP1β proteins were visualized by the HP1β-pS89 antibody, scale bar: 5 μm. (**F**) Characterization of HP1β-pS89 monoclonal antibody by western blot. GFP-HP1β purified from HEK293T cells was incubated with alkaline phosphatase and visualized with anti-HP1β-pS89 antibody. (**G**) HP1β-pS89 is upregulated in the metastable condition. Total cell lysates of mESCs from naive and metastable culturing conditions were separated and visualized by anti-HP1β antibody. The anti-Tubulin blot was used as a loading control. (**H**) Co-immunoprecipitation shows an interaction between GFP-CK2 and Ch-HP1β. Cherry alone and cherry-tagged HP1β were immunoprecipitated from HEK293T cells co-transfected with GFP-CK2 using a RFP-Trap. Bound fractions were separated and visualized with an anti-GFP antibody and ponceau staining. (**I**) HP1β-pS89 is downregulated in the CK2a1^{as} cell line treated with 1-NA-PP1. Total cell lysates from wt and CK2a1^{as} mESCs treated with DMSO or 1-NA-PP1 were separated and visualized with anti-HP1β-pS89 and anti-HP1β antibodies. The anti-Actin blot was used as a loading control. Intensities of HP1β-pS89 were measured with ImageJ and normalized to the corresponding intensities of Actin before intensity ratios (1-NA-PP1/DMSO) were calculated. Values represent mean ± SEM of four biological replicates and the *P*-value of a two-sided Student's t-test is indicated.

pS89 (Figure 2E and F). With this antibody, we stained mouse rod photoreceptor cells, which display three distinct and spatially separated classes of chromatin, to assay for altered binding preferences of HP1β-pS89 but found a similar heterochromatin distribution as for HP1β (Supplementary Figure S3D). We observed an increase of HP1β-pS89 in the metastable condition by western blot in the absence of transcriptional changes (Figure 2G). This phosphorylation might stabilize HP1β and, thus, in the absence of transcriptional changes, contribute to increased protein levels at the transition from naive to primed state.

The serine 89 residue is located within a sequence of S/TxxE/D that is the consensus recognition motif for casein kinase 2 (CK2, Supplementary Figure S3A). As we also found a physical interaction between CK2 and HP1β (Figure 2H), we introduced a CK2a1 analog sensitive mutation (CK2a1^{as}) into wt mESCs by CRISPR-Cas9 (Supplementary Figure S3E). This analog sensitive mutation allows for rapid and highly specific CK2a1 inhibition with the adenine analog 1-NA-PP1 (49), which does not affect other kinases and wt cells. Upon addition of the adenine analog, we observed a clear reduction of HP1β-pS89 level (Figure 2I). Additionally, we treated cells expressing GFP-HP1β protein with the specific CK2 inhibitor 4,5,6,7-tetrabromobenzotriazole (TBB). Analysis of the phosphorylated to unmodified HP1β ratio in a coomassie stained protein gel indicated a clear reduction with TBB treatment (Supplementary Figure S3F). These results suggest that the phosphorylation of HP1β is catalyzed by CK2.

Phosphorylation enhances the phase separation of HP1β *in vitro*

Phase separation of HP1 is involved in regulation of heterochromatin formation (8,9). We recently showed that the charge of IDR-H determines the phase separation of HP1 homologues. HP1β forms phase separated droplets in the presence of core histones *in vitro* (11). The phosphorylation of serine 89 adds additional negative charge to the IDR-H of HP1β and lowers the p*I* to 5.3. To investigate the function of HP1β phosphorylation in phase separation, we purified HP1β wt and its mutants including HP1β S89A, HP1β S89E and HP1β S89D and incubated different amounts of the HP1β proteins (from 6 to 25 μM) with 25 μM histones. We collected phase-separated droplets by centrifugation and quantified the precipitated HP1β and histones with coomassie stained gels (Figure 3A and B). In contrast to the HP1β wt and non-phosphorylatable mutant HP1β S89A, the mutants mimicking HP1β phosphorylation (HP1β S89D and HP1β S89E), were more efficient in forming phase-separated droplets at the concentration of 25 μM as more histone H3 was depleted from supernatants and enriched in the pellets (Figure 3B). These results suggest that the phosphorylation of HP1β at S89 enhances phase separation in the presence of histones, probably through weak interactions between the acidic IDR-H of HP1β and basic histones.

HP1β-pS89 promotes mESCs exit from naïve pluripotent state

To investigate the function of HP1β S89 phosphorylation, we generated mESCs expressing either the non-phosphorylatable HP1β S89A or the phosphomimetic HP1β S89E (Supplementary Figure S4A and B). Western blot and immunostaining indicated that the mutant mESCs express a similar HP1β level to wt cells (Supplementary Figure S4C and D). In mESC cultures we noticed that $HP1\beta^{-/-}$ and HP1β S89A cells formed dome-shape colonies under metastable culture condition, while wt and HP1β S89E mESCs cultures were heterogeneous with mixed dome-shape and differentiated colonies. To quantify the morphology changes, we performed colony-formation assays and observed that $HP1\beta^{-/-}$ and HP1β S89A mESCs formed more naïve-like compact dome-shaped colonies (Figure 4A). We also observed that under metastable culture conditions $HP1\beta^{-/-}$ and HP1β S89A mESCs maintained the lower proliferation rate typical for the naïve state, while HP1β wt and HP1β S89E ESCs more than doubled (Supplementary Figure S4E). The fact that $HP1\beta^{-/-}$ and HP1β S89A mESCs continue to resemble naïve ESCs in terms of morphology and proliferation under metastable culture conditions suggests a possible defect in the exit from pluripotency.

To further investigate the role of HP1β we performed RNA-seq analysis of wt E14 and mutant mESCs, including $HP1\beta^{-/-}$, HP1β S89A and HP1β S89E cells cultured in both naive and metastable conditions. Principal component analysis (PCA) of transcriptomes revealed a significant separation between these two culture conditions reflecting the extensive changes in gene expression at the exit from pluripotency (Figure 4B). While the phosphorylation status of HP1β did not seem to matter in the naïve state, extensive differences were observed under metastable culture conditions. mESCs with the non-phosphorylatable HP1β S89A were widely separated from phosphomimetic HP1β S89E cells and closely resembled $HP1\beta^{-/-}$ cells in the PCA (Figure 4B).

Thus, the phosphorylation mutation (HP1β S89A) significantly affected the expression of 178 genes (fold change >1.5) in metastable state but only 12 genes in naive state (Figure 4C and Supplementary Table S2). Interestingly, HP1β S89A and HP1β S89E affected gene expression in opposing ways (Figure 4D). Gene Ontology (GO) enrichment analyses of biological processes showed that axis specification and cell differentiation were observed in both $HP1\beta^{-/-}$ and HP1β S89A (Supplementary Figure S5A). Also, we found that the dysregulated genes in $HP1\beta^{-/-}$ and HP1β S89A cells overlapped with the pluripotency cell fate (PCF) genes identified previously (50) (Supplementary Figure S5B).

Next, we further differentiated wt E14 and HP1β mutant mESCs to NPCs and analyzed their transcriptomes at distinct stages of differentiation (Supplementary Table S3). Notably, the transcriptomes of $HP1\beta^{-/-}$ and HP1β S89A cells showed dramatic changes in contrast to HP1β S89E, especially at the D0 of differentiation (Figure 4E). In agreement with the colony formation assay (Figure 4A), we found pluripotency genes such as *Tfcp2l1*, *Esrrb* and *Nanog* marker for naïve pluripotency state (51,52), upregulated in both $HP1\beta^{-/-}$ and HP1β S89A cells and slightly downregulated in HP1β S89E cells (Figure 4F and Supplementary Table S3). Collectively, these morphology, proliferation and gene expression data indicate that phosphorylation of HP1β at S89 is necessary for mESCs exit from the naïve pluripotent state.

HP1β-pS89 binds and sequesters KAP1 in heterochromatin compartments

To investigate how S89 phosphorylation affects the HP1β interactome, we generated HEK293T cell lines stably expressing either GFP-HP1β wt or the non-phosphorylatable mutant GFP-HP1β S89A or the phosphomimetic GFP-HP1β S89D. Since the serine 91 residue is close to serine 89 and was identified as an alternative phosphorylation site previously (53) (Supplementary Figure S6), it was also mutated to alanine in this assay. Interacting proteins were co-immunoprecipitated from cell extracts and compared by coomassie stained gels. A protein band specific for the phosphomimetic GFP-HP1β S89D was cut out and identified by MS analysis, as KAP1 (Figure 5A and Supplementary Figure S7). To test whether KAP1, as an interacting protein, is recruited by HP1β, we added truncated GFP-tKAP1 (aa 114–834) and found that it was specifically enriched in

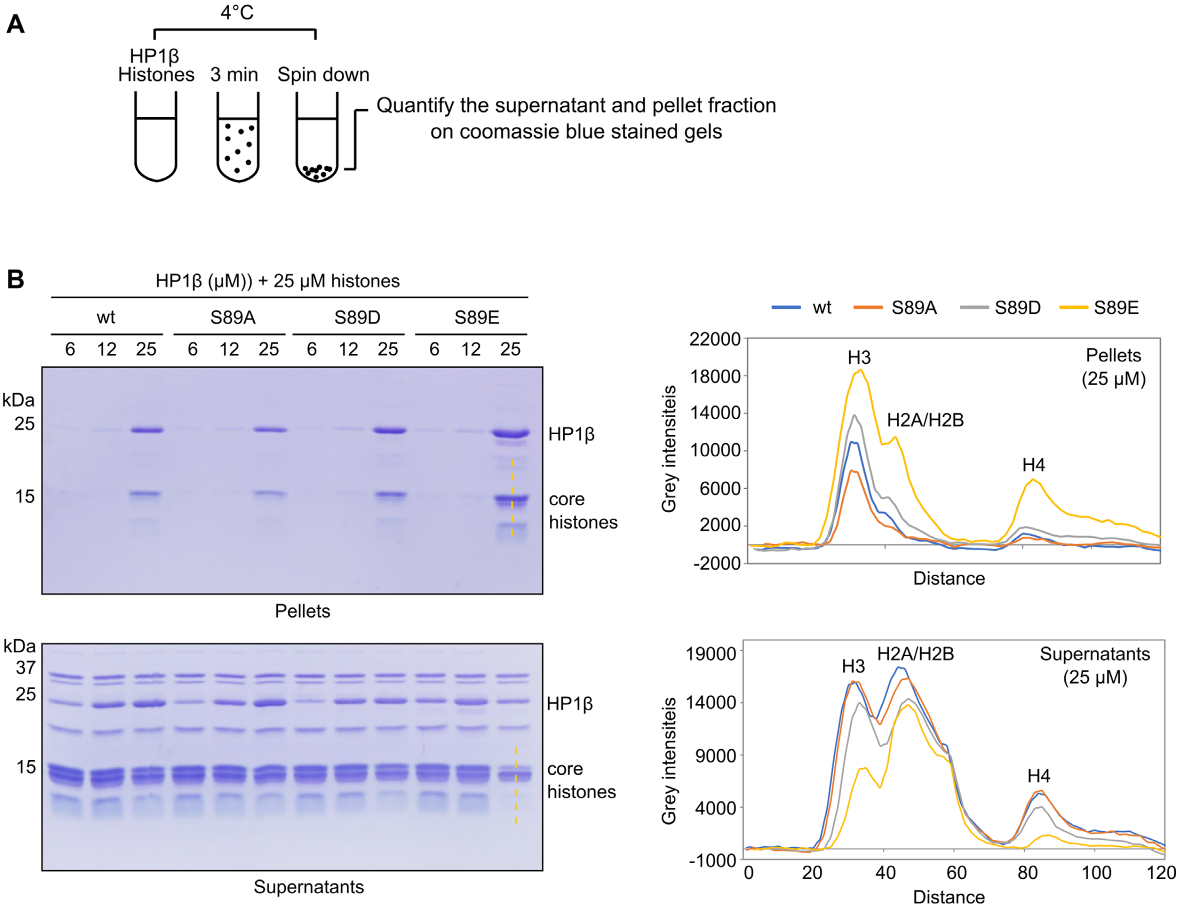


Figure 3. HP1β phosphorylation enhances its phase separation. (**A**) Illustration of the spin down assay to separate phase droplets from solution. (**B**) HP1β variants from 6 to 25 μM were incubated with 25 μM histones and phase-separated droplets were collected by centrifugation. Proteins in supernatants and pellets were visualized in coomassie stained gels. Line scans along the core histones in the supernatants and pellets of HP1β wt and mutant droplets at the concentration of 25 μM.

HP1β S89E phase-separated droplets *in vitro* (Figure 5B). Previously, the PxVxL motif of KAP1 (also known as HP1 box) was shown to bind the HP1 CSD (54,55). However, the observation that phosphorylation of S89 in the IDR-H of HP1β enhances the interaction with KAP1 (Supplementary Figure S7), suggests that KAP1 comprises a second site, besides the PxVxL motif, that specifically recognizes the phosphorylated HP1β.

In the amino acid sequence of mouse KAP1 the region from aa 247–376, known as coiled-coiled (CC) domain, stands out by its extreme basicity reaching a pI of 10.4 (Figure 5C) which makes it a good candidate to bind the acidic IDR-H of HP1β and to discriminate the phosphorylation at S89. A closer inspection of this CC domain revealed a striking similarity of the N-terminal aa 250–324 (CCN) with the C-terminus of H2B (helix α3 and helix αC) (Figure 5C). With a fluorescence three hybrid protein-protein interaction (F3H) assay (25,56), we could show that this CCN subdomain of KAP1 specifically binds HP1β-pS89 but not the non-phosphorylatable HP1β S89A (Figure 5D). These results fit well with the recent observation that sites within the IDR-H of HP1 interact with core histones (13).

To test the relative contribution of both, CCN and PxVxL, domains toward HP1β binding, we generated first an mESC line lacking KAP1 and expressing the phosphomimetic HP1β S89E (*KAP1*$^{-/-}$/$^{\mathrm{HP1\beta\ S89E}}$ cell line; Figure 6A). We, then, tested complementation with GFP-tKAP1 fusion proteins with mutated CCN (RH) and/or PxVxL (PVL) domains (Figure 5E) for enrichment at heterochromatic chromocenters. The comparison with wt GFP-tKAP1 shows that both single mutations reduce the enrichment at DAPI stained chromocenters and the double mutation (RH/PVL) mostly abolished KAP1 localization at chromocenters (Figure 5E and F). These results indicate that KAP1 CCN and PxVxL both contribute to HP1β binding and enrichment at chromocenters, whereby the CCN subdomain at the same time recognizes the S89 phosphorylation.

KAP1 contributes to pluripotency maintenance

As HP1β-pS89 regulates pluripotency exit and specifically interacts with KAP1, we further analyzed *KAP1*$^{-/-}$ mESCs. These cells were generated using a gRNA that targets the site after the first start codon (Figure 6A). PCR

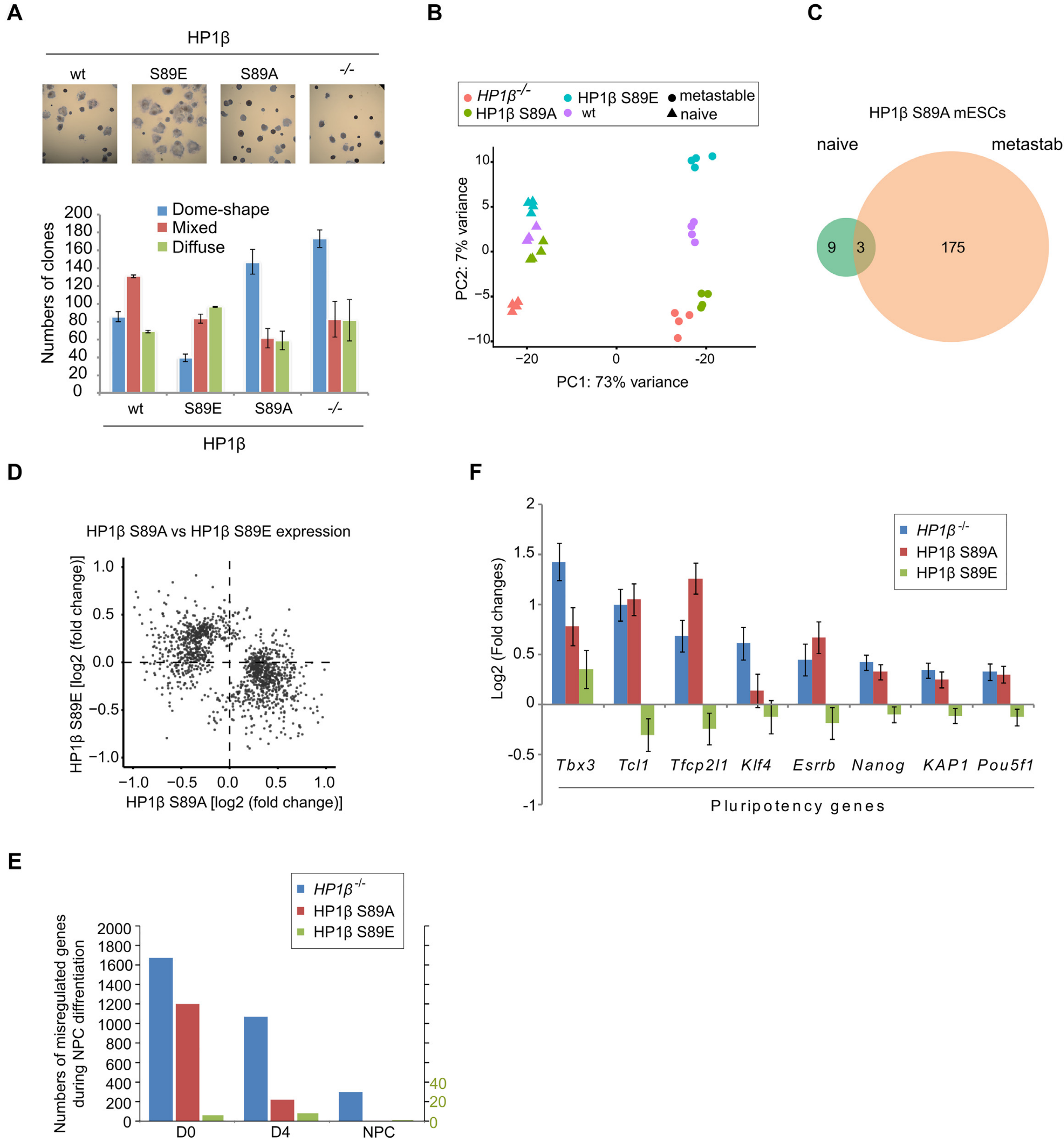


Figure 4. HP1β-pS89 promotes mESCs exit from naïve pluripotent state. (**A**) Representative images show alkaline phosphatase (AP) staining of wt E14 and HP1β mutant mESCs cultured in serum/LIF medium for 6 days. Numbers of dome-shape, diffuse and mixed colonies were counted, and values represent mean ± SEM from two different clones, each as a biological triplicate. (**B**) Principal component analysis (PCA) of whole transcriptome RNA-seq data from indicated cell lines in naive and metastable conditions. (**C**) Venn diagram showing dysregulated genes with fold changes >1.5 in HP1β S89A mESCs in naive and metastable conditions. (**D**) Scatter plot depicts overlapping dysregulated genes of HP1β S89A and HP1β S89E. (**E**) Bar plot showing the number of dysregulated genes from the transcriptomes of *HP1β*$^{-/-}$, HP1β S89A and HP1β S89E mESCs at the indicated stages of NPC differentiation. (**F**) Pluripotency genes found to be dysregulated in (E) were plotted for the respective cell lines.

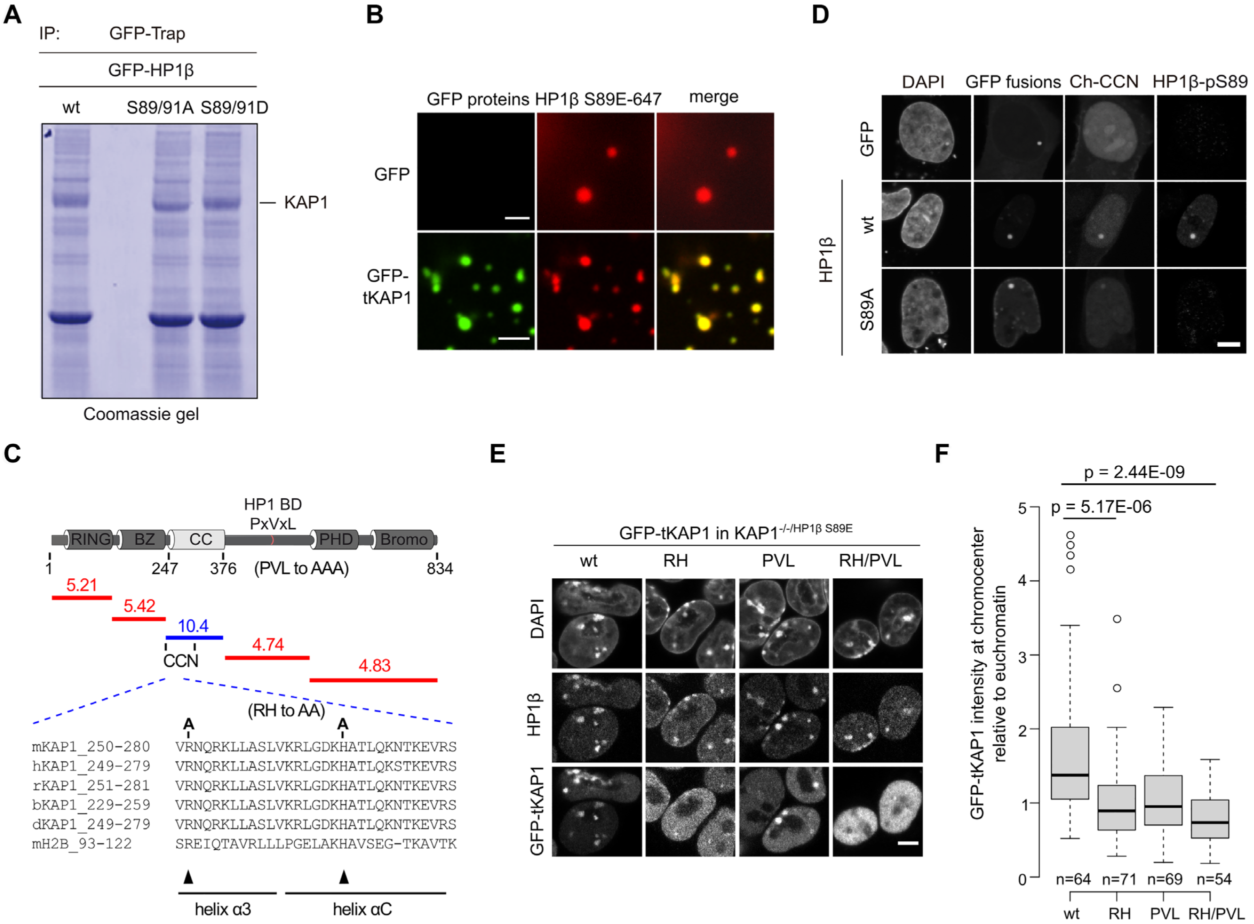


Figure 5. HP1β-pS89 interacts and recruits KAP1 to heterochromatin. (**A**) GFP-HP1β proteins immunoprecipitated using a GFP-Trap from HEK293T cells were separated and visualized by coomassie stained gels. A band showing more in GFP-HP1β wt and GFP-HP1β S89D, but less GFP-HP1β S89A, was cutted and sequenced by MS. (**B**) KAP1 is enriched in HP1β S89E phase-separated droplets *in vitro.* GFP and GFP-tKAP1 purified from HEK293T cells were incubated with 25 μM of HP1β S89E and histones in a buffer of 20 mM HEPES pH 7.2, 75 mM KCl and 1 mM DTT. 30 nM of HP1β S89E labeled with a NT-647 dye was added and phase-separated droplets were imaged using a 63x objective on a DeltaVision Personal Microscopy at 63 ×, scale bar: 5 μm. (**C**) Schematic illustration of KAP1 domains and their respective pI values. RING: really interesting new gene, BZ: B-box zinc finger, CC: Coiled-Coil, HP1 BD PxVxL: HP1 binding motif, PHD: plant homeodomains and Bromo domains. The N-terminus of CC (CCN) comprises a sequence (aa 250–280) that shares similarity with mouse histone H2B (aa 93–122). Conserved amino acids are highlighted in blue. (**D**) The CCN interacts with HP1β-pS89. To use the fluorescence three hybrid assay (F3H) (Herce *et al.*, 2013; Rothbauer *et al.*, 2008), GFP and GFP-HP1β fusion proteins as well as Ch-CCN were transiently expressed in BHK cells. GFP and GFP-HP1β proteins are anchored at a lac operator (*lacO*) array inserted in the BHK genome, thereby leading to a spot of enriched GFP fluorescence within the nucleus. While GFP-HP1β showed accumulation of Ch-CCN at the lacO spot, no or only weak interactions were detected for GFP and GFP-HP1β-SA, respectively. HP1β-pS89 was visualized with an anti-HP1β-pS89 antibody and nuclei were stained with DAPI, scale bar: 5 μm. (**E**) Images show $KAP1^{-/-}/^{\mathrm{HP1\beta\ S89E}}$ mESCs stably expressing either GFP-KAP1 wt or RH/PVL single or double mutation stained with an anti-HP1β antibody and DAPI, scale bar: 5 μm. (**F**) Quantification of chromocenter enrichment of GFP-tKAP1 wt and its mutations. GFP intensities in the chromocenters and euchromatic regions were measured with ImageJ and their ratio was calculated. Center lines depict the median; box limits indicate the 25th and 75th percentiles as determined by R software; whiskers extend 1.5x the interquartile range from the 25th and 75th percentiles; outliers are represented by circles. Individual chromocenters were analyzed (n = 64, 71, 69, 54 for GFP-tKAP1 wt, RH, PVL and RH/PVL, respectively). P values of a two-sided Student's t-test are indicated.

followed by sanger sequencing showed a 51 or 126 bp deletion within exon 1 of the KAP1 locus (Supplementary Figure S8A and B). Consistently, we did not detect KAP1 with an antibody against the N-terminus (aa 1–50) but observed faint shorter KAP1 bands with an antibody against the C-terminus of KAP1 (Figure 6B). Further analyses using mass spectrometry indicated that KAP1 protein, lacking the first N-terminal region, was present (Figure 6C). We only detected N-terminal peptides of KAP1 from wt ESCs (Figure 6C). In view of the very low level of KAP1 protein in the mutant cells (Figure 6B), they can be used as $KAP1^{-/-}$ mESCs. We next performed RNA-seq analysis of $KAP1^{-/-}$ mESCs and analyzed the GO term enrichment of upregulated and downregulated genes to biological processes (Supplementary Figure S9A and B). We observed cell differentiation in the GO terms of downregulated genes that is also found in both $HP1\beta^{-/-}$ and HP1β S89A. Among the misregulated genes in $KAP1^{-/-}$ mESCs (Supplementary Table S4), in particular naïve pluripotency genes, such as *Tfcp2l1*, *Tcl1, Esrrb* and *Nanog,* were downregulated, which is consistent with previous studies showing that KAP1 derepresses pluripotency genes (57). In-

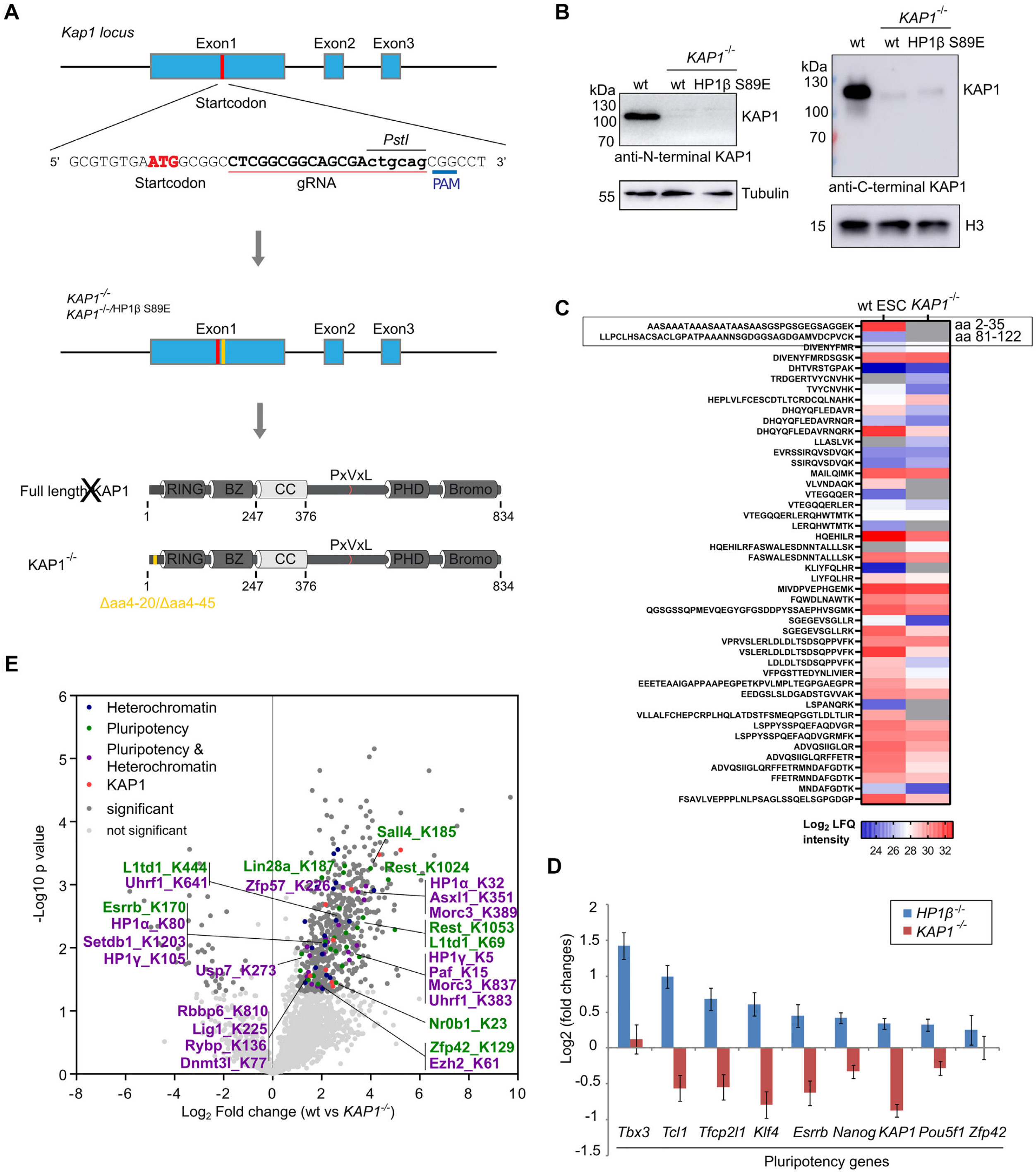


Figure 6. KAP1 relies on its ubiquitination activity to regulate pluripotency. (**A**) Schematic representation shows the CRISPR/Cas9 gene editing strategy used to generate *KAP1*$^{-/-}$ mESCs. gRNA target sequence and restriction enzyme recognition sites for screening are shown. (**B**) Western blot analysis of KAP1 protein levels in wt and *KAP1*$^{-/-}$ mESCs using antibodies against N- (left) and C-terminus (right) of KAP1. The tubulin and H3 blots were used as loading controls. (**C**) Mass spectrometry analyses of KAP1 expression in wt and *KAP1*$^{-/-}$ mESCs. (**D**) Volcano plot from diGly pulldowns in wt (n = 3 biological replicates) and *KAP1*$^{-/-}$ mESCs (n = 2 technical replicates). Dark gray dots: significantly enriched proteins. Blue dots: proteins involved in heterochromatin regulation. Green dots: proteins involved in pluripotency. Purple dots: proteins involved in both heterochromatin and pluripotency. Red dots: KAP1. Statistical significance determined by performing a Student's t test with a permutation-based FDR of 0.05 and an additional constant S0 = 1. (**E**) Plot of dysregulated pluripotency genes in the transcriptomes of *HP1β*$^{-/-}$ and *KAP1*$^{-/-}$ mESCs. Dark gray dots: significantly enriched proteins. Blue dots: proteins involved in heterochromatin regulation. Green dots: proteins involved in pluripotency. Purple dots: proteins involved in both heterochromatin and pluripotency. Red dots: KAP1 peptides.

terestingly, *KAP1*$^{-/-}$ mESCs show the opposite effect on gene expression as *HP1β*$^{-/-}$ cells (Figure 6D) and resemble mESCs with phosphomimetic HP1β S89E (Figure 4F).

To investigate the mechanism of KAP1 in pluripotency maintenance, we further identified its ubiquitin targets by comparing wt and *KAP1*$^{-/-}$ cells by performing diGly pulldowns and mass spectrometry analyses as KAP1 has ubiquitin E3 ligase activity (58). Among the ubiquitin targets identified, we found the proteins that regulate heterochromatin, for example Setdb1, ZFP57, MORC3 and HP1, and several (naive) transcription factors, such as Sall4 and Esrrb (Figure 6E). These results suggest that KAP1 ubiquitinates heterochromatin regulators or transcription factors to regulate pluripotency.

Sequestration of KAP1 in heterochromatin by HP1β-pS89 promotes pluripotency exit

We used ActivinA/FGF to induce mESCs transition from naïve (0 h) to epiblast state (48 hr.) (59) and analyzed the interactome of HP1β at these two states by ChIP-MS. We observed a stronger interaction of HP1β with KAP1 and also with other heterochromatin regulators such as SUV39H1, SUV420H2 and HP1α at the epiblast state (Figure 7A). Co-immunoprecipitation using HP1β-pS89 antibodies showed that HP1β-pS89 interacts with KAP1 at the metastable state as compared with the naïve state (Supplementary Figure S10A). We hypothesized that HP1β-pS89 binds and sequesters KAP1 in the heterochromatin compartment causing *de facto* its functional depletion. To test this hypothesis, we knocked GFP into the *KAP1* locus to create a C-terminal fusion gene product (Supplementary Figure S10B) and monitored enrichment of KAP1 at chromocenters during pluripotency exit. We observed increasing chromocenter enrichment of KAP1 in wt mESCs during differentiation to epiblast state (Figure 7B). HP1β S89E showed efficient sequestration of KAP1 at chromocenters in both naïve and epiblast states in contrast to HP1β S89A (Figure 7B). These results suggest that the displacement of KAP1 to chromocenters is HP1β-pS89 dependent.

To synthetically mimic this sequestration, we expressed GFP binding nanobodies (GBP) fused with either a methylcytosine binding domain (MBD) to tether KAP1-GFP to chromocenters (MBD-GBP) (25,45) or Lamin B1 for tethering to the nuclear membrane (56) (Figure 7C). The sequestration of KAP1-GFP at nuclear envelope and chromocenters was monitored by fluorescence microscopy and correlated with decreased levels of the pluripotency protein NANOG (Figure 7D and E). These results support our hypothesis that KAP1 sequestration at chromocenters by HP1β-pS89 causes a functional depletion and a downregulation of pluripotency genes.

Altogether, our results show that phosphorylation of HP1β at S89 generates a specific binding site for KAP1 and thereby captures this essential regulator of pluripotency (Figure 7F).

DISCUSSION

HP1 proteins bind H3K9me3 and regulate chromatin organization during cell differentiation. We found that the *HP1β*$^{-/-}$ mESCs are defective in NPC differentiation. This result is consistent with a previous finding showing an impaired neuronal precursor differentiation in mouse brain (17). We found that the pluripotency exit depends on a phosphorylation of HP1β at the serine 89 residue (HP1β-pS89), as we observed similar alterations in *HP1β*$^{-/-}$ and HP1β S89A cells at the D0 of NPC differentiation. However, only a few genes in HP1β S89A ESCs show altered expression at the NPC stage. These results suggest that HP1β-pS89 contributes to the pluripotency exit, but it is not required for the late stage of NPC differentiation.

With mutation analyses, we identified that the HP1β-pS89 is catalyzed by CK2 in cells, which is in line with *in vitro* phosphorylation assay following mass spectrometry analyses (60,61). This phosphorylation generates a specific binding site for KAP1 that provides a link to pluripotency as KAP1 has been identified as an essential factor that represses differentiation-inducible and derepresses pluripotency-associated genes (57,62–65). Consistent with this observation we found that deletion of KAP1 causes a downregulation of pluripotency genes. We identified ubiquitin targets of KAP1, such as MORC3 and HP1. Their ubiquitination may release these regulators from the promoter region that facilitates the expression of pluripotency genes. The key role of HP1β-pS89 phosphorylation in controlling this interaction with KAP1 and the exit from naive pluripotency becomes apparent from the opposite phenotypes of mESC lines with specific phosphorylation mutations. While the phospho-mimicking HP1β S89E promoted the exit from naive pluripotency, the non-phosphorylatable mutant HP1β S89A impairs this transition.

We also found that the binding of KAP1 requires the phosphorylation of HP1β at S89 in the IDR-H. In addition to the known PxVxL HP1 binding motif, which had been reported to be essential for early development (66) we identified the N-terminal part of the coiled-coil domain (CCN) of KAP1 as a second binding domain that discriminates the phosphorylation state of HP1β. Furthermore, we found that the binding of KAP1 to phosphorylated HP1β at heterochromatic chromocenters causes a depletion of free KAP1 in the nucleoplasm. We reproduced this KAP1 depletion by fusing KAP1 with GFP and captured the fusion protein at chromocenters and at the nuclear lamina with a GFP binding nanobody (GBP) fused to a methylcytosine binding domain (MBD) and lamin B, respectively. This synthetic capture caused a depletion of available KAP1-GFP and a concomitant downregulation of the NANOG pluripotency factor.

The naïve, formative and primed pluripotency states of stem cells are characterized and maintained by distinct transcriptional networks (48,50–52,59,67–69). We used 2i/LIF and serum/LIF to maintain mESCs at the naïve and metastable states, respectively. As most mESCs in metastable state exhibit an altered transcriptional and epigenetic profile relative to preimplantation epiblast cells (primed), we analyzed the cells from these two culture conditions to investigate the naïve pluripotency exit. Restricting the nuclear localization of one of these factors may destabilize the pluripotency network as was shown for the bHLH transcription factor Tfe3 (70). Our results suggest that the binding to HP1β-pS89 in chromocenters restricts

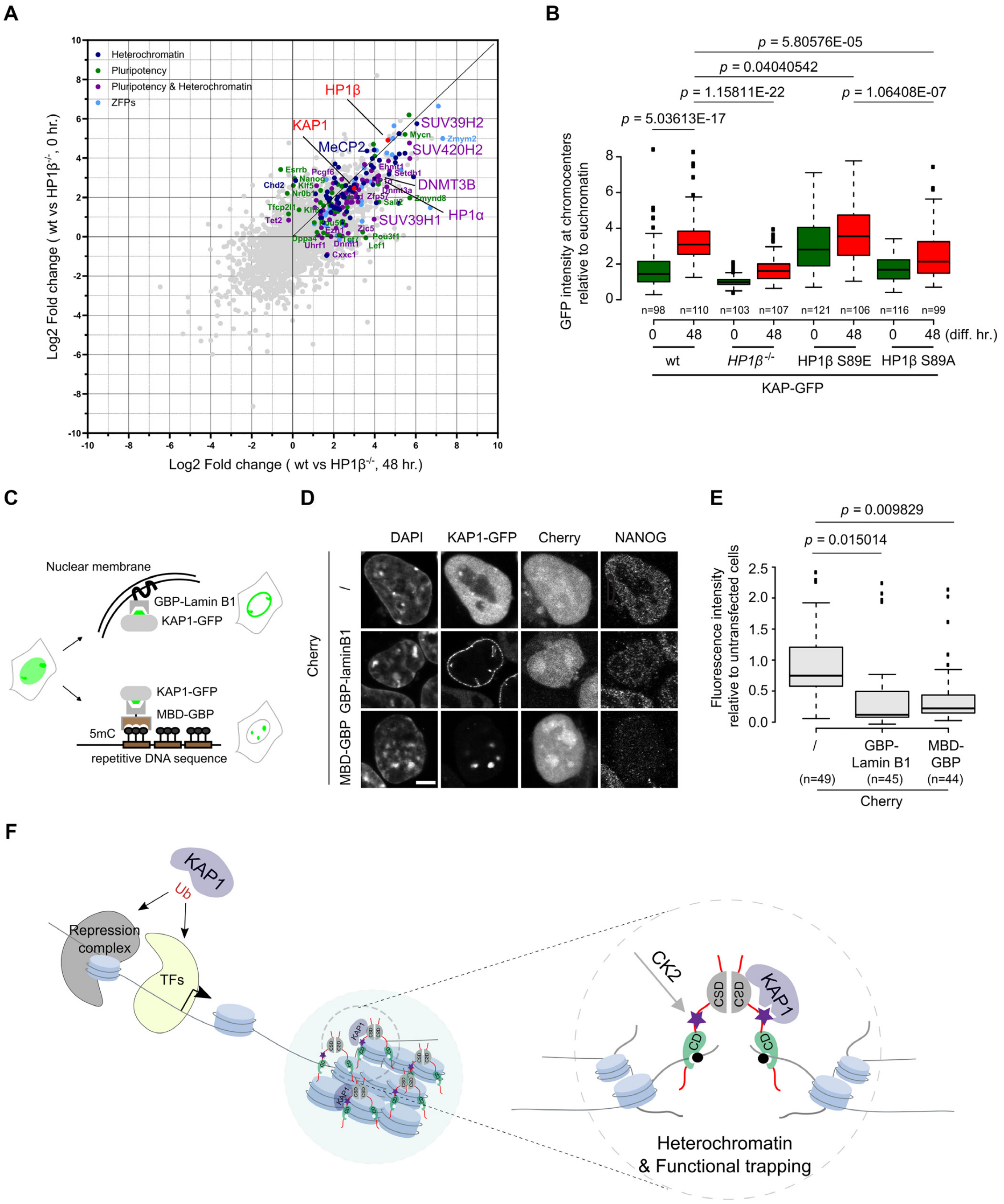


Figure 7. HP1β-pS89 sequesters KAP1 into heterochromatin to promote mESCs exit from pluripotency. (**A**) Comparison of the HP1β ChIP-MS under naive (0 h) and epiblast states (48 h). (**B**) KAP1 is recruited to chromocenters by HP1β-pS89 during pluripotency exit. Box plot depicts the intensity of KAP1-GFP at chromocenters relative to the signal at euchromatic regions in GFP knockin cell lines at the naïve (0h) and epiblast (48 h) state, respectively. Center lines show the medians; box limits indicate the 25th and 75th percentiles as determined by R software; whiskers extend 1.5x the interquartile range from the 25th and 75th percentiles, outliers are represented by dots. The number of chromocenters (n) analyzed for each sample is indicated. *P* values from a two-sided Student's t-test are indicated. (**C**) Schematic representation of tethering KAP1-GFP to the nuclear envelope and chromocenters by using GBP-Lamin B1 and MBD-GBP, respectively. (**D**) Representative images of *HP1*$\beta^{-/-}$ cells ectopically expressing Cherry in combination with GBP-Lamin B1 or MBD-GBP stained with NANOG and DAPI, scale bar: 5 μm. (**E**) Box plots depict relative levels of the pluripotency protein NANOG for cells showing nuclear envelope and chromocenter tethering of GFP-tagged KAP1. Fluorescence intensities in nuclei were measured with ImageJ and normalized to the signals for untransfected cells. Center lines show the medians; box limits indicate the 25th and 75th percentiles as determined by R software; whiskers extend 1.5× the interquartile range from the 25th and 75th percentiles, outliers are represented by dots. The number of cells (*n*) analyzed for each sample is indicated. Two-sided Student's *t*-test was performed, and p values are indicated. (**F**) HP1β dimerizes and binds H3K9me3 clustering chromatin to form heterochromatin compartments. In response to pluripotency exit, HP1β is phosphorylated at serine 89 residue (HP1β-pS89) by CK2, thereby sequestering KAP1 into heterochromatin compartments. KAP1 relies on its ubiquitination/sumoylation activity to regulate pluripotency. The sequestration of KAP1 leads to downregulation of pluripotency genes allowing mESCs to exit pluripotency.

the nuclear availability of KAP1 and thereby impairs the expression of pluripotency genes and promotes the exit from pluripotency.

Phase separation has been described as a novel mechanism to locally gather and enrich factors to activate genes and to enhance transcription (71–73). Our results now suggest an opposite mechanism to negatively regulate transcription. The phosphorylation of HP1β at chromocenters creates a specific binding site for the transcription regulator KAP1. This capture of an essential regulator of pluripotency genes promotes the exit from pluripotency. In addition, a previous publication suggests that the capture of KAP1 could enhance the phase separation of HP1β/nucleosomes and heterochromatin organization (15). These results also outline a new function of heterochromatin as a subnuclear compartment to capture regulatory factors and thereby remotely control gene activation and transcription at distant parts of the genome representing a novel form of remote control of transcriptional regulation.

DATA AVAILABILITY

Sequencing data reported in this paper are available at ArrayExpress (EMBL-EBI) under accessions ‘E-MTAB-8329’ (RNA-seq).

The raw mass spectrometry proteomics data have been deposited at the ProteomeXchange Consortium via the PRIDE partner repository with the dataset identifier ‘PXD025053’.

The flow cytometry data have been deposited to FlowRepository (https://flowrepository.org/) with repository ID: FR-FCM-Z3MZ.

Supplementary Table S2 contains the list of differentially expressed genes of HP1β S89A and S89E cells at naive and metastable conditions, related to Figure 4C and D. Supplementary Table S3 contains the list of differentially expressed genes in *HP1β*$^{-/-}$, HP1β S89A and S89E cells during NPC differentiation, related to Figure 4E and F. Supplementary Table S4 contains the list of differentially expressed genes of *KAP1*$^{-/-}$ cells at the metastable condition, related to Figure 6D.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank Geeta Narlikar and Gary Karpen (University of California, USA) for encouragement and comments. We thank Masahiro Muto and Haruhiko Koseki for mouse E14 wt ESCs; David L. Spector for providing BHK cells containing a lac operator repeat array. We thank Stephanie Link and Martina Antoniolli for help with the RNA isolation and plasmid constructs as well as Pin Zhao and Ningjun Duan for technical help with cell line generation. We thank Carina Trummer for comments on the manuscript. We thank Matthias Mann for access to mass spectrometers and Igor Paron and Christian Deiml for the technical maintenance of MS machines. E.U. is a fellow of the International Max Planck Research School for Molecular Life Sciences (IMPRS-LS).

Author contributions: W.Q. conceived the study, designed, performed and analyzed experiments and wrote the manuscript. C.M., S.B. and M.S. performed RNA-seq and analyzed data. C.M. generated KAP1-GFP knockin mESCs. I.S. performed immunostaining in the retina. E.U. performed ChIP-MS and diGly enrichments. E.U., M.W. and I.F. performed MS analysis. W.Q. performed phase separation experiments. M.M. performed FACS analysis. M.C.C. and A.I. discussed the project and contributed to the manuscript writing. H.L. supervised the study and wrote the manuscript. All authors discussed the data and edited the manuscript.

FUNDING

Deutsche Forschungsgemeinschaft (DFG) [SFB1064/A17, LE 721/18-1 to H.L., 213249687 and 425470807, SFB1064/A22 to S.B., 213249687 and CA 198/16-1 to M.C.C., 425470807]. Funding for open access charge: Deutsche Forschungsgemeinschaft (DFG).

Conflict of interest statement. None declared.

REFERENCES

1. Eissenberg,J.C. and Elgin,S.C. (2000) The HP1 protein family: getting a grip on chromatin. *Curr. Opin. Genet. Dev.*, **10**, 204–210.
2. Li,Y., Kirschmann,D.A. and Wallrath,L.L. (2002) Does heterochromatin protein 1 always follow code? *Proc. Natl. Acad. Sci. U.S.A.*, **99**, 16462–16469.
3. Bannister,A.J., Zegerman,P., Partridge,J.F., Miska,E.A., Thomas,J.O., Allshire,R.C. and Kouzarides,T. (2001) Selective recognition of methylated lysine 9 on histone H3 by the HP1 chromo domain. *Nature*, **410**, 120–124.
4. Jacobs,S.A. and Khorasanizadeh,S. (2002) Structure of HP1 chromodomain bound to a lysine 9-methylated histone H3 tail. *Science*, **295**, 2080–2083.
5. Nakayama,J., Rice,J.C., Strahl,B.D., Allis,C.D. and Grewal,S.I. (2001) Role of histone H3 lysine 9 methylation in epigenetic control of heterochromatin assembly. *Science*, **292**, 110–113.
6. Lavigne,M., Eskeland,R., Azebi,S., Saint-Andre,V., Jang,S.M., Batsche,E., Fan,H.Y., Kingston,R.E., Imhof,A. and Muchardt,C. (2009) Interaction of HP1 and Brg1/Brm with the globular domain of histone H3 is required for HP1-mediated repression. *PLoS Genet.*, **5**, e1000769.
7. Nielsen,A.L., Oulad-Abdelghani,M., Ortiz,J.A., Remboutsika,E., Chambon,P. and Losson,R. (2001) Heterochromatin formation in mammalian cells: interaction between histones and HP1 proteins. *Mol. Cell*, **7**, 729–739.
8. Larson,A.G., Elnatan,D., Keenen,M.M., Trnka,M.J., Johnston,J.B., Burlingame,A.L., Agard,D.A., Redding,S. and Narlikar,G.J. (2017) Liquid droplet formation by HP1alpha suggests a role for phase separation in heterochromatin. *Nature*, **547**, 236–240.
9. Strom,A.R., Emelyanov,A.V., Mir,M., Fyodorov,D.V., Darzacq,X. and Karpen,G.H. (2017) Phase separation drives heterochromatin domain formation. *Nature*, **547**, 241–245.
10. Keenen,M.M., Brown,D., Brennan,L.D., Renger,R., Khoo,H., Carlson,C.R., Huang,B., Grill,S.W., Narlikar,G.J. and Redding,S. (2021) HP1 proteins compact DNA into mechanically and positionally stable phase separated domains. *Elife*, **10**, e64563.
11. Qin,W., Stengl,A., Ugur,E., Leidescher,S., Ryan,J., Cardoso,M.C. and Leonhardt,H. (2021) HP1beta carries an acidic linker domain and requires H3K9me3 for phase separation. *Nucleus*, **12**, 44–57.
12. Erdel,F., Rademacher,A., Vlijm,R., Tunnermann,J., Frank,L., Weinmann,R., Schweigert,E., Yserentant,K., Hummert,J., Bauer,C. *et al.* (2020) Mouse heterochromatin adopts digital compaction states without showing hallmarks of HP1-driven liquid-liquid phase separation. *Mol. Cell*, **78**, 236–249.

13. Sanulli,S., Trnka,M.J., Dharmarajan,V., Tibble,R.W., Pascal,B.D., Burlingame,A.L., Griffin,P.R., Gross,J.D. and Narlikar,G.J. (2019) HP1 reshapes nucleosome core to promote phase separation of heterochromatin. *Nature*, **575**, 390–394.
14. Strickfaden,H., Tolsma,T.O., Sharma,A., Underhill,D.A., Hansen,J.C. and Hendzel,M.J. (2020) Condensed chromatin behaves like a solid on the mesoscale in vitro and in living cells. *Cell*, **183**, 1772–1784.
15. Wang,L., Gao,Y., Zheng,X., Liu,C., Dong,S., Li,R., Zhang,G., Wei,Y., Qu,H., Li,Y. *et al.* (2019) Histone modifications regulate chromatin compartmentalization by contributing to a phase separation mechanism. *Mol. Cell*, **76**, 646–659.
16. Eberhart,A., Feodorova,Y., Song,C., Wanner,G., Kiseleva,E., Furukawa,T., Kimura,H., Schotta,G., Leonhardt,H., Joffe,B. *et al.* (2013) Epigenetics of eu- and heterochromatin in inverted and conventional nuclei from mouse retina. *Chromosome Res.*, **21**, 535–554.
17. Aucott,R., Bullwinkel,J., Yu,Y., Shi,W., Billur,M., Brown,J.P., Menzel,U., Kioussis,D., Wang,G., Reisert,I. *et al.* (2008) HP1-beta is required for development of the cerebral neocortex and neuromuscular junctions. *J. Cell Biol.*, **183**, 597–606.
18. Mattout,A., Aaronson,Y., Sailaja,B.S., Raghu Ram,E.V., Harikumar,A., Mallm,J.P., Sim,K.H., Nissim-Rafinia,M., Supper,E., Singh,P.B. *et al.* (2015) Heterochromatin protein 1beta (HP1beta) has distinct functions and distinct nuclear distribution in pluripotent versus differentiated cells. *Genome Biol.*, **16**, 213.
19. Sharif,J., Muto,M., Takebayashi,S., Suetake,I., Iwamatsu,A., Endo,T.A., Shinga,J., Mizutani-Koseki,Y., Toyoda,T., Okamura,K. *et al.* (2007) The SRA protein Np95 mediates epigenetic inheritance by recruiting Dnmt1 to methylated DNA. *Nature*, **450**, 908–912.
20. Hayashi,K. and Saitou,M. (2013) Generation of eggs from mouse embryonic stem cells and induced pluripotent stem cells. *Nat. Protoc.*, **8**, 1513–1524.
21. Dambacher,S., Deng,W., Hahn,M., Sadic,D., Frohlich,J., Nuber,A., Hoischen,C., Diekmann,S., Leonhardt,H. and Schotta,G. (2012) CENP-C facilitates the recruitment of M18BP1 to centromeric chromatin. *Nucleus*, **3**, 101–110.
22. Mulholland,C.B., Smets,M., Schmidtmann,E., Leidescher,S., Markaki,Y., Hofweber,M., Qin,W., Manzo,M., Kremmer,E., Thanisch,K. *et al.* (2015) A modular open platform for systematic functional studies under physiological conditions. *Nucleic. Acids. Res.*, **43**, e112.
23. Ran,F.A., Hsu,P.D., Lin,C.Y., Gootenberg,J.S., Konermann,S., Trevino,A.E., Scott,D.A., Inoue,A., Matoba,S., Zhang,Y. *et al.* (2013) Double nicking by RNA-guided CRISPR Cas9 for enhanced genome editing specificity. *Cell*, **154**, 1380–1389.
24. Rottach,A., Kremmer,E., Nowak,D., Leonhardt,H. and Cardoso,M.C. (2008) Generation and characterization of a rat monoclonal antibody specific for multiple red fluorescent proteins. *Hybridoma (Larchmt)*, **27**, 337–343.
25. Herce,H.D., Deng,W., Helma,J., Leonhardt,H. and Cardoso,M.C. (2013) Visualization and targeted disruption of protein interactions in living cells. *Nat. Commun.*, **4**, 2660.
26. Shechter,D., Dormann,H.L., Allis,C.D. and Hake,S.B. (2007) Extraction, purification and analysis of histones. *Nat. Protoc.*, **2**, 1445–1457.
27. Bibel,M., Richter,J., Lacroix,E. and Barde,Y.A. (2007) Generation of a defined and uniform population of CNS progenitors and neurons from mouse embryonic stem cells. *Nat. Protoc.*, **2**, 1034–1043.
28. Marti,M., Mulero,L., Pardo,C., Morera,C., Carrio,M., Laricchia-Robbio,L., Esteban,C.R. and Izpisua Belmonte,J.C. (2013) Characterization of pluripotent stem cells. *Nat. Protoc.*, **8**, 223–253.
29. Ziegenhain,C., Vieth,B., Parekh,S., Reinius,B., Guillaumet-Adkins,A., Smets,M., Leonhardt,H., Heyn,H., Hellmann,I. and Enard,W. (2017) Comparative analysis of single-cell RNA sequencing methods. *Mol. Cell*, **65**, 631–643.
30. Renaud,G., Stenzel,U., Maricic,T., Wiebe,V. and Kelso,J. (2015) deML: robust demultiplexing of Illumina sequences using a likelihood-based approach. *Bioinformatics*, **31**, 770–772.
31. Martin,J.A. and Wang,Z. (2011) Next-generation transcriptome assembly. *Nat. Rev. Genet.*, **12**, 671–682.
32. Dobin,A., Davis,C.A., Schlesinger,F., Drenkow,J., Zaleski,C., Jha,S., Batut,P., Chaisson,M. and Gingeras,T.R. (2013) STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, **29**, 15–21.
33. Macosko,E.Z., Basu,A., Satija,R., Nemesh,J., Shekhar,K., Goldman,M., Tirosh,I., Bialas,A.R., Kamitaki,N., Martersteck,E.M. *et al.* (2015) Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell*, **161**, 1202–1214.
34. Parekh,S., Ziegenhain,C., Vieth,B., Enard,W. and Hellmann,I. (2018) zUMIs - a fast and flexible pipeline to process RNA sequencing data with UMIs. *Gigascience*, **7**, giy059.
35. Rau,A., Gallopin,M., Celeux,G. and Jaffrezic,F. (2013) Data-based filtering for replicated high-throughput transcriptome sequencing experiments. *Bioinformatics*, **29**, 2146–2152.
36. Love,M.I., Huber,W. and Anders,S. (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.*, **15**, 550.
37. Eden,E., Lipson,D., Yogev,S. and Yakhini,Z. (2007) Discovering motifs in ranked lists of DNA sequences. *PLoS Comput. Biol.*, **3**, e39.
38. Eden,E., Navon,R., Steinfeld,I., Lipson,D. and Yakhini,Z. (2009) GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC Bioinformatics*, **10**, 48.
39. Mulholland,C.B., Nishiyama,A., Ryan,J., Nakamura,R., Yigit,M., Gluck,I.M., Trummer,C., Qin,W., Bartoschek,M.D., Traube,F.R. *et al.* (2020) Recent evolution of a TET-controlled and DPPA3/STELLA-driven pathway of passive DNA demethylation in mammals. *Nat. Commun.*, **11**, 5972.
40. Rappsilber,J., Mann,M. and Ishihama,Y. (2007) Protocol for micro-purification, enrichment, pre-fractionation and storage of peptides for proteomics using StageTips. *Nat. Protoc.*, **2**, 1896–1906.
41. Cox,J. and Mann,M. (2008) MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.*, **26**, 1367–1372.
42. Cox,J., Neuhauser,N., Michalski,A., Scheltema,R.A., Olsen,J.V. and Mann,M. (2011) Andromeda: a peptide search engine integrated into the MaxQuant environment. *J. Proteome Res.*, **10**, 1794–1805.
43. Cox,J., Hein,M.Y., Luber,C.A., Paron,I., Nagaraj,N. and Mann,M. (2014) Accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction, termed MaxLFQ. *Mol. Cell. Proteomics*, **13**, 2513–2526.
44. Bertulat,B., De Bonis,M.L., Della Ragione,F., Lehmkuhl,A., Milden,M., Storm,C., Jost,K.L., Scala,S., Hendrich,B., D'Esposito,M. *et al.* (2012) MeCP2 dependent heterochromatin reorganization during neural differentiation of a novel Mecp2-deficient embryonic stem cell reporter line. *PLoS One*, **7**, e47848.
45. Brero,A., Easwaran,H.P., Nowak,D., Grunewald,I., Cremer,T., Leonhardt,H. and Cardoso,M.C. (2005) Methyl CpG-binding proteins induce large-scale chromatin reorganization during terminal differentiation. *J. Cell Biol.*, **169**, 733–743.
46. Dixon,J.R., Jung,I., Selvaraj,S., Shen,Y., Antosiewicz-Bourget,J.E., Lee,A.Y., Ye,Z., Kim,A., Rajagopal,N., Xie,W. *et al.* (2015) Chromatin architecture reorganization during stem cell differentiation. *Nature*, **518**, 331–336.
47. Hackett,J.A. and Surani,M.A. (2014) Regulatory principles of pluripotency: from the ground state up. *Cell Stem Cell*, **15**, 416–430.
48. Ying,Q.L., Wray,J., Nichols,J., Batlle-Morera,L., Doble,B., Woodgett,J., Cohen,P. and Smith,A. (2008) The ground state of embryonic stem cell self-renewal. *Nature*, **453**, 519–523.
49. Lopez,M.S., Kliegman,J.I. and Shokat,K.M. (2014) The logic and design of analog-sensitive kinases and their small molecule inhibitors. *Methods Enzymol.*, **548**, 189–213.
50. Fidalgo,M., Huang,X., Guallar,D., Sanchez-Priego,C., Valdes,V.J., Saunders,A., Ding,J., Wu,W.S., Clavel,C. and Wang,J. (2016) Zfp281 coordinates opposing functions of Tet1 and Tet2 in pluripotent states. *Cell Stem Cell*, **19**, 355–369.
51. Kinoshita,M., Barber,M., Mansfield,W., Cui,Y., Spindlow,D., Stirparo,G.G., Dietmann,S., Nichols,J. and Smith,A. (2021) Capture of mouse and human stem cells with features of formative pluripotency. *Cell Stem Cell*, **28**, 453–471.
52. Wang,X., Xiang,Y., Yu,Y., Wang,R., Zhang,Y., Xu,Q., Sun,H., Zhao,Z.A., Jiang,X., Wang,X. *et al.* (2021) Formative pluripotent stem cells show features of epiblast cells poised for gastrulation. *Cell Res.*, **31**, 526–541.
53. Yang,P., Humphrey,S.J., Cinghu,S., Pathania,R., Oldfield,A.J., Kumar,D., Perera,D., Yang,J.Y.H., James,D.E., Mann,M. *et al.*

(2019) Multi-omic Profiling Reveals Dynamics of the Phased Progression of Pluripotency. *Cell Syst.*, **8**, 427–445.

54. Nielsen,A.L., Ortiz,J.A., You,J., Oulad-Abdelghani,M., Khechumian,R., Gansmuller,A., Chambon,P. and Losson,R. (1999) Interaction with members of the heterochromatin protein 1 (HP1) family and histone deacetylation are differentially involved in transcriptional silencing by members of the TIF1 family. *EMBO J.*, **18**, 6385–6395.
55. Thiru,A., Nietlispach,D., Mott,H.R., Okuwaki,M., Lyon,D., Nielsen,P.R., Hirshberg,M., Verreault,A., Murzina,N.V. and Laue,E.D. (2004) Structural basis of HP1/PXVXL motif peptide interactions and HP1 localisation to heterochromatin. *EMBO J.*, **23**, 489–499.
56. Rothbauer,U., Zolghadr,K., Muyldermans,S., Schepers,A., Cardoso,M.C. and Leonhardt,H. (2008) A versatile nanotrap for biochemical and functional studies with fluorescent fusion proteins. *Mol. Cell. Proteomics*, **7**, 282–289.
57. Cheng,B., Ren,X. and Kerppola,T.K. (2014) KAP1 represses differentiation-inducible genes in embryonic stem cells through cooperative binding with PRC1 and derepresses pluripotency-associated genes. *Mol. Cell. Biol.*, **34**, 2075–2091.
58. Pineda,C.T., Ramanathan,S., Fon Tacer,K., Weon,J.L., Potts,M.B., Ou,Y.H., White,M.A. and Potts,P.R. (2015) Degradation of AMPK by a cancer-specific ubiquitin ligase. *Cell*, **160**, 715–728.
59. Hayashi,K., Ohta,H., Kurimoto,K., Aramaki,S. and Saitou,M. (2011) Reconstitution of the mouse germ cell specification pathway in culture by pluripotent stem cells. *Cell*, **146**, 519–532.
60. Munari,F., Gajda,M.J., Hiragami-Hamada,K., Fischle,W. and Zweckstetter,M. (2014) Characterization of the effects of phosphorylation by CK2 on the structure and binding properties of human HP1beta. *FEBS Lett.*, **588**, 1094–1099.
61. Sales-Gil,R. and Vagnarelli,P. (2020) How HP1 post-translational modifications regulate heterochromatin formation and maintenance. *Cells*, **9**, 1460.
62. Hu,G., Kim,J., Xu,Q., Leng,Y., Orkin,S.H. and Elledge,S.J. (2009) A genome-wide RNAi screen identifies a new transcriptional module required for self-renewal. *Genes Dev.*, **23**, 837–848.
63. Seki,Y., Kurisaki,A., Watanabe-Susaki,K., Nakajima,Y., Nakanishi,M., Arai,Y., Shiota,K., Sugino,H. and Asashima,M. (2010) TIF1beta regulates the pluripotency of embryonic stem cells in a phosphorylation-dependent manner. *Proc. Natl. Acad. Sci. USA*, **107**, 10926–10931.
64. Zhao,T. and Eissenberg,J.C. (1999) Phosphorylation of heterochromatin protein 1 by casein kinase II is required for efficient heterochromatin binding in Drosophila. *J. Biol. Chem.*, **274**, 15095–15100.
65. Zhao,T., Heyduk,T. and Eissenberg,J.C. (2001) Phosphorylation site mutations in heterochromatin protein 1 (HP1) reduce or eliminate silencing activity. *J. Biol. Chem.*, **276**, 9512–9518.
66. Herzog,M., Wendling,O., Guillou,F., Chambon,P., Mark,M., Losson,R. and Cammas,F. (2011) TIF1beta association with HP1 is essential for post-gastrulation development, but not for Sertoli cell functions during spermatogenesis. *Dev. Biol.*, **350**, 548–558.
67. Kalkan,T. and Smith,A. (2014) Mapping the route from naive pluripotency to lineage specification. *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, **369**, 20130540.
68. Kojima,Y., Kaufman-Francis,K., Studdert,J.B., Steiner,K.A., Power,M.D., Loebel,D.A., Jones,V., Hor,A., de Alencastro,G., Logan,G.J. *et al.* (2014) The transcriptional and functional properties of mouse epiblast stem cells resemble the anterior primitive streak. *Cell Stem Cell*, **14**, 107–120.
69. Smith,A. (2017) Formative pluripotency: the executive phase in a developmental continuum. *Development*, **144**, 365–373.
70. Betschinger,J., Nichols,J., Dietmann,S., Corrin,P.D., Paddison,P.J. and Smith,A. (2013) Exit from pluripotency is gated by intracellular redistribution of the bHLH transcription factor Tfe3. *Cell*, **153**, 335–347.
71. Boija,A., Klein,I.A., Sabari,B.R., Dall'Agnese,A., Coffey,E.L., Zamudio,A.V., Li,C.H., Shrinivas,K., Manteiga,J.C., Hannett,N.M. *et al.* (2018) Transcription factors activate genes through the phase-separation capacity of their activation domains. *Cell*, **175**, 1842–1855.
72. Lu,H., Yu,D., Hansen,A.S., Ganguly,S., Liu,R., Heckert,A., Darzacq,X. and Zhou,Q. (2018) Phase-separation mechanism for C-terminal hyperphosphorylation of RNA polymerase II. *Nature*, **558**, 318–323.
73. Sabari,B.R., Dall'Agnese,A., Boija,A., Klein,I.A., Coffey,E.L., Shrinivas,K., Abraham,B.J., Hannett,N.M., Zamudio,A.V., Manteiga,J.C. *et al.* (2018) Coactivator condensation at super-enhancers links phase separation and gene control. *Science*, **361**, eaar3958.

Kempf, J. M., Weser, S., Bartoschek, M. D., Metzeler, K. H., Vick, B., Herold, T., Völse, K., Mattes, R., Scholz, M., Wange, L. E., Festini, M., **Ugur, E.**, Roas, M., Weigert, O., Bultmann, S., Leonhardt, H., Schotta, G., Hiddemann, W., Jeremias, I., & Spiekermann, K. (**2021**). **Loss-of-function mutations in the histone methyltransferase EZH2 promote chemotherapy resistance in AML**. Scientific Reports, 11(1), 5838.

License: CC-BY-4.0

https://doi.org/10.1038/s41598-021-84708-6

www.nature.com/scientificreports

scientific reports

Check for updates

OPEN

Loss-of-function mutations in the histone methyltransferase EZH2 promote chemotherapy resistance in AML

Julia M. Kempf[1,10], Sabrina Weser[1,10], Michael D. Bartoschek[2], Klaus H. Metzeler[1], Binje Vick[3], Tobias Herold[1], Kerstin Völse[3], Raphael Mattes[1], Manuela Scholz[4], Lucas E. Wange[8], Moreno Festini[1], Enes Ugur[2], Maike Roas[1], Oliver Weigert[1], Sebastian Bultmann[2], Heinrich Leonhardt[2], Gunnar Schotta[5], Wolfgang Hiddemann[6,9], Irmela Jeremias[3,6,7] & Karsten Spiekermann[1,6,9]✉

Chemotherapy resistance is the main impediment in the treatment of acute myeloid leukaemia (AML). Despite rapid advances, the various mechanisms inducing resistance development remain to be defined in detail. Here we report that loss-of-function mutations (LOF) in the histone methyltransferase EZH2 have the potential to confer resistance against the chemotherapeutic agent cytarabine. We identify seven distinct EZH2 mutations leading to loss of H3K27 trimethylation via multiple mechanisms. Analysis of matched diagnosis and relapse samples reveal a heterogenous regulation of EZH2 and a loss of EZH2 in 50% of patients. We confirm that loss of EZH2 induces resistance against cytarabine in the cell lines HEK293T and K562 as well as in a patient-derived xenograft model. Proteomics and transcriptomics analysis reveal that resistance is conferred by upregulation of multiple direct and indirect EZH2 target genes that are involved in apoptosis evasion, augmentation of proliferation and alteration of transmembrane transporter function. Our data indicate that loss of EZH2 results in upregulation of its target genes, providing the cell with a selective growth advantage, which mediates chemotherapy resistance.

Acute myeloid leukaemia (AML) is a heterogeneous haematological malignancy, characterised by clonal expansion of abnormal, undifferentiated myeloid precursor cells. Even though many patients with AML respond well to induction chemotherapy, relapse and refractory disease are common, representing the major cause of treatment failure. Treatment with cytarabine (AraC) and daunorubicin (DNR) remains the standard care for AML patients, although several new therapeutic strategies have been implemented within the last years[1–3]. Epigenetic dysregulation of DNA methylation or histone modifications has been identified in many malignant tumors[4,5] and can be considered as a cause of cancer development and progression[6,7]. Since considerable insight concerning those epigenetic changes has been gained in recent years, many therapy concepts targeting the involved regulatory factors have been proposed and hold promise for novel treatment approaches[8,9].

Enhancer of zeste homolog 2 (EZH2) is a lysine methyltransferase found as the central core protein of the polycomb repressive complex 2 (PRC2)[10]. Comprising four subunits (SUZ12, EED, EZH2/EZH1 and RbAp46), this complex mediates transcriptional repression by catalysing the trimethylation of histone H3 at lysine 27 (H3K27me3)[11]. EZH2 has been found to serve a dual purpose, as either tumour suppressor or oncogene,

[1]Department of Medicine III, University Hospital, LMU Munich, Munich, Germany. [2]Department of Biology II and Center for Integrated Protein Science Munich (CIPSM), Human Biology and BioImaging, LMU Munich, Planegg Martinsried, Germany. [3]Research unit Apoptosis in Haematopoietic Stem Cells (AHS), Helmholtz Zentrum München, Munich, Germany. [4]Center for Human Genetics and Laboratory Diagnostic (AHC), Martinsried, Germany. [5]Biomedical Center and Center for Integrated Protein Science Munich, LMU Munich, Martinsried, Germany. [6]German Cancer Consortium (DKTK), Heidelberg, Germany. [7]Department of Pediatrics, Dr. von Hauner Children's Hospital, LMU Munich, Munich, Germany. [8]Anthropology and Human Genomics, Department of Biology II, Ludwig-Maximilians-University, Martinsried, Germany. [9]German Cancer Research Center (DKFZ), Heidelberg, Germany. [10]These authors contributed equally: Julia M. Kempf and Sabrina Weser. ✉email: karsten.spiekermann@med.uni-muenchen.de

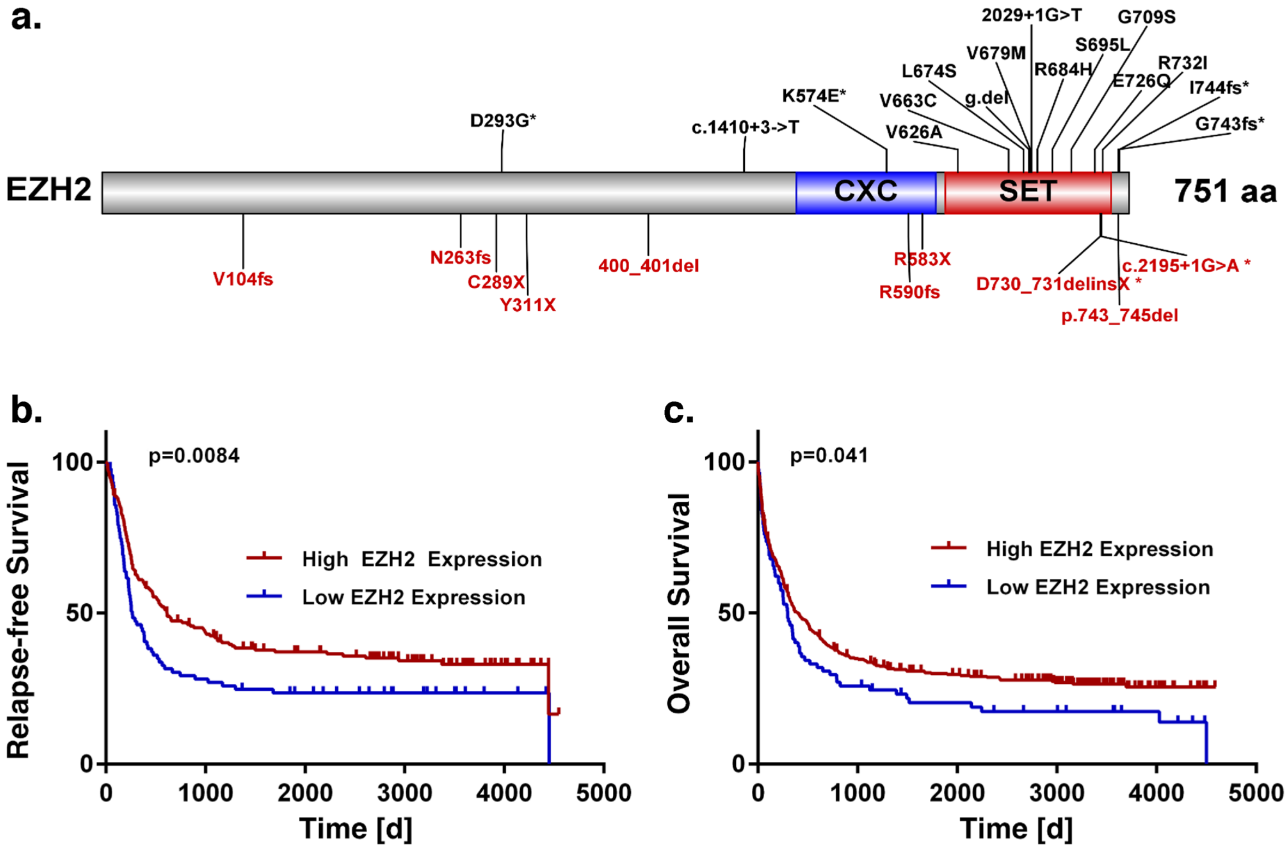


Figure 1. Recurrent *EZH2* mutations. (**a**) Schematic overview of EZH2 protein structure (NM_004456.4) and identified mutations (27 in total, c.2195+1G>A appeared twice) in a cohort of 664 AML patients at diagnosis. Functional domains are indicated at distinct locations and truncating mutations are displayed in red. Patients from Metzeler et al. 2016 (AMLCG-1999, AMLCG-2008). (**b**–**c**) Survival analysis of patients with low or high *EZH2* mRNA expression at the time point of diagnosis. *EZH2* high and low groups defined by the upper and lower quartile of *EZH2* mRNA expression, independent of mutation status. (**b**) Relapse-free survival (RFS). (**c**) Overall survival (OS). Patients from AMLCG 1999 (GSE37642), n = 517. 21 patients harboured an *EZH2* mutation. P-value calculated by log-rank test.

depending on the type of cancer[12–17]. In leukaemia, overexpression of EZH2 has been observed in CLL[18], paediatric T-ALL[19] and CML[20], while other studies reported EZH2 levels to be decreased in CMML[21] as well as ALL[18,19,22]. A recent study of Basheer et al. suggests opposing roles of EZH2 in initiation and maintenance of AML[23].

EZH1, an EZH2 homolog capable of partially compensating EZH2 function, holds an essential role in preserving pathological stem cells[24]. Therefore, it might contribute to the already complex role of EZH2 in hematopoietic malignancies[25–27]. Although EZH2 loss-of-function mutations seem to be rare in AML[28], loss of EZH2 by other mechanisms have been frequently reported and appear to play a major role in disease progression[29,30]. Absence of EZH2 in leukaemia cells was recently found to aberrantly activate BCAT1, resulting in enhanced mTOR signaling[27] and activation of the oncogene Hmga2 by causing an epigenetic switch from H3K27 trimethylation to H3K27 acetylation[31]. Furthermore, reduced disease-free survival was found to be associated with EZH2 mutations in myeloid malignancies[28,32,33] including AML[23]. In addition, chemoresistance was found in a recent study on AML patients with poor prognosis and downregulated EZH2[34].

In our previous study[35], examining diagnosis/relapse pairs of 50 cytogenetically normal (CN) AML patients, we found mutations in epigenetic modifiers, including EZH2, frequently gained at relapse, suggesting epigenetic mechanisms to be involved in disease progression in a subset of patients. The current study aims to evaluate the importance of chemotherapy resistance in AML. We investigated *EZH2* mutations and their functional loss of methyltransferase activity using patient samples, *in vivo* and *in vitro* patient-derived xenografts (PDX), and haematopoietic cell lines. We found *EZH2* loss-of-function mutations to be involved in the development of resistance against cytarabine and observed upregulation of EZH2 target genes due to loss of H3K27 trimethylation.

Results

Recurrent *EZH2* mutations at diagnosis.

In our previous work, we analysed 664 AML patients to study recurrently mutated genes, including *EZH2*[36]. In this cohort, 25 patients (4 %) carried an *EZH2* mutation at the time of diagnosis (27 mutations in total, Fig. 1a). Most of these mutations (n = 20, 74%) were located in the SET ([Su(var)3-9, Enhancer-of-zeste and Trithorax]) or CXC (cysteine-rich region, sometimes referred to as pre-SET) domain at the C-terminus of the protein and are responsible for the catalytic activity of the methyltransferase. Furthermore, 41% (11) of mutations cause a stop-gain or frameshift, resulting in a truncated protein. An additional two frameshift mutations result in an elongated protein variant. Mutations most frequently co-occurring with mutated *EZH2* were found in *RUNX1*, *ASXL1*, *DNMT3A* and *TET2* (44%, 40%, 20% and 20%, Supplementary Fig. 1a). Additionally, *RUNX1* and *ASXL1* mutations were found to occur more often in *EZH2* mutated patients (44% and 40%) than in *EZH2* wild type patients (14% and 10%, p = 4.6e−04, and p = 9.3 e−05, Fisher's Exact Test). In contrast, NPM1, the most frequently mutated gene in our cohort, was found to be mutated less often in *EZH2* mutated (12%) than in *EZH2* wild type patients (34%) (p = 2.8 e−02, Fisher's Exact Test). Interestingly, *KDM6A* and *EZH2* mutations were found to be mutually exclusive. Most patients with *EZH2* mutations (76%, n = 19) can be assigned to the adverse risk group (Supplementary Fig. 1a), according to the recent ELN classification[37].

In order to evaluate the prognostic importance of *EZH2,* we examined the survival of patients dependent on their *EZH2* mutation and expression status. The overall survival (OS) of patients harbouring *EZH2* mutations did not differ significantly from patients without mutation (Supplementary Fig. 1b). However, low *EZH2* mRNA expression was significantly associated with poor relapse-free survival (RFS) and OS in publicly available independent data sets of the AMLCG 1999 trial (GSE37642, Fig. 1b-c) and HOVON (GSE14468, Supplementary Fig. 1) study groups[38–40]. Additionally, monosomy 7, resulting in reduced *EZH2* expression, was associated with poor overall survival (Supplementary Fig. 1c).

Relevance of EZH2 status in AML relapse.

To further investigate the poor survival in patients with low *EZH2* mRNA expression, we compared protein expression in a set of matched diagnosis and relapse pairs of ten AML patients without *EZH2* mutations (Supplementary Table 2). In 50% of patients, we observed decreased levels of EZH2 protein expression, whereas the other half revealed increased protein expression levels in relapse (Fig 2a). An increase of at least 2-fold in protein expression was found in four patients, whereas a strong decrease (2-fold or more) in protein expression was observed in three patients. An additional analysis of *EZH2* mRNA expression in 32 CN-AML patients revealed a similar heterogenous picture. Downregulation of *EZH2* was found in 22% of patients, while upregulation was found in 53% (Fig. 2b). Additionally, we identified two relapse-associated *EZH2* mutations. EZH2/p.A692G found in the second relapse of patient CN-021 from the Greif et al. cohort[35] and EZH2/Y733LfsX6 found in the first relapse of a patient from the AML-CG cohort. Both mutations revealed subclonal outgrowth during the course of treatment and increasing variant allele frequencies (VAFs) in relapsed patients (Fig. 2c). Additionally, we found an increase of VAFs in the relapse of three other EZH2 mutations found in the Greif et al. cohort[35] (Supplementary Fig. 2).

Functional characterisation of *EZH2* mutations.

To evaluate the biochemical activity of the EZH2 variants, we measured global H3K27 trimethylation levels in a 293T/*EZH2*$^{-/-}$ model (Fig. 3a), which was established through CRISPR/Cas9 mediated genome editing, targeting exon 3 of *EZH2.* We found EZH2 protein expression levels to be strongly correlated with global H3K27me3 levels (Fig. 3b). In fact, global H3K27me3 was not detectable in any of the tested *EZH2*$^{-/-}$ clones, whereas *EZH2*$^{+/-}$ clones showed decreased EZH2 expression as well as reduced global H3K27me3 levels (Fig. 3a). Since EZH2 is only one part of the PRC2 complex, we additionally analysed protein expression of the remaining components SUZ12, RBAP46 and EED. We could not detect aberrant expression of these subunits in both 293T/*EZH2*$^{-/-}$ clones (Fig. 3c). Interestingly, both clones showed an increased resistance against AraC compared to the wild type clones (Fig. 3d, Supplementary Fig. 3a) and a slightly reduced colony count was observed in a colony formation assay (Supplementary Fig. 3c). Re-expression of seven different EZH2 variants, (Supplementary Fig. 3b) found in the AML-CG-1999 and AML-CG-2008 studies, could only partially rescue global H3K27me3 levels, indicating a LOF phenotype, while the re-expressed wildtype protein was able to restore complete activity (Fig. 3e).

In addition, co-expression of these variants with wild type *EZH2* led to a reduction of H3K27me3 levels in four mutations, suggesting a dominant-negative effect (Fig. 3f).

To validate the robustness of our 293T/*EZH2*$^{-/-}$ model, we performed the rescue experiment with two previously described EZH2 variants. EZH2/p.Y646N, a gain-of-function mutation found in lymphomas[16,17] and EZH2/Y731, a LOF mutation[41]. We were able to verify the functions of both mutations with our model (Supplementary Fig. 3d and e).

EZH2 depletion promotes resistance in K562 cells.

In order to study the impact of *EZH2* mutations on chemoresistance in a hematopoietic context, we screened 12 AML cell lines with *EZH2* mutant or wild type background (Supplementary Table 1). We identified two *EZH2* mutated cell lines, SKM-1 and KG-1a, that seem to be more resistant against cytarabine and daunorubicin, respectively (Fig. 4a). Notably, three of the *EZH2*^{wt} cell lines were harbouring *KDM6A* mutations, which can also affect drug resistance[42]. Furthermore, we found a strong positive correlation between H3K27me3 levels and EZH2 protein expression (Supplementary Fig. 3a). Next, we established seven *EZH2*$^{-/-}$ single cell (sc) knockout clones in the myeloid cell line K562, using CRISPR Cas9 genome editing (Fig. 4b, Supplementary Fig. 4e). Both knockout and control cells were treated for 72 h with either AraC or DNR. In K562/*EZH2*$^{-/-}$ clones, increased chemoresistance was found against AraC, while sensitivity against DNR was not affected (Fig. 4c, Supplementary Fig. 4c). Additionally, we observed reduced proliferation in K562/*EZH2*$^{-/-}$ clones compared to K562/*EZH2*$^{+/+}$ clones (Fig. 4d). Furthermore, the response towards AraC treatment was studied in a long-term proliferation assay, consequently treating single cell clones for 12 days with a low dose of AraC. In accordance with the short-term assay, also the K562/*EZH2*$^{-/-}$ sc clones of the long-term assay displayed higher resistance against AraC (Fig. 4e, Supplementary Fig. 4d).

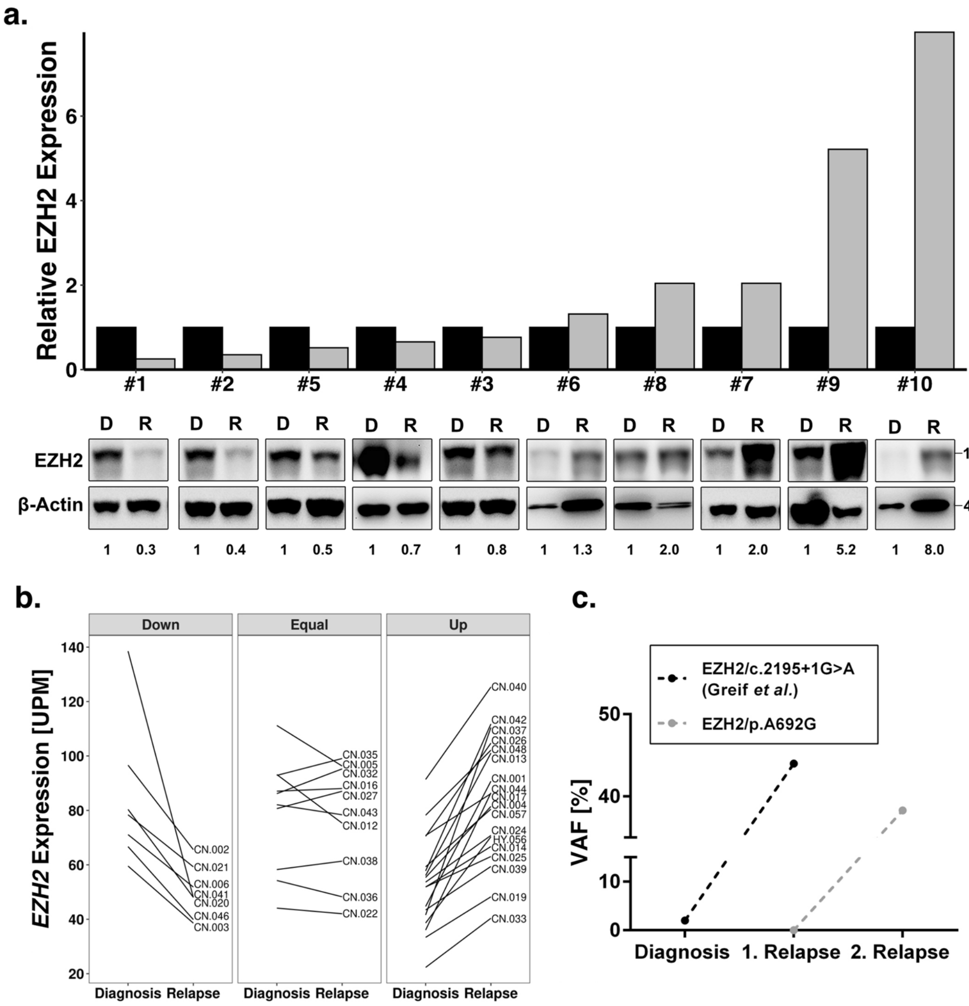


Figure 2. Relevance of EZH2 status in AML relapse. **(a)** Immunoblot for EZH2 protein expression in 10 AML patients at diagnosis and relapse. MW, molecular weight; β-actin, loading control. The ratio of EZH2 to β-actin expression is indicated below and presented in the histogram above. Each relapse value was normalized to the corresponding diagnosis sample. None of the patients carried an *EZH2* mutation. **(b)** *EZH2* mRNA expression between diagnosis and relapse of 32 CN-AML patients from Greif et al. cohort[35]. Up and down are defined as a change in mRNA expression of at least 20%. Three patients carried an *EZH2* mutation. **(c)** Variant allele frequency of the two relapse-associated *EZH2* mutations with outgrowth in first and second relapse.

EZH2 re-expression sensitises to AraC treatment in K562 cells. To investigate if re-expression of EZH2 can reconstitute baseline H3K27me3 levels and therefore sensitise cells to AraC treatment, we established a stable, doxycycline-inducible EZH2 expression system via the PiggyBac transposon system (Supplementary Fig. 5a–d). For this reason, DNA coding for $EZH2^{wt}$ (AA1-751) was introduced into K562/$EZH2^{-/-}$ cells (clone #7). Re-expression of wildtype *EZH2* was able to restore global H3K27me3 levels after 48 h (Fig. 5a and Supplementary Fig. 5e). Furthermore, sensitivity against AraC could be restored in a long-term low dose AraC treatment experiment (Fig. 5b,c). Additionally, we introduced DNA coding for the relapse-associated mutation EZH2/Y733LfsX6 (AA1-737, Supplementary Fig. 5c,d). Re-expression of this mutant after doxycycline induction did not result in restoration of H3K27me3 levels (Fig. 5d and Supplementary Fig. 5f). Likewise, the mutation was not able to restore sensitivity against AraC treatment, indicating an involvement of H3K27 trimethylation in the phenotype of chemoresistance (Fig. 5e,f). Doxycycline alone had no effect on the sensitivity of either $EZH2^{wt}$ or $EZH2^{-/-}$ cells towards AraC treatment (Supplementary Fig. 5g,h).

Upregulation of EZH2 target genes desensitises cells to AraC treatment. RNA sequencing and Proteome analysis was performed to uncover the molecular mechanism involved in EZH2-mediated chemoresistance. *EZH2* knockout in K562 cells resulted in aberrant gene and protein expression (Supplementary Fig. 7), visible in transcriptional upregulation of 216 genes and downregulation of 42 genes as well as translational upregulation of 375 genes and downregulation of 205 genes (Supplementary Table 3 and Supplementary Table 4). The change in protein and RNA expression was found to be correlated (R = 0.5, p = 2.2e-16, Pearson's correla-

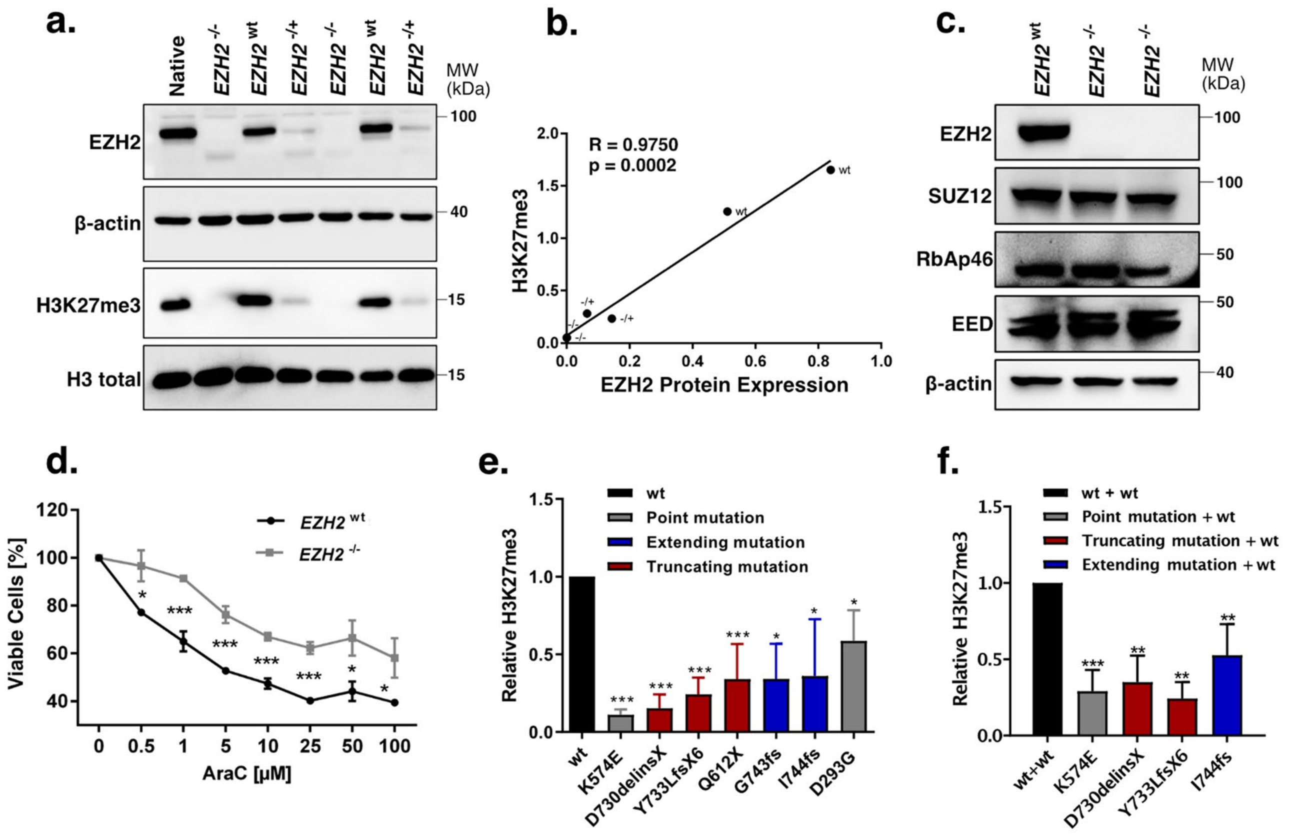


Figure 3. Evaluation of *EZH2* mutations found in AML patients at diagnosis. **(a)** Comparison of EZH2 expression and global H3K27me3 between *EZH2*$^{-/-}$, *EZH2*$^{+/-}$ and *EZH2*^{wt} sc clones in 293T cells. MW, molecular weight. β-actin and H3 total, loading controls. **(b)** Correlation between EZH2 protein expression and global H3K27me3 in 293T sc clones. Pearson's correlation. **(c)** Immunoblot for EZH2, SUZ12, RbAP46 and EED expression in 293T/*EZH2*$^{-/-}$ sc clones. MW, molecular weight; β-actin, loading control. **(d)** AraC resistance in one 293T/*EZH2*$^{-/-}$ and one 293T/*EZH2*^{wt} sc clone. Cells were treated for 72 h with different concentrations of AraC. Viable cells relative to untreated control. **(e)** H3K27me3 levels after re-expression of seven *EZH2* mutations, detected in patient diagnosis samples. Colours referring to protein structural changes caused by the mutation. 293T/*EZH2*$^{-/-}$ cells were transfected transiently with *EZH2* constructs 72 h before protein isolation, and global H3K27me3 was evaluated by immunoblot. Values relative to the wild type. **(f)** H3K27me3 levels after re-expression of four *EZH2* mutants in combination with wild type *EZH2*. 293T/*EZH2*$^{-/-}$ cells were transfected transiently with *EZH2* wildtype and *EZH2* mutant constructs 72 h before protein isolation. Values relative to the wild type. Unpaired, two-tailed Student's t-test; $*p < 0.05$; $**p < 0.01$; $***p < 0.001$. Error bars indicate mean ± s.d of at least three independent experiments.

tion, Fig. 6a) and 41 genes showed differential expression of both mRNA and protein. Most of these genes (37) were upregulated, and only two downregulated in both measures (Fig. 6b). Amongst the upregulated genes, we identified *FHL1* as well as *UBE2E1*, both involved in chemotherapy resistance and relapse in AML[43,44]. Additionally, upregulation of *CA2*, *CNN3* and *AKAP13* was found, which are suggested to be involved in chemotherapy resistance in glioblastoma, colon cancer and breast cancer, respectively. Furthermore, *PDK3*, *TPD52*, *MYO5A*, *AKT3* and *SPECC1* were upregulated, genes associated with poor prognosis in AML or involved in apoptosis. EZH2 ChIP-seq in K562 cells of a publicly available dataset (ENCSR000AQE, ENCSR000AKY) revealed peaks in the promoter region of *FHL1*, suggesting *FHL1* to be a direct target of EZH2. In the promoter region of *UBE2E1* no EZH2 peaks were found, but EZH2 binding was detected in a distal enhancer region (GeneHancer Accession: GH03J023748). Other potential direct targets of EZH2 are *CNN3*, *AKAP13*, *TPD52*, *MYO5A*, *AKT3* and *SPECC1* as EZH2 peaks were found in the respective promoter regions. Additionally, enhancers regulating *FHL1* and *TPD52* could be identified (GeneHancer Accession: GHOXJ136155 and GH08J080078). No peaks were assigned to the genes *CA2* and *PDK3*.

Resistance of *EZH2* mutated patient-derived xenografts (PDX).

To extend our findings of *EZH2* associated AraC resistance *in vitro*, we screened relapsed AML samples for clonal outgrowth of *EZH2* mutated cells. We identified a 54-year-old patient who gained an *EZH2* mutation (p.A692G) at second relapse. A summary of the patients' course of disease including bone marrow blast counts from 9 time points is shown in Figure 7a. We established a sensitive, custom designed digital droplet PCR (ddPCR) assay to monitor the abundance of p.A692G during disease progression from first to second relapse. We detected the mutation only in the

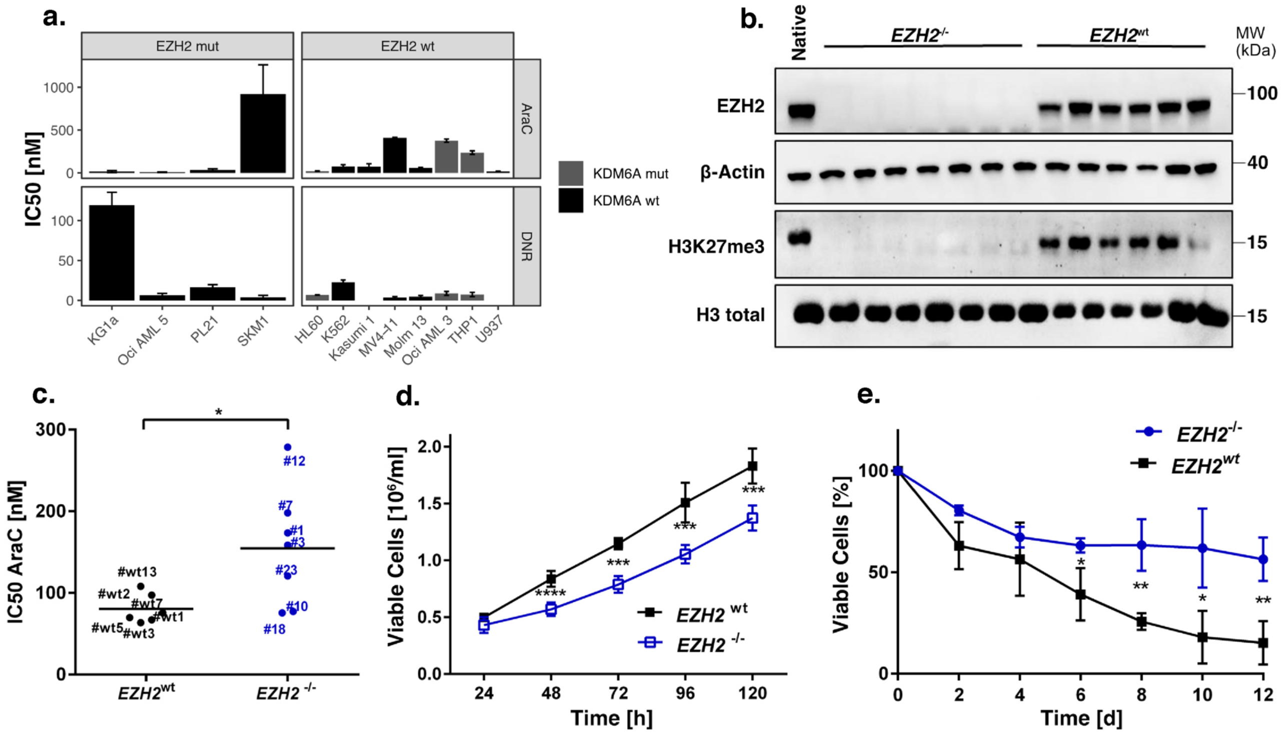


Figure 4. EZH2 depletion promotes resistance in the myeloid cell line K562. **(a)** Comparison of IC_{50} values for DNR and AraC in twelve haematopoietic cell lines. Cells were treated with AraC/DNR for 72 h. **(b)** Immunoblot for EZH2 expression and global H3K27me3 of seven *EZH2*$^{-/-}$ and six *EZH2*^{wt} sc clones in K562 cells. MW, molecular weight; β-actin and H3 total, loading controls. **(c)** Comparison of AraC IC_{50} values in *EZH2*^{wt} (n = 6) and *EZH2*$^{-/-}$ (n = 7) clones. Cells were treated with AraC/DMSO for 72 h. Each value represents the mean of three independent experiments. **(d)** Proliferation of *EZH2*^{wt} (n = 4) and *EZH2*$^{-/-}$ (n = 7) clones for 5 d. Medium was changed every 48 h. **(e),** Long-term low dose AraC treatment in *EZH2*^{wt} (n = 3) and *EZH2*$^{-/-}$ (n = 3) clones. Cells were treated with 30 nM AraC/DMSO for 12 d. Viable cells relative to untreated control. Unpaired, two-tailed Student's t-test; *$p < 0.05$; **$p < 0.01$; ***$p < 0.001$. Error bars indicate mean ± s.d of three independent experiments.

second relapse with a more than 20% increase within three months (Fig. 7a). Furthermore, the patient gained a heterozygous 7q deletion at first relapse, as analysed by MLPA (Fig. 7a).

Patient cells of the first and second relapse were serially transplanted into immune-deficient mice, establishing patient-derived xenografts PDX-AML-491 and PDX-AML-661 respectively (Fig. 7a)[45]. Leukaemic cells of the initial diagnosis did not engraft in the model. PDX cells were lentivirally transduced for transgenic expression of luciferase, enabling disease monitoring *in vivo*[36]. We treated the xenograft mice with a combination of cytarabine and daunorubicin and monitored the leukaemic burden over 80 days. We observed a drastic drop in leukaemic cells in the PDX-491 mice, with a complete cure of three out of four animals (Fig. 7b). The treatment of PDX-661 mice only had minimal effect (Fig. 7c).

Targeted sequencing of a panel of 68 recurrently mutated genes[36] of patient and PDX samples revealed a strong increase of the clone harbouring the *EZH2*/p.A692G mutation in the PDX-661 samples (VAF: 98.8%) in comparison to the second relapse of the patient (VAF: 39.2%, Fig. 7d). The only other mutation illustrating an increase in variant allele frequency was a subclonal *JAK1* mutation, detectable only in the PDX-661 cells. The majority of mutations (*BCOR*, *DNMT3A*, *ETV6*, *PTPN11* and *RUNX1*) remained stable at all time points. Furthermore, two subclonal mutations in *NRAS* and *KRAS* were detected. Both were absent in the patient's first relapse. KRAS was only detectable in the PDX-491 samples, while NRAS decreased during PDX-491 passaging but was detectable again in the PDX-661 as well as in the patient's second relapse.

Dose-response analysis of PDX-491 and PDX-661 cells in vitro confirmed an increased resistance of PDX-661 towards AraC (Fig. 8a). Moreover, global H3K37me3 levels were completely depleted in PDX-661 cells, while EZH2 protein expression was stable (Fig. 8b). Transient transfection of the p.A692G mutation into our 293T/*EZH2*$^{-/-}$ model revealed decreased global H3K27me3 compared to the wild type, further confirming a LOF phenotype (Fig. 8c). To examine if the observed chemoresistance can be caused by EZH2 depletion, we established an siRNA knockdown (kd) targeting wild type *EZH2* in the PDX-491 cells. EZH2 levels could thereby be reduced by approximately 40% (Fig. 8d). Treatment of these cells for 72 h with AraC resulted in lower proliferation (Fig. 8e) and an increased resistance (Fig. 8f).

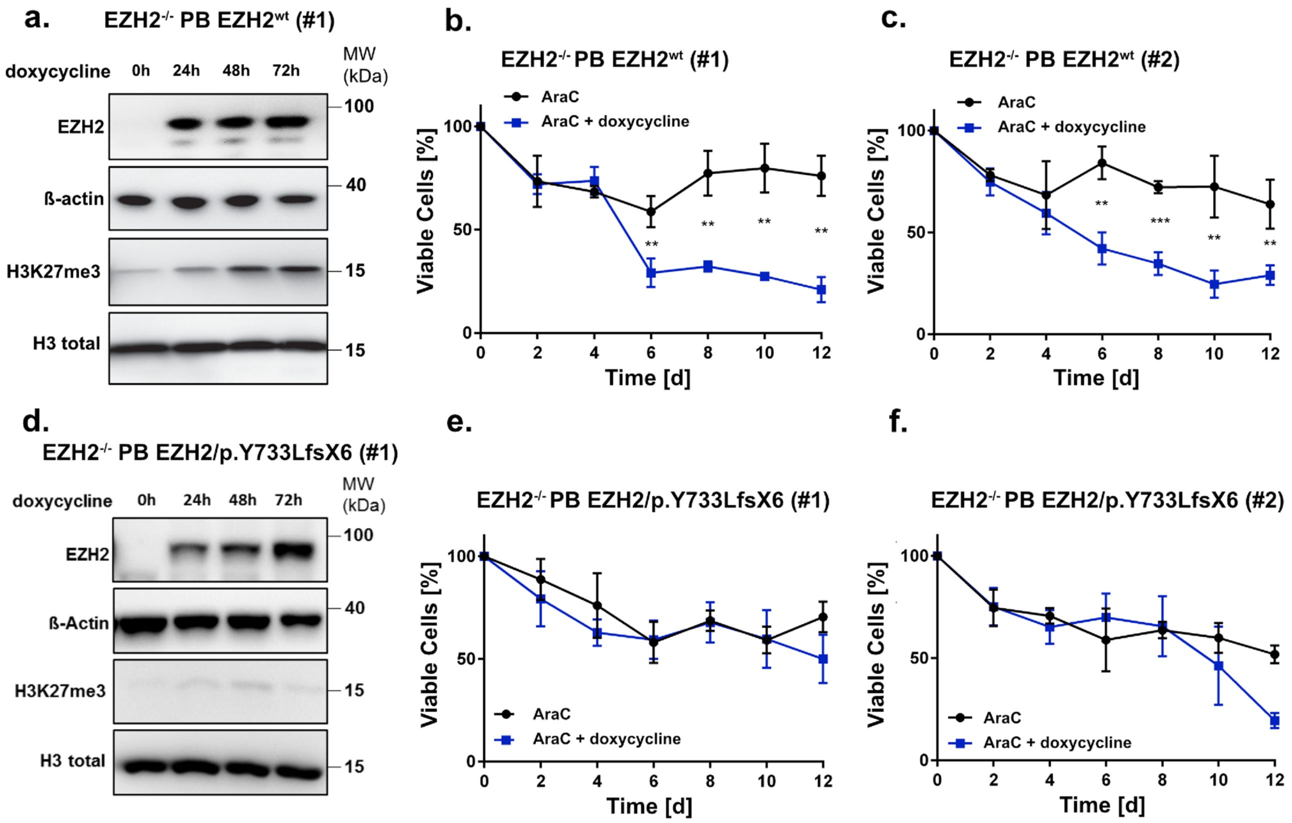


Figure 5. EZH2 re-expression sensitizes K562 cells to AraC treatment. **(a,d)** Immunoblot for EZH2 expression and global H3K27me3 in **(a)** *EZH2*$^{-/-}$ PB *EZH2*^{wt} cells (clone #1) and **(d)** *EZH2*$^{-/-}$ PB EZH2/p.Y733LfsX6 (clone #1) after 0 h, 24 h, 48 h and 72 h of doxycycline induction. Cells were treated with 1 µg/ml doxycycline every 24 h. MW, molecular weight; β-actin and H3 total, loading controls. **(b–c, e–f),** AraC low dose long-term treatment in **(b–c)** *EZH2*$^{-/-}$ PB *EZH2*^{wt} and **(e–f)** *EZH2*$^{-/-}$ PB EZH2/p.Y733LfsX6 cells. Cells were pre-treated for 3 d with doxycycline and then treated with 30 nM AraC/DMSO for 12 d. Cells were split and treated every 4 d and doxycycline was added every 48 h to ensure stable expression of EZH2. Error bars indicate mean ± s.d of three independent experiments. Unpaired, two-tailed Student's t-test; **p* < 0.05; ***p* < 0.01; ****p* < 0.001.

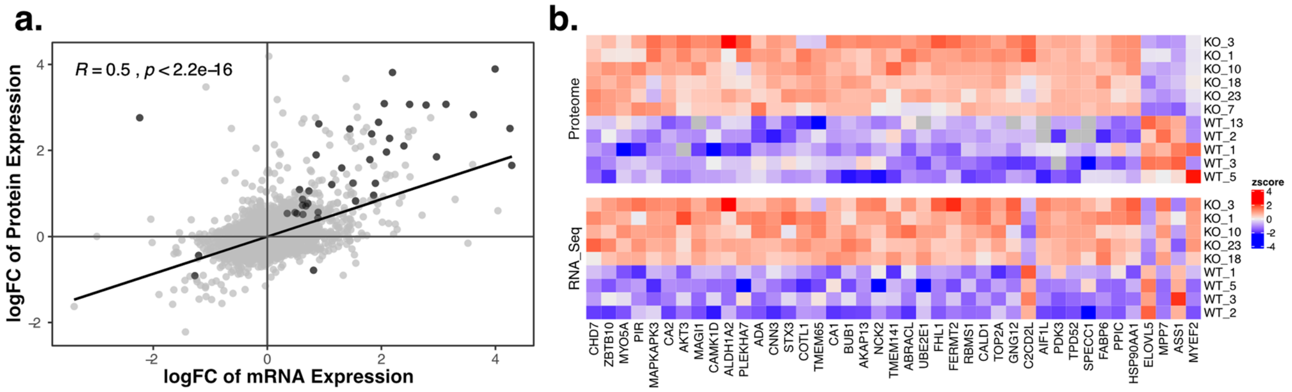


Figure 6. Upregulation of EZH2 target genes. **(a)** Correlation of protein and mRNA expression. Dark grey points representing genes differentially expressed in both measures (adj.P < 0.05). Pearson's correlation. **(b)** Heatmap of the 41 genes differentially expressed (adj.P < 0.05) between *EZH2*$^{-/-}$ and wild type clones, in both protein and mRNA. The colour gradient from red to blue represents high to low expression of genes. White indicates no change.

Figure 7. Resistance in an *EZH2* mutated PDX model. **(a)** Course of disease of an AML patient suffering from two relapses (indicated by dashed vertical lines). 7q deletion was confirmed with MLPA. Variant allele frequency (VAF) of the p.A692G mutation was monitored by digital droplet PCR (time points of samples indicated by red stars). Blast count was measured from bone marrow at the indicated time points (blue bars). Samples used for PDX engraftment are indicated with black triangles. In July 2015 two samples were taken in the same month. **(b–c)** In vivo treatment of PDX mice. NSG mice were injected with patient material of relapse 1 and 2, establishing **(b)** PDX-491 and **(c)** PDX-661. 21 d after injection, mice were treated with AraC (100 mg/kg) and DaunoXome (1 mg/kg) (treatment days indicated with red x). Leukaemic burden was monitored in vivo by bioluminescence imaging. Control mice treated only with PBS are shown in blue. **(d–e),** Variant allele frequencies in the course of first to second relapse of **(d)** EZH2/p.A692G and **(e)** other mutations identified by targeted sequencing in Patient and PDX samples. PDX samples of first engraftment (primograft) as well as first (1st re-Tx) second (2nd re-TX) and third (3rd re-TX) re-transplantation.

Discussion

Development of resistance against standard chemotherapeutics is common in AML and can be induced through various mechanisms[46]. In this study we report that loss-of-function mutations in the histone methyltransferase EZH2 are associated with increased resistance against the antimetabolite cytarabine (AraC).

EZH2 mutations in AML are a very rare event. In fact, only 4% of our patients harboured these mutations at diagnosis, with the majority located in the catalytic SET domain, a known hotspot for *EZH2* mutations[28,47]. Seemingly, the mutations induce loss of EZH2 function, independent of the type of mutation. (Fig. 3, Supplementary Fig. 3a). Apart from the SET domain (aa 605-725), also the post-SET domain (aa 725-746), which is essential for the formation of the cofactor S-adenosyl-L-methionine (SAM) binding pocket, was found crucial in maintaining enzymatic function[48]. Two of the mutations identified in our patients, D730_delinsX and Y733LfsX6, previously described in myelodysplastic syndromes (MDS)[49], caused almost complete elimination of the post-SET domain, while two others, I744fs[50,51] and G743fs, caused frame shifts that resulted in elongated protein variants, highlighting the importance of this domain. In contrast, the missense mutation K574E, located in the CXC domain, is likely to impair the domain's binding ability to the substrate nucleosome and thereby bringing the H3 tail out of reach[50].

We found that complete loss of EZH2 promotes AraC resistance in HEK293T cells as well as the myeloid cell line K562 (Fig. 3, Fig. 4). Furthermore, increased resistance was observed in K562 cells expressing the LOF mutation Y733LfsX6 (Fig. 5d–f) and in a PDX model of a patient who gained the LOF mutation A692G at second relapse (Fig. 7). Additionally, low *EZH2* mRNA expression correlated with poor overall and relapse-free survival (Fig. 1b–c). Our findings are therefore in concordance with the study of Göllner et al.[34], who described AraC resistance in a shRNA knockdown of EZH2 in MV4-11 cells. However, elevated *EZH2* expression has also been reported in AML patients[52,53], and dual inhibition of EZH1/2 was found to eliminate quiescent leukaemic stem cells (LSCs) to prevent relapse[25]. These combined findings suggest a dual role of *EZH2* as either tumour suppressor or oncogene. In our matched diagnosis/relapse pairs, EZH2 protein and mRNA expression levels were found to be highly patient specific, and in most cases, we observed up- or downregulation in relapse (Fig. 2). EZH2 therefore appears to bear an important function in disease progression, and close monitoring of expression and mutation status seems to be crucial in choosing the best treatment approach.

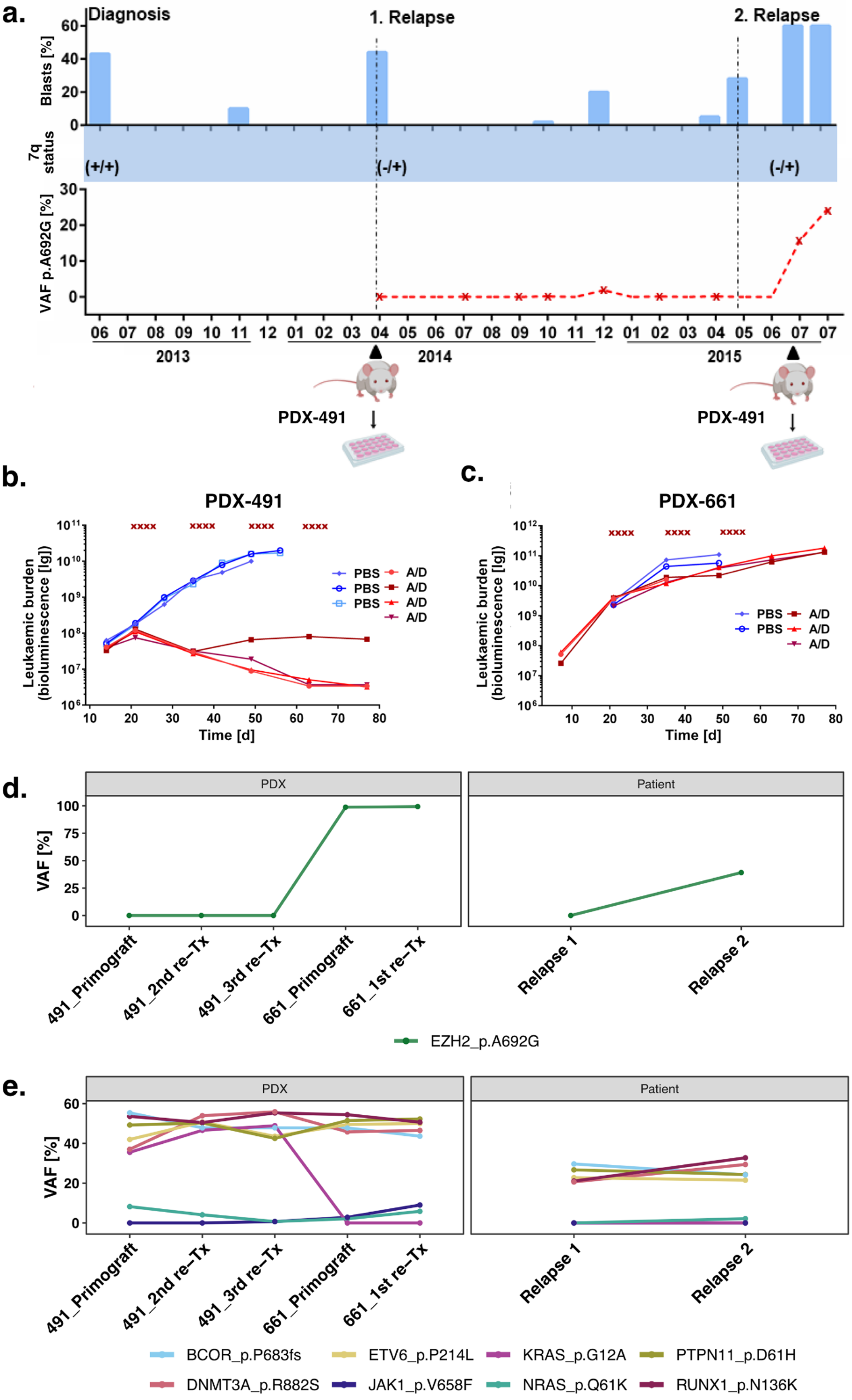
Interestingly, *RUNX1* and *ASXL1* mutations were significantly co-occurring with mutations in *EZH2*. Similar associations have been described before in myeloid malignancies including AML[28,54,55]. Therapy resistance was associated with frequent co-occurrence of *EZH2* and *RUNX1* LOF mutations[56], suggesting a cooperative role of these mutations. *ASXL1* LOF mutations on the other hand can establish an additive effect to EZH2 loss by additional reduction of H3K27 trimethylation through inhibition of PRC2 recruitment[29].

Mutations in other PRC2 subunits (EZH1, EED, SUZ12 or RbAp48) are extremely rare in AML. In the cohort of 50 AML patients of Greif et al. (2018) none could be detected, while in a study of 165 AML patients from Faber et al. (2016), only EED mutations were found with a frequency of 1.8 %. Co-occurrence of *EED* and *EZH2* mutations was found in only one of the patients. EZH2 requires direct interaction with EED to exert its enzymatic function[57]. Thus, also other mutations in the PRC2 complex like EED mutations harbour the potential to confer chemoresistance.

H3K27me3 levels can also be altered by the histone demethylase KDM6A. Loss-of-function mutations in *KDM6A* have been detected in AML and are associated with the development of chemoresistance[42]. Although we and other groups found mutations in both genes to be mutually exclusive[58], expression levels of KDM6A and EZH2 have an antagonistic effect on global H3K27 trimethylation (Supplementary Fig. 4d). Further research is needed to investigate common and specific EZH2 and KDM6A target sites.

EZH2 is responsible for the trimethylation of H3K27 and therefore inactivation of its target genes. Knockout of *EZH2* in K562 cells induced almost complete loss of H3K27me3 levels and resulted in the upregulation of 216 genes and 375 proteins (Fig. 6b). We identified *FHL1* and *UBE2E1* to be direct targets of EZH2. Overexpression of these genes has recently been described to be involved in resistance against cytarabine, and in relapse in AML patients[43,44].

FHL1 might be involved in the transmembrane transport of chemotherapeutic agents. Fu et al.[43] found upregulation of *ABCC1* and *ABCC4*, encoding for the unidirectional efflux transporter proteins MRP1 and MRP4, in AML patients with high FHL1 expression. A slight upregulation of ABCC1 protein could also be detected in our data. Interestingly, Fu et al. also found expression of *FHL1* to be negatively correlated to *SLC29A1* (ENT1) expression. ENT1 is an influx transporter that mediates the uptake of chemotherapeutics and is downregulated upon loss of KDM6A[42]. Since *EZH2* and *KDM6A* mutations were found to be mutually exclusive, those findings suggest an involvement of either EZH2 or KDM6A in the regulation of transmembrane transporter proteins,


a.
Diagnosis
1. Relapse
2. Relapse
Blasts [%]
60
40
20
0
7q status
(+/+)
(-/+)
(-/+)
VAF p.A692G [%]
30
20
10
0
06 07 08 09 10 11 12 01 02 03 04 05 06 07 08 09 10 11 12 01 02 03 04 05 06 07 07
2013
2014
2015
PDX-491
PDX-491
b.
PDX-491
xxxx xxxx xxxx xxxx
Leukaemic burden (bioluminescence [lg])
10^{11}
10^{10}
10^{9}
10^{8}
10^{7}
10^{6}
PBS A/D
PBS A/D
PBS A/D
A/D
10 20 30 40 50 60 70 80
Time [d]
c.
PDX-661
xxxx xxxx xxxx
Leukaemic burden (bioluminescence [lg])
10^{12}
10^{11}
10^{10}
10^{9}
10^{8}
10^{7}
10^{6}
PBS A/D
PBS A/D
A/D
10 20 30 40 50 60 70 80
Time [d]
d.
PDX
Patient
VAF [%]
100
75
50
25
0
491_Primograft
491_2nd re–Tx
491_3rd re–Tx
661_Primograft
661_1st re–Tx
Relapse 1
Relapse 2
EZH2_p.A692G
e.
PDX
Patient
VAF [%]
60
40
20
0
491_Primograft
491_2nd re–Tx
491_3rd re–Tx
661_Primograft
661_1st re–Tx
Relapse 1
Relapse 2
BCOR_p.P683fs
ETV6_p.P214L
KRAS_p.G12A
PTPN11_p.D61H
DNMT3A_p.R882S
JAK1_p.V658F
NRAS_p.Q61K
RUNX1_p.N136K

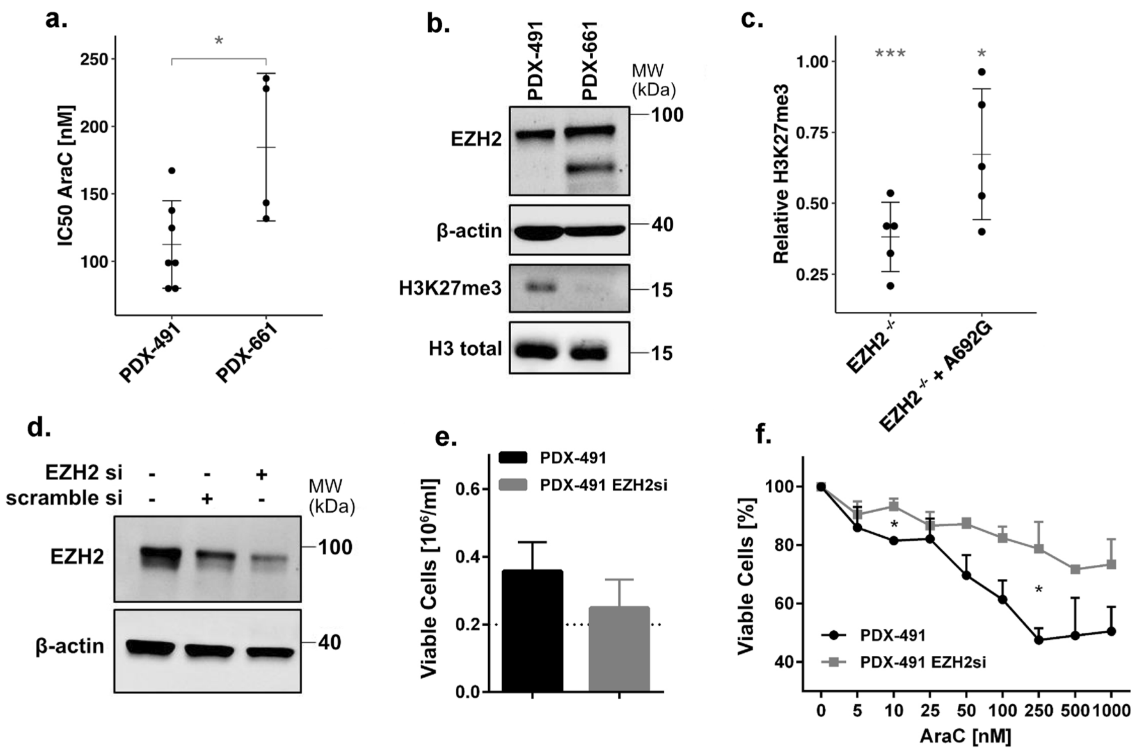


Figure 8. Knockdown of EZH2 in a patient-derived xenograft (PDX) model. **(a)** Comparison of IC_{50} AraC values for PDX-491 and PDX-661 in vitro. **(b)** Immunoblot of EZH2 expression and global H3K27me3 in PDX-491 and PDX-661. **(c)** H3K27me3 levels of EZH2/p.A692G in 293T/*EZH2*$^{-/-}$ cells. Values normalized to H3 loading control and relative to wild type. *indicates significant difference to the wild type. **(d)** Immunoblot of EZH2 expression in PDX-491 cells treated with 10 nM siRNA. Representative blot shown for two independent experiments. **(e)** Histogram showing the proliferation of PDX-491 cells with 10 nM siRNA. Cells were pre-treated for 2 d with siRNA and then incubated for another 3 d for the proliferation assay. **(f)** AraC treatment in PDX-491 cells with 10 nM siRNA. Cells were pre-treated for 2 d with siRNA and then treated for 72 h with AraC. Unpaired, two-tailed Student's t-test; $^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$. MW, molecular weight. β-actin and H3 total, loading controls. Error bars indicate mean ± s.d of three independent experiments.

responsible for the release or uptake of chemotherapeutic agents. The ubiquitin-conjugating enzyme UBE2E1 can regulate the expression of *HOX* genes by its ability to ubiquitinate histones[59]. Although we did not detect any aberrant expression of *HOX* genes, upregulation of HOXA9 as well as HOXB7 was reported by Göllner et al. in a resistant EZH2 negative AML cell line model. We furthermore identified upregulation of the direct EZH2 target genes *CNN3* and *AKAP13*, that are involved in chemotherapy resistance in colon cancer and breast cancer, respectively[60,61], and the genes *MYO5A*, *AKT3* and *SPECC1*, which are implicated in the evasion of apoptosis[62–64]. Additionally, upregulation of TPD52, involved in proliferation, migration, invasion and apoptosis, was found in many cancer types including AML[65].

We conclude that loss-of-function mutations in the histone methyltransferase EZH2 have the potential to confer resistance against the chemotherapeutic agent cytarabine and suggest an involvement of upregulated EZH2 target genes in apoptosis, proliferation and transmembrane transport.

Materials and methods

Cell culture and patient samples.

All cell lines (Supplementary Table 1) were acquired from DSMZ (Braunschweig, Germany) and cultured according to the supplier's recommendations. Patient-derived xenograft (PDX) AML samples were serially passaged in NSG mice and re-isolated for *in vitro* cultivation as previously described[23,45]. Exclusion of mycoplasma contamination was performed continuously during cell culture using the MycoAlert Mycoplasma detection kit (Lonza, Basel, Switzerland). Analysis of patient samples was based on material of AML patients from the AMLCG-99 trial (NCT00266136), AMLCG-2008 trial (NCT01382147), and the Department of Medicine III, University Hospital, LMU. Mononuclear cells were enriched from bone marrow or peripheral blood by Ficoll density gradient centrifugation. Written informed consent for scientific use of sample material was obtained from all patients. The study was performed in accordance with the ethical standards of the responsible committee on human experimentation (written approval by the Research Ethics Boards of the medical faculty of Ludwig-Maximilians-Universität, Munich, number 068-08 and 222-10) and with the Helsinki Declaration of 1975, as revised in 2000. All animal trials were performed in accordance with the current ethical standards of the official committee on animal experimentation (Regierung von Oberbayern, number 55.2-1-54-2531-95-2010 and ROB-55.2Vet-143 2532.Vet_02-16-7).

Proliferation assay. Suspension cells were treated with cytarabine (AraC, Selleck Chemicals, Houston, TX, USA), and daunorubicin (in-house). For short time assays, viable cells were treated once (d0) and counted after 72 h on Vi-Cell Cell Viability Analyzer (Beckman Coulter, Krefeld, Germany). For long-term proliferation assays, cells were treated three times (d0, d4, d8) and viable cells were counted every second day. Unpaired, two-tailed Student's *t*-test and calculation of IC_{50} values were performed using GraphPad Prism version 6.07 (GraphPad Software, La Jolla, CA, USA). PiggyBac[23,45,66] (PB)/*EZH2* cells were pre-cultured with or without doxycycline (1µg/mL) for 72 h followed by treatment with AraC +/− doxycycline, which was added every 48 h. For knockdown experiments in PDX cells, siRNA targeting EZH2 (#s4918, Thermo Fisher Scientific, Waltham, USA) was transiently transfected (10 nM) via nucleofection (Supplementary Methods). Cells were pre-incubated for 48 h and then treated with AraC for 72 h.

Immunoblotting. Immunoblotting was performed as described before[3]. The following antibodies were used: anti-EZH2 (#5246, Cell Signaling Technology, Danvers, USA), anti-β-actin (A5441, Sigma Aldrich, St. Louis, USA), anti-H3 (ab1791, Abcam, Cambridge, UK), anti-H3K27me3 (#9733, Cell Signaling Technology, Danvers, USA), anti-SUZ12 (#3737, Cell Signaling Technology, Danvers, USA), anti-RbAP46 (#4522, Cell Signaling Technology, Danvers, USA), anit-EED (ab113911, Abcam, Cambridge, UK), anti-EZH1 (#42088, Cell Signaling Technology, Danvers, USA). Western blots were quantified using ImageJ version 1.50d and levels were normalized to the associated loading control (β-actin for EZH2, total H3 for H3K27me3).

***In vivo* therapy trial.** Patient-derived xenograft (PDX) cells expressing enhanced firefly luciferase and mCherry were established as described previously[45]. For *in vivo* therapy trials, $1*10^5$ PDX-AML-491 or $8*10^5$ PDX-AML-661 luciferase-positive cells were injected intravenously into 11 or 16 week old male NSG mice (NOD.Cg-$Prkdc^{scid}$ $Il2rg^{tm1Wjl}$/SzJ, The Jackson Laboratory, Bar Harbour, ME, USA), and tumour growth was regularly monitored by bioluminescence imaging (BLI) as described previously[24]. 21 days after transplantation, mice were treated with a combination of Cytarabine (AraC; 100 mg/kg, i.p., days 1-4 of therapy weeks) and liposomal daunorubicin (DaunoXome; 1mg/kg, i.v., days 1 and 4 of therapy weeks) every second week for three (AML-661, n = 3) or four (AML-491, n = 4) cycles. Tumour burden was regularly monitored by BLI and compared to untreated control mice. In total, 13 mice were included in this study; one AML-661 control mouse was sacrificed 14 days after injection due to leukaemia unrelated illness. End point of the study was end-stage leukaemia. All animal trials were performed in accordance with the current ethical standards of the official committee on animal experimentation (Regierung von Oberbayern, number 55.2-1-54-2531-95-2010 and ROB-55.2Vet-143 2532.Vet_02-16-7) and in compliance with the ARRIVE guidelines.

Ethics approval. We hereby confirm that all experimental protocols were approved by the Department of Medicine III, University Hospital, LMU Munich, the Department of Biology III and Center for Integrated Protein Science Munich (CIPSM); Human Biology and BioImaging, LMU Munich, Planegg Martinsried, Germany and the Helmholtz Zentrum München, Munich.

Data availability

The RNA-seq data generated for this study is available at GEO under the accession number: GSE162623 The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE[67] partner repository with the dataset identifier PXD023139.

Received: 9 October 2019; Accepted: 15 February 2021
Published online: 12 March 2021

References

1. Bradstock, K. F. *et al.* A randomized trial of high-versus conventional-dose cytarabine in consolidation chemotherapy for adult de novo acute myeloid leukemia in first remission after induction therapy containing high-dose cytarabine. *Blood* **105**, 481–488 (2005).
2. Wang, J. *et al.* Meta-analysis of randomised clinical trials comparing idarubicin + cytarabine with daunorubicin + cytarabine as the induction chemotherapy in patients with newly diagnosed acute myeloid leukaemia. *PLoS One* **8**, e60699 (2013).
3. Hann, I. M. *et al.* Randomized Comparison of DAT Versus ADE as Induction Chemotherapy in Children and Younger Adults With Acute Myeloid Leukemia Results of the Medical Research Council's 10th AML Trial (MRC AML10). *Blood* **89**, 2311–2318 (1997).
4. Marcucci, G. *et al.* Age-related prognostic impact of different types of DNMT3A mutations in adults with primary cytogenetically normal acute myeloid leukemia. *J. Clin. Oncol.* **30**, 742–750 (2012).
5. Schenk, T. *et al.* Inhibition of the LSD1 (KDM1A) demethylase reactivates the all-trans-retinoic acid differentiation pathway in acute myeloid leukemia. *Nat. Med.* **18**, 605–611 (2012).
6. Kanwal, R., Gupta, K. & Gupta, S. Cancer Epigenetics: An Introduction. in *Cancer Epigenetics: Risk Assessment, Diagnosis, Treatment, and Prognosis* (ed. Verma, M.) 3–25 (Springer New York, 2015).
7. Hanahan, D. & Weinberg, R. A. Hallmarks of cancer: the next generation. *Cell* **144**, 646–674 (2011).
8. Stahl, M. *et al.* Hypomethylating agents in relapsed and refractory AML: outcomes and their predictors in a large international patient cohort. *Blood Adv* **2**, 923–932 (2018).
9. Cashen, A. F., Schiller, G. J., O'Donnell, M. R. & DiPersio, J. F. Multicenter, phase II study of decitabine for the first-line treatment of older patients with acute myeloid leukemia. *J. Clin. Oncol.* **28**, 556–561 (2010).
10. Margueron, R. & Reinberg, D. The Polycomb complex PRC2 and its mark in life. *Nature* **469**, 343–349 (2011).
11. Di Croce, L. & Helin, K. Transcriptional regulation by Polycomb group proteins. *Nat. Struct. Mol. Biol.* **20**, 1147–1155 (2013).
12. Chang, C.-J. *et al.* EZH2 promotes expansion of breast tumor initiating cells through activation of RAF1-β-catenin signaling. *Cancer Cell* **19**, 86–100 (2011).

13. Behrens, C. *et al.* EZH2 protein expression associates with the early pathogenesis, tumor progression, and prognosis of non-small cell lung carcinoma. *Clin. Cancer Res.* **19**, 6556–6565 (2013).
14. Varambally, S. *et al.* The polycomb group protein EZH2 is involved in progression of prostate cancer. *Nature* **419**, 624–629 (2002).
15. Zingg, D. *et al.* The epigenetic modifier EZH2 controls melanoma growth and metastasis through silencing of distinct tumour suppressors. *Nat. Commun.* **6**, 6051 (2015).
16. Morin, R. D. *et al.* Somatic mutations altering EZH2 (Tyr641) in follicular and diffuse large B-cell lymphomas of germinal-center origin. *Nat. Genet.* **42**, 181–185 (2010).
17. Bödör, C. *et al.* EZH2 mutations are frequent and represent an early event in follicular lymphoma. *Blood* **122**, 3165–3168 (2013).
18. Rabello, D. *et al.* Overexpression of EZH2 associates with a poor prognosis in chronic lymphocytic leukemia. *Blood Cells Mol. Dis.* **54**, 97–102 (2015).
19. D'Angelo, V. *et al.* EZH2 is increased in paediatric T-cell acute lymphoblastic leukemia and is a suitable molecular target in combination treatment approaches. *J. Exp. Clin. Cancer Res.* **34**, 83 (2015).
20. Nishioka, C., Ikezoe, T., Yang, J. & Yokoyama, A. BCR/ABL increases EZH2 levels which regulates XIAP expression via miRNA-219 in chronic myeloid leukemia cells. *Leuk. Res.* **45**, 24–32 (2016).
21. Jankowska, A. M. *et al.* Mutational spectrum analysis of chronic myelomonocytic leukemia includes genes associated with epigenetic regulation: UTX, EZH2, and DNMT3A. *Blood* **118**, 3932–3941 (2011).
22. Simon, C. *et al.* A key role for EZH2 and associated genes in mouse and human adult T-cell acute leukemia. *Genes Dev.* **26**, 651–656 (2012).
23. Basheer, F. *et al.* Contrasting requirements during disease evolution identify EZH2 as a therapeutic target in AML. *J. Exp. Med.* **216**, 966–981 (2019).
24. Mochizuki-Kashio, M. *et al.* Ezh2 loss in hematopoietic stem cells predisposes mice to develop heterogeneous malignancies in an Ezh1-dependent manner. *Blood* **126**, 1172–1183 (2015).
25. Fujita, S. *et al.* Dual inhibition of EZH1/2 breaks the quiescence of leukemia stem cells in acute myeloid leukemia. *Leukemia* **32**, 855–864 (2018).
26. Neff, T. *et al.* Polycomb repressive complex 2 is required for MLL-AF9 leukemia. *Proc. Natl. Acad. Sci. U. S. A.* **109**, 5028–5033 (2012).
27. Gu, Z. *et al.* Loss of EZH2 reprograms BCAA metabolism to drive leukemic transformation. *Cancer Discov.* **9**, 1228–1247 (2019).
28. Ernst, T. *et al.* Inactivating mutations of the histone methyltransferase gene EZH2 in myeloid disorders. *Nat. Genet.* **42**, 722–726 (2010).
29. Abdel-Wahab, O. *et al.* ASXL1 mutations promote myeloid transformation through loss of PRC2-mediated gene repression. *Cancer Cell* **22**, 180–193 (2012).
30. Kim, E. *et al.* SRSF2 mutations contribute to myelodysplasia by mutant-specific effects on exon recognition. *Cancer Cell* **27**, 617–630 (2015).
31. Sashida, G. *et al.* The loss of Ezh2 drives the pathogenesis of myelofibrosis and sensitizes tumor-initiating cells to bromodomain inhibition. *J. Exp. Med.* **213**, 1459–1477 (2016).
32. Wang, J. *et al.* TET2, ASXL1 and EZH2 mutations in Chinese with myelodysplastic syndromes. *Leuk. Res.* **37**, 305–311 (2013).
33. Grossmann, V. *et al.* Molecular profiling of chronic myelomonocytic leukemia reveals diverse mutations in >80% of patients with TET2 and EZH2 being of high prognostic relevance. *Leukemia* **25**, 877–879 (2011).
34. Göllner, S. *et al.* Loss of the histone methyltransferase EZH2 induces resistance to multiple drugs in acute myeloid leukemia. *Nat. Med.* **23**, 69–78 (2017).
35. Greif, P. A. *et al.* Evolution of cytogenetically normal acute myeloid leukemia during therapy and relapse: an exome sequencing study of 50 patients. *Clin. Cancer Res.* **24**, 1716–1726 (2018).
36. Metzeler, K. H. *et al.* Spectrum and prognostic relevance of driver gene mutations in acute myeloid leukemia. *Blood* **128**, 686–698 (2016).
37. Döhner, H. *et al.* Diagnosis and management of AML in adults: 2017 ELN recommendations from an international expert panel. *Blood* **129**, 424–447 (2017).
38. Li, Z. *et al.* Identification of a 24-gene prognostic signature that improves the European LeukemiaNet risk classification of acute myeloid leukemia: an international collaborative study. *J. Clin. Oncol.* **31**, 1172–1181 (2013).
39. Wouters, B. J. *et al.* Double CEBPA mutations, but not single CEBPA mutations, define a subgroup of acute myeloid leukemia with a distinctive gene expression profile that is uniquely associated with a favorable outcome. *Blood* **113**, 3088–3091 (2009).
40. Barrett, T. *et al.* NCBI GEO: archive for functional genomics data sets—update. *Nucleic Acids Res.* **41**, D991–D995 (2013).
41. Papaemmanuil, E. *et al.* Clinical and biological implications of driver mutations in myelodysplastic syndromes. *Blood* **122**, 3616–27 (2013).
42. Stief, S. M. *et al.* Loss of KDM6A confers drug resistance in acute myeloid leukemia. *Leukemia* https://doi.org/10.1038/s41375-019-0497-6 (2019).
43. Fu, Y. *et al.* Genome-wide identification of FHL1 as a powerful prognostic candidate and potential therapeutic target in acute myeloid leukaemia. *EBioMedicine* **52**, 102664 (2020).
44. Luo, H. *et al.* Microarray-based analysis and clinical validation identify ubiquitin-conjugating enzyme E2E1 (UBE2E1) as a prognostic factor in acute myeloid leukemia. *J. Hematol. Oncol.* **9**, 125 (2016).
45. Vick, B. *et al.* An advanced preclinical mouse model for acute myeloid leukemia using patients' cells of various genetic subgroups and in vivo bioluminescence imaging. *PLoS One* **10**, e0120925 (2015).
46. Zhang, J., Gu, Y. & Chen, B. Mechanisms of drug resistance in acute myeloid leukemia. *Onco. Targets. Ther.* **12**, 1937–1945 (2019).
47. Guglielmelli, P. *et al.* EZH2 mutational status predicts poor survival in myelofibrosis. *Blood* **118**, 5227–5234 (2011).
48. Wu, H. *et al.* Structure of the catalytic domain of EZH2 reveals conformational plasticity in cofactor and substrate binding sites and explains oncogenic mutations. *PLoS One* **8**, e83737 (2013).
49. Nikoloski, G. *et al.* Somatic mutations of the histone methyltransferase gene EZH2 in myelodysplastic syndromes. *Nat. Genet.* **42**, 665–667 (2010).
50. Poepsel, S., Kasinath, V. & Nogales, E. Cryo-EM structures of PRC2 simultaneously engaged with two functionally distinct nucleosomes. *Nat. Struct. Mol. Biol.* **25**, 154–162 (2018).
51. Papaemmanuil, E. *et al.* Genomic Classification and Prognosis in Acute Myeloid Leukemia. *N. Engl. J. Med.* **374**, 2209–2221 (2016).
52. Grubach, L. *et al.* Gene expression profiling of Polycomb, Hox and Meis genes in patients with acute myeloid leukaemia. *Eur. J. Haematol.* **81**, 112–122 (2008).
53. Wen, S. *et al.* Novel combination of histone methylation modulators with therapeutic synergy against acute myeloid leukemia in vitro and in vivo. *Cancer Lett.* **413**, 35–45 (2018).
54. Gaidzik, V. I. *et al.* RUNX1 mutations in acute myeloid leukemia are associated with distinct clinico-pathologic and genetic features. *Leukemia* **30**, 2160–2168 (2016).
55. Rinke, J. *et al.* Molecular characterization of EZH2 mutant patients with myelodysplastic/myeloproliferative neoplasms. *Leukemia* **31**, 1936–1943 (2017).
56. Booth, C. A. G. *et al.* Ezh2 and Runx1 mutations collaborate to initiate lympho-myeloid leukemia in early thymic progenitors. *Cancer Cell* **33**, 274-291.e8 (2018).

57. Ueda, T. *et al.* EED mutants impair polycomb repressive complex 2 in myelodysplastic syndrome and related neoplasms. *Leukemia* **26**, 2557–2560 (2012).
58. Khan, S. N. *et al.* Multiple mechanisms deregulate EZH2 and histone H3 lysine 27 epigenetic changes in myeloid malignancies. *Leukemia* **27**, 1301–1309 (2013).
59. Zhu, B. *et al.* Monoubiquitination of human histone H2B: the factors involved and their roles in HOX gene regulation. *Mol. Cell* **20**, 601–611 (2005).
60. Nair, V. A., Al-Khayyal, N. A., Sivaperumal, S. & Abdel-Rahman, W. M. Calponin 3 promotes invasion and drug resistance of colon cancer cells. *World J. Gastrointest. Oncol.* **11**, 971–982 (2019).
61. Bentin Toaldo, C. *et al.* Protein Kinase A-induced tamoxifen resistance is mediated by anchoring protein AKAP13. *BMC Cancer* **15**, 588 (2015).
62. Alves, C. P. *et al.* Myosin-Va contributes to manifestation of malignant-related properties in melanoma cells. *J. Invest. Dermatol.* **133**, 2809–2812 (2013).
63. Hinz, N. & Jücker, M. Distinct functions of AKT isoforms in breast cancer: a comprehensive review. *Cell Commun. Signal.* **17**, 154 (2019).
64. D'Agostino, L. & Giordano, A. A novel dual signaling axis for NSP 5a3a induced apoptosis in head and neck carcinoma. *Oncotarget* **2**, 1055–1074 (2011).
65. Ha, M. *et al.* Prognostic role of TPD52 in acute myeloid leukemia: a retrospective multicohort analysis. *J. Cell. Biochem.* **120**, 3672–3678 (2019).
66. Mulholland, C. B. *et al.* Recent evolution of a TET-controlled and DPPA3/STELLA-driven pathway of passive demethylation in mammals. *bioRxiv* 321604 (2020) https://doi.org/10.1101/321604.
67. Perez-Riverol Y, *et al.* The PRIDE database and related tools and resources in 2019: improving support for quantification data. *Nucleic Acids Res.* **47**(D1), D442–D450 (2019).

Acknowledgements

The authors thank all study participants. We thank M. Fritschle for animal handling. We thank Bianka Ksienzyk for cell sorting. We thank Karl-Peter Hopfner for the nice cooperation within mutation characterisation. We thank Dr. Igor Paron and Dr. Christian Deiml from the Proteomics and Signal Transduction Group and the Clinical Proteomics Group at the Max Planck Institute of Biochemistry, Martinsried, for their technical assistance and Michael Wierer und Matthias Mann for their support. K.H.M., P.A.G., G.S., H.L., I.J., and K.S. were supported by the German Research Council (DFG) within the SFB 1243 "Cancer Evolution". J.M.K., R.M., K.V., L.E.W., M.R. and S.W. are members of the IRTG-1243 within the SFB 1243 of the DFG. TH was supported by the Wilhelm-Sander-Stiftung (Grant 2013.086.2) and the Physician Scientists Grant (G-509200-004) from the Helmholtz Zentrum München. E.U. is a fellow of the International Max Planck Research School for Molecular Life Sciences (IMPRS-LS) and is supported by the research training group 1721 (RTG 1721, DFG).

Author contributions

K.S., J.M.K. and S.W. conceived the study. J.M.K. and S.W. wrote the manuscript and analysed the data. J.M.K. performed all experiments which are not mentioned below. M.D.B. and S.B. supported and supervised CRISPR/Cas9 mediated genome editing experiments. R.M., K.V., M.R., M.F. and H.L. supported the experiments. L.E.W. performed RNA-seq experiments. Mass spectrometry-based proteomics experiments were designed by S.W., M.R. and E.U., performed by E.U. and analysed by S.W. and E.U. B.V. supported and supervised all mice and PDX experiments. K.H.M. performed mutation analysis in AML diagnosis. T.H. performed survival analysis and provided supplementary Fig. 1d. M.S. performed and evaluated MLPA analysis. K.S., I.J., G.S., W.H. and O.W. interpreted the data and supported the project by coordinating the teams and experiments. All authors reviewed the manuscript.

Funding

Open Access funding enabled and organized by Projekt DEAL.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-021-84708-6.

Correspondence and requests for materials should be addressed to K.S.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2021

Bartoschek, M. D., **Ugur, E.**, Nguyen, T.-A., Rodschinka, G., Wierer, M., Lang, K., & Bultmann, S. (**2021**). **Identification of permissive amber suppression sites for efficient non-canonical amino acid incorporation in mammalian cells**. Nucleic Acids Research, 49(11), e62-e62.

License: CC-BY-4.0

https://doi.org/10.1093/nar/gkab132

Published online 3 March 2021

Nucleic Acids Research, 2021, Vol. 49, No. 11 **e62**
https://doi.org/10.1093/nar/gkab132

Identification of permissive amber suppression sites for efficient non-canonical amino acid incorporation in mammalian cells

Michael D. Bartoschek[1], Enes Ugur[1,2], Tuan-Anh Nguyen[3], Geraldine Rodschinka[1], Michael Wierer[2], Kathrin Lang[3,*] and Sebastian Bultmann[1,*]

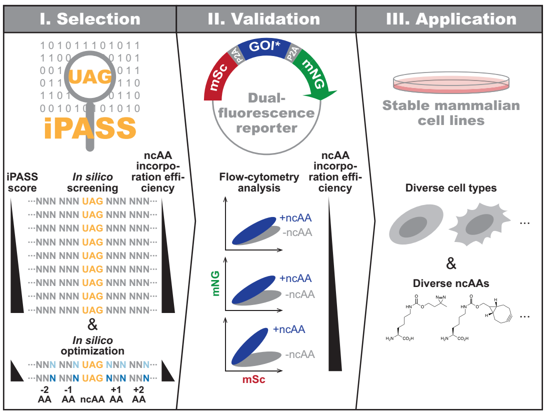
[1]Department of Biology II and Center for Molecular Biosystems (BioSysM), Human Biology and BioImaging, Ludwig-Maximilians-Universität München, Munich 81377, Germany, [2]Department of Proteomics and Signal Transduction, Max Planck Institute of Biochemistry, Martinsried 82152, Germany and [3]Department of Chemistry, Synthetic Biochemistry, Technical University of Munich, Garching 85748, Germany

Received June 22, 2020; Revised February 16, 2021; Editorial Decision February 17, 2021; Accepted February 17, 2021

ABSTRACT

The genetic code of mammalian cells can be expanded to allow the incorporation of non-canonical amino acids (ncAAs) by suppressing in-frame amber stop codons (UAG) with an orthogonal pyrrolysyl-tRNA synthetase (PylRS)/tRNA$^{Pyl}_{CUA}$ (PylT) pair. However, the feasibility of this approach is substantially hampered by unpredictable variations in incorporation efficiencies at different stop codon positions within target proteins. Here, we apply a proteomics-based approach to quantify ncAA incorporation rates at hundreds of endogenous amber stop codons in mammalian cells. With these data, we compute iPASS (Identification of Permissive Amber Sites for Suppression; available at www.bultmannlab.eu/tools/iPASS), a linear regression model to predict relative ncAA incorporation efficiencies depending on the surrounding sequence context. To verify iPASS, we develop a dual-fluorescence reporter for high-throughput flow-cytometry analysis that reproducibly yields context-specific ncAA incorporation efficiencies. We show that nucleotides up- and downstream of UAG synergistically influence ncAA incorporation efficiency independent of cell line and ncAA identity. Additionally, we demonstrate iPASS-guided optimization of ncAA incorporation rates by synonymous exchange of codons flanking the amber stop codon. This combination of *in silico* analysis followed by validation in living mammalian cells substantially simplifies identification as well as adaptation of sites within a target protein to confer high ncAA incorporation rates.

GRAPHICAL ABSTRACT



INTRODUCTION

Decoding of in-frame amber stop codons (UAG), generally referred to as amber suppression, enables the translational incorporation of non-canonical amino acids (ncAAs) into target proteins *in vitro* and *in vivo* (1,2). The pyrrolysyl-tRNA synthetase (PylRS, encoded by *PylS*)/tRNA$^{Pyl}_{CUA}$ (PylT, encoded by *PylT*) pair from *Methanosarcina* species is one of the most commonly used orthogonal translation systems (OTSs) to incorporate ncAAs at amber stop codons in bacteria (3–5), yeast (6), mammalian cells (7–9) and animals (10–13). This expansion of the genetic code allows site-specific introduction of unique moieties into proteins

---

[*]To whom correspondence should be addressed. Tel: +49 89 2180 74230; Fax: +49 89 2180 74236; Email: bultmann@bio.lmu.de
Correspondence may also be addressed to Kathrin Lang. Tel: +49 89 289 13836; Email: kathrin.lang@tum.de
Present addresses:
Geraldine Rodschinka, Max Planck Institute of Biochemistry, Martinsried 82152, Germany.
Michael Wierer, Proteomics Research Infrastructure, University of Copenhagen, Copenhagen 2200, Denmark.

© The Author(s) 2021. Published by Oxford University Press on behalf of Nucleic Acids Research.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

including bioorthogonal handles for chemical conjugation (14) and photocrosslinkers (15,16) to rationally probe and control protein structure, dynamics, and function in living cells. To date, >100 structurally and functionally diverse ncAAs have been added to the mammalian genetic code (17). The initially low efficiency of ncAA incorporation via amber suppression in mammalian cells has been progressively enhanced by engineering OTS components (18–21) and the eukaryotic release factor 1 (22) as well as by tuning OTS expression levels (22–27) and the generation of stable cell lines (28–30). However, depending on the UAG context, high variations in ncAA incorporation rates are frequently observed in bacteria and mammalian cells (1,2,31–37).

In eukaryotes, the nucleotide frequency around stop codons is non-random. Rather, the nucleotide following the stop codon (+4; stop codon corresponds to nucleotides +1, +2, +3), which together with the stop codon forms a tetranucleotide termination signal, is biased for purines (38–42). This purine bias is especially evident at highly expressed genes and hence has been proposed to promote efficient translational termination (38,42–46). Additionally, stop codon readthrough has been found in prokaryotes, eukaryotes, as well as plant and animal viral RNAs (47). Numerous studies have documented that basal stop codon readthrough by near-cognate tRNAs in eukaryotes is modulated by the flanking sequence context, with a clear influence of the nucleotides downstream of the stop codon (43,48–66). For example, cytosine at +4 (+4 C) is a hallmark of motifs that trigger translational readthrough in higher eukaryotes (43,48,55,58–61,63,64,66).

Previous studies have indicated that basal translational readthrough and the suppression of amber stop codons in mammalian cells are governed by similar (54,67,68) but not identical (63) context-specific effects. In prokaryotes, purines, especially at +4, have been found to boost ncAA incorporation at in-frame amber stop codons (33,34,69). Contradicting these reports, a recent study failed to identify features that reliably predict ncAA incorporation efficiency at amber stop codons in prokaryotes (70). Moreover, the influence of sequence context on the efficiency of translational termination and amber suppression differs between bacteria and mammalian cells (67,68,71). As a result, UAG contexts found to be favorable in prokaryotes cannot reliably be used for the selection of optimal mammal-specific UAG contexts. To date, literature on context effects in mammalian amber suppression is scarce. The few existing studies have been restricted to analysis of just the downstream nucleotide (67,68) or codon (63), at which +4 C was suggested to have a stimulatory effect. In fact, it remains largely unclear to what extent *cis*-acting sequence elements determine amber suppression and ncAA incorporation rates in mammalian cells with an expanded genetic code.

Due to their stringency in quantifying ncAA incorporation rates, dual-fluorescence reporters in combination with flow-cytometry analysis constitute an attractive high-throughput screening platform to analyze amber suppression efficiencies (72). In these systems, expression of two spectrally distinct fluorescent proteins or fluorophore epitopes is coupled via a linker region harboring the amber stop codon. By calculating the ratio between these two fluorescence intensities in the presence and absence of an ncAA, relative amber suppression efficiencies can be robustly quantified (73,74). Currently, dual-fluorescence reporters like *mCherry-TAG-EGFP* in mammalian cells (8) are routinely used to analyze and compare amber suppression efficiencies across different OTSs, ncAAs, and cell lines. However, their applicability for rapid screening of permissive ncAA incorporation sites in target proteins has not yet been explored.

In this study, we sought to streamline the identification of positions permitting high ncAA incorporation efficiencies in mammalian cells with an expanded genetic code. Applying CRISPR/Cas9 or PiggyBac (PB) transposase-mediated genome engineering, we first established mouse embryonic stem cell (mESC) and human embryonic kidney 293T (HEK293T) cell lines stably expressing the orthogonal *Methanosarcina mazei PylS/PylT* pair. Using these cell lines we then performed a novel variation of stochastic orthogonal recoding of translation with enrichment (SORT-E) (75) to characterize the entire amber suppressed proteome (amberome) of mammalian cells for the first time. After labeling amber suppressed proteins with a biotin probe and following enrichment by streptavidin pulldown, we used mass spectrometry-based proteomics to systematically assess the efficiency of ncAA incorporation at hundreds of endogenous amber stop codons. With this data, we built a linear regression model of UAG contexts to predict and adjust ncAA incorporation efficiencies *in silico*, which we call iPASS (Identification of Permissive Amber Sites for Suppression; available at www.bultmannlab.eu/tools/iPASS). The resulting iPASS consensus motif suggests that amber suppression efficiency is subject to synergistic context effects mediated by the nucleotides up- and downstream of UAG. To experimentally validate the robustness of iPASS predictions, we developed a dual-fluorescence reporter for the rapid and reproducible quantification of amber suppression efficiencies at individual sequence contexts within a chosen target protein. Using this reporter in flow-cytometry, we analyzed amber suppression at multiple positions within histones H2A and H3, the *de novo* DNA methyltransferase 3B (DNMT3B), as well as at selected synthetic sequence contexts. Our results demonstrate that overall iPASS reliably predicts relative ncAA incorporation efficiencies, which we show to be independent of ncAA as well as cell line identity. Furthermore, we validate iPASS to optimize amber suppression efficiencies at fixed ncAA incorporation sites by silently mutating the two codons following and preceding the amber stop codon. Collectively, iPASS in combination with our dual-fluorescence reporter provides a methodological framework for advancing the applicability of genetic code expansion technologies in mammalian cells.

MATERIALS AND METHODS

Cell culture

Cell lines. Human embryonic kidney 293T (HEK293T) cells were acquired from the Leibniz Institute – German Collection of Microorganisms and Cell Cultures (DSMZ #ACC635; Braunschweig, GER) and were not further authenticated. J1 mouse embryonic stem cells (mESCs) were a kind gift of En Li and Taiping Chen and were not further authenticated. Cells were cultured under standard conditions (5% CO_2, 90% humidity, 37°C). Cells were counted after Trypan Blue staining using a Countstar® BioTech Automated Cell Counter system (Alit Life Science). All cell lines regularly tested negative by PCR for Mycoplasma contamination.

HEK293T cells were maintained in Dulbecco's modified Eagle's medium (DMEM; D6429, Sigma-Aldrich) supplemented with 10% fetal bovine serum (FBS; Sigma-Aldrich) and 50 μg/ml gentamycin (47991.01, SERVA Electrophoresis).

J1 mESCs were maintained on 0.2% (w/v) gelatin-coated (G2500, Sigma-Aldrich) dishes in Dulbecco's modified Eagle's medium (DMEM; D6429, Sigma-Aldrich) supplemented with 16% fetal bovine serum (FBS; Sigma-Aldrich), 0.1 mM 2-mercaptoethanol (M3148, Sigma-Aldrich), 2 mM L-glutamine (G7513, Sigma-Aldrich), 1× MEM non-essential amino acids (M7145, Sigma-Aldrich), 100 U/ml penicillin, 100 μg/ml streptomycin (Pen/Strep; P4333, Sigma-Aldrich), homemade recombinant LIF tested for efficient self-renewal maintenance, and 2i (1 μM PD032591 and 3 μM CHIR99021; Axon Medchem).

To maintain expression of the respective transgenes, stable cell lines were continuously cultured under selection pressure using 1 μg/ml puromycin (A1113803, Thermo Fisher Scientific) and/or 1 mg/ml G418 (A2167, AppliChem).

Non-canonical amino acids. Three ncAA stock solutions were prepared for use in mammalian cell culture: (i) 100 mM BocK in 100 mM NaOH; (ii) 50 mM DiazK in 100 mM TFA (Trifluoroacetic acid); (iii) 100 mM BcnK in 200 mM NaOH, 15% (v/v) DMSO. All solutions were 0.2 μm sterile filtered and stored at −20°C.

Immediately before adding to cell culture medium, ncAA stock solutions were freshly diluted in 3 volumes of 1 M HEPES (15630056, Thermo Fisher Scientific) to neutralize pH. Within all cell culture experiments, a final concentration of 0.5 mM ncAA was used. For –ncAA control samples, cell culture medium was supplemented with the respective solvent only.

CRISPR/Cas9 genome engineering

To MIN-tag (*attP* site for Bxb1-mediated recombination; see (76)) the *Gt(ROSA)26Sor* (R26) locus (77,78) in mESCs, sgRNA (R26_sgRNA) targeting R26 exon 1 (NCBI ref. seq. NR_027008.1) was designed using the Benchling CRISPR design online tool (https://benchling.com [Biology Software]; accessed 2015) and cloned into a modified version of the plasmid pSpCas9(BB)-2A-GFP (PX458, a gift from Feng Zhang, Addgene plasmid #48138; (79)), where we fused a truncated form of human Geminin (hGem) to SpCas9 increasing homology-directed repair efficiency (80). A 200 nt ssDNA repair template (R26_oligo; 4 nmole Ultramer™ DNA Oligo, Standard Desalting, Integrated DNA Technologies) was designed with homology arms centered around the MIN-tag. Re-cleavage after repair template incorporation was prevented by co-delivering a CRISPR/Cas9-blocking mutation within the respective sgRNA PAM. To generate the homozygous R26^{MIN} mESC line, 500 000 cells were transfected in a six-well plate with 2.0 μg ssDNA repair template and 0.5 μg SpCas9 plasmid using Lipofectamine3000 (Thermo Fisher Scientific) according to the manufacturer's instructions. 48 h after transfection, GFP positive cells were enriched by fluorescence-activated cell sorting with a BD FACS Aria II (BD FACSDiva Software version 6.1.3, Firmware version 1.6, BD Biosciences). GFP-positive cells were pooled and plated at clonal density into a p100 cell culture dish. After 7 days, single colonies were picked manually using a 10 μl sterile pipette tip, transferred to a 96-well plate (flat bottom), and expanded. To screen for R26^{MIN} clones, 96-well plates were duplicated after cell outgrowth and genomic DNA isolated for screening by PCR and HincII restriction digest as described previously with minor modifications (76). Briefly, cells were washed two times with Dulbecco's PBS (D8537, Sigma-Aldrich), resuspended in 50 μl/well lysis buffer (50 mM TRIS/HCl pH 7.5, 10 mM $CaCl_2$, 1.7 μM SDS, 50 μg/ml Proteinase K), frozen at −80°C for 30 min, incubated at 56°C for 3 h, and finally Proteinase K heat inactivated at 85°C for 30 min. 2.5 μl/well of the resulting crude cell lysate were directly subjected to PCR (25 μl/rxn, 0.1 μl MyTaq™ DNA Polymerase, BIO-21107, Bioline) using the external screening primers R26_scr.fwd and R26_scr.rev and following cycling settings: 95°C/5 min – [95°C/30 s – 60°C/30 s – 72°C/30 s] × 45 – 72°C/40 s – 4°C/∞. MIN-tagged clones were identified by restriction fragment analysis of 7.5 μl PCR product using the HincII restriction site located within the MIN-tag (20 μl/rxn, 0.25 μl FastDigest HincII, Thermo Fisher Scientific). R26^{MIN} candidates were further verified by Sanger sequencing (Mix2Seq, Eurofins Genomics) of R26 exon 1 after genomic DNA isolation using the QIAamp DNA Mini Kit (QIAGEN) according to the manufacturer's instructions.

Generation of stable cell lines

Bxb1-mediated recombination. 500 000 R26^{MIN} mESCs were transfected in a six-well plate with 1.25 μg of the respective MIN-tag compatible vector harboring an *attB* site and 1.25 μg Bxb1 integrase plasmid pCAG-NLS-HA-Bxb1 (a gift from Pawel Pelczar, Addgene plasmid #51271; (81)) using Lipofectamine3000 (Thermo Fisher Scientific) according to the manufacturer's instructions. After 48 h, mESCs were plated at clonal density in a p100 cell culture dish and 1 mg/ml G418 (A2167, AppliChem) was added. After 7 days, single colonies were picked manually using a 10 μl sterile pipette tip, transferred to a 96-well plate, and expanded. To screen for Bxb1 recombined clones, 96-well plates were duplicated after cell outgrowth and genomic DNA isolated for screening by PCR as described previously with minor modifications (76). Briefly, cells were washed two times with Dulbecco's PBS (D8537, Sigma-Aldrich), resuspended in 50 μl/well lysis buffer (50 mM TRIS/HCl pH 7.5, 10 mM $CaCl_2$, 1.7 μM SDS, 50 μg/ml Proteinase K), frozen at −80°C for 30 min, incubated at 56°C for 3 h, and finally Proteinase K heat inactivated at 85°C for 30 min. 2.5 μl/well of the resulting crude cell lysate were directly subjected to PCR (25 μl/rxn, 0.1 μl MyTaq™ Red DNA Polymerase, BIO-21110, Bioline) using the external screening primers R26_scr.fwd and R26_scr.rev in combination with attL_scr.fwd and following cycling settings: 95°C/5 min – [95°C/30 s – 60°C/30 s – 72°C/30 s] × 45 – 72°C/40 s – 4°C/∞. Bxb1 recombined clones stably harboring the respective synthetase (R26^{RS}) are identified by the attL_scr.fwd and R26_scr.rev PCR product. To validate stable R26^{RS} clones, genomic DNA was isolated using the QIAamp DNA Mini Kit (QIAGEN) according to the manufacturer's instructions and the PCR repeated on 20 ng purified genomic DNA.

PiggyBac transposition. 500 000 HEK293T or mESCs were transfected in a six-well plate with 1.875 μg of the respective donor plasmid and 0.625 μg PiggyBac transposase vector (System Biosciences, #PB200PA-1) using Lipofectamine3000 (Thermo Fisher Scientific) according to the manufacturer's instructions. After 48 h, cells were plated at 40% confluency in a p100 cell culture dish and the respective selection antibiotic, 1 μg/ml puromycin (A1113803, Thermo Fisher Scientific) or 1 mg/ml G418 (A2167, AppliChem), was added. Cells were passaged at least two times under selection pressure to generate stable polyclonal pools before commencing experiments. PiggyBac transposition was used to establish HEK293T cells stably expressing the respective synthetase (HEK293T^{RS}) and R26^{RS} mESC clones stably expressing the mSc/mNG fluorescent reporter harboring an amber mutated *GOI** coding sequence or *context** (R26^{RS}/PB^{GOI*} or R26^{RS}/PB^{context*}).

Transient transfections

300 000 or 40 000 stable HEK293T^{RS} cells per 12- or 96-well, respectively, were seeded into ncAA containing medium 4 h before transfection. 250 000 stable R26^{RS} mESCs were plated per 12-well 2 h before transfection and ncAA was added at transfection. Cells were transfected with 1.0 μg (12-well) or 225 ng (96-well) of the respective plasmid using Lipofectamine3000 (Thermo Fisher Scientific) according to the manufacturer's instructions and incubated for 24 h before flow-cytometry.

Flow-cytometry data collection and analysis

Transiently transfected R26^{RS} mESCs and HEK293T^{RS} cells were analyzed 24 h after transfection. Stable R26^{RS}/PB^{GOI*} or R26^{RS}/PB^{context*} mESCs were seeded at 30% confluency into ncAA containing medium in 12- or 96-wells and analyzed after 24 h.

For flow-cytometry, cells grown in 12- or 96-wells were washed with 1 ml or 200 μl Dulbecco's PBS (D8537, Sigma-Aldrich), dissociated with 100 μl or 28 μl Trypsin-EDTA in PBS (T4299, Sigma-Aldrich), and resuspended in 500 or 100 μl FluoroBrite Dulbecco's modified Eagle's medium (DMEM; A1896701, Thermo Fisher Scientific) supplemented with 10% fetal bovine serum (FBS; Sigma-Aldrich) and 100 U/ml penicillin, 100 μg/ml streptomycin (Pen/Strep; P4333, Sigma-Aldrich). Before acquisition, cells from 12-wells were filtered through a 35 μm cell strainer (352235, Corning) and cells in 96-well plates were thoroughly resuspended using a multichannel pipette. Cells were recorded on a BD LSRFortessa (BD FACSDiva Software version 8.0.1, Firmware version 1.4, BD Biosciences) with a BD High Throughput Sampler (HTS, BD Biosciences) for loading of 96-well plates.

Flow-cytometry data of the mSc/mNG dual-fluorescence reporter were processed in three steps with FlowJo (version 10.6.1, BD Biosciences) by (i) gating for single cells excluding debris (FSC-A/SSC-A) and doublets (FSC-A/FSC-H and SSC-A/SSC-H), (ii) gating for transfected/stable cells by excluding mSc negative cells, and (iii) calculating mSc and mNG mean fluorescence intensities (MFIs, see Supplementary Figure S7 and S8 for representative flow-cytometry data). MFIs were further analyzed with RStudio (version1.3.1093, R version 3.6.1; RStudio: Integrated Development Environment for R; RStudio, PBC, Boston, MA; http://www.rstudio.com) using the *tidyverse* (version 1.3.0) (82) and *rstatix* (version 0.6.0; https://rpkgs.datanovia.com/rstatix) R packages. Relative readthrough efficiency (RRE) for samples + or – ncAA and incorporation efficiencies for each position were calculated according to equations from Figure 3B. Flow-cytometry raw data and analysis files are available via FlowRepository (83) with the repository identifier FR-FCM-Z2N3.

Purification of amber suppressed endogenous proteins by streptavidin pulldown (SORT-E)

Proteomic incorporation of BcnK at amber stop codons. 1.6 × 10^6 R26^{MIN} and R26$^{RS\_BcnK}$ mESCs or wtHEK293T and HEK293T$^{RS\_BcnK}$ cells were seeded per p150 plate and after 2 h 0.5 mM BcnK diluted in 3 volumes of 1 M HEPES (15630056, Thermo Fisher Scientific) was added. After 66–72 h, cells on p150 plates were washed once with Dulbecco's PBS (D8537, Sigma-Aldrich), incubated for 1 h with fresh medium, washed a second time with PBS, and incubated for another 3 h with fresh medium. For harvesting, cells on p150 plates were washed once with PBS, dissociated with 2 ml trypsin–EDTA solution (T3924, Sigma-Aldrich), resuspended in 10 ml fresh medium, and collected by centrifugation at 500 g and 4°C for 5 min. Cell pellets were washed on ice two times by resuspending in 10 ml ice cold PBS and centrifugation at 500 g and 4°C for 5 min. Pellets were flash frozen in liquid N_2 and stored at −80°C.

Full proteome samples. For each sample, 10% of flash-frozen cell pellet from one p150 plate (ca. 2.5 × 10^6 cells) were lysed in 200 μl lysis buffer (6 M guanidinium Chloride, 100 mM Tris–HCl pH 8.5, 2 mM DTT). Samples were homogenized by pipetting and boiled at 99°C for 10 min with constant shaking at 1400 rpm. After quickly spinning down, samples were sonicated at 4°C for 15 min in 1.5 ml tubes using a Bioruptor® Plus sonication device (Diagenode) with the following settings: high intensity, 30 s on/30 s off cycle. Protein concentrations were then determined using the Pierce™ BCA Protein Assay Kit (23225, Thermo Fisher Scientific) according to the manufacturer's instructions for microplate settings. Meanwhile, samples were alkylated with 40 mM chloroacetamide (CAA) for 20 min at room temperature. Afterwards, 30 μg of lysate was diluted in a total volume of 50 μl lysis buffer supplemented with 40 mM CAA and 2 mM dithiothreitol (DTT). Samples were then diluted 1:10 with digestion buffer (10% acetonitrile, 25 mM

Tris–HCl pH 8.5). To each sample 0.6 μg Trypsin (Pierce™ Trypsin Protease, 90058, Thermo Fisher Scientific) and 0.6 μg LysC (Pierce™ LysC Protease, 90051, Thermo Fisher Scientific) was added and proteins were digested overnight at 37°C with constant shaking at 1100 rpm. The next day, protease digestion was stopped by adding 1% (v/v) trifluoroacetic acid (TFA) and samples were loaded on StageTips containing three layers of SDB-RPS matrix (Empore) in a 200 μl pipette tip according to standard protocol (84). After one washing step with 0.1% (v/v) TFA, peptides were eluted into 60 μl of 80% acetonitrile and 2% ammonium hydroxide. Evaporation of the eluates was performed in a SpeedVac centrifuge and peptides were subsequently resuspended in 20 μl of A* buffer (0.1% TFA, 2% acetonitrile) and shook for 10 min at 2000 rpm at room temperature prior to peptide concentration estimations at 280 nm.

In vitro chemoselective labeling of BcnK tagged proteomes with biotin-tetrazine conjugate. For each sample, 90% of flash frozen cell pellet from one p150 plate (ca. 22.5 × 10^6 cells) were lysed on ice with 1 volume of ice cold RIPA buffer (50 mM Tris–HCl pH 8.0, 150 mM NaCl, 0.1% UltraPure™ SDS Solution (24730020, Invitrogen), 0.5% sodium deoxycholate detergent, 1% Triton X-100; freshly add 1× cOmplete™ EDTA-free Protease Inhibitor Cocktail (04693132001, Roche)) and sonicated at 4°C for 20 min in 1.5 ml tubes using a Bioruptor® Plus sonication device (Diagenode) with the following settings: high intensity, 30 s on/30 s off cycle. Lysates were subsequently cleared by centrifugation at 20 000 g and 4°C for 15 min and supernatants collected. Protein concentrations were then determined using the Pierce™ BCA Protein Assay Kit (23225, Thermo Fisher Scientific) according to the manufacturer's instructions for microplate settings. Cleared cell lysates were diluted with RIPA buffer to a final concentration of 3 mg/ml protein and 1 ml lysate was typically used. Therefore, 1 ml lysates (3 mg protein input) were reduced with 2 mM DTT for 30 min on ice and subsequently alkylated with 40 mM chloroacetamide (CAA) for 45 min on ice. 7.5 μM biotin-tetrazine conjugate (2.5 nmol biotin-tetrazine/1 mg protein input) was then added to lysates and incubated overnight at 4°C with end-over-end rotation. The next day, 50 μl aliquots (150 μg protein) were boiled for 10 min at 95°C in 1× Laemmli buffer supplemented with 20 mM DTT as input samples for analysis by western blot.

Streptavidin pulldown. 60 μl settled resin per sample (binding capacity: 160 μg biotinylated BSA/1 mg protein input) of Pierce™ High Capacity NeutrAvidin™ Agarose (29202, Thermo Fisher Scientific) were washed three times in 3 volumes of RIPA buffer, diluted in RIPA buffer to 120 μl slurry per sample, and added to biotin-tetrazine labeled lysates. Samples were then incubated for 2 h at room temperature with end-over-end rotation. After 2 h, 50 μl aliquots of supernatants were boiled for 10 min at 95°C in 1× Laemmli buffer supplemented with 20 mM DTT as unbound fraction for analysis by western blot. The remaining supernatant was aspirated and agarose beads were washed on ice by resuspending in 1 ml of the following buffers and centrifugation for 3 min at 500 g and 4°C: two times in RIPA buffer, once in 1 M KCl, once in 100 mM Na_2CO_3, and twice in urea buffer (2 M urea solution (U4883, Sigma-Aldrich), 50 mM ammonium bicarbonate). During the last washing step, agarose beads were transferred to a fresh 1.5 ml tube. For western blot analysis, 10% of agarose beads were washed two more times in RIPA buffer and proteins eluted by boiling for 10 min at 95°C in 1× Laemmli supplemented with 20 mM DTT and 2 mM biotin. For mass spectrometry analysis, peptides were eluted from beads by resuspending in 200 μl elution buffer (1 M urea solution (U4883, Sigma-Aldrich), 50 mM ammonium bicarbonate) and on-beads-digest with 1.5 μg Pierce™ trypsin protease (0.5 μg trypsin/1 mg protein input; 90058, Thermo Fisher Scientific) for 18–20 h shaking at 30°C and 1300 rpm. Trypsinization was stopped by adding 1% (v/v) trifluoroacetic acid (TFA) and samples were stored at −20°C. Eluted peptides were desalted and concentrated using C18 based StageTips according to standard protocol (84). Evaporation of the eluates was performed in a SpeedVac centrifuge and peptides were subsequently resuspended in 20 μl of A* buffer (0.1% TFA, 2% acetonitrile) and shook for 10 min at 2000 rpm and room temperature prior to peptide concentration estimations at 280 nm.

LC–MS/MS

Acquisition of full proteomes and SORT-E eluates. Each sample was loaded on a 50 cm C18-based reversed phase column (in-house packed with ReproSil-Pur C18-AQ 1.9 μm resin from Dr Maisch a total inner diameter of 75 μm), which was mounted on an EASY-nLC 1200 (Thermo Fisher Scientific) ultra-high pressure system and constantly kept at 60°C. The liquid chromatography was coupled to a Q Exactive HF-X Hybrid Quadrupole-Orbitrap Mass Spectrometer (Thermo Fisher Scientific) via a nano-electrospray source and operational parameters were monitored by SprayQc. Peptides were eluted constantly at around 300 nl/min during a 120 min non-linear ACN gradient. After each set of replicates ($R26^{MIN}$ + wtHEK293T in triplicates and $R26^{RS\_BcnK}$ + $HEK293T^{RS\_BcnK}$ in quadruplicates) an additional wash step was scheduled. Data-dependent acquisition was applied; after sequential full scans (maximum injection time: 20 ms, resolution: 60 000, target value 3 × 10^6) the most abundant 12 ions were addressed to MS/MS scans. The m/z range was limited to 400–1650 m/z.

The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE (85) partner repository with the dataset identifier PXD019815.

Computational analysis of raw MS data of full proteomes and SORT-E samples. Analysis of raw MS data was accomplished by the MaxQuant software package (version 1.6.11.0 (86)). The underlying FASTA files for peak list searches were derived from Uniprot by including both reviewed and unreviewed proteomes (mouse proteome, version October 2018; human proteome, version May 2020). An additional common contaminant list comprising 262 entries was applied using the Andromeda search engine (87). The 'Match between runs' option was enabled and the FDR was set to 1%, which applies on protein and peptide level (minimum of seven amino acids). Relative quantification of proteins was accomplished by the MaxLFQ algorithm (88). The cut-off was set to a minimal ratio count of two peptides.

For both full proteome and SORT-E samples the initial MaxQuant output was analyzed by Perseus (version 1.6.2.3). Here, common contaminants and protein groups measured less than twice within at least one set of replicates were filtered out and LFQ values were transformed into $\log_2$-values.

Statistical analysis of full proteomes. For full proteomes, imputation of missing values was based on a gaussian distribution relative to the standard deviations of measured values (width of 0.2 and a downshift of 1.8 standard deviations). Student's *t*-tests of $\mathrm{R26}^{\mathrm{RS\_BcnK}}$ versus $\mathrm{R26}^{\mathrm{MIN}}$ and $\mathrm{HEK293T}^{\mathrm{RS\_BcnK}}$ versus wtHEK293T were performed with a permutation-based FDR of 0.05 and a minimal $\log_2$ fold change of 1 (S0).

GO analysis of differentially expressed proteins according to both Student's *t*-tests was performed by using the Panther classification system (89). Here, the up- and down-regulated proteins were analyzed together due to the low amount of significantly changing proteins. GO terms with a lower fold change of 2 or a higher *P*-value than 0.05 were excluded.

Preprocessing of proteomics data for iPASS linear regression model. SORTE-E samples were normalized for differences in protein expression levels by subtracting the full proteome LFC (expression level) from their matched SORT-E LFC (enrichment in pulldown). The resulting normalized replicates of control ($\mathrm{R26}^{\mathrm{MIN}}$; wtHEK293T) and amber suppressed samples ($\mathrm{R26}^{\mathrm{RS\_BcnK}}$; $\mathrm{HEK293T}^{\mathrm{RS\_BcnK}}$) were tested for significant difference using a two-sided Student's *t*-test. Proteins with a *P*-value <0.01 were considered significantly enriched. Stop codon identity, sequence context, and GC content around the stop codon (positions −6 to +9) were extracted from coding sequences and cDNA assemblies of *Mus musculus* (GRCm38) and *Homo sapiens* (GRCh38) using custom Python scripts and assigned to the respective proteins. Isoforms of individual proteins were filtered by keeping only unique sequence contexts.

Linear regression model (iPASS model)

Linear regression analysis. Regression analysis was performed as described previously (61). In brief, to predict amber suppression efficiencies for a given UAG context we employed a linear regression model based on the sequence context (SC), the GC content, and their normalized fold changes obtained by SORT-E (see above). The SC included the stop codon itself (positions +1 to +3) and the nucleotide sequences surrounding the stop codon (positions −6 to -1, +4 to +9). Nucleotide sequences were represented by indicator vector coding. Here, 12×4 binary vector entries are used to indicate the presence [1] or absence [0] of a nucleotide (A, C, G, or U) at a particular position (−6 to −1, +4 to +9) surrounding the stop codon. Three further entries are reserved to indicate the type of stop codon (UAA, UAG, or UGA; positions +1 to +3) and a separate column for the GC content of the sequence from positions −6 to +9. The resulting feature vectors of all sequences were scaled using the *preProcess* function of the (v6.0–78) R package (90). Regularized ridge regression was performed using the *glmnet* (v2.0–13) R package.

For the estimation of the regression model coefficients, we performed a regularized least-squares ('ridge') regression (91). Let $\mathbf{X}$ be the $n \times d$ matrix of n sequence feature vectors with dimensionality d and $\mathbf{y}$ be the (n-dimensional) vector of readthrough values associated with the sequences. Then the weight vector $\mathbf{w} = (\mathbf{XTX} + k \times \mathbf{I}){-}1 \times \mathbf{XTy}$ represents the solution of the linear least-squares problem and $y = \mathbf{wTx}$ corresponds to the RTP value y for a sequence feature vector $\mathbf{x}$. The minimum loo-cv error (lambda) in terms of the sum of squared deviations of predictions from known readthrough values was 0.13 for $k = 10^{0.3}$ ($\sim$1.995).

The decoy model was created as described above keeping identical proteins and SCs but randomly reshuffling LFC values.

Feature elimination analysis. Starting from the complete iPASS model, we removed the variable corresponding to the minimum sum of squared regression coefficients. The residual error was then calculated for the remaining variables (including the stop codon) as described above. This procedure was repeated until only the stop codon variable was left.

Probability logo construction. To construct a probability logo (motif) reflecting sequence contexts for efficient amber suppression, we first generated all possible 12-mer sequence contexts (4^{12}) comprised of the nucleotides 6 bp up- and downstream of a central amber stop codon (nucleotides –6 to –1 and +4 to +9; stop codon at +1, +2, +3) *in silico*. After removal of sequences containing in-frame stop codons, we used the iPASS model to predict amber suppression efficiencies of all 13 845 841 *k*-mers. To construct a probability logo, this list of *k*-mers together with their iPASS scores was used as the input for kpLogo (v1.1) (92) with the options *-k 1 -weighted*.

Supplementary material and methods

Additional material and methods including plasmid construction, chemicals and chemical synthesis, and western blotting are available as supplementary material at NAR online. NMR spectra of synthesized chemicals are depicted in Supplementary Figure S13. Uncropped SDS-gel and blots are presented in Supplementary Figure S14 and S15. Plasmids used and cloned and oligonucleotides used in this study are listed in Supplementary Table S1 and S2 within Supplementary Material and Methods. Plasmids cloned in this study have been deposited at Addgene with the IDs 167491–99.

RESULTS

Step-wise generation of stable mESC and HEK293T lines with an expanded genetic code

Efficient amber suppression in mammalian cells has been reported to depend on high suppressor tRNA expression levels (22,23,25,27), whereas two *PylS* copies are sufficient to expand the genetic code of mice (13). We therefore speculated that efficient amber suppression in mESCs could be achieved by biallelic integration of a construct harboring one copy of *PylS* and four copies of *PylT* into the *Gt(ROSA)26Sor* (R26) genomic safe harbor locus (77,78). To this end, we applied our previously developed multifunctional integrase (MIN) tag genome engineering strategy (76), which leverages Bxb1-mediated recombination between *attB* and *attP* attachment sites (Figure 1A). In a first step, we established a monoclonal mESC line harboring the *attP* site (MIN-tag) within R26 (R26$^{\text{MIN}}$) using CRISPR/Cas9 gene editing. Homozygous integration of the MIN-tag into the R26 locus was confirmed by agarose gel electrophoresis and Sanger sequencing (Supplementary Figure S1A+B). In a second step, the MIN-tagged R26 serves as a genetic entry site for the rapid and selective integration of *PylS*/*PylT* pairs. To construct a targeting vector compatible with PB- as well as Bxb1-mediated genomic integration, we modified a previously reported 4x*PylT*/*PylS* vector (30) to include an *attB* attachment site (Figure 1A). The encoded wtPylRS transfers the pyrrolysine analog tert-butoxycarbonyl-L-lysine (BocK, Supplementary Figure S1C) onto PylT in mammalian cells (7). To expand the toolkit of ncAAs that can be incorporated, we also generated targeting vectors encoding two additional *PylS* variants: (i) *PylS_DiazK* to incorporate the ncAA methyl-diazirin-L-lysine (DiazK, Supplementary Figure S1C+D) bearing a diazirine moiety for site-directed photocrosslinking of proteins (6,93–95); (ii) *PylS_BcnK* to incorporate the ncAA bicyclo[6.1.0]nonyne-L-lysine (BcnK, Supplementary Figure S1C) bearing a strained alkyne motif for selective labeling with tetrazine conjugates via inverse electron-demand Diels–Alder cycloaddition (iEDDAC) (96–99). Additionally, we N-terminally tagged wild-type (wt) as well as engineered PylRS variants with a nuclear export signal (NES), which has been reported to enhance amber suppression efficiency up to 15-fold (19). After co-transfecting the targeting vector and a Bxb1 expression plasmid, we selected for stable *NES-PylS*/4x*PylT* integrants using the co-delivered neomycin resistance cassette. We verified that the three *NES-PylS*/4x*PylT* targeting vectors had been inserted into both R26 alleles (R26$^{\text{RS}}$, Figure 1B) in each of the newly generated cell lines: (i) *NES-wtPylS*/4x*PylT* (clone: R26$^{\text{wtRS}}$); (ii) *NES-PylS_DiazK*/4x*PylT* (clone: R26$^{\text{RS-DiazK}}$); and (iii) *NES-PylS_BcnK*/4x*PylT* (clone: R26$^{\text{RS-BcnK}}$). In addition, we generated three polyclonal HEK293T lines by PB-mediated stable integration (30) of *NES-PylS*/4x*PylT* cassettes and puromycin selection (HEK293T$^{\text{RS}}$, Figure 1A): (i) *NES-wtPylS*/4x*PylT* (cell line: HEK293T$^{\text{wtRS}}$); (ii) *NES-PylS_DiazK*/4x*PylT* (cell line: HEK293T$^{\text{RS-DiazK}}$); and (iii) *NES-PylS_BcnK*/4x*PylT* (cell line: HEK293T$^{\text{RS-BcnK}}$).

To validate amber suppression in stable R26$^{\text{RS}}$ cell lines, we transiently transfected a *4xPylT*/*mCherry-TAG-EGFP* reporter construct (30), which expresses full-length mCh-eGFP upon efficient decoding of the amber stop codon. After 24 h in the presence or absence of the respective ncAA, we analyzed mCh and eGFP fluorescence by flow-cytometry (Figure 1C, Supplementary Figure S2). In comparison to the incorporation of BocK in R26$^{\text{wtRS}}$ and DiazK in R26$^{\text{RS-DiazK}}$, BcnK was less efficiently incorporated in R26$^{\text{RS-BcnK}}$, indicated by the lower correlation between mCh and eGFP fluorescence. This difference may be attributable to the reduced PylRS aminoacylation activity of BcnK compared to BocK or DiazK (99). Furthermore, the transfection efficiency of stable R26$^{\text{RS}}$ clones was generally low (∼25%, Supplementary Figure S2B). Genomic integration of *PylS*/*PylT* via the MIN-tag strategy allowed us to use PB transposition in a second step to establish polyclonal pools stably co-expressing multiple copies of the gene of interest harboring an in-frame amber stop codon (*GOI**; cell line: R26$^{\text{RS}}$/PB$^{\text{GOI*}}$). As a proof of principle, we stably integrated a *4xPylT*/*sfGFP*^{N150*} reporter construct, which harbors the amber stop codon at position 150 of *sfGFP* (30), into R26$^{\text{wtRS}}$ and R26$^{\text{RS\_BcnK}}$ cells. After selection with puromycin, we analyzed expression of full-length sfGFP in the absence or presence of BocK or BcnK by flow-cytometry (Figure 1D). Both R26$^{\text{RS}}$/PB$^{\text{sfGFPN150*}}$ stable cell lines suppressed the amber stop codon within *sfGFP*^{N150*} upon induction with the respective ncAA. This demonstrates that two *PylS* copies expressed from the R26 genomic locus of mESCs are sufficient to direct efficient amber suppression. In summary, we established stable and defined mESC clones capable of amber suppression and compatible with PB transposition to genomically integrate *4xPylT*/*GOI** expression cassettes.

A linear regression model of amber stop codon contexts to predict ncAA incorporation efficiencies

Next, we wondered to what extent the nucleotide composition around UAG determines ncAA incorporation efficiencies. Genetic code expansion with *PylS*/*PylT* in HEK293T cells has been reported to suppress endogenous amber stop codons resulting in off-target labeling of the cellular proteome (25,99). In line with these studies, we observed widespread amber suppression of endogenous proteins in the stable R26$^{\text{RS\_BcnK}}$ mESC clone by in-gel fluorescence analysis of BcnK harboring proteins with a silicon rhodamine-tetrazine conjugate (SiR-Tet) (Supplementary Figure S1E, S3A). We hypothesized that assessment of ncAA incorporation rates in living cells at endogenous UAG contexts would allow us to identify sequence motifs that stimulate amber suppression in a target *GOI**. To this end, we leveraged the chemoselective iEDDAC reaction between BcnK and tetrazine-based probes (97,98) to implement a novel adaptation of the SORT-E strategy by Elliott *et al.* (75). Here, we tagged BcnK harboring endogenous proteins with a biotin–tetrazine (Biotin-Tet) probe (Supplementary Figure S1E), which enables amber suppressed proteins to be selectively enriched by streptavidin pulldown and subsequently identified by mass spectrometry (75). Since pulldown by streptavidin initially depends on ncAA incorporation and hence amber suppression, endogenous proteins with UAG contexts permitting high BcnK incorporation rates should be enriched and can be subsequently extracted for bioinformatic analysis.

We first verified labeling of whole cell lysates with Biotin-Tet by western blot after BcnK incorporation in R26$^{\text{RS\_BcnK}}$ mESCs (Supplementary Figure S3B). Furthermore, we observed a marked enrichment of Biotin-Tet-labeled endogenous proteins after streptavidin pulldown from R26$^{\text{RS\_BcnK}}$ lysates compared to those from the R26$^{\text{MIN}}$ entry cell line lacking *PylS*/*PylT* (Figure 2A). To specifically evaluate

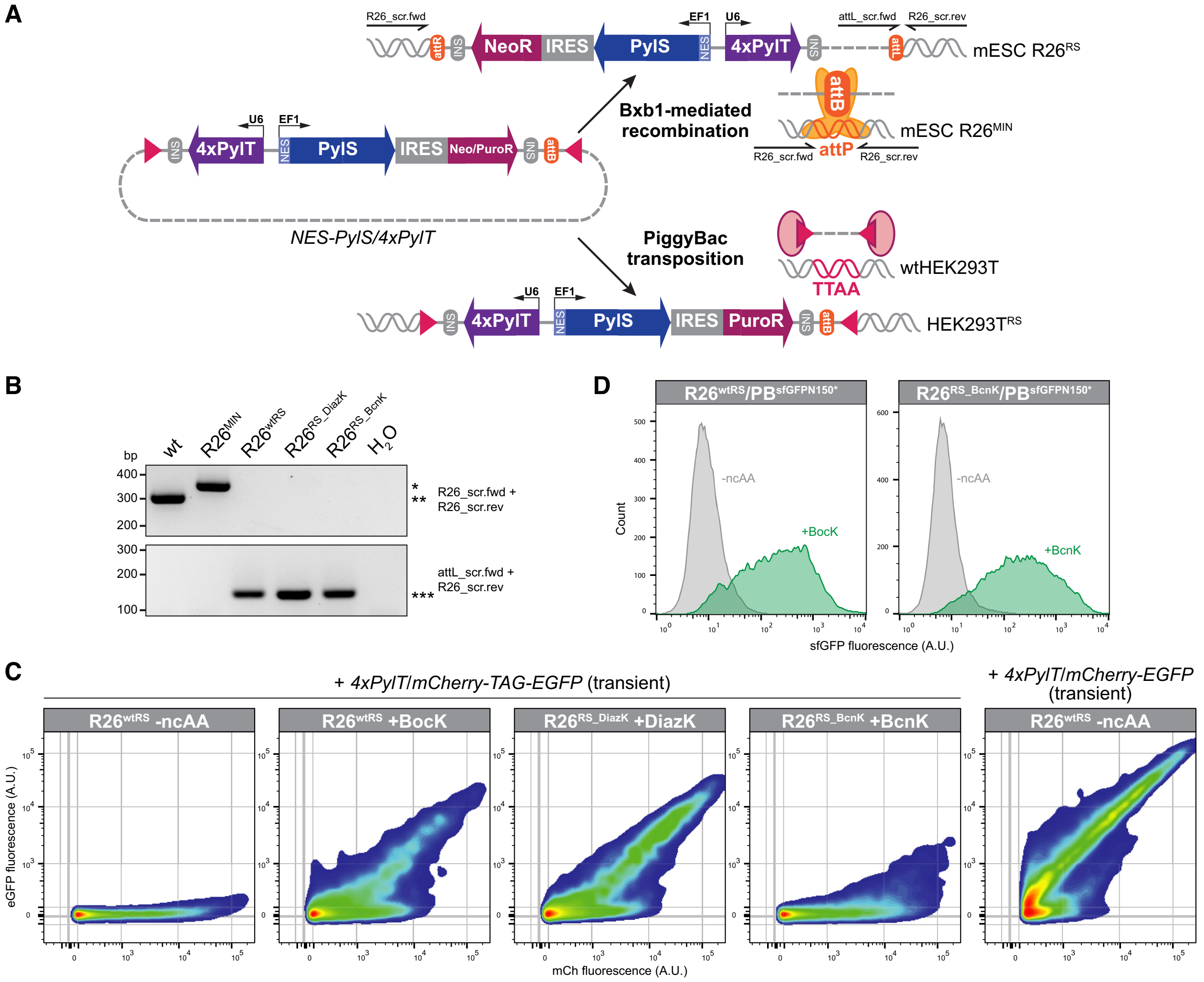


Figure 1. (**A**) Strategy to generate stable cell lines with an expanded genetic code. Bxb1-mediated recombination in mouse embryonic stem cells (mESCs) homozygous (only one allele is depicted) for the MIN-tag within the *Rosa26* locus (R26^{MIN}) or PiggyBac (PB) transposition in wild-type human embryonic kidney cells (wtHEK293T). The Bxb1 integrase specifically recombines the attachment sites *attP* and *attB* to generate *attR* and *attL* sites that flank the integrated vector, whereas the PB transposase integrates the cassette that is flanked by inverted terminal repeats (ITRs, indicated as rectangles) into TTAA chromosomal sites. Primer binding sites used for screening of stable R26^{RS} clones are indicated. Abbreviations: *Methanosarcina mazei* tRNA^{Pyl} synthetase (PylS) N-terminally fused to a nuclear export signal (NES), tRNA$^{Pyl}_{CUA}$ (PylT), internal ribosomal entry site (IRES), neomycin resistance (NeoR), puromycin resistance (PuroR), constitutive EF1α promoter (EF1), constitutive U6 promoter (U6), insulator (INS). (**B**) Sequential stable integration of the MIN-tag and *NES-PylS/4xPylT* into the *Rosa26* locus (R26) in mESCs. Agarose gel electrophoresis of screening PCRs using the indicated primers. Homozygous integration of the MIN-tag (*attP* site) into R26 results in a 48 bp shift (*) compared to wt (**). Subsequent Bxb1-mediated stable integration of *NES-wtPylS/4xPylT* (R26^{wtRS}), *NES-PylS_DiazK/4xPylT* (R26$^{RS\_DiazK}$), or *NES-PylS_BcnK/4xPylT* (R26$^{RS\_BcnK}$) generates an *attL* binding site (***). (**C**) Variable amber suppression efficiencies in stable mESC lines expressing PylRS variants. R26^{wtRS}, R26$^{RS\_DiazK}$, or R26$^{RS\_BcnK}$ mESCs transiently transfected with the *4xPylT/mCherry-TAG-EGFP* reporter construct were cultured for 24 h in the presence of the indicated ncAA (0.5 mM final concentration). The *4xPylT/mCherry-EGFP* reporter construct lacking the amber stop codon was transiently transfected as reference of the optimal mCherry/EGFP ratio. For flow-cytometry measurements, 9000 mCh positive single cells per condition were analyzed (for gating strategy and complete panel of dot plots see Supplementary Figure S2). Fluorescence intensities are indicated in arbitrary units (A.U.). (**D**) Efficient incorporation of ncAAs into target proteins in mESCs after stable integration of both *NES-PylS/4xPylT* and amber transgene. The *4xPylT/sfGFP*^{N150*} reporter construct harboring the amber stop codon within the *sfGFP* ORF was integrated in R26^{wtRS} or R26$^{RS\_BcnK}$ mESCs by PB transposition (R26^{wtRS}/PB^{sfGFPN150*} and R26$^{RS\_BcnK}$/PB^{sfGFPN150*}). Incorporation of the respective ncAA (0.5 mM final concentration) into sfGFP^{N150*} was verified after 48 h by flow-cytometry analysis of 20 000 single cells per condition. Fluorescence intensity of sfGFP is indicated in arbitrary units (A.U.).

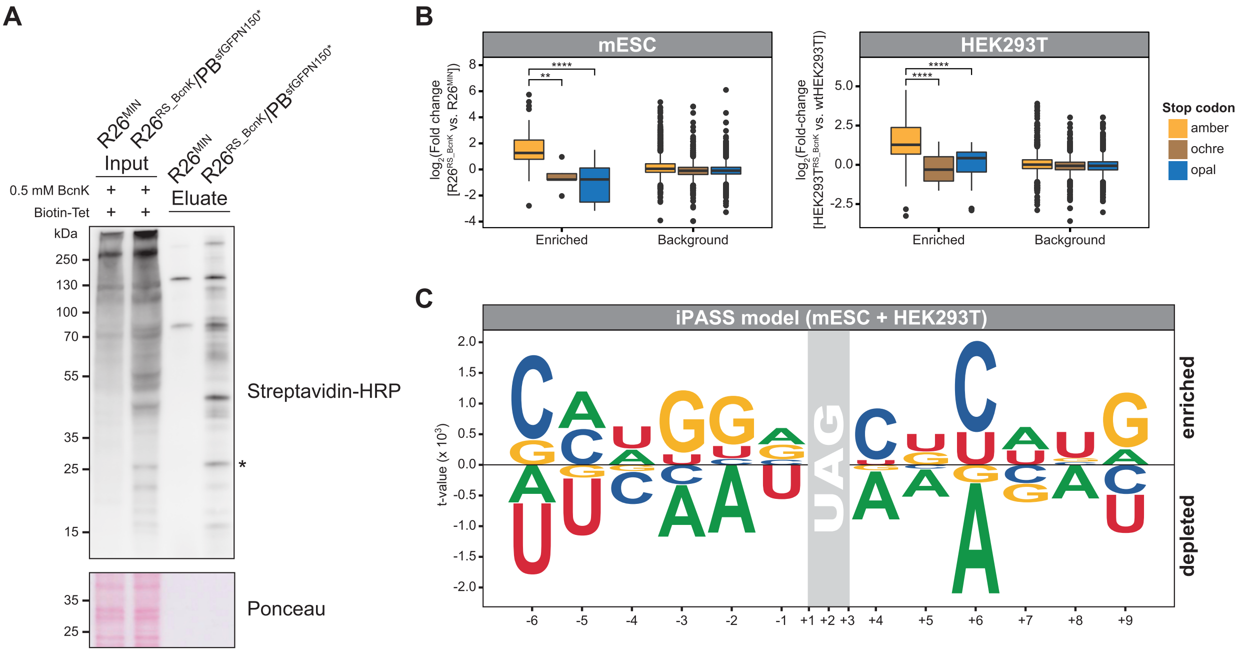


Figure 2. (**A**) Selective streptavidin pulldown of amber suppressed endogenous proteins in stable R26$^{RS\_BcnK}$ mESCs after covalent labeling with a biotin-tetrazine (Biotin-Tet) probe (SORT-E approach). Stable R26$^{RS\_BcnK}$/PB^{sfGFPN150*} and R26^{MIN} mESCs were cultured for 68 h in the presence of 0.5 mM BcnK, whole cell lysates labeled with Biotin-Tet, and biotinylated proteins captured by streptavidin pulldown. Input and eluate samples were subjected to western blotting using a streptavidin-HRP conjugate and Ponceau S staining as loading control. Amber suppressed sfGFP^{N150*} (*) is indicated. (B, C) After SORT-E and analysis of pulldowns by LC–MS/MS (Supplementary Figure S3C), significantly enriched proteins were identified and a linear regression model of enriched stop codon sequence contexts calculated (Supplementary Figure S5A). (**B**) Selective enrichment of amber suppressed endogenous proteins by SORT-E from R26$^{RS\_BcnK}$ mESCs and HEK293T$^{RS\_BcnK}$ cells cultured with BcnK. The fold change (PylRS_BcnK versus respective control cell line) of proteins harboring one of the three stop codons (amber, ochre, opal) is depicted for the enriched ($P < 0.01$) and background ($P > 0.01$) fraction in HEK293T and mESCs. PylRS_BcnK and control cell lines were analyzed by LC–MS/MS in biological quadruplicates and triplicates, respectively. ANOVA + Tukey's honestly significant difference post-hoc test: **$P < 0.01$, ****$P < 0.0001$. (**C**) Probability logo of 12-mer sequence contexts comprised of the nucleotides 6 bp up- and downstream of the amber stop codon (nucleotides –6 to –1 and +4 to +9; stop codon at +1, +2, +3). Each UAG sequence context (4^{12} sequences) was weighted by its computed iPASS score. Character height corresponds to the *t*-value (two-tailed unpaired two-sample Student's *t*-test) with positive and negative values indicating enriched and depleted sequence contexts, respectively. The iPASS model combines the linear regression analysis of significantly enriched stop codon sequence contexts after SORT-E from both mESCs and HEK293T cells.

ncAA incorporation rates at endogenous UAG contexts, we cultured R26^{MIN} and R26$^{RS\_BcnK}$ mESCs or wtHEK293T and HEK293T$^{RS\_BcnK}$ cells in the presence of BcnK for 3 days and, after labeling of cellular lysates with Biotin-Tet, performed streptavidin pulldowns (Supplementary Figure S3C–E). To control for protein expression levels, whole cell lysates were collected in parallel for subsequent full proteome measurements from the same samples (Supplementary Figure S3C). Eluates and full proteomes from PylRS_BcnK and respective control cell lines were then analyzed by LC–MS/MS. Principal component analyses (PCA) of LC–MS/MS data revealed that the global proteome was negligibly altered in response to amber suppression with BcnK (Supplementary Figure S4A), with only 61 and 51 proteins exhibiting altered expression in R26$^{RS\_BcnK}$ and HEK293T$^{RS\_BcnK}$ cells, respectively (Supplementary Figure S4B). Additionally, GO analysis suggested no significant translational perturbance or clear enrichment of a specific pathway, arguing against a directed cellular response to amber suppression with BcnK (Supplementary Figure S4C). In contrast to the full proteome, streptavidin pulldowns from R26$^{RS\_BcnK}$ and HEK293T$^{RS\_BcnK}$ cells clustered apart from their respective control cell lines in PCA, indicating enrichment of distinct proteins (Supplementary Figure S4A). To identify amber suppressed proteins, we first normalized the protein abundance measured in each pulldown to that measured in each corresponding full proteome, and then determined enrichment by calculating the mean fold changes in protein levels between PylRS_BcnK and respective control cell lines (Supplementary Figure S5A). Importantly, normalization of pulldowns to full proteomes enables the extent of amber suppression for each detected protein to be determined irrespective of its abundance. Using these parameters, we identified 123 and 101 proteins that were significantly enriched ($P < 0.01$) in mESC R26$^{RS\_BcnK}$ and HEK293T$^{RS\_BcnK}$ pulldowns, respectively (Supplementary Figure S5A, Supplementary Data 1). The majority of all significantly enriched proteins (69%) are terminated by UAG in contrast to the low proportion (27%) in the background fraction ($P > 0.01$) (Supplementary Figure S5A, Supplementary Data 1). This enrichment compares favorably to the theoretical proportion (23%) of proteins terminating at UAG in mammals (41,100), validating the specificity of the streptavidin pulldown for amber suppressed proteins. Furthermore, fold changes of these proteins were significantly higher compared to proteins containing ochre (UAA) or opal (UGA) stop codons (Figure 2B). Additionally, amber suppressed proteins were enriched independently of their cellular abundance as determined by LC–MS/MS analysis of full proteomes (Supplementary Figure S5B, Supplementary Data 1). Taken together, these data clearly demonstrate that the amberome of mammalian cell lines with an expanded genetic code can be specifically captured and identified by mass spectrometry.

To define the impact of UAG sequence context on amber suppression efficiency, we analyzed the base composition surrounding the termination codons of proteins significantly enriched in pulldowns upon amber suppression. In particular, we focused our analysis on nucleotides 6 base pairs (bp) up- and downstream of the stop codon (nucleotides –6 to –1 and +4 to +9; stop codon at +1, +2, +3) as nucleotides up to the +9 position have been reported to substantially influence translational termination in mammalian cells (60,63). To predict context-specific ncAA incorporation rates at amber stop codons *in silico*, we adapted a linear regression model that has been previously applied to assess genome-wide translational readthrough at stop codons (61). In this approach, significantly enriched stop codon contexts are encoded in a multi-dimensional binary vector space and correlated with their relative fold change determined by SORT-E. In addition to the sequence context, we also included the GC content, which has been reported as one of the most informative structural features governing eukaryotic translational readthrough *in vitro* (64). We first computed linear regression models of UAG contexts separately for mESCs and HEK293T cells. Importantly, HEK293T fold changes measured by SORT-E were linearly correlated with values calculated by the mESC regression model and *vice versa* (Supplementary Figure S5C). This reciprocal validation with data from mESCs or HEK293T cells both confirms the reliability of the regression models to predict relative ncAA incorporation efficiencies and also indicates that amber suppression efficiency is subject to similar context effects in mESCs and HEK293T cells. By combining SORT-E data from mESC and HEK293T cell lines, we then computed a mammal-specific regression model that predicts ncAA incorporation efficiency based on sequence context, which we refer to as iPASS (Identification of Permissive Amber Sites for Suppression).

To extract sequences permitting high amber suppression efficiencies according to iPASS, we weighted each 12-mer amber stop codon context (nucleotides –6 to –1 and +4 to +9) by its iPASS score and computed their probability logo using kpLogo (92) (Figure 2C). Whereas the UAGC tetranucleotide is largely underrepresented as a termination codon in mammalian genes (43), we detected an enrichment of +4 C. Interestingly, +4 C has been described as one of the strongest predictors of translational readthrough by near-cognate tRNAs (63,64,66). Moreover, in the presence of an amber suppressor tRNA, particularly the UAGC tetranucleotide has been reported to permit above-average amber suppression in mammalian cell lines (63,67,68). Additionally, we found purines depleted at the +4 position of efficiently amber suppressed proteins. Interestingly, +4 purines are thought to stabilize formation of the termination complex (101,102) and are generally associated with low translational readthrough in mammalian cells (61,63,66). Remarkably, we detected the strongest nucleotide enrichment at +6 for C as well as depletion of A. In general, distinct enrichment of nucleotides across all positions investigated indicate that ncAA incorporation efficiency is not determined by the identity of a single nucleotide, but rather modulated by a synergistic interplay of nucleotides surrounding the amber stop codon. To further characterize the relative impact of each position on ncAA incorporation efficiency, we performed feature selection by successively removing positions with the smallest contribution to the regression error (Supplementary Figure S5D). Eliminating positions that flank the stop codon gradually increased the regression error, with +4 and –3 having the strongest effect. Therefore, reducing the number of iPASS input values decreases the accuracy of prediction, which highlights the relative importance of flanking sequences on ncAA incorporation efficiency. In summary, by quantifying BcnK incorporation at endogenous amber stop codons in mESCs and HEK293T cells, we developed a linear regression model called iPASS revealing a synergistic influence of the surrounding sequence context on ncAA incorporation efficiency in mammalian cells.

Development of a dual-fluorescence reporter to identify permissive ncAA incorporation sites

To further validate the accuracy of the iPASS model in predicting UAG context-dependent ncAA incorporation efficiencies, we developed a dual-fluorescence reporter to experimentally assess the efficiency of amber suppression across multiple incorporation sites within a target protein in living cells. In combination with flow-cytometry, dual-fluorescence reporters harboring the amber stop codon within the linker region were recently evaluated in yeast to accurately measure ncAA incorporation efficiencies (72). In contrast to previously developed reporters, we constructed a PB and Bxb1-compatible *mSc-P2A-GOI*-P2A-mNG/4xPylT* reporter construct, in which 2A peptides cotranslationally cleave the ncAA-containing protein of interest (POI^{ncAA}) from two flanking monomeric fluorescent proteins, mScarlet (mSc) and mNeonGreen (mNG) (Figure 3A). As such, our reporter can be used to assess ncAA incorporation efficiency at a specific site within any user-defined POI^{ncAA}. Both mSc and mNG exhibit superior brightness and a favorable maturation speed compared to their conventional spectral counterparts mRFP and GFP/YFP, respectively (103,104). The physiological concentration of a fluorescent protein is directly proportional to its fluorescence intensity (105,106) and amber suppression of the *GOI** leads to the equimolar translation of mSc and mNG. Importantly, cotranslational cleavage by 2A peptides uncouples mSc and mNG from POI^{ncAA} turnover rates. Therefore, the ratio between mSc and mNG fluorescence can be calculated to compare amber suppression efficiency between different *GOI** mutants independently of the respective POI^{ncAA} stability. Furthermore, translational reinitiation downstream of an in-frame UAG has been reported to occur within the first 70 codons in yeast (107) and the first 160 codons in mammalian cells (108), leading to the leaky expression of N-terminally truncated proteins. By encoding *mSc* upstream of the *GOI**, the resulting transcript lacks an in-frame UAG within the first 250 codons follow-

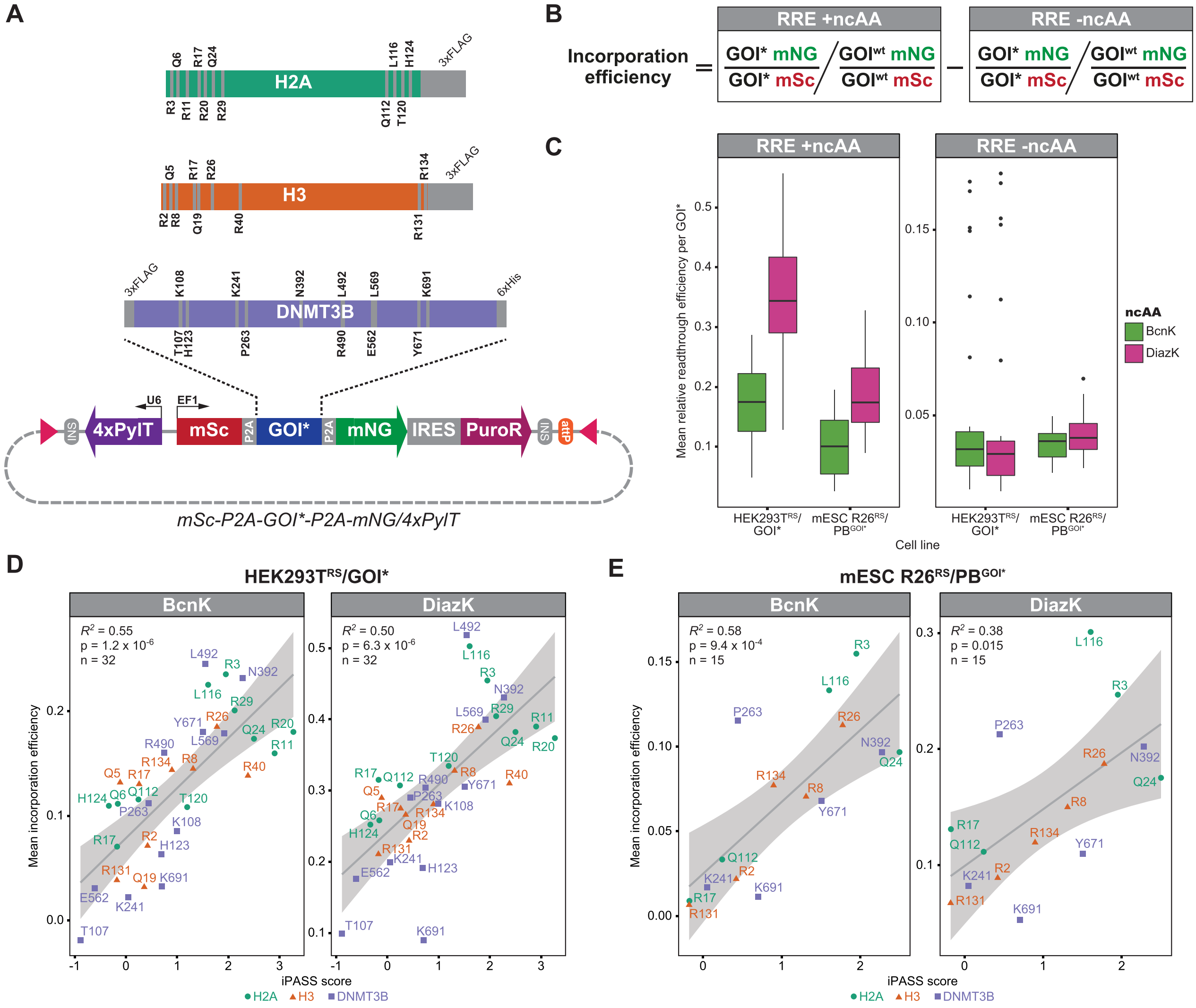


Figure 3. (**A**) Dual-fluorescence reporter construct *mSc-P2A-GOI*-P2A-mNG/4xPylT* to assess ncAA incorporation efficiencies across selected positions within target proteins by flow-cytometry. Expression of the gene of interest with an in-frame amber stop codon (*GOI**) is linked to mScarlet (mSc) and mNeonGreen (mNG) expression levels via the self-cleaving peptide 2A (P2A). Amber mutants of the *Mus musculus* histone H2A/H3 with a C-terminal 3xFLAG-tag or the *de novo* DNA methyltransferase 3b (DNMT3B) with an N-terminal 3xFLAG-tag and C-terminal 6xHis-tag were integrated as *GOI**. Analyzed amber stop codon positions within each *GOI** are indicated. The vector harbors the *attP* attachment site as well as inverted terminal repeats (ITRs, indicated as rectangles) for Bxb1 or PiggyBac (PB) mediated stable integration. Abbreviations: *Methanosarcina mazei* tRNA$^{\text{Pyl}}$$_{\text{CUA}}$ (PylT), internal ribosomal entry site (IRES), puromycin resistance (PuroR), constitutive EF1α promoter (EF1), constitutive U6 promoter (U6), insulator (INS). (**B**) Formula to calculate the incorporation efficiency of the respective ncAA at each amber stop codon position (*GOI**) relative to the wild-type codon (*GOI^{wt}*). For each construct, the mean fluorescence intensity (MFI, see also Supplementary Figure S7C and S8C for representative flow-cytometry data) of mNG (*GOI** or *GOI^{wt}* mNG) is normalized to the respective MFI of mSc (*GOI** or *GOI^{wt}* mSc). The relative readthrough efficiency (RRE) in the absence of an ncAA (-ncAA) is subtracted from the RRE +ncAA to account for basal translational readthrough over the stop codon that results in full-length peptides lacking the respective ncAA. (**C**) RREs are in total higher in stable HEK293T$^{\text{RS}}$ cells and with DiazK. Mean RRE ($n = 3$ biological replicates) at each analyzed *GOI** site for DiazK and BcnK (+ncAA) or the -ncAA control was calculated for HEK293T cells stably expressing the respective PylRS variant and transiently transfected with the mSc/mNG fluorescent reporter (HEK293T$^{\text{RS}}$/GOI*, $n = 32$) or mESCs stably expressing both PylRS variant and mSc/mNG fluorescent reporter construct (mESC R26$^{\text{RS}}$/PB$^{\text{GOI*}}$, $n = 15$). Per replicate, mNG and mSc MFIs from ca. 10 000 mSc positive single cells (mSc positive single cell counts are listed in Supplementary Data 2) were acquired by flow-cytometry 24 h after addition of 0.5 mM ncAA (for gating strategy and representative flow-cytometry data see Supplementary Figure S7 and S8). Horizontal black lines within boxes represent median values, boxes indicate the lower and upper quartiles, and whiskers indicate the 1.5 interquartile range. (**D, E**) iPASS reliably predicts relative ncAA incorporation efficiencies at *GOI** mutants in mammalian cells. iPASS scores of each *GOI** sequence context were correlated with experimentally determined mean incorporation efficiencies ($n = 3$ biological replicates, see also Supplementary Figure S9) of DiazK and BcnK using the *mSc-P2A-GOI*-P2A-mNG/4xPylT* fluorescent reporter in HEK293T$^{\text{RS}}$/GOI* (D) or mESC R26$^{\text{RS}}$/PB$^{\text{GOI*}}$ (E) lines. Coefficient of determination (R^2), *P*-value (*P*), and number (*n*) of analyzed mSc/mNG fluorescent reporters harboring different *GOI** are indicated. The 95% confidence interval of the regression line is marked.

ing the initial start codon. Hence, translational initiation at secondary start codons downstream of an in-frame UAG should be reduced. This decrease in the aberrant expression of N-terminally truncated POIs might improve yields of POI^{ncAA} that harbor ncAA incorporation sites close to the N-terminus.

We placed amber stop codons into the coding sequences of *H2A*, *H3*, and *Dnmt3b* so that they could also be used for ncAA-mediated crosslinking to identify interaction partners. For H2A and H3, we selected positions within the N- and C-terminal tails of each histone that are in close proximity to post-translationally modified lysine residues, like H3K9, H3K27 or H2AK119 (109). For *Dnmt3b*, we selected weakly conserved positions in the N-terminus and highly conserved residues within interaction surfaces (Supplementary Figure S6). To assess amber suppression at each position, mSc/mNG fluorescent reporter constructs harboring these *GOI** amber mutants (*H2A**, *H3**, or *Dnmt3b**) (Figure 3A) were transiently transfected into HEK293T^{RS} cell lines (denoted as HEK293T^{RS}/GOI*) or stably integrated into R26^{RS} mESC lines by PB-mediated transposition (denoted as R26^{RS}/PB^{GOI*}). Cell lines were subsequently cultured in the presence of either 0.5 mM BcnK or DiazK to also assess the incorporation efficiency of an ncAA with a chemical moiety distinct from BcnK. After 24 h, mSc and mNG fluorescence intensities were recorded by flow-cytometry in biological triplicates (Supplementary Figure S7 and S8). Using mSc and mNG mean fluorescence intensities, relative readthrough efficiencies (RREs) (72–74) were calculated for each position in the presence or absence of an ncAA by normalization to the respective wild-type *GOI* (*GOI^{wt}*) expression levels (Figure 3B). DiazK and BcnK were better incorporated in HEK293T^{RS}/GOI* compared to mESC R26^{RS}/PB^{GOI*}. In both cell lines, we found that DiazK yielded generally higher RREs than BcnK (Figure 3C), in line with the differences observed after transient transfection of the 4x*PylT*/*mCh-TAG-EGFP* reporter in R26^{RS} (Figure 1C). As readthrough events caused by near-cognate tRNAs might obscure the true efficiency of ncAA incorporation, we sought to account for this by calculating a corrected incorporation efficiency at each position, where the RRE –ncAA is subtracted from the RRE +ncAA (Figure 3B). Single incorporation efficiency measurements were highly reproducible between biological triplicates, confirming the stringency of the mSc/mNG fluorescent reporter assay (Supplementary Figure S9A+B). Within *H2A**, *H3**, and *Dnmt3b** amber mutants, measured incorporation efficiencies maximally varied between 4.2- and 33-fold for BcnK and 2.4- and 11-fold for DiazK (Supplementary Figure S9C). Additionally, these variations seemed to be independent of the distance between UAG position and PolyA tail. We observed incorporation efficiencies to differ more than 2-fold between proximal (e.g. *H2A^{Q112*}* and *H2A^{L116*}*) and even adjacent (e.g. *Dnmt3b^{T107*}* and *Dnmt3b^{K108*}*) UAG positions, highlighting the distinct influence of the immediate sequence context on amber suppression by PylT. The varying incorporation efficiencies at selected positions as well as cleavage of 2A peptides were also verified by western blot (Supplementary Figure S10A). Although P2A peptides have been reported to be efficiently cleaved (105), we detected a small fraction of uncleaved fusion proteins for all tested constructs. Furthermore, overall expression of N-terminal mSc in the absence of an ncAA was generally lower in amber mutants compared to wt coding sequences in both HEK293T^{RS}/GOI* and mESC R26^{RS}/PB^{GOI*} (Supplementary Figure S10B). This reduction in mSc levels might be due to exon-junction complex-independent nonsense-mediated decay (NMD) of these intron-free, nonsense transcripts (110). Conversely, after ncAA addition, amber mutant constructs exhibited up to 2-fold higher mSc levels compared to their respective *GOI^{wt}* constructs (Supplementary Figure S10B). These increases in mSc were linearly correlated with changes in mNG intensity and, as such, suppression of the amber stop codon (Supplementary Figure S10C), suggesting that PylT-mediated suppression of in-frame UAG and incorporation of ncAAs stabilize nonsense transcripts and/or enhance translational efficiency in mammalian cells. While the underlying molecular mechanisms remain unclear, this unexpected observation highlights the methodological advantage of dual- over C-terminal single-fluorescence reporters to accurately assess relative incorporation efficiencies by accounting for varying reporter construct expression levels between different conditions. Taken together, we demonstrate that our mSc/mNG fluorescent reporter used in conjunction with flow-cytometry analysis offers a rapid and reliable means for the high-throughput characterization of ncAA incorporation efficiency at different sites within a *GOI**.

Validation of the iPASS model with experimentally determined ncAA incorporation efficiencies

To verify the iPASS model, we directly compared the predicted iPASS score of amber stop codon contexts for each *GOI** with their experimental ncAA incorporation efficiencies measured with the mSc/mNG fluorescent reporter in living cells. For both HEK293T^{RS}/GOI* and mESC R26^{RS}/PB^{GOI*}, predicted and measured incorporation efficiencies are linearly correlated, confirming the accuracy and reliability of iPASS in identifying permissive ncAA incorporation sites based on UAG sequence contexts (Figure 3D and E). In general, low-scoring (e.g. *H2A^{R17*}*, *H3^{R131*}*, *Dnmt3b^{K241*}*) and high-scoring (e.g. *H2A^{R3*}*, *H3^{R26*}*, *Dnmt3b^{N392*}*) *GOI** contexts were also associated with relatively low or high ncAA incorporation efficiencies, respectively (Figure 3D and E). Of note, some *GOI** contexts with similar iPASS scores still varied several fold in their incorporation efficiencies (e.g. *Dnmt3b^{K691*}* and *Dnmt3b^{L492*}* in HEK293T^{RS}; Figure 3D, Supplementary Figure S9A), indicating the existence of additional factors influencing amber suppression that are not accounted for in the iPASS model. By randomly re-assigning fold changes determined by SORT-E to all detected sequence contexts, we also computed a decoy linear regression model. The decoy model fails to predict measured ncAA incorporation efficiencies in HEK293T and mESC lines (Supplementary Figure S11A and B). This negative control further highlights the specificity of iPASS to predict *bona fide* incorporation efficiencies according to favorable UAG contexts. Importantly, the capacity of iPASS to identify permissive ncAA sites for both BcnK and DiazK incorporation in mESC as well as HEK293T cell lines indicates that relative amber suppression with PylT is predominantly influenced by mRNA context rather than ncAA or cell line identity. Accordingly, incorporation efficiencies measured with the mSc/mNG fluorescent reporter were also linearly correlated between mESCs R26^{RS} and HEK293T^{RS} as well as between BcnK and DiazK (Supplementary Figure S11C and D). Hence, once identified, positions with high incorporation efficiencies can be readily modified with diverse ncAAs across different mammalian cell types. In conclusion, the iPASS model reliably predicts relative ncAA incorporation efficiencies at different UAG sequence contexts in mammalian cell lines expressing the *PylS/PylT* pair.

iPASS guided optimization of ncAA incorporation efficiencies by silent mutation of flanking codons

After generally assessing the predictive power of iPASS, we next wondered whether ncAA incorporation efficiencies can be improved by silently mutating the codons flanking the amber stop codon. To this end, we replaced the C-terminal 3xFLAG-tag of *H2A*^{wt} within the mSc/mNG dual-fluorescence reporter (Figure 3A) with selected nucleotide contexts spanning the 6 bp up- and downstream of the amber stop codon (*contexts**) (Figure 4A). As before, these *mSc-P2A-context*-P2A-mNG/4xPylT* reporters were transiently transfected into HEK293T^{RS} (denoted as HEK293T^{RS}/context*) or stably integrated into R26^{RS} mESCs by PB-mediated transposition (denoted as R26^{RS}/PB^{context*}) to identify *context** dependent variations in BcnK or DiazK incorporation efficiencies by flow-cytometry analysis.

We first asked whether the nucleotides preceding or following the amber stop codon determine incorporation efficiency. Within the ribosome, the codons preceding the amber stop codon have already been decoded into amino acids or are bound in the ribosomal P-site by a peptidyl-tRNA, whereas downstream codons have yet to be translated. This biomechanical difference between up- and downstream contexts might differentially affect PylT decoding at amber stop codons. However, the iPASS model suggests a synergistic influence of the surrounding sequence context on ncAA incorporation efficiency (Figure 2C). To better understand the relative importance of up- versus downstream nucleotides, we applied iPASS to extract the *context** with the lowest or highest iPASS score. Additionally, we exchanged either the up- or downstream sequence with the lowest iPASS score for the sequence with the highest iPASS score, thereby constructing two chimeric *contexts** with intermediate iPASS scores (Figure 4B). Compared to the *context** with the lowest iPASS score, incorporation efficiencies significantly increased for both chimeric *contexts**. In particular, replacing the preceding nucleotides improved incorporation efficiencies to a greater extent than the nucleotides following the amber stop codon. However, fully replacing the low-score with the high-score *context** increased incorporation efficiencies even further, up to 4.8- or 11.5-fold in HEK293T^{RS} or mESCs R26^{RS}, respectively. Although the preceding nucleotides may have a greater impact, these results confirm that both up- and downstream nucleotides synergistically influence ncAA incorporation efficiency.

We then tested whether iPASS can be applied to optimize amber suppression within a given amino acid sequence by silently mutating the two codons flanking the stop codon. This strategy would be useful in amber suppression applications where the ncAA incorporation site is usually fixed, such as the installation of post-translationally modified ncAAs or the incorporation of ncAAs to probe enzyme active sites. For this, we analyzed *context** pairs displaying a minimal difference of 2.4 in iPASS score after iPASS guided synonymous codon exchange. BcnK as well as DiazK incorporation significantly increased in 10 out of 13 iPASS optimized *contexts** in HEK293T^{RS}/context* cells, ranging from 1.4- up to 5.3-fold (Figure 4C). Furthermore, fold changes of five out of six selected *context** pairs in stable mESC R26^{RS}/PB^{context*} cell lines were generally higher than in HEK293T^{RS} cells, ranging between 5.0- and 22.9-fold (Figure 4D). Notably, even though iPASS optimization did not improve incorporation efficiencies for all *contexts**, we did not detect a reduced ncAA incorporation efficiency after iPASS guided optimization. Consistent with reports that +4 C permits above-average amber suppression in mammalian cells (63,67,68), we observed in HEK293T^{RS} and mESCs R26^{RS} the highest fold changes for iPASS optimized *contexts** in which +4 A or +4 U was exchanged with +4 C. Accordingly, the overall efficiency of DiazK incorporation into all *GOI** (Supplementary Figure S9A+B) and *context** (Figure 4B-D, Supplementary Figure S12A) mutants analyzed was significantly higher for +4 C than for the remaining three +4 nucleotides (Supplementary Figure S12B). At the same time, iPASS also successfully optimized *contexts** without altering the +4 base, further confirming that the influence of nucleotides on amber suppression extends over the +4 position. Taken together, we demonstrate that an approximate iPASS score difference of 2.5–3.0 after synonymous exchange of codons flanking the amber stop codon usually increases ncAA incorporation efficiencies.

To further validate the iPASS model, we compared the experimentally determined incorporation efficiencies of different *contexts** with their respective iPASS scores, revealing a maximal coefficient of determination (R^2) of 0.49 and 0.56 in HEK293T^{RS} and mESCs R26^{RS} lines, respectively (Figure 4E, Supplementary Figure S12C). Interestingly, also translational readthrough at *contexts** by near-cognate tRNAs in the absence of ncAAs was linearly correlated with suppression by PylT upon ncAA addition, although with in total lower R^2 values compared to iPASS, ranging from 0.24 to 0.43 (Supplementary Figure S12D+E). This correlation argues for overall similar context effects in amber suppression by PylT and near-cognate tRNAs. Additionally, incorporation efficiencies of both ncAAs as well as both analyzed cell lines were linearly correlated, further confirming that amber suppression efficiencies are largely independent of ncAA and cell line identity (Supplementary Figure S12F and G). Across all *GOI** and *contexts** analyzed with the mSc/mNG fluorescent reporter, an iPASS score >1.0 (approximate mean of iPASS score ranging from –2.3 to 3.7) is generally associated with significantly increased incorporation efficiencies (Figure 4F). Hence, an iPASS score cut-off of 1.0 should be applied when screening for permissive amber suppression sites in a target open

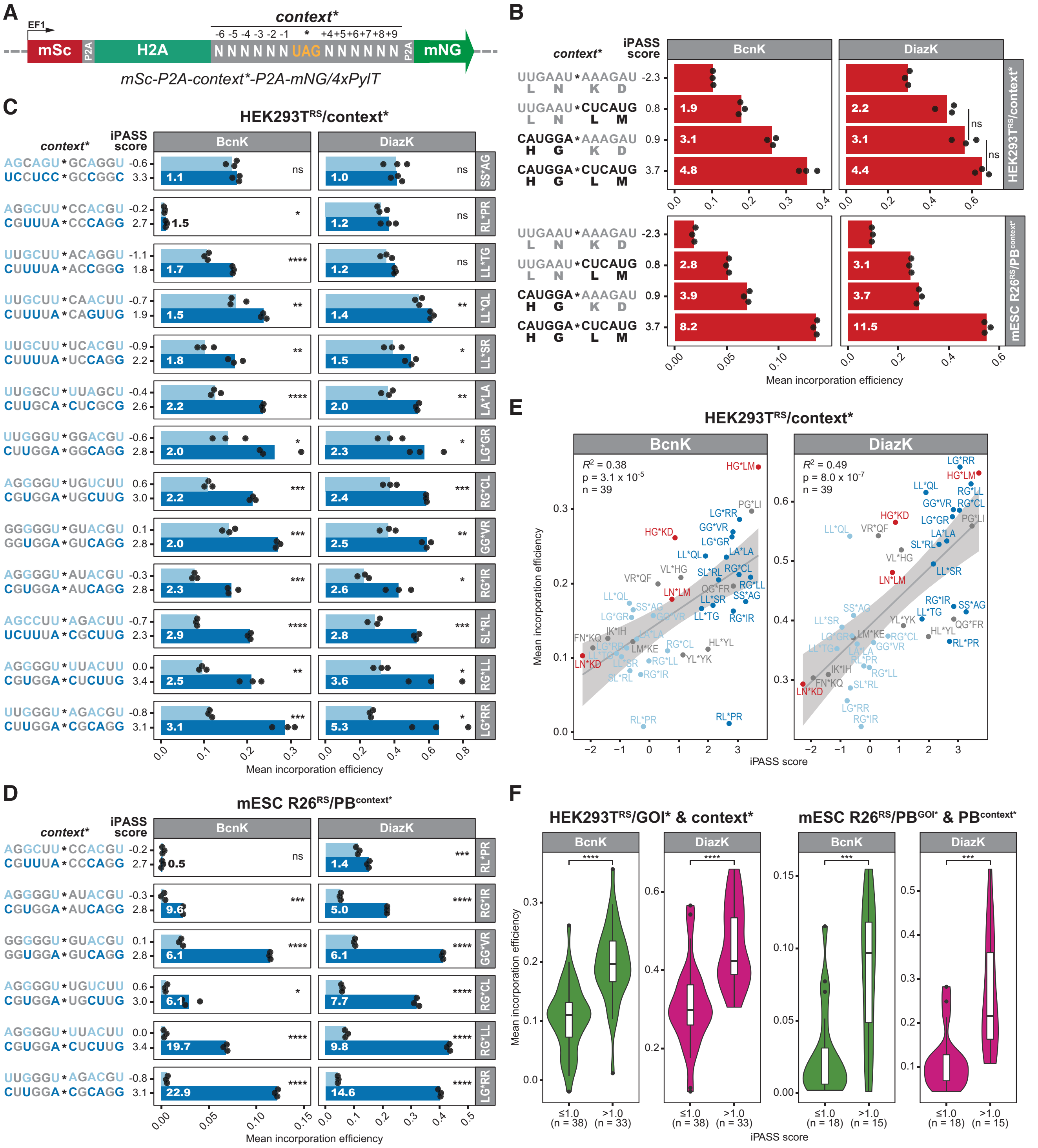


Figure 4. (**A**) Segment of mSc/mNG dual-fluorescence reporter construct *mSc-P2A-context*-P2A-mNG/4xPylT* (see also Figure 3A) to assess ncAA incorporation efficiencies at amber stop codons within selected sequence contexts (*context**) by flow-cytometry. The N-terminal 3xFLAG-tag of *H2A*^{wt} within the mSc/mNG fluorescent reporter is replaced with *context** composed of selected nucleotides (N) 6 bp up- and downstream of the amber stop codon (nucleotides –6 to +9; stop codon at +1, +2, +3). Incorporation efficiencies are calculated according to Figure 3B with *context** and *context*^{wt} replacing *GOI** and *GOI*^{wt}, respectively. In *context*^{wt} constructs the amber stop codon (UAG) is changed to the lysine codon (AAG) (not shown). Abbreviations: self-cleaving peptide 2A (P2A), constitutive EF1α promoter (EF1). (**B–E**) Mean incorporation efficiencies ($n = 3$ biological replicates) of DiazK or BcnK at *context** were calculated using the *mSc-P2A-context*-P2A-mNG/4xPylT* fluorescent reporter in HEK293T cells stably expressing the respective PylRS variant and transiently transfected with the fluorescent reporter (HEK293T^{RS}/context*) or mESCs stably expressing both PylRS variant and fluorescent reporter construct (mESC R26^{RS}/PB^{context*}). Per replicate, mNG and mSc mean fluorescence intensities from 5000 to 10 000 mSc positive single cells (mSc

reading frame. In summary, iPASS not only reliably aids the identification of permissive amber suppression sites, but also guides the silent mutation of flanking codons to optimize ncAA incorporation at a selected site.

DISCUSSION

Readthrough of amber stop codons in mammalian cells by near-cognate tRNAs is governed by the identity of flanking nucleotides (51,54,55,61,63,66) with a limited dataset indicating similar but not identical context effects in the decoding capacity of an amber suppressor tRNA (63). Depending on the location of the amber stop codon within a given sequence, these context effects lead to highly variable ncAA incorporation rates. Here, we have not only investigated to which extent UAG contexts influence ncAA incorporation rates but also provide a streamlined workflow that combines analysis *in silico* and in living mammalian cells to quickly and reliably identify permissive ncAA incorporation sites.

We combine two previously described strategies (30,76) to establish a vector system that is compatible with both site-specific recombination by Bxb1 as well as PB transposition. We then use this system to expand the genetic code of both murine and human cell lines. In particular, we leverage Bxb1-mediated recombination to integrate the *PylS*/*PylT* pair at the genomic safe harbor R26 in mESCs and subsequently apply PB transposition to genomically integrate multiple copies of the *GOI**. This stepwise approach has the advantage that the tRNA synthetase expression level should be lower in comparison to *GOI** and *PylT* expression levels, which has been reported to enhance amber suppression efficiency in mammalian cells (27). Collectively, we demonstrate that with this strategy stable and defined mESC clones that efficiently suppress amber stop codons can be established. Of note, in contrast to a recent report applying amber suppression with BocK (111), we observed no directed cellular response to amber suppression with BcnK. This discrepancy might be due to more stringent filtering of our full proteome data and the lower incorporation efficiency of BcnK compared to BocK.

To date, application of dual-fluorescence reporters has been limited to comparisons of amber suppression efficiencies among different OTSs and hosts (72–74) but not *GOI** mutants. Reporters such as *mCherry-TAG-EGFP* are useful to assess amber suppression efficiencies of newly developed synthetases and ncAAs. However, adaptation and application of these reporters in mammalian cells to quickly compare ncAA incorporation rates among defined sites within a target protein has not been explored. Our mSc/mNG fluorescent reporter bearing a P2A flanked *GOI** in combination with high-throughput analysis by flow-cytometry reproducibly yields these context-specific ncAA incorporation efficiencies. For all *GOI** and both ncAAs tested, the reporter detects several fold differences in ncAA incorporation efficiency even between close-by amber sites. Importantly, turnover rates of fluorescent proteins are uncoupled via self-cleaving 2A peptides to selectively evaluate ncAA incorporation independently of POI^{ncAA} stability. Furthermore, this vector system can be readily used to generate mammalian cell lines stably expressing a tag-free POI^{ncAA} for downstream applications. Taken together, the sensitive and consistent quantification of ncAA incorporation efficiencies renders the mSc/mNG fluorescent reporter a valuable tool to rapidly identify highly permissive ncAA incorporation sites.

To date, optimization of genetic code expansion in mammalian cells has been mostly focused on intrinsic properties of the OTS itself, for instance by engineering OTS components or tuning their expression levels. UAG context-dependent variations in ncAA incorporation efficiencies, on the other hand, require time-consuming cloning and screening of multiple amber mutants for each individual POI^{ncAA}. We reasoned that quantifying ncAA incorporation rates at endogenous amber stop codons might provide generalizable insights into the relationship between UAG flanking sequences and ncAA incorporation efficiencies. By adapting SORT-E (75) with BcnK and Biotin-Tet to amber codons, we endogenously probe BcnK incorporation at hundreds of potential sequence contexts in mESCs and HEK293T cells. As our regression model encompasses the six nucleotides up- and downstream of significantly enriched amber stop codon contexts, the potential contribution of extended RNA secondary structures on ncAA incorporation efficiency is omitted. However, both regression models calculated for HEK293T or mESCs reciprocally correlate with experimental mESC or HEK293T SORT-E data, justifying this focus on close-by nucleotides. Furthermore, this restric-

positive single cell counts are listed in Supplementary Data 2) were acquired by flow-cytometry 24 h after addition of 0.5 mM ncAA. (B–D) According to Smith and Yarus (114), the fold change between incorporation efficiencies (IEs) at two different amber stop codon sequence contexts is calculated as $\{\text{IE(max)} \times [1\text{-IE(min)}]\}/\{\text{IE(min)} \times [1 - \text{IE(max)}]\}$. For each *context** the nucleotide sequence ±6 bp flanking the amber stop codon (*) and its respective iPASS score are presented. Two-tailed unpaired two-sample Student's *t*-test: ns not significant, $^{*}P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$, $^{****}P < 0.0001$. (**B**) Incorporation efficiency is synergistically influenced by the six nucleotides up- as well as downstream of the amber stop codon and varies several fold between *context** with the lowest and highest iPASS score. White number within bars indicates fold change compared to the LN*KD context with the lowest iPASS score. Two-tailed unpaired two-sample Student's *t*-test: ns not significant, all other *context** at least $P < 0.01$ (not indicated). (**C, D**) The iPASS tool guides silent mutation of *contexts** improving incorporation efficiencies several fold in $\text{HEK293T}^{\text{RS}}$/context* (C) or mESC $\text{R26}^{\text{RS}}/\text{PB}^{\text{context*}}$ (D) lines. White number within dark blue bars (optimized *context**) indicates fold change compared to the same amino acid context with a lower iPASS score (light blue bars). (**E**) iPASS reliably predicts relative ncAA incorporation efficiencies at *context** mutants in mammalian cells. iPASS scores of each *context** target site were correlated with experimentally determined mean incorporation efficiencies of DiazK or BcnK in $\text{HEK293T}^{\text{RS}}$/context* cell lines. Coefficient of determination (R^2), *P*-value (*P*), and number (*n*) of analyzed mSc/mNG fluorescent reporters harboring different *context** are indicated. The 95% confidence interval of the regression line is marked. Color coding according to (B)–(D) (gray: additional *context** from Supplementary Figure S12A). (**F**) An iPASS score >1.0 generally indicates higher incorporation efficiencies compared to an iPASS score ≤1.0 (Mean[iPASS score] ≈ 1.0). Mean incorporation efficiencies at each analyzed *GOI** and *context** site for DiazK and BcnK in $\text{HEK293T}^{\text{RS}}$ or mESC R26^{RS} are grouped according to their iPASS score. Total number (*n*) of analyzed mSc/mNG fluorescent reporters in each group are indicated. Inside violin plots horizontal black lines within boxes represent median values, boxes indicate the lower and upper quartiles, and whiskers indicate the 1.5 interquartile range. Two-tailed unpaired two-sample Student's *t*-test: $^{***}P < 0.001$, $^{****}P < 0.0001$.

tion to the proximal nucleotides widely facilitates the selective modification of UAG contexts to adopt high ncAA incorporation efficiencies. By combining the HEK293T and mESC regression model, we formulate iPASS to predict relative ncAA incorporation efficiencies at amber stop codons based on the surrounding nucleotide context. Probability logo representation of all possible sequence contexts weighted by their iPASS score indicates that nucleotides up- as well as downstream of the stop codon affect ncAA incorporation. Removing nucleotide positions within iPASS gradually reduced the accuracy of relative ncAA incorporation efficiency prediction, which excluded potential overfitting of the iPASS model. By separately optimizing up- and downstream UAG contexts using iPASS and measuring their suppression efficiencies with the mSc/mNG fluorescent reporter, we experimentally confirm that nucleotides on both sides of UAG govern the efficiency of amber suppression. Thus, the identity of nucleotides flanking the amber stop codon synergistically influences ncAA incorporation efficiencies in mammalian cells.

It is tempting to speculate that UAG contexts with high translational readthrough also boost ncAA incorporation rates. However, the iPASS probability logo, despite encompassing the readthrough promoting feature +4 C, widely differs at the majority of nucleotide positions from similar linear regression analyses of readthrough motifs in human cells (61,66). Additionally, these previously reported stop codon contexts are enriched for the opal codon and hence motifs that might not promote translational readthrough in the context of amber stop codons. The iPASS motif also differs from bacterial UAG contexts with high amber suppression efficiencies (33), confirming distinct context effects between pro- and eukaryotes. An established amber stop codon context that confers strong translational readthrough *in vitro* and *in vivo* in eukaryotes is the consensus sequence +1 UAG CAR YYA (48–51,53,55,64), a readthrough motif which has been originally identified in the tobacco mosaic virus (TMV) as +1 UAG CAA UUA (112). However, the iPASS motif, although enriched for +4 C and +7 UUA, is depleted of +5 A and especially +6 R. Interestingly, translational readthrough at UAG contexts in the absence of an ncAA was yet roughly correlated with their suppression efficiency. Consistent with a previous report (63), this result indicates that amber suppression by PylT and translational readthrough by near-cognate tRNAs are influenced by similar but not identical flanking sequence preferences. Hence, the relative ncAA incorporation efficiency at target sites might benefit from sequence features permitting strong translational readthrough. Importantly, the identity of eukaryotic near-cognate tRNAs and their preferences for specific stop tetranucleotides have been reported to be interdependent (62,65). This observation implies that context-specific differences in the capacity of PylT to decode in-frame UAGs might be at least partially attributable to intrinsic PylT properties. Therefore, stop codon contexts would have to be specifically adapted to promote either readthrough by near-cognate tRNAs or suppression by PylT. Accordingly, iPASS might not reliably predict ncAA incorporation efficiencies at amber stop codons by OTSs other than the *PylS*/*PylT* pair in mammalian cells. However, our strategy of regression analysis after SORT-E together with the fluorescent reporter assay can easily be expanded for other OTSs as well as to opal, ochre, or quadruplet codons to identify favorable sequence contexts for efficient decoding.

Using incorporation efficiencies measured with the mSc/mNG fluorescent reporter, we validate the iPASS model to predict relative ncAA incorporation rates *in silico* depending on the UAG context. In particular, we experimentally determined ncAA incorporation efficiencies across multiple sites within a *GOI** as well as at a fixed *context** site with varying nucleotide compositions. In both experimental setups, analysis of UAG contexts with iPASS accurately reveals approximately half of the variation in amber suppression efficiency by PylT in mammalian cells, suggesting that efficient suppression is largely governed by sequence context. However, we detected some positions at which iPASS did not accurately predict ncAA incorporation efficiencies. Hence, additional factors not covered by iPASS might influence context-specific efficiency of amber suppression by PylT, such as mRNA abundance, translational speed and pausing, or structural and more distant sequence features surrounding UAG. For instance, besides close-by nucleotides also RNA secondary structures (64) and sequence features more than 6 bp downstream of an in-frame stop codon (60) were reported to promote eukaryotic translational readthrough. In general, however, UAG contexts with an iPASS score >1.0 confer significantly higher ncAA incorporation rates than contexts with a lower iPASS score. This iPASS guided pre-selection considerably reduces the number of amber mutants that must be cloned and screened to identify permissive ncAA incorporation sites.

Moreover, iPASS can be used to reliably optimize amber suppression efficiencies by synonymous codon exchange. In general, only three amino acids (Arg, Leu, Ser) allow for synonymous codon exchange by modifying the first base and only one (Ser) by modifying the second base. This inherent limitation in synonymous codons available for exchange might restrict the capacity of iPASS to improve amber suppression efficiencies at some fixed sites. However, in 6 out of 13 *contexts** tested in HEK293T^{RS} cells, optimization by iPASS resulted in significantly improved ncAA incorporation efficiencies without modifying the +4 or +5 base, further highlighting that ncAA incorporation efficiency is influenced by nucleotides beyond these positions. Notably, iPASS also optimized incorporation efficiencies in three out of five *contexts** without altering +4 purines, which are thought to stabilize formation of the termination complex (101,102) and are generally associated with reduced translational readthrough (63,66) and suppression (63,67) of an in-frame UAG. Here, we demonstrate that an iPASS score difference >2.5 after synonymous codon exchange generally results in an up to several fold increase in ncAA incorporation efficiency.

Lastly, both measured and predicted relative incorporation efficiencies seem to be independent of cell line and ncAA identity. This finding is consistent with the hypothesis that context effects in amber suppression depend on fundamental properties of mammalian translation (67). Additionally, in *Escherichia coli* the relative ncAA incorporation rate at defined mRNA contexts has been reported to be generally independent of ncAA size and chemical reactivity (31,33,113). Therefore, predicted incorporation efficiencies can be first validated using standard ncAAs and cell lines before continuing with specialized ncAAs and sophisticated cell lines like mESCs.

To the best of our knowledge, the iPASS regression model provides the first characterization of flanking sequences that permit high amber suppression and hence ncAA incorporation rates in mammalian cells. In combination with the mSc/mNG dual-fluorescence reporter, our pipeline streamlines the identification of permissive ncAA incorporation sites. This will greatly facilitate ncAA-based crosslinking and labeling experiments, enabling researchers to select the most efficient sites while reducing the total number of amber mutants that have to be cloned and tested. To assist with the preselection of ncAA incorporation sites in mammalian cell lines expressing the *PylS*/*PylT* pair, we developed the iPASS web-tool that can be accessed at www.bultmannlab.eu/tools/iPASS. The tool additionally guides the design of silent mutations of the nucleotide positions flanking the amber stop codon to improve ncAA incorporation rates. This functionality will be very useful in applications where the incorporation site is fixed, such as the selective installation of ncAAs mimicking post-translational modifications.

DATA AVAILABILITY

Plasmids cloned in this study have been deposited at Addgene with the IDs 167491–99. Mass spectrometry proteomics data are available via ProteomeXchange with the dataset identifier PXD019815. Flow-cytometry raw data are available via FlowRepository with the dataset identifier FR-FCM-Z2N3. Significantly enriched proteins in the streptavidin pulldown, their relative expression level according to LC–MS/MS full proteome data, encoded stop codon, and molecular weight are available in Supplementary_data_1.xlsx. Data from flow-cytometry of the mSc/mNG dual-fluorescence reporter constructs and respective analysis by iPASS (nucleotide context, iPASS score, mSc positive single cell count, mean fluorescence intensity, (mean) relative readthrough efficiency, (mean) incorporation efficiency) are available in Supplementary_data_2.xlsx. Significantly changed proteins in full proteomes used for GO enrichment analysis are available in Supplementary_data_3.csv. Data to compute iPASS (mean LFC, *P*-value, sequence context) are available in Supplementary_data_4.csv.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank Dr Heinrich Leonhardt for support, advice, and guidance. We thank Jeannette Koch and Ruzica Barisic for excellent technical assistance. We thank Dr Neda Tafrishi and Joshua Kie (Gene Center Flow Cytometry Core Facility, LMU Munich, Germany) for technical support with flow-cytometry. We thank Dr Matthias Mann (Department of Proteomics and Signal Transduction, Max Planck Institute of Biochemistry, Germany) for providing instrument time at mass spectrometers and Dr Igor Paron and Dr Christian Deiml for great technical support. We would like to thank Christopher B. Mulholland (MFMF) for helpful discussions and constructive criticism on the manuscript. M.D.B. and E.U. gratefully acknowledge the International Max Planck Research School for Molecular Life Sciences (IMPRS-LS) and E.U. the research training group 1721 (RTG 1721) for training and support.

Author contributions: M.D.B. and S.B. designed and conceived the study. S.B. and K.L. supervised the study. M.D.B., E.U., K.L. and S.B. interpreted the data. M.D.B. performed molecular and cellular experiments. G.R. assisted with molecular cloning. M.D.B. generated stable cell lines with help from G.R.. M.D.B. performed flow-cytometry and analysis. M.D.B. and E.U. conducted proteomic experiments. E.U. performed mass spectrometry and analysis under the supervision of M.W.. S.B. performed linear regression analysis. T.A.N. performed chemical synthesis of DiazK and the Biotin-tetrazine conjugate. M.D.B. and S.B. wrote the manuscript and prepared figures. All authors read, discussed, and approved the manuscript.

FUNDING

Deutsche Forschungsgemeinschaft [283300243 to S.B., K.L., SPP1623 to S.B., K.L., 213249687 to S.B., SFB1064 to S.B.]. Funding for open access charge: Deutsche Forschungsgemeinschaft.

Conflict of interest statement. None declared.

REFERENCES

1. Liu,C.C. and Schultz,P.G. (2010) Adding new chemistries to the genetic code. *Annu. Rev. Biochem.*, **79**, 413–444.
2. Chin,J.W. (2017) Expanding and reprogramming the genetic code. *Nature*, **550**, 53–60.
3. Blight,S.K., Larue,R.C., Mahapatra,A., Longstaff,D.G., Chang,E., Zhao,G., Kang,P.T., Green-Church,K.B., Chan,M.K. and Krzycki,J.A. (2004) Direct charging of tRNA(CUA) with pyrrolysine in vitro and in vivo. *Nature*, **431**, 333–335.
4. Neumann,H., Peak-Chew,S.Y. and Chin,J.W. (2008) Genetically encoding N(epsilon)-acetyllysine in recombinant proteins. *Nat. Chem. Biol.*, **4**, 232–234.
5. Yanagisawa,T., Ishii,R., Fukunaga,R., Kobayashi,T., Sakamoto,K. and Yokoyama,S. (2008) Multistep engineering of pyrrolysyl-tRNA synthetase to genetically encode N(epsilon)-(o-azidobenzyloxycarbonyl) lysine for site-specific protein modification. *Chem. Biol.*, **15**, 1187–1197.
6. Hancock,S.M., Uprety,R., Deiters,A. and Chin,J.W. (2010) Expanding the genetic code of yeast for incorporation of diverse unnatural amino acids via a pyrrolysyl-tRNA synthetase/tRNA pair. *J. Am. Chem. Soc.*, **132**, 14819–14824.
7. Mukai,T., Kobayashi,T., Hino,N., Yanagisawa,T., Sakamoto,K. and Yokoyama,S. (2008) Adding l-lysine derivatives to the genetic code of mammalian cells with engineered pyrrolysyl-tRNA synthetases. *Biochem. Biophys. Res. Commun.*, **371**, 818–822.
8. Gautier,A., Nguyen,D.P., Lusic,H., An,W., Deiters,A. and Chin,J.W. (2010) Genetically encoded photocontrol of protein localization in mammalian cells. *J. Am. Chem. Soc.*, **132**, 4086–4088.
9. Chen,P.R., Groff,D., Guo,J., Ou,W., Cellitti,S., Geierstanger,B.H. and Schultz,P.G. (2009) A facile system for encoding unnatural amino acids in mammalian cells. *Angew. Chem. Int. Ed Engl.*, **48**, 4052–4055.
10. Greiss,S. and Chin,J.W. (2011) Expanding the genetic code of an animal. *J. Am. Chem. Soc.*, **133**, 14196–14199.
11. Bianco,A., Townsley,F.M., Greiss,S., Lang,K. and Chin,J.W. (2012) Expanding the genetic code of Drosophila melanogaster. *Nat. Chem. Biol.*, **8**, 748–750.

12. Liu,J., Hemphill,J., Samanta,S., Tsang,M. and Deiters,A. (2017) Genetic code expansion in Zebrafish embryos and its application to optical control of cell signaling. *J. Am. Chem. Soc.*, **139**, 9100–9103.
13. Han,S., Yang,A., Lee,S., Lee,H.-W., Park,C.B. and Park,H.-S. (2017) Expanding the genetic code of Mus musculus. *Nat. Commun.*, **8**, 14568.
14. Lang,K. and Chin,J.W. (2014) Cellular incorporation of unnatural amino acids and bioorthogonal labeling of proteins. *Chem. Rev.*, **114**, 4764–4806.
15. Coin,I. (2018) Application of non-canonical crosslinking amino acids to study protein-protein interactions in live cells. *Curr. Opin. Chem. Biol.*, **46**, 156–163.
16. Nguyen,T.-A., Cigler,M. and Lang,K. (2018) Expanding the genetic code to study Protein-Protein interactions. *Angew. Chem. Int. Ed Engl.*, **57**, 14350–14361.
17. Nödling,A.R., Spear,L.A., Williams,T.L., Luk,L.Y.P. and Tsai,Y.-H. (2019) Using genetically incorporated unnatural amino acids to control protein functions in mammalian cells. *Essays Biochem.*, **63**, 237–266.
18. Takimoto,J.K., Xiang,Z., Kang,J.-Y. and Wang,L. (2010) Esterification of an unnatural amino acid structurally deviating from canonical amino acids promotes its uptake and incorporation into proteins in mammalian cells. *ChemBioChem*, **11**, 2268–2272.
19. Nikić,I., Estrada Girona,G., Kang,J.H., Paci,G., Mikhaleva,S., Koehler,C., Shymanska,N.V., Ventura Santos,C., Spitz,D. and Lemke,E.A. (2016) Debugging eukaryotic genetic code expansion for site-specific click-PAINT super-resolution microscopy. *Angew. Chem. Int. Ed Engl.*, **55**, 16172–16176.
20. Reinkemeier,C.D., Girona,G.E. and Lemke,E.A. (2019) Designer membraneless organelles enable codon reassignment of selected mRNAs in eukaryotes. *Science*, **363**, eaaw2644.
21. Serfling,R., Lorenz,C., Etzel,M., Schicht,G., Böttke,T., Mörl,M. and Coin,I. (2018) Designer tRNAs for efficient incorporation of non-canonical amino acids by the pyrrolysine system in mammalian cells. *Nucleic Acids Res.*, **46**, 1–10.
22. Schmied,W.H., Elsässer,S.J., Uttamapinant,C. and Chin,J.W. (2014) Efficient multisite unnatural amino acid incorporation in mammalian cells via optimized pyrrolysyl tRNA synthetase/tRNA expression and engineered eRF1. *J. Am. Chem. Soc.*, **136**, 15577–15583.
23. Wang,W., Takimoto,J.K., Louie,G.V., Baiga,T.J., Noel,J.P., Lee,K.-F., Slesinger,P.A. and Wang,L. (2007) Genetically encoding unnatural amino acids for cellular and neuronal studies. *Nat. Neurosci.*, **10**, 1063–1072.
24. Chatterjee,A., Xiao,H., Bollong,M., Ai,H.-W. and Schultz,P.G. (2013) Efficient viral delivery system for unnatural amino acid mutagenesis in mammalian cells. *Proc. Natl. Acad. Sci. U.S.A.*, **110**, 11803–11808.
25. Uttamapinant,C., Howe,J.D., Lang,K., Beránek,V., Davis,L., Mahesh,M., Barry,N.P. and Chin,J.W. (2015) Genetic code expansion enables live-cell and super-resolution imaging of site-specifically labeled cellular proteins. *J. Am. Chem. Soc.*, **137**, 4602–4605.
26. Cohen,S. and Arbely,E. (2016) Single-plasmid-based system for efficient noncanonical amino acid mutagenesis in cultured mammalian cells. *Chem. Biol. Chem*, **17**, 1008–1011.
27. Zheng,Y., Lewis,T.L. Jr, Igo,P., Polleux,F. and Chatterjee,A. (2017) Virus-Enabled optimization and delivery of the genetic machinery for efficient unnatural amino acid mutagenesis in mammalian cells and tissues. *ACS Synth. Biol.*, **6**, 13–18.
28. Shen,B., Xiang,Z., Miller,B., Louie,G., Wang,W., Noel,J.P., Gage,F.H. and Wang,L. (2011) Genetically encoding unnatural amino acids in neural stem cells and optically reporting voltage-sensitive domain changes in differentiated neurons. *Stem Cells*, **29**, 1231–1240.
29. Tian,F., Lu,Y., Manibusan,A., Sellers,A., Tran,H., Sun,Y., Phuong,T., Barnett,R., Hehli,B., Song,F. *et al.* (2014) A general approach to site-specific antibody drug conjugates. *Proc. Natl. Acad. Sci. U.S.A.*, **111**, 1766–1771.
30. Elsässer,S.J., Ernst,R.J., Walker,O.S. and Chin,J.W. (2016) Genetic code expansion in stable cell lines enables encoded chromatin modification. *Nat. Methods*, **13**, 158–164.
31. Young,T.S., Ahmad,I., Yin,J.A. and Schultz,P.G. (2010) An enhanced system for unnatural amino acid mutagenesis in E. coli. *J. Mol. Biol.*, **395**, 361–374.
32. Liu,C.C., Cellitti,S.E., Geierstanger,B.H. and Schultz,P.G. (2009) Efficient expression of tyrosine-sulfated proteins in E. coli using an expanded genetic code. *Nat. Protoc.*, **4**, 1784–1789.
33. Pott,M., Schmidt,M.J. and Summerer,D. (2014) Evolved sequence contexts for highly efficient amber suppression with noncanonical amino acids. *ACS Chem. Biol.*, **9**, 2815–2822.
34. Xu,H., Wang,Y., Lu,J., Zhang,B., Zhang,Z., Si,L., Wu,L., Yao,T., Zhang,C., Xiao,S. *et al.* (2016) Re-exploration of the codon context effect on amber codon-guided incorporation of noncanonical amino acids in *Escherichia coli* by the blue-white screening assay. *ChemBioChem*, **17**, 1250–1256.
35. Sakin,V., Hanne,J., Dunder,J., Anders-Össwein,M., Laketa,V., Nikić,I., Kräusslich,H.-G., Lemke,E.A. and Müller,B. (2017) A versatile tool for live-cell imaging and super-resolution nanoscopy studies of HIV-1 Env distribution and mobility. *Cell Chem. Biol.*, **24**, 635–645.
36. Schvartz,T., Aloush,N., Goliand,I., Segal,I., Nachmias,D., Arbely,E. and Elia,N. (2017) Direct fluorescent-dye labeling of α-tubulin in mammalian cells for live cell and superresolution imaging. *Mol. Biol. Cell*, **28**, 2747–2756.
37. Schwark,D.G., Schmitt,M.A. and Fisk,J.D. (2018) Dissecting the contribution of release factor interactions to amber stop codon reassignment efficiencies of the *Methanocaldococcus jannaschii* orthogonal pair. *Genes*, **9**, 546.
38. Brown,C.M., Stockwell,P.A., Trotman,C.N. and Tate,W.P. (1990) Sequence analysis suggests that tetra-nucleotides signal the termination of protein synthesis in eukaryotes. *Nucleic Acids Res.*, **18**, 6339–6345.
39. Cavener,D.R. and Ray,S.C. (1991) Eukaryotic start and stop translation sites. *Nucleic Acids Res.*, **19**, 3185–3192.
40. Arkov,A.L., Korolev,S.V. and Kisselev,L.L. (1995) 5′ contexts of *Escherichia coli* and human termination codons are similar. *Nucleic Acids Res.*, **23**, 4712–4716.
41. Shabalina,S.A., Ogurtsov,A.Y., Rogozin,I.B., Koonin,E.V. and Lipman,D.J. (2004) Comparative analysis of orthologous eukaryotic mRNAs: potential hidden functional signals. *Nucleic Acids Res.*, **32**, 1774–1782.
42. Cridge,A.G., Major,L.L., Mahagaonkar,A.A., Poole,E.S., Isaksson,L.A. and Tate,W.P. (2006) Comparison of characteristics and function of translation termination signals between and within prokaryotic and eukaryotic organisms. *Nucleic Acids Res.*, **34**, 1959–1973.
43. McCaughan,K.K., Brown,C.M., Dalphin,M.E., Berry,M.J. and Tate,W.P. (1995) Translational termination efficiency in mammals is influenced by the base following the stop codon. *Proc. Natl. Acad. Sci. U.S.A.*, **92**, 5431–5435.
44. Kochetov,A.V., Ischenko,I.V., Vorobiev,D.G., Kel,A.E., Babenko,V.N., Kisselev,L.L. and Kolchanov,N.A. (1998) Eukaryotic mRNAs encoding abundant and scarce proteins are statistically dissimilar in many structural features. *FEBS Lett.*, **440**, 351–355.
45. Trotta,E. (2013) Selection on codon bias in yeast: a transcriptional hypothesis. *Nucleic Acids Res.*, **41**, 9382–9395.
46. Trotta,E. (2016) Selective forces and mutational biases drive stop codon usage in the human genome: a comparison with sense codon usage. *BMC Genomics*, **17**, 366.
47. Rodnina,M.V., Korniy,N., Klimova,M., Karki,P., Peng,B.-Z., Senyushkina,T., Belardinelli,R., Maracci,C., Wohlgemuth,I., Samatova,E. *et al.* (2020) Translational recoding: canonical translation mechanisms reinterpreted. *Nucleic Acids Res.*, **48**, 1056–1067.
48. Skuzeski,J.M., Nichols,L.M., Gesteland,R.F. and Atkins,J.F. (1991) The signal for a leaky UAG stop codon in several plant viruses includes the two downstream codons. *J. Mol. Biol.*, **218**, 365–373.
49. Valle,R.P., Drugeon,G., Devignes-Morch,M.D., Legocki,A.B. and Haenni,A.L. (1992) Codon context effect in virus translational readthrough. A study in vitro of the determinants of TMV and Mo-MuLV amber suppression. *FEBS Lett.*, **306**, 133–139.
50. Zerfass,K. and Beier,H. (1992) Pseudouridine in the anticodon G psi A of plant cytoplasmic tRNA(Tyr) is required for UAG and UAA suppression in the TMV-specific context. *Nucleic Acids Res.*, **20**, 5911–5918.

51. Stahl,G., Bidou,L., Rousset,J.P. and Cassan,M. (1995) Versatile vectors to study recoding: conservation of rules between yeast and mammalian cells. *Nucleic Acids Res.*, **23**, 1557–1560.
52. Bonetti,B., Fu,L., Moon,J. and Bedwell,D.M. (1995) The efficiency of translation termination is determined by a synergistic interplay between upstream and downstream sequences in *Saccharomyces cerevisiae*. *J. Mol. Biol.*, **251**, 334–345.
53. Namy,O., Hatin,I. and Rousset,J.P. (2001) Impact of the six nucleotides downstream of the stop codon on translation termination. *EMBO Rep.*, **2**, 787–793.
54. Cassan,M. and Rousset,J.P. (2001) UAG readthrough in mammalian cells: effect of upstream and downstream stop codon contexts reveal different signals. *BMC Mol. Biol.*, **2**, 3.
55. Harrell,L., Melcher,U. and Atkins,J.F. (2002) Predominance of six different hexanucleotide recoding signals 3′ of read-through stop codons. *Nucleic Acids Res.*, **30**, 2011–2017.
56. Tork,S., Hatin,I., Rousset,J.-P. and Fabret,C. (2004) The major 5′ determinant in stop codon read-through involves two adjacent adenines. *Nucleic Acids Res.*, **32**, 415–421.
57. Williams,I., Richardson,J., Starkey,A. and Stansfield,I. (2004) Genome-wide prediction of stop codon readthrough during translation in the yeast Saccharomyces cerevisiae. *Nucleic Acids Res.*, **32**, 6605–6616.
58. Pacho,F., Zambruno,G., Calabresi,V., Kiritsi,D. and Schneider,H. (2011) Efficiency of translation termination in humans is highly dependent upon nucleotides in the neighbourhood of a (premature) termination codon. *J. Med. Genet.*, **48**, 640–644.
59. Jungreis,I., Lin,M.F., Spokony,R., Chan,C.S., Negre,N., Victorsen,A., White,K.P. and Kellis,M. (2011) Evidence of abundant stop codon readthrough in Drosophila and other metazoa. *Genome Res.*, **21**, 2096–2113.
60. Loughran,G., Chou,M.-Y., Ivanov,I.P., Jungreis,I., Kellis,M., Kiran,A.M., Baranov,P.V. and Atkins,J.F. (2014) Evidence of efficient stop codon readthrough in four mammalian genes. *Nucleic Acids Res.*, **42**, 8928–8938.
61. Schueren,F., Lingner,T., George,R., Hofhuis,J., Dickel,C., Gärtner,J. and Thoms,S. (2014) Peroxisomal lactate dehydrogenase is generated by translational readthrough in mammals. *Elife*, **3**, e03640.
62. Beznosková,P., Gunišová,S. and Valášek,L.S. (2016) Rules of UGA-N decoding by near-cognate tRNAs and analysis of readthrough on short uORFs in yeast. *RNA*, **22**, 456–466.
63. Cridge,A.G., Crowe-McAuliffe,C., Mathew,S.F. and Tate,W.P. (2018) Eukaryotic translational termination efficiency is influenced by the 3′ nucleotides within the ribosomal mRNA channel. *Nucleic Acids Res.*, **46**, 1927–1944.
64. Anzalone,A.V., Zairis,S., Lin,A.J., Rabadan,R. and Cornish,V.W. (2019) Interrogation of eukaryotic stop codon readthrough signals by in vitro RNA selection. *Biochemistry*, **58**, 1167–1178.
65. Beznosková,P., Pavlíková,Z., Zeman,J., Echeverría Aitken,C. and Valášek,L.S. (2019) Yeast applied readthrough inducing system (YARIS): an invivo assay for the comprehensive study of translational readthrough. *Nucleic Acids Res.*, **47**, 6339–6350.
66. Wangen,J.R. and Green,R. (2020) Stop codon context influences genome-wide stimulation of termination codon readthrough by aminoglycosides. *Elife*, **9**, e52611.
67. Phillips-Jones,M.K., Hill,L.S., Atkinson,J. and Martin,R. (1995) Context effects on misreading and suppression at UAG codons in human cells. *Mol. Cell. Biol.*, **15**, 6593–6600.
68. Phillips-Jones,M.K., Watson,F.J. and Martin,R. (1993) The 3′ codon context effect on UAG suppressor tRNA is different in Escherichia coli and human cells. *J. Mol. Biol.*, **233**, 1–6.
69. Schinn,S.-M., Bradley,W., Groesbeck,A., Wu,J.C., Broadbent,A. and Bundy,B.C. (2017) Rapid in vitro screening for the location-dependent effects of unnatural amino acids on protein expression and activity. *Biotechnol. Bioeng.*, **114**, 2412–2417.
70. Hostetler,Z.M., Ferrie,J.J., Bornstein,M.R., Sungwienwong,I., Petersson,E.J. and Kohli,R.M. (2018) Systematic evaluation of soluble protein expression using a fluorescent unnatural amino acid reveals no reliable predictors of tolerability. *ACS Chem. Biol.*, **13**, 2855–2861.
71. Martin,R., Mogg,A.E., Heywood,L.A., Nitschke,L. and Burke,J.F. (1989) Aminoglycoside suppression at UAG, UAA and UGA codons in *Escherichia coli* and human tissue culture cells. *Mol. Gen. Genet.*, **217**, 411–418.
72. Potts,K.A., Stieglitz,J.T., Lei,M. and Van Deventer,J.A. (2020) Reporter system architecture affects measurements of noncanonical amino acid incorporation efficiency and fidelity. *Mol. Syst. Des. Eng.*, **5**, 573–588.
73. Monk,J.W., Leonard,S.P., Brown,C.W., Hammerling,M.J., Mortensen,C., Gutierrez,A.E., Shin,N.Y., Watkins,E., Mishler,D.M. and Barrick,J.E. (2017) Rapid and inexpensive evaluation of nonstandard amino acid incorporation in *Escherichia coli*. *ACS Synth. Biol.*, **6**, 45–54.
74. Stieglitz,J.T., Kehoe,H.P., Lei,M. and Van Deventer,J.A. (2018) A robust and quantitative reporter system to evaluate noncanonical amino acid incorporation in yeast. *ACS Synth. Biol.*, **7**, 2256–2269.
75. Elliott,T.S., Bianco,A., Townsley,F.M., Fried,S.D. and Chin,J.W. (2016) Tagging and enriching proteins enables cell-specific proteomics. *Cell Chem Biol*, **23**, 805–815.
76. Mulholland,C.B., Smets,M., Schmidtmann,E., Leidescher,S., Markaki,Y., Hofweber,M., Qin,W., Manzo,M., Kremmer,E., Thanisch,K. *et al.* (2015) A modular open platform for systematic functional studies under physiological conditions. *Nucleic Acids Res.*, **43**, e112.
77. Friedrich,G. and Soriano,P. (1991) Promoter traps in embryonic stem cells: a genetic screen to identify and mutate developmental genes in mice. *Genes Dev.*, **5**, 1513–1523.
78. Zambrowicz,B.P., Imamoto,A., Fiering,S., Herzenberg,L.A., Kerr,W.G. and Soriano,P. (1997) Disruption of overlapping transcripts in the ROSA beta geo 26 gene trap strain leads to widespread expression of beta-galactosidase in mouse embryos and hematopoietic cells. *Proc. Natl. Acad. Sci. U.S.A.*, **94**, 3789–3794.
79. Ran,F.A., Hsu,P.D., Wright,J., Agarwala,V., Scott,D.A. and Zhang,F. (2013) Genome engineering using the CRISPR-Cas9 system. *Nat. Protoc.*, **8**, 2281–2308.
80. Gutschner,T., Haemmerle,M., Genovese,G., Draetta,G.F. and Chin,L. (2016) Post-translational Regulation of Cas9 during G1 Enhances Homology-Directed Repair. *Cell Rep.*, **14**, 1555–1566.
81. Hermann,M., Stillhard,P., Wildner,H., Seruggia,D., Kapp,V., Sánchez-Iranzo,H., Mercader,N., Montoliu,L., Zeilhofer,H.U. and Pelczar,P. (2014) Binary recombinase systems for high-resolution conditional mutagenesis. *Nucleic Acids Res.*, **42**, 3894–3907.
82. Wickham,H., Averick,M., Bryan,J., Chang,W., McGowan,L.D., François,R., Grolemund,G., Hayes,A., Henry,L., Hester,J. *et al.* (2019) Welcome to the Tidyverse. *J. Open Source Softw.*, **4**, 1686.
83. Spidlen,J., Breuer,K., Rosenberg,C., Kotecha,N. and Brinkman,R.R. (2012) FlowRepository: a resource of annotated flow cytometry datasets associated with peer-reviewed publications. *Cytometry A*, **81**, 727–731.
84. Rappsilber,J., Mann,M. and Ishihama,Y. (2007) Protocol for micro-purification, enrichment, pre-fractionation and storage of peptides for proteomics using StageTips. *Nat. Protoc.*, **2**, 1896–1906.
85. Perez-Riverol,Y., Csordas,A., Bai,J., Bernal-Llinares,M., Hewapathirana,S., Kundu,D.J., Inuganti,A., Griss,J., Mayer,G., Eisenacher,M. *et al.* (2019) The PRIDE database and related tools and resources in 2019: improving support for quantification data. *Nucleic Acids Res.*, **47**, D442–D450.
86. Cox,J. and Mann,M. (2008) MaxQuant enables high peptide identification rates, individualized ppb-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.*, **26**, 1367.
87. Cox,J., Neuhauser,N., Michalski,A., Scheltema,R.A., Olsen,J.V. and Mann,M. (2011) Andromeda: a peptide search engine integrated into the MaxQuant environment. *J. Proteome Res.*, **10**, 1794–1805.
88. Cox,J., Hein,M.Y., Luber,C.A., Paron,I., Nagaraj,N. and Mann,M. (2014) Accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction, termed MaxLFQ. *Mol. Cell. Proteomics*, **13**, 2513–2526.
89. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Michael Cherry,J., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T. *et al.* (2000) Gene Ontology: tool for the unification of biology. *Nat. Genet.*, **25**, 25–29.
90. Brooks,A.N., Yang,L., Duff,M.O., Hansen,K.D., Park,J.W., Dudoit,S., Brenner,S.E. and Graveley,B.R. (2011) Conservation of an RNA regulatory map between Drosophila and mammals. *Genome Res.*, **21**, 193–202.

91. Hoerl,A.E. and Kennard,R.W. (1970) Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, **12**, 55–67.
92. Wu,X. and Bartel,D.P. (2017) kpLogo: positional k-mer analysis reveals hidden specificity in biological sequences. *Nucleic Acids Res.*, **45**, W534–W538.
93. Chou,C., Uprety,R., Davis,L., Chin,J.W. and Deiters,A. (2011) Genetically encoding an aliphatic diazirine for protein photocrosslinking. *Chem. Sci.*, **2**, 480–483.
94. Ai,H.-W., Shen,W., Sagi,A., Chen,P.R. and Schultz,P.G. (2011) Probing protein-protein interactions with a genetically encoded photo-crosslinking amino acid. *Chem. Bio. Chem*, **12**, 1854–1857.
95. Cigler,M., Müller,T.G., Horn-Ghetko,D., von Wrisberg,M.-K., Fottner,M., Goody,R.S., Itzen,A., Müller,M.P. and Lang,K. (2017) Proximity-triggered covalent stabilization of low-affinity protein complexes in vitro and in vivo. *Angew. Chem. Int. Ed Engl.*, **56**, 15737–15741.
96. Blackman,M.L., Royzen,M. and Fox,J.M. (2008) Tetrazine ligation: fast bioconjugation based on inverse-electron-demand Diels-Alder reactivity. *J. Am. Chem. Soc.*, **130**, 13518–13519.
97. Lang,K., Davis,L., Wallace,S., Mahesh,M., Cox,D.J., Blackman,M.L., Fox,J.M. and Chin,J.W. (2012) Genetic encoding of bicyclononynes and trans-cyclooctenes for site-specific protein labeling in vitro and in live mammalian cells via rapid fluorogenic Diels-Alder reactions. *J. Am. Chem. Soc.*, **134**, 10317–10320.
98. Borrmann,A., Milles,S., Plass,T., Dommerholt,J., Verkade,J.M.M., Wiessler,M., Schultz,C., van Hest,J.C.M., van Delft,F.L. and Lemke,E.A. (2012) Genetic encoding of a bicyclo[6.1.0]nonyne-charged amino acid enables fast cellular protein imaging by metal-free ligation. *Chem. Bio. Chem*, **13**, 2094–2099.
99. Mideksa,Y.G., Fottner,M., Braus,S., Weiß,C.A.M., Nguyen,T.-A., Meier,S., Lang,K. and Feige,M.J. (2020) Site-specific protein labeling with fluorophores as a tool to monitor protein turnover. *Chem. Biol. Chem*, **21**, 1861–1867.
100. Sun,J., Chen,M., Xu,J. and Luo,J. (2005) Relationships among stop codon usage bias, its context, isochores, and gene expression level in various eukaryotes. *J. Mol. Evol.*, **61**, 437–444.
101. Brown,A., Shao,S., Murray,J., Hegde,R.S. and Ramakrishnan,V. (2015) Structural basis for stop codon recognition in eukaryotes. *Nature*, **524**, 493–496.
102. Matheisl,S., Berninghausen,O., Becker,T. and Beckmann,R. (2015) Structure of a human translation termination complex. *Nucleic Acids Res.*, **43**, 8615–8626.
103. Bindels,D.S., Haarbosch,L., van Weeren,L., Postma,M., Wiese,K.E., Mastop,M., Aumonier,S., Gotthard,G., Royant,A., Hink,M.A. *et al.* (2017) mScarlet: a bright monomeric red fluorescent protein for cellular imaging. *Nat. Methods*, **14**, 53–56.
104. Shaner,N.C., Lambert,G.G., Chammas,A., Ni,Y., Cranfill,P.J., Baird,M.A., Sell,B.R., Allen,J.R., Day,R.N., Israelsson,M. *et al.* (2013) A bright monomeric green fluorescent protein derived from Branchiostoma lanceolatum. *Nat. Methods*, **10**, 407–409.
105. Lo,C.-A., Kays,I., Emran,F., Lin,T.-J., Cvetkovska,V. and Chen,B.E. (2015) Quantification of protein levels in single living cells. *Cell Rep.*, **13**, 2634–2644.
106. Furtado,A. and Henry,R. (2002) Measurement of green fluorescent protein concentration in single cells by image analysis. *Anal. Biochem.*, **310**, 84–92.
107. Kalstrup,T. and Blunck,R. (2015) Reinitiation at non-canonical start codons leads to leak expression when incorporating unnatural amino acids. *Sci. Rep.*, **5**, 11866.
108. Cohen,S., Kramarski,L., Levi,S., Deshe,N., Ben David,O. and Arbely,E. (2019) Nonsense mutation-dependent reinitiation of translation in mammalian cells. *Nucleic Acids Res.*, **47**, 6330–6338.
109. Bannister,A.J. and Kouzarides,T. (2011) Regulation of chromatin by histone modifications. *Cell Res.*, **21**, 381–395.
110. Kurosaki,T., Popp,M.W. and Maquat,L.E. (2019) Quality and quantity control of gene expression by nonsense-mediated mRNA decay. *Nat. Rev. Mol. Cell Biol.*, **20**, 406–420.
111. Kramarski,L. and Arbely,E. (2020) Translational read-through promotes aggregation and shapes stop codon identity. *Nucleic Acids Res.*, **48**, 3747–3760.
112. Goelet,P., Lomonossoff,G.P., Butler,P.J., Akam,M.E., Gait,M.J. and Karn,J. (1982) Nucleotide sequence of tobacco mosaic virus RNA. *Proc. Natl. Acad. Sci. U.S.A.*, **79**, 5818–5822.
113. Kipper,K., Lundius,E.G., Ćurić,V., Nikić,I., Wiessler,M., Lemke,E.A. and Elf,J. (2017) Application of noncanonical amino acids for protein labeling in a genomically recoded *Escherichia coli*. *ACS Synth. Biol.*, **6**, 233–255.
114. Smith,D. and Yarus,M. (1989) Transfer RNA structure and coding specificity. I. Evidence that a D-arm mutation reduces tRNA dissociation from the ribosome. *J. Mol. Biol.*, **206**, 489–501.

Qin, W., Stengl, A., **Ugur, E.**, Leidescher, S., Ryan, J., Cardoso, M. C., & Leonhardt, H. (**2021**). HP1**β** carries an acidic linker domain and requires H3K9me3 for phase separation. Nucleus (Austin, Tex.), 12(1), 44–57.

License: CC-BY-4.0

https://doi.org/10.1080/19491034.2021.1889858

Nucleus

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/kncl20

HP1β carries an acidic linker domain and requires H3K9me3 for phase separation

Weihua Qin, Andreas Stengl, Enes Ugur, Susanne Leidescher, Joel Ryan, M. Cristina Cardoso & Heinrich Leonhardt

To cite this article: Weihua Qin, Andreas Stengl, Enes Ugur, Susanne Leidescher, Joel Ryan, M. Cristina Cardoso & Heinrich Leonhardt (2021) HP1β carries an acidic linker domain and requires H3K9me3 for phase separation, Nucleus, 12:1, 44-57, DOI: 10.1080/19491034.2021.1889858

To link to this article: https://doi.org/10.1080/19491034.2021.1889858

© 2021 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group.

View supplementary material

Published online: 04 Mar 2021.

Submit your article to this journal

Article views: 2742

View related articles

View Crossmark data

Citing articles: 8 View citing articles

Full Terms & Conditions of access and use can be found at
https://www.tandfonline.com/action/journalInformation?journalCode=kncl20

NUCLEUS
2021, VOL. 12, NO. 1, 44–57
https://doi.org/10.1080/19491034.2021.1889858

Taylor & Francis
Taylor & Francis Group

RESEARCH PAPER

OPEN ACCESS Check for updates

HP1β carries an acidic linker domain and requires H3K9me3 for phase separation

Weihua Qin[a], Andreas Stengl[a], Enes Ugur[a,b], Susanne Leidescher[a], Joel Ryan[a], M. Cristina Cardoso [c], and Heinrich Leonhardt [a]

[a]Center for Molecular Biosystems (BioSysM), Faculty of Biology, Ludwig-Maximilians-Universität München, Munich, Germany; [b]Department of Proteomics and Signal Transduction, Max Planck Institute for Biochemistry, Martinsried, Germany; [c]Cell Biology and Epigenetics, Department of Biology, Technical University of Darmstadt, Darmstadt, Germany

ABSTRACT

Liquid-liquid phase separation (LLPS) mediated formation of membraneless organelles has been proposed to coordinate biological processes in space and time. Previously, the formation of phase-separated droplets was described as a unique property of HP1α. Here, we demonstrate that the positive net charge of the intrinsically disordered hinge region (IDR-H) of HP1 proteins is critical for phase separation and that the exchange of four acidic amino acids is sufficient to confer LLPS properties to HP1β. Surprisingly, the addition of mono-nucleosomes promoted H3K9me3-dependent LLPS of HP1β which could be specifically disrupted with methylated but not acetylated H3K9 peptides. HP1β mutants defective in H3K9me3 binding were less efficient in phase separation*in vitro* and failed to accumulate at heterochromatin *in vivo*. We propose that multivalent interactions of HP1β with H3K9me3-modified nucleosomes via its chromodomain and dimerization via its chromoshadow domain enable phase separation and contribute to the formation of heterochromatin compartments *in vivo*.

ARTICLE HISTORY
Received 10 October 2020
Revised 6 February 2021
Accepted 8 February 2021

KEYWORDS
Phase separation; chromatin structure; heterochromatin; histone posttranslational modification; heterochromatin binding protein HP1

Introduction

Liquid-liquid phase separation (LLPS) has recently emerged as a novel form of the cellular organization [1–4]. In addition to canonical membrane-bound organelles, phase separation forms membraneless organelles within cells to compartmentalize complex biological reactions in space and time. The formation of membraneless organelles is driven by intrinsically disordered proteins or disordered protein regions (IDR) [5,6]. Those proteins or protein domains are characterized by a low content of hydrophobic amino acids, biased amino acid composition, and low sequence complexity [5,7–10]. The cellular abundance of disordered proteins is tightly regulated and mutations in those proteins or changes in their cellular abundance are often associated with disease [11,12].

Heterochromatin binding protein HP1 is a non-histone chromosome binding protein and has a function in nuclear organization, chromosome segregation, telomere maintenance, DNA repair, and gene silencing [13,14]. In mammals, there are three homologs of HP1, termed HP1α, HP1β, and HP1γ, encoded by the genes *Cbx5*, *Cbx1*, and *Cbx3*, respectively. HP1 homologs have two conserved functional domains, an N-terminal chromodomain (CD) and a C-terminal chromoshadow domain (CSD), linked by a hinge region. The CD domain mediates recognition of di- and trimethylated K9 of histone H3 (H3K9me2 and H3K9me3) [15–17], while the CSD domain is responsible for interaction with other proteins and also mediates homo- and hetero-dimerization [18,19]. The intrinsically disordered regions and posttranslational modifications are likely responsible for the unique functions of HP1 homologs. Recent studies testing the capacity of HP1 to induce phase separation revealed that only HP1α formed phase-separated droplets [20–22]. This phase separation is initiated through intermolecular interaction of the phosphorylated N-terminus with the hinge region and correlates with the formation of heterochromatin and chromocenters in the nucleus.

CONTACT Weihua Qin weihua@zi.biologie.uni-muenchen.de; Heinrich Leonhardt h.leonhardt@lmu.de Center for Molecular Biosystems (BioSysM), Faculty of Biology, Ludwig-Maximilians-Universität München, Butenandtstraße 1, Munich D-81377, Germany

Supplemental data for this article can be accessed here.

© 2021 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Although HP1β also predominantly accumulates at pericentromeric heterochromatin (chromocenters), it does not form phase-separated droplets under the conditions described for HP1α. It is though not clear how much LLPS mechanisms contribute to heterochromatin formation and clustering and, indeed, a model polymer-polymer /liquid-phase separation (PPPS or PLPS) has been recently proposed [23,24].

Chromatin organization undergoes dramatic changes during mammalian cell differentiation and proliferation. In proliferating cells, heterochromatin clusters (chromocenters) are disrupted during mitosis as they contain clustered centromeric and pericentromeric DNA from several chromosomes and then fuse again throughout interphase reaching the highest clustering in G2 and in terminally differentiated post-mitotic cells [25]. This fusion of chromocenters *in vivo* resembles the formation of phase-separated droplets *in vitro* and depends on the presence and concentration of heterochromatin proteins like HP1α and MeCP2 [20,22,25]. At the transition from pluripotent to differentiated stages, heterochromatin foci become more clustered and spherical [25,26], which correlates with lower exchange rates of chromatin proteins. HP1 proteins form homo- or hetero-dimers and have often been considered to play a rather equivalent role in heterochromatin organization. However, several lines of evidence suggest that the different HP1 proteins have specific functions in heterochromatin organization. For example, it has been shown that HP1α plays an important role in heterochromatin organization, while HP1β functionally associates with H4K20me3 [27,28]. HP1β has been suggested to act as a bridge linking H3K9me3 enriched condensed chromatin [29]. In addition, HP1α and HP1β likely play distinct roles during early embryo development, as they show different expression patterns [30].

To dissect functional differences of HP1 homologs in phase separation and chromatin organization, we compared the amino acid composition of HP1 proteins at disordered regions. We found that the charge of the IDR-H is a distinctive feature of HP1 homologs and plays a decisive role in LLPS and that HP1β undergoes phase separation in a histone H3K9me3 dependent manner. Hence, an HP1β mutant defective in H3K9me3 binding was deficient in phase separation and showed faster binding kinetics *in vivo*.

Materials and methods

Cell culture and transfection

Mouse E14 ESCs, cells were cultured in gelatinized flasks in DMEM supplemented with 16% fetal calf serum, 0.1 mM β-mercaptoethanol (ThermoFisher Scientific, Invitrogen), 2 mM L-glutamine, 1× MEM non-essential amino acids, 100 U/ml penicillin, 100 μg/ml streptomycin (Sigma-Aldrich, Germany), 2i (1 μM PD032591 and 3 μM CHIR99021 (Axon Medchem, Netherlands) and 1000 U/ml recombinant leukemia inhibitory factor LIF (Millipore). Human embryonic kidney (HEK) 293 T cells were cultured in DMEM supplemented with 10% fetal calf serum and 50 μg/ml gentamycin (PAA).

Mouse ESCs were transfected with Lipofectamine 3000 Reagent (ThermoFisher Scientific, Invitrogen) according to the manufacturer's instructions.

CRISPR/Cas-mediated gene editing and generation of stable cell lines

For the generation of GFP-HP1β WT and KW cell lines, the MINtag strategy was used as described previously [31]. In brief, HP1β specific gRNA was cloned into a vector expressing GFP and SpCas9 (px458: F. Zhang Lab). Mouse ESCs were transfected with the Cas9-gRNA vector and a 200 nt donor oligo coding for the MINtag. Two days after transfection, GFP positive cells were separated using FACS (Becton Dickinson, Germany) and plated at clonal density (2000 cells per p100 plate). After one-week, single clones were picked manually and transferred into two 96-well plates. Cell lysis in 96-well plates, PCR on lysates, and restriction digest were performed. To generate WT and KW GFP-HP1β cell lines, we used our MIN-tagged HP1β mESCs and inserted the WT or the KW GFP-HP1β coding sequence into the N-terminus of the endogenous *HP1β*$^{attP/attP}$ locus by Bxb1 mediated recombination (Figure 5d). PCR primers for screening are as follows:

HP1β-ext F: 5′-GATTTCCCTGGGCTCCTCAC-3′

HP1β-ext R: 5′-ATGCCCATCACAGAACTGCT-3′

AttL-F: 5′- CCGGCTTGTCGACGACG-3′.

Protein purification, histone, and mononucleosome isolation

HP1 cDNA was cloned into a pET28a (+) expression vector (Merck KGaA, Novagen), mutants were made using overlap extension PCR, and proteins were subsequently expressed in *E. coli*. For protein purification, BL21 cells were grown to OD 0.6–0.8 at 37°C, then IPTG was added to a final concentration of 0.5 mM and cultures were incubated at 18°C overnight. Harvested cells were resuspended in lysis buffer (20 mM Tris-HCl pH 8.2, 250 mM NaCl, 20 mM Imidazole, 3 mM β-Mercaptoethanol, 1 mM PMSF, 25 µg/ml DNase I and 100 µg/ml Lysozyme) and incubated at 4°C under constant rotation for 1–2 h. Following sonication, cell debris was removed by centrifugation at 20,000 x g for 30 min at 4°C. Clarified lysate was injected into an ÄKTA Purifier system (GE Life Sciences, Germany) equipped with a Ni-NTA column and His-tagged proteins were finally eluted in elution buffer (20 mM Tris/HCl pH 8.2, 250 mM NaCl, 500 mM Imidazole, and 3 mM β–Mercaptoethanol). The fractions with the highest purity were mixed and concentrated to about 1 µg/µl using Amicon® Ultra 4 mL centrifugal filter (Merck, Germany) in the buffer (20 mM HEPES pH 7.2, 200 mM KCl, 1 mM DTT, 10% glycerol) before flash freezing in liquid nitrogen. Protein concentrations were measured with the Pierce™ 660 nm protein assay kit (ThermoFisher Scientific) according to the manual.

Histone isolation was conducted as previously described with minor changes of the protocol [33]. In brief, 15 p100 plates of HEK293T cells were harvested and cell pellets were resuspended in a hypotonic buffer (10 mM Tris-HCl pH 8, 10 mM KCl, 1.5 mM $MgCl_2$, 1 mM DTT, and 1x Protease Inhibitor, 2 mM PMSF). To obtain pure nuclei, cells were disrupted using a homogenizer and nuclei were subsequently incubated in a chromatin dissociation buffer (10 Tris-HCl pH 8.0, 20 mM EDTA, and 400 mM NaCl) for 30 min on ice. This chromatin dissociation step was repeated 4x. Afterward, nuclei were resuspended in 0.4 N H_2SO_4 and incubated on a rotator at 4°C overnight. After centrifugation, histones in the supernatant were transferred into a fresh reaction tube and precipitated using 33% trichloroacetic acid (TCA). After washing 3x with cold acetone, histones were dissolved in H_2O and centrifuged at 2000 rpm for 5 min to remove precipitates. Histone concentrations were measured using the Pierce™ 660 nm protein assay kit.

For isolation of mononucleosomes, 3×10^7 HEK293T cells were resuspended in 1 ml of hypotonic buffer containing 0.1% Triton-X 100, homogenized with 20 strokes in a Glass Teflon homogenizer and centrifuged at 1000 x g at 4°C to obtain intact nuclei. Nuclei were then resuspended in 800 µl of MNase digestion buffer (10 mM Tris-HCl, pH 7.4, 10 mM NaCl, 3 mM $CaCl_2$, 0.1% NP-40, and protease inhibitors) supplemented with 40 U/ml MNase and incubated at 37°C for 5 min. The digestion was inactivated by a 5× stop buffer containing 10 mM Tris-HCl, pH7.4, 710 mM NaCl, and 7.5 mM EDTA. Mononucleosome extracts were cleared by centrifugation at 2 000 x g for 15 min at 4°C and the quality of the preparation was determined on an agarose gel after isolating DNA from the mononucleosome extracts.

In vitro droplet assays

For the droplet assay, proteins were concentrated to ~10 µg/µl using Amicon® Ultra 4 mL centrifugal filter (Merck, Germany). After the concentration step, the buffer was exchanged to 20 mM HEPES pH 7.2, 75 mM KCl, 1 mM DTT with Zeba™ Spin Desalting Columns, and 1.4 nmol of HP1β were mixed with 1.4 nmol of histones in a total of 30 µl buffer at 4°C. 20 µl of the turbid solution was imaged in a 15 µ-Slide 18 Well ibidi chamber. Differential interference contrast (DIC) images were acquired on a DeltaVision Personal widefield microscope (GE Life Sciences) equipped with a 60 × 1.42 NA objective (Olympus), LED epi-illumination, and a CoolSnap ES2 camera (Photometrics).

For the spin-down assay, 30 µl of the turbid solution was spun down at 2000 rpm for 5 min and 29 µl of supernatant was transferred into a Protein LoBind Tube (Eppendorf). The supernatant and droplets were boiled in 250 µl Laemmli loading buffer at 95°C for 10 min. 5 µl of supernatant and droplets were loaded into an SDS-PAGE gel followed by either detection using coomassie staining or western blot analysis.

For HP1β phase separation with mononucleosome extracts, 28 µl of extract were incubated with 30 µg of HP1β in 30 µl solution for 5 min and spun down at 12,000 rpm for 5 min. 29 µl of supernatant was transferred into a Protein LoBind Tube and both supernatant and droplets were boiled in 40 µl Laemmli loading buffers at 95°C for 10 min. 20 µl of supernatant and droplets were again loaded into an SDS-PAGE gel followed by either detection using Coomassie staining or western blot analysis.

For comparison of histones and H3 peptides in HP1β phase separation, H3 peptides (aa 1–20) carrying H3K9me3 and biotinylated at the C-terminus were purchased from PSL GmbH, Heidelberg.

For the peptide competition assay, 25 µM of HP1β was incubated for 1 h with C-terminal TAMRA labeled histone H3 peptides (aa 1–20), containing H3K9me3, H3K9me1 or H3K9ac (PSL GmbH, Heidelberg) in a ratio of 1:5 or 1:50 in 30 µl buffers in Protein LoBind Tubes at 4°C. Then, 25 µM of histones were added to the solution and incubated at 4°C for 3 min. Droplets were separated by centrifugation at 2000 rpm for 5 min and 29 µl of supernatant were transferred into a new microfuge tube. Supernatant and droplets were boiled in 200 µl Laemmli loading buffers at 95°C for 10 min and 6 µl of each sample was loaded into an SDS-PAGE gel for detection by coomassie staining and TAMRA fluorescence.

Analytic ion-exchange chromatography (IEX) and size-exclusion chromatography (SEC)

The surface charge of HP1 variants was analyzed by anion exchange chromatography. 50 µg HP1α, HP1β, or HP1γ were diluted in 500 µL buffer A (20 mM Tris-HCl, pH 8.0) and loaded on a 1 mL Resource Q column at room temperature and 4 ml/min flow rate using a Äkta Pure FPLC system. Samples were eluted with a linear gradient over 20 column volumes (CV) to 50% buffer B (20 mM Tris-HCl, 1 M NaCl, pH 8.0) followed by 10 CV 100% buffer B. Absorption at 280 nm was recorded.

250 µg of extracted histones were diluted in 50 µL SEC running buffers (20 mM Tris-HCl, 300 mM NaCl, pH 7.4). The sample was separated on an equilibrated Superdex 200 Increase 10/300 GL column at room temperature and 0.75 ml/min flow rate using a Äkta Pure FPLC system. Absorption at 280 nm was recorded. For size comparison, a protein gel filtration marker mix (Sigma-Aldrich) including carbonic anhydrase (29 kDa), bovine serum albumin (66 kDa), alcohol dehydrogenase (150 kDa), beta-amylase (200 kDa), apoferritin (443 kDa), thyroglobulin (669 kDa) was analyzed under identical conditions.

Antibodies for western blot analysis

Primary antibodies used for western blot, including the polyclonal rabbit anti-H3 (Cat # ab1791), anti-H3K9me3 (Cat # ab8898), and anti-HP1β (Cat #10478) antibodies, were purchased from Abcam and the secondary antibody, anti-rabbit-IgG AF647 (Cat # A32733), from Invitrogen. The primary mouse monoclonal anti-H1 antibody (H-2) was purchased from Santa Cruz Biotechnology (Cat # sc-393358) and the secondary antibodies, anti-mouse IgG-HRP (Cat # A9044), and anti-rabbit IgG-HRP (Cat # A6154) were purchased from Sigma-Aldrich.

Immunofluorescence staining

mESCs were washed with phosphate-buffered saline (PBS) and fixed with 3.7% formaldehyde in PBS, permeabilized with 0.5% Triton X-100 in PBS, and then blocked with 3% BSA. Cells were then incubated with a rabbit polyclonal anti-H3K9me3 antibody (Abcam, Cat # ab8898) or a rabbit polyclonal anti-HP1β antibody (Abcam, Cat #10478) for 1 hour at RT. After washing, cells were incubated with Alexa594-conjugated donkey anti-rabbit IgG secondary antibody (Invitrogen, Cat # A21207) for H3K9me3 and Alexa488-conjugated goat anti-rabbit IgG (Invitrogen, Cat # A11034) for HP1β 1 hour at RT. Nuclei were stained with 4′,6-diamidino-2-phenylindole (DAPI) and mounted on coverslips with Vectashield (Vector Laboratories). Images were taken using an SP5 Leica confocal microscope equipped with Plan Apo 63x/1.4 NA oil immersion objective and lasers with excitation lines: 405 nm for DAPI, 488 nm for HP1β and GFP-HP1β, and 594 nm for H3K9me3.

FRAP analysis

FRAP experiments were performed on an UltraVIEW VoX spinning disc microscope with an integrated FRAP PhotoKinesis accessory (PerkinElmer) assembled onto an Axio Observer D1 inverted stand (Zeiss) and using a 100×/1.4 NA Plan-Apochromat oil immersion objective. The microscope was equipped with a heated environmental chamber set to 37°C. Fluorophores were excited with a 488 nm solid-state diode laser line. Confocal image series were typically recorded with 16-bit image depth, a frame size of 512 × 512 or 256 × 256 pixels, and a pixel size of 69 nm. The bleach regions, typically with a diameter of 2 µm, were manually chosen to cover chromocenters. Photobleaching was performed using one iteration with the acousto-optical tunable filter (AOTF) of the 488 nm laser line set to 100% transmission. Twenty prebleach images were acquired at maximum speed, then 60 post-bleach frames were recorded at maximum speed followed by 30 frames at a rate of 3 s per frame. Data correction, normalization, and quantitative evaluations were performed by processing with ImageJ (http://rsb.info.nih.gov/ij/) followed by calculations in Excel. For normalization, the average intensity of five prebleach images was used.

Results

HP1β differs from HP1α and HP1γ in that it contains an acidic linker domain

Although the three HP1 homologs are very similar in their overall structure, only HP1α was reported to undergo LLPS [20]. As LLPS involves intrinsically disordered regions (IDRs) of proteins, we scrutinized and compared the disordered regions of HP1 proteins (Figure 1a). The

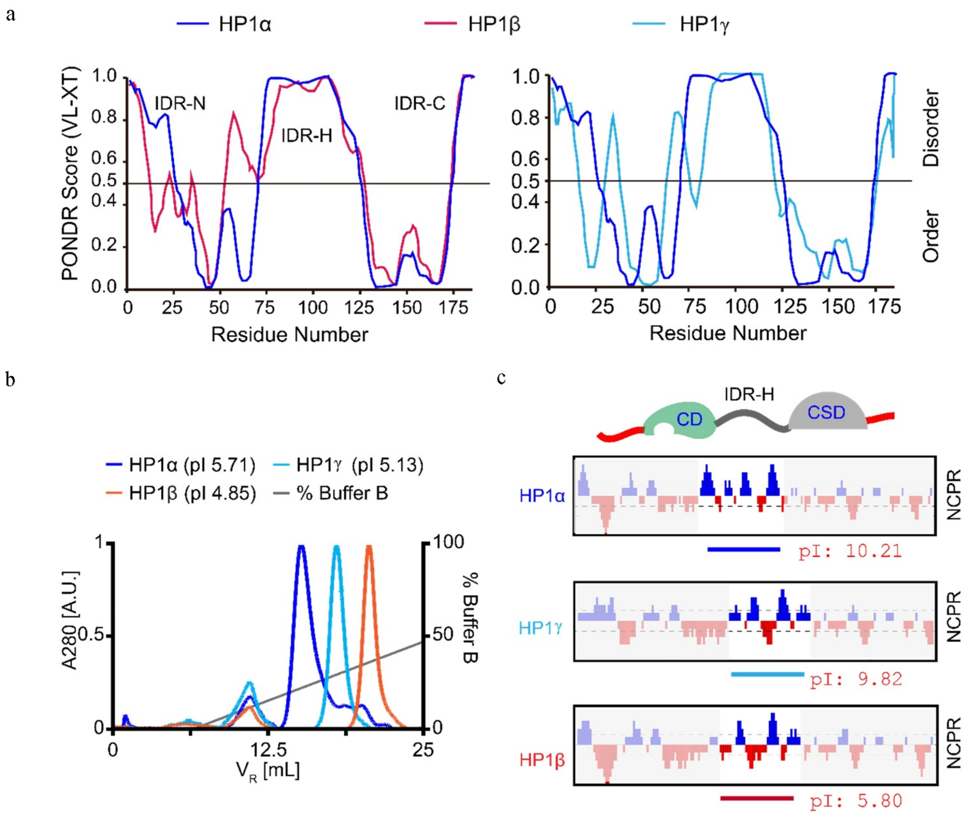


Figure 1. HP1β differs from HP1α and HP1γ in that it contains an acidic linker domain. (a) Comparison of order/disorder prediction of HP1 homologs by the PONDR algorithm, a website tool (http://www.pondr.com/). VLXT scores are shown on the y-axis, amino acid positions are shown on the x-axis. (b) Ion exchange chromatography analysis of HP1 proteins. 50 µg of HP1 proteins were diluted in 500 µL buffer B (20 mM Tris-HCl, pH 8.0) and loaded on a 1 mL Resource Q column and analyzed by using a Äkta Pure FPLC system. (c) Net charge distribution per residue (NCPR) of HP1 proteins (CIDER, pappulab.wustl.edu). Negatively charged amino acids are marked in red, positively charged amino acids in blue. The pI of IDR-H in HP1 proteins is indicated.

C-terminal disordered region (IDR-C) was relatively conserved and only minor differences were observed at the N-terminus (IDR-N) and hinge region (IDR-H) (Figure 1a). However, HP1 homologs have different theoretical isoelectric points (pI). To investigate the HP1 proteins *in vitro*, we induced the expression of His tagged HP1s in *E. coli* and purified them using a Ni-NTA column. Purified HP1 proteins were checked by coomassie blue-stained SDS-PAGE gels (Figure S1A) and showed the expected protein sizes, HP1α (24.3 kDa), HP1β (23.7 kDa), and HP1γ (23.0 kDa). Then, we performed ion-exchange chromatography analysis and confirmed the expected pI of HP1 proteins (Figure 1b). Among the three homologs, HP1β is the most acidic protein (pI 4.85), followed by HP1γ (pI 5.13) and HP1α (pI 5.71) (Figure 1b). Further analysis revealed that this difference between HP1 homologs was most pronounced in the IDR-H. Whereas HP1α and HP1γ contain more positively than negatively charged residues in their IDR-H (15/6 and 13/8, respectively), HP1β has relatively equal numbers of positively and negatively charged residues (11/12) in the IDR-H resulting in a much lower local pI of 5.80 (Figure 1c).

HP1β cannot self-phase separate because of its acidic linker domain

We next systematically compared the property of HP1 homologs in phase separation. In the absence of IDR-N phosphorylation and DNA, we observed LLPS with HP1α at 200 μM and to a lesser extent at 50 μM, both at 4°C (Figure 2a). While HP1γ underwent LLPS at a higher concentration (900 μM at 4°C); HP1β did not at any of these conditions (Figure 2a). As the most distinguishing feature of HP1β is the acidic rather than basic IDR-H, we next replaced four acidic amino acids in IDR-H, including aspartic acid (D) 88, D90, glutamic acid (E) 92, and D93, with lysine (K) or arginine (R) (HP1β RKRK). Indeed, this engineered HP1β RKRK could form phase-separated droplets at concentrations as low as 170 μM at 4°C (Figures 2b and S1B) underscoring the decisive role of the basic IDR-H in LLPS. The size of the HP1β RKRK droplets was comparable with the HP1α droplets at the concentration of 200 μM (Figure 2a). As the self-phase separation of HP1 proteins is mediated by the interaction of IDR-N and IDR-H [20,22], we analyzed the correlation between the charge ratio of IDRs and HP1 protein concentration at which phase separation was observed. With a linear function fitting, we

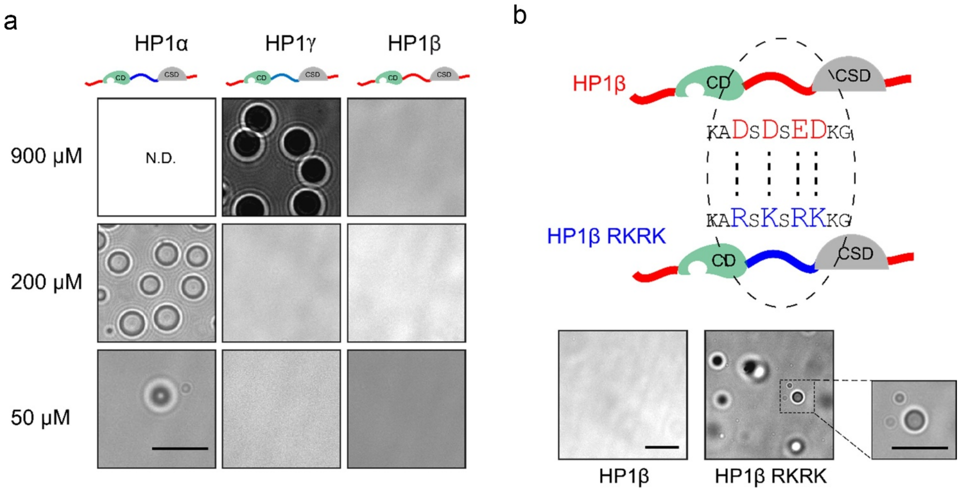


Figure 2. HP1β cannot self-phase separate because of its acidic linker domain. (a) DIC images of HP1 droplets at 4°C in a buffer containing 20 mM HEPES pH 7.2, 75 mM KCl and 1 mM DTT using the 63x objective of a DeltaVision Personal Microscope (scale bar: 10 μm). Protein concentrations are as indicated. N.D.: not done. (b) Phase separation of engineered HP1β at 170 μM and 4°C with four amino acid substitutions in the IDR-H changing it from acidic to basic (HP1β RKRK). A zoomed-in image is shown with the same magnification as in (a). Scale bar: 10 μm.

obtained an estimated concentration of HP1β self-phase separation of ~ 1.736 mM (Figure S2).

HP1β can form phase-separated droplets in the presence of histones

These results show that HP1β by itself hardly undergoes LLPS *in vitro*, but then again it interacts with numerous cellular proteins, which will likely affect and modulate its properties. As the most prominent known interactors are histone tails, we isolated mononucleosomes from HEK293T cells by MNase digestion (Figures 3a and S3). To isolate pure mononucleosomes, we first titrated the MNase concentration from 1.25 to 160 U/ml and used 40 U/ml for the preparation of mononucleosomes (Figure S3A and S3B). We incubated HP1β with isolated mononucleosomes, collected phase-separated droplets by centrifugation, and analyzed the precipitates by coomassie stained SDS-PAGE gel and western blot (Figures 3b and S4A). These results suggest that mononucleosomes promote HP1β phase separation as evidenced by an enrichment together with core histones in the pellet fraction.

To further examine the histone mediated phase separation, we prepared histones from human HEK293T cells by following an acid-extraction protocol [32]. We directly compared the three HP1 homologs and found that at low concentrations (50 μM) HP1α and HP1γ did not form phase-separated droplets with histones (Figure 3c). However, HP1β mixed with histones yielded an opalescent solution containing spherical droplets (Figure 3c) that fused over time, which is a central criterion for LLPS (Video S1).

Toward a mechanistic understanding of HP1β phase separation, we investigated the influence of protein and salt concentration on droplet formation in the presence of histones. To do so, we incubated different concentrations of HP1β protein (3 to 100 μM) with 100 μM of histones at 4°C in a buffer containing 20 mM HEPES pH 7.2, 75 mM KCl and 1 mM DTT. HP1β phase-separated droplets were then separated by centrifugation for visualization by coomassie stained SDS-PAGE gels (Figures 3d, S4B and S4C). As a control, we incubated BSA with histones at the same conditions and did not observe phase-separated droplets at any of the conditions (Figure S4D and S4E). However, in the presence of histones, HP1β solutions became turbid starting at concentrations as low as 25 μM, showing characteristic phase-separated droplets (Figures 3d, S4B and S4C). Droplets with 50 μM HP1β and stoichiometric amounts of histones formed up to 400 mM NaCl became smaller with increasing salt concentrations and disappeared at 600 mM NaCl (Figures 3e, S5A, S5B and summarized in Figure 3f). These results indicate that HP1β undergoes LLPS under physiological salt and protein concentrations.

Trimethylation of K9 of histone H3 and histone dimerization are required for HP1β phase separation

Previously it was reported that linker histone H1 forms LLPS with DNA and nucleosomes [33–35]. To investigate the contribution of histone H1 to HP1β phase separation, we analyzed phase-separated droplets by western blot. We clearly detected histone H1 in the supernatants, but not in pellets of HP1β phase-separated droplets (Figure S6). This result suggests that histone H1 is not required for HP1β phase separation.

When incubating increasing concentrations of HP1β with purified core histones, we found first histone H3 (in particular the trimethylated K9 form, H3K9me3) in droplets starting at 25 μM with a corresponding depletion from the supernatants (Figures 3d, Figures 4a and S7A), while at higher concentrations also the other core histones (H2A, H2B, and H4) were present (Figure 3d). While core histones were sufficient for HP1β LLPS, we found that H3K9me3 peptides encompassing amino acids 1–20 (aa 1–20), the binding substrate of the HP1β CSD, did not cause turbidity and droplet formation (Figure 4b). The fact that H3K9me3 histone tails were not sufficient for HP1β LLPS suggests that the remainder of the H3 histone, in particular the histone fold domains, and their ability to dimerize are required for LLPS. Indeed, size-exclusion chromatography (SEC) of histone preparations showed a major peak between 29 and 66 kDa, likely corresponding to a histone dimer (Figure 4c). We, next, performed a competition assay using H3 peptides containing either K9me3, or K9me1, or K9ac modifications added to the HP1β and histones (Figure 4d and

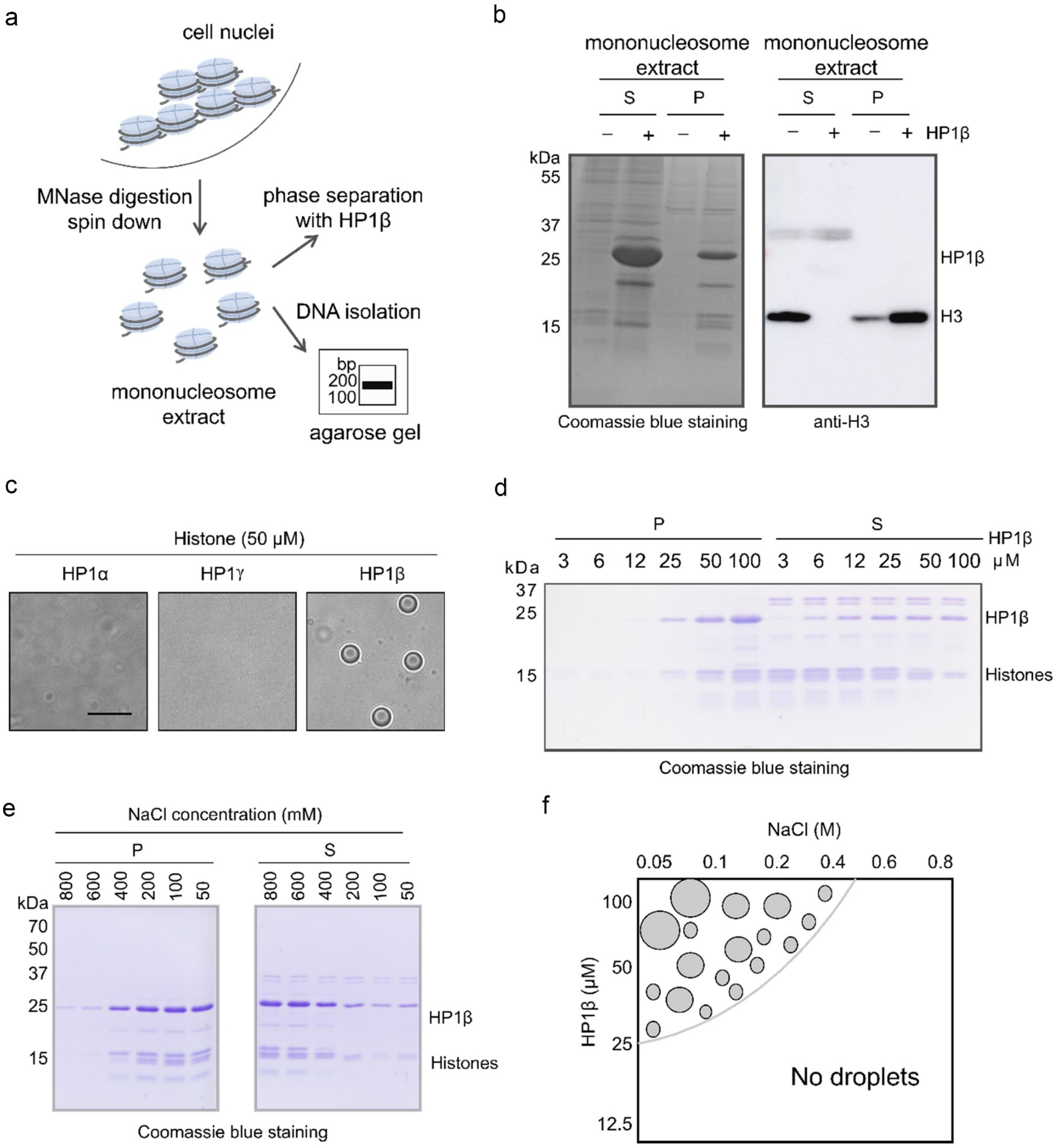


Figure 3. HP1β can form phase-separated droplets in the presence of histones. (a) Illustration of isolating mononucleosomes by MNase treatment (left). (b) Mononucleosome solution was incubated with or without 30 μg of HP1β at 4°C in a buffer containing 20 mM Tris-HCl pH 7.5, 150 mM NaCl, 3 mM $CaCl_2$, 0.1% NP-40 and 1.5 mM EDTA. Phase-separated droplets were pelleted by centrifugation. Proteins in the supernatant (S) and phase-separated droplets (P) were separated and visualized by coomassie blue SDS-PAGE gels and western blotting with an anti-H3 antibody. (c-e) HP1 phase separation in the presence of histones isolated by acid-extraction from HEK293T cells in a buffer of 20 mM HEPES pH 7.2, 75 mM KCl and 1 mM DTT. 50 μM of HP1 homologs were incubated with 50 μM of histones (scale bar: 10 μm) (c). 3 to 100 μM of HP1β was incubated with 100 μM of histones. HP1β phase-separated droplets were separated and visualized as above (d). 50 μM of HP1β was incubated with 50 μM of histones in a buffer with NaCl concentrations ranging from 50 to 800 mM. Proteins in the P and S fractions were analyzed as above (e). (f) Phase diagram of HP1β with protein and salt concentration as order parameters. Phase separation was scored by the presence or absence of droplets in the sample.

S7B). Notably, only H3K9me3 peptides, but not H3K9me1 and H3K9ac peptides, efficiently disrupted HP1β-histone dependent LLPS.

The amino acids in the CD domain, including tyrosine (Y) 21, tryptophan (W) 42, and phenylalanine (F) 45, form an aromatic cage for H3K9me3

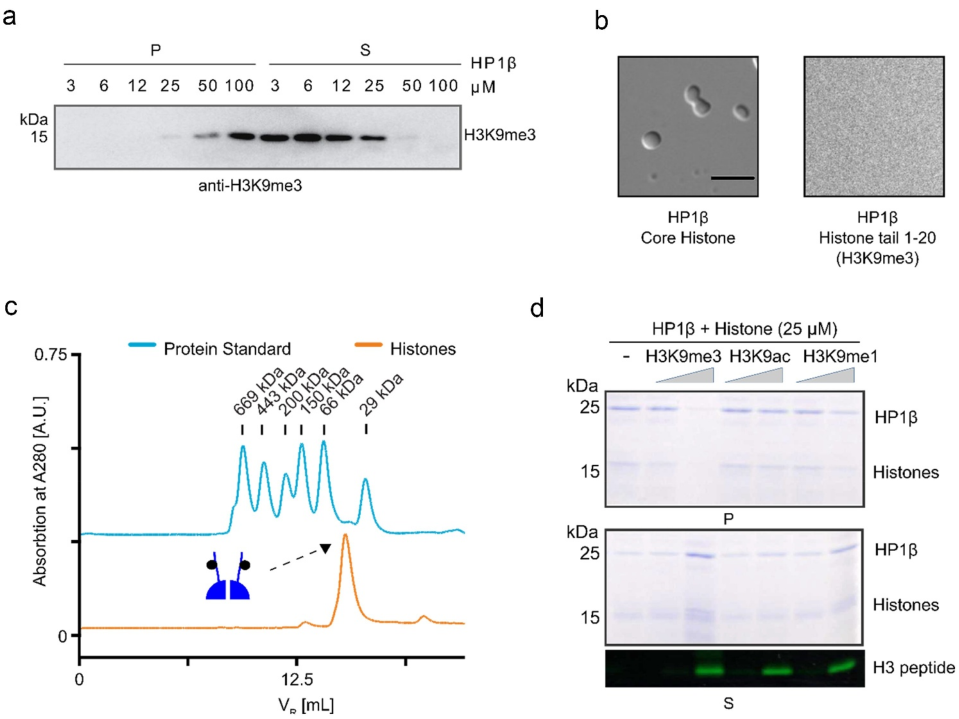


Figure 4. Trimethylation of K9 of histone H3 (H3K9me3) and histone dimerization are required for HP1β phase separation. (a) HP1β protein from 3 to 100 μM was incubated with 100 μM of histones at 4°C in a buffer containing 20 mM HEPES pH 7.2, 75 mM KCl and 1 mM DTT. HP1β phase-separated droplets were separated by spin down. Proteins in P and S fractions were analyzed by SDS-PAGE gels and visualized western blot with anti-H3K9me3 antibody. (b) Representative DIC images show HP1β phase separation assay outcome in the presence of histones or histone H3 peptide (aa 1–20) carrying H3K9me3. 25 μM of HP1β was incubated with either 25 μM core histones or H3K9me3 peptide (aa 1–20). (c) Analysis of histones by size exclusion chromatography (SEC). 250 μg of histones were diluted in a buffer of 20 mM Tris-HCl, 300 mM NaCl, pH 7.4 and separated on an equilibrated Superdex 200 Increase 10/300 GL column. For size comparison a protein marker mix including carbonic anhydrase (29 kDa), bovine serum albumin (66 kDa), alcohol dehydrogenase (150 kDa), beta-amylase (200 kDa), apoferritin (443 kDa), and thyroglobulin (669 kDa) was analyzed under identical conditions. (d) Histone H3 peptide (aa 1–20) carrying H3K9me3, or H3K9me1 or H3K9ac was incubated with 25 μM of HP1β and histones. Proteins in S and P fractions were analyzed and visualized by coomassie stained SDS-PAGE gels and H3 peptides by fluorescent imaging.

binding (Figure 5a). The replacement of K41 and W42 with alanine (HP1β KW) is sufficient to abolish the H3K9me3 binding of HP1β [15,16]. We purified HP1β KW and incubated different concentrations of the mutant proteins (6 to 25 μM) with 25 μM histones. By analyzing coomassie stained SDS-PAGE gels, we found that almost half of histone H3 was still detected in the supernatant of phase-separated droplets of HP1β KW at the concentration (12 μM), while H3 was nearly completely depleted from the supernatant into the pellet of HP1β WT droplets (Figure S8A and S8B). This concentration corresponds to the physiological HP1β concentration measured at heterochromatin [36]. At the higher concentration (25 μM), HP1β KW formed phase-separated droplets similar to HP1β WT, which may be due to the unspecific binding with histones (Figure S8A and S8B). These results indicate that HP1β KW, which is deficient in binding H3K9me3, is less efficient in forming phase-separated droplets at physiological concentrations.

To study the function of HP1β phase separation *in vivo*, we generated a mouse embryonic stem cell (mESC) line carrying the GFP-HP1β KW mutant as well as a wild type using the MIN tag genome engineering strategy, called MINtool [31]. The MINtool allows to replace the endogenous gene of interest with the mini gene products that carry mutations or tags. With this strategy, a multifunctional integrase (MIN) tag sequence was first inserted into the open reading frame of

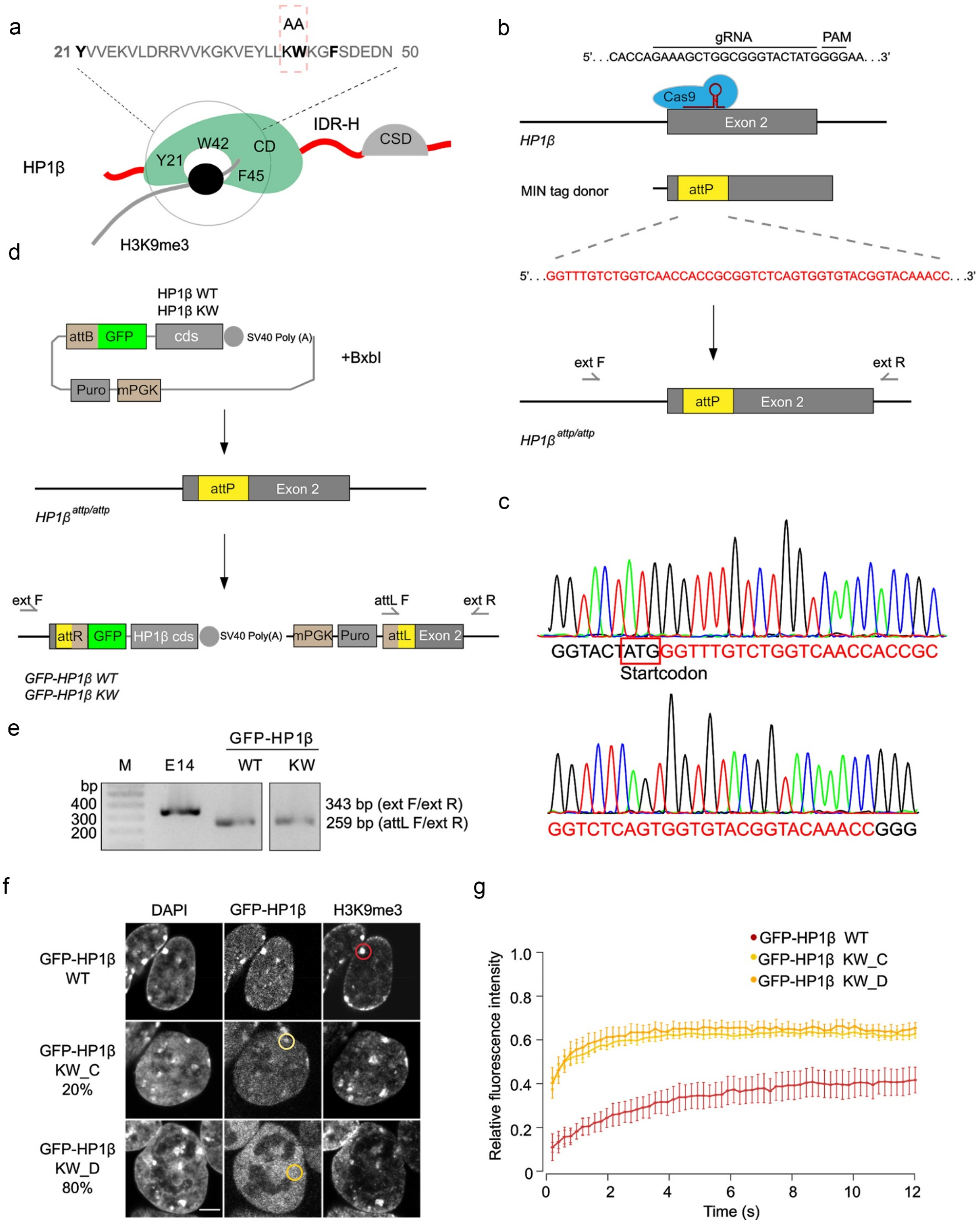


Figure 5. HP1β phase separation contributes to heterochromatin formation *in vivo*. (a) Illustration of the binding of H3K9me3 and the CD domain of HP1β. The amino acids, tyrosine (Y) 21, tryptophan (W) 42 and phenylalanine (F) 45, form an aromatic cage for H3K9me3 peptide that is abolished by the replacement of K41W42 with alanine (A) [15]. (b and c) Schematic representations show the CRISPR/Cas9 gene-editing strategy used to generate MIN tagged HP1β mESCs. The donor harbors the MIN tag sequence (*attP*) and homology arms to the genomic sequence 5′ and 3′ of the translational start site. The targeting region was amplified with primers as indicated and assessed by Sanger sequencing. (d) Schematic representation shows the strategy to generate GFP-HP1β WT and KW mESC lines with Bxb1 mediated recombination. (e) Gel electrophoresis of the multiplex PCR for validation of GFP-HP1β mESCs with primers as indicated in (d). 343 bp and 259 bp sequences were amplified from E14 and GFP-HP1β cells, respectively. (f) Representative images of GFP-HP1β WT and KW mESCs stained with an anti-H3K9me3 antibody. Scale bar: 5 µm. See overview images in Figure S10. (g) FRAP quantification of GFP-HP1β WT and GFP-HP1β KW. Curves show average GFP signal relative to the fluorescence signal prior to bleaching (WT, n = 20 and KW_C (chromocenter), n = 6 and KW_D (diffuse), n = 6). The areas used for FRAP are indicated by circles in (f).

HP1β directly downstream of the start codon by the CRISPR/Cas9 genome editing tool (Figure 5b and c). By Bxb1-mediated recombination, the coding sequences for GFP-HP1β WT and GFP-HP1β KW were subsequently integrated into the locus (Figure 5d). With specific primers, 343 bp and 259 bp sequences were amplified from the MIN tagged and GFP tagged HP1β cell lines, respectively (Figure 5e). We performed western blot analysis and found that the levels of GFP-HP1β WT and KW in the engineered cells are higher than the endogenous HP1β levels in WT mESCs (Figure S9). In line with previous publications, GFP-HP1β WT is predominantly localized at the chromocenters (Figure 5f). GFP-HP1β KW, on the other hand, showed a dispersed nuclear distribution (KW_D) in 80% of the mutant cells with no accumulation at heterochromatin compartments, while it slightly accumulated at the chromocenters (KW_C) in 20% of the cells (Figures 5f and S10). To measure the kinetics of binding in living cells, fluorescence recovery after photobleaching (FRAP) analyses were performed and evaluated. These showed a similar kinetics of recovery of GFP-HP1β KW_C and KW_D that is substantially faster than GFP-HP1β WT (Figure 5g).

Altogether, our results show that all three HP1 proteins can in principle form phase-separated droplets *in vitro* but require different conditions. While LLPS of HP1α/HP1γ mostly relies on the interaction of IDR-N and IDR-H (Figure 6), HP1β phase separation requires the binding of H3K9me3 nucleosomes (Figure 6). These multivalent interactions are required for the formation of oligomeric structures and phase-separated droplets *in vitro*. HP1β dimerization and binding of two H3K9me3 histone tails thus contribute to heterochromatin clustering *in vivo*.

Discussion

The three HP1 homologs are considered important regulators of heterochromatin formation and spreading. HP1α, but not HP1β and HP1γ, was shown to form LLPS driving heterochromatin formation [20,22], raising the question of which molecular determinants are responsible for these differences. A comparison shows that all three HP1s share a common overall structure but differ in the net charge of their IDR-H (Figure 1). We found that HP1γ, similar to HP1α, contains a basic IDR-H and indeed forms phase-separated droplets albeit at high concentrations of about 0.9 mM, which is, however, four times higher than the 0.2 mM used for HP1α and way beyond the reported physiological concentrations of about 10 μM [36,37]. Here, we showed that HP1β has a slightly acidic IDR-H in contrast to the very basic one of HP1α and HP1γ. Our further finding

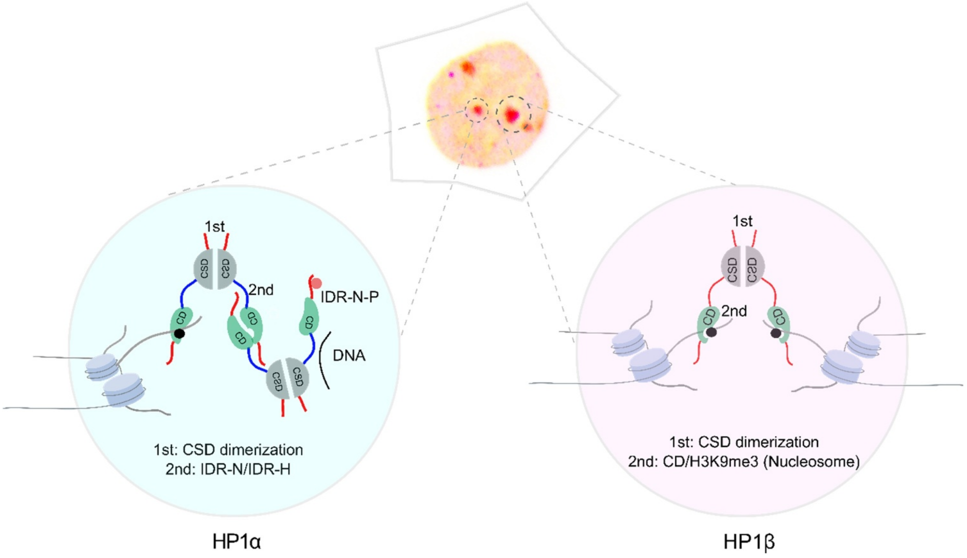


Figure 6. Model of HP1α/γ and HP1β phase separation contributing to heterochromatin formation *in vivo*. The interaction of IDR-N and IDR-H is an essential valency for HP1α and HP1γ phase separation (left). Although HP1α and HP1γ contain basic IDR-H, the minor difference leads to a threshold phase separation concentration higher than the physiological concentration for HP1γ. The negatively charged DNA and phosphorylation (P) of IDR-N can promote HP1α phase separation. In contrast, HP1β phase separation is more complex and requires the CSD mediated dimerization and the binding of the CD domain to the H3K9me3 nucleosome (right).

that HP1β WT does not form phase-separated droplets, but could be engineered to do so simply by changing four acidic to basic amino acids in the IDR-H, supports the notions that HP1α (and HP1γ) LLPS relies on interactions between their acidic IDR-N and their basic IDR-H. Previously, it was shown that the addition of negative-charged DNA promotes the phase separation of HP1α but not of HP1β [20,24]. Considering the difference of HP1α and HP1β IDR-H regions, we added positive-charged histones and found that HP1β showed phase-separated droplets even at concentrations as low as 25 µM. In line with our findings, it was shown that HP1β together with SUV39H1 forms phase-separated droplets in the presence of nuclear extracts [38]. The mode of HP1β LLPS differs and requires the binding of H3K9me3 nucleosomes (Figure 6). Interestingly, HP1α and HP1γ do not phase separate under these conditions, although they have functionally similar CD domains binding H3K9me3 and a CSD for dimerization. We speculate that the interaction of their acidic IDR-N and basic IDR-H antagonizes oligomerization via histone H3K9me3 binding [39]. In any case, our study identified the net charge of the IDR-H as a critical feature controlling LLPS of HP1 *in vitro*. The observation that the simple addition of histones promotes LLPS with HP1β indicates that the situation *in vivo*, with its numerous direct and indirect interactions, is much more complex.

As diverse as the phase separating properties of HP1s are *in vitro*, so are their subcellular distribution and function *in vivo*. While HP1γ is predominantly localized in euchromatin, HP1α and HP1β are mostly associated with heterochromatin [40]. Whereas HP1α plays a central role in the formation of satellite heterochromatin, HP1β is involved in chromocenter formation by bridging H3K9me3 containing nucleosomes [29,41–43]. Interestingly, histone acetylation was recently described to drive LLPS and chromatin organization [33]. These results suggest that histone tail modifications in combination with specific reader proteins may encode the establishment of functionally distinct chromatin domains in the nucleus. The recent observation of HP1 independent formation of heterochromatin in cultured cell lines [24] serves as a reminder that there are several mechanisms that may cooperate or compete in the establishment of heterochromatin states in vivo. Future comprehensive studies are needed to dissect their relative contributions in different cell types throughout differentiation.

Acknowledgments

We thank Geeta Narlikar and Gary Karpen (University of California, USA) for their encouragement and comments. We thank Masahiro Muto and Haruhiko Koseki for mouse E14 WT ESCs. We thank Carina Trummer for her comments on the manuscript and Hartmann Harz for help with microscopy.

Disclosure statement

The authors declare no competing interests.

Funding

This work was supported by grants from the Deutsche Forschungsgemeinschaft [DFG grants SFB1064/A17 and LE 721/18-1 to H.L. and CA 198/16-1 to M.C.C.].

ORCID

M. Cristina Cardoso http://orcid.org/0000-0001-8427-8859
Heinrich Leonhardt http://orcid.org/0000-0002-5086-6449

References

[1] Bergeron-Sandoval LP, Safaee N, Michnick SW. Mechanisms and consequences of macromolecular phase separation. Cell. 2016;165:1067–1079.
[2] Boeynaems S, Alberti S, Fawzi NL, et al. Protein phase separation: a new phase in cell biology. Trends Cell Biol. 2018;28:420–435.
[3] Brangwynne CP, Eckmann CR, Courson DS, et al. Germline P granules are liquid droplets that localize by controlled dissolution/condensation. Science. 2009;324:1729–1732.
[4] Hyman AA, Weber CA, Julicher F. Liquid-liquid phase separation in biology. Annu Rev Cell Dev Biol. 2014;30:39–58.
[5] Kato M, Han TW, Xie S, et al. Cell-free formation of RNA granules: low complexity sequence domains form dynamic fibers within hydrogels. Cell. 2012;149:753–767.
[6] Wright PE, Dyson HJ. Intrinsically disordered proteins in cellular signalling and regulation. Nat Rev Mol Cell Biol. 2015;16:18–29.
[7] Alberti S, Halfmann R, King O, et al. A systematic survey identifies prions and illuminates sequence features of prionogenic proteins. Cell. 2009;137:146–158.
[8] Malinovska L, Palm S, Gibson K, et al. Dictyostelium discoideum has a highly Q/N-rich proteome and shows

an unusual resilience to protein aggregation. Proc Natl Acad Sci U S A. 2015;112:E2620–2629.

[9] Pak CW, Kosno M, Holehouse AS, et al. Sequence determinants of intracellular phase separation by complex coacervation of a disordered protein. Mol Cell. 2016;63:72–85.

[10] Wang J, Choi JM, Holehouse AS, et al. A molecular grammar governing the driving forces for phase separation of prion-like RNA binding proteins. Cell. 2018;174:688–699 e616.

[11] Aguzzi A, Altmeyer M. Phase separation: linking cellular compartmentalization to disease. Trends Cell Biol. 2016;26:547–558.

[12] Shin Y, Brangwynne CP. Liquid phase condensation in cell physiology and disease. Science. 2017;357, eaaf4382.

[13] Eissenberg JC, Elgin SC. The HP1 protein family: getting a grip on chromatin. Curr Opin Genet Dev. 2000;10: 204–210.

[14] Li Y, Kirschmann DA, Wallrath LL. Does heterochromatin protein 1 always follow code? Proc Natl Acad Sci U S A. 2002;99(Suppl 4):16462–16469.

[15] Bannister AJ, Zegerman P, Partridge JF, et al. Selective recognition of methylated lysine 9 on histone H3 by the HP1 chromo domain. Nature. 2001;410:120–124.

[16] Jacobs SA, Khorasanizadeh S. Structure of HP1 chromodomain bound to a lysine 9-methylated histone H3 tail. Science. 2002;295:2080–2083.

[17] Nakayama J, Rice JC, Strahl BD, et al. Role of histone H3 lysine 9 methylation in epigenetic control of heterochromatin assembly. Science. 2001;292:110–113.

[18] Lavigne M, Eskeland R, Azebi S, et al. Interaction of HP1 and Brg1/Brm with the globular domain of histone H3 is required for HP1-mediated repression. PLoS Genet. 2009;5:e1000769.

[19] Nielsen AL, Oulad-Abdelghani M, Ortiz JA, et al. Heterochromatin formation in mammalian cells: interaction between histones and HP1 proteins. Mol Cell. 2001;7:729–739.

[20] Larson AG, Elnatan D, Keenen MM, et al. Liquid droplet formation by HP1alpha suggests a role for phase separation in heterochromatin. Nature. 2017;547:236–240.

[21] Sanulli S, Trnka MJ, Dharmarajan V, et al. HP1 reshapes nucleosome core to promote phase separation of heterochromatin. Nature. 2019;575:390–394.

[22] Strom AR, Emelyanov AV, Mir M, et al. Phase separation drives heterochromatin domain formation. Nature. 2017;547:241–245.

[23] Erdel F, Rippe K. Formation of chromatin subcompartments by phase separation. Biophys J. 2018;114: 2262–2270.

[24] Erdel F, Rademacher A, Vlijm R, et al. Mouse heterochromatin adopts digital compaction states without showing hallmarks of HP1-driven liquid-liquid phase separation. Mol Cell. 2020;78:236–249 e237.

[25] Brero A, Easwaran HP, Nowak D, et al. Methyl CpG-binding proteins induce large-scale chromatin reorganization during terminal differentiation. J Cell Biol. 2005;169:733–743.

[26] Bertulat B, De Bonis ML, Della Ragione F, et al. MeCP2 dependent heterochromatin reorganization during neural differentiation of a novel Mecp2-deficient embryonic stem cell reporter line. PLoS One. 2012;7:e47848.

[27] Bosch-Presegue L, Raurell-Vila H, Thackray JK, et al. Mammalian HP1 isoforms have specific roles in heterochromatin structure and organization. Cell Rep. 2017;21:2048–2057.

[28] Raurell-Vila H, Bosch-Presegue L, Gonzalez J, et al. An HP1 isoform-specific feedback mechanism regulates Suv39h1 activity under stress conditions. Epigenetics. 2017;12:166–175.

[29] Hiragami-Hamada K, Soeroes S, Nikolov M, et al. Dynamic and flexible H3K9me3 bridging via HP1beta dimerization establishes a plastic state of condensed chromatin. Nat Commun. 2016;7:11310.

[30] Wongtawan T, Taylor JE, Lawson KA, et al. Histone H4K20me3 and HP1alpha are late heterochromatin markers in development, but present in undifferentiated embryonic stem cells. J Cell Sci. 2011;124:1878–1890.

[31] Mulholland CB, Smets M, Schmidtmann E, et al. A modular open platform for systematic functional studies under physiological conditions. Nucleic Acids Res. 2015;43:e112

[32] Gibson BA, Doolittle LK, Schneider MWG, et al. Organization of chromatin by intrinsic and regulated phase separation. Cell. 2019;179:470–484 e421.

[33] Shechter D, Dormann HL, Allis CD, et al. Extraction, purification and analysis of histones. Nat Protoc. 2007;2:1445–1457.

[34] Shakya A, Park S, Rana N, et al. Liquid-liquid phase separation of histone proteins in cells: role in chromatin organization. Biophys J. 2020;118:753–764.

[35] Turner AL, Watson M, Wilkins OG, et al. Highly disordered histone H1-DNA model complexes and their condensates. Proc Natl Acad Sci U S A. 2018;115:11964–11969.

[36] Muller-Ott K, Erdel F, Matveeva A, et al. Specificity, propagation, and memory of pericentric heterochromatin. Mol Syst Biol. 2014;10:746.

[37] Muller KP, Erdel F, Caudron-Herger M, et al. Multiscale analysis of dynamics and interactions of heterochromatin protein 1 by fluorescence fluctuation microscopy. Biophys J. 2009;97:2876–2885.

[38] Wang L, Gao Y, Zheng X, et al. Histone modifications regulate chromatin compartmentalization by contributing to a phase separation mechanism. Mol Cell. 2019;76:646–659 e646.

[39] Canzio D, Liao M, Naber N, et al. A conformational switch in HP1 releases auto-inhibition to drive heterochromatin assembly. Nature. 2013;496:377–381.

[40] Eberhart A, Feodorova Y, Song C, et al. Epigenetics of eu- and heterochromatin in inverted and conventional nuclei from mouse retina. Chromosome Res. 2013; 21:535–554.

[41] Machida S, Takizawa Y, Ishimaru M, et al. Structural basis of heterochromatin formation by human HP1. Mol Cell. 2018;69:385–397 e388.

[42] Maison C, Bailly D, Roche D, et al. SUMOylation promotes de novo targeting of HP1alpha to pericentric heterochromatin. Nat Genet. 2011;43:220–227.

[43] Probst AV, Okamoto I, Casanova M, et al. A strand-specific burst in transcription of pericentric satellites is required for chromocenter formation and early mouse development. Dev Cell. 2010;19:625–638.

Mulholland, C. B., Traube, F. R., **Ugur, E.**, Parsa, E., Eckl, E.-M., Schönung, M., Modic, M., Bartoschek, M. D., Stolz, P., Ryan, J., Carell, T., Leonhardt, H., & Bultmann, S. (**2020**). **Distinct and stage-specific contributions of TET1 and TET2 to stepwise cytosine oxidation in the transition from naive to primed pluripotency**. Scientific Reports, 10(1), 12066.

License: CC-BY-4.0

https://doi.org/10.1038/s41598-020-68600-3

www.nature.com/scientificreports

SCIENTIFIC REPORTS

natureresearch

Check for updates

OPEN

Distinct and stage-specific contributions of TET1 and TET2 to stepwise cytosine oxidation in the transition from naive to primed pluripotency

Christopher B. Mulholland[1], Franziska R. Traube[2], Enes Ugur[1], Edris Parsa[2], Eva-Maria Eckl[1], Maximilian Schönung[1], Miha Modic[3], Michael D. Bartoschek[1], Paul Stolz[1], Joel Ryan[1], Thomas Carell[2], Heinrich Leonhardt[1]✉ & Sebastian Bultmann[1]✉

Cytosine DNA bases can be methylated by DNA methyltransferases and subsequently oxidized by TET proteins. The resulting 5-hydroxymethylcytosine (5hmC), 5-formylcytosine (5fC), and 5-carboxylcytosine (5caC) are considered demethylation intermediates as well as stable epigenetic marks. To dissect the contributions of these cytosine modifying enzymes, we generated combinations of *Tet* knockout (KO) embryonic stem cells (ESCs) and systematically measured protein and DNA modification levels at the transition from naive to primed pluripotency. Whereas the increase of genomic 5-methylcytosine (5mC) levels during exit from pluripotency correlated with an upregulation of the de novo DNA methyltransferases DNMT3A and DNMT3B, the subsequent oxidation steps turned out to be far more complex. The strong increase of oxidized cytosine bases (5hmC, 5fC, and 5caC) was accompanied by a drop in TET2 levels, yet the analysis of KO cells suggested that TET2 is responsible for most 5fC formation. The comparison of modified cytosine and enzyme levels in *Tet* KO cells revealed distinct and differentiation-dependent contributions of TET1 and TET2 to 5hmC and 5fC formation arguing against a processive mechanism of 5mC oxidation. The apparent independent steps of 5hmC and 5fC formation suggest yet to be identified mechanisms regulating TET activity that may constitute another layer of epigenetic regulation.

DNA methylation plays critical roles in the epigenetic regulation of gene expression and genome stability in mammals[1]. During mammalian development, methylated cytosine (5mC) serves as a critical epigenetic barrier to guide cell fate decisions and restrict developmental potential[2]. Genomic 5mC patterns are established by the de novo DNA methyltransferases DNMT3A and DNMT3B and maintained through subsequent cell divisions by DNMT1[3]. The mitotic inheritance of 5mC constitutes a form of epigenetic memory enabling the long term maintenance of cell identity. Extinguishing such memory requires extensive epigenetic reprogramming and is key for the acquisition of naive pluripotency (i.e. the capacity of cells to contribute to all lineages in the embryo) during development[4]. In mammals, genome-wide erasure of 5mC accompanies the restoration of developmental potential following fertilization, reaching a nadir in the naive pluripotent inner cell mass (ICM) of the pre-implantation blastocyst[5–7]. In turn, the transition from a naive pluripotent state to one “primed” for lineage commitment upon implantation coincides with the establishment of global DNA methylation patterns[8–10].

The cellular landscape of 5mC can be altered by the inhibition of maintenance DNA methylation and/or via the action of the Ten-eleven Translocation (TET) family of dioxygenases[11]. The three mammalian

[1]Department of Biology II and Center for Integrated Protein Science Munich (CIPSM), Ludwig-Maximilians-Universität München, Planegg-Martinsried, Germany. [2]Department of Chemistry and Center for Integrated Protein Science Munich (CIPSM), Ludwig-Maximilians-Universität München, Munich, Germany. [3]Department of Neuromuscular Disease, UCL Queen Square Institute of Neurology, London, UK. ✉email: h.leonhardt@lmu.de; bultmann@bio.lmu.de

homologs, TET1, TET2, and TET3, share a conserved dioxygenase domain and catalyze the stepwise oxidation from 5mC to 5-hydroxymethylcytosine (5hmC), 5-formylcytosine (5fC), and 5-carboxylcytosine (5caC) (Fig. 1a)[12–15]. These oxidized cytosine derivatives have been described as intermediates of passive and active DNA demethylation[14,16–18], yet may also serve as stable epigenetic marks[19,20]. Moreover, their largely separate genomic distributions and reader proteins imply distinct epigenetic regulatory functions for 5hmC, 5fC, and 5caC[21,22].

TET-mediated cytosine oxidation is indispensable for mammalian development[23–26], as evidenced by the failure of TET-deficient mice to develop beyond gastrulation[25,26]. However, single *Tet* mutants exhibit less severe albeit distinct phenotypes, suggesting each enzyme can partially compensate for loss of the other[27–29]. While all TETs oxidize 5mC, the three TETs are not entirely functionally redundant. Individual TET family members exhibit distinct cellular localization patterns and genome-wide binding profiles, which appear to confer them with discrete functions during development[30–32].

Despite extensive research into the differing functions of TETs, the precise roles of the three TET proteins in the stepwise oxidation of 5mC in vivo remains to be elucidated. Clearly, the observed stable cellular levels of oxidized cytosine derivatives and their distinct genome-wide distributions seem to require dedicated regulatory mechanisms for each oxidation step[19–21,33]. Interestingly, the three TET proteins differ in their large, unstructured N-terminal domains, possibly enabling divergent contributions to stage and cell-type specific DNA modification[12]. While TET1/2/3 have all been demonstrated to mediate iterative cytosine oxidation in vitro, whether these proteins equally contribute to the levels of the three oxidized cytosine derivatives in a cellular context is unclear[13,14]. Moreover, currently available biochemical data do not conclusively resolve whether TET proteins oxidize 5mC in a chemically processive manner or in a rather distributive mode with independent steps[34–36].

Due to fast kinetics and limited material, studying the dynamics of DNA modifications during mammalian peri-implantation development remains experimentally intractable. The naive pluripotent state of the preimplantation mouse embryo can be captured and maintained in vitro by culturing murine embryonic stem cells (ESCs) in the presence of leukemia inhibitory factor (LIF) and inhibitors of MEK and GSK3 (2i)[37]. These naive ESCs feature closely similar transcriptional and epigenetic characteristics of the E3.75-E4.5 ICM from which they are derived[38], including global DNA hypomethylation[39–41]. The transition from naive to primed pluripotency accompanying peri-implantation development can be recapitulated in vitro by differentiating naive ESCs into epiblast-like cells (EpiLCs) by exposure to fibroblast growth factor 2 and Activin A. After 48 h of differentiation, EpiLCs exhibit both a transcriptional profile and genome-wide DNA hypermethylation that closely resembles that of the post-implantation pre-gastrulation epiblast (E5.75-E6.5)[10,42,43]. As such, this in vitro system offers an ideal model for uncovering basic principles of oxidized cytosine regulation.

Here, we combine quantitative proteomics and global DNA modification measurements to dissect the individual contributions of TET enzymes to cytosine modification dynamics during the transition from naive to primed pluripotency. We find that TET1 and TET2 distinctly contribute to global oxidized cytosine levels in naive ESCs as well as EpiLCs. While TET2 is required for the formation of 5hmC in the naive state, TET1 is responsible for most of the global 5hmC wave during the transition to primed pluripotency. Most notably, despite a strong downregulation during differentiation, TET2 accounts for the majority of 5fC in both stages of pluripotency.

Results

We first set out to characterize DNA modification dynamics in the naive to primed transition. To this end, we used ultra-high performance liquid chromatography coupled to tandem mass spectrometry (UHPLC-MS/MS) to quantitatively assess the levels of 5mC, 5hmC, 5fC, and 5caC in genomic DNA isolated from wild-type (wt) mouse naive ESCs and EpiLCs differentiated for 48 h (Fig. 1b). As previous studies have shown[12–15], we found that cytosine modifications become exceedingly less abundant with higher oxidation states. In ESCs 5hmC constitutes ~ 3% of modified cytosines whereas 5fC and 5caC make up only 0.02% and 0.01%, respectively. In EpiLCs, a similar distribution is observable, albeit with 5hmC and 5fC accounting for a larger fraction of modified cytosines than in ESCs (Fig. 1c).

Global DNA methylation (5mC) increased over the course of differentiation with 5mC levels in naive ESCs and EpiLCs reminiscent of those in their respective in vivo counterparts, the E3.5 ICM and E6.5 epiblast[6,8] (Fig. 1d; Supplementary Table S1). The precise quantification of cytosine derivatives demonstrated that, along with 5mC, the levels of 5hmC, 5fC, and 5caC increased from ESCs to EpiLCs (Fig. 1e–g; Supplementary Table S1). While 5mC and 5caC levels doubled, 5hmC and 5fC displayed a five- and three-fold increase, respectively. This overproportional increase of 5hmC and 5fC suggests that the oxidation of 5mC may occur in successive steps subjected to independent regulation during exit from naive pluripotency.

In search of possible mechanisms for the uncoupled levels of cytosine derivatives, we examined the protein abundance of cytosine modifying enzymes (DNMTs and TETs) during the naive to primed transition. Mass spectrometry (MS)-based quantitative proteomics showed the global wave of DNA methylation during differentiation to coincide with a substantial increase in the levels of the de novo DNA methyltransferases DNMT3A and DNMT3B (Fig. 1h; Supplementary Table S2), consistent with similar changes observed during peri-implantation development[9,44,45]. Protein levels of the ubiquitous maintenance DNA methyltransferase DNMT1 as well as its essential regulator and cofactor UHRF1 remained constant in the naive to primed transition (Fig. 1h). Despite even larger gains in oxidized cytosine levels, we did not detect corresponding increases in TET protein levels during the transition from naive ESCs to EpiLCs. On the contrary, while TET1 levels remained relatively constant, we measured a significant reduction (~ 4.5 fold) in TET2 peptides in EpiLCs and failed to detect TET3 in either cell type (Fig. 1h). These changes in TET protein levels were directly confirmed by independent Western blot analyses and reflected similar trends in *Tet* mRNA levels as determined by qPCR (Supplementary Fig. S1a,b).

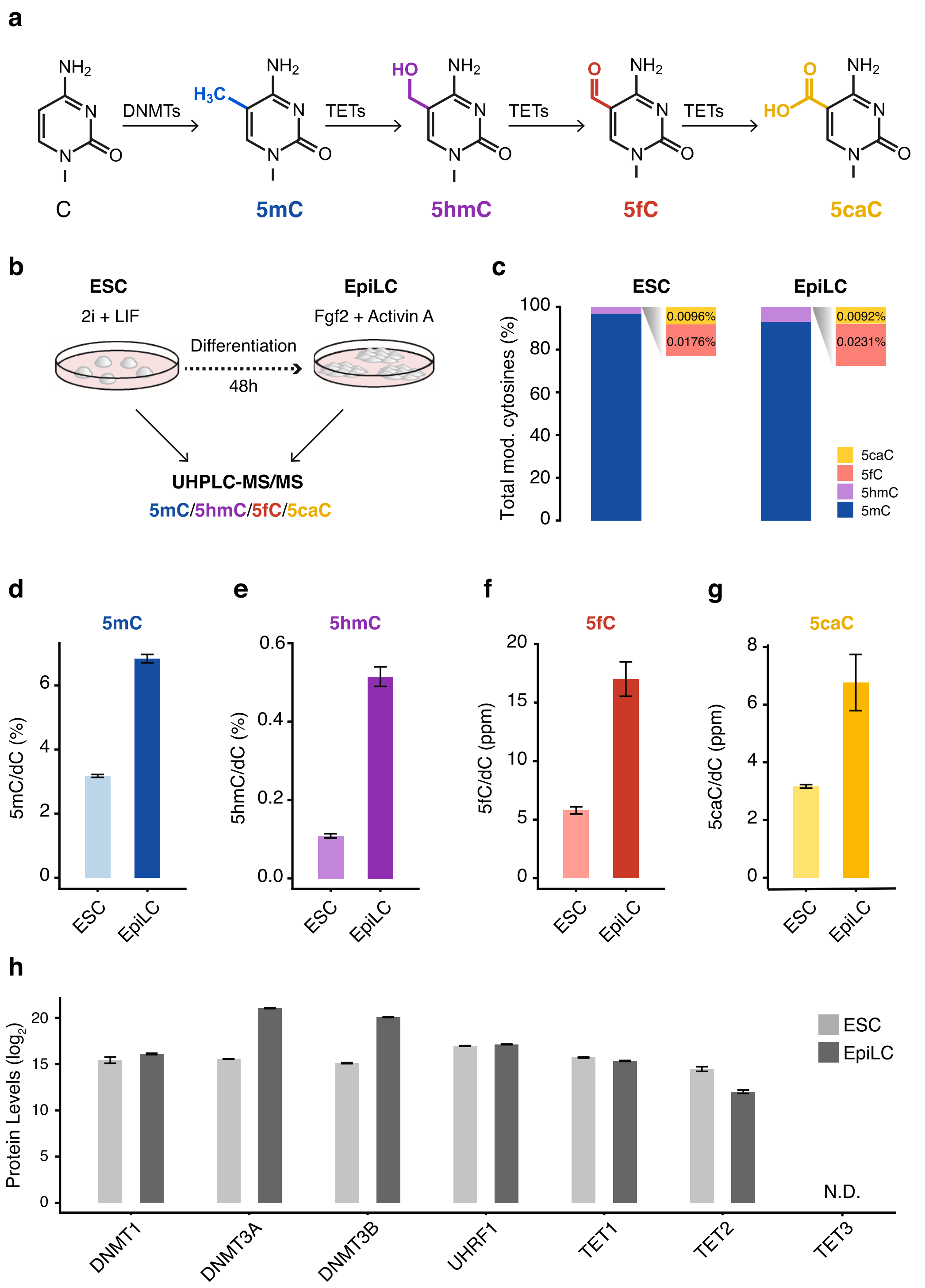


Figure 1. Global increases in cytosine modifications accompany the transition from naive to primed pluripotency. (**a**) Cytosine modifications depicted with the enzymes responsible for their generation. (**b**) Schematic overview of experimental design. DNA modifications were measured in murine naive embryonic stem cells (ESC) and epiblast-like stem cells (EpiLC) using UHPLC-MS/MS. (**c**) Abundance of genomic 5mC, 5hmC, 5fC, and 5caC in wild-type ESCs and EpiLCs shown as the fraction of total modified (mod.) cytosines. Due to their relative scarcity, 5fC and 5caC are depicted with a zoomed-in view. *n* = 6 (ESCs) and *n* = 12 (EpiLCs) biological replicates. (**d**–**g**) Global levels of (**d**) 5mC, (**e**) 5hmC, (**f**) 5fC, and (**g**) 5caC in wild-type ESCs and EpiLCs as determined by mass spectrometry (UHPLC-MS/MS). DNA modification levels are expressed as a percentage (%) or parts per million (ppm: 1 ppm = 0.0001%) of total cytosine (dC). Error bars indicate mean ± SD calculated from *n* = 6 (ESCs) and *n* = 12 (EpiLCs) biological replicates. (**h**) Protein abundance of DNA modifying enzymes in wild-type ESCs and EpiLCs as determined by LC–MS/MS-based whole proteome profiling. Shown are log2-transformed protein levels. Error bars indicate mean ± SD calculated from *n* = 3 (ESCs) and *n* = 3 (EpiLCs) biological replicates. N.D.: no peptides of protein detected.

Moreover, these data are consistent with the expression profile of TETs during in vivo peri-implantation development, where TET1 and TET2 but not TET3 are expressed[46–48].

As the overall abundance of TET family members decreases during the naive to primed transition, we considered whether the increase in oxidized cytosine derivatives might be attributable to expression changes in the Base Excision Repair (BER) pathway. Genomic 5fC and 5caC can be specifically recognized and excised by thymine DNA glycosylase (TDG), and ultimately replaced by unmodified cytosine via the BER pathway[49]. As such, the abundance of modified cytosines, especially 5fC and 5caC, in genomic DNA is subject to influence from the BER pathway[50]. However, our proteomics data from ESCs and EpiLCs indicated that levels of the BER proteins (e.g. APEX1, LIG3, PNKP, XRCC1, and PARP1) remained largely unchanged (Supplementary Fig. S1c). To assess the expression of additional BER factors undetected in our proteomics analysis, we profiled the transcriptomes of ESCs and EpiLCs using RNA-seq (Supplementary Table S3). In line with our proteomics measurements, most BER genes exhibited mostly static transcript levels in the naive to primed transition, whereas the expression of *Tdg* even increased (Supplementary Fig. S1c). These data argue against reduced removal of oxidized cytosine derivatives by the BER pathway as an explanation for the observed increase in 5hmC and 5fC levels during the naive to primed pluripotency transition. Additionally, we assessed the expression profile of factors involved in alternative base modification pathways, such as the AID/APOBEC family of cytosine deaminases. However, we failed to detect the majority of these deaminases, including AID (AICDA), in either our proteome or transcriptome data from ESCs and EpiLCs (Supplementary Fig. S1c). Together with our previous work demonstrating the deamination pathway to negligibly influence 5hmC levels in ESCs[51], these results indicate that the AID/APOBEC enzymes do not appreciably contribute to the global increase in oxidized cytosine levels in the transition from naive to primed pluripotency.

We next sought to dissect and identify the specific contributions of TET proteins to cytosine modifications during the naive to primed transition. To this end, we used CRISPR/Cas-mediated mutagenesis to generate *Tet1* and *Tet2* single knockout (KO) and *Tet1*/*Tet2* double KO (DKO) ESC lines (Supplementary Fig. S2a,b) and confirmed loss of TET1 and TET2 by Western blot analyses (Supplementary Fig. S2c,d). Using two independent clones for each genotype, we quantified the levels of 5mC, 5hmC, 5fC, and 5caC in ESCs and EpiLCs by LC–MS/MS. In parallel, we used RNA-seq and MS-based proteomics to monitor how loss of TET1 and/or TET2 affected the transcriptome and proteome of ESCs and EpiLCs (Supplementary Table S2). Elimination of either TET1 or TET2, or both TET1 and TET2 resulted in modest yet significant increases in 5mC in both naive ESCs and primed EpiLCs (Fig. 2a,b; Supplementary Table S1 and S4). The expression levels of DNMT1, DNMT3A/B, and UHRF1 in *Tet* KO ESCs and EpiLCs were similar to those in their wild-type counterparts, suggesting the 5mC gains were not a result of upregulated DNA methylating enzymes (Supplementary Fig. S3). Double *Tet1*/*Tet2* KO resulted in the loss of practically all oxidized cytosine derivatives, with levels of 5hmC, 5fC, and 5caC reduced to near or below the detection limit in ESCs and EpiLCs (Fig. 2c–f; Supplementary Fig. S4a, b; Supplementary Table S1 and S4). Together with our expression data (Fig. 1h; Supplementary Fig. S1a,b and S3) these results argue for major roles of TET1 and TET2 in 5mC oxidation during naive pluripotency exit with little to no contribution from TET3.

Analysis of the individual *Tet* KOs revealed stark, stage-specific differences in each enzyme's functional contribution to the consecutive steps of cytosine oxidation. Genomic 5hmC levels were significantly decreased in *Tet1* KO (50% of wt 5hmC) as well as *Tet2* KO (30% of wt 5hmC) ESCs demonstrating that, despite both being highly expressed, TET1 and TET2 are not redundant (Fig. 2c; Supplementary Table S4). The comparatively severe 5hmC depletion in *Tet2* KO ESCs indicates the majority of 5mC to 5hmC conversion in naive pluripotency to require TET2. Strikingly, *Tet2* KO 5hmC levels substantially increased upon exit from pluripotency, recovering from ~ 0.03% (30% of wt ESC 5hmC) to ~ 0.3% of genomic cytosines (60% of wt EpiLC 5hmC). As 5hmC increases in the absence of TET2 in *Tet2* KO EpiLCs, this suggests that the majority of 5hmC newly acquired during differentiation is generated by TET1, which remains highly expressed in EpiLCs (Fig. 2c,d, Supplementary Fig. S3, Supplementary Table S1 and S4). Supporting this notion was the finding that *Tet1* KOs fail to acquire 5hmC upon exit from naive pluripotency, with 5hmC levels remaining essentially unchanged between naive and primed pluripotency (~ 0.05% in *Tet1* KO ESCs and ~ 0.06% in *Tet1* KO EpiLCs versus ~ 0.5% of genomic cytosines in wt EpiLCs) (Fig. 2c,d and Supplementary Table S1).

Notably, EpiLCs express TET1 at levels similar to naive ESCs (Fig. 1h) and possess higher 5hmC levels (~ 0.5% versus ~ 0.1% of genomic cytosines). However, TET1, even in the absence of TET2 (in *Tet2* KO), is able to generate 60% of cellular 5hmC in EpiLCs (Fig. 2c, d; Supplementary Table S1 and S4). In other words, comparable amounts of TET1 produce ten-times more 5hmC in EpiLCs versus ESCs (~ 0.3% versus ~ 0.03% of genomic cytosines in *Tet2* KOs). Taken together, TET1 and TET2 possess distinct, stage-specific roles in the oxidation of 5mC, in which the responsibility of 5hmC formation passes from TET2 to TET1 upon differentiation.

To investigate whether similar stage-dependent preferences apply for the subsequent oxidation step, i.e. the conversion of 5hmC to 5fC, we compared 5fC levels in ESCs and EpiLCs. Analysis of 5fC levels in KO lines revealed an unexpectedly prominent role of TET2 in ESCs and even EpiLCs (Fig. 2e,f; Supplementary Table S1 and S4). In naive ESCs, *Tet2* KO caused an ~ 87% reduction in 5fC levels, almost reaching the background levels of the *Tet1*/*Tet2* DKO, whereas only 50% of 5fC was lost in *Tet1* KO ESCs (Fig. 2e). As the reduction of 5fC in *Tet1* KO ESCs was proportional to the loss of its precursor, 5hmC, the overall 5fC/5hmC ratio remained similar to that of wild-type ESCs (Fig. 2g, i). In striking contrast, the large reduction of 5fC in naive *Tet2* KO ESCs did not correlate with a decrease in 5hmC, with TET2 loss leading to a much lower ratio of 5fC/5hmC than in wt or *Tet1* KO ESCs (Fig. 2g, i). Thus, TET2 is required for the majority of global cytosine oxidation in naive pluripotency, with TET1 unable to compensate for TET2 loss in naive ESCs.

In EpiLCs, 5fC levels dropped to ~ 18% and ~ 26% of their wt levels in *Tet1* KO and *Tet2* KO cells, respectively (Fig. 2f). The similarity of 5fC levels in both, *Tet1* and *Tet2* KO EpiLCs stands in stark contrast to their 5hmC levels (Fig. 2d). As in naive ESCs, the stark reduction of 5fC in *Tet1* KO EpiLCs was accompanied by a strong

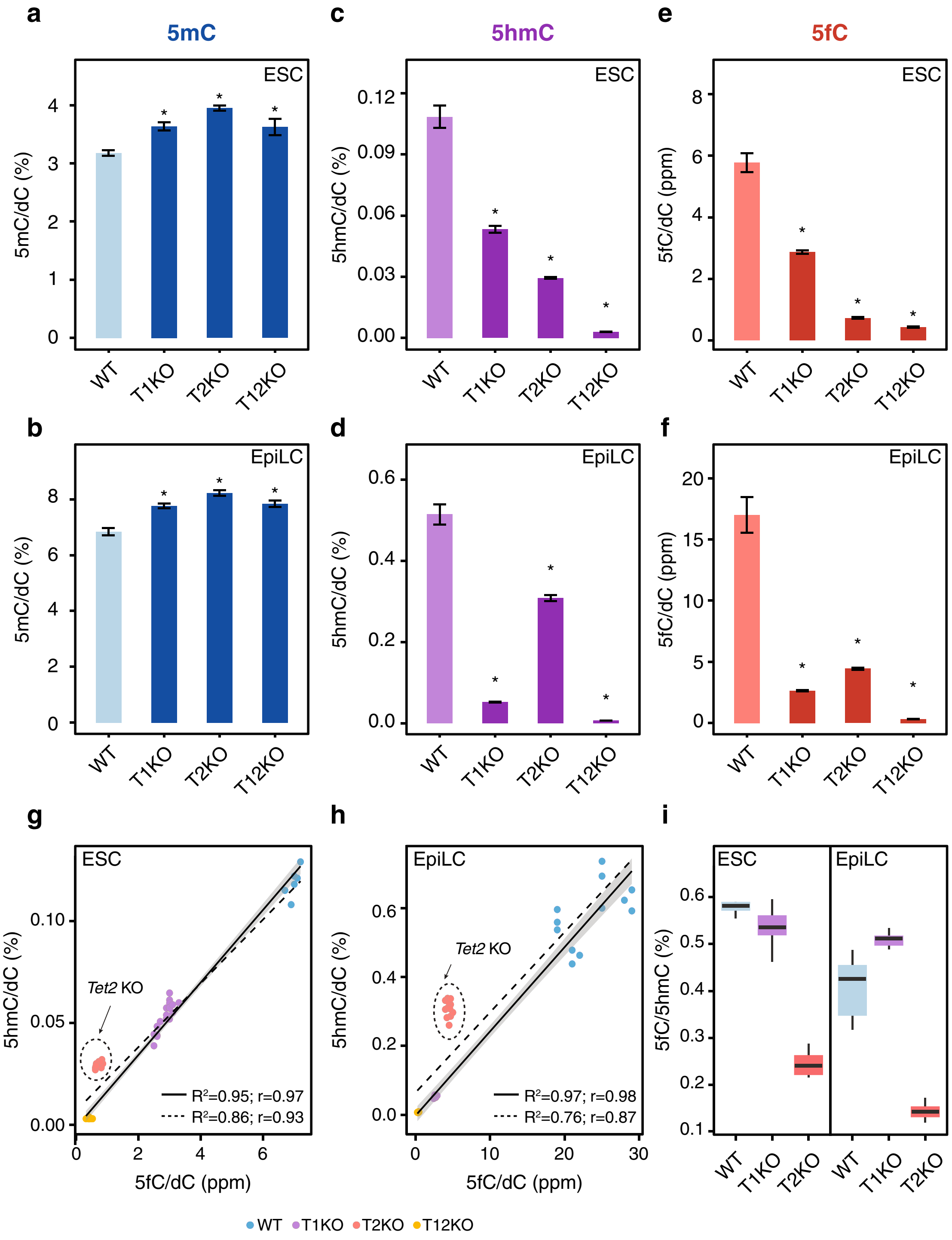


Figure 2. Quantification of cytosine modifications in *Tet1* and *Tet2* knockout ESCs and EpiLCs. (**a**–**f**) Global levels of (**a**, **b**) 5mC, (**c**, **d**) 5hmC, and (**e**, **f**) 5fC, in wild-type (WT), *Tet1* KO (T1KO), *Tet2* KO (T2KO), and *Tet1/Tet2* DKO (T12KO) ESCs and EpiLCs as determined by mass spectrometry (UHPLC-MS/MS). DNA modification levels are expressed as a percentage (%) or parts per million (ppm: 1 ppm = 0.0001%) of total cytosine (dC) and shown as the mean ± SD of biological replicates as follows: WT (ESCs: *n* = 18; EpiLCs: *n* = 24), T1KO (ESCs: *n* = 18; EpiLCs: *n* = 12), T2KO (ESCs: *n* = 12; EpiLCs: *n* = 12), and T12KO (ESCs: *n* = 12; EpiLCs: *n* = 12). * $p < 0.005$ to wt as determined using a one-way ANOVA followed by a post-hoc Tukey HSD test. (**g**–**h**) Correlations between 5hmC and 5fC levels in wt and *Tet* KO (**g**) ESCs and (**h**) EpiLCs. The dashed regression line was generated using the full data set, the solid regression line was generated by excluding *Tet2* KO data. Depicted are values from the individual replicates presented in **c**–**f**. R^2: coefficient of determination; r: Pearson correlation coefficient. (**i**) Box plots of the ratio of 5fC to 5hmC in wt, *Tet1* KO and *Tet2* KO ESCs and EpiLCs. Unlike the *Tet1* KO, *Tet2* KO drastically affects the 5fC/5hmC ratio. The median is represented by the central bold line. The lower and upper hinges correspond to the first and third quartiles (the 25th and 75th percentiles). The upper and lower whisker extend from the hinge to the largest and lowest value, respectively, no further than 1.5 * interquartile range (IQR).

decrease in 5hmC. However, the loss of TET2 in EpiLCs led to a disproportionate decrease in 5fC compared to 5hmC (Fig. 2g–i). The significant global depletion of 5fC resulting from TET2 loss in EpiLCs is particularly striking considering that TET2 is expressed at lower levels at this particular stage compared to ESCs (Fig. 1h, Supplementary Fig. S1a).

As 5fC can be excised from DNA by the BER pathway, we investigated whether the decrease in 5fC in *Tet2* KOs might be an indirect consequence resulting from upregulation of DNA repair enzymes upon TET2 loss. We assessed the expression levels of BER pathway proteins by RNA-seq and full proteome mass spectrometry at both time points (Supplementary Fig. S3). Neither the loss of TET2 nor TET1 significantly affected the expression of these genes in ESCs or EpiLCs. Therefore, the decrease of 5fC in *Tet2* KOs appears to be a direct effect of TET2 loss. The disproportionate loss of 5fC in both stages, naive and primed, reveals a previously unappreciated prominence of TET2 in the formation of 5fC in pluripotent stem cells.

Due to the extremely low abundance of 5caC in comparison to the other cytosine modifications (Fig. 1c,g), loss of TET activity resulted in levels below the detection limit (Supplementary Fig. S4). We were only able to clearly detect 5caC in wt and *Tet2* KO EpiLCs, but not *Tet1* KO EpiLCs, suggesting that the more abundant TET1 is responsible for most 5caC formation in EpiLCs.

Discussion

In summary, the systematic quantification of cytosine derivatives and their respective enzymes in this defined cellular differentiation system leads to a number of unexpected findings (Fig. 3). Whereas the increase of 5mC during naive pluripotency exit correlated with the growing abundance of the de novo DNA methyltransferases, DNMT3A and DNMT3B, the rising levels of oxidized cytosine derivatives, 5hmC and 5fC, were accompanied by stable TET1 and diminishing TET2 levels. In these cells, TET3 seems to play little to no role given its undetectable expression and the practically complete loss of genomic 5hmC, 5fC, and 5caC in cells lacking TET1 and TET2.

Our analysis of global cytosine modification levels in *Tet1* and *Tet2* KO ESCs and EpiLCs revealed both enzymes to have profound stage-specific contributions to cytosine oxidation, which cannot be fully compensated by the other enzyme. In ESCs, the oxidation of 5mC to 5hmC relies primarily on TET2, whereas the global increase in 5hmC during differentiation is almost exclusively catalyzed by TET1. Thus, the distinct, stage-specific contributions of TET1 and TET2 to 5hmC generation might underlie their opposing roles in controlling the transition between naive and primed pluripotency[52].

The previously observed downregulation had argued against any role of TET2 in peri-implantation development[28,53]. We also observed downregulation of *Tet2* expression in EpiLCs but still detected TET2 protein by mass spectrometry and Western blot analysis. In fact, our KO data identified a rather distinct role of TET2 in naive and primed pluripotency. Remarkably, *Tet2* KO ESCs and EpiLCs show an unexpected loss of 5fC, arguing that TET2 governs the formation of 5fC in ESCs as well as the increase of 5fC during the naive to primed transition. Our KO data clearly demonstrate that the residual amounts of TET2 proteins in EpiLCs have a prominent role in the oxidation of 5hmC to 5fC, which cannot be compensated by the much more abundant TET1.

Detailed analysis of the different KO lines also showed that 5hmC and 5fC levels respond independently. Since TET1 can rescue the majority of 5hmC but not 5fC upon loss of TET2 (especially in EpiLCs), we propose that in vivo stepwise oxidation largely follows a distributive model in line with previous in vitro findings[35,36] with the caveat that the in vivo distributive oxidation is shaped by an additional layer of regulation, one in which different TET paralogs preferentially catalyze separate steps. Our results suggest TET1 preferentially oxidizes 5mC to 5hmC, then dissociates, leaving the subsequent oxidation step of 5hmC to 5fC to be catalyzed by TET2. In line with this hypothesis, a similar division between TET1 and TET2 activities has been described for SALL4A-bound enhancers[32].

In addition to the apparently differing substrate proclivities of TET1 and TET2, we observe a differentiation-dependent, dynamic regulation of oxidative potential, especially for TET1. Despite maintaining comparable protein levels in the naive to primed transition, TET1 drives the differentiation-dependent quintupling of 5hmC almost exclusively and independent of TET2, yet can only contribute to 30% of 5hmC in naive ESCs.

Thus, not only the underlying mechanisms regulating the individual TET-specific contribution to distributive oxidation deserve further investigation, but also those controlling the dynamics of substrate oxidation. It remains to be seen to what extent modulation of the catalytic activity of the three TET enzymes by differential isoform expression, posttranslational modifications, interacting factors, and site-specific recruitment could constitute an additional layer of epigenetic regulation. Interestingly, 5fC was revealed to possess novel characteristics, such as the ability to distort the DNA double helix[54] and directly mediate DNA–protein crosslinks[55,56], with potentially far reaching consequences on transcriptional regulation and chromatin remodeling[57,58]. In this context, our observation that 5fC formation appears to be largely TET2-dependent might also have novel implications for understanding how *Tet2* mutations contribute to cancerogenesis.

Methods

Cell culture. Naive J1 mESCs were cultured and differentiated into EpiLCs as described previously[59,60]. In brief, for both naive ESCs and EpiLCs defined media was used, consisting of: N2B27 (50% neurobasal medium (Life Technologies), 50% DMEM/F12 (Life Technologies)), 2 mM L-glutamine (Life Technologies), 0.1 mM β-mercaptoethanol (Life Technologies), N2 supplement (Life Technologies), B27 serum-free supplement (Life Technologies), and 100 U/mL penicillin, 100 µg/mL streptomycin (Sigma). Naive ESCs were maintained on flasks treated with 0.2% gelatin in defined media containing 2i (1 µM PD032591 and 3 µM CHIR99021 (Axon Medchem, Netherlands)), 1,000 U/mL recombinant leukemia inhibitory factor (LIF, Millipore), and 0.3% BSA (Gibco) for at least three passages before commencing differentiation.

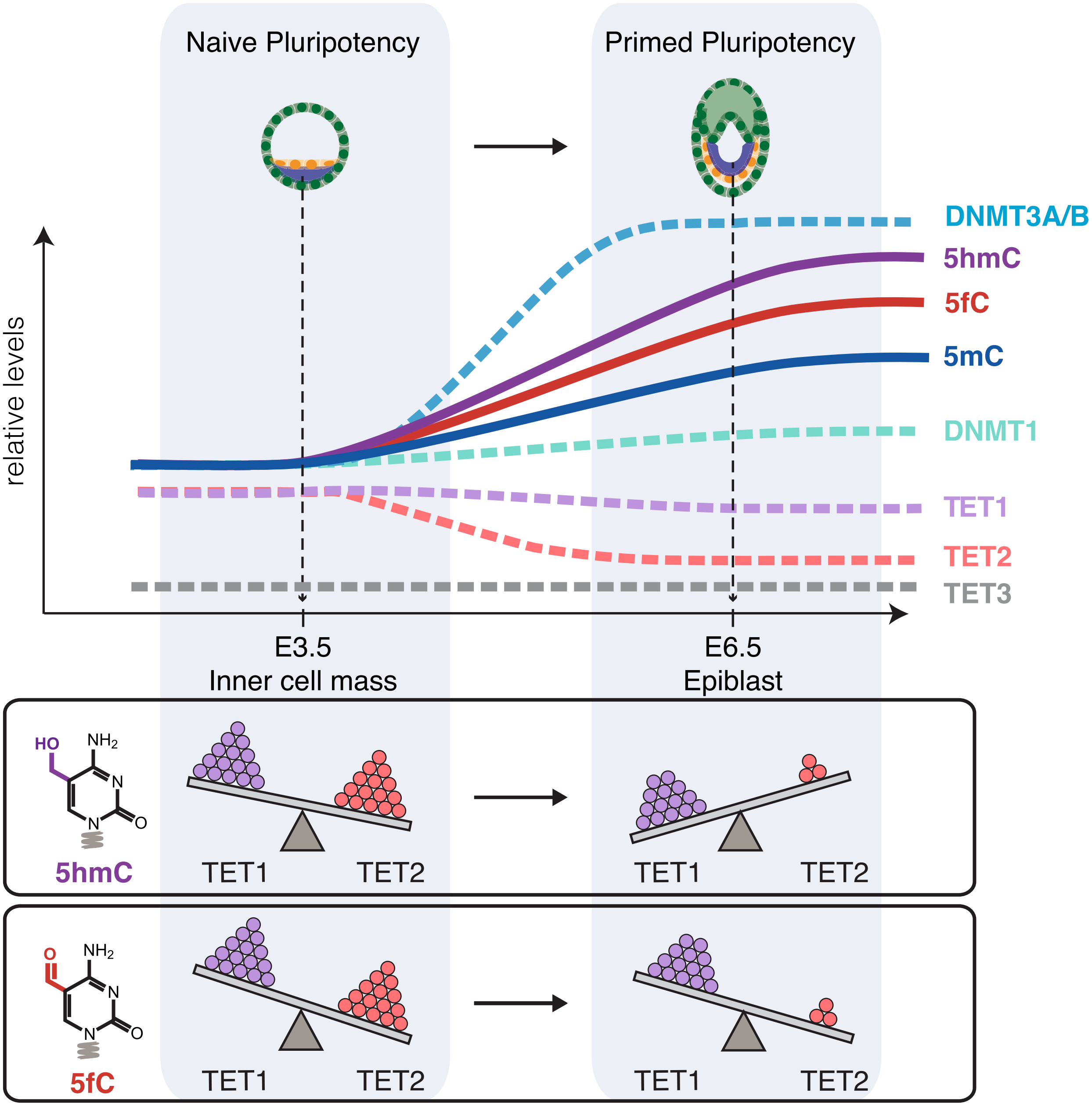


Figure 3. Epigenetic changes and distinct contributions of different DNA modifying enzymes during the transition from naive to primed pluripotency. Graphical summary depicting changes in cellular levels of cytosine modifications and their respective DNA modifying enzymes in the transition from naive to primed pluripotency. The relative contributions of TET1 and TET2 to the generation of 5hmC and 5fC as estimated from observations in *Tet* KO ESCs and EpiLCs are illustrated at the bottom; the number of spheres and tilt of the balance represent the protein abundance of each TET and the contribution of each TET to the levels of the depicted cytosine derivative, respectively. TET1 gains importance in the oxidation of 5mC to 5hmC during differentiation as TET2 abundance decreases. Most remarkably, despite drastic downregulation TET2 remains critical for the formation of 5fC in primed pluripotency.

For CRISPR-assisted cell line generation mESCs were maintained on 0.2% gelatin-coated dishes in Dulbecco's modified Eagle's medium (Sigma) supplemented with 16% fetal bovine serum (FBS, Biochrom), 0.1 mM ß-mercaptoethanol (Invitrogen), 2 mM L-glutamine (Sigma), 1 × MEM Non-essential amino acids (Sigma), 100 U/mL penicillin, 100 µg/mL streptomycin (Sigma), homemade recombinant LIF tested for efficient self-renewal maintenance, and 2i (1 µM PD032591 and 3 µM CHIR99021 (Axon Medchem, Netherlands)).

To differentiate naive ESCs into epiblast-like cells, cells were plated on flasks treated with Geltrex (Life Technologies) diluted 1:100 in DMEM/F12 (Life Technologies) in defined medium containing 10 ng/mL Fgf2 (R&D Systems), 20 ng/mL Activin A (R&D Systems) and 0.1 × Knockout Serum Replacement (KSR) (Life Technologies). Media was changed after 24 h and EpiLCs were harvested after 48 h.

Cells were regularly tested for Mycoplasma contamination by PCR.

CRISPR/Cas-mediated gene knockout and Western blot. For the generation of *Tet1* and *Tet2* knockouts, *Tet1* and *Tet2*-specific gRNAs (Supplementary Table S3) were cloned into puromycin-selectable vector expressing both SpCas9 and gRNA (px459: F. Zhang Lab). mESCs were transfected with Cas9-gRNA vectors using Lipofectamine 3000 (Invitrogen) according to manufacturer's protocol. Two days after transfection, J1 mESCs were plated at clonal density in ESC media supplemented with 1 µg/mL puromycin (Gibco). Selection media was removed after 48 h, replaced with normal ESC media, and colonies were allowed to grow for an additional 4–5 days. Single ESC colonies were transferred into 96-well plates and the plates were duplicated after 2 days. Enrichment for mutated clones was accomplished by amplifying the CRISPR/Cas targeted region via PCR (oligonucleotides in Supplementary Table S5) and performing restriction-fragment length polymorphism (RFLP) analysis[61] with SacI or EcoRV (FastDigest; Thermo Scientific) for *Tet1* or *Tet2*, respectively (see also Supplementary Fig. S2a). Cell lysis in 96-well plates, PCR on lysates, and restriction digest were performed as previously described[60].

Clones harboring biallelic mutations were then assessed for loss of TET1 or TET2 via Western blot. Western blots for both *Tet* KOs were performed as described previously[60] using monoclonal antibodies (rat anti-TET1 5D6, rat anti-TET2 9F7, and rat anti-TET3 23B9)[62] and polyclonal rabbit anti-H3 (ab1791, Abcam) as loading control. Blots were probed with secondary antibodies anti-rat (112-035-068, Jackson ImmunoResearch) and anti-rabbit (170–6515, Bio-Rad) conjugated to horseradish peroxidase (HRP) and visualized using an ECL detection kit (Thermo Scientific Pierce).

Quantitative real-time PCR (qRT-PCR) Analysis. Total RNA was isolated using the NucleoSpin Triprep Kit (Macherey-Nagel) according to the manufacturer's instructions. cDNA synthesis was performed with the High-Capacity cDNA Reverse Transcription Kit (with RNase Inhibitor; Applied Biosystems) using 500 ng of total RNA as input. Oligonucleotides used in qRT-PCR assays are listed in Supplementary Table S5 were performed in 10 µL reactions with 5 ng of cDNA used as input. For TaqMan and SYBR green detection, TaqMan Universal Mastermix (Applied Biosystems) and FastStart Universal SYBR Green Master Mix (Roche) were used, respectively. The reactions were run on a LightCycler480 (Roche).

RNA-seq. Digital gene expression libraries for RNA-seq were prepared using the single-cell RNA barcoding sequencing (SCRB-seq) method as described previously[63–65], with minor modifications to accommodate bulk cell populations. In brief, RNA was extracted and purified from ~ 1×10^6 cells using the NucleoSpin Triprep Kit (Machery-Nagel) according to the manufacturer's instructions. In the initial cDNA synthesis step, purified, bulk RNA (70 ng) from individual samples were subjected to reverse transcription in 10 µL reactions containing 25 units of Maxima H Minus reverse transcriptase (ThemoFisher Scientific), 1 × Maxima RT Buffer (ThemoFisher Scientific), 1 mM dNTPs (ThermoFisher Scientific), 1 µM oligo-dT primer with a sample-specific barcode (IDT), and 1 µM template-switching oligo (IDT). Reverse transcription reactions were incubated 90 min at 42 °C. Next, the barcoded cDNAs from individual samples were pooled together and then purified using the DNA Clean & Concentrator-5 Kit (Zymo Research) according to the manufacturer's instructions. Purified pooled cDNA was eluted in 18 µL DNase/RNase-Free Distilled Water (Thermo Fisher) and then, to remove residual primers, incubated with 1 µL Exonuclease I Buffer (NEB) and 1 µL Exonuclease I (NEB) (final reaction volume: 20 µL) at 37 C for 30 min followed by heat-inactivation at 80 C for 20 min. Full-length cDNA was then amplified via PCR using KAPA HiFi HotStart ReadyMix (KAPA Biosystems) and SINGV6 primer (IDT). The pre-amplification PCR was performed using the following conditions: 3 min at 98 °C for initial denaturation, 10 cycles of 15 s at 98 °C, 30 s at 65 °C, and 6 min at 68 °C, followed by 10 min at 72 °C for final elongation. After purification using CleanPCR SPRI beads (CleanNA), the pre-amplified cDNA pool concentration was quantified using the Quant-iT PicoGreen dsDNA Assay Kit (Thermo Fisher). A Bioanalzyer run using the High-sensitivity DNA Kit (Agilent Technologies) was then performed to confirm the concentration and assess the size distribution of the amplified cDNA pool (Agilent Technologies). Next, 0.8 ng of the pure, amplified cDNA pool was used as input for generating a Nextera XT DNA library (Illumina) following the Manufacturer's instructions with the exception that a custom P5 primer (P5NEXTPT5) (IDT) was used to preferentially enrich for 3′ cDNA ends in the final Nextera XT Indexing PCR[63–65]. After an initial purification step using a 1:1 ratio of CleanPCR SPRI beads (CleanNA), the amplified Nextera XT Library the 300–800 bp range of the library was size-selected using a 2% E-Gel Agarose EX Gels (Life Technologies) and then extracted from the gel using the MinElute Gel Extraction Kit (Qiagen, Cat. No. 28606) according to manufacturer's recommendations. The final concentration, size distribution, and quality of Nextera XT library were assessed with a Bioanalyzer (Agilent Technologies) using a High-sensitivity DNA Kit (Agilent Technologies). The Nextera XT RNA-seq library was paired-end sequenced using a high output flow cell on an Illumina HiSeq 1500. In read 1, sample-specific barcodes were obtained by sequencing the first 16 bases, while the sequence of the cDNA fragment was obtained by the 50 bases in read 2. An additional 8 base i7 barcode read was performed to distinguish the library from others sequenced in parallel on the same flow cell.

RNA-seq processing and analysis. Raw RNA-seq data was processed and mapped to the mouse genome (mm10) using the zUMIs pipeline[66]. Gene annotations were obtained from Ensembl (GRCh38.84 or GRCm38.75). UMI count tables were filtered for low counts using HTSFilter[67]. Differential expression analysis was performed in R using DESeq2[68] and genes with an adjusted $P < 0.05$ were considered to be differentially expressed.

UHPLC-MS/MS analysis of DNA samples. *DNA digestion.* Isolation of genomic DNA was performed according to earlier published work[51].

1.0–5 µg of genomic DNA in 35 µL H_2O were digested as follows: an aqueous solution (7.5 µL) of 480 µM $ZnSO_4$, containing 18.4 U nuclease S1 (Aspergillus oryzae, Sigma-Aldrich), 5 U Antarctic phosphatase (New England BioLabs) and labeled internal standards were added ($[^{15}N_2]$-cadC 0.04301 pmol, $[^{15}N_2,D_2]$-hmdC 7.7 pmol, $[D_3]$-mdC 51.0 pmol, $[^{15}N_5]$-8-oxo-dG 0.109 pmol, $[^{15}N_2]$-fdC 0.04557 pmol) and the mixture was incubated at 37 °C for 3 h. After addition of 7.5 µl of a 520 µM $[Na]_2$-EDTA solution, containing 0.2 U snake venom phosphodiesterase I (Crotalus adamanteus, USB corporation), the sample was incubated for 3 h at 37 °C and then stored at − 20 °C. Prior to LC/MS/MS analysis, samples were filtered by using an AcroPrep Advance 96 filter plate 0.2 µm Supor (Pall Life Sciences).

UHPLC-MS/MS analysis. Quantitative UHPLC-MS/MS analysis of digested DNA samples was performed using an Agilent 1290 UHPLC system equipped with a UV detector and an Agilent 6490 triple quadrupole mass spectrometer. Natural nucleosides were quantified with the stable isotope dilution technique. An improved method, based on earlier published work[51,69] was developed, which allowed the concurrent analysis of all nucleosides in one single analytical run. The source-dependent parameters were as follows: gas temperature 80 °C, gas flow 15 L/min (N_2), nebulizer 30 psi, sheath gas heater 275 °C, sheath gas flow 15 L/min (N_2), capillary voltage 2,500 V in the positive ion mode, capillary voltage − 2,250 V in the negative ion mode and nozzle voltage 500 V. The fragmentor voltage was 380 V/ 250 V. Delta EMV was set to 500 V for the positive mode. Compound-dependent parameters are summarized in Supplementary Table S6. Chromatography was performed by a Poroshell 120 SB-C8 column (Agilent, 2.7 µm, 2.1 mm × 150 mm) at 35 °C using a gradient of water and MeCN, each containing 0.0085% (v/v) formic acid, at a flow rate of 0.35 mL/min: 0 → 4 min; 0 → 3.5% (v/v) MeCN; 4 → 6.9 min; 3.5 → 5% MeCN; 6.9 → 7.2 min; 5 → 80% MeCN; 7.2 → 10.5 min; 80% MeCN; 10.5 → 11.3 min; 80 → 0% MeCN; 11.3 → 14 min; 0% MeCN. The effluent up to 1.5 min and after 9 min was diverted to waste by a Valco valve. The autosampler was cooled to 4 °C. The injection volume amounted to 39 µL. Data were processed according to earlier published work[51].

MS-based quantitative proteomics. *Full proteome sample preparation.* For full proteome measurements flash-frozen cells were lysed in 200 µL of the lysis buffer (6 M Guanidinium Chloride, 100 mM Tris–HCl pH 8.5 and freshly added 2 mM DTT). By thoroughly pipetting, samples were homogenized and subsequently boiled at 99 °C for 10 min in a thermal shaker at 1,700 rpm. To get rid of bubbles and to collect the evaporated liquid, samples were quickly spun down. After sonication for 15 min (30 s on/off interval, Bioruptor Plus by Diagenode) protein concentrations were estimated by a BCA assay in a TECAN reader. Chloroacetamide was added to the samples (40 mM final concentration) and samples were incubated at room temperature for 20 min. For the protein digestion, 30 µg of the lysate was diluted in 30 µL of the lysis buffer already including 2 mM DTT and 40 mM CAA. Then, samples were diluted 1:10 in the digestion buffer (25 mM Tris–HCl pH 8.5 and 10% acetonitrile) containing trypsin and LysC in a 1:50 protease to protein ratio. Digestion was performed overnight at 37 °C and 100 rpm. After acidifying samples with 1% trifluoroacetic acid (TFA), peptides were cleaned up on three layers of SDB-RPS matrix[70]. Eluted and speedvac dried peptides were resuspended in 20 µL of A* buffer (0.1% TFA and 2% acetonitrile) and peptide concentrations were estimated by nanodrop at 280 nm.

Full proteome measurements based on data-independent acquisition method. Mass spectrometric analysis of peptides was performed on a quadrupole Orbitrap mass spectrometer (Q Exactive HF-X, ThermoFisher Scientific, Bremen, Germany) after prior nanoflow liquid chromatography on an Easy-nLC 1200 (ThermoFisher Scientific). The injection was mediated under high-pressure conditions via a nano-electrospray ion source. For this purpose, in-house packed 50 cm columns of ReproSil-Pur C18-AQ 1.9-µm resin (Dr. Maisch GmbH) were used to elute approximately 400 ng peptides of each sample in an acetonitrile gradient for 120 min. The flow rate was kept constantly at around 300 nL/min and the column oven temperature at 60 °C.

The peptides were analyzed following a data-independent acquisition (DIA) method (MS1 scan: resolution 60,000, 300 to 1,650 m/z, maximum injection time 60 ms and AGC target 3E6, MS2 scan: resolution 30,000, 32 segments at varying isolation windows ranging from 14.4 m/z to 562.8 m/z, maximum injection time 54 ms and AGC target 3E6). For MS2 scans the default charge state was set to 2. The stepped normalized collision energy was set to 25, 27.5 and 30.

Processing of DIA data. The DIA raw files were analyzed with the Spectronaut Pulsar X software package (Biognosys, version 13.15.200430.43655) applying the default Biognosys factory settings for DIA analysis. To get a deeper proteome a hybrid spectral library strategy[71] was followed using the DIA measurements as a project-specific library harboring 55,697 precursors (4,108 protein groups) and an ESC/EpiLC-specific Data-dependent acquisition (DDA) library with in total 230,581 precursors and 9,158 protein groups.

Data availability

Full proteome data generated in this study can be found in Supplementary Table S2. RNA-seq data generated in this study are available under the accession number E-MTAB-6797 at ArrayExpress https://www.ebi.ac.uk/arrayexpress/ and Supplementary Table S3.

Received: 27 January 2020; Accepted: 29 June 2020
Published online: 21 July 2020

References

1. Smith, Z. D. & Meissner, A. DNA methylation: Roles in mammalian development. *Nat. Rev. Genet.* **14**, 204–220 (2013).
2. Seisenberger, S. *et al.* Reprogramming DNA methylation in the mammalian life cycle: Building and breaking epigenetic barriers. *Philos. Trans. R. Soc. Lond. B Biol Sci.* **368**, 20110330 (2013).
3. Greenberg, M. V. C. & Bourc'his, D. The diverse roles of DNA methylation in mammalian development and disease. *Nat. Rev. Mol. Cell Biol.* https://doi.org/10.1038/s41580-019-0159-6 (2019).
4. Lee, H. J., Hore, T. A. & Reik, W. Reprogramming the methylome: erasing memory and creating diversity. *Cell Stem Cell* **14**, 710–719 (2014).
5. Monk, M., Boubelik, M. & Lehnert, S. Temporal and regional changes in DNA methylation in the embryonic, extraembryonic and germ cell lineages during mouse embryo development. *Development* **99**, 371–382 (1987).
6. Smith, Z. D. *et al.* A unique regulatory phase of DNA methylation in the early mammalian embryo. *Nature* **484**, 339–344 (2012).
7. Wang, L. *et al.* Programming and inheritance of parental DNA methylomes in mammals. *Cell* **157**, 979–991 (2014).
8. Borgel, J. *et al.* Targets and dynamics of promoter DNA methylation during early mouse development. *Nat. Genet.* **42**, 1093–1100 (2010).
9. Auclair, G., Guibert, S., Bender, A. & Weber, M. Ontogeny of CpG island methylation and specificity of DNMT3 methyltransferases during embryonic development in the mouse. *Genome Biol.* **15**, 545 (2014).
10. Seisenberger, S. *et al.* The dynamics of genome-wide DNA methylation reprogramming in mouse primordial germ cells. *Mol. Cell* **48**, 849–862 (2012).
11. Tahiliani, M. *et al.* Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1. *Science* **324**, 930–935 (2009).
12. Iyer, L. M., Tahiliani, M., Rao, A. & Aravind, L. Prediction of novel families of enzymes involved in oxidative and other complex modifications of bases in nucleic acids. *Cell Cycle* **8**, 1698–1710 (2009).
13. Ito, S. *et al.* Tet proteins can convert 5-methylcytosine to 5-formylcytosine and 5-carboxylcytosine. *Science* **333**, 1300–1303 (2011).
14. He, Y.-F. *et al.* Tet-mediated formation of 5-carboxylcytosine and its excision by TDG in mammalian DNA. *Science* **333**, 1303–1307 (2011).
15. Pfaffeneder, T. *et al.* The discovery of 5-formylcytosine in embryonic stem cell DNA. *Angew. Chem. Int. Ed Engl.* **50**, 7008–7012 (2011).
16. Maiti, A. & Drohat, A. C. Thymine DNA glycosylase can rapidly excise 5-formylcytosine and 5-carboxylcytosine: Potential implications for active demethylation of CpG sites. *J. Biol. Chem.* **286**, 35334–35338 (2011).
17. Hashimoto, H. *et al.* Recognition and potential mechanisms for replication and erasure of cytosine hydroxymethylation. *Nucleic Acids Res.* **40**, 4841–4849 (2012).
18. Otani, J. *et al.* Cell cycle-dependent turnover of 5-hydroxymethyl cytosine in mouse embryonic stem cells. *PLoS ONE* **8**, e82961 (2013).
19. Bachman, M. *et al.* 5-Hydroxymethylcytosine is a predominantly stable DNA modification. *Nat. Chem.* **6**, 1049–1055 (2014).
20. Bachman, M. *et al.* 5-Formylcytosine can be a stable DNA modification in mammals. *Nat. Chem. Biol.* **11**, 555–557 (2015).
21. Wu, H., Wu, X., Shen, L. & Zhang, Y. Single-base resolution analysis of active DNA demethylation using methylase-assisted bisulfite sequencing. *Nat. Biotechnol.* **32**, 1231–1240 (2014).
22. Spruijt, C. G. *et al.* Dynamic readers for 5-(hydroxy)methylcytosine and its oxidized derivatives. *Cell* **152**, 1146–1159 (2013).
23. Dawlaty, M. M. *et al.* Loss of Tet enzymes compromises proper differentiation of embryonic stem cells. *Dev. Cell* **29**, 102–111 (2014).
24. Kang, J. *et al.* Simultaneous deletion of the methylcytosine oxidases Tet1 and Tet3 increases transcriptome variability in early embryogenesis. *Proc. Natl. Acad. Sci. USA* **112**, E4236–E4245 (2015).
25. Dai, H.-Q. *et al.* TET-mediated DNA demethylation controls gastrulation by regulating Lefty-Nodal signalling. *Nature* **538**, 528 (2016).
26. Li, X. *et al.* Tet proteins influence the balance between neuroectodermal and mesodermal fate choice by inhibiting Wnt signaling. *Proc. Natl. Acad. Sci. USA* **113**, E8267–E8276 (2016).
27. Dawlaty, M. M. *et al.* Tet1 is dispensable for maintaining pluripotency and its loss is compatible with embryonic and postnatal development. *Cell Stem Cell* **9**, 166–175 (2011).
28. Khoueiry, R. *et al.* Lineage-specific functions of TET1 in the postimplantation mouse embryo. *Nat. Genet.* https://doi.org/10.1038/ng.3868 (2017).
29. Moran-Crusio, K. *et al.* Tet2 loss leads to increased hematopoietic stem cell self-renewal and myeloid transformation. *Cancer Cell* **20**, 11–24 (2011).
30. Zhang, Q. *et al.* Differential regulation of the ten-eleven translocation (TET) family of dioxygenases by O-linked β-N-acetylglucosamine transferase (OGT). *J. Biol. Chem.* **289**, 5986–5996 (2014).
31. Huang, Y. *et al.* Distinct roles of the methylcytosine oxidases Tet1 and Tet2 in mouse embryonic stem cells. *Proc. Natl. Acad. Sci. USA* **111**, 1361–1366 (2014).
32. Xiong, J. *et al.* Cooperative action between SALL4A and TET proteins in stepwise oxidation of 5-methylcytosine. *Mol. Cell* **64**, 913–925 (2016).
33. Shen, L. *et al.* Genome-wide analysis reveals TET- and TDG-dependent 5-methylcytosine oxidation dynamics. *Cell* **153**, 692–706 (2013).
34. Crawford, D. J. *et al.* Tet2 catalyzes stepwise 5-methylcytosine oxidation by an iterative and de novo mechanism. *J. Am. Chem. Soc.* **138**, 730–733 (2016).
35. Tamanaha, E., Guan, S., Marks, K. & Saleh, L. Distributive processing by the iron(II)/α-ketoglutarate-dependent catalytic domains of the TET enzymes is consistent with epigenetic roles for oxidized 5-methylcytosine bases. *J. Am. Chem. Soc.* **138**, 9345–9348 (2016).
36. Xu, L. *et al.* Pyrene-based quantitative detection of the 5-formylcytosine loci symmetry in the CpG duplex content during TET-dependent demethylation. *Angew. Chem. Int. Ed Engl.* **53**, 11223–11227 (2014).
37. Ying, Q.-L. *et al.* The ground state of embryonic stem cell self-renewal. *Nature* **453**, 519–523 (2008).
38. Weinberger, L., Ayyash, M., Novershtern, N. & Hanna, J. H. Dynamic stem cell states: naive to primed pluripotency in rodents and humans. *Nat. Rev. Mol. Cell Biol.* **17**, 155–169 (2016).
39. Leitch, H. G. *et al.* Naive pluripotency is associated with global DNA hypomethylation. *Nat. Struct. Mol. Biol.* **20**, 311–316 (2013).
40. Ficz, G. *et al.* FGF signaling inhibition in ESCs drives rapid genome-wide demethylation to the epigenetic ground state of pluripotency. *Cell Stem Cell* **13**, 351–359 (2013).
41. Habibi, E. *et al.* Whole-genome bisulfite sequencing of two distinct interconvertible DNA methylomes of mouse embryonic stem cells. *Cell Stem Cell* **13**, 360–369 (2013).
42. Hayashi, K., Ohta, H., Kurimoto, K., Aramaki, S. & Saitou, M. Reconstitution of the mouse germ cell specification pathway in culture by pluripotent stem cells. *Cell* **146**, 519–532 (2011).
43. Shirane, K. *et al.* Global landscape and regulatory principles of DNA methylation reprogramming for germ cell specification by mouse pluripotent stem cells. *Dev. Cell* **39**, 87–103 (2016).

44. Okano, M., Bell, D. W., Haber, D. A. & Li, E. DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development. *Cell* **99**, 247–257 (1999).
45. Watanabe, D., Suetake, I., Tada, T. & Tajima, S. Stage- and cell-specific expression of Dnmt3a and Dnmt3b during embryogenesis. *Mech. Dev.* **118**, 187–190 (2002).
46. Ito, S. *et al.* Role of Tet proteins in 5mC to 5hmC conversion, ES-cell self-renewal and inner cell mass specification. *Nature* **466**, 1129–1133 (2010).
47. Wossidlo, M. *et al.* 5-Hydroxymethylcytosine in the mammalian zygote is linked with epigenetic reprogramming. *Nat. Commun.* **2**, 241 (2011).
48. Boroviak, T. *et al.* Lineage-specific profiling delineates the emergence and progression of naive pluripotency in mammalian embryogenesis. *Dev. Cell* **35**, 366–382 (2015).
49. Kohli, R. M. & Zhang, Y. TET enzymes, TDG and the dynamics of DNA demethylation. *Nature* **502**, 472–479 (2013).
50. Wu, X. & Zhang, Y. TET-mediated active DNA demethylation: Mechanism, function and beyond. *Nat. Rev. Genet.* https://doi.org/10.1038/nrg.2017.33 (2017).
51. Pfaffeneder, T. *et al.* Tet oxidizes thymine to 5-hydroxymethyluracil in mouse embryonic stem cell DNA. *Nat. Chem. Biol.* **10**, 574–581 (2014).
52. Fidalgo, M. *et al.* Zfp281 coordinates opposing functions of Tet1 and Tet2 in pluripotent states. *Cell Stem Cell* **19**, 355–369 (2016).
53. Sohni, A. *et al.* Dynamic switching of active promoter and enhancer domains regulates Tet1 and Tet2 expression during cell state transitions between pluripotency and differentiation. *Mol. Cell. Biol.* **35**, 1026–1042 (2015).
54. Raiber, E.-A. *et al.* 5-Formylcytosine alters the structure of the DNA double helix. *Nat. Struct. Mol. Biol.* **22**, 44–49 (2015).
55. Ji, S., Shao, H., Han, Q., Seiler, C. L. & Tretyakova, N. Y. Reversible DNA-protein cross-linking at epigenetic DNA marks. *Angew. Chem. Int. Ed. Engl.* **56**, 14130–14134 (2017).
56. Li, F. *et al.* 5-Formylcytosine yields DNA-protein cross-links in nucleosome core particles. *J. Am. Chem. Soc.* **139**, 10617–10620 (2017).
57. Raiber, E. A. *et al.* 5-Formylcytosine controls nucleosome positioning through covalent histone-DNA interaction. *bioRxiv* https://doi.org/10.1101/224444 (2017).
58. Raiber, E.-A. *et al.* 5-Formylcytosine organizes nucleosomes and forms Schiff base interactions with histones in mouse embryonic stem cells. *Nat. Chem.* **10**, 1258–1266 (2018).
59. Hayashi, K. & Saitou, M. Generation of eggs from mouse embryonic stem cells and induced pluripotent stem cells. *Nat. Protoc.* **8**, 1513–1524 (2013).
60. Mulholland, C. B. *et al.* A modular open platform for systematic functional studies under physiological conditions. *Nucleic Acids Res.* **43**, e112 (2015).
61. Wang, H. *et al.* One-step generation of mice carrying mutations in multiple genes by CRISPR/Cas-mediated genome engineering. *Cell* **153**, 910–918 (2013).
62. Bauer, C. *et al.* Phosphorylation of TET proteins is regulated via O-GlcNAcylation by the O-linked N-acetylglucosamine transferase (OGT). *J. Biol. Chem.* **290**, 4801–4812 (2015).
63. Soumillon, M., Cacchiarelli, D., Semrau, S., van Oudenaarden, A. & Mikkelsen, T. S. Characterization of directed differentiation by high-throughput single-cell RNA-Seq. *bioRxiv* https://doi.org/10.1101/003236 (2014).
64. Ziegenhain, C. *et al.* Comparative analysis of single-cell RNA sequencing methods. *Mol. Cell* **65**, 631-643.e4 (2017).
65. Bagnoli, J. W. *et al.* Sensitive and powerful single-cell RNA sequencing using mcSCRB-seq. *Nat. Commun.* **9**, 2937 (2018).
66. Parekh, S., Ziegenhain, C., Vieth, B., Enard, W. & Hellmann, I. zUMIs—A fast and flexible pipeline to process RNA sequencing data with UMIs. *Gigascience* **7**, giy059 (2018).
67. Rau, A., Gallopin, M., Celeux, G. & Jaffrézic, F. Data-based filtering for replicated high-throughput transcriptome sequencing experiments. *Bioinformatics* **29**, 2146–2152 (2013).
68. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
69. Wagner, M. *et al.* Age-dependent levels of 5-methyl-, 5-hydroxymethyl-, and 5-formylcytosine in human and mouse brain tissues. *Angew. Chem. Int. Ed. Engl.* **54**, 12511–12514 (2015).
70. Rappsilber, J., Mann, M. & Ishihama, Y. Protocol for micro-purification, enrichment, pre-fractionation and storage of peptides for proteomics using StageTips. *Nat. Protoc.* **2**, 1896–1906 (2007).
71. Muntel, J. *et al.* Comparison of protein quantification in a complex background by DIA and TMT workflows with fixed instrument time. *J. Proteome Res.* **18**, 1340–1351 (2019).

Acknowledgments

We thank Dr. Feng Zhang for providing the pSpCas9(BB)-2A-Puro (PX459) V2.0 (Addgene plasmid # 62988). We thank Dr. Weihua Qin, V. Burak Ozan, and Terrence Tigney for help with experiments. E.U., M.M., M.D.B., and J.R. are fellows of the International Max Planck Research School for Molecular Life Sciences (IMPRS-LS) and P.S. of the Integrated Research Training Group (IRTG) of the SFB1064. F.R.T. thanks the Boehringer Ingelheim Fonds for a PhD fellowship. E.U. is supported by the research training group 1721 (RTG 1721). This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation)—Project-ID 213249687—SFB 1064 A17 to H.L. and A22 to S.B.

Author contributions

C.B.M. conceived and designed the study, performed experiments, analyzed the data, and wrote the paper. S.B. conceived, designed, and supervised the study, analyzed the data, and wrote the paper. H.L. designed and supervised the study and wrote the paper. F.R.T., E.P., and T.C. performed UHPLC-MS/MS measurements and evaluated the data. E.M.E. and M.S. generated and characterized the cell lines. M.M. and E.U. performed the LC–MS/MS proteomics and data analysis. M.D.B. helped characterize the cell lines. P.S. and J.R. performed validation experiments. All authors read, discussed, and approved the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41598-020-68600-3.

Correspondence and requests for materials should be addressed to H.L. or S.B.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2020

Supplementary Material

Distinct and stage-specific contributions of TET1 and TET2 to stepwise cytosine oxidation in the transition from naive to primed pluripotency

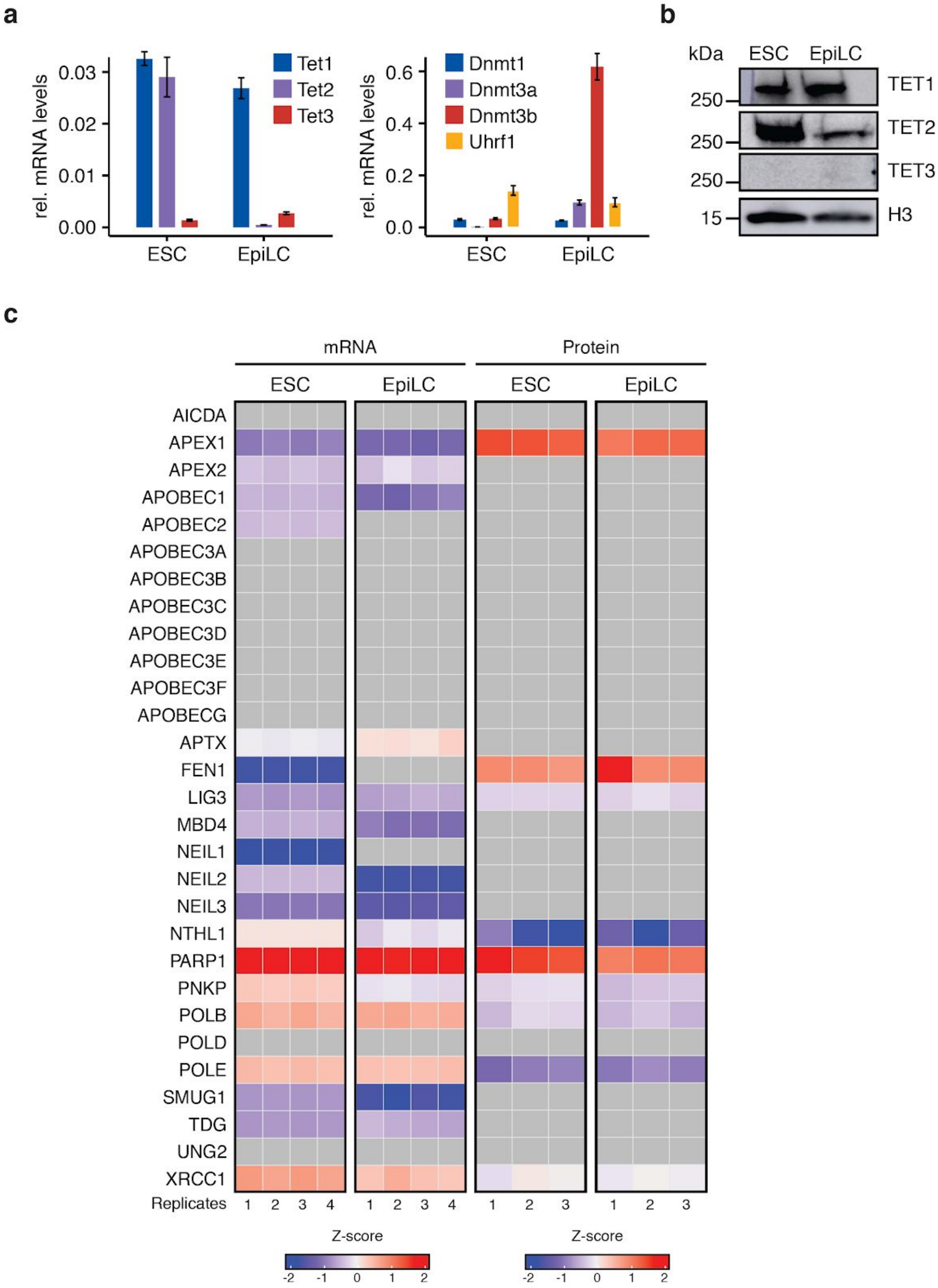
Christopher B. Mulholland[1], Franziska R. Traube[2], Enes Ugur[2], Edris Parsa[2], Eva-Maria Eckl[1], Maximilian Schönung[1], Miha Modic[3], Michael D. Bartoschek[1], Paul Stolz[1], Joel Ryan[1], Thomas Carell[2], Heinrich Leonhardt*[1] and Sebastian Bultmann*[1]

[1]Department of Biology II and Center for Integrated Protein Science Munich (CIPSM), Ludwig-Maximilians-Universität München, Planegg-Martinsried, Germany

[2]Center for Integrated Protein Science (CIPSM) at the Department of Chemistry, Ludwig-Maximilians-Universität München, Munich, Germany

[3]The Francis Crick Institute, London NW1 1AT, United Kingdom

* correspondence can be addressed to Sebastian Bultmann (bultmann@bio.lmu.de) and Heinrich Leonhardt (h.leonhardt@lmu.de)

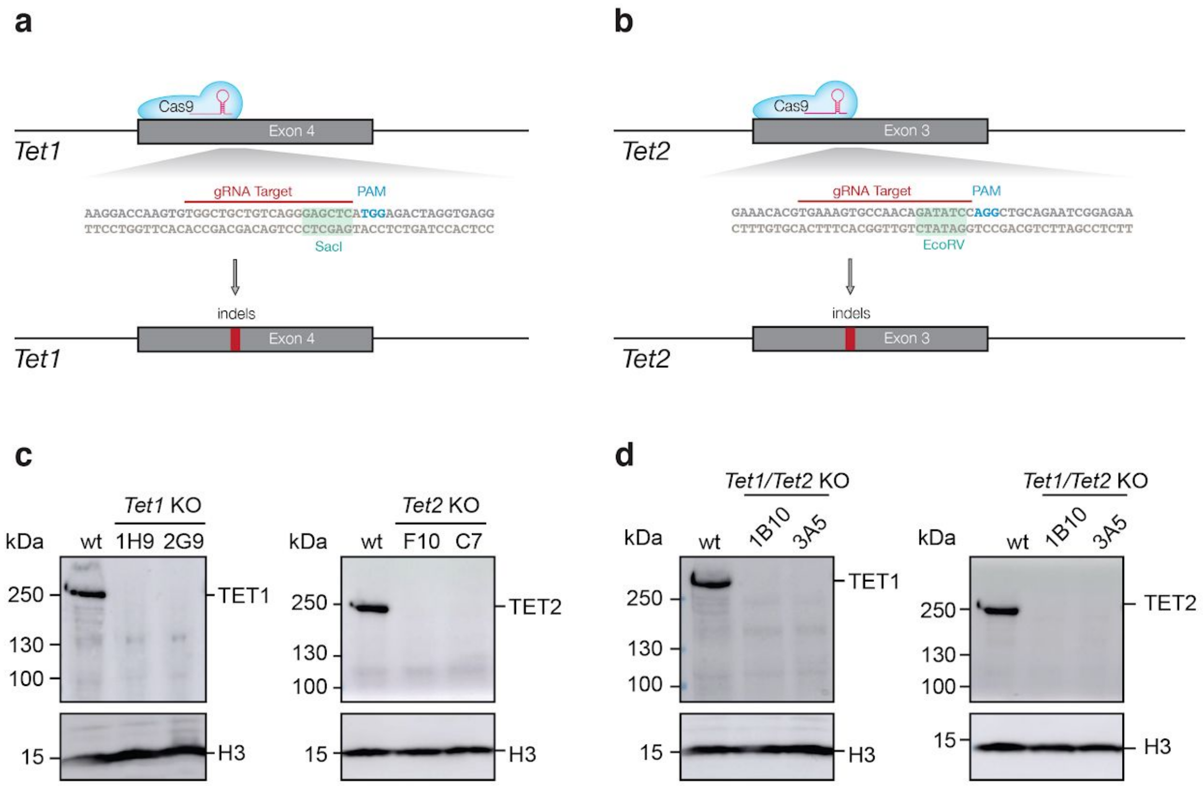

a
rel. mRNA levels
0.03
0.02
0.01
0.00
Tet1
Tet2
Tet3
ESC
EpiLC
rel. mRNA levels
0.6
0.4
0.2
0.0
Dnmt1
Dnmt3a
Dnmt3b
Uhrf1
ESC
EpiLC
b
kDa
ESC
EpiLC
250
TET1
250
TET2
250
TET3
15
H3
c
mRNA
Protein
ESC
EpiLC
ESC
EpiLC
AICDA
APEX1
APEX2
APOBEC1
APOBEC2
APOBEC3A
APOBEC3B
APOBEC3C
APOBEC3D
APOBEC3E
APOBEC3F
APOBECG
APTX
FEN1
LIG3
MBD4
NEIL1
NEIL2
NEIL3
NTHL1
PARP1
PNKP
POLB
POLD
POLE
SMUG1
TDG
UNG2
XRCC1
Replicates
1 2 3 4
1 2 3 4
1 2 3
1 2 3
Z-score
-2 -1 0 1 2
Z-score
-2 -1 0 1 2

Supplementary Figure S1: DNA modification dynamics during the naive to primed transition are accompanied by changes in the expression of DNA modifying enzymes.

(a) Expression of DNA modifying enzymes in mESCs and mEpiLCs shown as the relative mRNA levels as a proportion of *Gapdh* at each stage of pluripotency. Error bars indicate mean ± SD calculated from technical triplicate reactions from n = 3 biological replicates.

(b) Western blot analysis of TET1, TET2, TET3 protein levels in wild-type ESCs and EpiLCs with histone H3 as loading control. Immunoblots were repeated 3 times with similar results obtained.

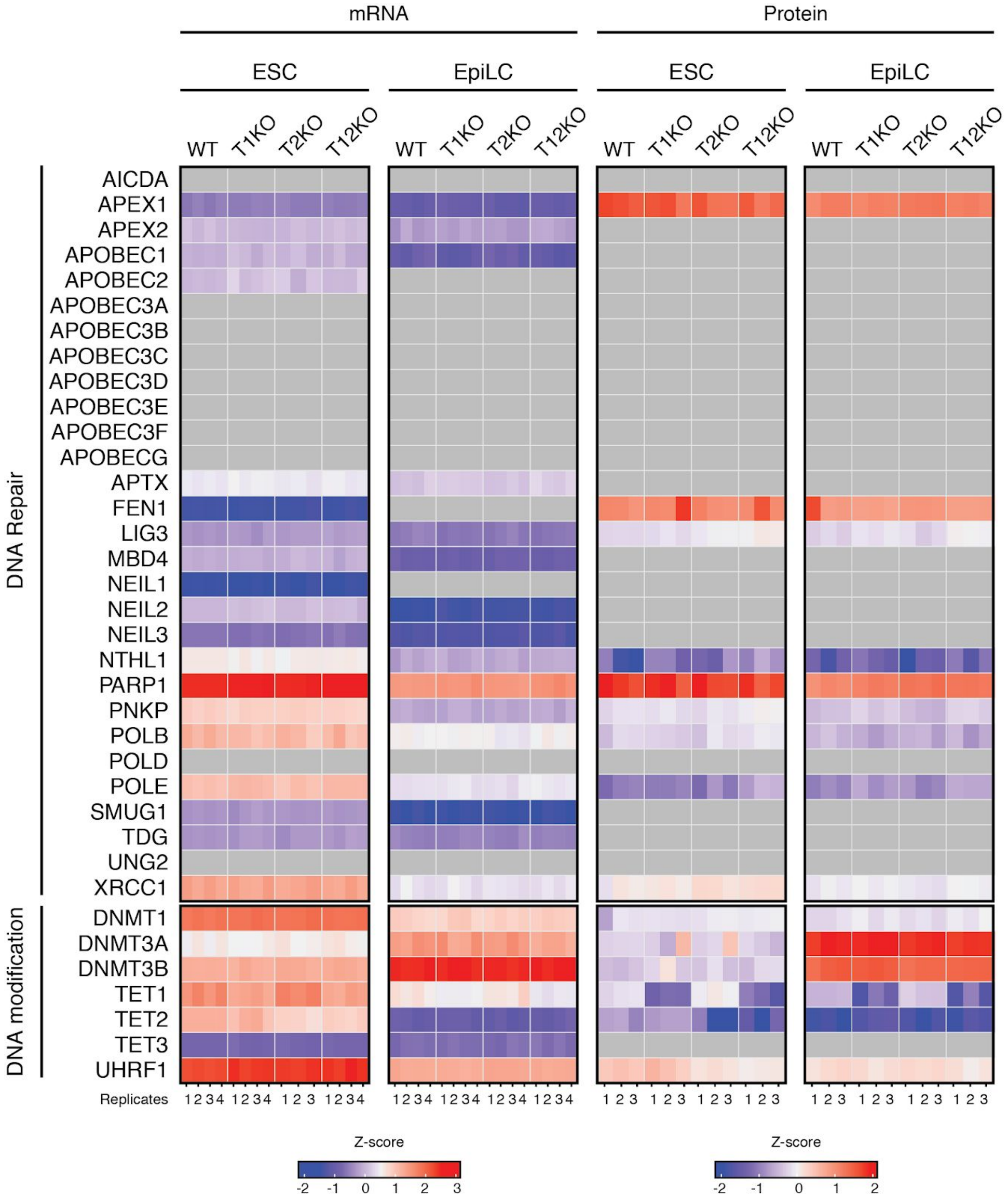
(c) Heatmaps depicting the mRNA levels (left) and protein abundance (right) of DNA repair factors in wild-type ESCs and EpiLCs. Z-scored (Z-score) transcript and protein levels are shown for individual biological replicates (*n* indicated at the bottom of the plots). Gray boxes are used for transcripts and proteins not detected in individual samples by RNA-seq or proteomics measurements, respectively.



Supplementary Figure S2: Generation and characterization of *Tet* KO cell lines

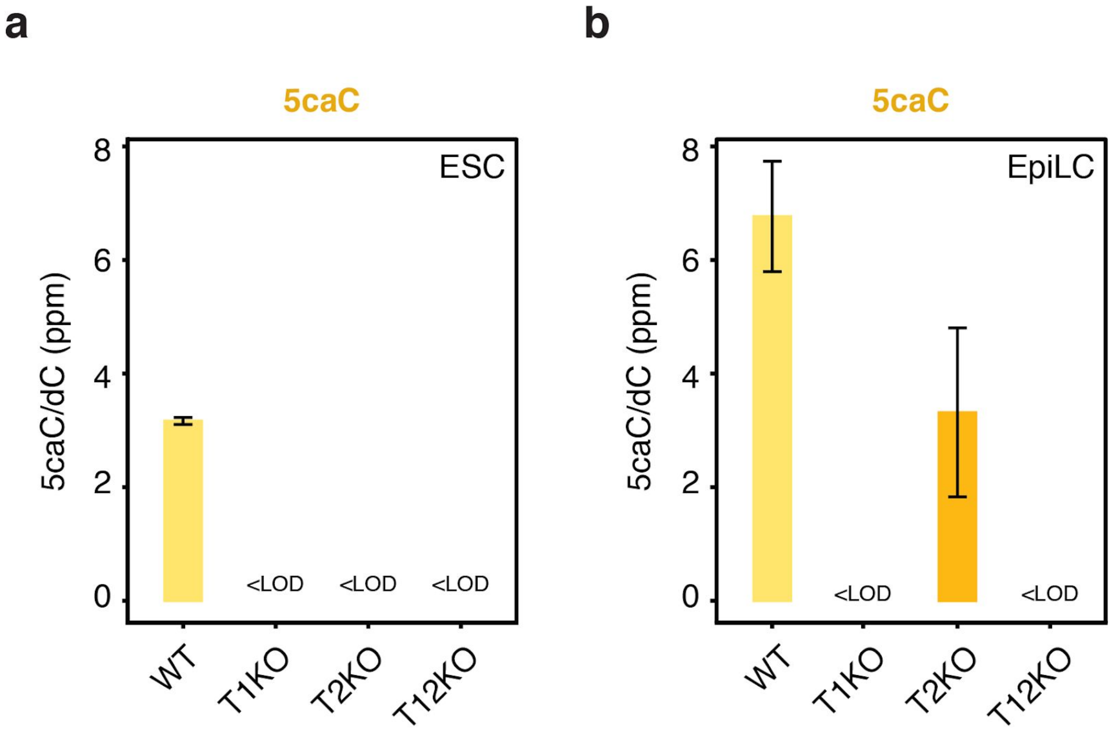
(a-b) Schematic representation of CRISPR/Cas9 targeting of *Tet1* (**a**) and *Tet2* (**b**) loci for KO generation using gRNA target sequences from [61]. For each gRNA, the PAM (NGG) and specific target sequence are indicated, as well as the location of the restriction enzyme recognition sites used for restriction fragment length polymorphism (RFLP) screening.

(**c-d**) Western blot analysis of TET1 and TET2 protein levels in *Tet1* KO and *Tet2* KO (**c**) and *Tet1*/*Tet2* DKO (**d**) cell lines with histone H3 as loading control. For each *Tet* KO, two independent clones were validated and used in all subsequent experiments. The clones validated via Western blot are as follows: *Tet1* KO (1H9 and 2G9), *Tet2* KO (F10 and C7), and *Tet1*/*Tet2* DKO (1B10 and 3A5). Immunoblots were repeated 3 times with similar results obtained.


mRNA
Protein
ESC
EpiLC
ESC
EpiLC
WT
T1KO
T2KO
T12KO
DNA Repair
DNA modification
AICDA
APEX1
APEX2
APOBEC1
APOBEC2
APOBEC3A
APOBEC3B
APOBEC3C
APOBEC3D
APOBEC3E
APOBEC3F
APOBECG
APTX
FEN1
LIG3
MBD4
NEIL1
NEIL2
NEIL3
NTHL1
PARP1
PNKP
POLB
POLD
POLE
SMUG1
TDG
UNG2
XRCC1
DNMT1
DNMT3A
DNMT3B
TET1
TET2
TET3
UHRF1
Replicates
1 2 3 4
1 2 3
Z-score
-2 -1 0 1 2 3
Z-score
-2 -1 0 1 2

Supplementary Figure S3: Comparison of mRNA and protein levels for DNA repair factors and DNA modifying enzymes among Tet KO ESCs and EpiLCs

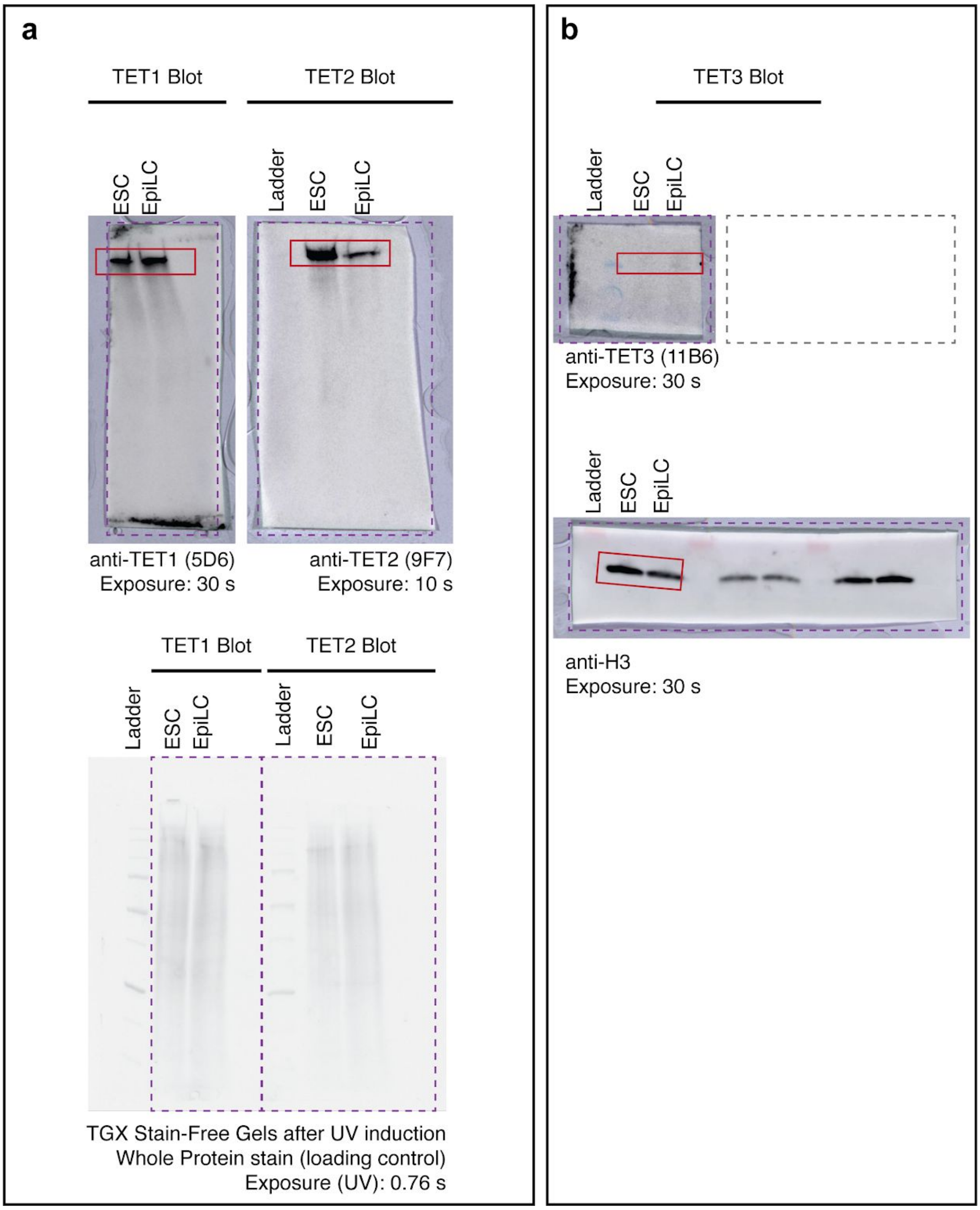
Heatmaps depicting the mRNA levels (left) and protein abundance (right) of DNA repair and modification factors in wild-type (WT), *Tet1 KO* (T1KO), *Tet2 KO* (T2KO), and *Tet1*/*Tet2* DKO (T12KO) ESCs and EpiLCs. Z-scored (Z-score) transcript and protein levels are shown for individual biological replicates (*n* indicated at the bottom of the plots). Gray boxes are used for transcripts and proteins not detected in individual samples by RNA-seq or proteomics measurements, respectively.



Supplementary Figure S4: Global 5caC levels in *Tet* KO ESCs and EpiLCs

(a,b) Global levels of 5caC in wild-type (WT), *Tet1* KO (T1KO), *Tet2* KO (T2KO), and *Tet1*/*Tet2* DKO (T12KO) ESCs **(a)** and EpiLCs **(b)** as determined by mass spectrometry (UHPLC-MS/MS). 5caC levels are expressed as parts per million (ppm: 1 ppm = 0.0001%) of total cytosine (dC). Error bars indicate mean ± SD calculated from $n = 6$ biological replicates for each genotype. LOD, limit of detection.

Supplementary Figure S5

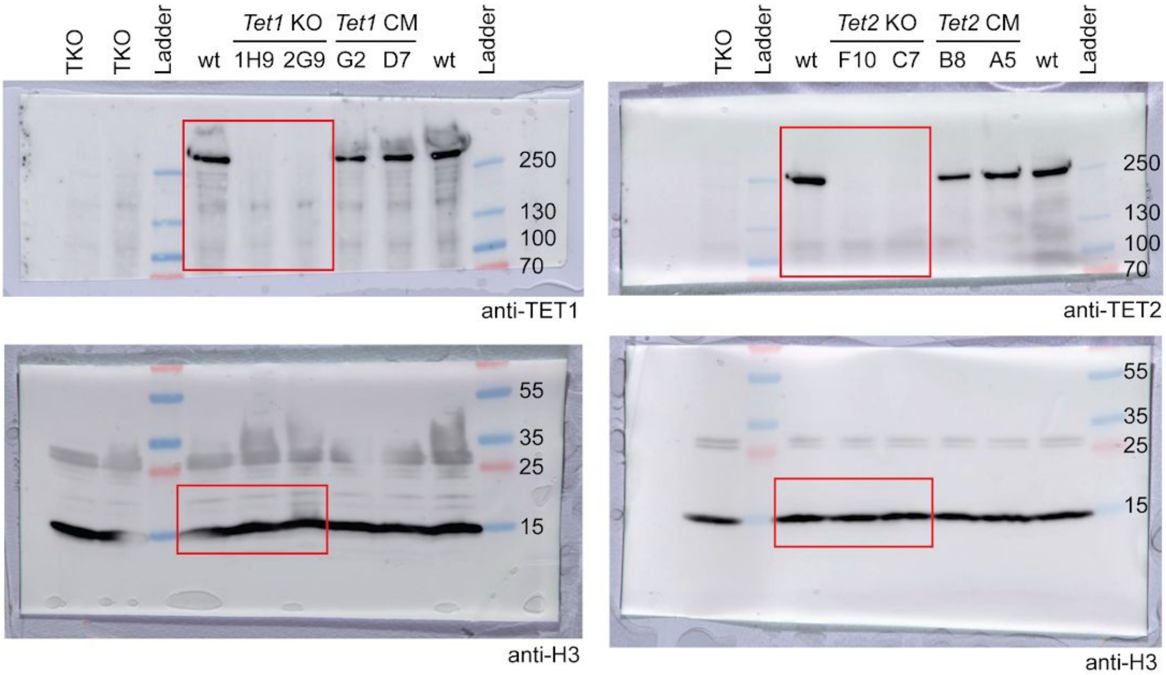


Supplementary Figure S5: Original, uncropped Western Blots from Supplementary Fig. S1b

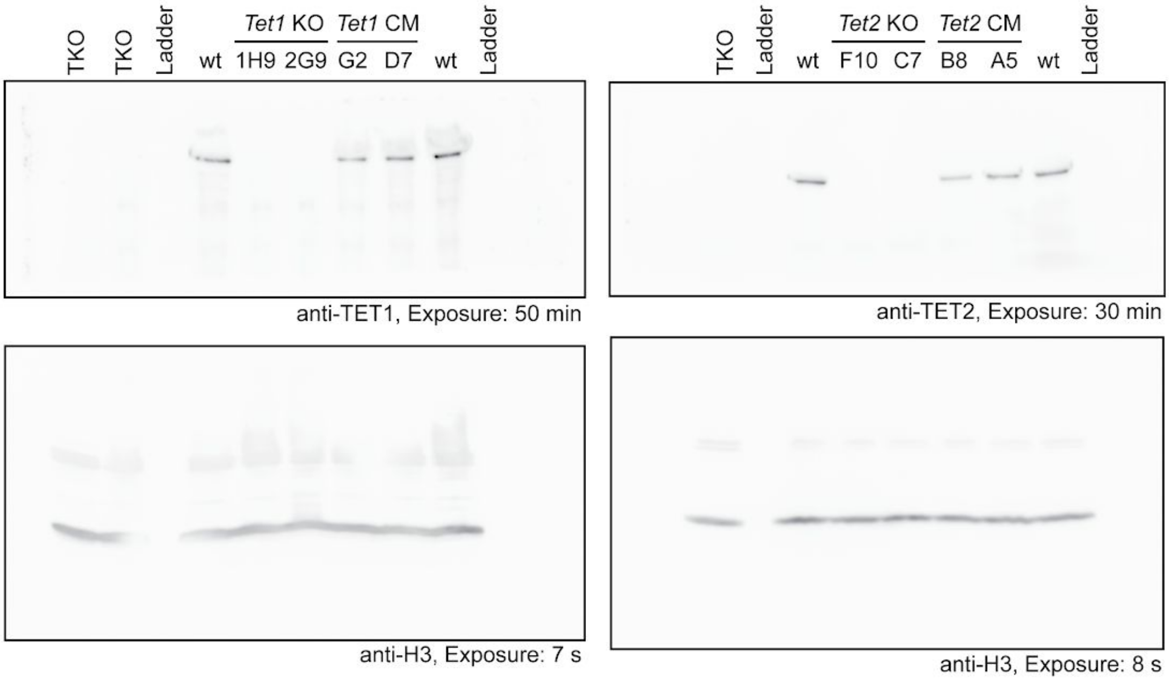
Original, uncropped Western blots of TET1, TET2, and TET3 protein levels in wild-type ESCs and EpiLCs displayed in Supplementary Fig. S1b with whole protein stain (Tet1 and Tet2) or histone H3 (Tet3) serving as loading controls. Cropped areas are indicated by red boxes. Dotted purple lines delineate the relationship of cut blots. Grey dotted line indicates part of blot removed containing unrelated samples.

Supplementary Figure S6

a **Overlay**



b **Unprocessed**

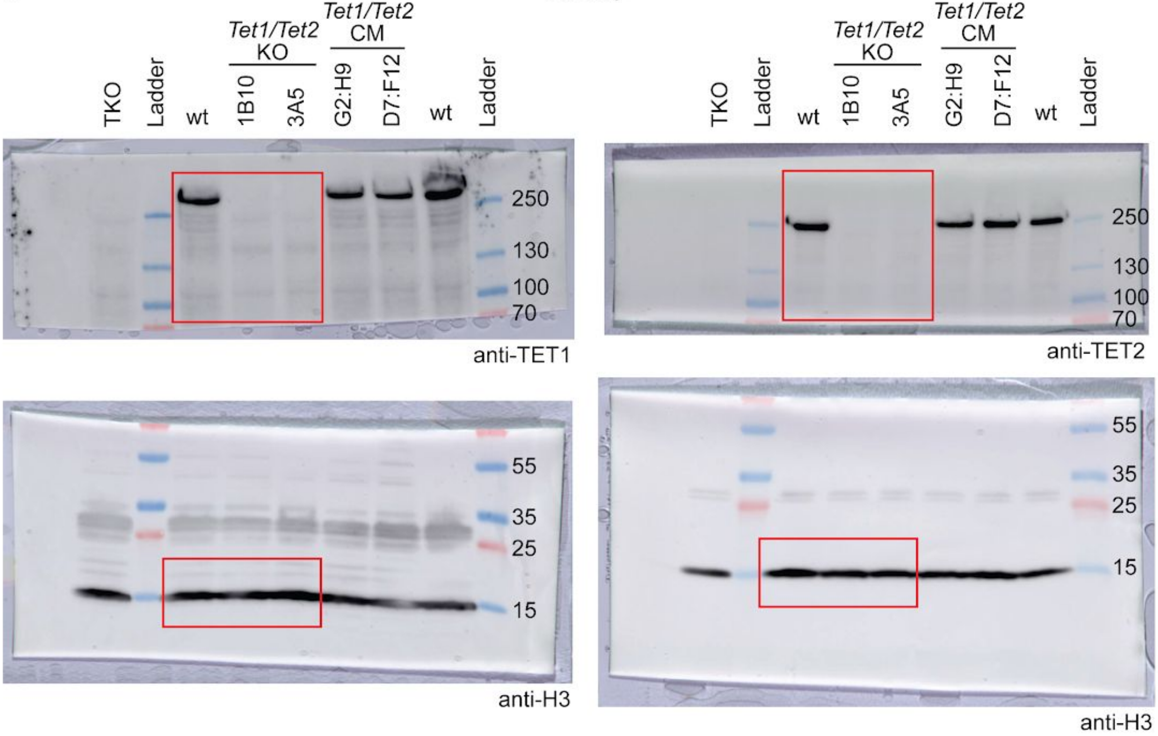


Supplementary Figure S6: Original, uncropped Western Blots from Supplementary Fig. S2c

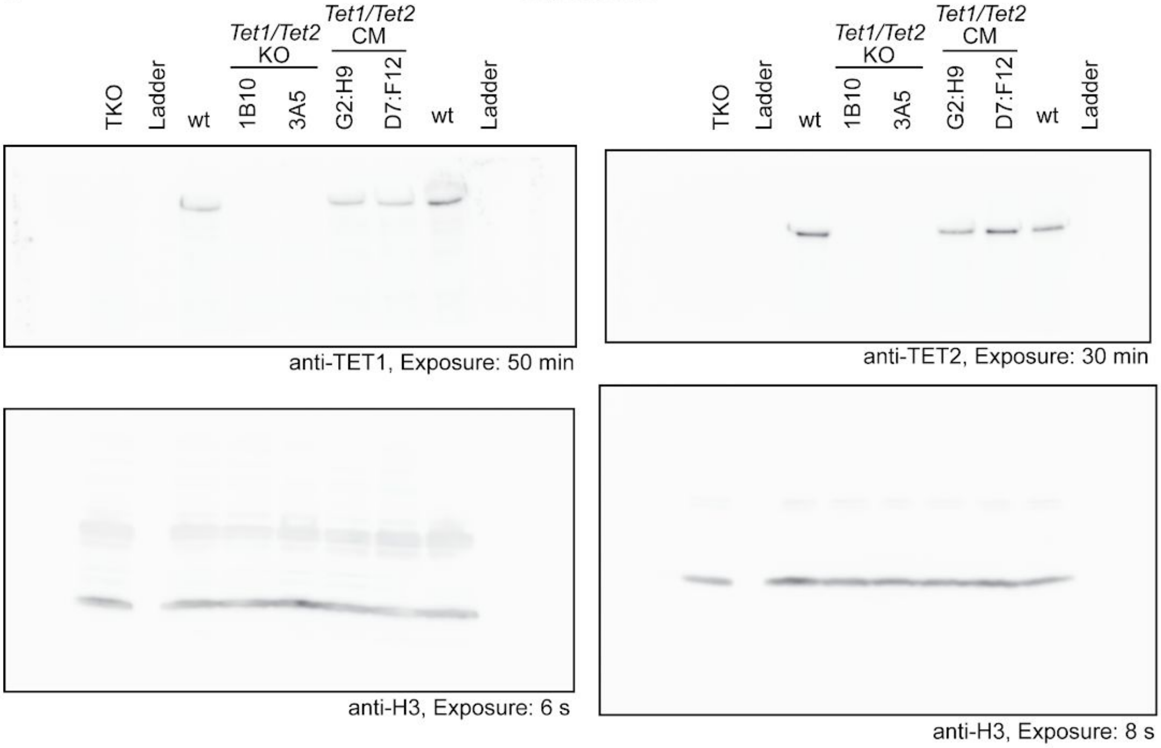
Overlay (**a**) and unprocessed Western blots with indicated exposure times (**b**) of *Tet1* or *Tet2* single knockout (KO) ESCs displayed in Supplementary Figure S2c with histone H3 as loading control. Cropped areas are indicated by red boxes. Lysates from *Tet1*/*Tet2*/*Tet3* triple knockout (TKO) ESCs [23] are loaded as negative controls. Lysates from wild-type (wt) and *Tet1* or *Tet2* single catalytic mutant (CM) ESCs (unpublished) are loaded as positive controls.

Supplementary Figure S7

a **Overlay**



b **Unprocessed**



Supplementary Figure S7: Uncropped Western Blots from Supplementary Fig. S2d

Overlay (**a**) and unprocessed Western blots with indicated exposure times (**b**) of *Tet1*/*Tet2* double knockout (KO) ESCs displayed in Supplementary Figure S2d with histone H3 as loading control. Cropped areas are indicated by red boxes. Lysates from *Tet1*/*Tet2*/*Tet3* triple knockout (TKO) ESCs [23] are loaded as a negative control. Lysates from wild-type (wt) and *Tet1*/*Tet2* double catalytic mutant (CM) ESCs (unpublished) are loaded as positive controls.

Supplementary Table S1: Modified Cytosine Level Quantification

condition	genotype	n	5mC/dC (%)	5mC sd	hmC/dC (%)	5hmC sd	fC/dC (ppm)	5fC sd	n	5caC/dC (ppm)	5caC sd
ESC	wt	18	3.1778	0.11707	0.1084	0.00022	5.7778	0.018	6	3.1700	0.0615
ESC	T1KO	18	3.6346	0.08240	0.0533	0.00096	2.8722	0.043	6	<L.O.D	<L.O.D
ESC	T2KO	12	3.9500	0.09779	0.0294	0.00699	0.7250	0.088	6	<L.O.D	<L.O.D
ESC	T12KO	12	3.6250	0.13198	0.0031	0.02495	0.4275	1.467	6	<L.O.D	<L.O.D
EpiLC	wt	24	6.8375	0.13878	0.5148	8.3E-05	17.0000	0.015	6	6.7700	0.9740
EpiLC	T1KO	12	7.7614	0.07211	0.0525	0.00169	2.6480	0.055	6	<L.O.D	<L.O.D
EpiLC	T2KO	12	8.2250	0.04174	0.3087	0.00042	4.4500	0.025	6	3.3200	1.4900
EpiLC	T12KO	12	7.8417	0.04468	0.0073	0.00549	0.3308	0.311	6	<L.O.D	<L.O.D

Ugur, E., Bartoschek, M. D., & Leonhardt, H. (**2020**). Locus-Specific Chromatin Proteome Revealed by Mass Spectrometry-Based CasID. Methods in Molecular Biology , 2175, 109–121.

https://doi.org/10.1007/978-1-0716-0763-3_9

Mulholland, C. B., Nishiyama, A., Ryan, J., Nakamura, R., Yiğit, M., Glück, I. M., Trummer, C., Qin, W., Bartoschek, M. D., Traube, F. R., Parsa, E., **Ugur, E.**, Modic, M., Acharya, A., Stolz, P., Ziegenhain, C., Wierer, M., Enard, W., Carell, T., Bultmann, S., Leonhardt, H. (**2020**). **Recent evolution of a TET-controlled and DPPA3/STELLA-driven pathway of passive DNA demethylation in mammals**. Nature Communications, 11(1), 5972.

License: CC-BY-4.0

https://doi.org/10.1038/s41467-020-19603-1

ARTICLE

https://doi.org/10.1038/s41467-020-19603-1 OPEN

Recent evolution of a TET-controlled and DPPA3/STELLA-driven pathway of passive DNA demethylation in mammals

Christopher B. Mulholland[1], Atsuya Nishiyama[2,9], Joel Ryan[1,9], Ryohei Nakamura[3], Merve Yiğit[1], Ivo M. Glück[4], Carina Trummer[1], Weihua Qin[1], Michael D. Bartoschek[1], Franziska R. Traube[5], Edris Parsa[5], Enes Ugur[1,6], Miha Modic[7], Aishwarya Acharya[1], Paul Stolz[1], Christoph Ziegenhain[8], Michael Wierer[6], Wolfgang Enard[8], Thomas Carell[5], Don C. Lamb[4], Hiroyuki Takeda[3], Makoto Nakanishi[2], Sebastian Bultmann[1,10]✉ & Heinrich Leonhardt[1,10]✉

Genome-wide DNA demethylation is a unique feature of mammalian development and naïve pluripotent stem cells. Here, we describe a recently evolved pathway in which global hypomethylation is achieved by the coupling of active and passive demethylation. TET activity is required, albeit indirectly, for global demethylation, which mostly occurs at sites devoid of TET binding. Instead, TET-mediated active demethylation is locus-specific and necessary for activating a subset of genes, including the naïve pluripotency and germline marker *Dppa3* (*Stella, Pgc7*). DPPA3 in turn drives large-scale passive demethylation by directly binding and displacing UHRF1 from chromatin, thereby inhibiting maintenance DNA methylation. Although unique to mammals, we show that DPPA3 alone is capable of inducing global DNA demethylation in non-mammalian species (Xenopus and medaka) despite their evolutionary divergence from mammals more than 300 million years ago. Our findings suggest that the evolution of *Dppa3* facilitated the emergence of global DNA demethylation in mammals.

---

[1] Department of Biology II and Center for Integrated Protein Science Munich (CIPSM), Human Biology and BioImaging, Ludwig-Maximilians-Universität München, Planegg-Martinsried, Germany. [2] Division of Cancer Cell Biology, The Institute of Medical Science, The University of Tokyo, 4-6-1 Shirokanedai, Minato-ku, Tokyo 108-8639, Japan. [3] Department of Biological Sciences, Graduate School of Science, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033, Japan. [4] Physical Chemistry, Department of Chemistry, Center for Nanoscience, Nanosystems Initiative Munich and Center for Integrated Protein Science Munich, Ludwig-Maximilians-Universität München, Munich, Germany. [5] Center for Integrated Protein Science (CIPSM) at the Department of Chemistry, Ludwig-Maximilians-Universität München, Munich, Germany. [6] Department of Proteomics and Signal Transduction, Max Planck Institute for Biochemistry, Am Klopferspitz 18, 82152 Martinsried, Germany. [7] The Francis Crick Institute and UCL Queen Square Institute of Neurology, London, UK. [8] Department of Biology II, Anthropology and Human Genomics, Ludwig-Maximilians-Universität München, Planegg-Martinsried, Germany. [9] These authors contributed equally: Atsuya Nishiyama, Joel Ryan. [10] These authors jointly supervised: Sebastian Bultmann, Heinrich Leonhardt. ✉email: bultmann@bio.lmu.de; h.leonhardt@lmu.de

During early embryonic development the epigenome undergoes massive changes. Upon fertilization, the genomes of highly specialized cell types—sperm and oocyte—need to be reprogrammed in order to obtain totipotency. This process entails decompaction of the highly condensed gametic genomes and global resetting of chromatin states to confer the necessary epigenetic plasticity required for the development of a new organism[1]. At the same time, the genome needs to be protected from the activation of transposable elements (TEs) abundantly present in vertebrate genomes[2]. Activation and subsequent transposition of TEs result in mutations that can have deleterious effects and are passed onto offspring if they occur in the germline during early development[2,3]. The defense against these genomic parasites has shaped genomes substantially[4,5].

Cytosine DNA methylation (5-methylcytosine (5mC)) is a reversible epigenetic mark essential for cellular differentiation, genome stability, and embryonic development in vertebrates[6]. Predominantly associated with transcriptional repression, DNA methylation has important roles in gene silencing, genomic imprinting, and X inactivation[7]. However, the most basic, conserved function of DNA methylation is the stable repression of TEs and other repetitive sequences[8]. Accordingly, the majority of genomic 5mC is located within these highly abundant repetitive elements. Global DNA methylation loss triggers the derepression of transposable and repetitive elements, which leads to genomic instability and cell death, highlighting the crucial function of vertebrate DNA methylation[9–14]. Hence, to ensure continuous protection against TE reactivation, global DNA methylation levels remain constant throughout the lifetime of non-mammalian vertebrates[15–18]. Paradoxically, mammals specifically erase DNA methylation during preimplantation development[19,20], a process that would seemingly expose the developing organism to the risk of genomic instability through the activation of TEs. DNA methylation also acts as an epigenetic barrier to restrict and stabilize cell fate decisions and thus constitutes a form of epigenetic memory. The establishment of pluripotency in mammals requires the erasure of epigenetic memory and as such, global hypomethylation is a defining characteristic of pluripotent cell types including naïve embryonic stem cells (ESCs), primordial germ cells (PGCs), and induced pluripotent stem cells (iPSCs)[21].

In animals, DNA methylation can be reversed to unmodified cytosine by two mechanisms; either actively by Ten-eleven translocation (TET) dioxygenase-mediated oxidation of 5mC in concert with the base excision repair machinery[22–25] or passively by a lack of functional DNA methylation maintenance during the DNA replication cycle[26,27]. Both active and passive demethylation pathways have been implicated in the genome-wide erasure of 5mC accompanying mammalian preimplantation development[28–34]. Despite the extensive conservation of the TET enzymes and DNA methylation machinery throughout metazoa[35], developmental DNA demethylation appears to be unique to placental mammals[19,36–43]. In contrast, 5mC patterns have been found to remain constant throughout early development in all non-mammalian vertebrates examined to date[15,44–48]. This discrepancy implies the existence of yet-to-be-discovered mammalian-specific pathways that orchestrate the establishment and maintenance of global hypomethylation.

Here, we use mouse embryonic stem cells (ESCs) cultured in conditions promoting naïve pluripotency[49–51] as a model to study global DNA demethylation in mammals. By dissecting the contribution of the catalytic activity of TET1 and TET2 to global hypomethylation, we find that TET-mediated active demethylation drives the expression of the Developmental pluripotency-associated protein 3 (DPPA3/PGC7/STELLA). We show that DPPA3 directly binds UHRF1 and triggers its release from chromatin, thereby inhibiting maintenance methylation and causing global passive demethylation. Although DPPA3 is only found in mammals, we found that DPPA3 can also potently induce global demethylation when introduced into non-mammalian vertebrates. In summary, our study uncovers a novel TET-controlled and DPPA3-driven pathway for passive demethylation in naïve pluripotency in mammals.

Results

TET1 and TET2 indirectly protect the naïve genome from hypermethylation. Mammalian TET proteins, TET1, TET2, and TET3, share a conserved catalytic domain and the ability to oxidize 5mC but exhibit distinct expression profiles during development[52]. Naïve ESCs and the inner cell mass (ICM) of the blastocyst from which they are derived feature high expression of *Tet1* and *Tet2* but not *Tet3*[29,53–55]. To dissect the precise contribution of TET-mediated active DNA demethylation to global DNA hypomethylation in naïve pluripotency we generated isogenic *Tet1* (T1CM) and *Tet2* (T2CM) single as well as *Tet1/Tet2* (T12CM) double catalytic mutant mouse ESC lines using CRISPR/Cas-assisted gene editing (Supplementary Fig. 1). We derived two independent clones for each mutant cell line and confirmed the inactivation of TET1 and TET2 activity by measuring the levels of 5-hydroxymethylcytosine (5hmC), the product of TET-mediated oxidation of 5mC[22] (Supplementary Fig. 1i). While the loss of either *Tet1* or *Tet2* catalytic activity significantly reduced 5hmC levels, inactivation of both TET1 and TET2 resulted in the near total loss of 5hmC in naïve ESCs (Supplementary Fig. 1i) indicating that TET1 and TET2 account for the overwhelming majority of cytosine oxidation in naïve ESCs. We then used reduced representation bisulfite sequencing (RRBS) to determine the DNA methylation state of T1CM, T2CM, and T12CM ESCs as well as wild-type (wt) ESCs. All *Tet* catalytic mutant (T1CM, T2CM, and T12CM) cell lines exhibited severe DNA hypermethylation throughout the genome including promoters, gene bodies, and repetitive elements (Fig. 1a, b and Supplementary Fig. 2a). The increase in DNA methylation was particularly pronounced at LINE-1 (L1) elements of which 97%, 98%, and 99% were significantly hypermethylated in T1CM, T2CM, and T12CM ESCs, respectively (Supplementary Fig. 2b). This widespread DNA hypermethylation was reminiscent of the global increase in DNA methylation accompanying the transition of naïve ESCs to primed epiblast-like cells (EpiLCs)[54,56,57], which prompted us to investigate whether the DNA methylation signature in T1CM, T2CM, and T12CM ESCs resembles that of more differentiated cells. In line with this hypothesis, *Tet* catalytic mutant ESCs displayed DNA methylation levels similar to or higher than those of wt EpiLCs (Supplementary Fig. 2c). Moreover, hierarchical clustering and principal component analyses (PCA) of the RRBS data revealed that ESCs from *Tet* catalytic mutants clustered closer to wt EpiLCs than wt ESCs (Fig. 1c and Supplementary Fig. 2d). In fact, the vast majority of significantly hypermethylated CpGs in *Tet* catalytic mutant ESCs overlapped with those normally gaining DNA methylation during the exit from naïve pluripotency (Fig. 1d). In contrast, T1CM, T2CM, and T12CM transcriptomes are clearly clustered by differentiation stage, indicating that the acquisition of an EpiLC-like methylome was not due to premature differentiation (Supplementary Fig. 2e). When comparing our data to that of TET knockout ESCs[58], we found that the catalytic inactivation of the TET proteins caused a far more severe hypermethylation phenotype than the complete removal of the TET proteins (Supplementary Fig. 2f). Intriguingly, whereas TET1 and TET2 prominently associate with sites of active demethylation (Supplementary Fig. 2g), we found that the majority of sites hypermethylated in *Tet* catalytic mutant ESCs

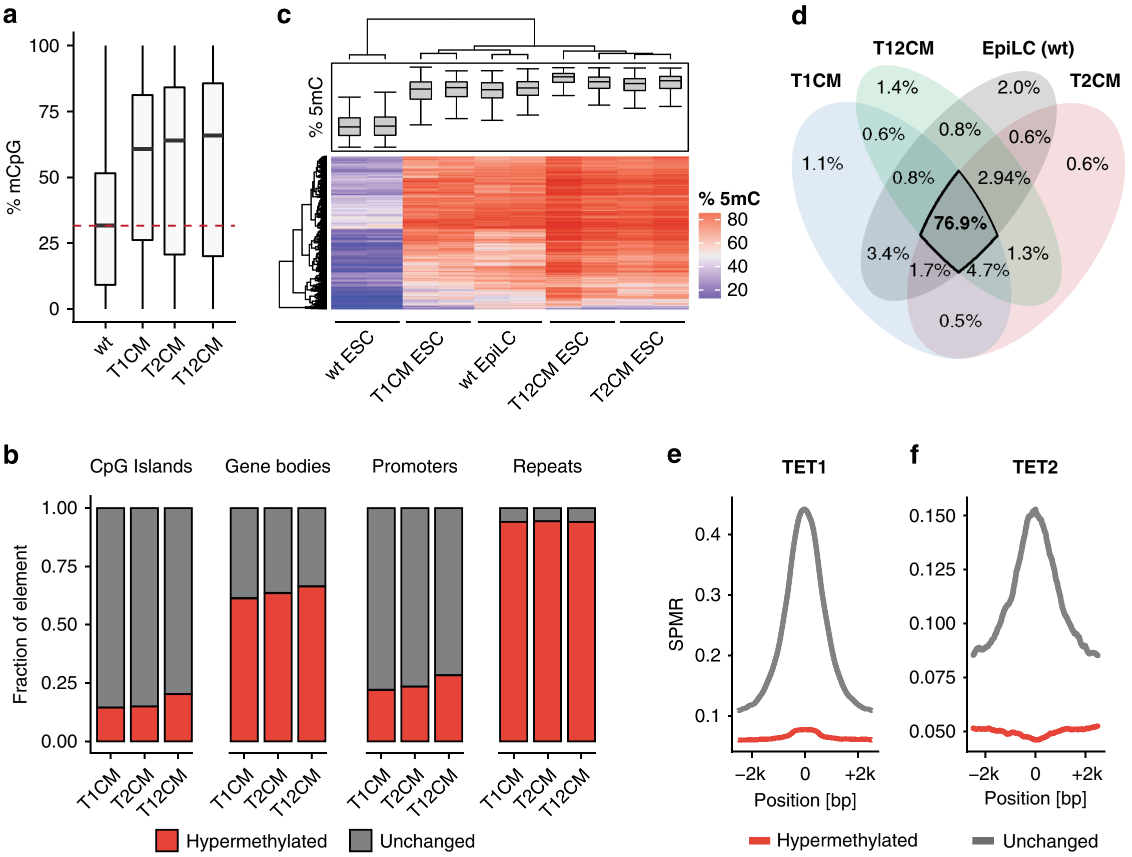
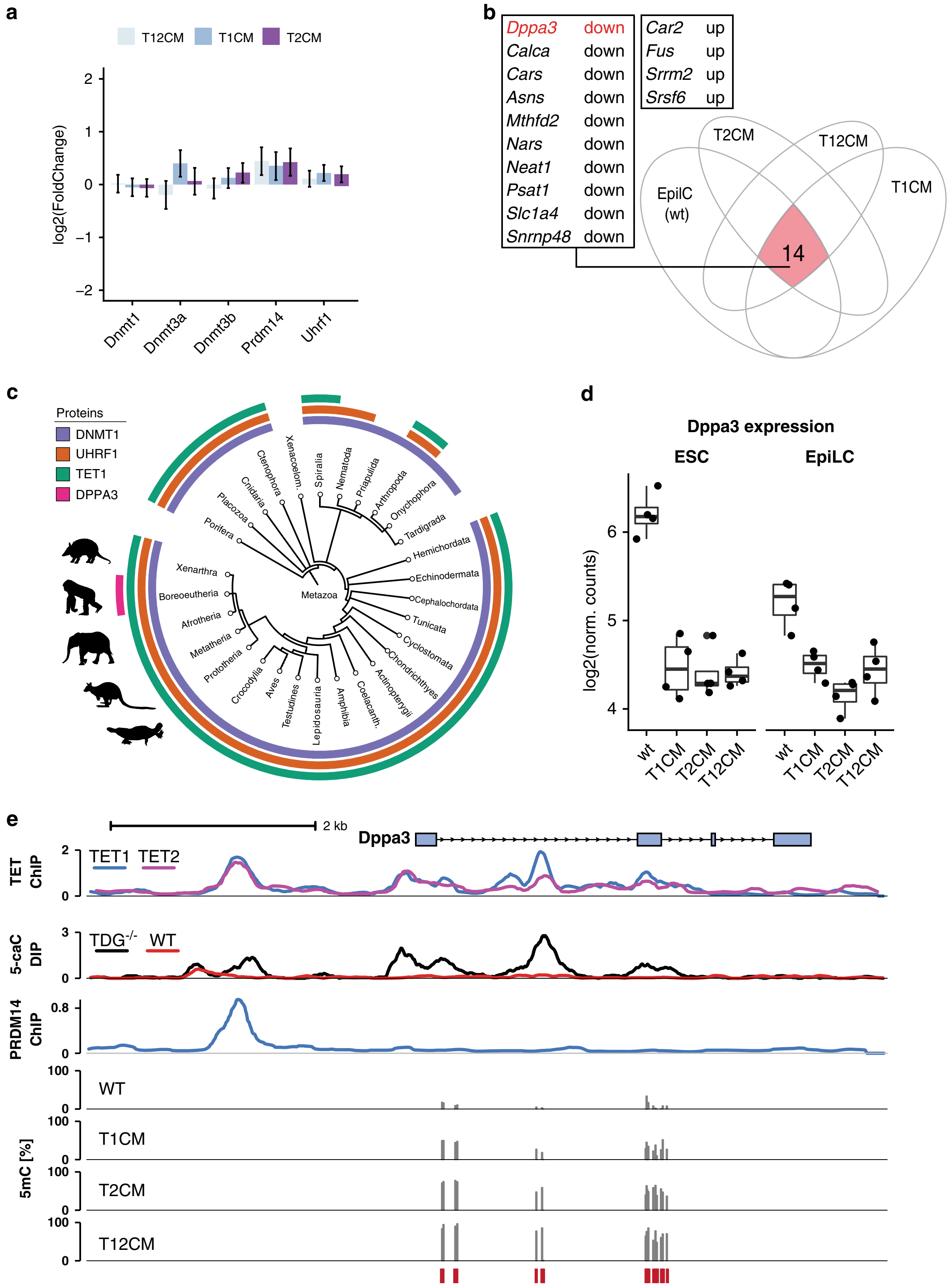


Fig. 1 TET1 and TET2 prevent hypermethylation of the naïve genome. a Loss of TET catalytic activity leads to global DNA hypermethylation. Percentage of total 5mC as measured by RRBS. For each genotype, $n = 2$ biologically independent samples per condition. **b** Loss of TET catalytic activity leads to widespread DNA hypermethylation especially at repetitive elements. Relative proportion of DNA hypermethylation (q value < 0.05; absolute methylation difference >20%) at each genomic element in T1CM, T2CM, and T12CM ESCs compared to wt ESCs. **c** Heat map of the hierarchical clustering of the RRBS data depicting the top 2000 most variable 1 kb tiles during differentiation of wt ESCs to EpiLCs with $n = 2$ biologically independent samples per genotype and condition. **d** Venn diagram depicting the overlap of hypermethylated (compared to wt ESCs; q value < 0.05; absolute methylation difference >20%) sites among T1CM, T2CM, and T12CM ESCs and wt EpiLCs. **e, f** TET binding is not associated with DNA hypermethylation in TET mutant ESCs. Occupancy of (**e**) TET1[66] and (**f**) TET2[67] over 1 kb tiles hypermethylated (dark red) or unchanged (dark gray) in T1CM and T2CM ESCs, respectively (SPMR: Signal per million reads). In the boxplots in (**a**) and (**c**), horizontal black lines within boxes represent median values, boxes indicate the upper and lower quartiles, and whiskers extend to the most extreme value within 1.5 x the interquartile range from each hinge. In (**b**) and (**d**), the q-values were calculated with a two-sided Wald test followed by p-value adjustment using SLIM[209].

are not bound by either enzyme (Fig. 1e, f) suggesting that TET1 and TET2 maintain the hypomethylated state of the naïve methylome by indirect means.

TET1 and TET2 control *Dppa3* expression in a catalytically dependent manner. To explore how TET1 and TET2 might indirectly promote demethylation of the naïve genome, we first examined the expression of the enzymes involved in DNA methylation. Loss of TET catalytic activity was not associated with changes in the expression of *Dnmt1*, *Uhrf1*, *Dnmt3a*, and *Dnmt3b* nor differences in UHRF1 protein abundance, indicating the hypermethylation in *Tet* catalytic mutant ESCs is not caused by aberrant upregulation of DNA methylation machinery components (Fig. 2a, Supplementary Fig. 2h). To identify candidate factors involved in promoting global hypomethylation, we compared the transcriptome of hypomethylated wild-type ESCs with those of hypermethylated cells, which included wt EpiLCs as well as T1CM, T2CM, and T12CM ESCs (Fig. 2b). Among the 14 genes differentially expressed in hypermethylated cell lines, the naïve pluripotency factor, *Dppa3* (also known as *Stella* and *Pgc7*), stood out as an interesting candidate due to its reported involvement in the regulation of global DNA methylation in germ cell development and oocyte maturation[59–62]. In contrast to the core components of the DNA (de)methylation machinery (DNMTs, UHRF1, TETs), which are conserved throughout metazoa, *Dppa3* is only present in mammals, suggesting it might also contribute to the mammal-specific hypomethylation in naïve pluripotency (Fig. 2c).

While normally highly expressed in naïve ESCs and only downregulated upon differentiation[63,64], *Dppa3* was prematurely repressed in T1CM, T2CM, and T12CM ESCs (Fig. 2d). The strongly reduced expression of *Dppa3* in TET mutant ESCs was accompanied by significant hypermethylation of the *Dppa3* promoter (Fig. 2e), consistent with reports demonstrating *Dppa3* to be one of the few pluripotency factors downregulated by promoter methylation upon differentiation in vitro and in vivo[51,63–65]. In contrast to the majority of genomic sites gaining methylation in TET mutant ESCs (Fig. 1e, f), hypermethylation at the *Dppa3* locus occurred at sites bound by both TET1 and TET2 (Fig. 2e)[66,67]. This hypermethylation overlapped with regions at which the TET oxidation product 5-carboxylcytosine (5caC) accumulates in Thymine DNA glycosylase (TDG)-knockdown ESCs (Fig. 2e)[68], indicating that the



Dppa3 locus is a direct target of TET/TDG-mediated active DNA demethylation in ESCs. To test whether Dppa3 transcription can be induced by DNA demethylation, we analyzed RNA-seq data from conditional Dnmt1 KO ESCs[69]. In the absence of genome-wide DNA methylation, *Dppa3* levels more than doubled, thus confirming our results that the *Dppa3* promoter is sensitive to DNA methylation (Supplementary Fig. 2i).

In addition, *Dppa3* is also a direct target of PRDM14, a PR domain-containing transcriptional regulator known to promote the DNA hypomethylation associated with naïve pluripotency[50,70–72] (Fig. 2e). PRDM14 has been shown to recruit TET1 and TET2 to sites of active demethylation and establish global hypomethylation in naïve pluripotency[50,54,71–73]. As the expression of *Prdm14* was not altered in *Tet* catalytic mutant ESCs (Fig. 2a), we analyzed PRDM14

Fig. 2 TET1 and TET2 catalytic activity is necessary for *Dppa3* expression. a Expression of genes involved in regulating DNA methylation levels in T1CM, T2CM, and T12CM ESCs as assessed by RNA-seq. Expression is given as the $\log_2$ fold-change compared to wt ESCs. Error bars indicate mean ± SD, $n = 4$ biological replicates. No significant changes observable (Likelihood ratio test). **b** *Dppa3* is downregulated upon loss of TET activity and during differentiation. Venn diagram depicting the overlap (red) of genes differentially expressed (compared to wt ESCs; adjusted $p < 0.05$) in T1CM, T2CM, T12CM ESCs, and wt EpiLCs. **c** Phylogenetic tree of TET1, DNMT1, UHRF1, and DPPA3 in metazoa. **d** *Dppa3* expression levels as determined by RNA-seq in the indicated ESC and EpiLC lines ($n = 4$ biological replicates). **e** TET proteins bind and actively demethylate the *Dppa3* locus. Genome browser view of the *Dppa3* locus with tracks of the occupancy (Signal pileup per million reads; (SPMR)) of TET1[66], TET2[67], and PRDM14[71] in wt ESCs, 5caC enrichment in wt vs. TDG$^{-/-}$ ESCs[68], and 5mC (%) levels in wt, T1CM, T2CM, and T12CM ESCs (RRBS). Red bars indicate CpGs covered by RRBS. In the boxplots in (**d**), horizontal black lines within boxes represent median values, boxes indicate the upper and lower quartiles, and whiskers extend to the most extreme value within 1.5 x the interquartile range from each hinge.

occupancy at the *Dppa3* locus using publicly available ChIP-seq data[71]. This analysis revealed that PRDM14 binds the same upstream region of *Dppa3* occupied by TET1 and TET2 (Fig. 2e). Taken together, these data suggest that TET1 and TET2 are recruited by PRDM14 to maintain the expression of *Dppa3* by active DNA demethylation.

DPPA3 acts downstream of TET1 and TET2 and is required to safeguard the naïve methylome. DPPA3 has been reported to both prevent and promote DNA demethylation depending on the cellular and developmental context[59,61,62,74–78]. However, the function of DPPA3 in naïve pluripotency, for which it is a well-established marker gene[63], remains unclear. To investigate the relationship between *Dppa3* expression and DNA hypomethylation in naïve pluripotency, we established *Dppa3* knockout (Dppa3KO) mouse ESCs (Supplementary Fig. 3a–c) and profiled their methylome by RRBS. Deletion of *Dppa3* led to severe global hypermethylation (Fig. 3a), with substantial increases in DNA methylation observed across all analyzed genomic features, including promoters, repetitive sequences, and imprinting control regions (ICRs) (Supplementary Fig. 3d–f). In particular, transposable elements experienced the most extensive gains in DNA methylation, with >90% of detected LINE and ERVs found hypermethylated in Dppa3KO ESCs (Supplementary Fig. 3e).

A principal component analysis of the RRBS data revealed that Dppa3KO ESCs clustered closer to wt EpiLCs and *Tet* catalytic mutant ESCs rather than wt ESCs (Fig. 3b). Furthermore, we observed a striking overlap of hypermethylated CpGs between *Tet* catalytic mutant and Dppa3KO ESCs (Fig. 3c), suggesting that DPPA3 and TETs promote demethylation at largely the same targets. A closer examination of the genomic distribution of overlapping hypermethylation in *Tet* catalytic mutant and Dppa3KO ESCs revealed that the majority (~90%) of hypermethylated events within repetitive elements are common to both cell lines (Fig. 3d and Supplementary Fig. 3g–j) and are globally correlated with heterochromatic histone modifications (Supplementary Fig. 3k). In contrast, only half of the observed promoter hypermethylation among all cell lines was dependent on DPPA3 (classified as “common”, Fig. 3d and Supplementary Fig. 3h–j). This allowed us to identify a set of strictly TET-dependent promoters (N = 1573) (Fig. 3d, Supplementary Fig. 3i and Supplementary Data 1), which were enriched for developmental genes (Fig. 3e and Supplementary Data 2). Intriguingly, these TET-specific promoters contained genes (such as *Pax6*, *Foxa1*, and *Otx2*) that were recently shown to be conserved targets of TET-mediated demethylation during *Xenopus*, zebrafish, and mouse development[79].

DPPA3 appeared to act downstream of TETs as the global increase in DNA methylation in Dppa3KO ESCs was not associated with a reduction in 5hmC levels nor with a downregulation of TET family members (Fig. 3f and Supplementary Fig. 3l). In support of this notion, inducible overexpression of *Dppa3* (Supplementary Fig. 3m–o) completely rescued the observed hypermethylation phenotype at LINE-1 elements in T1CM as well as T2CM ESCs and resulted in a significant reduction of hypermethylation in T12CM cells (Fig. 3g). Strikingly, prolonged induction of *Dppa3* resulted in hypomethylation in wild-type as well as T1CM ESCs (Fig. 3g). Collectively, these results show that TET1 and TET2 activity contributes to genomic hypomethylation in naïve pluripotency by both direct and indirect pathways. Whereas direct and active demethylation protects a limited but key set of promoters, global DNA demethylation occurs as an indirect effect of *Dppa3* activation.

TET-dependent expression of DPPA3 regulates UHRF1 subcellular distribution and controls DNA methylation maintenance in embryonic stem cells. To investigate the mechanism underlying the regulation of global DNA methylation patterns by DPPA3, we first generated an endogenous DPPA3-HALO fusion ESC line to monitor the localization of DPPA3 throughout the cell cycle (Supplementary Fig. 4a, c). Previous studies have shown that DPPA3 binds H3K9me2[77] and that in oocytes its nuclear localization is critical to inhibit the activity of UHRF1[62], a key factor for maintaining methylation. Expecting a related mechanism to be present in ESCs, we were surprised to find that DPPA3 primarily localized to the cytoplasm of ESCs (Fig. 4a). Although present in the nucleus, DPPA3 was far more abundant in the cytoplasmic fraction (Supplementary Fig. 4e). Furthermore, DPPA3 did not bind to mitotic chromosomes indicating a low or absent chromatin association of DPPA3 in ESCs (Fig. 4a). To further understand the mechanistic basis of DPPA3-dependent DNA demethylation in ESCs, we performed FLAG-DPPA3 pulldowns followed by liquid chromatography tandem mass spectrometry (LC-MS/MS) to profile the DPPA3 interactome in naïve ESCs. Strikingly, among the 303 significantly enriched DPPA3 interaction partners identified by mass spectrometry, we found both UHRF1 and DNMT1 (Fig. 4b and Supplementary Data 3), the core components of the DNA maintenance methylation machinery[80,81]. A reciprocal immunoprecipitation of UHRF1 confirmed its interaction with DPPA3 in ESCs (Supplementary Fig. 4g). Moreover, GO analysis of the top 131 interactors of DPPA3 in ESCs showed the two most enriched GO terms to be related to DNA methylation (Supplementary Data 4). These findings are consistent with previous studies implicating DPPA3 in the regulation of maintenance methylation in other cellular contexts[60,62]. We also detected multiple members of the nuclear transport machinery in our DPPA3 interactome (highlighted in purple, Fig. 4b and Supplementary Data 3), which prompted us to investigate whether DPPA3 influences the subcellular localization of UHRF1. Surprisingly, biochemical fractionation experiments revealed UHRF1 to be present in both the nucleus and cytoplasm of naïve wt ESCs (Supplementary Fig. 4f). Despite comparable total UHRF1 protein levels in wt and Dppa3KO ESCs (Supplementary Fig. 4h), loss of DPPA3 completely abolished the cytoplasmic fraction of UHRF1 (Supplementary Fig. 4f).

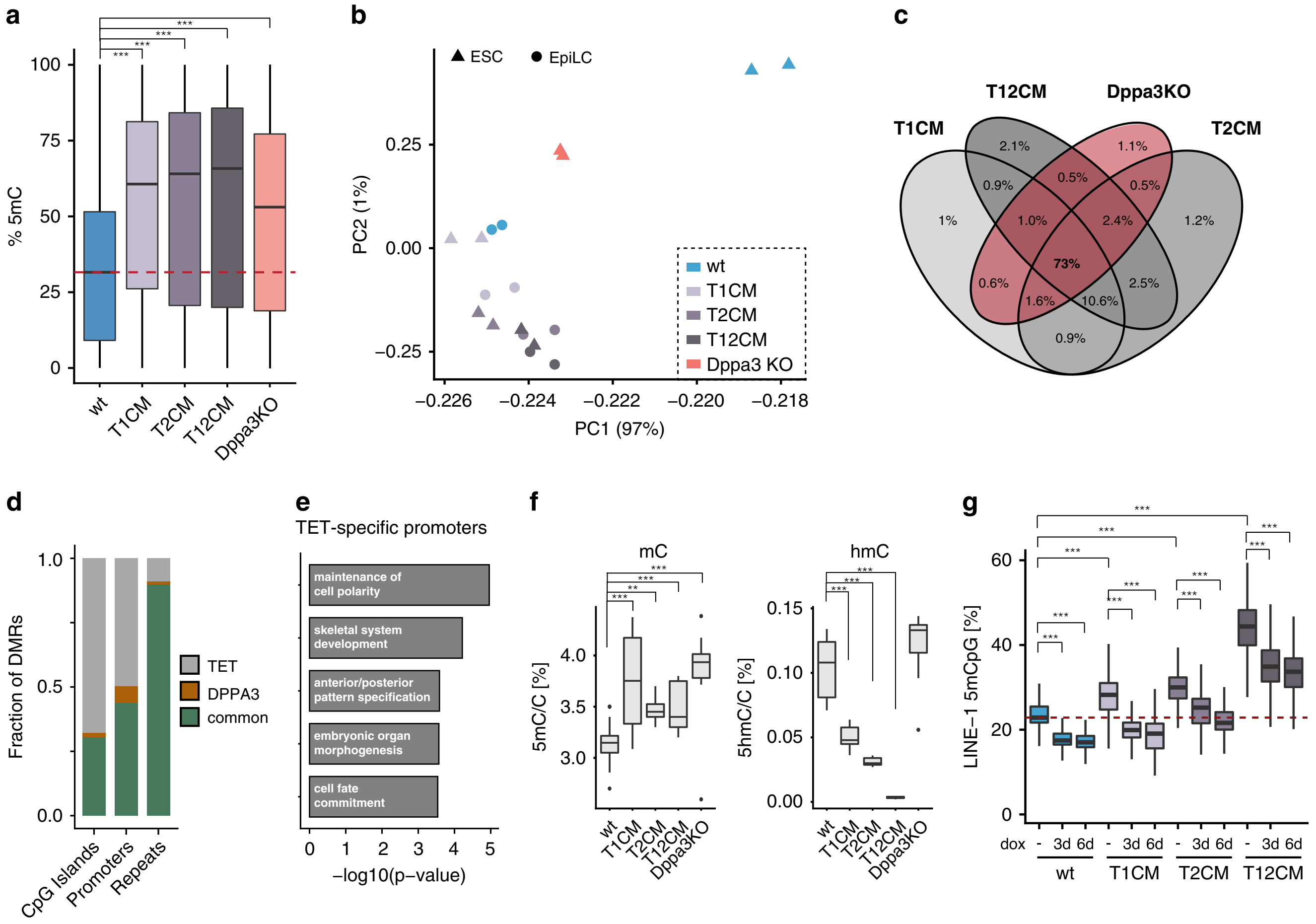


Fig. 3 DPPA3 acts downstream of TET1 and TET2 to establish and preserve global hypomethylation. a *Dppa3* loss results in global hypermethylation. Percentage of total 5mC as measured by RRBS using $n = 2$ biologically independent samples per condition. **b** *Dppa3* prevents the premature acquisition of a primed methylome. Principal component (PC) analysis of RRBS data from wt, T1CM, T2CM, and T12CM ESCs and EpiLCs and Dppa3KO ESCs. **c** DPPA3 and TET proteins promote demethylation of largely similar targets. Venn Diagram depicting the overlap of hypermethylated sites among T1CM, T2CM, T12CM, and Dppa3KO ESCs. **d** *Dppa3* protects mostly repeats from hypermethylation. Fraction of hypermethylated genomic elements classified as TET-specific (only hypermethylated in TET mutant ESCs), DPPA3-specific (only hypermethylated in Dppa3KO ESCs), or common (hypermethylated in TET mutant and Dppa3KO ESCs). **e** Gene ontology (GO) terms associated with promoters specifically dependent on TET activity; adjusted p-values calculated using a two-sided Fisher's exact test followed by Benjamini-Hochberg correction for multiple testing. **f** TET activity remains unaffected in Dppa3KO ESCs. DNA modification levels for 5-methylcytosine (5mC) and 5-hydroxymethylcytosine (5hmC) as measured by mass spectrometry (LC-MS/MS) in wt ($n = 24$), T1CM ($n = 8$), T2CM ($n = 12$), T12CM ($n = 11$), Dppa3KO ($n = 12$) mESC biological replicates. **g** *Dppa3* expression can rescue the hypermethylation in TET mutant ESCs. DNA methylation levels at LINE-1 elements (%) as measured by bisulfite sequencing 0, 3, or 6 days after doxycycline (dox) induction of *Dppa3* expression using $n = 2$ replicates per condition. The dashed red line indicates the median methylation level of wt ESCs. In the boxplots in (**a**, **f** and **g**), horizontal black lines within boxes represent median values, boxes indicate the upper and lower quartiles, and whiskers extend to the most extreme value within 1.5 x the interquartile range from each hinge. In (**a**, **f**, and **g**), *p*-values were calculated using Welch's two-sided *t*-test: ***$p < 2\text{e}{-16}$. Source data are provided as a Source Data file.

As maintenance DNA methylation critically depends on the correct targeting and localization of UHRF1 within the nucleus[82–85], we asked whether TET-dependent regulation of DPPA3 might affect the subnuclear distribution of UHRF1. To this end, we tagged endogenous UHRF1 with GFP in wild-type (U1G/wt) as well as Dppa3KO and T12CM ESCs (U1G/Dppa3KO and U1G/T12CM, respectively) enabling us to monitor UHRF1 localization dynamics in living cells (Supplementary Fig. 4b, d). Whereas UHRF1-GFP localized to both the nucleus and cytoplasm of wt ESCs, UHRF1-GFP localization was solely nuclear in Dppa3KO and T12CM ESCs (Supplementary Fig. 4i, j). In addition, UHRF1 appeared to display a more diffuse localization in wt ESCs compared to Dppa3KO and T12CM ESCs, in which we observed more focal patterning of UHRF1 particularly at heterochromatic foci (Supplementary Fig. 4i). To quantify this observation, we calculated the coefficient of variation (CV) of nuclear UHRF1-GFP among wt, Dppa3KO, and T12CM ESCs. The CV of a fluorescent signal correlates with its distribution, with low CV values reflecting more homogenous distributions and high CV values corresponding to more heterogeneous distributions[86,87]. Indeed, the pronounced focal accumulation of UHRF1-GFP observed in Dppa3KO and T12CM ESCs corresponded with a highly significant increase in the CV values of nuclear UHRF1-GFP compared with wt ESCs (Supplementary Fig. 4i, j).

To assess whether these differences in nuclear UHRF1 distribution reflected altered chromatin binding, we used fluorescence recovery after photobleaching (FRAP) to study the dynamics of nuclear UHRF1-GFP in wt, Dppa3KO, and T12CM ESCs. Our FRAP analysis revealed markedly increased UHRF1

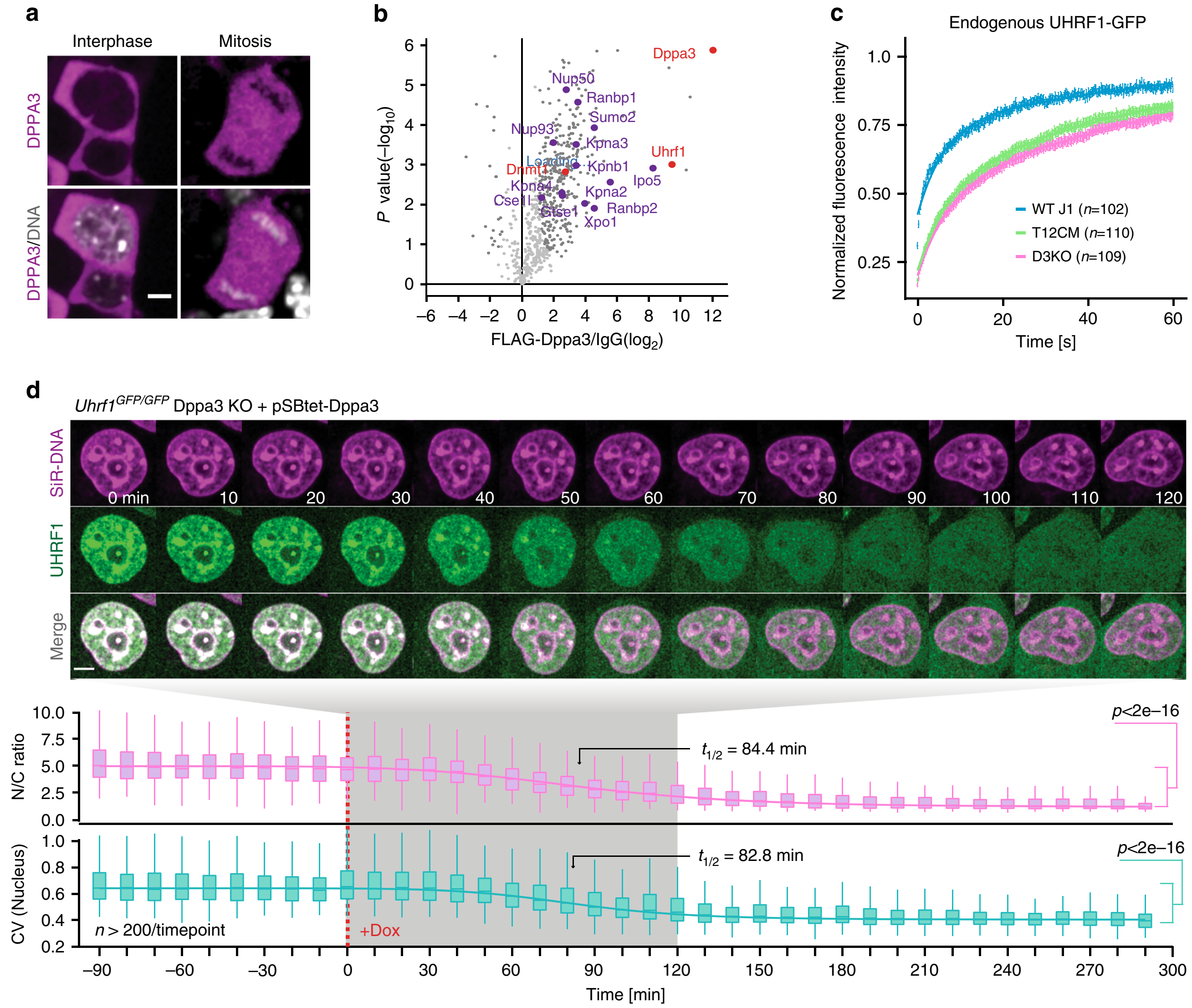


Fig. 4 TET-dependent expression of DPPA3 alters UHRF1 localization and chromatin binding in naïve ESCs. a Localization of endogenous DPPA3-HALO in live ESCs counterstained with SiR-Hoechst (DNA). Representative result, $n \geq 4$. Scale bar: 5 µm. **b** Volcano plot from DPPA3-FLAG pulldowns in ESCs. Dark gray dots: significantly enriched proteins. Red dots: proteins involved in DNA methylation regulation. Purple dots: proteins involved in nuclear transport. anti-FLAG antibody: $n = 3$ biological replicates, IgG control antibody: $n = 3$ biological replicates. Statistical significance determined by performing a Student's t test with a permutation-based FDR of 0.05 and an additional constant $S0 = 1$. **c** FRAP analysis of endogenous UHRF1-GFP. Each genotype comprises the combined single-cell data from two independent clones acquired in two independent experiments. **d** Localization dynamics of endogenous UHRF1-GFP in response to *Dppa3* induction in U1G/D3KO + pSBtet-D3 ESCs with confocal timelapse imaging over 8 h (10 min intervals). $t = 0$ corresponds to start of *Dppa3* induction with doxycycline (+Dox). (*top panel*) Representative images of UHRF1-GFP and DNA (SiR-Hoechst stain) throughout confocal timelapse imaging. Scale bar: 5 µm. (middle panel) Nucleus to cytoplasm ratio (N/C ratio) of endogenous UHRF1-GFP signal. (bottom panel) Coefficient of variance (CV) of endogenous UHRF1-GFP intensity in the nucleus. (*middle and bottom panel*) N/C ratio and CV values: measurements in $n > 200$ single cells per time point (precise values can be found in the Source Data file), acquired at $n = 16$ separate positions. Curves represent fits of four parameter logistic (4PL) functions to the N/C ratio (pink line) and CV (green line) data. Live-cell imaging was repeated three times with similar results. In (**c**), the mean fluorescence intensity of n cells (indicated in the plots) at each timepoint are depicted as shaded dots. Error bars indicate mean ± SEM. Curves (solid lines) indicate double-exponential functions fitted to the FRAP data. In the boxplots in (**d**), darker horizontal lines within boxes represent median values. The limits of the boxes indicate upper and lower quartiles, and whiskers extend to the most extreme value within 1.5 x the interquartile range from each hinge. P-values based on Welch's two-sided t test. Source data are provided as a Source Data file.

chromatin binding in both Dppa3KO and T12CM ESCs as demonstrated by the significantly slower recovery of UHRF1-GFP in these cell lines compared to wt ESCs (Fig. 4c and Supplementary Fig. 4k, l). These data confirmed the notion that the more pronounced focal patterning of nuclear UHRF1 observed in Dppa3KO and T12CM ESCs (Supplementary Fig. 4i, j) was indeed a consequence of increased UHRF1 chromatin binding. Interestingly, although strongly reduced compared to wt ESCs, UHRF1 mobility was slightly higher in T12CM ESCs than Dppa3KO ESCs, consistent with a severe but not total loss of *Dppa3* in the absence of TET activity (Supplementary Fig. 4m). Induction of ectopic *Dppa3* rescued the cytoplasmic fraction of UHRF1 (N/C ratio: Fig. 4d) as well as the diffuse localization of nuclear UHRF1 in Dppa3KO ESCs (CV: Fig. 4d), which reflected a striking increase in the mobility of residual nuclear UHRF1-GFP as assessed by FRAP (Supplementary Figs. 4n and 5a, b).

Our analysis also revealed that the nuclear export of UHRF1 and the inhibition of UHRF1 chromatin binding caused by *Dppa3* induction occur with almost identical kinetics (N/C $t_{1/2} = 84.4$ min; CV $t_{1/2} = 82.8$) (Fig. 4d). UHRF1 is required for the proper targeting of DNMT1 to DNA replication sites and therefore essential for DNA methylation maintenance[80,81]. We observed a marked reduction of both UHRF1 and DNMT1 at replication foci upon induction of *Dppa3*, indicating that DPPA3 promotes hypomethylation in naïve ESCs by impairing DNA methylation maintenance (Supplementary Fig. 5c, d). Ectopic expression of DPPA3 not only altered the subcellular distribution of endogenous UHRF1 in mouse ESCs (Fig. 4d and Supplementary Fig. 5e) but also in human ESCs suggesting evolutionary conservation of this mechanism among mammals (Supplementary Fig. 5f, g). Collectively, our results demonstrate that TET proteins control both the subcellular localization and chromatin binding of UHRF1 in naïve ESCs via the regulation of DPPA3 levels. Furthermore, these data show that DPPA3 is both necessary and sufficient for ensuring the nucleocytoplasmic translocation, diffuse nuclear localization, and attenuated chromatin binding of UHRF1 in ESCs.

DPPA3-mediated demethylation is achieved via inhibition of UHRF1 chromatin binding and attenuated by nuclear export. Our results demonstrated that *Dppa3* induction causes UHRF1 to be released from chromatin and exported to the cytoplasm near simultaneously (Fig. 4d, Supplementary Figs. 4n and 5a, b). In principle, either a reduction in the nuclear concentration of UHRF1 or the impairment of UHRF1 chromatin binding alone would suffice to compromise effective maintenance DNA methylation[84,88]. To dissect the contribution of these distinct modes of disrupting UHRF1 activity to DPPA3-mediated DNA demethylation in naïve ESCs, we generated inducible *Dppa3*-mScarlet expression cassettes (Supplementary Fig. 6a) harboring mutations to residues described to be critical for its nuclear export (ΔNES)[61] and the interaction with UHRF1 (KRR and R107E)[62], as well as truncated forms of DPPA3 found in zygotes, 1-60 and 61-150[78] (Fig. 5a). After introducing these *Dppa3* expression cassettes into U1GFP/Dppa3KO ESCs, we used live-cell imaging to track each DPPA3 mutant's localization and ability to rescue the Dppa3KO phenotype (Fig. 5b). DPPA3-ΔNES and DPPA3 61-150, which both lacked a functional nuclear export signal, were retained in the nucleus (Fig. 5b). In contrast DPPA3-WT as well as the DPPA3-KRR, DPPA3-R107E, and DPPA3 1-60 mutants localized primarily to the cytoplasm (Fig. 5b), closely mirroring the localization of endogenous DPPA3 in naïve ESCs (Fig. 4a). However, all tested DPPA3 mutants failed to efficiently reestablish nucleocytoplasmic translocation of UHRF1 (Fig. 5b and Supplementary Fig. 6b), indicating that the DPPA3-UHRF1 interaction and nuclear export of DPPA3 are both required for the shuttling of UHRF1 from the nucleus to the cytoplasm in naïve ESCs.

Nevertheless, DPPA3-ΔNES and DPPA3 61-150 managed to significantly disrupt the focal pattern and heterochromatin association of UHRF1 within the nucleus, with DPPA3-ΔNES causing a more diffuse localization of nuclear UHRF1 than DPPA3-WT (Fig. 5b and Supplementary Fig. 6c). In contrast, the loss or mutation of residues critical for its interaction with UHRF1 compromised DPPA3's ability to effectively restore the diffuse localization of nuclear UHRF1 (Fig. 5b and Supplementary Fig. 6c). FRAP analysis revealed that the disruption or deletion of the UHRF1 interaction interface (DPPA3-KRR, DPPA3-R107E, DPPA3 1-60) severely diminished the ability of DPPA3 to release UHRF1 from chromatin (Fig. 5c and Supplementary Fig. 6f–k). On the other hand, the C-terminal half of DPPA3, lacking a nuclear export signal but retaining UHRF1 interaction, came close to fully restoring the mobility of UHRF1 (Fig. 5c and Supplementary Fig. 6i–k). DPPA3-ΔNES mobilized UHRF1 to a greater extent than DPPA3-WT (Fig. 5c and Supplementary Fig. 6d, e, j, k), suggesting that active nuclear export might antagonize DPPA3-mediated inhibition of UHRF1 chromatin binding. Supporting this notion, chemical inhibition of nuclear export using leptomycin-B (LMB) significantly enhanced the inhibition of UHRF1 chromatin binding in U1G/D3KO ESCs expressing DPPA3-WT (Supplementary Fig. 5h–k). Taken together, our data show that the efficiency of DPPA3-dependent release of UHRF1 from chromatin requires its interaction with UHRF1 but not its nuclear export.

To further address the question whether the nucleocytoplasmic translocation of UHRF1 and impaired UHRF1 chromatin binding both contribute to DPPA3-mediated inhibition of DNA methylation maintenance, we assessed the ability of each DPPA3 mutant to rescue the hypermethylation of LINE-1 elements in Dppa3KO ESCs (Fig. 5d). Strikingly, DPPA3-ΔNES fully rescued the hypermethylation and achieved a greater loss of DNA methylation than DPPA3-WT, whereas DPPA3 mutants lacking the residues important for UHRF1 binding failed to restore low methylation levels (Fig. 5d). Overall, the ability of each DPPA3 mutant to reduce DNA methylation levels closely mirrored the extent to which each mutant impaired UHRF1 chromatin binding (Fig. 5c and Supplementary Fig. 6d–k). In line with the high mobility of UHRF1 achieved by the DPPA3-ΔNES, (Fig. 5c, Supplementary Figs. 5h–k and 6d, e, j, k), nuclear export is not only dispensable for DPPA3-mediated demethylation, but attenuates the ability of DPPA3 to inhibit maintenance methylation (Fig. 5d). Collectively, our findings demonstrate the inhibition of UHRF1 chromatin binding, as opposed to nucleocytoplasmic translocation of UHRF1, to be the primary mechanism by which DPPA3 drives hypomethylation in naïve ESCs.

DPPA3 binds nuclear UHRF1 with high affinity prompting its release from chromatin in ESCs. Next, we set out to investigate the mechanistic basis of DPPA3's ability to inhibit UHRF1 chromatin binding in naïve ESCs. DPPA3 has been reported to specifically bind H3K9me2[77], a histone modification critical for UHRF1 targeting[84,89,90]. These prior findings led us to consider two possible mechanistic explanations for DPPA3-mediated UHRF1 inhibition in naïve ESCs: (1) DPPA3 blocks access of UHRF1 to chromatin by competing in binding to H3K9me2, (2) DPPA3 directly or indirectly binds to UHRF1 and thereby prevents it from accessing chromatin.

To simultaneously assess the dynamics of both UHRF1 and DPPA3 under physiological conditions in live ES cells, we employed raster image correlation spectroscopy with pulsed interleaved excitation (PIE-RICS) (Fig. 6a). RICS is a confocal imaging method that measures the diffusive properties of fluorescently labeled molecules, and thereby also their binding, in living cells. Using images acquired on a laser scanning confocal microscope, spatiotemporal information of fluorescently labeled proteins can be extracted from the shape of the spatial autocorrelation function (SACF). A diffusive model is fitted to the SACF which yields the average diffusion coefficient, the concentration, and the fraction of quickly diffusing and slowly diffusing (in this case, bound) molecules[91]. If two proteins are labeled with distinct fluorophores and imaged simultaneously with separate detectors, the extent of their interaction can be extracted from the cross-correlation of their fluctuations using cross-correlation RICS (ccRICS) (Fig. 6a)[92].

We first measured the mobility of DPPA3-mScarlet variants expressed in U1GFP/D3KOs (Supplementary Fig. 7a, b). The RICS analysis revealed that, over the timescale of the

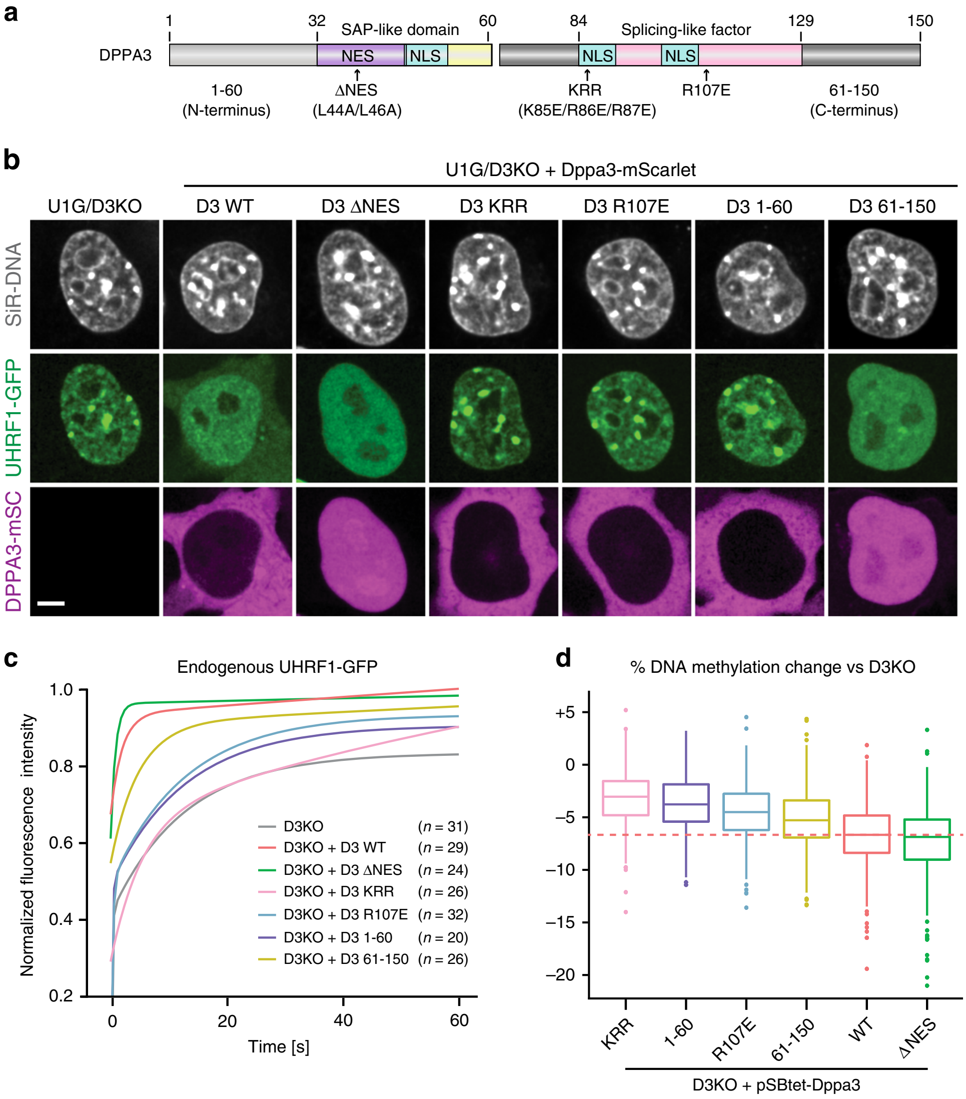


Fig. 5 DPPA3-mediated demethylation is achieved via inhibition of UHRF1 chromatin binding and attenuated by nuclear export. a Schematic illustration of murine DPPA3 with the nuclear localization signals (NLS), nuclear export signal (NES), and predicted domains (SAP-like and splicing factor-like[210]) annotated. For the DPPA3 mutant forms used in this study, point mutations are indicated with arrows (ΔNES, KRR, R107E) and the two truncations are denoted by the middle break (1–60, left half; 61–150, right half). **b**, **c** Nuclear export and the N-terminus of DPPA3 are dispensable for disrupting focal UHRF1 patterning and chromatin binding in ESCs. **b** Representative confocal images illustrating the localization of endogenous UHRF1-GFP and the indicated mDPPA3-mScarlet fusions in live U1G/D3KO + pSB-D3-mSC ESCs after doxycycline induction. DNA counterstain: SiR-Hoechst. Scale bar: 5 µm. **c** FRAP analysis of endogenous UHRF1-GFP in U1G/D3KO ESCs expressing the indicated mutant forms of DPPA3. FRAP Curves (solid lines) indicate double-exponential functions fitted to the FRAP data acquired from *n* cells (shown in the plots). For single-cell FRAP data and additional quantification, see Supplementary Fig. 6d–k. **d** DPPA3-mediated inhibition of UHRF1 chromatin binding is necessary and sufficient to promote DNA demethylation. Percentage of DNA methylation change at LINE-1 elements (%) in D3KO ESCs after induction of the indicated mutant forms of *Dppa3* as measured by bisulfite sequencing of *n* = 4 biological replicates. In the boxplot in (**d**), horizontal lines within boxes represent median values, boxes indicate the upper and lower quartiles, whiskers extend to the most extreme value within 1.5 x the interquartile range from each hinge, and dots indicate outliers. Source data are provided as a Source Data file.

measurements, nuclear DPPA3-WT was predominantly unbound from chromatin and freely diffusing through the nucleus at a rate of 7.18 ± 1.87 µm^2/s (Supplementary Fig. 7f). The fraction of mobile DPPA3-mScarlet molecules was measured to be 88.4 ± 5.2% (Fig. 6f), validating the globally weak binding inferred from ChIP-Seq profiles[76]. These mobility parameters were largely unaffected by disruption of the UHRF1 interaction, with the DPPA3-KRR mutant behaving similarly to wild-type DPPA3 (Fig. 6f and Supplementary Fig. 7f). To rule out a potential competition between UHRF1 and DPPA3 for H3K9me2 binding, we next used RICS to determine if DPPA3 dynamics are altered in the absence of UHRF1. For this purpose, we introduced the DPPA3-WT-mScarlet cassette into Uhrf1KO (U1KO) ESCs[93], in which free eGFP is expressed from the endogenous *Uhrf1* promoter (Supplementary Fig. 7c). However, neither the diffusion rate nor the mobile fraction of DPPA3 were appreciably altered in cells devoid of UHRF1, suggesting the high fraction of unbound DPPA3 to be unrelated to the presence of UHRF1 (Fig. 6f and Supplementary Fig. 7f). Overall, our RICS data demonstrate that, in contrast to zygotes[77], DPPA3 in ESCs lacks a strong capacity for chromatin binding, and, as such, is not engaged in competition with UHRF1 for chromatin binding.

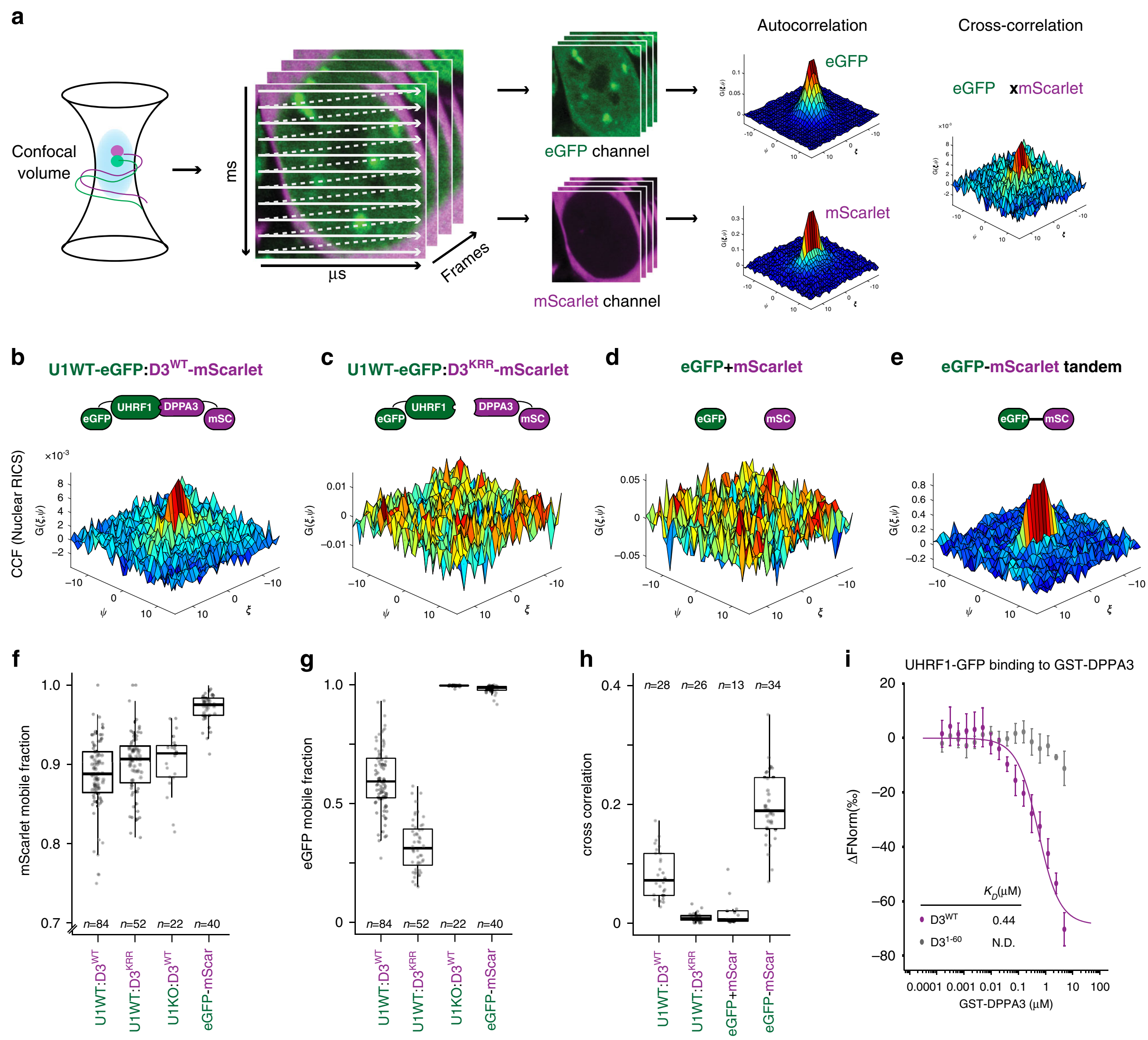


Fig. 6 DPPA3 binds nuclear UHRF1 with high affinity prompting its release from chromatin in naïve ESCs. a Overview of RICS and ccRICS. Confocal image series are acquired on a laser scanning confocal microscope, containing spatiotemporal fluorescence information on the microsecond and millisecond timescales. A spatial autocorrelation function (SACF) is calculated from the fluorescence image and fit to a diffusive model. The cross-correlation of intensity between two channels is used to estimate the co-occurrence of two fluorescent molecules in live cells. The mean cross-correlation of the fluctuations is calculated and shown in the 3D plot color-coded according to the correlation value. **b**–**e** Representative plots of the spatial cross-correlation function (SCCF) between the depicted fluorescent molecules in cells from each cell line measured: (**b**) wild-type (U1WT:D3^{WT}) and (**c**) K85E/R85E/R87E DPPA3 mutant (U1WT:D3^{KRR}), and control ESCs expressing (**d**) free eGFP, free mScarlet (eGFP + mScarlet) and (**e**) an eGFP-mScarlet tandem fusion (eGFP-mScarlet). **f**, **g** Mobile fraction of (**f**) mScarlet and (**g**) eGFP species in the cell lines depicted in (**b**, **c**, and **e**) and in Uhrf1KO ESCs expressing free eGFP and wild-type DPPA3-mScarlet (U1KO:D3^{WT}). The mobile fraction was derived from a two-component model fit of the autocorrelation function. Data are pooled from three (U1WT:D3^{WT}, U1WT:D3^{KRR}) or two (U1KO:D3^{WT}, eGFP-mScar) independent experiments. **h** Mean cross-correlation values of mobile eGFP and mScarlet measured in the cell lines depicted in (**b**–**e**). The spatial lag in the x-dimension (sensitive to fast fluctuations) is indicated by ξ, and the spatial lag in the y-dimension (sensitive to slower fluctuations) is indicated by ψ. Data are pooled from two independent experiments. **i** Microscale thermophoresis measurements of UHRF1-eGFP binding to GST-DPPA3 WT (D3^{WT}) or GST-DPPA3 1-60 (D3$^{1\text{-}60}$). Error bars indicate the mean ± SEM of $n = 2$ technical replicates from $n = 4$ independent experiments. In (**f**–**h**), each data point represents the measured and fit values from a single cell where n = number of cells measured (indicated in the plots). In the boxplots, darker horizontal lines within boxes represent median values. The limits of the boxes indicate the upper and lower quartiles; the whiskers extend to the most extreme value within 1.5 x the interquartile range from each hinge. Source data are provided as a Source Data file.

We next used RICS to analyze the dynamics of UHRF1-GFP in response to DPPA3 induction (Fig. 6a). In cells expressing DPPA3-KRR, RICS measurements revealed that only 32.4 ± 10% of UHRF1 is mobile, indicating that the majority of UHRF1 is chromatin-bound (Fig. 6g). In contrast, expression of wild-type DPPA3 leads to a dramatic increase in the mobile fraction of UHRF1 (60.6 ± 13.7% mobile fraction for UHRF1) (Fig. 6g and Supplementary Fig. 7g, h). Furthermore, the mobile fraction of

UHRF1 increased as a function of the relative abundance of nuclear DPPA3 to UHRF1 (Supplementary Fig. 7i), thereby indicating a stoichiometric effect of DPPA3 on UHRF1 chromatin binding, consistent with physical interaction. Thus, these results demonstrate that DPPA3 potently disrupts UHRF1 chromatin binding in live ESCs and suggest its interaction with UHRF1 to be critical to do so.

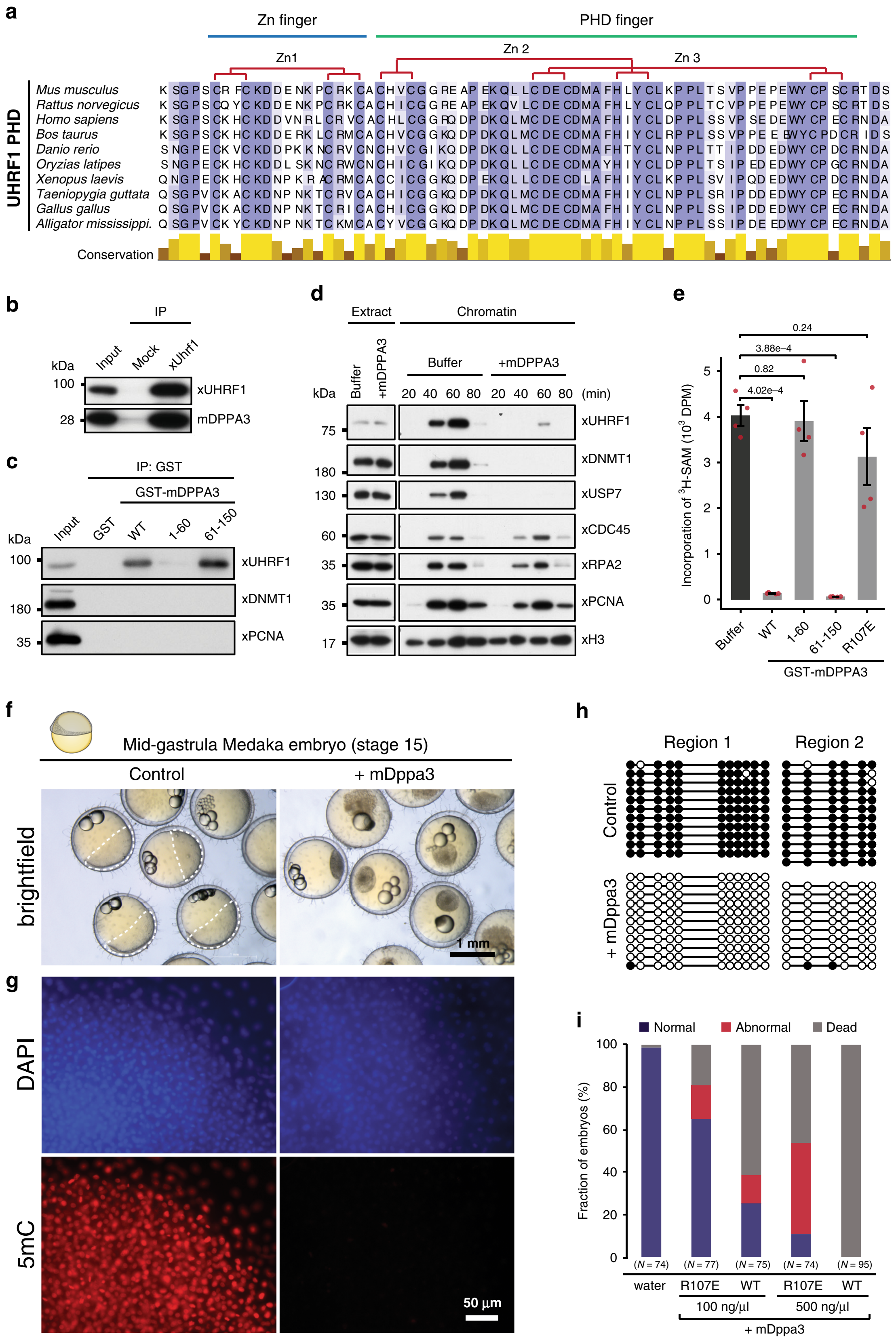
To determine whether such an interaction is indeed present in the nuclei of live ESCs, we performed cross-correlation RICS (ccRICS) (Fig. 6a). We first validated ccRICS in ESCs by analyzing live cells expressing a tandem eGFP-mScarlet fusion (Fig. 6e and Supplementary Fig. 7d), or expressing both freely diffusing eGFP and mScarlet (Fig. 6d and Supplementary Fig. 7e). For the tandem eGFP-mScarlet fusion, we observed a clear positive cross-correlation indicative of eGFP and mScarlet existing in the same complex (Fig. 6e, h), as would be expected for an eGFP-mScarlet fusion. On the other hand, freely diffusing eGFP and mScarlet yielded no visible cross-correlation (Fig. 6d, h), consistent with two independent proteins that do not interact. Upon applying ccRICS to nuclear UHRF1-GFP and DPPA3-mScarlet, we observed a prominent cross-correlation between wild-type DPPA3 and the primarily unbound fraction of UHRF1 (Fig. 6b, h), indicating that mobilized UHRF1 exists in a high affinity complex with DPPA3 in live ESCs. In marked contrast, DPPA3-KRR and UHRF1-GFP failed to exhibit detectable cross-correlation (Fig. 6c, h), consistent with the DPPA3-KRR mutant's diminished capacity to bind[62] and mobilize UHRF1 (Fig. 5c and Supplementary Fig. 6f, j, k). Overall, these findings demonstrate that nuclear DPPA3 interacts with UHRF1 to form a highly mobile complex in naïve ESCs which precludes UHRF1 chromatin binding.

To determine whether the DPPA3-UHRF1 complex identified in vivo (Fig. 6h) corresponds to a high affinity direct interaction, we performed microscale thermophoresis (MST) measurements using recombinant UHRF1-GFP and DPPA3 proteins. MST analysis revealed a direct and high affinity (K_D: 0.44 µM) interaction between the DPPA3 WT and UHRF1 (Fig. 6i). No binding was observed for DPPA3 1-60, lacking the residues essential for interaction with UHRF1 (Fig. 6i). In line with the results obtained by ccRICS, these data support the notion that DPPA3 directly binds UHRF1 in vivo. Interestingly, the affinity of the UHRF1-DPPA3 interaction was comparable or even greater than that reported for the binding of UHRF1 to H3K9me3 or unmodified H3 peptides, respectively[94,95].

To better understand how UHRF1 chromatin loading is impaired by its direct interaction with DPPA3, we applied a fluorescent-three-hybrid (F3H) assay to identify the UHRF1 domain bound by DPPA3 in vivo (Supplementary Fig. 7j, k). In short, this method relies on a cell line harboring an array of *lac* operator binding sites in the nucleus at which a GFP-tagged "bait" protein can be immobilized and visualized as a spot. Thus, the extent of recruitment of an mScarlet-tagged "prey" protein to the nuclear GFP spot offers a quantifiable measure of the interaction propensity of the "bait" and "prey" proteins in vivo (Supplementary Fig. 7k)[96]. Using UHRF1-GFP domain deletions as the immobilized bait (Supplementary Fig. 7j, k), we assessed how the loss of each domain affected the recruitment of DPPA3-mScarlet to the GFP spot. In contrast to the other UHRF1 domain deletions, removal of the PHD domain essentially abolished recruitment of DPPA3 to the *lac* spot, demonstrating DPPA3 binds UHRF1 via its PHD domain in vivo (Supplementary Fig. 7l, m). The PHD of UHRF1 is essential for its recruitment to chromatin[88,95,97], ubiquitination of H3 and recruitment of DNMT1 to replication foci[82,83]. Thus, our in vivo results suggest that the high affinity interaction of DPPA3 with UHRF1's PHD domain precludes UHRF1 from binding chromatin in ESCs, which is also supported by a recent report demonstrating that DPPA3 specifically binds the PHD domain of UHRF1 to competitively inhibit H3 tail binding in vitro[98].

DPPA3 can inhibit UHRF1 function and drive global DNA demethylation in distantly related, non-mammalian species. Whereas UHRF1 and TET proteins are widely conserved throughout plants and vertebrates[99,100], both early embryonic global hypomethylation[101] and the *Dppa3* gene are unique to mammals. Consistent with UHRF1's conserved role in maintenance DNA methylation, a multiple sequence alignment of UHRF1's PHD domain showed that the residues critical for the recognition of histone H3 are completely conserved from mammals to invertebrates (Fig. 7a). This prompted us to consider the possibility that DPPA3 might be capable of modulating the function of distantly related UHRF1 homologs outside of mammals. To test this hypothesis, we used amphibian (*Xenopus laevis*) egg extracts to assess the ability of mouse DPPA3 (mDPPA3) to interact with a non-mammalian form of UHRF1. Despite the 360 million years evolutionary distance between mouse and *Xenopus*[102], mDPPA3 not only bound *Xenopus* UHRF1 (xUHRF1) with high affinity (Fig. 7b, c and Supplementary Fig. 8a, b) it also interacted with xUHRF1 specifically via its PHD domain (Supplementary Fig. 8c–e). Moreover, the first 60 amino acids of DPPA3 were dispensable for its interaction with UHRF1 (Supplementary Fig. 8a, b). Interestingly, mutation to R107, reported to be critical for DPPA3's binding with mouse UHRF1[62], diminished but did not fully disrupt the interaction (Supplementary Fig. 8b, e). The R107E mutant retained the ability to bind the xUHRF1-PHD domain but exhibited decreased binding to xUHRF1-PHD-SRA under high-salt conditions (Supplementary Fig. 8e), suggesting that R107E changes the binding mode of mDPPA3 to xUHRF1, rather than inhibiting the complex formation. Considering the remarkable similarity between DPPA3's interaction with mouse and *Xenopus* UHRF1, we reasoned that the ability of DPPA3 to inhibit UHRF1 chromatin binding and maintenance DNA methylation might be transferable to *Xenopus*. To address this, we took advantage of a cell-free system derived from interphase *Xenopus* egg extracts to reconstitute DNA maintenance methylation[82]. Remarkably, recombinant mDPPA3 completely disrupted chromatin binding of both *Xenopus* UHRF1 and DNMT1 without affecting the loading of replication factors such as xCDC45, xRPA2, and xPCNA (Fig. 7d). We determined that the inhibition of xUHRF1 and xDNMT1 chromatin loading only requires DPPA3's C-terminus (61-150 a.a.) and is no longer possible upon mutation of R107 (R107E) (Supplementary Fig. 8h), in line with our results in mouse ESCs (Fig. 5d). Moreover, DPPA3-mediated inhibition of xUHRF1 chromatin loading resulted in the severe perturbation of histone H3 dual-monoubiquitylation (H3Ub2), which is necessary for the recruitment of DNMT1[82,83,103] (Supplementary Fig. 8f). To determine whether mDPPA3 can displace xUHRF1 already bound to chromatin, we first depleted *Xenopus* egg extracts of xDNMT1 to stimulate the hyper-accumulation of xUHRF1 on chromatin[82,104] and then added recombinant mDPPA3 after S-phase had commenced (Supplementary Fig. 8g). Under these conditions, both wild-type mDPPA3 and the 61-150 fragment potently displaced xUHRF1 from chromatin, leading to suppressed H3 ubiquitylation (Supplementary Fig. 8g).

We next assessed the effect of DPPA3 on *Xenopus* maintenance DNA methylation. Consistent with the severe disruption of xDNMT1 chromatin loading, both DPPA3 wild-type and 61–150 effectively abolished replication-dependent DNA methylation in *Xenopus* egg extracts (Fig. 7e). In contrast, DPPA3 1-60 and DPPA3 R107E, which both failed to suppress xUHRF1 and



xDNMT1 binding, did not significantly alter maintenance DNA methylation activity (Fig. 7e and Supplementary Fig. 8d, e). Taken together, our data demonstrate DPPA3 to be capable of potently inhibiting maintenance DNA methylation in a non-mammalian system.

These findings raised the question whether a single protein capable of inhibiting UHRF1 function like DPPA3 could establish a mammalian-like global hypomethylation during the early embryonic development of a non-mammalian organism. To explore this possibility we turned to the biomedical model fish

Fig. 7 DPPA3 evolved in boreoeutherian mammals but also functions in lower vertebrates. a Protein sequence alignment of the PHD domain of the UHRF1 family. **b** Endogenous xUHRF1 binds mDPPA3. IPs were performed on *Xenopus* egg extracts incubated with FLAG-mDPPA3 using either a control (Mock) or anti-xUHRF1 antibody and then analyzed by immunoblotting using the indicated antibodies. Representative of $n = 3$ independent experiments. **c** GST-tagged mDPPA3 wild-type (WT), point mutant R107E, and truncations (1–60 and 61–150) were immobilized on GSH beads and incubated with *Xenopus* egg extracts. Bound proteins were analyzed using the indicated antibodies. Representative of $n = 3$ independent experiments. **d** Sperm chromatin was incubated with interphase *Xenopus* egg extracts supplemented with buffer (+buffer) or GST-mDPPA3 (+mDPPA3). Chromatin fractions were isolated and subjected to immunoblotting using the antibodies indicated. Representative of $n = 3$ independent experiments. **e** The efficiency of maintenance DNA methylation was assessed by the incorporation of radiolabelled methyl groups from S-[methyl-^{3}H]-adenosyl-L-methionine (^{3}H-SAM) into DNA purified from egg extracts. Disintegrations per minute (DPM). Error bars indicate mean ± SD calculated from $n = 4$ independent experiments. Depicted p-values based on Welch's two-sided t-test. **f** Representative images of developing mid-gastrula stage embryos (control injection) and arrested, blastula stage embryos injected with *mDppa3*. Injections were performed on one-cell stage embryos and images were acquired ~18 h after fertilization. **g** Immunofluorescence staining of 5mC in control and *mDppa3*-injected medaka embryos at the late blastula stage (~8 h after fertilization). Images are representative of $n = 3$ independent experiments. DNA counterstain: DAPI,4′,6-diamidino-2-phenylindole. **h** Bisulfite sequencing of two intergenic regions (Region 1: chr20:18,605,227-18,605,449, Region 2: chr20:18,655,561-18,655,825) in control and *mDppa3*-injected medaka embryos at the late blastula stage. **i** Percentage of normal, abnormal, or dead medaka embryos. Embryos were injected with wild-type *mDppa3* (WT) or *mDppa3* R107E (R107E) at two different concentrations (100 ng/µl or 500 ng/µl) or water at the one-cell stage and analyzed ~18 h after fertilization. N = number of embryos from $n = 3$ independent injection experiments. Source data are provided as a Source Data file.

medaka (*Oryzias latipes*), which does not exhibit genome-wide erasure of DNA methylation[105] and diverged from mammals 450 million years ago[102]. We injected medaka embryos with *Dppa3* mRNA at the one-cell stage and then tracked their developmental progression. Remarkably, medaka embryos injected with *Dppa3* failed to develop beyond the blastula stage (Fig. 7f) and exhibited a near-complete elimination of global DNA methylation as assessed by immunofluorescence and bisulfite sequencing (Fig. 7g, h). DPPA3-mediated DNA methylation loss was both dose dependent and sensitive to the R107E mutation, which induced only partial demethylation (Supplementary Fig. 8i). Interestingly, medaka embryos injected with DPPA3 R107E showed far fewer developmental defects than those injected with wild-type DPPA3 (Fig. 7i), suggesting that the embryonic arrest resulting from DPPA3 expression is truly a consequence of the global loss of DNA methylation. Taken together, these results demonstrate that mammalian DPPA3 can inhibit UHRF1 to drive passive demethylation in distant, non-mammalian contexts.

Discussion

In this study, we aimed to identify the mechanistic basis for the formation of genome-wide DNA hypomethylation unique to mammals. As the role of TET enzymes in active demethylation is well documented[106], we investigated their contribution to the hypomethylated state of naïve ESCs. Mutation of the catalytic core of TET enzymes caused—as expected—a genome-wide increase in DNA methylation but mostly at sites where TET proteins do not bind suggesting a rather indirect mechanism. Among the few genes depending on TET activity for expression in naïve ESCs and downregulated at the transition to EpiLCs was *Dppa3*. Demethylation at the *Dppa3* locus coincides with TET1 and TET2 binding and TDG-dependent removal of oxidized cytosine residues via base excision repair. DPPA3 in turn binds and displaces UHRF1 from chromatin and thereby prevents the recruitment of DNMT1 and the maintenance of DNA methylation in ESCs (see graphic summary in Fig. 8).

Despite long recognized as a marker of naïve ESCs resembling the inner cell mass[63,107], we provide, to our knowledge, the first evidence that DPPA3 directly promotes the genome-wide DNA hypomethylation characteristic of mammalian naïve pluripotency. This unique pathway, in which TET proteins indirectly cause passive demethylation, is based upon two uniquely mammalian innovations: the expression of TET genes in pluripotent cell types[53,79,108] and the evolution of the novel *Dppa3* gene, which is positioned within a conserved pluripotency gene cluster[109] and dependent on TET activity for expression. In support of this novel pathway for passive demethylation, we found that TET mutant ESCs show a similar phenotype as Dppa3KO cells with respect to UHRF1 chromatin binding and hypermethylation and can be rescued by ectopic expression of *Dppa3*.

Our findings also provide the missing link to reconcile previous, apparently conflicting reports. To date, three distinct mechanisms have been proposed for the global hypomethylation accompanying naïve pluripotency: TET-mediated active demethylation[51,54,58], impaired maintenance DNA methylation[58], and PRDM14-dependent suppression of methylation[50,51,71]. As a downstream target of both TETs and PRDM14 as well as a direct inhibitor of maintenance DNA methylation, DPPA3 mechanistically connects and integrates these three proposed pathways of demethylation (see graphic summary in Fig. 8).

Our mechanistic data showing DPPA3 to displace UHRF1 and DNMT1 from chromatin provide a conclusive explanation for the previous observation that global hypomethylation in naïve ESCs was accompanied by reduced levels of UHRF1 at replication foci[58]. The hypomethylated state of naïve ESCs has also been reported to be dependent on PRDM14[50,71], which has been suggested to promote demethylation by repressing de novo DNA methyltransferases[50,54,71,73]. However, recent studies have demonstrated that the loss of de novo methylation only marginally affects DNA methylation levels in mouse and human ESCs[58,110]. Interestingly, while the loss of *Prdm14* leads to global hypermethylation, it also causes downregulation of *Dppa3*[71,73,111]. Our results suggest that the reported ability of PRDM14 to promote hypomethylation in naïve ESCs largely relies on its activation of the *Dppa3* gene ultimately leading to an inhibition of maintenance methylation.

Of note, other epigenetic pathways such as suppression of H3K9me2 by MAD2L2 as well as eRNA dependent enhancer regulation also have been shown to positively regulate the transcription of *Dppa3*[109,112], and silencing of *Dppa3* has been shown to depend on Lin28a, TBX3, and intact DNA methylation maintenance[113–115]. Taken together, these findings suggest that *Dppa3* is regulated by a complex network of pathways to ensure proper timing of its expression in order to prevent unwanted global DNA demethylation.

The comparison of TET catalytic mutants and Dppa3KO ESCs allows us to distinguish TET-dependent passive DNA demethylation mediated by DPPA3 from *bona fide* active demethylation. We show that TET activity is indispensable for the active demethylation of a subset of promoters in naïve ESCs, especially those of developmental genes. These findings uncover two evolutionary and mechanistically distinct functions of TET catalytic activity.

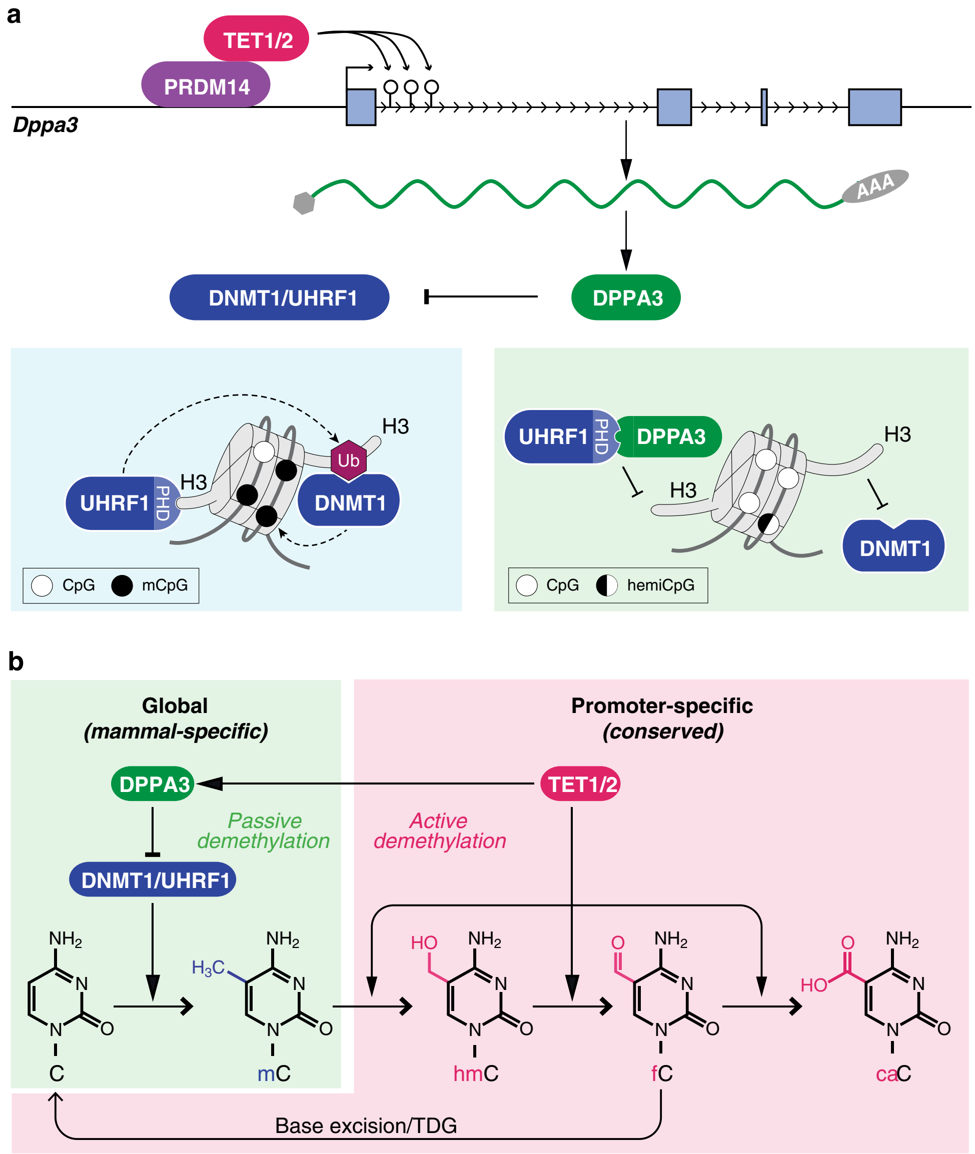


Fig. 8 Recent evolution of a TET-controlled and DPPA3-mediated pathway of DNA demethylation in boreoeutherian mammals. a In mammals, TET1 and TET2 are recruited by PRDM14 to the promoter of *Dppa3* where they promote active DNA demethylation and transcription of *Dppa3*. In most cellular contexts, high fidelity maintenance DNA methylation is guaranteed by the concerted activities of UHRF1 and DNMT1 at newly replicated DNA. Both the recruitment and activation of DNMT1 critically depend on the binding and ubiquitination of H3 tails by UHRF1. In naïve pluripotent cells, DPPA3 is expressed and inhibits maintenance DNA methylation by directly binding UHRF1 via its PHD domain and releasing it from chromatin. **b** TET1 and TET2 control DNA methylation levels by two evolutionary and mechanistically distinct pathways. TET-mediated active demethylation regulates focal DNA methylation states e.g. developmental genes and is evolutionarily conserved among vertebrates. The use of TET proteins to promote global demethylation appears to be specific to mammalian pluripotency and mediated by the recently evolved *Dppa3*.

Whereas TET-mediated active demethylation of developmental genes is evolutionarily conserved among vertebrates[79,116–118], the use of TET proteins to promote global demethylation appears to be specific to mammalian pluripotency[51,54,58] and mediated by the recently evolved *Dppa3* (Figs. 2c, 8).

In contrast to our findings in TET catalytic mutant ESCs, TET knockout ESCs do not appear to exhibit global hypermethylation[58]. This discrepancy might be explained by recent findings demonstrating that TET proteins influence global DNA methylation not only via their catalytic activity but also by their genomic binding[119,120]. Knockout of TET proteins results in a seemingly paradoxical loss of DNA methylation at repetitive elements like LINEs and LTRs due to a global redistribution of DNMT3A from heterochromatin to euchromatic sites previously occupied by TETs. In contrast to TET KOs, disruption of TET catalytic activity would not be expected to affect global TET occupancy, presumably leaving DNMT3A genomic occupancy intact. Thus, the extensive hypermethylation occurring upon TET inactivation, but not TET knockout, could be attributable to both the preservation of TET binding as well as the enhanced loading of the DNA methylation machinery on chromatin in TET CM ESCs.

To date, our understanding of DPPA3's function in the regulation of DNA methylation has been clouded by seemingly conflicting reports from different developmental stages and cell types. DPPA3's ability to modulate DNA methylation was first described in the context of zygotes[61], where it was demonstrated to specifically protect the maternal genome from TET3-dependent demethylation[29,74,77]. In contrast, DPPA3 was later shown to prevent aberrant DNA hypermethylation during PGC specification[59], iPSC reprogramming[75], and oocyte maturation[62,121]. Whereas DPPA3 was shown to disrupt UHRF1 function by sequestering it to the cytoplasm in oocytes[62], we demonstrate that DPPA3-mediated nucleocytoplasmic translocation of UHRF1 is not only dispensable but actually attenuates DPPA3's promotion of hypomethylation in ESCs. Another example of development- and context-specific function of DPPA3 is its role in the regulation of imprinting. While DPPA3 has no impact on ICR methylation in oocytes[61,62], it is required to prevent the loss of both paternal and maternal imprints in zygotes[77]. In naïve ESCs, we found that the Dppa3 KO results in a gain of DNA methylation at ICRs. Although contrary to its zygotic role in protecting imprints from demethylation, our data is consistent with previous findings examining the effect of Dppa3 loss on iPSC generation, where imprints also became hypermethylated[75].

In light of our data from naïve ESCs, *Xenopus*, and medaka, DPPA3's capacity to directly bind UHRF1's PHD domain and thereby inhibit UHRF1 chromatin binding appears to be its most basal function. Considering that DPPA3 localization is highly dynamic during the different developmental time periods at which it is expressed[59,78,122], it stands to reason that its role in modulating DNA methylation might also be dynamically modulated by yet-to-be determined regulatory mechanisms. For example, immediately following fertilization, full-length DPPA3 is cleaved and its C-terminal domain is specifically degraded[78]. Interestingly, we identified this exact C-terminal stretch of DPPA3 to be necessary and sufficient for DPPA3's inhibition of maintenance DNA methylation. Thus, the precisely timed destruction of this crucial domain might offer an explanation for the differing roles of DPPA3 in regulating DNA methylation between oocytes and zygotes[62,74,77,121].

As the most basic and evolutionarily conserved function of DNA methylation is the repression of TEs[6], the post-fertilization wave of DNA demethylation found in mammals raises several fundamental questions. Considering the mutational risks associated with TE activity, why have mammals come to dispense with such a central genomic defense mechanism during early development? Whereas derepression of TEs leads to genomic instability and ultimately cell death in most cell types[9,10,13,14], TE activity is not only tolerated but increasingly appreciated to fulfill key roles in early mammalian development[123–129]. The activation of TEs, in particular endogenous retroviruses (ERVs), appears to be a conserved feature of early mammalian embryos[130], beginning after fertilization and continuing for the duration of gestation in the cells of the trophoblast and the placenta[131,132]. During mammalian evolution, the placenta emerged in the common ancestor of therian mammals, after the divergence from the egg-laying monotremes[133,134]. Accumulating evidence suggests ERVs facilitated the complex, network level changes necessary for the evolutionary emergence and diversification of placental viviparity[135–137]. By enabling embryos to directly regulate the allocation of maternal resources, placental viviparity creates unique evolutionary challenges absent in egg-laying species[138]. At the fetal–maternal interface, the interests of the mother and her offspring as well as those of the paternal and maternal genomes within the embryo are brought into conflict, unleashing a coevolutionary arms race for control of maternal resources and provisioning[139]. The existence of such an evolutionary struggle is perhaps best exemplified by the emergence of genomic imprinting, or parent-of-origin-specific gene expression, in therian mammals[140]. Transposons, particularly ERVs, have played an important role in the evolution of genomic imprinting as an adaption to parental conflict; many of the cis-elements controlling imprinting status and, in some cases, even the imprinted genes themselves are derived from ERV insertions[141–143]. The retroviral origins of genomic imprinting are further illustrated by the use of conserved vertebrate host defense systems, namely DNA methylation and KRAB-ZFPs, to maintain imprint status[144,145]. In agreement with the parental conflict hypothesis, the evolution of more elaborated and invasive placentation has been accompanied by the expansion of genomic imprinting, with only 6 genes imprinted in marsupials compared with >100 in eutherians[146]. Indeed, the progressive co-option of retrotransposons over evolutionary time appears to have been a key driver in the transformation of a marsupial-like reproductive mode to the invasive and extended pregnancy of eutherians by facilitating the emergence of many of the unique, defining features of eutherian development such as the early allocation of the trophoblast cell lineage, invasive placentation, and suppression of the maternal immune response provoked by implantation[124,147–150]. Despite the importance of ERVs in eutherian development, the majority of ERV-derived regulatory elements, genes, and cis-elements controlling genomic imprinting are the result of evolutionarily recent and largely species-specific insertions[123,125,128,151–153].

How did eutherians come to rely on ERVs for so many aspects of their unique development? Such prolific ERV co-option among eutherians is proposed to have been a consequence of the evolution of precocious zygotic genome activation (ZGA) and an epigenetically permissive environment during early embryonic development[154,155]. It is tempting to speculate that post-fertilization demethylation was an important event in Eutherian evolution that contributed to the emergence and expansion of ERV/TE-based developmental regulation, including genomic imprints. Once ERV-derived genes and, in particular, regulatory networks acquired essential roles, mammalian preimplantation and placental development would have become "addicted" to the active transcription of ERVs[156]. Likewise, proper host development would require the establishment and maintenance of epigenetic states permissive for global ERV activity. In both mice and humans, the onset of ERV-dependent regulation coincides with a wave of genome-wide DNA demethylation, which commences upon fertilization and reaches its nadir in the ICM and trophectoderm of the blastocyst[19,40,157]. Whereas ERVs are silenced in the cells of the embryo proper by the wave of global de novo DNA methylation accompanying implantation, ERV activity and DNA hypomethylation persist in the trophoblast lineage throughout development[123,126,157–161]. Indeed, hypomethylation of the placenta relative to somatic cells appears to be conserved throughout Eutheria, despite dramatic differences in the embryonic and placental development among taxa[162].

As genome-wide DNA methylation is static throughout the lifecycle of most vertebrates, the evolution of novel mechanisms would have been required for the emergence of global DNA methylation erasure in the early embryonic development of eutherian mammals. DPPA3 may have arisen as a means to facilitate the early embryonic exposure of ERVs by neutralizing the host defense system of an ancestral eutherian mammal. In line with this notion, mouse embryos lacking *Dppa3* exhibit extensive genome-wide hypermethylation and undergo developmental arrest before the blastocyst stage as a result of impaired ERV activation and ZGA failure[62,76]. As *Dppa3* orthologs exhibit similar patterns of early embryonic expression in mice, humans, marmosets, cows, sheep, and pigs[163–167], it is plausible that function of DPPA3 during development is broadly conserved among mammals. However, our analysis identified *Dppa3* orthologs to be present in only a single clade of placental mammals, namely Boreoeutheria (Fig. 2c).

This raises the question whether eutherian lineages that lack DPPA3 also erase their methylomes and if so how? Pre-implantation DNA demethylation has been documented in every boreoeutherian species tested to date (e.g. mice, humans, monkeys, pigs, cows, sheep, rabbits)[19,36–43], however early embryonic DNA methylation dynamics have not been investigated in Eutherian lineages other than Boreoeutheria, i.e Afrotheria and Xenarthra, not to mention the more distant marsupial and monotreme groups. Likewise, the functional importance of ERV activity in early developmental and placental gene expression programs has also only been demonstrated in boreoutherian species. Thus, it is currently wholly unclear whether global DNA demethylation and ERV-dependent regulatory networks are even present, let alone important for early embryonic and trophoblast development outside of Boreoeutheria. Follow-up studies that investigate the origins of *Dppa3* and whether a similar ERV-based rewiring of early development may have occurred in other, not yet studied branches of vertebrates, are needed to understand how global DNA demethylation shaped the evolution of placental mammals.

Methods

Cell culture. Naïve J1 mouse ESCs were cultured and differentiated into EpiLCs using an established protocol[168,169]. In brief, for both naïve ESCs and EpiLCs defined media was used, consisting of N2B27: 50% neurobasal medium (Life Technologies), 50% DMEM/F12 (Life Technologies), 2 mM L-glutamine (Life Technologies), 0.1 mM β-mercaptoethanol (Life Technologies), N2 supplement (Life Technologies), B27 serum-free supplement (Life Technologies), 100 U/mL penicillin, and 100 µg/mL streptomycin (Sigma). Naïve ESCs were maintained on flasks treated with 0.2% gelatin in defined media containing 2i (1 µM PD032591 and 3 µM CHIR99021 (Axon Medchem, Netherlands)), 1000 U/mL recombinant leukemia inhibitory factor (LIF, Millipore), and 0.3% BSA (Gibco) for at least three passages before commencing differentiation. To differentiate naïve ESCs into Epiblast-like cells (EpiLCs), flasks were first pre-treated with Geltrex (Life Technologies) diluted 1:100 in DMEM/F12 (Life Technologies) and incubated at 37 °C overnight. Naïve ESCs were plated on Geltrex-treated flasks in defined medium containing 10 ng/mL Fgf2 (R&D Systems), 20 ng/mL Activin A (R&D Systems) and 0.1× Knockout Serum Replacement (KSR) (Life Technologies). Media was changed after 24 h and EpiLCs were harvested for RRBS and RNA-seq experiments after 48 h of differentiation.

For CRISPR-assisted cell line generation, mouse ESCs were maintained on 0.2% gelatin-coated dishes in Dulbecco's modified Eagle's medium (Sigma) supplemented with 16% fetal bovine serum (FBS, Sigma), 0.1 mM ß-mercaptoethanol (Invitrogen), 2 mM L-glutamine (Sigma), 1× MEM Non-essential amino acids (Sigma), 100 U/mL penicillin, 100 µg/mL streptomycin (Sigma), homemade recombinant LIF tested for efficient self-renewal maintenance, and 2i (1 µM PD032591 and 3 µM CHIR99021 (Axon Medchem, Netherlands)).

Human ESCs (line H9) were maintained in mTeSR1 medium (05850, STEMCELL Technologies) on Matrigel-coated plates (356234, Corning) prepared by 1:100 dilution, and 5 ml coating of 10 cm plates for 1 h at 37 °C. Colonies were passaged using the gentle cell dissociation reagent (07174, StemCell Technologies).

All cell lines were regularly tested for Mycoplasma contamination by PCR.

Sleeping beauty constructs. To generate the sleeping beauty donor vector with an N-terminal 3xFLAG tag and a fluorescent readout of doxycycline induction, we first used primers with overhangs harboring SfiI sites to amplify the IRES-DsRed-Express from pIRES2-DsRed-Express (Clontech)(Supplementary Data 5). This fragment was then cloned into the NruI site in pUC57-GentR via cut-ligation to generate an intermediate cloning vector pUC57-SfiI-IRES-DsRed-Express-SfiI. A synthesized gBlock (IDT, Coralville, IA, USA) containing Kozak-BIO-3XFLAG-AsiSI-NotI-V5 was cloned into the Eco47III site of the intermediate cloning vector via cut-ligation. The luciferase insert from pSBtet-Pur[170] (Addgene plasmid #60507) was excised using SfiI. The SfiI-flanked Kozak-BIO-3XFLAG-AsiSI-NotI-V5-IRES-DsRed-Express cassette was digested out of the intermediate cloning vector using SfiI and ligated into the pSBtet-Pur vector backbone linearized by SfiI. The end result was the parental vector, pSBtet-3xFLAG-IRES-DsRed-Express-PuroR. The pSBtet-3x-FLAG-mScarlet-PuroR vector was constructed by inserting a synthesized gBlock (IDT, Coralville, IA, USA) containing the SfiI-BIO-3XFLAG-AsiSI-NotI-mScarlet sequence into the SfiI-linearized pSBtet-Pur vector backbone using Gibson assembly[171]. For *Dppa3* expression constructs, the coding sequence of wild-type and mutant forms of *Dppa3* were synthesized as gBlocks (IDT, Coralville, IA, USA) and inserted into the pSBtet-3xFLAG-IRES-DsRed-Express-PuroR vector (linearized by AsiSI and NotI) using Gibson assembly. To produce the *Dppa3*-mScarlet fusion expression constructs, wild-type and mutant forms of *Dppa3* were amplified from pSBtet-3xFLAG-Dppa3-IRES-DsRed-Express-PuroR constructs using primers with overhangs homologous to the AsiSI and NotI restriction sites of the pSBtet-3x-FLAG-mScarlet-PuroR vector (Supplementary Data 5). Wild-type and mutant *Dppa3* amplicons were subcloned into the pSBtet-3x-FLAG-mScarlet-PuroR vector (linearized with AsiSI and NotI) using Gibson assembly.

For experiments involving the SBtet-3xFLAG-*Dppa3* cassette, all inductions were performed using 1 µg/mL doxycycline (Sigma-Aldrich). The DPPA3-WT construct was able to rescue the cytoplasmic localization and chromatin association of UHRF1 indicating that C-terminally tagged DPPA3 remains functional (Fig. 5b–d).

CRISPR/Cas9 genome engineering. For the generation of *Tet1*, *Tet2*, and *Tet1*/*Tet2* catalytic mutants, specific gRNAs targeting the catalytic center of *Tet1* and *Tet2* (Supplementary Data 5) were cloned into a modified version of the SpCas9-T2A-GFP/gRNA plasmid (px458[172], Addgene plasmid #48138), where we fused a truncated form of human Geminin (hGem) to SpCas9 in order to increase homology-directed repair efficiency[173] generating SpCas9-hGem-T2A-GFP/gRNA.

To generate *Tet1* and *Tet2* catalytic mutant targeting donors, 200 bp single-stranded DNA oligonucleotides carrying the desired HxD mutations (*Tet1*: H1652Y and D1654A, *Tet2*: H1304Y and D1306A) and ~100 bp homology arms were synthesized (IDT, Coralville, IA, USA) (Supplementary Data 5). For targetings in wild-type J1 ESCs, cells were transfected with a 4:1 ratio of donor oligo and SpCas9-hGem-T2A-GFP/gRNA construct. Positively transfected cells were isolated based on GFP expression using fluorescence-activated cell sorting (FACS) and plated at clonal density in ESC media 2 days after transfection. After 5–6 days, single colonies were picked and plated on 96-well plates. These plates were then duplicated 2 days later and individual clones were screened for the desired mutation by PCR followed by restriction fragment length polymorphism (RFLP) analysis. Cell lysis in 96-well plates, PCR on lysates, and restriction digests were performed as previously described[169]. The presence of the desired *Tet1* and/or *Tet2* catalytic mutations in putative clones was confirmed by Sanger sequencing.

As C-terminally tagged GFP labeled UHRF1 transgenes were shown to be able to rescue U1KO[83], the tagging of endogenous *Uhrf1* was also performed at the C-terminus. For insertion of the HALO or eGFP coding sequence into the endogenous *Dppa3* and *Uhrf1* loci, respectively, *Dppa3* and *Uhrf1* specific gRNAs were cloned into SpCas9-hGem-T2A-Puromycin/gRNA vector, which is a modified version of SpCas9-T2A-Puromycin/gRNA vector (px459;[172], Addgene plasmid #62988) similar to that described above. To construct the homology donors plasmids, gBlocks (IDT, Coralville, IA, USA) were synthesized containing either the HALO or eGFP coding sequence flanked by homology arms with ~200-400 bp homology upstream and downstream of the gRNA target sequence at the *Dppa3* or *Uhrf1* locus, respectively, and then cloned into the NruI site of pUC57-GentR via cut-ligation. ESCs were transfected with equimolar amounts of gRNA and homology donor vectors. Two days after transfection, cells were plated at clonal density and subjected to a transient puromycin selection (1 µg/mL) for 40 h. After 5-6 days, ESCs positive for HALO or eGFP integration were isolated via fluorescence-activated cell sorting (FACS) and plated again at clonal density in ESC media. After 4–5 days, colonies were picked and plated on Optical bottom µClear 96-well plates and re-screened for the correct expression and localization of eGFP or HALO using live-cell spinning-disk confocal imaging. Clones were subsequently genotyped using the aforementioned cell lysis strategy and further validated by Sanger sequencing[169].

To generate *Dppa3* knockout cells, the targeting strategy entailed the use of two gRNAs with target sites flanking the *Dppa3* locus to excise the entire locus on both alleles. gRNA oligos were cloned into the SpCas9-T2A-PuroR/gRNA vector (px459) via cut-ligation (Supplementary Data 5). ESCs were transfected with an equimolar amount of each gRNA vector. Two days after transfection, cells were plated at clonal density and subjected to a transient puromycin selection (1 µg/mL) for 40 h. Colonies were picked 6 days after transfection. The triple PCR strategy used for screening is depicted in Supplementary Fig. 3a. Briefly, PCR primers 1F and 4R were used to identify clones in which the *Dppa3* locus had been removed, resulting in the appearance of a ~350 bp amplicon. To identify whether the *Dppa3* locus had been removed from both alleles, PCRs were performed with primers 1F and 2R or 3F and 4R (Supplementary Data 5) to amplify upstream or downstream ends of the *Dppa3* locus, which would only be left intact in the event of mono-allelic locus excision. Removal of the *Dppa3* locus was confirmed with Sanger sequencing and loss of *Dppa3* expression was assessed by qRT-PCR.

For CRISPR/Cas gene editing, all transfections were performed using Lipofectamine 3000 (Thermo Fisher Scientific) according to the manufacturer's instructions. All DNA oligos used for gene editing and screening are listed in Supplementary Data 5.

Bxb1-mediated recombination and Sleeping Beauty transposition. To generate stable mESC lines carrying doxycycline-inducible forms of *Dppa3* or *Dppa3-mScarlet*, mES cells were first transfected with equimolar amounts of the pSBtet-3xFLAG-Dppa3-IRES-DsRed-PuroR or pSBtet-3xFLAG-Dppa3-mScarlet-PuroR and the Sleeping Beauty transposase, pCMV(CAT)T7-SB100[174] (Addgene plasmid #34879) vector using Lipofectamine 3000 (Thermo Fisher Scientific) according to manufacturer's instructions. Two days after transfection, cells were plated at clonal density and subjected to puromycin selection (1 µg/mL) for 5–6 days. To ensure comparable levels of *Dppa3* induction, cells were first treated for 18 h with doxycycline (1 µg/mL) and then sorted with FACS based on thresholded levels of DsRed or mScarlet expression, the fluorescent readouts of successful induction. Post sorting, cells were plated back into media without doxycycline for 7 days before commencing experiments.

To generate stable doxycycline-inducible *Dppa3* hESC lines, hES cells were first transfected with equimolar amounts of the pSBtet-3xFLAG-Dppa3-IRES-DsRed-PuroR and Sleeping Beauty transposase pCMV(CAT)T7-SB100[175] (Addgene plasmid #34879) vector using using the P3 Primary Cell 4D-NucleofectorTM Kit (V4XP-3012 Lonza) and the 4D-Nucleofector™ Platform (Lonza), program CB-156. Two days after nucleofection, cells were subjected to puromycin selection (1 µg/mL) for subsequent two days, followed by an outgrowth phase of 4 days. At this stage, cells were sorted with FACS based on thresholded levels of DsRed expression to obtain two bulk populations of positive stable hESC lines with inducible *Dppa3*.

For the generation of the *Uhrf1*$^{GFP/GFP}$ cell line, we used our previously described ESC line with a C-terminal MIN-tag (*Uhrf1*$^{attP/attP}$; Bxb1 *attP* site) and inserted the GFP coding sequence as described previously[169]. Briefly, attB-GFP-Stop-PolyA (Addgene plasmid #65526) was inserted into the C-terminal of the endogenous *Uhrf1*$^{attP/attP}$ locus by transfection with equimolar amounts of Bxb1 and attB-GFP-Stop-PolyA construct, followed by collection of GFP-positive cells with FACS after 6 days.

Cellular fractionation. Cell fractionation was performed as described previously with minor modifications[175]. Approximately 1×10^7 ESCs were resuspended in 250 µL of buffer A (10 mM HEPES pH 7.9, 10 mM KCl, 1.5 mM $MgCl_2$, 0.34 M sucrose, 10% glycerol, 0.1% Triton X-100, 1 mM DTT, 1 mM phenylmethylsulfonyl fluoride (PMSF), 1x mammalian protease inhibitor cocktail (PI; Roche)) and incubated for 5 min on ice. Nuclei were collected by centrifugation (4 min, 1300 × *g*, 4 °C) and the cytoplasmic fraction (supernatant) was cleared again by centrifugation (15 min, 20,000 × *g*, 4 °C). Nuclei were washed once with buffer A, and then lysed in buffer B (3 mM EDTA, 0.2 mM EGTA, 1 mM DTT, 1 mM PMSF, 1× PI). Insoluble chromatin was collected by centrifugation (4 min, 1700 × *g*, 4 °C) and washed once with buffer B. Chromatin fraction was lysed with 1× Laemmli buffer and boiled (10 min, 95 °C).

Western blot. Western blots were performed as described previously[82,169]. The following antibodies were used:

Rabbit anti-UHRF1 (polyclonal; 1:250; Cell Signalling, D6G8E), mouse anti-alpha-Tubulin (monoclonal; 1:500; Sigma, T9026), rabbit anti-H3 (polyclonal; 1:1000; Abcam, ab1791), mouse anti-GFP (monoclonal; 1:1000; Roche), mouse anti-FLAG M2 (monoclonal; 1:1000; Sigma, F3165), rabbit anti-xDNMT1 (polyclonal;[82]), rabbit anti-xUHRF1 (polyclonal;[82]), rabbit anti-USP7 (polyclonal; Bethyl Lab., A300-033A), rabbit anti-H3 (polyclonal; Abcam, ab1791), rat anti-TET1 (monoclonal; 1:10;[176]), rat anti-alpha-Tubulin (monoclonal; 1:250; Abcam, ab6160). goat anti-rat HRP (polyclonal; 1:1000; Jackson ImmunoResearch), goat anti-rabbit HRP (polyclonal; 1:1000; BioRad), mouse anti-xCDC45 (monoclonal;[177]), mouse anti-xRPA2 (monoclonal;[178]), and mouse anti-PCNA (monoclonal; Santa Cruz, sc56). Uncropped and unprocessed scans of blots can be found in the Source Data file.

Quantitative real-time PCR (qRT-PCR) analysis. Total RNA was isolated using the NucleoSpin Triprep Kit (Macherey-Nagel) according to the manufacturer's instructions. cDNA synthesis was performed with the High-Capacity cDNA Reverse Transcription Kit (with RNase Inhibitor; Applied Biosystems) using 500 ng of total RNA as input. qRT-PCR assays with oligonucleotides listed in Supplementary Data 5 were performed in 8 µL reactions with 1.5 ng of cDNA used as input. FastStart Universal SYBR Green Master Mix (Roche) was used for SYBR green detection. The reactions were run on a LightCycler480 (Roche).

LC-MS/MS analysis of DNA samples. Isolation of genomic DNA was performed according to earlier published work[57]. 1.0–5 µg of genomic DNA in 35 µL H_2O were digested as follows: An aqueous solution (7.5 µL) of 480 µM $ZnSO_4$, containing 18.4 U nuclease S1 (Aspergillus oryzae, Sigma-Aldrich), 5 U Antarctic phosphatase (New England BioLabs) and labeled internal standards were added ([$^{15}N_2$]-cadC 0.04301 pmol, [$^{15}N_2$,D_2]-hmdC 7.7 pmol, [D_3]-mdC 51.0 pmol, [$^{15}N_5$]-8-oxo-dG 0.109 pmol, [$^{15}N_2$]-fdC 0.04557 pmol) and the mixture was incubated at 37 °C for 3 h. After addition of 7.5 µl of a 520 µM $[Na]_2$-EDTA solution, containing 0.2 U snake venom phosphodiesterase I (Crotalus adamanteus, USB corporation), the sample was incubated for 3 h at 37 °C and then stored at −20 °C. Prior to LC/MS/MS analysis, samples were filtered by using an AcroPrep Advance 96 filter plate 0.2 µm Supor (Pall Life Sciences).

Quantitative UHPLC-MS/MS analysis of digested DNA samples was performed using an Agilent 1290 UHPLC system equipped with a UV detector and an Agilent 6490 triple quadrupole mass spectrometer. Natural nucleosides were quantified with the stable isotope dilution technique. An improved method, based on earlier published work[57,179] was developed, which allowed the concurrent analysis of all nucleosides in one single analytical run. The source-dependent parameters were as follows: gas temperature 80 °C, gas flow 15 L/min (N_2), nebulizer 30 psi, sheath gas heater 275 °C, sheath gas flow 15 L/min (N_2), capillary voltage 2,500 V in the positive ion mode, capillary voltage −2,250 V in the negative ion mode and nozzle voltage 500 V. The fragmentor voltage was 380 V/ 250 V. Delta EMV was set to 500 V for the positive mode. Chromatography was performed by a Poroshell 120 SB-C8 column (Agilent, 2.7 µm, 2.1 mm × 150 mm) at 35 °C using a gradient of water and MeCN, each containing 0.0085% (v/v) formic acid, at a flow rate of 0.35 mL/min: 0 → 4 min; 0 → 3.5% (v/v) MeCN; 4 → 6.9 min; 3.5 → 5% MeCN; 6.9 → 7.2 min; 5 → 80% MeCN; 7.2 → 10.5 min; 80% MeCN; 10.5 → 11.3 min; 80 → 0% MeCN; 11.3 → 14 min; 0% MeCN. The effluent up to 1.5 min and after 9 min was diverted to waste by a Valco valve. The autosampler was cooled to 4 °C. The injection volume amounted to 39 µL. Data were processed according to earlier published work[57].

RNA-seq library preparation. Digital gene expression libraries for RNA-seq were prepared using the single-cell RNA barcoding sequencing (SCRB-seq) method as described previously[180–182], with minor modifications to accommodate bulk cell populations. In brief, RNA was extracted and purified from ~1×10^6 cells using the NucleoSpin Triprep Kit (Machery-Nagel) according to the manufacturer's instructions. In the initial cDNA synthesis step, purified, bulk RNA (70 ng) from individual samples were subjected to reverse transcription in 10 µL reactions containing 25 units of Maxima H Minus reverse transcriptase (ThemoFisher Scientific), 1× Maxima RT Buffer (ThemoFisher Scientific), 1 mM dNTPs (ThermoFisher Scientific), 1 µM oligo-dT primer with a sample-specific barcode (IDT), and 1 µM template-switching oligo (IDT). Reverse transcription reactions were incubated 90 min at 42 °C. Next, the barcoded cDNAs from individual samples were pooled together and then purified using the DNA Clean & Concentrator-5 Kit (Zymo Research) according to the manufacturer's instructions. Purified pooled cDNA was eluted in 18 µL DNase/RNase-Free Distilled Water (Thermo Fisher) and then, to remove residual primers, incubated with 1 µL Exonuclease I Buffer (NEB) and 1 µL Exonuclease I (NEB) (final reaction volume: 20 µL) at 37 °C for 30 min followed by heat-inactivation at 80 °C for 20 min. Full-length cDNA was then amplified via PCR using KAPA HiFi HotStart ReadyMix (KAPA Biosystems) and SINGV6 primer (IDT). The pre-amplification PCR was performed using the following conditions: 3 min at 98 °C for initial denaturation, 10 cycles of 15 s at 98 °C, 30 s at 65 °C, and 6 min at 68 °C, followed by 10 min at 72 °C for final elongation. After purification using CleanPCR SPRI beads (CleanNA), the pre-amplified cDNA pool concentration was quantified using the Quant-iT PicoGreen dsDNA Assay Kit (Thermo Fisher). A Bioanalzyer run using the High-sensitivity DNA Kit (Agilent Technologies) was then performed to confirm the concentration and assess the size distribution of the amplified cDNA pool (Agilent Technologies). Next, 0.8 ng of the pure, amplified cDNA pool was used as input for generating a Nextera XT DNA library (Illumina) following the Manufacturer's instructions with the exception that a custom P5 primer (P5NEXTPT5) (IDT) was used to preferentially enrich for 3′ cDNA ends in the final Nextera XT Indexing PCR[180–182]. After an initial purification step using a 1:1 ratio of CleanPCR SPRI beads (CleanNA), the amplified Nextera XT Library the 300–800 bp range of the library was size-selected using a 2% E-Gel Agarose EX Gels (Life Technologies) and then extracted from the gel using the MinElute Gel Extraction Kit (Qiagen, Cat. No. 28606) according to manufacturer's recommendations. The final concentration, size distribution, and quality of Nextera XT library were assessed with a Bioanalyzer (Agilent Technologies) using a High-sensitivity DNA Kit (Agilent Technologies). The Nextera XT RNA-seq library was paired-end sequenced using a high output flow cell on an Illumina HiSeq 1500.

Reduced representation bisulfite sequencing (RRBS) library preparation. For RRBS library preparation, genomic DNA was isolated using the QIAamp DNA Mini Kit (QIAGEN), after an overnight lysis and proteinase K treatment. RRBS library preparation was performed as described previously[183], with slight modifications. In brief, once purified, genomic DNA (100 ng) from each sample was used as starting material and first digested with 60 units of MspI (New England Biolabs) in a 30 µl reaction volume at 37 °C overnight. Digested DNA ends were then repaired and A-tailed by adding a 2 µl of a mixture containing 10 mM dATP, 1 mM dCTP, 1 mM dGTP and Klenow fragment (3′→5′ exo-) (New England Biolabs) to the unpurified digestion reaction and incubated first at 30 °C for 20 min followed by 37 °C for 20 min. Individual end-repaired and A-tailed DNA samples were purified using a 2:1 ratio of CleanPCR SPRI beads (CleanNA) and eluted in 20 µl elution buffer (10 mM Tris-HCl, pH 8.5). Next, barcoded adapters were ligated to the eluted DNA fragments in a 30 µl reaction containing 1× T4 Ligase Buffer (New England Biolabs), 2000 units of T4 Ligase (New England Biolabs), and 0.8 µM sample-specific TruSeq adapters (Illuminas) and incubated at 16 °C overnight. After adapter ligation, individual samples were first pooled before being purified with a 2:1 ratio of CleanPCR SPRI beads (CleanNA) and then eluted using 4 µl elution buffer times the number of samples in the pool. Pooled samples were then bisulfite converted using the EZ DNA Methylation-Gold™ Kit (Zymo Research) according to the manufacturer's instructions with the exception that libraries were eluted 2 × 20 µL M-elution buffer (Zymo Research). After bisulfite conversion, libraries were amplified in a 200 µl large-scale PCR reaction containing, 1x PfuTurbo Cx Reaction Buffer (Agilent Technologies), 10 units of PfuTurbo Cx Hotstart DNA Polymerase (Agilent Technologies), 1 mM dNTPs (New England Biolabs), 0.3 µM TruSeq Primers (Illumina), and 20 µl of pooled, bisulfite-converted DNA samples. After dividing the reaction into 4 wells of a 96-well plate (each containing 50 µl), the PCR was performed using the following cycling conditions: 2 min at 95 °C for initial denaturation and Polymerase activation, 16 cycles of 30 s at 95 °C, 30 s at 65 °C, and 45 s at 72 °C, followed by 7 min at 72 °C for final elongation. After amplification, the samples are pooled together again, subjected to a final round of purification using a 1.2:1 ratio of CleanPCR SPRI beads (CleanNA), and eluted in 40 µl of elution buffer. For an initial assessment of quality and yield, purified RRBS libraries were first analyzed on 2% E-Gel Agarose EX Gels (Life Technologies) and the concentrations then measured using the Quant-iT™ PicoGreen™ dsDNA Assay-Kit (ThermoFisher). The final concentration, size distribution, and quality of each RRBS library was then assessed with a Bioanalyzer (Agilent Technologies) using a High-sensitivity DNA Kit (Agilent Technologies). RRBS libraries were then sequenced on an Illumina HiSeq 1500.

Targeted bisulfite amplicon (TaBA) sequencing. Genomic DNA was isolated from 10^6 cells using the PureLink Genomic DNA Mini Kit (Thermo Fisher Scientific) according to the manufacturer's instructions. The EZ DNA Methylation-Gold Kit (Zymo Research) was used for bisulfite conversion according to the manufacturer's instructions but with the following alterations: 500 ng of genomic DNA was used as input and bisulfite converted DNA was eluted in 2 × 20 µL Elution Buffer (10 mM Tris-HCl, pH 8.5).

TaBA-seq library preparation entailed two sequential PCRs to first amplify a specific locus and then index sample-specific amplicons. For the first PCR, the locus specific primers were designed with Illumina TruSeq and Nextera compatible overhangs (Supplementary Data 5). The amplification of bisulfite converted DNA was performed in 25 µL PCR reaction volumes containing 0.4 µM each of forward and reverse primers, 2 mM Betaiinitialne (Sigma-Aldrich, B0300-1VL), 10 mM Tetramethylammonium chloride solution (Sigma-Aldrich T3411-500ML), 1x MyTaq Reaction Buffer, 0.5 units of MyTaq HS (Bioline, BIO-21112), and 1 µL of the eluted bisulfite converted DNA (~12.5 ng). The following cycling parameters were used: 5 min for 95 °C for initial denaturation and activation of the polymerase, 40 cycles (95 °C for 20 s, 58 °C for 30 s, 72 °C for 25 s) and a final elongation at 72 °C for 3 min. Agarose gel electrophoresis was used to determine the quality and yield of the PCR. For purifying amplicon DNA, PCR reactions were incubated with 1.8x volume of CleanPCR beads (CleanNA, CPCR-0005) for 10 min. Beads were immobilized on a DynaMag™-96 Side Magnet (Thermo Fisher, 12331D) for 5 min, the supernatant was removed, and the beads washed 2× with 150 µL 70% ethanol. After air drying the beads for 5 min, DNA was eluted in 15 µL of 10 mM Tris-HCl pH 8.0. Amplicon DNA concentration was determined using the Quant-iT™ PicoGreen™ dsDNA Assay Kit (Thermo Fisher, P7589) and then diluted to 0.7 ng/µL.

Thereafter, indexing PCRs were performed in 25 µL PCR reaction volumes containing 0.08 µM (1 µL of a 2 µM stock) each of i5 and i7 Indexing Primers (Supplementary Data 5), 1x MyTaq Reaction Buffer, 0.5 units of MyTaq HS (Bioline, BIO-21112), and 1 µL of the purified PCR product from the previous step. The following cycling parameters were used: 5 min for 95 °C for initial denaturation and activation of the polymerase, 40 cycles (95 °C for 10 s, 55 °C for 30 s, 72 °C for 40 s) and a final elongation at 72 °C for 5 min. Agarose gel electrophoresis was used to determine the quality and yield of the PCR. An aliquot from each indexing reaction (5 µL of each reaction) was then pooled and purified with CleanPCR magnetic beads as described above and eluted in 1 µL × Number of pooled reactions. Concentration of the final library was determined using PicoGreen and the quality and size distribution of the library was assessed with a Bioanalyzer. Dual indexed TaBA-seq libraries were sequenced on an Illumina MiSeq in 2 × 300 bp output mode.

RNA-seq processing and analysis. RNA-seq libraries were processed and mapped to the mouse genome (mm10) using the zUMIs pipeline[184]. UMI count tables were filtered for low counts using HTSFilter[185]. Differential expression analysis was performed in R using DESeq2[186] and genes with an adjusted $P < 0.05$ were considered to be differentially expressed. Hierarchical clustering was performed on genes differentially expressed in TET mutant ESCs respectively, using k-means clustering ($k = 4$) in combination with the ComplexHeatmap (v 1.17.1) R-package[187]. Principal component analysis was restricted to genes differentially expressed during wild-type differentiation and performed using all replicates of wild-type, TET mutant, and Dppa3KO ESCs.

RRBS alignment and analysis. Raw RRBS reads were first trimmed using Trim Galore (v.0.3.1) with the "-rrbs" parameter. Alignments were carried out to the mouse genome (mm10) using bsmap (v.2.90) using the parameters "-s 12 -v 10 -r 2 -I 1". Summary statistics of the RRBS results are provided in Supplementary Data 6 and sample reproducibility information is shown in Supplementary Fig. 9. CpG-methylation calls were extracted from the mapping output using bsmaps methratio.py. Analysis was restricted to CpG with a coverage >10. methylKit[188] was used to identify differentially methylated regions between the respective contrasts for the following genomic features: (1) all 1-kb tiles (containing a minimum of three CpGs) detected by RRBS; (2) Repeats (defined by Repbase); (3) gene promoters (defined as gene start sites −2 kb/+2 kb); and (4) gene bodies (defined as longest isoform per gene) and CpG islands (as defined by Ilingworth et al.[189]). Differentially methylated regions were identified as regions with $P < 0.05$ and a difference in methylation means between two groups greater than 20%. Principal component analysis of global DNA methylation profiles was performed on single CpGs using all replicates of wild-type, T1KO and T1CM ESCs and EpiLCs.

Chromatin immunoprecipitation (ChIP) and Hydroxymethylated-DNA immunoprecipitation (hMeDIP) alignment and analysis. ChIP-seq reads for TET1 binding in ESCs and EpiLCs were downloaded from GSE57700[67] and PRJEB19897[66], respectively. hMeDIP reads for wild-type ESCs and T1KO ESCs were download from PRJEB13096[66]. Reads were aligned to the mouse genome (mm10) with Bowtie (v.1.2.2) with parameters "-a -m 3 -n 3 -best -strata". Subsequent ChIP-seq analysis was carried out on data of merged replicates. Peak calling and signal pileup was performed using MACS2 callpeak[190] with the parameters "-extsize 150" for ChIP, "-extsize 220" for hMeDIP, and "-nomodel -B -nolambda" for all samples. Tag densities for promoters and 1 kb Tiles were calculated using the deepTools2 computeMatrix module[191]. TET1 bound genes were defined by harboring a TET1 peak in the promoter region (defined as gene start sites −2 kb/+2 kb).

Immunofluorescence staining. For immunostaining, naïve ESCs were grown on coverslips coated with Geltrex (Life Technologies) diluted 1:100 in DMEM/F12 (Life Technologies), thereby allowing better visualization of the cytoplasm during microscopic analysis. All steps during immunostaining were performed at room temperature. Coverslips were rinsed two times with PBS (pH 7.4; 140 mM NaCl, 2.7 mM KCl, 6.5 mM Na_2HPO_4, 1.5 mM KH_2PO_4) prewarmed to 37 °C, cells fixed for 10 min with 4% paraformaldehyde (pH 7.0; prepared from paraformaldehyde powder (Merck) by heating in PBS up to 60 °C; store at −20 °C), washed three times for 10 min with PBST (PBS, 0.01% Tween20), permeabilized for 5 min in PBS supplemented with 0.5% Triton X-100, and washed two times for 10 min with PBS. Primary and secondary antibodies were diluted in blocking solution (PBST, 4% BSA). Coverslips were incubated with primary and secondary antibody solutions in dark humid chambers for 1 h and washed three times for 10 min with PBST after primary and secondary antibodies. For DNA counterstaining, coverslips were incubated 6 min in PBST containing a final concentration of 2 µg/mL DAPI (Sigma-Aldrich) and washed three times for 10 min with PBST. Coverslips were mounted in antifade medium (Vectashield, Vector Laboratories) and sealed with colorless nail polish.

The following antibodies were used: rabbit anti-DPPA3 (polyclonal; 1:200; Abcam, ab19878), mouse anti-UHRF1 (monoclonal; 1:250; Santa Cruz, sc373750), goat anti-mouse A488 (polyclonal; 1:500; used in IF; Invitrogen, A11029), donkey anti-rabbit Dylight594 (polyclonal; 1:500; Dianova, 711-516-152), anti-GFP-Booster ATTO488 (1:200; Chromotek), mouse anti-5mC (monoclonal; 1:200; Active Motif, 39649), donkey anti-anti-rabbit A555 (polyclonal; 1:500; Invitrogen, A31572), and donkey anti-anti-rabbit A488 (polyclonal; 1:500; Dianova, 711-547-003).

Immunofluorescence and Live-cell imaging. For immunofluorescence, stacks of optical sections were collected on a Nikon TiE microscope equipped with a Yokogawa CSU-W1 spinning-disk confocal unit (50 µm pinhole size), an Andor Borealis illumination unit, Andor ALC600 laser beam combiner (405 nm/488 nm/561 nm/640 nm), Andor IXON 888 Ultra EMCCD camera, and a Nikon 100×/1.45 NA oil immersion objective. The microscope was controlled by software from Nikon (NIS Elements, ver. 5.02.00). DAPI or fluorophores were excited with 405 nm, 488 nm, or 561 nm laser lines and bright-field images acquired using Nikon differential interference contrast optics. Confocal image z-stacks were recorded with a step size of 200 nm, 16-bit image depth, 1 × 1 binning, a frame size of 1024 × 1024 pixels, and a pixel size of 130 nm. Within each experiment, cells were imaged using the same settings on the microscope (camera exposure time, laser power, and gain) to compare signal intensities between cell lines.

For live-cell imaging, cells were plated on Geltrex-coated glass bottom 2-well imaging slides (Ibidi). Both still and timelapse images were acquired on the Nikon spinning-disk system described above equipped with an environmental chamber maintained at 37 °C with 5% CO_2 (Oko Labs), using a Nikon 100x/1.45 NA oil immersion objective and a Perfect Focus System (Nikon). Images were acquired with the 488, 561, and 640 nm laser lines, full-frame (1024 × 1024) with 1 × 1 binning, and with a pixel size of 130 nm. Transfection of a RFP-PCNA vector[192] was used to identify cells in S-phase. For DNA staining in live cells, cells were exposed to media containing 200 nM SiR-DNA (Spirochrome) for at least 1 h before imaging. For imaging endogenous DPPA3-HALO in live cells, cells were treated with media containing 50 nM HaloTag-TMR fluorescent ligand (Promega) for 1 h. After incubation, cells were washed 3× with PBS before adding back normal media. Nuclear export inhibition was carried out using media containing 20 nM leptomycin-B (Sigma-Aldrich). Live-cell imaging data was acquired with NIS Elements ver. 4.5 (Nikon). NIS Elements ver. 5.02.00 (Nikon) and Volocity (PerkinElmer) were used for acquiring FRAP data. RICS measurements were acquired using FABSurf (v 1.0).

Image analysis. For immunofluorescence images, Fiji software (ImageJ 1.51j)[193,194] was used to analyze images and create RGB stacks. For analysis of live-cell imaging data, CellProfiler Software (version 3.0)[195] was used to quantify fluorescence intensity in cells stained with SiR-DNA. CellProfiler pipelines used in this study are available upon request. In brief, the SiR-DNA signal was used to segment ESC nuclei. Mean fluorescence intensity of GFP was measured both inside the segmented area (nucleus) and in the area extending 4–5 pixels beyond the segmented nucleus (cytoplasm). GFP fluorescence intensity was normalized by subtracting the experimentally-determined mean background intensity and background-subtracted GFP intensities were then used for all subsequent quantifications shown in Fig. 4 and Supplementary Figs. 4h, 5h, and 6b, c.

Fluorescence recovery after photobleaching (FRAP). For FRAP assays, cells cultivated on Geltrex-coated glass bottom 2-well imaging slides (Ibidi) were imaged in an environmental chamber maintained at 37 °C with 5% CO_2 either using the Nikon system mentioned above equipped with a FRAPPA photobleaching module (Andor) or on an Ultraview-Vox spinning-disk system (Perkin-Elmer) including a FRAP Photokinesis device mounted to an inverted Axio Observer D1 microscope (Zeiss) equipped with an EMCCD camera (Hamamatsu) and a 63x/1.4 NA oil immersion objective, as well as 405, 488 and 561 nm laser lines.

For endogenous UHRF1-GFP FRAP, eight pre-bleach images were acquired with the 488 nm laser, after which an area of 4 × 4 pixels was irradiated for a total of 16 ms with a focused 488 nm laser (leading to a bleached spot of ~1 µm) to bleach a fraction of GFP-tagged molecules within cells, and then recovery images were acquired every 250 ms for 1-2 min. Recovery analysis was performed in Fiji. Briefly, fluorescence intensity at the bleached spot was measured in background-subtracted images, then normalized to pre-bleach intensity of the bleached spot, and normalized again to the total nuclear intensity in order to account for acquisition photobleaching. Images of cells with visible drift were discarded.

***Xenopus* egg extracts**. The interphase extracts (low-speed supernatants (LSS)) were prepared as described previously[82]. After thawing, LSS were supplemented with an energy regeneration system (5 µg/ml creatine kinase, 20 mM creatine phosphate, 2 mM ATP) and incubated with sperm nuclei at 3000–4000 nuclei per µl. Extracts were diluted 5-fold with ice-cold CPB (50 mM KCl, 2.5 mM MgCl2, 20 mM HEPES-KOH, pH 7.7) containing 2% sucrose, 0.1% NP-40 and 2 mM NEM, overlaid onto a 30% sucrose/CPB cushion, and centrifuged at 15,000 *g* for 10 min. The chromatin pellet was resuspended in SDS sample buffer and analyzed by SDS-PAGE. GST-mDPPA3 was added to egg extracts at 50 ng/µl at final concentration.

Monitoring DNA methylation in *Xenopus* egg extracts. DNA methylation was monitored by the incorporation of S-[methyl-^{3}H]-adenosyl-L-methionine, incubated at room temperature, and the reaction was stopped by the addition of CPB containing 2% sucrose up to 300 µl. Genomic DNA was purified using a Wizard Genomic DNA purification kit (Promega) according to the manufacturer's instructions. Incorporation of radioactivity was quantified by liquid synchillation counter.

Plasmid construction for recombinant mDPPA3. To generate GST-tagged mDPPA3 expression plasmids, mDPPA3 fragment corresponding to full-length protein was amplified by PCR using mouse DPPA3 cDNA and specific primers (Supplementary Data 5). The resulting DNA fragment was cloned into pGEX4T-3 vector digested with EcoRI and SalI using an In-Fusion HD Cloning Kit.

Protein expression and purification. For protein expression in *Escherichia coli* (BL21-CodonPlus), the mDPPA3 genes were transferred to pGEX4T-3 vector as described above. Protein expression was induced by the addition of 0.1 mM Isopropyl β-D-1-thiogalactopyranoside (IPTG) to media followed by incubation for 12 h at 20 °C. For purification of Glutathione S transferase (GST) tagged proteins, cells were collected and resuspended in Lysis buffer (20 mM HEPES-KOH (pH 7.6), 0.5 M NaCl, 0.5 mM EDTA, 10% glycerol, 1 mM DTT) supplemented with 0.5% NP40 and protease inhibitors, and were then disrupted by sonication on ice. After centrifugation, the supernatant was applied to Glutathione Sepharose (GSH) beads (GE Healthcare) and rotated at 4 °C for 2 h. Beads were then washed three times with Wash buffer 1 (20 mM Tris-HCl (pH 8.0), 150 mM NaCl, 1% TritonX-100, 1 mM DTT) three times and with Wash buffer 2 (100 mM Tris-HCl (pH 7.5), 100 mM NaCl) once. Bound proteins were eluted in Elution buffer (100 mM Tris-HCl (pH 7.5), 100 mM NaCl, 5% glycerol, 1 mM DTT) containing 42 mM reduced Glutathione and purified protein was loaded on PD10 desalting column equilibrated with EB buffer (10 mM HEPES/KOH at pH 7.7, 100 mM KCl, 0.1 mM $CaCl_2$, 1 mM $MgCl_2$) containing 1 mM DTT, and then concentrated by Vivaspin (Millipore).

Data collection for the presence of TET1, UHRF1, DNMT1, and DPPA3 throughout metazoa. Reference protein sequences of TET1 (Human Q8NFU7, Mouse Q3URK3, *Naegleria gruberi* D2W6T1), DNMT1 (Rat Q9Z330, Human P26358, Mouse P13864, Chicken Q92072, Cow Q92072), UHRF1 (Mouse Q8VDF2, Rat Q7TPK1, Zebra fish E7EZF3, Human Q96T88, Cow A7E320, Xenopus laevis F6UA42) and DPPA3 (Mouse Q8QZY3, Human Q6W0C5, Cow A9Q1J7) were downloaded from the Universal Protein Resource (UniProt). Orthologous were identified with *hmmsearch* of the HMMER (http://hmmer.org/) toolkit using default parameters. Presence of the proteins throughout metazoa was visualized using iTOL[196].

Chromatin immunoprecipitation coupled to Mass Spectrometry and Proteomics data analysis. For Chromatin immunoprecipitation coupled to Mass Spectrometry (ChIP-MS), whole cell lysates of the doxycycline-inducible *Dppa3*-FLAG mES cells were used by performing three separate immunoprecipitations with an anti-FLAG antibody and three samples with a control IgG. Trypsinized cells were washed twice by PBS and subsequently diluted to 15*10^6 cells per 10 mL PBS. Paraformaldehyde (PFA) was added to a final concentration of 1% and crosslinking was performed at room temperature on an orbital shaker for 10 min. Free PFA was quenched by 125 mM Glycine for 5 min and crosslinked cells were washed twice by ice-cold PBS before cell lysis. Proteins were digested on the beads after the pulldown and desalted subsequently on StageTips with three layers of C18[197]. Here, peptides were separated by liquid chromatography on an Easy-nLC 1200 (Thermo Fisher Scientific) on in-house packed 50 cm columns of ReproSil-Pur C18-AQ 1.9-µm resin (Dr. Maisch GmbH). Peptides were then eluted successively in an ACN gradient for 120 min at a flow rate of around 300 nL/min and were injected through a nanoelectrospray source into a Q Exactive HF-X Hybrid Quadrupole-Orbitrap Mass Spectrometer (Thermo Fisher Scientific). After measuring triplicates of a certain condition, an additional washing step was scheduled. During the measurements, the column temperature was constantly kept at 60 °C while after each measurement, the column was washed with 95% buffer B and subsequently with buffer A. Real time monitoring of the operational parameters was established by SprayQc[198] software. Data acquisition was based on a top10 shotgun proteomics method and data-dependent MS/MS scans. Within a range of 400-1650 m/z and a max. injection time of 20 ms, the target value for the full scan MS spectra was 3×10^6 and the resolution at 60,000.

The raw MS data was then analyzed with the MaxQuant software package (version 1.6.0.7)[199]. The underlying FASTA files for peak list searches were derived from Uniprot (UP000000589_10090.fasta and UP000000589_10090 additional.fasta, version June 2015) and an additional modified FASTA file for the FLAG-tagged *Dppa3* in combination with a contaminants database provided by the Andromeda search engine[200] with 245 entries. During the MaxQuant-based analysis the "Match between runs" option was enabled and the false discovery rate was set to 1% for both peptides (minimum length of 7 amino acids) and proteins. Relative protein amounts were determined by the MaxLFQ algorithm[201], with a minimum ratio count of two peptides.

For the downstream analysis of the MaxQuant output, the software Perseus[202] (version 1.6.0.9) was used to perform two-sided Student's *t*-test with a permutation-based FDR of 0.05 and an additional constant $S0 = 1$ in order to calculate fold enrichments of proteins between triplicate chromatin immunoprecipitations of anti-FLAG antibody and control IgG. The result was visualized in a scatter plot. The complete catalog of proteins interacting with FLAG-DPPA3 in ESCs including statistics can be found in Supplementary Data 3.

For GO analysis of biological processes the Panther classification system was used[203]. For the analysis, 131 interactors of DPPA3 were considered after filtering the whole amount of 303 significant interactors for a *p*-value of at least 0.0015 and 3 or more identified peptides. The resulting GO groups (determined by a two-sided Fisher's exact test) were additionally filtered for a fold enrichment of observed over expected amounts of proteins of at least 4 and a *p*-value of 5.30 E−08. The result can be found in Supplementary Data 4.

***Dppa3* overexpression in medaka embryos and immunostaining**. Medaka d-rR strain was used. Medaka fish were maintained and raised according to standard protocols. Developmental stages were determined based on a previous study[204]. *Dppa3* and mutant *Dppa3* (R107E) mRNA were synthesized using HiScribe T7 ARCA mRNA kit (NEB, E2060S), and purified using RNeasy mini kit (QIAGEN, 74104). *Dppa3* or mutant *Dppa3* (R107E) mRNA was injected into the one-cell stage (stage 2) medaka embryos. After 7 h of incubation at 28 °C, the late blastula (stage 11) embryos were fixed with 4% PFA in PBS for 2 h at room temperature, and then at 4 °C overnight. Embryos were dechorionated, washed with PBS, and permeabilized with 0.5% Triton X-100 in PBS for 30 min at room temperature. DNA was denatured in 4 M HCl for 15 min at room temperature, followed by neutralization in 100 mM Tris-HCl (pH 8.0) for 20 min. After washing with PBS, embryos were blocked in blocking solution (2% BSA, 1%DMSO, 0.2% Triton X-100 in PBS) for 1 h at room temperature, and then incubated with 5-methylcytosine antibody (1:200; Active Motif #39649) at 4 °C overnight. The embryos were washed with PBSDT (1% DMSO, 0.1% Triton X-100 in PBS), blocked in blocking solution for 1 h at room temperature, and incubated with Alexa Fluor 555 goat anti-mouse 2nd antibody (1:500; ThermoFisher Scientific #A21422) at 4 °C overnight. After washing with PBSDT, cells were mounted on slides and examined under a fluorescence microscope.

Fluorescence three hybrid (F3H) assay. The F3H assay was performed as described previously[96]. In brief, BHK cells containing multiple lac operator repeats were transiently transfected with the respective GFP- and mScarlet-constructs on coverslips using PEI and fixed with 3.0% formaldehyde 24 h after transfection. For DNA counterstaining, coverslips were incubated in a solution of DAPI (200 ng/ml) in PBS-T and mounted in Vectashield. Images were collected using a Leica TCS SP5 confocal microscope. To quantify the interactions within the lac spot, the following intensity ratio was calculated for each cell in order to account for different expression levels: $\text{mScarlet}_{\text{spot}} - \text{mScarlet}_{\text{background}})/(\text{GFP}_{\text{spot}} - \text{GFP}_{\text{background}})$.

Microscale thermophoresis (MST). For MST measurements, mUHRF1 C-terminally tagged with GFP- and 6xHis-tag was expressed in HEK 293 T cells and then purified using Qiagen Ni-NTA beads (Qiagen #30230). Recombinant mDPPA3 WT and 1-60 were purified as described above. Purified UHRF1 (200 nM) was mixed with different concentrations of purified DPPA3 (0.15 nM to 5 µM) followed by a 30 min incubation on ice. The samples were then aspirated into NT.115 Standard Treated Capillaries (NanoTemper Technologies) and placed into the Monolith NT.115 instrument (NanoTemper Technologies). Experiments were conducted with 80% LED and 80% MST power. Obtained fluorescence signals were normalized (F_{norm}) and the change in F_{norm} was plotted as a function of the concentration of the titrated binding partner using the MO. Affinity Analysis software version 2.1 (NanoTemper Technologies). For fluorescence normalization ($F_{\text{norm}} = F_{\text{hot}}/F_{\text{cold}}$), the manual analysis mode was selected and cursors were set as follows: $F_{\text{cold}} = -1$ to 0 s, $F_{\text{hot}} = 10$ to 15 s. The Kd was obtained by fitting the mean F_{norm} of eight data points (four independent replicates, each measured as a technical duplicate).

RICS. Data for Raster Image Correlation Spectroscopy (RICS) was acquired on a home-built laser scanning confocal setup equipped with a 100x NA 1.49 NA objective (Nikon) pulsed interleaved excitation (PIE) as used elsewhere[205]. Samples were excited using pulsed lasers at 470 (Picoquant) and 561 nm (Toptica Photonics), synchronized to a master clock, and then delayed ~20 ns relative to one another to achieve PIE. Laser excitation was separated from descanned fluorescence emission by a Di01-R405/488/561/635 polychroic mirror (Semrock, AHF Analysentechnik) and eGFP and mScarlet fluorescence emission was separated by a 565 DCXR dichroic mirror (AHF Analysentechnik) and collected on avalanche photodiodes, a Count Blue (Laser Components) and a SPCM-AQR-14 (Perkin-Elmer) with 520/40 and a 630/75 emission filters (Chroma, AHF Analysentechnik). Detected photons were recorded by time-correlated single-photon counting.

The alignment of the system was verified prior to each measurement session by performing FCS with PIE on a mixture of Atto-488 and Atto565 dyes excited with pulsed 470 and 561 nm lasers set to 10 µW (measured in the collimated space before entering the galvo-scanning mirror system), 1 µm above the surface of the coverslip[206]. Cells were plated on Ibidi two-well glass bottom slides, and induced with doxycycline overnight prior to measurements. Scanning was performed in cells maintained at 37 °C using a stage top incubator, with a total field-of-view of 12 µm × 12 µm, composed of 300 pixels × 300 lines (corresponding to a pixel size of 40 nm), a pixel dwell time of 11 µs, a line time of 3.33 ms, at one frame per second, for 100–200 s. Pulsed 470 and 561 nm lasers were adjusted to 4 and 5 µW, respectively.

Image analysis was done using the Pulsed Interleaved Excitation Analysis with Matlab (PAM) software[207]. Briefly, time gating of the raw photon stream was performed by selecting only photons collected on the appropriate detector after the corresponding pulsed excitation, thereby allowing cross-talk free imaging for each channel. Then, using the Microtime Image Analysis (MIA) analysis program, slow fluctuations were removed by subtracting a moving average of 3 frames and a region of interest corresponding to the nucleus was selected, excluding nucleoli and dense aggregates. The spatial autocorrelation and cross-correlation functions (SACF and SCCF) were calculated as done previously[208] using arbitrary region RICS:

$$G(\xi,\psi) = \frac{\langle I_{\text{RICS},1}(x,y) I_{\text{RICS},2}(x+\xi,y+\psi)\rangle_{XY}}{\langle I_{\text{RICS},1}\rangle_{XY}\langle I_{\text{RICS},2}\rangle_{XY}} \tag{1}$$

where ξ and ψ are the correlation lags in pixel units along the *x*- and *y*-axis scan directions. The correlation function was then fitted to a two-component model (one mobile and one immobile component) in MIAfit:

$$G_{\text{fit}}(\xi,\psi) = A_{\text{mob}} G_{\text{fit, mob}}(\xi,\psi) + A_{\text{imm}} \exp\left(-\delta r^2 \omega_{\text{imm}}^{-2}(\xi^2+\psi^2)\right) + y_0, \tag{2}$$

where:

$$G_{\text{fit, mob}}(\xi,\psi) = \left(1 + \frac{4D(\tau_p\xi + \tau_l\psi)}{\omega_r^2}\right)^{-1}\left(1 + \frac{4D(\tau_p\xi + \tau_l\psi)}{\omega_z^2}\right)^{-1/2} \cdot \exp\left(-\frac{\delta r^2(\xi^2+\psi^2)}{\omega_r^2 + 4D(\tau_p\xi + \tau_l\psi)}\right) \tag{3}$$

which yields parameters such as the diffusion coefficient (D) and the amplitudes of the mobile and immobile fractions (A_{mob} and A_{imm}). The average number of mobile molecules per excitation volume on the RICS timescale was determined by

$$N_{\text{mob}} = \left(\frac{\gamma}{A_{\text{mob}}}\right)\left(\frac{2\Delta F}{2\Delta F + 1}\right), \tag{4}$$

where γ is a factor pertaining to the 3D Gaussian shape of the PSF, and $2\Delta F/(2\Delta F + 1)$ is a correction factor when using a moving average subtraction prior to calculating the SACF. The immobilized molecules (i.e. bound fraction) is the contribution of particles that remain visible without significant motion during the acquisition of 5–10 lines of the raster scan, corresponding to ~30 ms. The cross-correlation model was fitted to the cross-correlation function and the extent of cross-correlation was calculated from the amplitude of the mobile fraction of the cross-correlation fit divided by the amplitude of the mobile fraction of the autocorrelation fit of DPPA3-mScarlet.

Statistics and reproducibility. No statistical methods were used to predetermine sample size, the experiments were not randomized, and the investigators were not blinded to allocation during experiments and outcome assessment. Blinding was not implemented in this study as analysis was inherently objective in the overwhelming majority of experiments. For microscopy analysis, where possible, experimenter bias was avoided by selecting fields of view (or individual cells) for acquisition of UHRF1-GFP or DNMT1-GFP signal using the DNA stain (or another marker not being directly assessed in the experiment e.g. DsRed/mScarlet as a readout of Dppa3 induction or RFP-PCNA). To further reduce bias, imaging analysis was subsequently performed indiscriminately on all acquired images using semi-automated analysis pipelines (either with CellProfiler or Fiji scripts). All the experimental findings were reliably reproduced in independent experiments as indicated in the Figure legends. In general, all micrographs from immunofluorescence and live cell imaging, immunoblots, and DNA gel images depicted in this study are representative of $n \geq 2$ independent experiments. The number of replicates used in each experiment are described in the figure legends and/or in the Methods section, as are the Statistical tests used. *P* values or adjusted *P* values are given where possible. Unless otherwise indicated, all statistical calculations were performed using R Studio 1.2.1335. Next-generation sequencing experiments include at least two independent biological replicates. RNA-seq experiments include $n = 4$ biological replicates comprised of $n = 2$ independently cultured samples from two clones (for T1CM, T2CM, T12CM ESCs and EpiLCs) or four independently cultured samples (for wild-type ESCs and EpiLCs). For RRBS experiments, data are derived from $n = 2$ biological replicates. For bisulfite sequencing of LINE-1 elements $n = 2$ biological replicates were analyzed from two independent clones for T1CM, T2CM, T12CM, and Dppa3KO ESCs or two independent cultures for wt ESCs. LC-MS/MS quantification was performed on at least four biological replicates comprising at least two independently cultured samples (usually even more) from $n = 2$ independent clones (T1CM, T2CM, T12CM, and Dppa3KO ESCs) or four independently cultured samples (wild-type ESCs and cell lines shown in Fig. 5d).

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

Sequencing data reported in this paper are available at ArrayExpress (EMBL-EBI) under accessions "E-MTAB-6785" (wild-type and Tet catalytic mutants RRBS), "E-MTAB-6797" (RNA-seq), "E-MTAB-6800" (Dppa3KO RRBS), "E-MTAB-9654" (TaBA-seq of Tet catalytic mutants during Dppa3 induction) and "E-MTAB-9653" (TaBA-seq of Dppa3KO cells expressing *Dppa3* mutant constructs). The raw mass spectrometry proteomics data from the FLAG-DPPA3 pulldown have been deposited at the ProteomeXchange Consortium via the PRIDE partner repository with the dataset identifier "PXD019794". Publically available data sets used in this study can be found here: "GSE77420" (RRBS of TET triple knockout ESCs), "GSE42616" (PRDM14 ChIP-seq), "GSE46111" (5caC-DIP in TDK knockout ESCs), "GSE57700" (TET1 and TET2 ChIP-seq).

Supplementary Data 1 contains the entire list of differentially methylated promoters classified as either "TET-specific", "DPPA3-specific" or "common", which are summarized in Supplementary Fig. 3i. Supplementary Data 2 contains the extended gene ontology analysis of TET-specific promoters with the five most significant terms displayed in Fig. 3e. Supplementary Data 3 contains the complete catalog of proteins interacting with FLAG-DPPA3 in ESCs, which are plotted in Fig. 4b. Supplementary Data 4 contains the full gene ontology analysis of significant DPPA3 interactors. Source data are provided with this paper.

Code availability

The PAM and MIA software is available as source code, requiring MATLAB, or as a precompiled, standalone distribution for Windows or MacOS at http://www.cup.uni-muenchen.de/ pc/lamb/software/pam.html or hosted in Git repositories under http://www.gitlab.com/PAM-PIE/PAM and http://www.gitlab.com/PAM-PIE/PAMcompiled.

Received: 24 March 2020; Accepted: 22 October 2020;
Published online: 24 November 2020

References

1. Ladstätter, S. & Tachibana, K. Genomic insights into chromatin reprogramming to totipotency in embryos. *J. Cell Biol.* **218**, 70–82 (2019).

2. Warren, I. A. et al. Evolutionary impact of transposable elements on genomic diversity and lineage-specific innovation in vertebrates. *Chromosome Res.* **23**, 505–531 (2015).
3. Arkhipova, I. R. Neutral theory, transposable elements, and eukaryotic genome evolution. *Mol. Biol. Evol.* **35**, 1332–1337 (2018).
4. Jacobs, F. M. J. et al. An evolutionary arms race between KRAB zinc-finger genes ZNF91/93 and SVA/L1 retrotransposons. *Nature* **516**, 242–245 (2014).
5. Friedli, M. & Trono, D. The developmental control of transposable elements and the evolution of higher species. *Annu. Rev. Cell Dev. Biol.* **31**, 429–451 (2015).
6. Schmitz, R. J., Lewis, Z. A. & Goll, M. G. DNA methylation: shared and divergent features across eukaryotes. *Trends Genet.* **35**, 818–827 (2019).
7. Schübeler, D. Function and information content of DNA methylation. *Nature* **517**, 321–326 (2015).
8. Rowe, H. M. & Trono, D. Dynamic control of endogenous retroviruses during development. *Virology* **411**, 273–287 (2011).
9. Walsh, C. P., Chaillet, J. R. & Bestor, T. H. Transcription of IAP endogenous retroviruses is constrained by cytosine methylation. *Nat. Genet.* **20**, 116–117 (1998).
10. Jackson-Grusby, L. et al. Loss of genomic methylation causes p53-dependent apoptosis and epigenetic deregulation. *Nat. Genet.* **27**, 31–39 (2001).
11. Chernyavskaya, Y. et al. Loss of DNA methylation in zebrafish embryos activates retrotransposons to trigger antiviral signaling. *Development* **144**, 2925–2939 (2017).
12. Iida, A. et al. Targeted reduction of the DNA methylation level with 5-azacytidine promotes excision of the medaka fish Tol2 transposable element. *Genet. Res.* **87**, 187–193 (2006).
13. Chiappinelli, K. B. et al. Inhibiting DNA methylation causes an interferon response in cancer via dsRNA including endogenous retroviruses. *Cell* **169**, 361 (2017).
14. Roulois, D. et al. DNA-demethylating agents target colorectal cancer cells by inducing viral mimicry by endogenous transcripts. *Cell* **162**, 961–973 (2015).
15. Veenstra, G. J. & Wolffe, A. P. Constitutive genomic methylation during embryonic development of Xenopus. *Biochim. Biophys. Acta* **1521**, 39–44 (2001).
16. Stancheva, I., El-Maarri, O., Walter, J., Niveleau, A. & Meehan, R. R. DNA methylation at promoter regions regulates the timing of gene activation in Xenopus laevis embryos. *Dev. Biol.* **243**, 155–165 (2002).
17. Ortega-Recalde, O., Day, R. C., Gemmell, N. J. & Hore, T. A. Zebrafish preserve global germline DNA methylation while sex-linked rDNA is amplified and demethylated during feminisation. *Nat. Commun.* **10**, 3053 (2019).
18. Skvortsova, K. et al. Retention of paternal DNA methylome in the developing zebrafish germline. *Nat. Commun.* **10**, 3054 (2019).
19. Monk, M., Boubelik, M. & Lehnert, S. Temporal and regional changes in DNA methylation in the embryonic, extraembryonic and germ cell lineages during mouse embryo development. *Development* **99**, 371–382 (1987).
20. Sanford, J. P., Clark, H. J., Chapman, V. M. & Rossant, J. Differences in DNA methylation during oogenesis and spermatogenesis and their persistence during early embryogenesis in the mouse. *Genes Dev.* **1**, 1039–1046 (1987).
21. Lee, H. J., Hore, T. A. & Reik, W. Reprogramming the methylome: erasing memory and creating diversity. *Cell Stem Cell* **14**, 710–719 (2014).
22. Tahiliani, M. et al. Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1. *Science* **324**, 930–935 (2009).
23. He, Y.-F. et al. Tet-mediated formation of 5-carboxylcytosine and its excision by TDG in mammalian DNA. *Science* **333**, 1303–1307 (2011).
24. Cortellino, S. et al. Thymine DNA glycosylase is essential for active DNA demethylation by linked deamination-base excision repair. *Cell* **146**, 67–79 (2011).
25. Ito, S. et al. Tet proteins can convert 5-methylcytosine to 5-formylcytosine and 5-carboxylcytosine. *Science* **333**, 1300–1303 (2011).
26. Rougier, N. et al. Chromosome methylation patterns during mammalian preimplantation development. *Genes Dev.* **12**, 2108–2113 (1998).
27. Howlett, S. K. & Reik, W. Methylation levels of maternal and paternal genomes during preimplantation development. *Development* **113**, 119–127 (1991).
28. Carlson, L. L., Page, A. W. & Bestor, T. H. Properties and localization of DNA methyltransferase in preimplantation mouse embryos: implications for genomic imprinting. *Genes Dev.* **6**, 2536–2541 (1992).
29. Wossidlo, M. et al. 5-Hydroxymethylcytosine in the mammalian zygote is linked with epigenetic reprogramming. *Nat. Commun.* **2**, 241 (2011).
30. Gu, T.-P. et al. The role of Tet3 DNA dioxygenase in epigenetic reprogramming by oocytes. *Nature* **477**, 606–610 (2011).
31. Wang, L. et al. Programming and inheritance of parental DNA methylomes in mammals. *Cell* **157**, 979–991 (2014).
32. Guo, F. et al. Active and passive demethylation of male and female pronuclear DNA in the mammalian zygote. *Cell Stem Cell* **15**, 447–459 (2014).
33. Shen, L. et al. Tet3 and DNA replication mediate demethylation of both the maternal and paternal genomes in mouse zygotes. *Cell Stem Cell* **15**, 459–471 (2014).
34. Amouroux, R. et al. De novo DNA methylation drives 5hmC accumulation in mouse zygotes. *Nat. Cell Biol.* **18**, 225–233 (2016).
35. Iyer, L. M., Abhiman, S. & Aravind, L. Natural history of eukaryotic DNA methylation systems. *Prog. Mol. Biol. Transl. Sci.* **101**, 25–104 (2011).
36. Dean, W. et al. Conservation of methylation reprogramming in mammalian development: aberrant reprogramming in cloned embryos. *Proc. Natl Acad. Sci. USA* **98**, 13734–13738 (2001).
37. Beaujean, N. et al. Effect of limited DNA methylation reprogramming in the normal sheep embryo on somatic cell nuclear transfer. *Biol. Reprod.* **71**, 185–193 (2004).
38. Fulka, H., Mrazek, M., Tepla, O. & Fulka, J. Jr. DNA methylation pattern in human zygotes and developing embryos. *Reproduction* **128**, 703–708 (2004).
39. Chen, T. et al. The DNA methylation events in normal and cloned rabbit embryos. *FEBS Lett.* **578**, 69–72 (2004).
40. Smith, Z. D. et al. A unique regulatory phase of DNA methylation in the early mammalian embryo. *Nature* **484**, 339–344 (2012).
41. Guo, H. et al. The DNA methylation landscape of human early embryos. *Nature* **511**, 606–610 (2014).
42. Gao, F. et al. De novo DNA methylation during monkey pre-implantation embryogenesis. *Cell Res.* **27**, 526–539 (2017).
43. Ivanova, E. et al. DNA methylation changes during preimplantation development reveal inter-species differences and reprogramming events at imprinted genes. *Clin. Epigenetics* **12**, 64 (2020).
44. Macleod, D., Clark, V. H. & Bird, A. Absence of genome-wide changes in DNA methylation during development of the zebrafish. *Nat. Genet.* **23**, 139–140 (1999).
45. Bogdanović, O., Long, S. W. & van Heeringen, S. J. Temporal uncoupling of the DNA methylome and transcriptional repression during embryogenesis. *Genome* Res. **21**, 1313–1327 (2011).
46. Andersen, I. S., Reiner, A. H., Aanes, H., Aleström, P. & Collas, P. Developmental features of DNA methylation during activation of the embryonic zebrafish genome. *Genome Biol.* **13**, R65 (2012).
47. de Mendoza, A. et al. Convergent evolution of a vertebrate-like methylome in a marine sponge. *Nat. Ecol. Evol.* **3**, 1464–1473 (2019).
48. Xu, X. et al. Evolutionary transition between invertebrates and vertebrates via methylation reprogramming in embryogenesis. *Natl Sci. Rev.* **6**, 993–1003 (2019).
49. Marks, H. et al. The transcriptional and epigenomic foundations of ground state pluripotency. *Cell* **149**, 590–604 (2012).
50. Leitch, H. G. et al. Naive pluripotency is associated with global DNA hypomethylation. *Nat. Struct. Mol. Biol.* **20**, 311–316 (2013).
51. Hackett, J. A. et al. Synergistic mechanisms of DNA demethylation during transition to ground-state pluripotency. *Stem Cell Rep.* **1**, 518–531 (2013).
52. Rasmussen, K. D. & Helin, K. Role of TET enzymes in DNA methylation, development, and cancer. *Genes Dev.* **30**, 733–750 (2016).
53. Ito, S. et al. Role of Tet proteins in 5mC to 5hmC conversion, ES-cell self-renewal and inner cell mass specification. *Nature* **466**, 1129–1133 (2010).
54. Ficz, G. et al. FGF signaling inhibition in ESCs drives rapid genome-wide demethylation to the epigenetic ground state of pluripotency. *Cell Stem Cell* **13**, 351–359 (2013).
55. Boroviak, T. et al. Lineage-specific profiling delineates the emergence and progression of naive pluripotency in mammalian embryogenesis. *Dev. Cell* **35**, 366–382 (2015).
56. Habibi, E. et al. Whole-genome bisulfite sequencing of two distinct interconvertible DNA methylomes of mouse embryonic stem cells. *Cell Stem Cell* **13**, 360–369 (2013).
57. Pfaffeneder, T. et al. Tet oxidizes thymine to 5-hydroxymethyluracil in mouse embryonic stem cell DNA. *Nat. Chem. Biol.* **10**, 574–581 (2014).
58. von Meyenn, F. et al. Impairment of DNA methylation maintenance is the main cause of global demethylation in naive embryonic stem cells. *Mol. Cell* **62**, 848–861 (2016).
59. Nakashima, H. et al. Effects of dppa3 on DNA methylation dynamics during primordial germ cell development in mice. *Biol. Reprod.* **88**, 125 (2013).
60. Funaki, S. et al. Inhibition of maintenance DNA methylation by Stella. *Biochem. Biophys. Res. Commun.* **453**, 455–460 (2014).
61. Nakamura, T. et al. PGC7/Stella protects against DNA demethylation in early embryogenesis. *Nat. Cell Biol.* **9**, 64–71 (2007).
62. Li, Y. et al. Stella safeguards the oocyte methylome by preventing de novo methylation mediated by DNMT1. *Nature* **564**, 136–140 (2018).
63. Hayashi, K., Lopes, S. M. C., de, S., Tang, F. & Surani, M. A. Dynamic equilibrium and heterogeneity of mouse pluripotent stem cells with distinct functional and epigenetic states. *Cell Stem Cell* **3**, 391–401 (2008).
64. Kalkan, T. et al. Tracking the embryonic stem cell transition from ground state pluripotency. *Development* **144**, 1221–1234 (2017).

65. Auclair, G., Guibert, S., Bender, A. & Weber, M. Ontogeny of CpG island methylation and specificity of DNMT3 methyltransferases during embryonic development in the mouse. *Genome Biol.* **15**, 545 (2014).
66. Khoueiry, R. et al. Lineage-specific functions of TET1 in the postimplantation mouse embryo. *Nat. Genet.* **49**, 1061–1072 (2017).
67. Xiong, J. et al. Cooperative action between SALL4A and TET proteins in stepwise oxidation of 5-methylcytosine. *Mol. Cell* **64**, 913–925 (2016).
68. Shen, L. et al. Genome-wide analysis reveals TET- and TDG-dependent 5-methylcytosine oxidation dynamics. *Cell* **153**, 692–706 (2013).
69. Sharif, J. et al. Activation of endogenous retroviruses in Dnmt1−/− ESCs involves disruption of SETDB1-mediated repression by NP95 Binding to Hemimethylated DNA. *Cell Stem Cell* **19**, 81–94 (2016).
70. Magnúsdóttir, E. et al. A tripartite transcription factor network regulates primordial germ cell specification in mice. *Nat. Cell Biol.* **15**, 905–915 (2013).
71. Yamaji, M. et al. PRDM14 ensures naive pluripotency through dual regulation of signaling and epigenetic pathways in mouse embryonic stem cells. *Cell Stem Cell* **12**, 368–382 (2013).
72. Okashita, N. et al. PRDM14 promotes active DNA demethylation through the ten-eleven translocation (TET)-mediated base excision repair pathway in embryonic stem cells. *Development* **141**, 269–280 (2014).
73. Grabole, N. et al. Prdm14 promotes germline fate and naive pluripotency by repressing FGF signalling and DNA methylation. *EMBO Rep.* **14**, 629–637 (2013).
74. Han, L. et al. Embryonic defects induced by maternal obesity in mice derive from Stella insufficiency in oocytes. *Nat. Genet.* **50**, 432–442 (2018).
75. Xu, X. et al. Dppa3 expression is critical for generation of fully reprogrammed iPS cells and maintenance of Dlk1-Dio3 imprinting. *Nat. Commun.* **6**, 6008 (2015).
76. Huang, Y. et al. Stella modulates transcriptional and endogenous retrovirus programs during maternal-to-zygotic transition. *Elife* **6**, e22345 (2017).
77. Nakamura, T. et al. PGC7 binds histone H3K9me2 to protect against conversion of 5mC to 5hmC in early embryos. *Nature* **486**, 415–419 (2012).
78. Shin, S.-W., John Vogt, E., Jimenez-Movilla, M., Baibakov, B. & Dean, J. Cytoplasmic cleavage of DPPA3 is required for intracellular trafficking and cleavage-stage development in mice. *Nat. Commun.* **8**, 1643 (2017).
79. Bogdanović, O. et al. Active DNA demethylation at enhancers during the vertebrate phylotypic period. *Nat. Genet.* **48**, 417–426 (2016).
80. Sharif, J. et al. The SRA protein Np95 mediates epigenetic inheritance by recruiting Dnmt1 to methylated DNA. *Nature* **450**, 908–912 (2007).
81. Bostick, M. et al. UHRF1 plays a role in maintaining DNA methylation in mammalian cells. *Science* **317**, 1760–1764 (2007).
82. Nishiyama, A. et al. Uhrf1-dependent H3K23 ubiquitylation couples maintenance DNA methylation and replication. *Nature* **502**, 249–253 (2013).
83. Qin, W. et al. DNA methylation requires a DNMT1 ubiquitin interacting motif (UIM) and histone ubiquitination. *Cell Res.* **25**, 911–929 (2015).
84. Rothbart, S. B. et al. Association of UHRF1 with methylated H3K9 directs the maintenance of DNA methylation. *Nat. Struct. Mol. Biol.* **19**, 1155–1160 (2012).
85. Zhao, Q. et al. Dissecting the precise role of H3K9 methylation in crosstalk with DNA maintenance methylation in mammals. *Nat. Commun.* **7**, 12464 (2016).
86. Weihs, F. et al. Heterogeneous localisation of membrane proteins in Staphylococcus aureus. *Sci. Rep.* **8**, 3657 (2018).
87. Osswald, M., Santos, A. F. & Morais-de-Sá, E. Light-Induced Protein Clustering for Optogenetic Interference and Protein Interaction Analysis in Drosophila S2 Cells. *Biomolecules* **9**, 61 (2019).
88. Rothbart, S. B. et al. Multivalent histone engagement by the linked tandem Tudor and PHD domains of UHRF1 is required for the epigenetic inheritance of DNA methylation. *Genes Dev.* **27**, 1288–1298 (2013).
89. Citterio, E. et al. Np95 is a histone-binding protein endowed with ubiquitin ligase activity. *Mol. Cell. Biol.* **24**, 2526–2535 (2004).
90. Karagianni, P., Amazit, L., Qin, J. & Wong, J. ICBP90, a novel methyl K9 H3 binding protein linking protein ubiquitination with heterochromatin formation. *Mol. Cell. Biol.* **28**, 705–717 (2008).
91. Digman, M. A. et al. Measuring fast dynamics in solutions and cells with a laser scanning microscope. *Biophys. J.* **89**, 1317–1327 (2005).
92. Digman, M. A., Wiseman, P. W., Horwitz, A. R. & Gratton, E. Detecting protein complexes in living cells from laser scanning confocal image sequences by the cross correlation raster image spectroscopy method. *Biophys. J.* **96**, 707–716 (2009).
93. Karg, E. et al. Ubiquitome analysis reveals PCNA-associated factor 15 (PAF15) as a specific ubiquitination target of UHRF1 in embryonic stem cells. *J. Mol. Biol.* **429**, 3814–3824 (2017).
94. Fang, J. et al. Hemi-methylated DNA opens a closed conformation of UHRF1 to facilitate its histone recognition. *Nat. Commun.* **7**, 11197 (2016).
95. Harrison, J. S. et al. Hemi-methylated DNA regulates DNA methylation inheritance through allosteric activation of H3 ubiquitylation by UHRF1. *Elife* **5**, e17101 (2016).
96. Herce, H. D., Deng, W., Helma, J., Leonhardt, H. & Cardoso, M. C. Visualization and targeted disruption of protein interactions in living cells. *Nat. Commun.* **4**, 2660 (2013).
97. Arita, K. et al. Recognition of modification status on a histone H3 tail by linked histone reader modules of the epigenetic regulator UHRF1. *Proc. Natl Acad. Sci. USA* **109**, 12950–12955 (2012).
98. Du, W. et al. Stella protein facilitates DNA demethylation by disrupting the chromatin association of the RING finger-type E3 ubiquitin ligase UHRF1. *J. Biol. Chem.* **294**, 8907–8917 (2019).
99. Feng, S. et al. Conservation and divergence of methylation patterning in plants and animals. *Proc. Natl Acad. Sci. USA* **107**, 8689–8694 (2010).
100. Iyer, L. M., Tahiliani, M., Rao, A. & Aravind, L. Prediction of novel families of enzymes involved in oxidative and other complex modifications of bases in nucleic acids. *Cell Cycle* **8**, 1698–1710 (2009).
101. Wu, S. C. & Zhang, Y. Active DNA demethylation: many roads lead to Rome. *Nat. Rev. Mol. Cell Biol.* **11**, 607–620 (2010).
102. Kumar, S. & Hedges, S. B. A molecular timescale for vertebrate evolution. *Nature* **392**, 917–920 (1998).
103. Ishiyama, S. et al. Structure of the Dnmt1 reader module complexed with a unique two-mono-ubiquitin mark on histone H3 reveals the basis for DNA methylation maintenance. *Mol. Cell* **68**, 350–360.e7 (2017).
104. Yamaguchi, L. et al. Usp7-dependent histone H3 deubiquitylation regulates maintenance of DNA methylation. *Sci. Rep.* **7**, 55 (2017).
105. Walter, R. B., Li, H.-Y., Intano, G. W., Kazianis, S. & Walter, C. A. Absence of global genomic cytosine methylation pattern erasure during medaka (Oryzias latipes) early embryo development. *Comp. Biochem. Physiol. B Biochem. Mol. Biol.* **133**, 597–607 (2002).
106. Wu, X. & Zhang, Y. TET-mediated active DNA demethylation: mechanism, function and beyond. *Nat. Rev. Genet.* https://doi.org/10.1038/nrg.2017.33 (2017).
107. Singer, Z. S. et al. Dynamic heterogeneity and DNA methylation in embryonic stem cells. *Mol. Cell* **55**, 319–331 (2014).
108. Almeida, R. D. et al. 5-hydroxymethyl-cytosine enrichment of non-committed cells is not a universal feature of vertebrate development. *Epigenetics* **7**, 383–389 (2012).
109. Blinka, S., Reimer, M. H. Jr, Pulakanti, K. & Rao, S. Super-enhancers at the nanog locus differentially regulate neighboring pluripotency-associated genes. *Cell Rep.* **17**, 19–28 (2016).
110. Liao, J. et al. Targeted disruption of DNMT1, DNMT3A and DNMT3B in human embryonic stem cells. *Nat. Genet.* **47**, 469–478 (2015).
111. Ma, Z., Swigut, T., Valouev, A., Rada-Iglesias, A. & Wysocka, J. Sequence-specific regulator Prdm14 safeguards mouse ESCs from entering extraembryonic endoderm fates. *Nat. Struct. Mol. Biol.* **18**, 120–127 (2011).
112. Rahjouei, A., Pirouz, M., Di Virgilio, M., Kamin, D. & Kessel, M. MAD2L2 promotes open chromatin in embryonic stem cells and derepresses the Dppa3 locus. *Stem Cell Rep.* **8**, 813–821 (2017).
113. Sang, H. et al. Dppa3 is critical for Lin28a-regulated ES cells naïve-primed state conversion. *J. Mol. Cell Biol.* **11**, 474–488 (2019).
114. Mochizuki, K., Tachibana, M., Saitou, M., Tokitake, Y. & Matsui, Y. Implication of DNA demethylation and bivalent histone modification for selective gene regulation in mouse primordial germ cells. *PLoS ONE* **7**, e46036 (2012).
115. Waghray, A. et al. Tbx3 controls Dppa3 levels and exit from pluripotency toward mesoderm. *Stem Cell Rep.* **5**, 97–110 (2015).
116. Li, X. et al. Tet proteins influence the balance between neuroectodermal and mesodermal fate choice by inhibiting Wnt signaling. *Proc. Natl Acad. Sci. USA* **113**, E8267–E8276 (2016).
117. Dai, H.-Q. et al. TET-mediated DNA demethylation controls gastrulation by regulating Lefty–Nodal signalling. *Nature* **538**, 528 (2016).
118. Verma, N. et al. TET proteins safeguard bivalent promoters from de novo methylation in human embryonic stem cells. *Nat. Genet.* **50**, 83–95 (2018).
119. Gu, T. et al. DNMT3A and TET1 cooperate to regulate promoter epigenetic landscapes in mouse embryonic stem cells. *Genome Biol.* **19**, 88 (2018).
120. López-Moyado, I. F. et al. Paradoxical association of TET loss of function with genome-wide DNA hypomethylation. *Proc. Natl Acad. Sci. USA* **116**, 16933–16942 (2019).
121. Han, L., Ren, C., Zhang, J., Shu, W. & Wang, Q. Differential roles of Stella in the modulation of DNA methylation during oocyte and zygotic development. *Cell Disco.* **5**, 9 (2019).
122. Payer, B. et al. Generation of stella-GFP transgenic mice: a novel tool to study germ cell development. *Genesis* **44**, 75–83 (2006).
123. Macfarlan, T. S. et al. Embryonic stem cell potency fluctuates with endogenous retrovirus activity. *Nature* **487**, 57–63 (2012).
124. Dupressoir, A., Lavialle, C. & Heidmann, T. From ancestral infectious retroviruses to bona fide cellular genes: role of the captured syncytins in placentation. *Placenta* **33**, 663–671 (2012).
125. Chuong, E. B., Rumi, M. A. K., Soares, M. J. & Baker, J. C. Endogenous retroviruses function as species-specific enhancer elements in the placenta. *Nat. Genet.* **45**, 325–329 (2013).

126. Grow, E. J. et al. Intrinsic retroviral reactivation in human preimplantation embryos and pluripotent cells. *Nature* **522**, 221–225 (2015).
127. Fu, B., Ma, H. & Liu, D. Endogenous retroviruses function as gene expression regulatory elements during mammalian pre-implantation embryo development. *Int. J. Mol. Sci.* **20**, 790 (2019).
128. Hendrickson, P. G. et al. Conserved roles of mouse DUX and human DUX4 in activating cleavage-stage genes and MERVL/HERVL retrotransposons. *Nat. Genet.* **49**, 925–934 (2017).
129. Percharde, M. et al. A LINE1-nucleolin partnership regulates early development and ESC identity. *Cell* **174**, 391–405.e19 (2018).
130. Rodriguez-Terrones, D. & Torres-Padilla, M.-E. Nimble and ready to mingle: transposon outbursts of early development. *Trends Genet* **34**, 806–820 (2018).
131. Gill, T. J. III, Johnson, P. M., Lyden, T. W. & Mwenda, J. M. Endogenous retroviral expression in the human placenta. *Am. J. Reprod. Immunol.* **23**, 115–120 (1990).
132. Harris, J. R. Placental endogenous retrovirus (ERV): structural, functional, and evolutionary significance. *Bioessays* **20**, 307–316 (1998).
133. Wildman, D. E. et al. Evolution of the mammalian placenta revealed by phylogenetic analysis. *Proc. Natl Acad. Sci. USA* **103**, 3203–3208 (2006).
134. Selwood, L. & Johnson, M. H. Trophoblast and hypoblast in the monotreme, marsupial and eutherian mammal: evolution and origins. *Bioessays* **28**, 128–145 (2006).
135. Emera, D. & Wagner, G. P. Transposable element recruitments in the mammalian placenta: impacts and mechanisms. *Brief. Funct. Genomics* **11**, 267–276 (2012).
136. Frank, J. A. & Feschotte, C. Co-option of endogenous viral sequences for host cell function. *Curr. Opin. Virol.* **25**, 81–89 (2017).
137. Villarreal, L. P. Persistent virus and addiction modules: an engine of symbiosis. *Curr. Opin. Microbiol.* **31**, 70–79 (2016).
138. Zeh, D. W. & Zeh, J. A. Reproductive mode and speciation: the viviparity-driven conflict hypothesis. *Bioessays* **22**, 938–946 (2000).
139. Jangam, D., Feschotte, C. & Betrán, E. Transposable element domestication as an adaptation to evolutionary conflicts. *Trends Genet.* **33**, 817–831 (2017).
140. Renfree, M. B., Hore, T. A., Shaw, G., Graves, J. A. M. & Pask, A. J. Evolution of genomic imprinting: insights from marsupials and monotremes. *Annu. Rev. Genomics Hum. Genet.* **10**, 241–262 (2009).
141. Pask, A. J. et al. Analysis of the platypus genome suggests a transposon origin for mammalian imprinting. *Genome Biol.* **10**, R1 (2009).
142. Kaneko-Ishino, T. & Ishino, F. The role of genes domesticated from LTR retrotransposons and retroviruses in mammals. *Front. Microbiol.* **3**, 262 (2012).
143. Bogutz, A. B. et al. Evolution of imprinting via lineage-specific insertion of retroviral promoters. *Nat. Commun.* **10**, 5674 (2019).
144. Takahashi, N. et al. ZNF445 is a primary regulator of genomic imprinting. *Genes Dev.* **33**, 49–54 (2019).
145. Imbeault, M., Helleboid, P.-Y. & Trono, D. KRAB zinc-finger proteins contribute to the evolution of gene regulatory networks. *Nature* **543**, 550–554 (2017).
146. Edwards, C. A., Takahashi, N., Corish, J. A. & Ferguson-Smith, A. C. The origins of genomic imprinting in mammals. *Reprod. Fertil. Dev.* **31**, 1203–1218 (2019).
147. Lynch, V. J., Leclerc, R. D., May, G. & Wagner, G. P. Transposon-mediated rewiring of gene regulatory networks contributed to the evolution of pregnancy in mammals. *Nat. Genet.* **43**, 1154–1159 (2011).
148. Lynch, V. J. et al. Ancient transposable elements transformed the uterine regulatory landscape and transcriptome during the evolution of mammalian pregnancy. *Cell Rep.* **10**, 551–561 (2015).
149. Frankenberg, S. R., de Barros, F. R. O., Rossant, J. & Renfree, M. B. The mammalian blastocyst. *Wiley Interdiscip. Rev.: Developmental Biol.* **5**, 210–232 (2016).
150. Chavan, A. R., Griffith, O. W. & Wagner, G. P. The inflammation paradox in the evolution of mammalian pregnancy: turning a foe into a friend. *Curr. Opin. Genet. Dev.* **47**, 24–32 (2017).
151. Kunarso, G. et al. Transposable elements have rewired the core regulatory network of human embryonic stem cells. *Nat. Genet.* **42**, 631–634 (2010).
152. Jacques, P.-É., Jeyakani, J. & Bourque, G. The majority of primate-specific regulatory sequences are derived from transposable elements. *PLoS Genet.* **9**, e1003504 (2013).
153. Esnault, C., Cornelis, G., Heidmann, O. & Heidmann, T. Differential evolutionary fate of an ancestral primate endogenous retrovirus envelope gene, the EnvV syncytin, captured for a function in placentation. *PLoS Genet.* **9**, e1003400 (2013).
154. Haig, D. A. Going retro: transposable elements, embryonic stem cells, and the mammalian placenta. *Bioessays* 37, 1154 (2015).
155. Haig, D. Transposable elements: Self-seekers of the germline, team-players of the soma. *Bioessays* **38**, 1158–1166 (2016).
156. Chuong, E. B. Retroviruses facilitate the rapid evolution of the mammalian placenta. *Bioessays* **35**, 853–861 (2013).
157. Smith, Z. D. et al. DNA methylation dynamics of the human preimplantation embryo. *Nature* **511**, 611–615 (2014).
158. Sanford, J. P., Chapman, V. M. & Rossant, J. DNA methylation in extraembryonic lineages of mammals. *Trends Genet.* **1**, 89–93 (1985).
159. Smith, Z. D. et al. Epigenetic restriction of extraembryonic lineages mirrors the somatic transition to cancer. *Nature* **549**, 543–547 (2017).
160. Peaston, A. E. et al. Retrotransposons regulate host genes in mouse oocytes and preimplantation embryos. *Dev. Cell* **7**, 597–606 (2004).
161. Meyer, T. J., Rosenkrantz, J. L., Carbone, L. & Chavez, S. L. Endogenous retroviruses: with us and against us. *Front Chem.* **5**, 23 (2017).
162. Schroeder, D. I. et al. Early developmental and evolutionary origins of gene body DNA methylation patterns in mammalian placentas. *PLoS Genet.* **11**, e1005442 (2015).
163. Sato, M. et al. Identification of PGC7, a new gene expressed specifically in preimplantation embryos and germ cells. *Mech. Dev.* **113**, 91–94 (2002).
164. Thélie, A. et al. Differential regulation of abundance and deadenylation of maternal transcripts during bovine oocyte maturation in vitro and in vivo. *BMC Dev. Biol.* **7**, 125 (2007).
165. Masala, L. et al. Delay in maternal transcript degradation in ovine embryos derived from low competence oocytes. *Mol. Reprod. Dev.* **85**, 427–439 (2018).
166. Wasielak, M., Więsak, T., Bogacka, I., Jalali, B. M. & Bogacki, M. Zygote arrest 1, nucleoplasmin 2, and developmentally associated protein 3 mRNA profiles throughout porcine embryo development in vitro. *Theriogenology* **86**, 2254–2262 (2016).
167. Boroviak, T. et al. Single cell transcriptome analysis of human, marmoset and mouse embryos reveals common and divergent features of preimplantation development. *Development* **145**, dev.167833 (2018).
168. Hayashi, K. & Saitou, M. Generation of eggs from mouse embryonic stem cells and induced pluripotent stem cells. *Nat. Protoc.* **8**, 1513–1524 (2013).
169. Mulholland, C. B. et al. A modular open platform for systematic functional studies under physiological conditions. *Nucleic Acids Res.* **43**, e112 (2015).
170. Kowarz, E., Löscher, D. & Marschalek, R. Optimized Sleeping Beauty transposons rapidly generate stable transgenic cell lines. *Biotechnol. J.* **10**, 647–653 (2015).
171. Gibson, D. G. et al. Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nat. Methods* **6**, 343–345 (2009).
172. Ran, F. A. et al. Genome engineering using the CRISPR-Cas9 system. *Nat. Protoc.* **8**, 2281–2308 (2013).
173. Gutschner, T., Haemmerle, M., Genovese, G., Draetta, G. F. & Chin, L. Post-translational regulation of Cas9 during G1 enhances homology-directed repair. *Cell Rep.* **14**, 1555–1566 (2016).
174. Mátés, L. et al. Molecular evolution of a novel hyperactive Sleeping Beauty transposase enables robust stable gene transfer in vertebrates. *Nat. Genet.* **41**, 753–761 (2009).
175. Méndez, J. & Stillman, B. Chromatin association of human origin recognition complex, cdc6, and minichromosome maintenance proteins during the cell cycle: assembly of prereplication complexes in late mitosis. *Mol. Cell. Biol.* **20**, 8602–8612 (2000).
176. Bauer, C. et al. Phosphorylation of TET proteins is regulated via O-GlcNAcylation by the O-linked N-acetylglucosamine transferase (OGT). *J. Biol. Chem.* **290**, 4801–4812 (2015).
177. Shintomi, K. & Hirano, T. Releasing cohesin from chromosome arms in early mitosis: opposing actions of Wapl–Pds5 and Sgo1. *Genes Dev.* 23, 2224–2236 (2009).
178. Françon, P. et al. A hypophosphorylated form of RPA34 is a specific component of pre-replication centers. *J. Cell Sci.* **117**, 4909–4920 (2004).
179. Wagner, M. et al. Age-dependent levels of 5-methyl-, 5-hydroxymethyl-, and 5-formylcytosine in human and mouse brain tissues. *Angew. Chem. Int. Ed. Engl.* **54**, 12511–12514 (2015).
180. Soumillon, M., Cacchiarelli, D., Semrau, S., van Oudenaarden, A. & Mikkelsen, T. S. Characterization of directed differentiation by high-throughput single-cell RNA-Seq. *bioRxiv* https://doi.org/10.1101/003236 (2014).
181. Ziegenhain, C. et al. Comparative analysis of single-cell RNA sequencing methods. *Mol. Cell* **65**, 631–643.e4 (2017).
182. Bagnoli, J. W. et al. Sensitive and powerful single-cell RNA sequencing using mcSCRB-seq. *Nat. Commun.* **9**, 2937 (2018).
183. Boyle, P. et al. Gel-free multiplexed reduced representation bisulfite sequencing for large-scale DNA methylation profiling. *Genome Biol.* **13**, R92 (2012).
184. Parekh, S., Ziegenhain, C., Vieth, B., Enard, W. & Hellmann, I. zUMIs—a fast and flexible pipeline to process RNA sequencing data with UMIs. *Gigascience* **7**, giy059 (2018).
185. Rau, A., Gallopin, M., Celeux, G. & Jaffrézic, F. Data-based filtering for replicated high-throughput transcriptome sequencing experiments. *Bioinformatics* **29**, 2146–2152 (2013).
186. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
187. Gu, Z., Eils, R. & Schlesner, M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics* **32**, 2847–2849 (2016).

188. Akalin, A. et al. methylKit: a comprehensive R package for the analysis of genome-wide DNA methylation profiles. *Genome Biol.* **13**, R87 (2012).
189. Illingworth, R. S. et al. Orphan CpG islands identify numerous conserved promoters in the mammalian genome. *PLoS Genet.* **6**, e1001134 (2010).
190. Zhang, Y. et al. Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* **9**, R137 (2008).
191. Ramírez, F. et al. deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res.* **44**, W160–W165 (2016).
192. Sporbert, A., Domaing, P., Leonhardt, H. & Cardoso, M. C. PCNA acts as a stationary loading platform for transiently interacting Okazaki fragment maturation proteins. *Nucleic Acids Res.* **33**, 3521–3528 (2005).
193. Schindelin, J. et al. Fiji: an open-source platform for biological-image analysis. *Nat. Methods* **9**, 676–682 (2012).
194. Schneider, C. A., Rasband, W. S. & Eliceiri, K. W. NIH Image to ImageJ: 25 years of image analysis. *Nat. Methods* **9**, 671–675 (2012).
195. McQuin, C. et al. CellProfiler 3.0: Next-generation image processing for biology. *PLoS Biol.* **16**, e2005970 (2018).
196. Letunic, I. & Bork, P. Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. *Bioinformatics* **23**, 127–128 (2007).
197. Rappsilber, J., Mann, M. & Ishihama, Y. Protocol for micro-purification, enrichment, pre-fractionation and storage of peptides for proteomics using StageTips. *Nat. Protoc.* **2**, 1896–1906 (2007).
198. Scheltema, R. A. & Mann, M. SprayQc: a real-time LC–MS/MS quality monitoring system to maximize uptime using off the shelf components. *J. Proteome Res.* **11**, 3458–3466 (2012).
199. Cox, J. & Mann, M. MaxQuant enables high peptide identification rates, individualized ppb-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.* **26**, 1367 (2008).
200. Cox, J. et al. Andromeda: a peptide search engine integrated into the MaxQuant environment. *J. Proteome Res.* **10**, 1794–1805 (2011).
201. Cox, J., Hein, M. Y., Luber, C. A., Paron, I. & Nagaraj, N. MaxLFQ allows Accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction, termed MaxLFQ. *Mol. Cell. Proteomics* **13**, 2513–2526 (2014).
202. Tyanova, S. et al. The Perseus computational platform for comprehensive analysis of (prote)omics data. *Nat. Methods* **13**, 731 (2016).
203. Ashburner, M. et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25**, 25–29 (2000).
204. Iwamatsu, T. Stages of normal development in the medaka Oryzias latipes. *Mech. Dev.* **121**, 605–618 (2004).
205. Hendrix, J. et al. Live-cell observation of cytosolic HIV-1 assembly onset reveals RNA-interacting Gag oligomers. *J. Cell Biol.* **210**, 629–646 (2015).
206. Müller, B. K., Zaychikov, E., Bräuchle, C. & Lamb, D. C. Pulsed interleaved excitation. *Biophys. J.* **89**, 3508–3522 (2005).
207. Schrimpf, W., Barth, A., Hendrix, J. & Lamb, D. C. PAM: a framework for integrated analysis of imaging, single-molecule, and ensemble fluorescence data. *Biophys. J.* **114**, 1518–1528 (2018).
208. Hendrix, J., Dekens, T., Schrimpf, W. & Lamb, D. C. Arbitrary-region raster image correlation spectroscopy. *Biophys. J.* **111**, 1785–1796 (2016).
209. Wang, H.-Q., Tuominen, L. K. & Tsai, C.-J. SLIM: a sliding linear model for estimating the proportion of true null hypotheses in datasets with dependence structures. *Bioinformatics* **27**, 225–231 (2011).
210. Payer, B. et al. Stella is a maternal effect gene required for normal early development in mice. *Curr. Biol.* **13**, 2110–2117 (2003).

Acknowledgements

The authors would like to thank Dr. H. Blum, Dr. S. Krebs (Laboratory for Functional Genome Analysis, LMU Munich), and Dr. A. Brachmann (Faculty of Biology, Department of Genetics, LMU Munich) for next-generation sequencing services. We further thank Dr. Ralf Heermann (Bioanalytics core facility, LMU Munich) for help with the MST measurements. We thank V. Laban, C. Kirschner, T.J. Fischer, D. Nixdorf, and V.B. Ozan for help with experiments, J. Koch for technical assistance, Dr. M. Mann and Dr. I. Paron for mass spectrometry, and Dr. I. Hellmann for advice on data analysis and providing computational infrastructure. We would like to thank Dr. S. Stricker and Dr. I. Solovei for helpful discussions and constructive criticism on the manuscript, E. Ntouliou for critical reading of the manuscript, and G. Mahler for inspiration. We thank Dr. Feng Zhang for providing the pSpCas9(BB)-2A-Puro (PX459) and SpCas9-T2A-GFP (PX458) (Addgene plasmids #62988 and #48138) plasmids, Dr. Eric Kowarz for the gift of the pSB-tetPur (Addgene plasmid #60507) construct, and Dr. Zsuzsanna Izsvak for the gift of the pCMV(CAT)T7-SB100 (Addgene plasmid #34879) construct. J.R., C.T., E.U., M.M., and M.D.B. are fellows of the International Max Planck Research School for Molecular Life Sciences (IMPRS-LS). C.B.M. gratefully acknowledges the support of the Fulbright Commission and the late Dr. Glenn Cuomo. F.R.T. thanks the Boehringer Ingelheim Fonds and J.R. the Fonds de Recherche du Québec en Santé for a PhD fellowship. This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project-ID 213249687 – SFB 1064 to S.B. (A22) and H.L. (A17), by SFB1243/A01 to H.L., and SFB1243/A14 to W.E.

Author contributions

C.B.M. and S.B. designed and conceived the study. S.B. and H.L. supervised the study. C.B.M., S.B., and H.L. prepared the manuscript with the help of M.D.B. C.B.M. performed cellular and molecular experiments. C.B.M. generated cell lines with help from M.Y., C.B.M. performed RRBS and RNA-Seq with help and supervision from C.Z., S.B., and W.E., J.R. and C.B.M. performed live-cell microscopy and photobleaching analyses. I.G., J.R., and C.B.M. performed RICS experiments under the supervision of D.C.L., C.T. performed MST and F3H assays. A.N. performed Xenopus experiments under the supervision of M.N., R.N. performed the experiments in medaka embryos under the supervision of H.T. M.D.B., and P.S. helped with cell line validation and performed fluorescence microscopy analysis. W.Q. performed the biochemical analyses with assistance from A.A., M.M. performed hESC experiments. E.U. conducted proteomics experiments and analyses under the guidance of M.W. F.R.T., and E.P. quantified modified cytosines by LC-MS/MS with the supervision by T.C. S.B. performed data analysis. All authors read, discussed, and approved the manuscript.

Funding

Open Access funding enabled and organized by Projekt DEAL.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41467-020-19603-1.

Correspondence and requests for materials should be addressed to S.B. or H.L.

Peer review information *Nature Communications* thanks Maxim Greenberg, Michael Weber and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2020, corrected publication 2020

5.2 List of Abbreviations

2C stage	Two-cell stage
5caC	5-carboxylcytosine
5fC	5-formylcytosine
5hmC	5-hydroxymethylcytosine
5mC	Cytosine DNA methylation
aa	Amino acids
Ac	Acetylation
AI	Artificial intelligence
AML	Acute myeloid leukemia
AP	Affinity purification
bFGF	Basic fibroblast growth factor
BirA	Biotin ligase
bp	Base pair
C_{18}	Octadecylsilane
ChAC	Chromatin Aggregation Capture
ChAP	Chromatin Affinity Purification
CHAPS	3-((3-cholamidopropyl) dimethylammonio)-1-propanesulfonate
ChEP	Chromatin Enrichment for Proteomics
ChIP	Chromatin Immunoprecipitation
ChroP	Chromatin proteomics
CID	Collisional induced dissociation
CRISPR	Clustered regularly interspaced short palindromic repeats
Da	Dalton
dCas9	Catalytically inactive (dead) Cas9
DDA	Data-dependent acquisition
DIA	Data-independent acquisition
Dm-ChP	DNA-mediated Chromatin Pull-down
DNA	Deoxyribonucleic acid
DNaseI	Deoxyribonuclease I
DNMT1	DNA methyltransferase 1
DNN	Deep neural network
DREX	*Drosophila* preblastoderm embryo chromatin assembly extract
DSG	Disuccinimidyl glutarate
DSSO	Disuccinimidyl sulfoxide
E	Embryonic day
EdU	5-Ethynyl-2'-deoxyuridine

EGS	Ethylene glycol bis(succinimidyl succinate)
EpiLC	Epiblast-like cell
EpiSC	Epiblast stem cell
ERV	Endogenous retrovirus
ESI	Electrospray ionization
FA	Formaldehyde
FDR	False discovery rate
FSC	Formative stem cell
GdCl	Guanidinium hydrochloride
GO	Gene ontology
HCD	Higher energy collisional dissociation
HPLC	High-performance liquid chromatography
hPSC	Human pluripotent stem cell
hPTM	Histone post-translational modifications
iCLASPI	*In vivo* Crosslinking-Assisted and Stable Isotope Labeling by Amino acids in Cell culture
ICM	Inner cell mass
IP	Immunoprecipitation
iPOND	Isolation of Proteins on Nascent DNA
iPSC	Induced pluripotent stem cell
K	Lysine
KO	Knockout
LC	Liquid chromatography
LC-MS/MS	Liquid chromatography coupled with tandem mass spectrometry
LIF	Leukemia inhibitory factor
LINE	Long interspersed nuclear element
lncRNA	Long non-coding ribonucleic acid
MALDI	Matrix-assisted laser desorption/ionization
me(1/2/3)	Mono-, di- or trimethylation
mESC	Mouse embryonic stem cell
MNase	Micrococcal nuclease
mRNA	Messenger ribonucleic acid
MS	Mass spectrometry
MS1 scan	Full mass scan of all incoming ions
MS2 scan	MS scan of fragmented peptide ions
MZT	Maternal-to-zygotic transition
ncAA	Non-canonical amino acids
NEBD	Nuclear envelope breakdown
NeXO	Network-extracted ontology
NPC	Neural progenitor cell
PAC	Protein aggregation capture

PAGE	Polyacrylamide gel electrophoresis
PICh	Proteomics of isolated chromatin segments
PRC	Polycomb repressive complex
PSC	Pluripotent stem cell
PSM	Peptide spectrum match
PTM	Post-translational modifications
RIME	Rapid Immunoprecipitation Mass Spec of Endogenous proteins
RNaseA	Ribonuclease A
SAH	S-adenosylhomocysteine
SAM	S-adenosylmethionine
SAX	Strong anion exchange
SCX	Strong cation exchange
SDC	Sodium deoxycholate
SDS	Sodium dodecyl sulfate
SICAP	Selective Isolation of Chromatin-Associated Proteins
StageTip	Stop And Go Extraction tips
synchro-PASEF	Synchronized parallel accumulation - serial fragmentation
TDG	Thymine DNA glycosylase
TE	Transposable element
TET	Ten-eleven Translocation
Th	Thomson (unit of m/z)
TMT	Tandem mass tag
TOF	Time-of-flight
ZGA	Zygotic genome activation
m/z	The mass-to-charge ratio

5.3 Curriculum Vitae

PROFESSIONAL EXPERIENCE

Since 08/2022	**University lecturer** Ludwig-Maximilians-Universität München (LMU) Munich, Leonhardt lab
Since 07/2018	**PhD candidate** LMU Munich, Leonhardt lab
Since 07/2018	**Guest Scientist** Max Planck Institute (MPI) of Biochemistry, Mann lab
09/2017 – 06/2018	**Scientific Assistant** MPI of Biochemistry, Mann lab
09/2016 – 03/2018	**Scientific Assistant** LMU Munich, Biomedical center, Becker lab

EDUCATION

07/2018 – 09/2023	**Doctorate (Dr. rer. nat.) in biology** LMU Munich Graduate program: International Max Planck Research School for Molecular Life Sciences, Martinsried, Germany
10/2015 – 03/2018	**Master of Science (M.Sc.) in biochemistry** LMU Munich
10/2011 – 02/2015	**Bachelor of Science (B.Sc.) molecular medicine** Georg-August University Göttingen 2-year fellow of German National Academic Foundation (Studienstiftung des deutschen Volkes)
08/2011	**Abitur – German qualification for university** Ruhrtal-Gymnasium Schwerte

5.4 Acknowledgements

Over the past five years, I spent time on research that captivated my interest. This journey wouldn't have been possible without the support and companionship of many individuals.

Foremost, I would like to thank Heinrich Leonhardt. Your mentorship has been unparalleled, and the "professional playground" you provided became an instrumental setting for my growth. Your belief in my work and guidance on communicating my results have been invaluable.

Michael Wierer, you have my sincere gratitude for your guidance, especially during the early phases of my PhD. Michi, your teachings have equipped me with foundational tools every scientist needs.

My special gratitude goes to Matthias Mann who kindly hosted me as a guest in his department. Thank you, Matthias, for all your support throughout my PhD studies! You inspired and motivated me on many occasions which truly shaped my scientific career.

My scientific journey in Munich started in Peter Becker's lab; Peter, your unwavering support and inspiration have been crucial to my growth.

Throughout my PhD studies I had amazing collaboration partners; thank you Micha Drukker, Kathrin Lang, Gunnar Schotta, David Suter, Till Bartke, Cristina Cardoso, Karsten Spiekermann, and Pierre-Antoine Defossez! Working alongside you and your teams has been an incredible experience.

The most scientifically stimulating moments during my PhD studies were certainly during our “evening” Müsli-Forschung sessions with Michael Bartoschek, Evangelia Ntouliou and Christopher Mulholland. I cherish our friendships; Chris, thank you for the endless scientific discussions. I cannot find any fitting superlatives to describe how I enjoyed these moments. Evi, you continue impressing me with your amazing kindness, intellect, and friendship – thank you for everything! Michi, I cannot thank you without sounding too awkward, but I am truly grateful to have you as my friend.

Sebastian Bultmann and Weihua Qin, I would like to thank you both for always supporting and motivating me. You both gave me a lot of orientation and inspiration, steering my course many times! Sebastian, it was a huge honor to be a guest of the Bultmann lab! Weihua, I also really enjoyed being your desk neighbor, thanks for always helping me out and the shared supervisions of students ;-) ! Deis Haxholli, thank you for your enduring friendship and support. I will never forget the legendary stem cell course :-) Jeannette Koch, Ruzica Barisic, Susanne Breitsameter and Andrea Mathieu thank you for your help with all the technical and organizational issues. Hans C. Mescheder, Irina Solovei, Clemens Steinek, David Hörl, Hartmann Harz, Andreas Stengl, Paul Stolz and the whole Leonhardt lab, thanks for being such amazing colleagues.

I am deeply grateful to all the support I received from the Mann lab. Especially thanks to Igor Paron, Tim Heymann and Bianca Splettstoesser for the highly competent technical support as well as Alison Dalfovo, Theresa Schneider, Medini Steger, Ute Sackers, and Mario Oroshi for providing great organizational help. Special thanks for all the support I received from Ankit Sinha, Maria Tanzer, Johannes Müller-Reif, Sophia Steigerwald and Marvin Thielert!

I also want to thank my amazing students, especially Matthias Mack, Anna Semenova, Alina Ivanova and Anja Michl. I am grateful for our time together and I wish you a bright future!

Thank you to all my friends who supported me during my PhD studies, in particular Leonie Witte and Matthias Münchhalfen; thank you both for helping me making the right decisions!

I want to thank my parents Serpil and Ahmet Ugur. You are amazing parents, true pioneers, and heroes, thank you for everything! I further want to thank my beautiful family: my grandparents, my brothers Ilker and Emin, my sister Özge, my niece Elâ as well as my whole family, thank you for everything!

Sümeyye Ugur, my deepest gratitude goes to you. Thank you for supporting me and accompanying me through the colorful streets of life. Lastly, as a dad of two, I cannot stop this acknowledgement without expressing my gratitude to be the father of two beautiful human beings. Dear Rahel, dear Malik, I wish you the best possible family, societal, political, and cultural environment that nurtures and strengthens your aspirations. I believe in you!