Elucidating cell-fate decision-making of mammalian pluripotent cells through single-cell transcriptomics

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

vorgelegt von Elmir Mahammadov
München, den 18.08.2022

Diese Dissertation wurde angefertigt
unter der Leitung von **Dr. Antonio Scialdone**
am Institut für Epigenetik und Stammzellen
des Helmholtz Zentrum Münchens

Erstgutachter: Prof. Dr. Maria Elena Torres-Padilla

Zweitgutachter: Prof. Dr. Wolfgang Enard

Tag der Einreichung: 18.08.2022

Tag der mündlichen Prüfung: 20.07.2023

Eidesstattliche Erklärung
Satutory declaration

Ich versichere hiermit an Eides statt, dass die vorgelegte Dissertation von mir selbständig und ohne unerlaubte Hilfe angefertigt ist.
I hereby declare on oath that the thesis submitted is my own work and that I have not sought or used inadmissible help of third parties to produce this work.

München, den 22.02,2024 .... Elmir Mahammadov ....
Munich,
(Unterschrift/signature)

Erklärung
Declaration

Hiermit erkläre ich, *
Hereby I declare

☒ dass die Dissertation nicht ganz oder in wesentlichen Teilen einer anderen Prüfungskommission vorgelegt worden ist.
that this work, complete or in parts, has not yet been submitted to another examination institution

☒ dass ich mich anderweitig einer Doktorprüfung ohne Erfolg **nicht** unterzogen habe.
that I did **not** undergo another doctoral examination without success

☐ dass ich mich mit Erfolg der Doktorprüfung im Hauptfach ....
that I successfully completed a doctoral examination in the main subject

und in den Nebenfächern ....
and in the minor subjects

bei der Fakultät für .... der ....
at the faculty of at
(Hochschule/University)

unterzogen habe.

☐ dass ich ohne Erfolg versucht habe, eine Dissertation einzureichen oder mich der Doktorprüfung zu unterziehen.
that I submitted a thesis or did undergo a doctoral examination without success

München, den 22.02.2024 .... Elmir Mahammadov ....
Munich,
(Unterschrift/signature)

*) Nichtzutreffendes streichen/delete where not applicable

Table of contents

1. LIST OF PUBLICATIONS 5

2. STATEMENT OF CONTRIBUTIONS 6

3. SUMMARY 10

4. AIMS OF THE THESIS 11

5. INTRODUCTION 12

5.1 UNDERSTANDING CELLULAR IDENTITY 14
5.1.1 An evolutionary perspective 14
5.1.2 Why can cells not be “rigid”? 14
5.1.3 Studying cellular identity – a brief history 15
5.2 SINGLE-CELL TRANSCRIPTOMICS 16
5.2.1 Utilization of single-cell transcriptomics 17
5.2.2 Computational analysis of scRNA-seq data 18
5.3 SCOPE OF THE THESIS 21

6. RESULTS 25

6.1 PUBLICATION 1 25
6.2 PUBLICATION 2 54
6.3 PUBLICATION 3 94
6.4 PUBLICATION 4 99
6.5 MANUSCRIPT 1 127

7. DISCUSSION 160

7.1 SINGLE-CELL TRANSCRIPTOMIC CHARACTERIZATION OF A HUMAN GASTRULA 160
7.2 EFFECT OF SEMAPHORIN 4A ON MOUSE HEMATOPOIETIC STEM CELL DORMANCY 163
ESTIMATION OF HSC PROLIFERATION THROUGH CELL CYCLE ANALYSIS 163
7.3 DETECTING HETEROPLASMY FROM SCRNA-SEQ OF MOUSE EPIBLAST CELLS 164
7.4 FUTURE OUTLOOK 164

8. REFERENCES 165

9. APPENDICES 171

9.1 COPYRIGHT STATEMENTS 171
9.2 CURRICULUM VITAE 172

10. ACKNOWLEDGMENTS 174

1. List of publications

- Tyser, R. C. V., **Mahammadov, E*.,** Nakanoh, S., Vallier, L., Scialdone, A., & Srinivas, S. (2021). Single-cell transcriptomic characterization of a gastrulating human embryo. *Nature*, *600*(7888), 285–289. https://doi.org/10.1038/s41586-021-04158-y

- Lima, A., Lubatti, G., Burgstaller, J., Hu, D., Green, A. P., Di Gregorio, A., Zawadzki, T., Pernaute, B., **Mahammadov, E.,** Perez-Montero, S., Dore, M., Sanchez, J. M., Bowling, S., Sancho, M., Kolbe, T., Karimi, M. M., Carling, D., Jones, N., Srinivas, S., … Rodriguez, T. A. (2021). Cell competition acts as a purifying selection to eliminate cells with mitochondrial defects during early mouse development. *Nature Metabolism*, *3*(8), 1091–1108. https://doi.org/10.1038/s42255-021-00422-7

- Lubatti, G., **Mahammadov, E.,** & Scialdone, A. (2022). MitoHEAR: an R package for the estimation and downstream statistical analysis of the mitochondrial DNA heteroplasmy calculated from single-cell datasets. *Journal of Open Source Software*, *7*(74), 4265. https://doi.org/10.21105/joss.04265

- Nakatani, T., Lin, J., Ji, F., Ettinger, A., Pontabry, J., Tokoro, M., Altamirano-Pacheco, L., Fiorentino, J., **Mahammadov, E.,** Hatano, Y., Van Rechem, C., Chakraborty, D., Ruiz-Morales, E. R., Arguello Pascualli, P. Y., Scialdone, A., Yamagata, K., Whetstine, J. R., Sadreyev, R. I., & Torres-Padilla, M.-E. (2022). DNA replication fork speed underlies cell fate changes and promotes reprogramming. *Nature Genetics*, *54*(3), 318–327. https://doi.org/10.1038/s41588-022-01023-0

*co-first author

List of unpublished manuscripts

- Toghani D., Zeng S., **Mahammadov E.,** Crosse E., Seyedhassantehrani N., Burns C., Gravano D., Radtke S., Kiem HP., Rodriguez S., Carlesso N., Pradeep N., Wilson N., Kinston S., Gottgens B., Nerlov C., Pietras E., Maes C., Mesnieres M., Kumanogoh A., Worzfeld T., Scadden D., Scialdone A., Spencer J., Silberstein L. (Unpublished manuscript). Myeloid-biased HSC require Semaphorin 4a from the bone marrow niche for self-renewal under stress and life-long persistence

2. Statement of contributions

Publication 1

I hereby state my contribution to the following publication:

Tyser, R.C.V*., **Mahammadov, E***., Nakanoh, S., Vallier, L., Scialdone, A.,
Srinivas, S. Single-cell transcriptomic characterization of a gastrulating
human embryo. *Nature* 600**,** 285–289 (2021).

mainly consisted of computational analysis of single-cell RNA-seq sequencing data from a human embryo in the gastrula stage, including cell type discovery, trajectory construction, and inter-sample comparison. Additionally, I created a web app for interactive data visualization (human-gastrula.net). Dr. Richard Tyser carried out the cell collection from the embryo, single-cell sequencing, and cell type annotation, as well as interpretation and illustration of the results.

Elmir Mahammadov

München, July 29th, 2022

Confirmation of contribution

We hereby confirm that the statement of contribution reproduced above is both truthful and accurate and represents a substantial enough contribution to warrant co-first authorship.

Prof. Dr. Maria-Elena Torres-Padilla

München, August 16th, 2022

Dr. Richard Tyser 29/07/2022

Publication 2

I hereby state my contribution to the following publication:

Lima, A., Lubatti, G., Burgstaller, J., Hu, D., Green, A. P., Di Gregorio, A., Zawadzki, T., Pernaute, B., **Mahammadov, E.,** Perez-Montero, S., Dore, M., Sanchez, J. M., Bowling, S., Sancho, M., Kolbe, T., Karimi, M. M., Carling, D., Jones, N., Srinivas, S., … Rodriguez, T. A. (2021). Cell competition acts as a purifying selection to eliminate cells with mitochondrial defects during early mouse development. *Nature Metabolism*, *3*(8), 1091–1108. https://doi.org/10.1038/s42255-021-00422-7

consisted of helping to develop an analysis pipeline to measure heteroplasmy levels in mouse epiblast cells.

Elmir Mahammadov

München, August 16th, 2022

Confirmation of contribution

I hereby confirm that the statement of contribution reproduced above is both truthful and accurate.

Prof. Dr. Maria-Elena Torres-Padilla
München, August 16th, 2022

Publication 3

I hereby state my contribution to the following publication:

> Lubatti, G., **Mahammadov, E.,** & Scialdone, A. (2022). MitoHEAR: an R package for the estimation and downstream statistical analysis of the mitochondrial DNA heteroplasmy calculated from single-cell datasets. *Journal of Open Source Software*, *7*(74), 4265. https://doi.org/10.21105/joss.04265

consisted of helping to develop the MitoHEAR analysis software and manuscript editing.

Elmir Mahammadov

München, August 16th, 2022

Confirmation of contribution

I hereby confirm that the statement of contribution reproduced above is both truthful and accurate.

Prof. Dr. Maria-Elena Torres-Padilla
München, August 16th, 2022

Publication 4

I hereby state my contribution to the following publication:

Nakatani, T., Lin, J., Ji, F., Ettinger, A., Pontabry, J., Tokoro, M., Altamirano-Pacheco, L., Fiorentino, J., **Mahammadov, E.,** Hatano, Y., Van Rechem, C., Chakraborty, D., Ruiz-Morales, E. R., Arguello Pascualli, P. Y., Scialdone, A., Yamagata, K., Whetstine, J. R., Sadreyev, R. I., & Torres-Padilla, M.-E. (2022). DNA replication fork speed underlies cell fate changes and promotes reprogramming. *Nature Genetics*, *54*(3), 318–327. https://doi.org/10.1038/s41588-022-01023-0

consisted of helping to develop a mathematical model to describe the state transition between pluripotency and totipotency and its relation to the cell cycle.

Elmir Mahammadov

München, August 16th , 2022

Confirmation of contribution

I hereby confirm that the statement of contribution reproduced above is both truthful and accurate.

Prof. Dr. Maria-Elena Torres-Padilla

München, August 16th, 2022

3. Summary

Understanding cellular identity, heterogeneity and differentiation in mammals is crucial for solving many long-standing questions regarding regenerative medicine, developmental biology, and evolution. Recent cutting-edge molecular profiling methods have been developed to explore cell identity in different biological systems. Particularly, investigating gene expression patterns in individual cells with single-cell transcriptomics has provided significant opportunities for understanding complex tissues. By using single-cell transcriptomics, it has been possible to extract a large amount of information from cells in various tissues, such as embryonic and cancer tissues. Having access to such an extensive molecular profile from single cells paves the way to understanding factors that shape cell identity in a data-driven manner. However, to achieve this aim, the development of new and tailored computational tools are required to extract biologically meaningful information. In this dissertation, I discuss how I have explored the molecular factors that contribute to regulating cellular fate decision in different types of mammalian pluripotent cells by analysing single-cell RNA-seq (scRNA-seq) data. Specifically, I show my contributions to understanding early human and mouse development, as well as hematopoiesis in adult mice. By using state-of-the-art as well as novel computational tools and algorithms, I contributed to the first-ever single-cell characterization of a human embryo in the gastrula stage. Afterward, I demonstrate my work on elucidating stem cell state transition in early mouse development in both *in vivo* and *in vitro* models. Finally, I present my contribution to investigating the effect of the Sema4a signaling molecule on regulating hematopoietic stem cells in adult mice by computationally comparing the scRNA-seq data from wild-type and mutant mice. Overall, the studies present in the thesis demonstrate the power of single-cell transcriptomics in characterizing cellular heterogeneity and its link with cell fate decision, as well as elucidating possible mechanisms of cell differentiation in different model systems and organisms.

4. Aims of the thesis

- **Understanding human gastrulation through single-cell RNA-seq**
    - Characterize the cell atlas of the human embryo in the gastrula stage by using clustering methods
    - Explore endoderm, mesoderm, and ectoderm germ layers through trajectory reconstruction and sub-clustering
    - Compare the human data to the mouse and non-human primate in equivalent stages
    - Test *in vitro* models that are used to study gastrulation
    - Provide a valuable resource for scientists in the field by sharing the data in an accessible way and creating an app to explore the data

- **Investigating cell state transition in early mouse development *in vivo* and *in vitro***
    - Measuring mitochondrial heteroplasmy levels of mouse epiblast cells from scRNA-seq data and checking their role in cellular competition, in the transition from a "winner" to a "loser" state
    - Developing a computational pipeline for the above-mentioned aim
    - Helping to develop a mathematical model to describe state transition between pluripotency and totipotency and its relation to cell cycle in mouse embryonic stem cells

- **Hematopoiesis in adult mice through the lens of single-cell RNA-seq**
    - Describe the effect of the Semaphorin 4a (Sema4a) secreted molecule on myeloid-biased hematopoietic stem cell (myHSC) self-renewal and dormancy
    - Compare samples from wild-type and *Sema4a* knockout experiments by inferring their cell cycle phases and trajectories
    - Investigate possible molecular pathways involved in myHSC dormancy

5. Introduction

If we imagine the history of all organisms on earth as the result of countless cell divisions, life can *a priori* be envisioned as a single-cell genealogy. This genealogy entails both unicellular and multicellular organisms to arise and thrive in various environments. The success of unicellular organisms during evolution, however, did not prevent multicellularity to make an appearance in multiple lineages (Parfrey & Lahr, 2013). The cells that comprise these new multicellular organisms had to become more diverse and specialized as the size, complexity, and needs of these organisms grew. This diversity has been the source of survival of these species and the rich phenotypic variation among them. It has helped them to solve many problems regarding their changing circumstances during evolution by efficiently sharing different functions in order to survive (Goldsby et al., 2012). This cellular diversity would not have been possible without a cell's ability to change its internal state and function if the conditions demand it. The process of cell state shifting towards a functionally more mature state is called *cellular differentiation*. Although there are many hypotheses regarding the origin of differentiation during evolution, it is agreed that this ability existed in the cells before multicellularity emerged (Sogabe et al., 2019). In multicellular organisms, cell differentiation acquired a new meaning, as their functions were tied to organismal level properties, such as reproduction and homeostasis. The efficient and creative use of resources across cells to solve such problems for the organism has caused cellular differentiation and heterogeneity to become crucial characteristics of multicellular organisms during evolution.

We can observe this phenomenon in a wide number of contexts in nature. During the embryonic development of an animal (embryogenesis), stem cells can self-renew, as well as give rise to new cell types in a space- and time-dependent manner. Cells exhibit a wide range of behaviors and characteristics, such as migration, morphological variation, apoptosis, and communication. These eventually lead to the formation of tissues and organs with specified functions (organogenesis).

Embryos are not the only place where one can observe cell differentiation. Adult organisms of certain species have the ability to replace lost tissues. In adult vertebrate organisms, the liver can regenerate itself in the event of mass loss, in order to maintain its crucial function for homeostasis (Michalopoulos 2009). Blood cells also need to be constantly renewed through hematopoiesis. Axolotl has served as the main model organism to study the regeneration capabilities of different body parts. They can regrow both inner organs, such as kidneys and heart, as well as outer extremities and limbs (Vieira et al., 2020). Exciting studies have also come out from other model organisms, such as *Xenopus laevis*. They can correct their craniofacial structures after deformation by using the ability of the cells to differentiate and migrate into their correct positions (Vandenberg et al., 2012). These kinds of studies can be pivotal for improving the effectiveness of stem cell therapy in medicine.

Many important questions arise regarding cell differentiation in different contexts when they are observed. How do the cells know when to stop differentiating? Why do they move to the correct position in the tissue? How is proper tissue size achieved? These are some of the most sought-out questions in biological research. Because of the complexity of the cellular differentiation process, researchers have investigated several aspects of it with various methods. For instance, many studies involve confirming or tracking the identities of the cells as they differentiate, which also allows researchers to perform perturbation studies to uncover the mechanism of differentiation. A lot of cell types exhibit well-defined morphological features. This can help to confirm or predict the trajectory of these cell types using microscopy. However, because cell differentiation can happen gradually, morphological changes might not be so obvious. In this case, a more resolved approach is necessary. One method has been to use lineage-specific markers and track its presence with, for example, fluorescent tags. Other approaches like DNA barcoding have also been used for lineage tracing (Pei et al., 2017). Despite their various degrees of success, these methods often suffer from the limitation of not being able to capture detailed cell identity, due to a lack of information on the underlying gene expression (VanHorn & Morris, 2021). As a cell goes through a state change, its molecular portrait gets altered. This is because of the cell's ever-changing and context-dependent needs as it differentiates and acquires new roles. In other words, the cell needs to use different resources (i.e. molecules) within it in new circumstances. Therefore, if we can observe the distribution of these molecules within each cell at a given moment, we can assign an identity to the cell. To tackle the question of pinpointing cell identity during differentiation, highly resolved techniques have been introduced. They are able to capture different molecular content of a single cell, such as DNA, RNA, and protein. For this thesis, I will focus on the transcriptomic (mRNA) characterization of single cells.

Having an access to the transcriptomic profile of a cell can especially be advantageous. The gene expression pattern of a cell can provide valuable and detailed insights into the cell's identity and state. Single-cell RNA-seq (scRNA-seq) has been developed and used for this purpose since its advent. Aside from the single-cell resolution it provides (unlike bulk RNA-seq), its power also lies in the unbiased quantification of all the mRNA molecules present in a cell.

Of course, it would not be feasible to exploit the potential of scRNA-seq without using appropriate computational techniques. As the amount, complexity, and resolution of the data have increased as a result of advancing technologies, developing and using dedicated tools and algorithms have become a necessity. Extracting and distinguishing relevant signals from scRNA-seq data with computational tools have become a crucial part of such studies since the beginning (Stegle et al., 2015).

This thesis is devoted to the utilization of the power of scRNA-seq to characterize cellular differentiation in mammalian stem cells and embryos. There are various computational techniques that have been developed to investigate stem cell differentiation including the cell type identification (Traag et al., 2019), novel marker gene discovery (Delaney et al., 2019), cell cycle assignment (Scialdone et al., 2015), and identification of differentiation trajectories (Saelens et al., 2019).

The systems studied for this thesis include human and mouse embryonic development, as well as hematopoietic stem cells of adult mice.

First, I analyzed single-cell transcriptomics data from a complete embryo in the gastrula stage, roughly in embryonic day 16-19, (Publication 1), to get the first-ever glimpse of this crucial stage of development in humans.

In the project on adult mice hematopoiesis, the focus was on a molecule (Sema4a) that prevents hematopoietic stem cells (HSCs) from differentiating (Manuscript 1). The aim was to understand the cellular heterogeneity within HSCs, particularly cells that are myeloid lineage biased (myHSCs) and the effect of the loss of Sema4a on their quiescence.

I also contributed to two projects concerning mouse embryonic development. The first project entailed an in vitro model of pluripotency to totipotency transition in mouse embryonic stem cells (Publication 4, (Nakatani et al., 2022)). In the second project, we investigated cellular competition and the transition between a "winner" and a "loser" state in epiblast cells of mouse embryos before gastrulation (Publication 2, (Lima et al., 2021)). For this purpose, a new R package was also developed (Publication 3, (Lubatti et al., 2022)).

5.1 Understanding cellular identity

5.1.1 An evolutionary perspective

In biology, it is widely agreed that the fundamental unit of life is a cell. It is also one of the fundamental units of selection during the evolution of all lifeforms. Evolutionary divergence has always been accompanied by the transformation of cellular features. This was necessary because cells needed to make use of their environments and adapt to them at the same time under various conditions. Some of the most ancient cells (unicellular organisms) were able to survive in the harshest conditions. Studies have shown that all organisms have emerged from these single-celled *extremophiles* that lived near hydrothermal vents without an oxygen (Weiss et al., 2016). They are considered the Last Universal Common Ancestor (LUCA), a concept that was first introduced by Charles Darwin in his book *On the Origin of Species*. After 3.5 billion years of evolution since LUCA, we can now observe tremendous diversity among lifeforms. A major source of this diversity can be traced back to the emergence of multicellular organisms.

Despite the ability of ancient unicellular organisms to live in such habitats for a long time, evolutionary selection has also resulted in multiple origins of the multicellularity (Knoll, 2011). This has meant that the cells together were now able to perform certain tasks which could not perform by themselves. For example, in rudimentary multicellular aggregates, groups of cells could have improved motility to move to more favorable environments, increased resistance to stress, or longer memory capacity (Tong et al., 2022). Performing these tasks requires all the single cells to contribute to the overall multicellular organism to increase their fitness. Over time, the complexity and magnitude of objectives that needed to be achieved by the primitive multicellular organisms have led to sharing different responsibilities across cells. Thus, different cells performing various functions have emerged as a result, also known as *cell types*. Existence of somatic and germ cells in metazoan can be used as a simple example of such division of labor. Indeed, the complexity of an organism has been attributed to the number of cell types or cellular diversity that it possesses (Valentine, 2003), which would give the organism greater flexibility as it navigates through a hostile world.

5.1.2 Why can cells not be “rigid”?

For the organism to reach certain objectives in the face of changing circumstances, the cells themselves need to exhibit flexibility or plasticity. Understanding this phenomenon is not only crucial to appreciate the phenotypic diversity among species, but also the changes in the cell types that they carry with them throughout their life cycle. This life cycle starts with many cell divisions that end up forming the adult body. While dividing, the cells also differentiate and acquire specific characteristics. The cells that possess this type of ability are called *stem cells.* They are also able to maintain their existence by proliferating if needed (self-renewal). In many multicellular organisms during their development, these stem cells act as a reservoir to supply the right number of cells that differentiate into tissues and organs. They have varying capabilities of giving rise to different cell types and states. They can generally be classified as totipotent, pluripotent, and multipotent stem cells. While totipotent cells can give rise to both embryonic and extra-embryonic tissues, pluripotent stem cells can generate only embryonic tissues that give rise to the adult body. Multipotent cells are also able to generate more than one lineage of cells but are more restricted than pluripotent cells. For example, a zygote can be considered totipotent and the epiblast cells can be considered pluripotent. Blood stem cells can be given as an example of multipotent stem cells. As the cells differentiate during embryogenesis, their capability to produce cell types gradually decreases.

After the adult body formation, however, not all the cells lose plasticity. There are many instances of cell plasticity in adult organisms as well. For example, because vertebrates require a constant supply of different types of blood cells for their organs, a reservoir of stem cells needs to be present. These cells are called hematopoietic stem cells (HSC), and they give rise to various blood cell types with specific functions throughout an organism's life cycle. Some organisms can regenerate other types of tissues as well. Starfish are amazingly good at regenerating large chunks of their bodies. Not only they can replace a lost arm from the central disk, but they can also regenerate the entire body from just a part of an arm as well. This is accomplished by a group of cells that are capable of proliferating and differentiating to generate the lost body parts (Carnevali, 2006). Other members of phylum Echinodermata, as well as the members of phyla Cnidaria and Annelida can perform similar types of regeneration (Zattara et al., 2019).

Not all cases of cell plasticity and differentiation are considered to have beneficial outcomes. When certain cells in adult tissues receive carcinogenic signals, they can change their identity to acquire the ability to give rise to bigger and more invasive tissues. In other words, they act like the stem cells in an embryo to supply the tumor with heterogeneous cell populations (Takahashi & Yamanaka, 2006). They can achieve this by the process of changing their epithelial state towards mesenchymal one, also called EMT (epithelial-mesenchymal transition), as they can migrate and form metastases.

This type of state transition where cells acquire migratory characteristics is one of many ways for cells to achieve objectives related to their differentiation. EMT is also an essential process that many cells go through during embryogenesis. For such a transition to happen, there must be a complex and robust interplay between cell-intrinsic and extrinsic factors. Epigenetic modifications, gene regulatory networks, and intracellular signaling can be shown as factors that participate in cell differentiation. We also have to consider the context in which a cell undergoes these changes. During embryogenesis, the timing of the internal cellular events has to be reckoned with. For example, the pluripotency marker OCT4/POU5F1 exhibits differences across stages in its splicing isoform expression during pre-implantation embryo development (Cauffman et al., 2006). Furthermore, the question of "when" an event happens has to be accompanied by "where". A cell's spatial context also influences its identity. So, the internal molecular events are also influenced by the cell's location relative to other cells. For example, in an early developmental stage of mice (16-cell stage), the relative position of cells to each other can influence their identities in the following stage (Lorthongpanich et al., 2012).

5.1.3 Studying cellular identity – a brief history

The phenomenon of self-repair in animals has captured the attention of philosophers and scientists for many centuries. The first published study dates back to the 18th century by a French naturalist, who investigated the regeneration of crayfish claws (Réaumur, 1712). A few decades later, regeneration in freshwater polyps' arms was detected (Trembley, 1744). The field has immensely evolved since, allowing a new potential for its application in medicine (Ntege et al., 2020). However, the mechanisms and factors underlying this fascinating process are not completely clear. To study regeneration and cellular plasticity on a cellular and molecular level, researchers have employed various methods.

Traditionally, a cell's location in the body was used as an indication of its identity, such as a brain cell or muscle cell. Additionally, microscopy has been used to distinguish cells based on their morphological features. Some early works were conducted by pioneers like Golgi, who used microscopy and dye-staining to elegantly visualize neurons (Golgi, 1883). Conklin used these new techniques to study cell differentiation and construct a lineage tree during ascidian embryogenesis (Conklin, 1905). Later in the 20th century, developments in microscopy and molecular biology allowed the detection of specific molecules in a cell through immunofluorescence. This could be for example proteins (Coons et al., 1941), or nucleic acids (Pardue & Gall, 1969). All of these culminated in the complete fate mapping of *C.elegans* cells, the first such study was done in an animal (Sulston et al., 1983).

Fate mapping studies eventually moved from using cells to their genetic material, since using a dye to trace cells had limitations in the dye diffusion to neighboring cells after a certain number of cell divisions (Kretzschmar & Watt, 2012). Because the progeny cells need to inherit information from their parent cells to be tracked, using the gene coding for the green fluorescent protein (GFP) has been an advantageous approach. However, they need to be inserted into the cell through various methods, such as lentiviral transduction. Using specific markers, it's possible to detect cardiac or neural lineages both *in vivo* and *in vitro* (Nguyen et al., 2010). Recently, tracing the cells using DNA scarring methods has been exploited. They usually involved CRISPR/Cas9-based system to edit the genome and trace the introduced scars in progeny cells. The first application of this approach was GESTALT, which was successfully used to illustrate lineage relationships between the cells during the zebrafish development (McKenna et al., 2016).

Aside from experimental limitations, the methods described above also suffer from either introducing biases that are driven by prior knowledge or the inability to capture the cell identity in the right context (VanHorn & Morris, 2021). Thus, a more comprehensive and unbiased approach needs to be integrated into determining the cell identity and its relationship to the other cells.

5.2 Single-cell transcriptomics

One of the technologies that have especially influenced all areas of biological research is RNA-seq. Being able to measure global RNA quantity in a given biological sample has opened doors to many intriguing findings. For example, this has meant that scientists have been able to find transcriptomic differences across samples using differential gene expression (DGE) analysis. DGE analysis has been the main goal of RNA-seq ever since its successful application to various organisms and systems in its early stages (Stark et al., 2019). Numerous technologies and sequencing platforms have been committed to making the RNA measurements more and more precise to achieve a better sensitivity in DGE analysis.

Like with any technology in biological research, scientists have, in many cases, run into limitations with traditional bulk RNA-seq after employing it to make critical observations in their research. Because any sample had to be sequenced as a whole to measure its average global transcriptome across a population of cells, tissue and cellular heterogeneity within the sample were overlooked. This, for example, was an important limitation in the study of cancer tissues, which, in some cases, are affected only by an aberrant rare cell population (Gyanchandani et al., 2017). Studying the importance of such rare populations is not possible with the bulk RNA-seq methods. In general, it becomes difficult to define what should be selected as a homogenous sample for an RNA-seq experiment. This is especially important in developmental biology, which studies how cellular heterogeneity arises in tissues and organisms.

These limitations have been overcome by single-cell RNA-seq (scRNA-seq), for which the first protocol was published in 2009 (Tang et al., 2009). Thanks to the advent of scRNA-seq, it has been possible to quantify the amount of RNA of individual cells in a sample. Like traditional bulk RNA-seq, scRNA-seq changed the course of research in many areas of biology and medicine. Unsurprisingly, this new technology has revolutionized developmental and stem cell biology as well. It has revealed cellular heterogeneity within certain cell types that were previously considered to be homogeneous. Because the transcriptional state of a cell can define its identity, developmental biologists have been able to take a glimpse at how cell identity changes during various stages of development in different model organisms, as well as *in vitro* systems. Single-cell atlases of mice (Pijuan-Sala et al., 2019), zebrafish (Wagner et al., 2018), *Xenopus* (Briggs et al., 2018), and macaque (Ma et al., 2019) embryos have revealed many insightful findings on the nature of cellular heterogeneity in the early developmental stages of these organisms. However, these types of analyses have required the use of tailored algorithms and software tools to extract relevant information, as briefly described below.

5.2.1 Utilization of single-cell transcriptomics

Since the first study to generate the transcriptomic profile of a few cells through scRNA-seq (Tang et al., 2009), the capacity to sequence more cells has grown immensely. As the complexity and depth of the datasets obtained through scRNA-seq increased, its implementation in many different contexts emerged. In addition to the possible applications of bulk RNA-seq, having the transcriptomic information at a single-cell level paves the way for advancing knowledge in many areas of biological and medical research.

Although other modalities of single-cell data, such as chromatin accessibility, can give useful information about cell state, having access to a cell's transcriptomics provides a more convenient interpretation. This convenience mainly arises from the fact that features measured in transcriptomics methods correspond to the expression of well-annotated transcripts and genes, whereas signals obtained from epigenomics methods might correspond to single or multiple genomic regions. (Lähnemann et al., 2020). However, there have recently been advances in merging these two technologies to obtain better mechanistic insights into the cell function (Yao et al., 2021). The changes in cell identity will be to a large degree reflected in its transcriptome. Whether a cell receives an environmental stimulus (e.g signal from other cells) or gains a new spatiotemporal context, the underlying gene expression pattern will be adjusted accordingly. Hence observing the gene expression patterns of cells can help reveal their identities both independently and in relation to each other.

Below, I discuss some of the biological questions that scRNA-seq can help answer. Some of them are also illustrated in Figure 1.

Tissue heterogeneity

scRNA-seq is particularly suitable for studying tissues with a high degree of heterogeneity. For example, cancer tissues have various cell types, and the different subpopulations can be identified through their transcriptomic differences. This is also true for an animal embryo, which at any given time contains various cell types that change over the course of development. Thus, scRNA-seq is an excellent method to uncover the cellular heterogeneity in developing tissues, as well as discover novel and rare cell types.

Novel marker genes

Annotating cell states in data obtained through scRNA-seq usually requires examining the expression of known marker genes. However, once a cell type is assigned to the cells through this technique, it is also possible to find novel genes that define the cell population of interest. This can facilitate cell type identification for future studies and could give info on previously unknown functions that the cells might perform.

Cell-type specific comparative analysis

After establishing the various cell types within a tissue, one can compare cell type composition between tissues. This can be, for instance, comparing tissues from healthy and unhealthy patients to see which specific cell type is affected and how. Additionally, other types of cross-sample analyses are possible, such as cross-species comparisons.

Cell cycle phase inference

Part of the cellular heterogeneity comes from the cells that are in different phases of the cell cycle. While some cells are dividing or proliferating, others are in a dormant state. Identifying the cell cycle phase can help interpret the differences between cellular states. In certain cases, the transcriptional differences due to the cell cycle can compound the identification of signals related to other processes, like the cellular differentiation (Buettner et al., 2015). Thus, correctly assigning cells to their cell cycle phase can provide important biological insights and remove a possible confounding factor during the analysis.

Trajectory inference

As mentioned above, a cell's transcriptomic profile gets altered when it goes through a state change. In a given tissue where cells differentiate towards different fates, single-cell transcriptomics can capture cells in different stages of differentiation. Taking advantage of these transcriptomic differences in the sample, it is possible to predict the differentiation trajectory of the cells within a sample. This can reveal interesting trajectory shapes, such as bifurcating or cyclic. In a sample where stem cells give rise to various tissues, one can observe a tree-like trajectory. Aside from finding the overall course of differentiation, one can also extrapolate the expression trend of individual genes along a given trajectory.

5.2.2 Computational analysis of scRNA-seq data

Growing aims and ambitions of molecular biology have necessitated the utilization of knowledge and techniques from other scientific fields. Hence integrating the expertise of scientists with different backgrounds has become a quintessential part of molecular biology studies. It is possible to observe this integrative approach in recent years, as technologies have vastly improved. Although *The Human Genome Project* in 2003 has not immediately brought its promises of solving many standing problems in biology, it has paved a way for scientists to discover molecular components that play a role in many biological systems. Thanks to the advent of sequencing technologies, especially Next-Generation Sequencing (NGS), many omics fields have emerged, producing a breadth of crucial data to help understand complex molecular processes underlying disease progression, cellular function, and drug response. Hence another wave of the vast use of computers has been brought to biology, after the success of computational analysis of protein structure half a century before (Gauthier et al., 2019).

Producing data through NGS technologies has brought many challenges with it. Many analyses, such as variant calling could be likened to “finding a needle in the haystack”. Before scRNA-seq was invented, computational biologists dealt with many challenges as well, such as noise and data quality. Already many tools were made to process and analyze bulk RNA-seq. So, the scientists that started dealing with scRNA-seq data had tools at their disposal. However, scRNA-seq data have additional intrinsic challenges, such as dropouts, the number of cells analyzed, and various biological confounders. This has demanded the creation of novel statistical and computational tools.

Today, there are already hundreds of scRNA-seq tools existing to process and extract information for various biological questions (Zappia et al., 2018). I will discuss relevant computational methods to the uses of scRNA-seq mentioned above and their limitations.

Clustering and cell-type identification

Grouping cells based on their transcriptomic profile is usually a quintessential step of scRNA-seq workflows (Trapnell, 2015). This is needed to establish the level of heterogeneity in the dataset so that the downstream analysis can be performed accordingly. In machine learning terms, clustering is an unsupervised learning method to divide data points into groups or clusters based on their similarity. For scRNA-seq data, this means that cells with similar transcriptomic profiles should fall into the same clusters. This is achieved without having prior knowledge about cell types. Many tools have been adopted or implemented for this purpose.

Some of the most popular clustering algorithms are k-means clustering, hierarchical clustering, and graph-based community detection. Hierarchical clustering and k-means have scalability problem, having issues handling large datasets (Kiselev et al., 2019). Considering the size of scRNA-seq data, $O(n^2)$ time complexity for these two methods might not always be suitable depending on the number of cells being analyzed.

Clustering tools from popular libraries for single-cell RNA-seq data analysis like scanpy (Wolf et al., 2018) and Seurat (Stuart et al., 2019) use graph-based methods. The most popular one is the Louvain algorithm (Blondel et al., 2008). It can overcome the scalability issue of k-means and hierarchical clustering by relying on a procedure called *modularity optimization* (with time complexity of $O(n*\log(n))$). Modularity is a measure of network structure, so it is possible to detect communities with different structures and sizes within a graph by optimizing the modularity value. However, it has been shown that modularity optimization suffers from *resolution limit*, resulting in communities that might include sub-networks within (Fortunato and Barthelemy, 2007). For scRNA-seq data, it might be a problem when some small populations cannot be detected. Leiden algorithm was shown to be an improvement, by solving the resolution limit inherent to Louvain (Traag et al., 2019). For this reason, Leiden was chosen as a primary way of detecting cell populations for all the studies in this thesis, whenever clustering was performed.

Overall, all methods rely on one or more parameters that define the number of clusters detected in the data. For k-means, the parameter k represents the number of clusters that the algorithm will identify. The graph-based methods rely on the resolution parameter to tune the number of clusters. Additionally, the graph is constructed through the k-nearest neighbor method, where the k parameter needs to be set. Higher values of k will result in fewer clusters. As there is no straightforward way of choosing these parameter values, the outcomes can be very distinct from each other. This issue regarding parameter choices makes mapping clusters to biologically meaningful cell populations a challenging task. In this case, some prior knowledge of the cell types and their markers might be necessary. Although judgment from the researcher is ultimately necessary to interpret clustering results, there are ways to computationally predict the correct number of clusters. One way is the bootstrapping technique, whereby the same clustering algorithm is run on multiple subsets of data and the consistency of the clustering results is assessed with different parameter values. As the scRNA-seq databases (also known as "transcriptional atlases") are populated more with various samples and studies, the annotation of cell clusters can also make use of these databases.

Cross-sample comparison

Often in the scRNA-seq analysis workflow, more than one sample will be involved. The task could be merging different datasets, mapping one onto another, or doing comparisons between conditions. Because these data typically come from separate sources or batches, correcting this difference is ultimately a computational problem. In any data integration task, there might be more than one possible source of difference. Some of them are technical and others biological. These sources can be different laboratories, sequencing platforms, or species. Therefore, it is computationally necessary to remove these effects while preserving the biological variation within and between the datasets.

To date, there are dozens of methods available to integrate scRNA-seq data. Some of the popular methods are MNN (Haghverdi et al., 2018), scanorama (Hie et al., 2019), Harmony (Korsunsky et al., 2019), and ComBat (Büttner et al., 2019). There are fundamental differences in terms of their performance, output, and data handling. Additionally, they do not perform well on various criteria of batch removal and conservation of biological signals (Luecken et al., 2022). Hence, there is no single method that will perform well for all datasets and integration tasks. A method can remove a batch effect well but might also remove interesting biological differences between datasets. For example, according to Luecken *et al.,* Harmony performs well when it comes to batch correction but might perform poorly when it comes to conservation of biological signals. There are also differences in speed and scalability to consider. For example, MNN is less scalable compared to scVI (Lopez et al., 2018). Overall, the methods recommended by the authors were Scanorama, scVI and scGen (Lotfollahi et al., 2019).

One of the biggest goals of data integration is to incorporate existing transcriptional atlases to help with annotation. One such atlas is the Human Cell Atlas (Regev et al., 2017). However, it becomes computationally intensive to make use of an evergrowing atlas. The methods mentioned above are not suitable to overcome this task. Transfer learning-based methods can be employed to transfer relevant knowledge from the atlas to the query data under study. They are able to perform the transfer without losing important biological variation between all relevant datasets (Lotfollahi et al., 2022). However, the main challenge of these models is the lack of useable output (corrected count matrix) for downstream analysis.

Trajectory inference

As many biological processes involve dynamic transcriptomic change during cellular state transition, it is possible to capture this computationally in scRNA-seq data. This allows putting cells in pseudotemporal order since tracking single cells in a laboratory setting might not be straightforward. Like other computational problems in scRNA-seq, many methods have been developed for trajectory inference analysis. The main challenge is to estimate the topology of the trajectory, whether it is linear, bifurcating, cyclic, etc. However, the choice of the method usually involves prior information on the type of trajectory, since not all methods can estimate all types of topology (Saelens et al., 2019). Depending on the method chosen, defining a progenitor cell might be necessary. However, if a user wants to perform a less biased approach, it is possible to add additional information such as RNA velocity (La Manno et al., 2018). This method estimates the trajectory of a cell using the information on spliced and unspliced mRNA counts. There are also potential issues regarding RNA velocity, such as a lack of generalization for all types of biological processes and systems (Bergen et al., 2021).

After inferring a biologically reasonable trajectory, the typical downstream analysis involves estimating the gene expression trend. Because many genes gradually get either downregulated or upregulated in the cells that are changing states, detecting them can give crucial insights into the dynamics of cell differentiation. The comparison analysis mentioned above can be extended here too. For example, it is possible to compare the trajectories and gene expression trends between species, as has been done in some studies before (Kanton et al., 2019). The study in Publication 1 (Tyser et al., 2021) makes use of such comparison too, detecting genes that share similar and different trends between human and mouse gastrulation.

Differential gene expression and marker gene detection

The main objective of bulk RNA-seq was to find differential expressed genes between conditions. This is of course an important objective of scRNA-seq analysis as well, often used to find DEGs between computationally and biologically defined cell types and states. Differential expression analysis is also a crucial part of cluster annotation. Although there are differences in technical considerations of bulk and single-cell RNA-seq, methods developed to detect DEGs for bulk RNA-seq perform very well for scRNA-seq data (Soneson & Robinson, 2018; Luecken & Theis, 2019). However, there have also been tools designed specifically for scRNA-seq, including COMET (Delaney et al., 2019), scDD (Korthauer et al., 2016) , and MAST (Finak et al., 2015).

Due to inherent technical noise, intercellular variation, as well as bimodality of gene expression in scRNA-seq data, methods for DEGs detection devised for bulk RNA-seq might not always be suitable. It was shown that these approaches might overpredict the number of genes differentially expressed between cell populations in scRNA-seq data, leading to false positives (Finak et al., 2015). However, tools designed specifically for scRNA-seq have also possible limitations, such as the lack of consistency across single-cell experimental methods, and the inability to take dropouts and zero-inflation into account (Das et al., 2021). It is also very common to use well-known statistical tests such as the Wilcoxon-rank-sum test, t-test, and logistic regression for DE testing. However, they are unable to differentiate between biological and technical variation, leading to false positive results (Squair et al., 2021). For these reasons, it is important to be careful when inspecting top marker genes for cell type annotation. Additionally, defining top genes based on different metrics, such as false discovery rate (FDR) and log-fold change might be necessary.

Top marker genes obtained through the algorithms above can be used to annotate a cluster that represents a biologically meaningful cell population. For the human gastrula study (Publication 1), using Wilcoxon-rank-sum test was sufficient to annotate the populations. This was possible thanks to the literature on marker genes available from mouse and *in vitro* studies. Because only one embryo was analyzed and the data included only one batch, using a method that takes batch effect into account was not necessary. This was not the case when investigating mouse hematopoietic stem cells (Manuscript I), where data from two conditions were compared using DEseq2 and the batches had to be considered by specifying a design matrix (Love et al., 2014).

5.3 Scope of the thesis

As we have seen, single-cell transcriptomics studies can be a powerful method for studying cellular identity and heterogeneity. There are also many tools and resources to make use of the data and gain insightful knowledge on complex biological processes involving a dynamic change in cellular state. The projects described in this thesis used scRNA-seq to study cell differentiation across different systems and organisms.

Although the systems studied have unique characteristics, several general questions can be asked about the systems studied. At what point in time and space do the cells start changing their identities? What genes are involved during this process? How is the cell cycle related to cellular differentiation? Are there any distinct or intermediate cellular states that were not characterized before? Below, I briefly introduce the biological systems that I have worked on.

Human Gastrulation

Gastrulation is a fundamental process during development that is conserved across all animals. It is one of the earliest stages of development during which the body plan and the germ layers are formed. These germ layers eventually give rise to all the tissues and organs in an adult body. Therefore, it is an extremely important stage of animal development that starts with massive expansion in cellular heterogeneity. Gastrulation has been studied in model organisms, such as mouse (Pijuan-Sala et al., 2019) and zebrafish (Wagner et al., 2018). To get insights into human gastrulation, some *in vitro* studies have been done (Moris et al., 2020). However, there has never been a deep molecular look into human gastrulation *in utero*. In my work (Publication 1, (Tyser et al., 2021)), by analysing a single-cell RNA-seq dataset from a human embryo at the gastrula stage, many useful insights could be gained. In particular, the focus was on characterizing the cellular heterogeneity and differentiation paths. Additionally, comparisons were made with mouse and macaque embryos, as well as with *in vitro* studies (Messmer et al., 2019) published before. These findings will be a resource of paramount importance for investigations of human gastrulation. I performed all the computational analyses and data visualization in this project, in addition to creating a web app for interactive data exploration (http://human-gastrula.net).

State transition of pluripotent cells during early mouse development

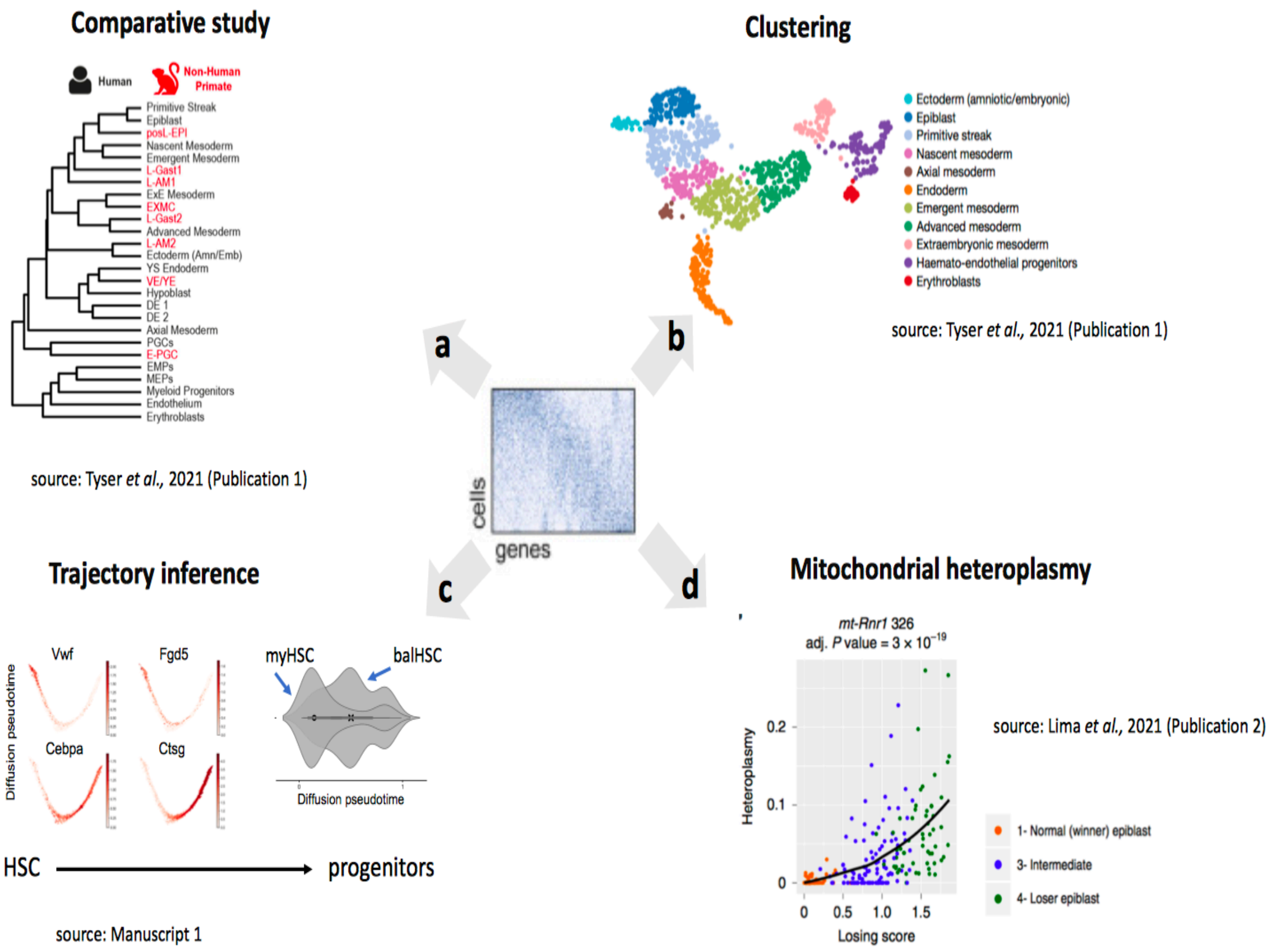
Mice have been powerful model organisms for studying complex processes that govern mammalian development. Many perturbation studies on mouse development can now be accompanied by computational analysis of scRNA-seq data. For example, we used scRNA-seq to investigate cell competition in mouse epiblast cells in pre-gastrulating human embryos (Publication 2, Lima et al., 2021). In this paper, we showed that the cells that are eliminated by competition possess defects in their mitochondrial DNA, which we identified from scRNA-seq data. For this project, I contributed to the development of a novel computational pipeline to extract mitochondrial heteroplasmy information from single-cell RNA-seq datasets (Publication 3, Lubatti et al., 2022).

I also contributed to a study investigating cell state transition in mouse embryonic stem cells (mESCs). In particular, the relationship between cell cycle and state transition was studied with a mathematical model in publication 4 (Nakatani et al., 2022). This mathematical model was used to estimate the transition probabilities of mESCs from a pluripotent to a totipotent state in each cell cycle stage.

Mouse hematopoiesis

Hematopoiesis is another dynamic system where applying single-cell transcriptomics can be extremely useful (Watcham et al., 2019). In a system where constant turnover of various cell types can be observed, grouping them based on their gene expression profiles can quantify the level of transcriptional heterogeneity of haematopoietic stem cells (HSCs) and how such heterogeneity can influence cell fate decision (Sanjuan-Pla et al., 2013). For example, some stem cells are more biased towards the myeloid lineage and their dormancy must be ensured to prevent fast exhaustion. This is important because they only need to proliferate in case of stress. There are extracellular and intracellular signaling cascades that regulate the balance between dormancy and proliferation. The study included in Manuscript 1 investigates the role of one of such signaling molecules that regulate myeloid-biased HSC proliferation, Semaphorin 4a (Sema4a). Using single-cell measurements from myeloid biased HSC (myHSC) in both WT and Sema4a KO cells, differences were explored to discover the role of Sema4a in maintaining myHSC dormancy. My contribution consisted in comparing single-cell RNA-seq data generated from

WT and Sema4a KO HSC to uncover which transcriptional differences arise in the mutant cells and what signaling pathways might act downstream of Sema4a.


Comparative study

source: Tyser *et al.*, 2021 (Publication 1)

Clustering

source: Tyser *et al.*, 2021 (Publication 1)

Trajectory inference

source: Manuscript 1

Mitochondrial heteroplasmy

source: Lima *et al.*, 2021 (Publication 2)

Figure 1. Types of analysis performed in this thesis from scRNA-seq data. a, Comparison between human CS7 gastrula and non-human primate (E16) clusters. **b,** Identification of clusters in human gastrula. **c,** Trajectory comparison of myeloid-biased HSCs and balanced HSCs **d,** Heteroplasmy levels of an mtDNA location located in gene mt-Rnr1 in mouse epiblast cells.

6. Results

6.1 Publication 1

Single-cell transcriptomic characterization of a gastrulating human embryo

https://doi.org/10.1038/s41586-021-04158-y

Received: 28 July 2020

Accepted: 5 October 2021

Published online: 17 November 2021

Check for updates

Richard C. V. Tyser[1,6], Elmir Mahammadov[2,3,4,6], Shota Nakanoh[5], Ludovic Vallier[5], Antonio Scialdone[2,3,4,7] ✉ & Shankar Srinivas[1,7] ✉

Gastrulation is the fundamental process in all multicellular animals through which the basic body plan is first laid down[1–4]. It is pivotal in generating cellular diversity coordinated with spatial patterning. In humans, gastrulation occurs in the third week after fertilization. Our understanding of this process in humans is relatively limited and based primarily on historical specimens[5–8], experimental models[9–12] or, more recently, in vitro cultured samples[13–16]. Here we characterize in a spatially resolved manner the single-cell transcriptional profile of an entire gastrulating human embryo, staged to be between 16 and 19 days after fertilization. We use these data to analyse the cell types present and to make comparisons with other model systems. In addition to pluripotent epiblast, we identified primordial germ cells, red blood cells and various mesodermal and endodermal cell types. This dataset offers a unique glimpse into a central but inaccessible stage of our development. This characterization provides new context for interpreting experiments in other model systems and represents a valuable resource for guiding directed differentiation of human cells in vitro.

Human gastrulation starts approximately 14 days after fertilization and continues for slightly over a week. Donations of human fetal material at these early stages are rare, making it nearly impossible to study directly. Our understanding of human gastrulation is therefore based almost entirely on extrapolation from model systems, historical collections of fixed samples[5–8] and more recently, several in vitro models. These include human embryonic stem (ES) cells cultured on circular micropatterns[9], human ES cell colonies engrafted into chick embryos[10] or 3D cellular models derived from human ES cells[11,12]. The stages just preceding gastrulation have also been studied using human embryos cultured in vitro[13–16]. There is currently no transcriptional data of in utero human gastrulation with which to compare such in vitro models. Here we present a morphological and spatially resolved single-cell transcriptomic characterisation of a single human gastrulating embryo at Carnegie stage (CS) 7, equivalent to 16–19 days post-fertilization, providing a detailed description of cell types present at this fundamental stage of human embryonic development.

Characterization of a CS7 human gastrula

We obtained a gastrulation stage human embryo through the Human Developmental Biology Resource, from a donor who provided informed consent for the use in research of embryonic material arising from the termination of her pregnancy. The embryo was karyotypically normal, male and staged as gestational week 4 plus 5 days, which corresponds to between 2 and 3 post-conception weeks (pcw).

The sample was completely intact and morphologically normal, comprising an embryonic disc with amniotic cavity, connecting stalk and yolk sac with pigmented cells (Fig.1a). We micro-dissected away the yolk sac and connecting stalk to isolate the embryonic disk with overlying amnion. Dorsal and ventral views of the disk showed the primitive streak extending approximately half the diameter of the disk along the long, rostral–caudal axis (Fig.1b, Extended Data Fig. 1a). The primitive node was visible at the rostral end of the streak. The length of the primitive streak relative to the embryonic disk, the presence of prechordal plate and the node at the middle of the disk enabled us to stage the embryo[17] as CS7. To retain anatomical information when disaggregating cells for the single-cell RNA sequencing (scRNA-seq), we sub-dissected the embryo into the yolk sac, rostral embryonic disk and caudal embryonic disk (Fig. 1d, Extended Data Fig. 1b).

After stringent quality filtering, we generated a library of 1,195 single cells (665 caudal, 340 rostral and 190 yolk sac cells), with a median of 4,000 genes detected per cell (Extended Data Fig. 1c). All cells showed expression of Y-chromosome genes and *XIST* transcript was largely undetectable (Extended Data Fig. 1d), confirming that there was no maternal cell contamination. All cell cycle stages could be detected, suggesting that normal cell cycling was occurring (Extended Data Fig. 1e). The genomic integrity of the sample was normal, with the number of indels identified falling in the same range as other human transcriptomic datasets (Extended Data Fig. 1f). These analyses, alongside the karyotyping (Methods) and morphology of the sample (Fig. 1a, b), suggest that this sample is representative of normal human gastrulation.

---

[1]Department of Physiology, Anatomy and Genetics, South Parks Road, University of Oxford, Oxford, UK. [2]Institute of Epigenetics and Stem Cells, Helmholtz Zentrum München–German Research Center for Environmental Health, Munich, Germany. [3]Institute of Functional Epigenetics, Helmholtz Zentrum München–German Research Center for Environmental Health, Neuherberg, Germany. [4]Institute of Computational Biology, Helmholtz Zentrum München–German Research Center for Environmental Health, Neuherberg, Germany. [5]Wellcome–MRC Cambridge Stem Cell Institute, Jeffrey Cheah Biomedical Centre, Cambridge, UK. [6]These authors contributed equally: Richard C. V. Tyser, Elmir Mahammadov. [7]These authors jointly supervised this work: Antonio Scialdone, Shankar Srinivas. ✉e-mail: antonio.scialdone@helmholtz-muenchen.de; shankar.srinivas@dpag.ox.ac.uk

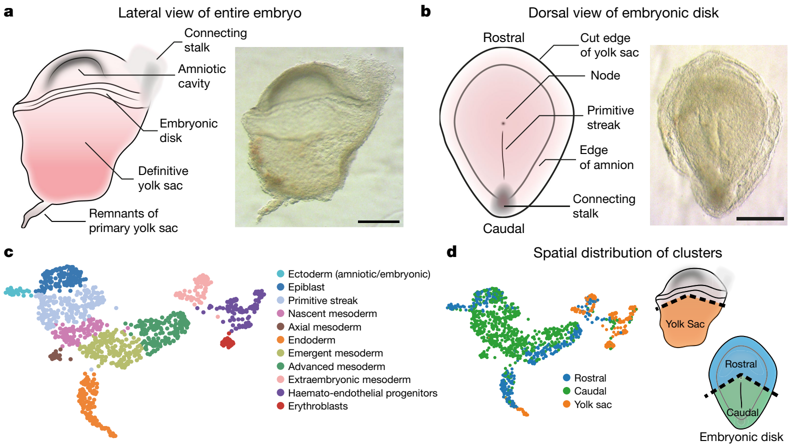


Fig. 1 | Morphological and transcriptional characterization of a CS7 human gastrula. a, Lateral view of the intact CS7 human embryo. Scale bar, 500 μm. **b**, Dorsal view of the dissected embryonic disk showing the primitive streak and node. Scale bar, 500 μm. **c**, Uniform manifold approximation and projection (UMAP) of all the cells computed from genes with highly variable expression. **d**, UMAP and schematics highlighting the anatomical region that cells were collected from (see also Extended Data Fig. 1b).

We detected 11 different cell populations with unsupervised clustering (Fig. 1c). Using a combination of anatomical location and marker genes (Supplementary Note 1), we annotated them as: epiblast, ectoderm (amniotic/embryonic), primitive streak, nascent mesoderm, axial mesoderm, emergent mesoderm, advanced mesoderm, extraembryonic mesoderm, endoderm, haemato-endothelial progenitors (HEP) and erythroblasts (Fig. 1c, Extended Data Fig. 2a, b, Supplementary Tables 1, 2). This annotation was supported by comparison with cell types described in the mouse[18] and the non-human primate cynomolgus macaque[19] (Extended Data Fig. 2c, d). The Smart-seq2 protocol also enabled us to differentiate between transcript isoforms and detect the cluster-specific expression of gene isoforms (Extended Data Fig. 2e, Supplementary Table 3).

We have created a web interface to interactively explore these data as a user-friendly community resource, accessible at http://www.human-gastrula.net.

Cell-type diversification

The identification of the CS7 epiblast cluster offered the opportunity to transcriptionally define the human primed pluripotent state as it exists in utero. To generate anchors of the in vivo primed and naive states, we first combined our epiblast data with existing pre-implantation human embryo scRNA-seq data[20] that captures the in vivo naive state. Cells showed an ordered pattern according to their developmental stage (Fig. 2a, Extended Data Fig. 3a). We next projected the transcriptomes of naive and primed in vitro cultured human ES cells[21] onto this representation. We found that naive human ES cells were closest to embryonic day (E) 6–E7 cells, whereas primed human ES cells partially overlapped with CS7 epiblast, verifying that at the global transcriptome level, the primed state captured in vitro in human ES cells closely represents the in vivo primed state. A comparison of the naive and primed state in vivo and in vitro showed some differences (Extended Data Fig. 3b, Supplementary Table 4), which could suggest ways to further refine in vitro models. Similar approaches could be adopted to evaluate in vitro models of human gastrulation, such as gastruloids (Extended Data Fig. 3c; details in Supplementary Note 2).

Diffusion maps and RNA velocity analysis[22,23] (Fig. 2b, Extended Data Fig. 3d) revealed trajectories from the epiblast along two broad streams corresponding to mesoderm and endoderm, separated along the second diffusion component (DC2). The first diffusion component (DC1) corresponded closely to cell type and spatial location, reflecting the extent of the differentiation and the 'age' of the cells, based on how far in the past of this sample they had emerged from the epiblast (Fig. 2b, Extended Data Fig. 3d). For example, extra-embryonic mesoderm cells, which emerge relatively early during gastrulation, were further from the epiblast than axial mesoderm cells, which emerge later. The cells that we annotated as nascent, emergent and advanced mesoderm showed overlapping expression of markers of established mesodermal sub-types, such as paraxial or lateral plate mesoderm. This suggests that at this stage, these clusters do not yet represent specified mesodermal subtypes and correspond to transitional states (Extended Data Fig. 4, Supplementary Notes 1, 3, Supplementary Table 16).

To probe changes in the epiblast during gastrulation, we computed RNA velocity vectors with cells belonging to the epiblast, primitive streak, nascent mesoderm and ectoderm (amniotic/embryonic) clusters. This supported the existence of a bifurcation from epiblast, towards mesoderm via the primitive streak on one side and towards ectoderm on the other (Fig. 2c). Ordering cells using diffusion pseudotime provided a method to infer the changes in gene expression as epiblast cells differentiate into ectoderm or enter the primitive streak and begin to delaminate into nascent mesoderm (Fig. 2c, Extended Data Fig. 5). Whereas we could detect robust upregulation of markers common to the amniotic and embryonic ectoderm[24] (*DLX5*, *TFAP2A* and *GATA3*), markers of early neural induction (*SOX1*, *SOX3* and *PAX6*) and differentiated neurons (*TUBB3*, *OLIG2* and *NEUROD1*) were undetectable or expressed at very low levels[25,26] (Extended Data Fig. 5c). In particular, we could not detect any cells expressing two or more of the markers *SOX3*, *PAX6* or *TUBB3*. Together, these data suggest that in this CS7 embryo, neural differentiation had not yet commenced.

The mouse is the predominant model used for research into mammalian gastrulation. To unbiasedly test similarities and differences between human and mouse gastrulation, we used pseudotime analyses to compare the transition from epiblast to nascent mesoderm in the human gastrula with the equivalent populations from the Mouse Gastrula Single Cell Atlas[18] (Extended Data Fig. 6a, Supplementary Tables 5, 6). We identified 662 genes common to both species that

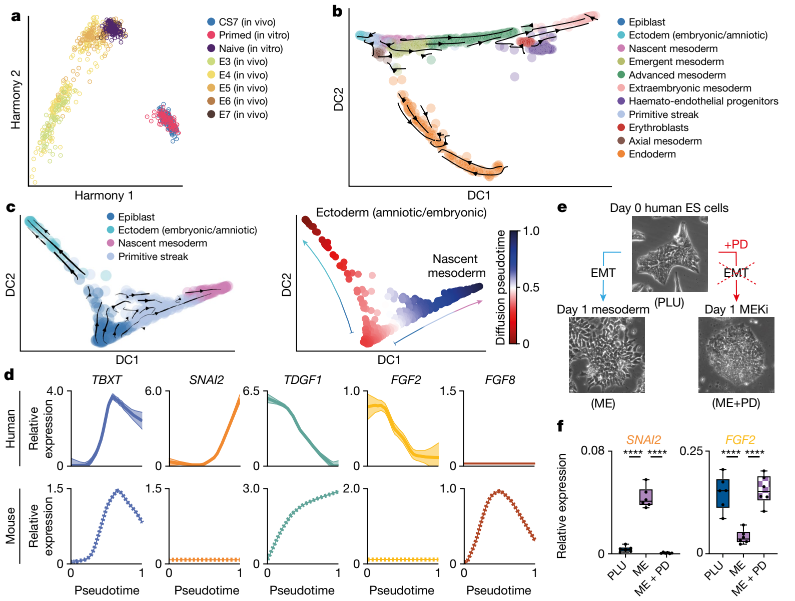


Fig. 2 | State transitions during gastrulation. a, Harmony representation of the transcriptomic profiles of CS7 epiblast cells compared with cells from pre-implantation human embryos[20], and primed and naive human ES cells[21]. **b**, RNA velocity vectors overlaid on diffusion map of cells from all 11 clusters. **c**, Diffusion maps with RNA velocity vectors (left) and diffusion pseudotime coordinates (right). The two differentiation trajectories from epiblast towards ectoderm (amniotic/embryonic) or mesoderm are shown. **d**, Comparison of primitive streak and nascent mesoderm formation in human and mouse[18]. Mean expression profile and standard error along pseudotime is plotted for selected human (top) and mouse (bottom) genes. **e**, In vitro model for epithelial-to-mesenchymal transition (EMT) during gastrulation. Human ES cells (day 0, PLU) are differentiated towards mesendoderm (ME) and undergo EMT. Inhibition of the MEK pathway (MEKi) prevents EMT (ME + PD). **f**, Quantification of selected transcripts (relative to housekeeping genes; Methods) across the three conditions: PLU, ME, ME + PD. Quantitative PCR results are consistent with in vivo data shown in **d**. $n = 6$ from three different experiments. In box plots, centre line shows median, box limits indicate upper and lower quartiles, whiskers extend to minimum and maximum values and dots show mean value per experiment. ****$P < 0.0001$; ordinary one-way ANOVA after Shapiro-Wilk normality test. Source data are presented in Supplementary Table 17.

were differentially expressed along this developmental trajectory (Extended Data Fig. 6b, Supplementary Table 7). The majority of these (531) shared the same trend across pseudotime, either increasing (117) or decreasing (414). For example, in both mouse and human, during the transition from epiblast to nascent mesoderm, *CDH1* expression decreased, *TBXT* was transiently expressed, and *SNAI1* continuously increased (Fig. 2d, Extended Data Fig. 6c). In addition, we also found some genes with trends that differed between the two species, such as *SNAI2* (upregulated only in human), *TDGF1* (opposing trends), *FGF8* (transient expression in mouse only) and *FGF2* (expression downregulated in human, but not expressed at all in mouse). To experimentally validate these human-specific transcriptional trends, we used a human ES cell-based in vitro model of the transition from epiblast to nascent mesoderm and found similar trends during human ES cell differentiation (Fig. 2e, f, Extended Data Fig. 7). We extended this comparison to include the closest available stages of gastrulation of the cynomolgus monkey[19]. An analysis of expression trends of signalling molecules across the three species again revealed broad similarities, as well some specific differences (Extended Data Fig. 8; details in Supplementary Note 4).

Cluster subtypes

The ectoderm (amniotic/embryonic) cluster expresses markers common to the embryonic ectoderm at the rostral boundary of the neural plate, which will generate surface ectoderm, and the amniotic ectoderm[24,27]. To explore this population further, we performed sub-clustering, which revealed two subpopulations, one of which represented amniotic ectoderm, indicated by high expression of *VTCN1* and *GABRP*[28] (Fig. 3a, Supplementary Table 8). The other subpopulation (NNE) represents either embryonic non-neural ectoderm at the rostral boundary of the forming neural plate[27] or immature amnion.

A crucial population of cells to originate from the early epiblast are the primordial germ cells (PGCs). In the mouse, PGCs emerge at approximately E7.25[29,30]. Recent work has shown that cells expressing some PGC markers can be identified at E11[31] in non-human primates and in ex vivo cultured human embryos[13]. Consistent with this, we were able to detect a small population of PGCs in the primitive streak cluster (Fig. 3b, Supplementary Table 9). A comparison of the transcriptional profile of early human PGCs with that of mouse and non-human primate identified markers shared between these species and others that differed, such as *DND1* and *PDPN* (Fig. 3b, Supplementary Table 10).

The endoderm cluster showed a higher order of substructure based on gene expression and anatomical origin of cells. Subclustering revealed four spatially distinct subpopulations: hypoblast, yolk sac (YS) endoderm and two definitive endoderm (DE1 and DE2) groups (Fig. 3c, Extended Data Fig. 9, Supplementary Table 11). A comparison of these cells with mouse endodermal subtypes at E7.25 confirmed our annotation (Fig. 3c). The two definitive endoderm clusters had the largest proportion of cells collected from the caudal region (Extended

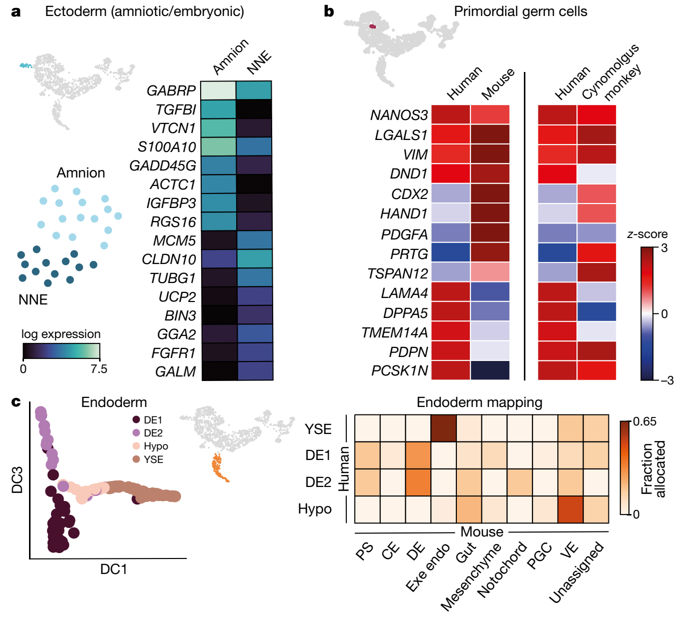


Fig. 3 | Identification of cell subtypes. a, Subclustering of ectoderm (amniotic/embryonic), highlighted in the UMAP (top left), into amnion and non-neural ectoderm (NNE) (bottom left, UMAP representation). Right, heat map of log expression of the top eight upregulated genes in the two subclusters. **b**, PGC population subclustered from the primitive streak cluster, highlighted in the UMAP (top). Heat maps comparing gene expression in human PGCs with those from cultured E7.5 mouse embryos (left) and cynomolgus monkey (right). **c**, Left, diffusion map of endodermal cluster, showing four subclusters: DE1 and DE2; hypoblast (Hypo); and yolk sac endoderm (YSE). Right, heat map showing the fraction of cells from the human endodermal subclusters allocated to mouse cell types at E7.25. CE, caudal epiblast; DE, definitive endoderm; ExE Endo, extraembryonic endoderm; PS, primitive streak; VE, visceral endoderm.

Data Fig. 9b). One of the main differences between them was in the distribution of cells across the phases of the cell cycle, with DE1 being more proliferative compared with DE2 (Extended Data Fig. 9c). DE2 also showed increased expression of the anterior endoderm markers *HHEX*, *OTX2*, *SHISA2* and *CER1* (Extended Data Fig. 9f). Analysis of transcript isoforms also revealed further differences between these endoderm clusters in markers such as *APOA2* and *TTR* (Extended Data Fig. 9i, Supplementary Table 12).

Maturation of haemogenic progenitors

Our initial analysis revealed two blood-related clusters, erythroblasts and haemato-endothelial progenitors (HEP). The identification of primitive erythroblasts was consistent with pigmented cells in the yolk sac and the expression of embryonic globin genes (Fig. 4a, Extended Data Fig. 10f). This was unexpected, given the absence of pigmented blood cells at the equivalent stage in mouse embryos (approximately E7.25).

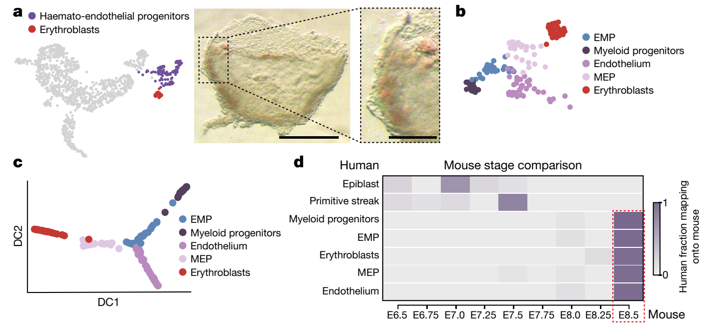


Fig. 4 | Identification of early blood progenitor types in the human. a, Bright-field image of the yolk sac, highlighting pigmented cells (scale bar, 500 µm). The boxed region is magnified on the right (scale bar, 150 µm). **b**, UMAP of the HEP and erythroblast clusters showing four subclusters within the HEP. **c**, Diffusion maps of HEP subclusters and erythroblasts. **d**, Estimation of equivalent mouse stage for selected human clusters. The heat map shows the fraction of human cells from each cluster that maps onto the equivalent mouse cell type at different stages. Epiblast and primitive streak cells are most similar to their mouse counterparts at E7.0 and E7.5, respectively, but blood-related cells are all equivalent to E8.5 mouse cells.

The expression of *XIST* and Y-chromosome specific genes (Extended Data Fig. 10a) ruled out the possibility of maternal origin of these cells.

Unsupervised clustering of the HEP revealed four subpopulations with distinct transcriptional and isoform signatures (Fig. 4b, Extended Data Fig. 10d, Supplementary Tables 13, 14). These represented endothelium, megakaryocyte-erythroid progenitors (MEP) (expressing both megakaryocyte and erythroid markers), myeloid progenitors and an erythro-myeloid progenitor (EMP) population. Diffusion analysis revealed a separation of trajectories based on HEP subtype (Fig. 4c, Extended Data Fig. 10e).

The existence of haemoglobinizing cells and multiple haematopoietic progenitor populations suggest that haematopoiesis in humans had progressed further in comparison to equivalent stage mouse embryos (E6.75–E7.5). To examine this in an unbiased manner, we compared the sequence of the human clusters to the equivalent populations from the Mouse Gastrula Single Cell Atlas[18], which span E6.5–E8.5. In contrast to the human Epiblast and Primitive Streak that correspond to mouse cells from E7.0 and E7.5 respectively, all the human haematopoietic populations most closely correlated with cells from stage E8.5 in the mouse (Fig. 4d, Extended Data Fig. 10g, h), further suggesting that haematopoiesis is more advanced in the human compared to the equivalent stage in mouse.

Discussion

The singular nature of the sequenced specimen means that care must be taken when making generalizations about human gastrulation in utero. Ethically obtained human samples at these early stages are exceptionally rare—thus, in this context, it will be informative to compare this human gastrula transcriptome with those from stage-matched non-human primates. For now, our characterization of this human sample provides some reassurance that it reflects normal development on the basis of gross morphology, karyotype, distribution and frequency of indels, and broad agreement of its single-cell transcriptome with established paradigms of gastrulation from model organisms.

Our characterization reveals that the embryo at this stage already had PGCs and red blood cells, but had not yet initiated neural specification. The differentiation trajectory and signalling pathways of gastrulating cells transitioning from epiblast to mesoderm was broadly conserved between humans and the mouse, indicating that the mouse represents a good model of human gastrulation. However, some notable differences suggest that the process of EMT may be regulated differently at the level of specific signalling family members. These human-specific details of differentiation will be a valuable resource for refining approaches towards directed differentiation of human embryonic stem cells. Furthermore, they will help in interpreting experimental results on gastrulation from model organisms such as the mouse or in vitro gastruloid systems. The human and mouse gastrula are morphologically very different, with the human gastrula forming a disc and the the mouse gastrula being cylindrical. This profound difference in morphology alters the migratory path of cells during gastrulation and therefore the inductive signals that the cells might be subject to from neighbouring germ layers. It will therefore be important to compare this human gastrula single-cell transcriptome with stage-matched gastrulae of other organisms with a similar embryonic disc, such as the rabbit, chick and non-human primates. This will enable us to address the extent to which specific differences between human and mouse transcriptomes are simply a result of evolutionary divergence or, instead, reflect differences in morphology.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-021-04158-y.

1. Stern, C. D. *Gastrulation: From Cells to Embryo* (CSHL Press, 2004).
2. Tam, P. P. L. & Loebel, D. A. F. Gene function in mouse embryogenesis: get set for gastrulation. *Nat. Rev. Genet.* **8**, 368–381 (2007).
3. Bardot, E. S. & Hadjantonakis, A. K. Mouse gastrulation: coordination of tissue patterning, specification and diversification of cell fate. *Mech. Dev.* **163**, 103617 (2020).
4. Arnold, S. J. & Robertson, E. J. Making a commitment: cell lineage allocation and axis patterning in the early mouse embryo. *Nat. Rev. Mol. Cell Biol.* **10**, 91–103 (2009).
5. O'Rahilly, R. & Müller, F. Developmental stages in human embryos: Revised and new measurements. *Cells Tissues Organs* **192**, 73–84 (2010).
6. Yamaguchi, Y. & Yamada, S. The Kyoto collection of human embryos and fetuses: History and recent advancements in modern methods. *Cells Tissues Organs* **205**, 314–319 (2019).
7. Florian, J. & Hill, J. P. An early human embryo (no. 1285, Manchester Collection), with capsular attachment of the connecting stalk. *J. Anat.* **69**, 399–411 (1935).
8. De Bakker, B. S. et al. An interactive three-dimensional digital atlas and quantitative database of human development. *Science* **354**, aag0053 (2016).
9. Warmflash, A., Sorre, B., Etoc, F., Siggia, E. D. & Brivanlou, A. H. A method to recapitulate early embryonic spatial patterning in human embryonic stem cells. *Nat. Methods* **11**, 847–854 (2014).
10. Martyn, I., Kanno, T. Y., Ruzo, A., Siggia, E. D. & Brivanlou, A. H. Self-organization of a human organizer by combined Wnt and Nodal signaling. *Nature* **558**, 132–135 (2018).
11. Simunovic, M. et al. A 3D model of a human epiblast reveals BMP4-driven symmetry breaking. *Nat. Cell Biol.* **21**, 900–910 (2019).
12. Moris, N. et al. An in vitro model of early anteroposterior organization during human development. *Nature* **582**, 410–415 (2020).
13. Chen, D. et al. Human primordial germ cells are specified from lineage-primed progenitors. *Cell Rep.* **29**, 4568–4582.e5 (2019).
14. Molè, M. A. et al. A single cell characterisation of human embryogenesis identifies pluripotency transitions and putative anterior hypoblast centre. *Nat. Commun.* **12**, 3769 (2021).
15. Xiang, L. et al. A developmental landscape of 3D-cultured human pre-gastrulation embryos. *Nature* **577**, 537–542 (2020).
16. Zhou, F. et al. Reconstituting the transcriptome and DNA methylome landscapes of human implantation. *Nature* **572**, 660–664 (2019).
17. O'Rahilly, R. & Müller, F. eds. *Developmental Stages in Human Embryos*. (Carnegie Institute of Washington, 1987).
18. Pijuan-Sala, B. et al. A single-cell molecular map of mouse gastrulation and early organogenesis. *Nature* **566**, 490–495 (2019).
19. Ma, H. et al. In vitro culture of cynomolgus monkey embryos beyond early gastrulation. *Science* **366**, eaax7890 (2019).
20. Petropoulos, S. et al. Single-cell RNA-seq reveals lineage and X chromosome dynamics in human preimplantation embryos. *Cell* **165**, 1012–1026 (2016).
21. Messmer, T. et al. Transcriptional heterogeneity in naive and primed human pluripotent stem cells at single-cell resolution. *Cell Rep.* **26**, 815–824.e4 (2019).
22. Haghverdi, L., Büttner, M., Wolf, F. A., Buettner, F. & Theis, F. J. Diffusion pseudotime robustly reconstructs lineage branching. *Nat. Methods* **13**, 845–848 (2016).
23. La Manno, G. et al. RNA velocity of single cells. *Nature* **560**, 494–498 (2018).
24. Streit, A. The preplacodal region: an ectodermal domain with multipotential progenitors that contribute to sense organs and cranial sensory ganglia. *Int. J. Dev. Biol.* **51**, 447–461 (2007).
25. Trevers, K. E. et al. Neural induction by the node and placode induction by head mesoderm share an initial state resembling neural plate border and ES cells. *Proc. Natl Acad. Sci. USA* **115**, 355–360 (2017).
26. Delile, J. et al. Single cell transcriptomics reveals spatial and temporal dynamics of gene expression in the developing mouse spinal cord. *Development* **146**, dev1738078 (2019).
27. Yang, L. et al. An early phase of embryonic Dlx5 expression defines the rostral boundary of the neural plate. *J. Neurosci.* **18**, 8322–8330 (1998).
28. Roost, M. S. et al. KeyGenes, a tool to probe tissue differentiation using a human fetal transcriptional atlas. *Stem Cell Rep.* **4**, 1112–1124 (2015).
29. Chiquoine, A. D. The identification, origin, and migration of the primordial germ cells in the mouse embryo. *Anat. Rec.* **118**, 135–146 (1954).
30. Magnúsdóttir, E. & Surani, A. M. How to make a primordial germ cell. *Development* **141**, 245–252 (2014).
31. Sasaki, K. et al. The germ cell fate of cynomolgus monkeys is specified in the nascent amnion. *Dev. Cell* **39**, 169–185 (2016).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2021

Methods

Collection of human gastrula cells

The CS7 embryo was provided by the Human Developmental Biology Resource (HDBR) (https://www.hdbr.org/general-information). HDBR has approval from the UK National Research Ethics Service (London Fulham Research Ethics Committee (18/LO/0822) and the Newcastle and North Tyneside NHS Health Authority Joint Ethics Committee (08/H0906/21+5)) to function as a Research Tissue Bank for registered projects. The HDBR is monitored by The Human Tissue Authority for compliance with the Human Tissue Act (2004). This work was done as part of project #200295 registered with the HDBR. The material was collected after appropriate informed written consent from the donor by medical termination. The sample was collected and transported in cold L15 medium. It was then transferred to M2 medium and imaged on a Leica Stereo microscope. The sample was micro-dissected using tungsten needles and dissociated into single cells using 200 µl Accutase (ThermoFisher, catalogue (cat.) no. A1110501) for 12 min at 37 °C, with agitation every 2 min, before adding 200 µl heat-inactivated FBS (ThermoFisher, cat. no. 10500) to quench the reaction. Cells were then centrifuged at 1,000 rpm for 3 min at 4 °C before being suspended in 100 µl HBSS (ThermoFisher, cat. no. 14025) + 1% FBS, and stored on ice. Single cells were collected using a Sony SH800 fluorescence-activated cell sorter with a stringent single-cell collection protocol and sorted into 384-well plates containing SMART-seq2 lysis buffer[32] plus ERCC spike-ins (1:10 M). To ensure we collected good quality cells, a live/dead dye (Abcam, Cat No. ab115347) was used; 100 µl was added to the cell suspension at a 2× concentration in HBSS 10 min before collection, and live cells were collected on the basis of their FITC intensity. Once cells were collected, plates were sealed, spun down, and frozen using dry ice before being stored at −80 °C. This complete process, from dissection to single-cell collection, took approximately 2–3 h. The embryo was karyotypically normal (region-specific assay: (13, 15, 16, 18, 21, 22) × 2, (X, Y) × 1).

Single-cell RNA sequencing

mRNA from single cells was isolated and amplified (21 PCR cycles) using the SMART-seq2 protocol[32]. Multiplexed sequencing libraries were generated from cDNA using the Illumina Nextera XT protocol and 125 bp paired-end sequencing was performed on an Illumina HiSeq 2500 instrument (V4 chemistry).

Raw data processing and normalization

To quantify the abundance of transcripts from 1,719 cells, Salmon v0.17[33] was used. After indexing the human transcriptome (GRCh38.p13) in quasi-mapping-based mode, we quantified the transcripts with Salmon using the −seqBias and −gcBias flags. We combined the transcript level abundances to the corresponding gene level counts, which were aggregated into a gene-count matrix. Then, for downstream analyses, we only retained cells with more than 2,000 detected genes, with overall mapping rate greater than 55% and with relatively low mapping rate to mitochondrial genes (<0.02) and to ERCC spike-ins (<0.2). After this step, we obtained 1,195 good quality cells. The data were normalized using the quickcluster and normalize functions from the scran package in R[34]. This was followed by pseudocount addition of 1 and natural-log transformation of the count matrix.

Clustering and cell type identification

To identify clusters of cells, we applied a graph-based algorithm. First, we selected the top 4,000 highly variable genes (HVGs) using the high_variable_genes function from scanpy v1.4.4[35]. We constructed the cell–cell distance matrix as $\sqrt{(1-\rho)/2}$, where ρ is the Spearman's correlation coefficient between cells. Next, a k-nearest neighbour graph was built with the first 30 principal components and $k = 50$. This was accomplished by the 'neighbors' function in scanpy, which computes the connectivity between cells based on UMAP (method = 'umap')[36]. To identify clusters, we applied the Leiden algorithm for community detection to the resulting graph (with a resolution of 0.75), as it has been shown to be a superior alternative to Louvain[37]. The same algorithm and resolution were used for subclustering the endoderm, the ectoderm and the haemogenic endothelial progenitors clusters with top 2,000 HVGs in each. However, in this case the k-nearest neighbour graph was built with the first 10 principal components and $k = 20$. We visualized the resulting clusters in two dimensions by computing a UMAP representation with default parameters in scanpy (tl.umap function). To check the robustness of the clustering, we also computed the shared nearest neighbour (SNN)[38] graph and applied Leiden to it (resolution = 1.75), which produced very similar clusters (the adjusted mutual information score calculated with Python's sklearn module was 0.8).

We identified marker genes for the clusters with the Wilcoxon rank-sum test in scanpy (rank_genes_groups function), by comparing the gene-expression levels in a given cluster with the rest of the cells in the dataset. The genes were ranked according to their false discovery rate (FDR), after P values were corrected with the Benjamini–Hochberg method. We visualized the expression values of marker genes on a heat map, after scaling the log-normalized counts between 0 and 1 by using the 'standard_scale = var' option in the scanpy heat map plotting function sc.pl.heatmap.

Isoform analysis

We obtained the isoform-level count matrix from Salmon, considering transcripts per million-normalized counts and the ENSEMBL database (GRCh38.p13) for annotation. We compared transcript levels between pairs of clusters. First, we removed genes with more than 80% counts mapped to a single isoform. Then, for each gene, we built a contingency table including the average normalized levels of each isoform in the two clusters being compared. A chi-squared test was then used to check whether the isoform abundances differ between the two clusters of cells for a given gene, as in ref. [39].

Trajectory analysis using diffusion pseudotime and RNA velocity

For the whole-embryo diffusion map, we built the k-nearest neighbour graph as described above (with $k = 50$ and using the first 30 principal components) to find the connectivity kernel width. We then used the diffmap function to build the diffusion map.

To estimate the trajectory of epiblast differentiation, we took 2,000 HVGs from epiblast, primitive streak, ectoderm and nascent mesoderm clusters combined. Finally, the diffusion components were computed from the first 15 principal components with $k = 15$.

To illustrate the estimated direction of differentiation of epiblast cells, we embedded the RNA velocities[23] of single cells on the above diffusion map. For this task, we aligned reads from each cell using STAR v2.7[40] to the human reference genome (GRCh38.p13), which was obtained from ENSEMBL. The aligned bam files were processed with velocyto v0.17.17[23] with the default run-smartseq2 mode, to create a count matrix made of spliced and unspliced read counts.

After filtering genes with less than 10 spliced and un-spliced counts from this matrix, we calculated the moments for velocity estimation by utilizing a built-in function from scVelo Python module v01.20[41]. Subsequently, we inferred the splicing kinetic dynamics of the genes by applying the recover_dynamics function. The velocity of each gene was estimated by solving splicing kinetics in the dynamical mode with the velocity function. Finally, we embedded the resulting velocities on the diffusion space calculated above by means of the velocity_embedding function from the scVelo module. The diffusion map and the RNA velocities for the mesoderm specification analysis were computed in the same way.

We defined a diffusion pseudotime (dpt) coordinate on the diffusion map of epiblast differentiation in order to visualize gene-expression trends. First, we fixed the cell with the highest value of the first diffusion component (DC1) as root, so that the middle point of pseudotime would fall roughly into epiblast. We fitted the expression levels of the genes as a function of the pseudotime with a generalized additive model using the gam package in R (v1.16.1). For visualisation purpose, we transformed the pseudotime values as (1 − dpt), so that ectoderm cells would fall onto the left side and primitive streak and nascent mesoderm would be on the right side of the pseudotime plot (Extended Data Fig. 5a–c). Both fitted and unfitted values of the genes were scaled by dividing each by its maximum value of expression.

Human and mouse EMT comparison

For this analysis, we considered published single-cell RNA-seq data from mouse embryos during mid-streak stage (E7.25)[18], but we also checked that the results remain largely unaffected if data from E7.0 or E7.5 are used. Epiblast, primitive streak and nascent mesoderm clusters were selected from the human and the mouse datasets for downstream analysis, and they were analysed separately as detailed below.

After constructing diffusion maps as described above with default parameters, we defined pseudotime starting from the cell with lowest DC1 value in both cases (Extended Data Fig. 6a). After fitting gene-expression values along pseudotime with generalized additive models (see above), we calculated the *P* values using the analysis of variance (ANOVA) non-parametric test from the gam R package and we then obtained the FDR values (Benjamini–Hochberg method). Genes with FDR <0.1 were clustered according to their expression pattern. This was achieved by hierarchical clustering with Spearman's correlation distance as described above (hclust function in R). For estimating the number of clusters, the dynamic hybrid cut method was used (cutreeDynamic function, in the package dynamicTreeCut, version 1.63, with 'deepslit' = 0 and 'minclustersize' = 50). In both human and mouse, we found three clusters of genes, two of which were characterized by a clear upward or downward average trend with an absolute $\log_2$ fold change greater than 1 between the fitted values at the end and at the beginning of the trajectory.

For the human–mouse comparison, we converted mouse genes to human equivalents (one-to-one homologous genes only) with the biomaRt R package[42]. We compared the trends of genes in human and mouse, and in particular we looked at genes coding for signalling molecules, as listed in the curated database of the CellPhoneDB package[43]. To visualize the trend of selected genes, we normalized the expression values by the maximum in both mouse and human. We set fitted values to zero for the genes that were expressed in fewer than 10 cells.

Mouse cluster comparison and blood staging analysis

We mapped the cells from the human gastrula against the mouse clusters at E7.25 available from ref. [18]. To do this, we took the median levels of genes as a representation of the typical expression pattern of a given mouse cluster, and then, for each cell in the human gastrula, we used the "scmapCluster" function from the "scmap" R package[44] (with 1,000 genes and similarity threshold parameter set to 0) to identify the mouse cluster that was most similar to it. We performed the same procedure for human endoderm (Fig. 3c) and HEP (Extended Data Fig. 10g) subclusters.

For staging analysis, we selected epiblast, primitive streak, endothelium, blood progenitors (1 and 2), and erythroid (1, 2 and 3) mouse clusters across the 9 stages, from E6.5 until E8.5. We merged the two blood progenitor clusters as well as three erythroid clusters and we obtained four mouse blood-related clusters that were used in downstream analyses. After verifying that the human blood-related clusters map onto the corresponding mouse clusters, we built a representative expression pattern for mouse for each cluster and stage, by calculating the median expression value of the genes per cluster and stage. Cells from human gastrula blood (erythroblasts, myeloid progenitors, endothelium, blood progenitors and EMPs), epiblast and primitive streak clusters were projected onto the corresponding mouse clusters (human erythroblasts to mouse erythroid; human myeloid progenitors, blood progenitors and EMPs to mouse blood progenitors; human endothelium to mouse endothelium; human primitive streak to mouse primitive streak; human epiblast to mouse epiblast) using scmap with the same parameters specified above.

Human and non-human primate gastrulation comparison

We considered single-cell non-human primate (NHP) gastrulation data[19] at 16 days post fertilization (dpf), since PGCs were only identified at that stage. Seurat integration method was applied to human and NHP single-cell data with 3,000 features used to find anchors (anchor.features parameter) and 70 neighbours to filter anchors (k.filter parameter). After obtaining the corrected expression values, we calculated the mean expression level of each gene per cluster. Finally, we performed hierarchical clustering with Spearman's correlation-based distance (see above) and the average aggregation method, using the linkage function from Python's scipy module (v 1.5.2).

PGC identification and cross-species comparison

To single out the PGCs, we ran the RaceID algorithm (RaceID package v0.1.5)[45], which can identify rare cell types, on the cells in the primitive streak cluster. We used these parameter values: $k = 1$, outlg = 8 and probthr = 0.005. This resulted in the identification of 9 subclusters of outlier cells. Among these, the PGCs were identified as the only cluster of outlier cells that had a median expression of PGC marker genes (*NANOS3*, *SOX17*, *DND1*, *LAMA4* and *DPPA5*) above 0.

To perform cross-species comparison of PGCs, we considered epiblast, primitive streak and PGC cells from human and mouse (E7.5 stage), and late epiblast (L-epi), late gastrulating cells 1 (L-gast1) and PGC clusters from non-human primate (16 dpf stage) single-cell datasets. *Z*-scores were calculated for each gene per species by using rank_genes_groups function from scanpy with Wilcoxon rank-sum test (method = 'wilcoxon') applied to PGC versus all others. The genes shown in the heat maps of Fig. 3d were selected from the top differentially expressed genes between PGC and the other clusters.

Cross-species signalling comparison

We obtained the gene sets for FGF, WNT and BMP signalling pathways from MSigDB database[46]. Here, we considered epiblast, primitive streak and nascent mesoderm clusters from mouse and human gastrula data, and L-epi, L-gast1 and L-gast2 clusters from the non-human primate dataset. We computed the *z*-scores for each cluster per organism separately with Wilcoxon rank-sum test as described above. The genes that were expressed in fewer than ten cells across all clusters in a species were labelled as undetected (Extended Data Fig. 8).

Cell cycle prediction

We estimated the cell cycle phase of each cell by applying the pairs algorithm described in[47]. A Python implementation of this algorithm, pypairs v3.1.1 was used in this analysis (https://pypairs.readthedocs.io/en/latest/documentation.html). After determining marker pairs from a training dataset[48] with the sandbag function, we applied the function cyclone to assign a cell cycle phase to each cell.

Indel analysis

Using our transcriptomic data, we estimated the sizes of genomic insertions and deletions (indels) in our data as well as in a dataset from human fetal liver cells[49]. This dataset was also processed with SMART-seq2 protocol and paired-end sequencing, although read lengths (75 bp) were smaller than in our data (125 bp). Hence, to minimize confounding effects in the results, we trimmed the reads in our data before processing it for this analysis. We aligned the data to the reference genome (GRCh38.p13), using bwa-mem v0.6[50] with default parameters. We then merged the aligned data from each single cell into one bam file and performed indel calling with a pipeline for insertion and deletion detection from RNA-seq data called transIndel v0.1[51]. We kept the parameters at default values, except the minimum deletion length to be detected, which was set to 1 (-L flag set to 1).

Differential gene-expression analysis between rostral and caudal mesoderm

We used the R packages DESeq2 v3.11[52] and Seurat v3.0[53] to identify the genes differentially expressed between rostral and caudal parts of the mesoderm cluster. After creating a Seurat object with the mesoderm cells, their anatomical and plate information, we converted it to DESeq2 object with convertTo function. We found differentially expressed genes (with FDR<0.1) between caudal and rostral parts of the mesoderm with DESeqDataSet and DESeq functions, while controlling for the plate effect.

Human embryonic stem cells comparison

For this comparison, we considered previously published single-cell RNA-seq data from pre-implantation human embryos[20] and from human ES cells[21]. In the pre-implantation embryo data, we removed cells from extra-embryonic tissues, from immunosurgery samples and with unannotated stage. Moreover, we only kept cells with a $\log_{10}$ total number of reads greater than 5.5. This resulted in 442 cells distributed between E3 and E7 stages.

In the human ES cell dataset, only cells in batch 1 (including both primed and naive human ES cells) that passed the quality test performed in the original publication were taken.

These data from pre-implantation embryos and human ES cells were combined with the epiblast cells in our dataset, and count per million (CPM) normalization was performed. To assess the relationship between the datasets, we also used two different integration methods: Harmony[54] (with the same HVGs and default parameters) and Seurat (using the same procedure as in the comparison with NHP data described above).

To compare changes in gene-expression levels between the naive and primed state in epiblast and in human ES cells, we took cells from E6 stage, given that they were closest to the naive (Fig. 2a, Extended Data Fig. 3a). Then, the log-fold changes of the previously identified HVGs (after removal of genes with less than mean log count of 1) were calculated between CS7 vs E6 cells and primed versus naive human ES cells, after adding a pseudocount of 0.1 to the mean expression values. The line in Extended Data Fig. 3b is obtained through a linear regression (LinearRegression function from sklearn Python module).

Human gastruloid comparison

Recently published spatial transcriptomic data from human gastruloids were considered for the comparison[12]. Specifically, we took the z-scores of the genes that were found to be reproducible across the two replicates of the spatial transcriptomic experiment (source data of figure 3c in ref. [12]). For these genes, we also calculated z-scores in each cluster of our human gastrula data using rank_genes_groups function from scanpy with Wilcoxon rank-sum test (method = 'wilcoxon') applied to cells in a given cluster versus all other cells.

Then, we compared the human gastrula with the gastruloid data by computing

$$\rho_{ij} = \mathrm{corr}\left(\mathbf{G}_i, \mathbf{S}_j\right),$$

that is the Pearson's correlation coefficient between the z-scores of the ith gastrula cluster G_i and the z-scores of a gastruloid slice taken at the jth position along the anterior–posterior axis S_j. A null distribution for ρ_{ij}, $\mathcal{P}(\rho_{ij})$, was estimated by computing the Pearson's correlation coefficient after shuffling the z-scores of the gastruloid dataset across slices 500 times. We estimated a P value p_{ij} as:

$$p_{ij} = \sum_{\rho_{ij}^{*} > \rho_{ij}} \mathcal{P}\left(\rho_{ij}^{*}\right)$$

Maintenance and differentiation of human ES cells

Human ES cells (H9/WA09 line; WiCell) were cultured on plates coated with 10 µg ml^{-1} vitronectin (Stem Cells Technologies) at 37 °C with 5% CO_2. Pluripotent human ES cells were plated as single cells at 4.0×10^4–5.0×10^4 cells per cm^2 using accutase (Gibco) and 10 µM Y27632 (Selleck), and maintained for two days in E6 medium[55] supplemented with 2 ng ml^{-1} TGF-β (bio-techne) and 25 ng ml^{-1} FGF2 (M. Hyvönen, Cambridge University). These cells were sampled as 'D0 PLU'. Then, the cells were cultured for one day in CDM/PVA medium[56], 1 mg ml^{-1} polyvinyl alcohol (Sigma) (instead of BSA) with 100 ng ml^{-1} activin A (M. Hyvönen, Cambridge University), 80 ng ml^{-1} FGF2, 10 ng ml^{-1} BMP4 (bio-techne), 10 µM LY294002 (Promega) and 3 mM CHIR99021 (Tocris), and sampled as 'D1 ME' or 'D1 ME + PD'. PD0325901 (Stem Cell Institute) was added at 1 µM. Bright-field images were taken with an Axiovert microscope (200M, Zeiss). Authentication of the H9/WA09 cell line was conducted by fingerprinting and the cells were confirmed negative for mycoplasma.

Immunocytochemistry

Cells plated on vitronectin-coated round coverslips (Scientific Laboratory Supplies) were washed once with PBS, and fixed with 4% paraformaldehyde (Alfa Aesar) in PBS at room temperature for 10 min. Following another PBS wash, cells were incubated with 0.25% Triton in PBS at 4 °C for 15–20 min, 0.5% BSA (Sigma) in PBS at room temperature for 30 min, primary antibodies at 4 °C overnight and secondary antibodies at room temperature for one hour. Anti-E-cadherin antibody (3195, Cell Signaling Technology; 1:200), and anti-Rabbit IgG–Alexa Fluoro 568 (A10042, Invitrogen; 1:1,000) together with 10 µg ml^{-1} Hoechst33258 (B2883) were diluted in 0.5% BSA in PBS and each staining was followed by three washes with 0.5% BSA in PBS. Coverslips were preserved on slide glasses (Corning) with ProLong Gold Antifade Mountant (Life Technologies) and nail polish, and observed with a Zeiss inverted confocal system (LSM 710, Zeiss).

Quantitative real-time PCR for human ES cell samples

Total RNA was extracted from cells using the GenElute Mammalian Total RNA Miniprep Kit (Sigma-Aldrich) and the On-Column DNase I Digestion set (Sigma-Aldrich). Complementary DNA was synthesized from the RNA using random primers (Promega), dNTPs (Promega), RNAseOUT (Invitrogen) and SuperScript II (Invitrogen). Real-time PCR was performed with KAPA SYBR FAST qPCR Master Mix (Kapa Biosystems) on QuantStudio 12K Flex Real-Time PCR System machine (Thermo Fisher Scientific). Molecular grade water (Thermo Fisher Scientific) was used when necessary. Each gene-expression level was normalized by the average expression level of *PBGD* and *RPLP0*. Primer sequences are shown in Supplementary Table 15 and source data are provided in Supplementary Table 17. Statistical analysis was performed using GraphPad Prism.

Mouse strains, husbandry and embryo collection

All animal experiments complied with the UK Animals (Scientific Procedures) Act 1986, approved by the local Biological Services Ethical Review Process and were performed under UK Home Office project licenses PPL 30/3420 and PCB8EF1B4. To obtain wild-type embryos, C57BL/6 males (in house) were crossed with 8- to 16-week-old CD1 females (Charles River). All mice were maintained in a 12-h light:dark cycle. Noon of the day when a vaginal plug was found was designated E0.5. To dissect the embryos, the pregnant females were killed by cervical dislocation in accordance with Schedule 1 of the Animals (Scientific Procedures) Act. Embryos of the appropriate stage were dissected in M2 medium (Sigma-Aldrich, cat. no. M7167).

In situ hybridization chain reaction

In situ hybridization chain reaction (HCR) kit (version 3) containing amplifier set, hybridization, amplification, wash buffers and DNA probe sets, were purchased from Molecular Instruments (http://molecularinstruments.org) and the protocol described[57] was followed with slight modifications[58]. Probe libraries were designed and manufactured by Molecular Instruments using *Mus musculus* sequences from the NCBI database. Following HCR embryos were then placed into 87% glycerol solution and imaged on a Zeiss 880 confocal microscope with a 40× oil (1.36 NA) objective. Images were captured at 512 × 512 pixels using multiple tiles with a z-step of 1.5 μm. Each HCR was repeated on at least 3 embryos.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

The raw data from our study can be downloaded from ArrayExpress under accession code E-MTAB-9388. The processed data can be downloaded from http://www.human-gastrula.net. Datasets used as references include mouse gastrula data (E-MTAB-6967); pre-implantation embryo data: E-MTAB-3929. Source data are provided with this paper.

Code availability

All data were analysed with standard programs and packages, as detailed in Methods. The code used to create the human gastrula shiny app is available at https://github.com/ScialdoneLab/human-gastrula-shiny.

32. Picelli, S. et al. Full-length RNA-seq from single cells using Smart-seq2. *Nat. Protoc.* **9**, 171–181 (2014).
33. Patro, R., Duggal, G., Love, M. I., Irizarry, R. A. & Kingsford, C. Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods* **14**, 417–419 (2017).
34. Lun, A. T. L., Bach, K. & Marioni, J. C. Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. *Genome Biol.* **17**, 75 (2016).
35. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: Large-scale single-cell gene expression data analysis. *Genome Biol.* **19**, 15 (2018).
36. McInnes, L., Healy, J., Saul, N. & Großberger, L. UMAP: uniform manifold approximation and projection. *J. Open Source Softw.* **3**, 861 (2018).
37. Traag, V. A., Waltman, L. & van Eck, N. J. From Louvain to Leiden: guaranteeing well-connected communities. *Sci. Rep.* **9**, 5233 (2019).
38. Patrick, E. A. Clustering using a similarity measure based on shared near neighbors. *IEEE Trans.* **C-22**, 1025–1034 (1973).
39. Froussios, K., Mourão, K., Simpson, G., Barton, G. & Schurch, N. Relative abundance of transcripts (RATs): identifying differential isoform abundance from RNA-seq. *F1000Research* **8**, 213 (2019).
40. Dobin, A. et al. STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
41. Bergen, V., Lange, M., Peidli, S., Wolf, F. A. & Theis, F. J. Generalizing RNA velocity to transient cell states through dynamical modeling. *Nat. Biotechnol.* **38**, 1408–1414 (2020).
42. Durinck, S., Spellman, P. T., Birney, E. & Huber, W. Mapping identifiers for the integration of genomic datasets with the R/ Bioconductor package biomaRt. *Nat. Protoc.* **4**, 1184–1191 (2009).
43. Efremova, M., Vento-Tormo, M., Teichmann, S. A. & Vento-Tormo, R. CellPhoneDB: inferring cell–cell communication from combined expression of multi-subunit ligand–receptor complexes. *Nat. Protoc.* **15**, 1484–1506 (2020).
44. Kiselev, V. Y., Yiu, A. & Hemberg, M. scmap: Projection of single-cell RNA-seq data across data sets. *Nat. Methods* **15**, 359–362 (2018).
45. Grün, D. et al. Single-cell messenger RNA sequencing reveals rare intestinal cell types. *Nature* **525**, 251–255 (2015).
46. Subramanian, A. et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA* **102**, 15545–15550 (2005).
47. Scialdone, A. et al. Computational assignment of cell-cycle stage from single-cell transcriptome data. *Methods* **85**, 54–61 (2015).
48. Leng, N. et al. Oscope identifies oscillatory genes in unsynchronized single-cell RNA-seq experiments. *Nat. Methods* **12**, 947–950 (2015).
49. Segal, J. M. et al. Single cell analysis of human foetal liver captures the transcriptional profile of hepatobiliary hybrid progenitors. *Nat. Commun.* **10**, 3350 (2019).
50. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. Preprint at https://arxiv.org/abs/1303.3997 (2013).
51. Yang, R., Van Etten, J. L. & Dehm, S. M. Indel detection from DNA and RNA sequencing data with transIndel. *BMC Genomics* **19**, 270 (2018).
52. 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).
53. Stuart, T. et al. Comprehensive integration of single-cell data. *Cell* **177**, 1888–1902.e21 (2019).
54. Korsunsky, I. et al. Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat. Methods* **16**, 1289–1296 (2019).
55. Chen, G. et al. Chemically defined conditions for human iPSC derivation and culture. *Nat. Methods* **8**, 424–429 (2011).
56. Johansson, B. M. & Wiles, M. V. Evidence for involvement of activin A and bone morphogenetic protein 4 in mammalian mesoderm and hematopoietic development. *Mol. Cell. Biol.* **15**, 141–151 (1995).
57. Choi, H. M. T. et al. Third-generation in situ hybridization chain reaction: multiplexed, quantitative, sensitive, versatile, robust. *Develpoment* **145**, dev165753 (2018).
58. Tyser, R. C. V. et al. Characterization of a common progenitor pool of the epicardium and myocardium. *Science* **371**, eabb2986 (2021).

Acknowledgements Human embryonic material was provided by the MRC–Wellcome Trust funded (grant no. 099175/Z/12/Z and MR/R006237/1) Human Developmental Biology Resource (www.hdbr.org). We thank N. Ashley (Oxford MRC single-cell facility) for help with sequencing, M. De Bruijn, B. Gottgens, J. Palis, L. Robertson, T. Rodriguez and M.-E. Torres-Padilla for helpful comments. This work was funded by British Heart Foundation Immediate Postdoctoral Basic Science Research Fellowship no. FS/18/24/33424 to R.C.V.T., JSPS Overseas Research Fellowship to S.N., European Research Council advanced grant ERC: 741707 to L.V., funding from the Helmholtz Association to A.S. and Wellcome Awards 105031/C/14/Z, 108438/Z/15/Z, 215116/Z/18/Z and 103788/Z/14/Z to S.S.

Author contributions Human gastrula processing: R.C.V.T. and S.S. Human ES cell in vitro experiments: S.N. Computational analyses of sequence data: E.M. and A.S. Gastrula single-cell annotation and analyses: R.C.V.T., E.M., A.S. and S.S. Human ES cell in vitro data analysis: R.C.V.T. and S.N. Preparation of illustrations and figures: R.C.V.T. Preparation of manuscript draft: R.C.V.T., E.M., S.N., A.S. and S.S. Editing and review of final manuscript: R.C.V.T., E.M., S.N., L.V., A.S. and S.S. Study coordination: A.S. and S.S.

Competing interests The authors declare no competing interests.

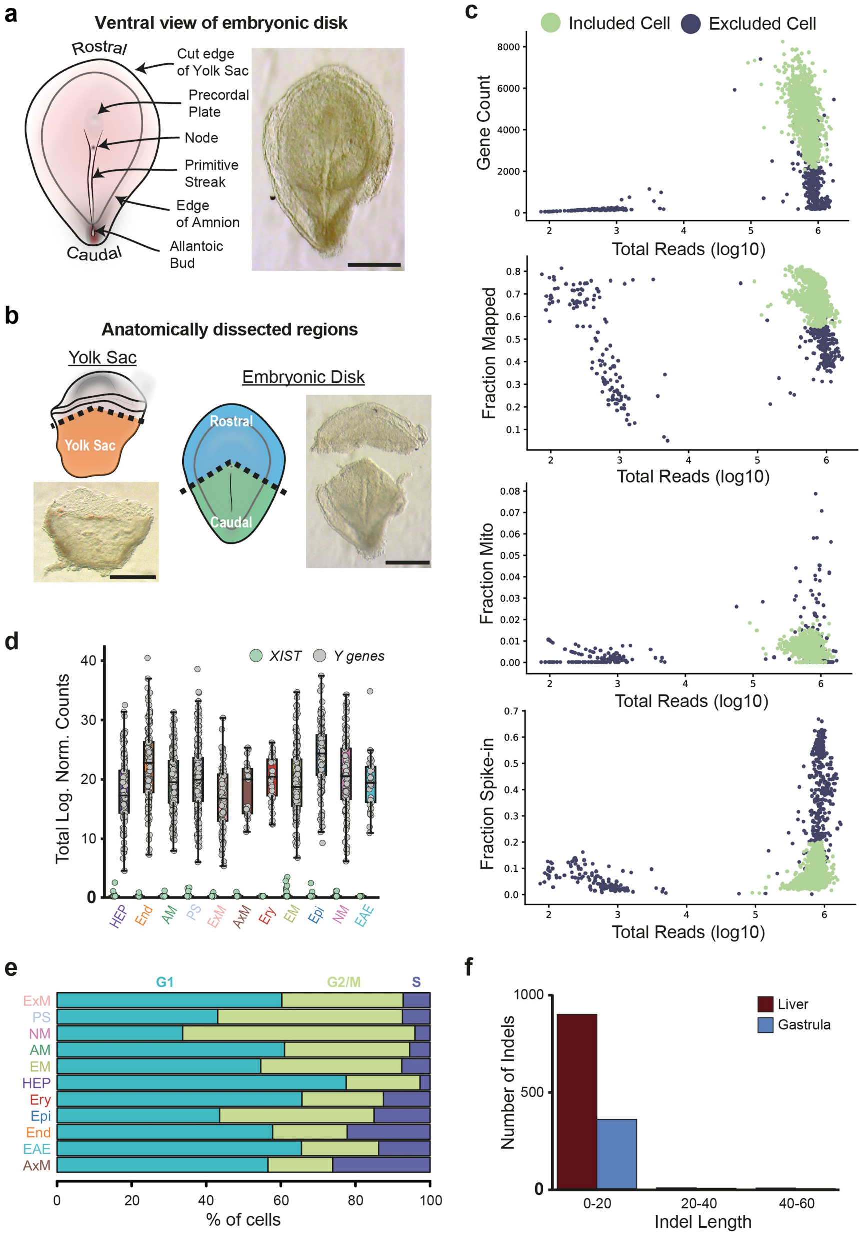
Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41586-021-04158-y.

Correspondence and requests for materials should be addressed to Antonio Scialdone or Shankar Srinivas.

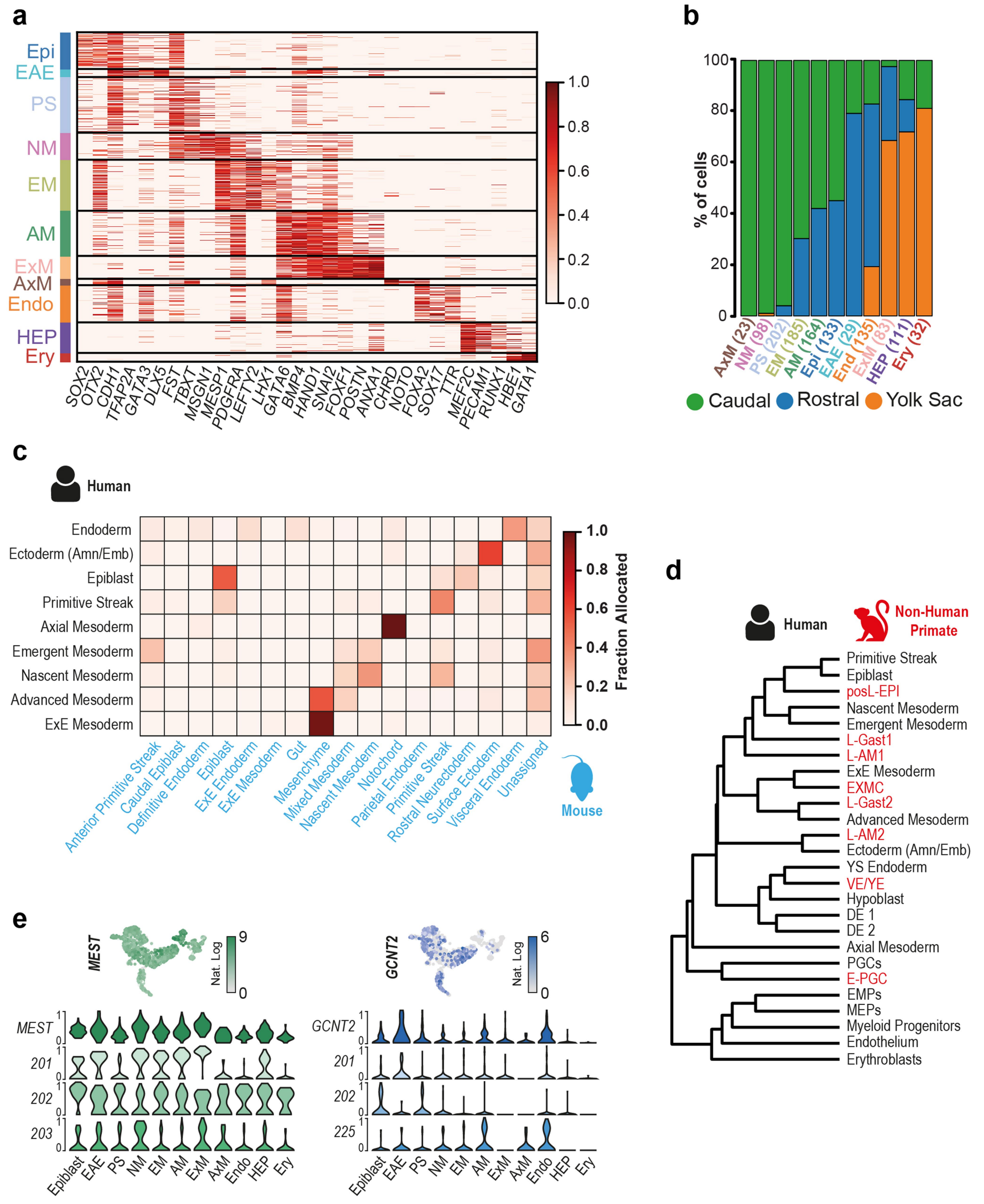
Peer review information *Nature* thanks Joshua Welch and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at http://www.nature.com/reprints.



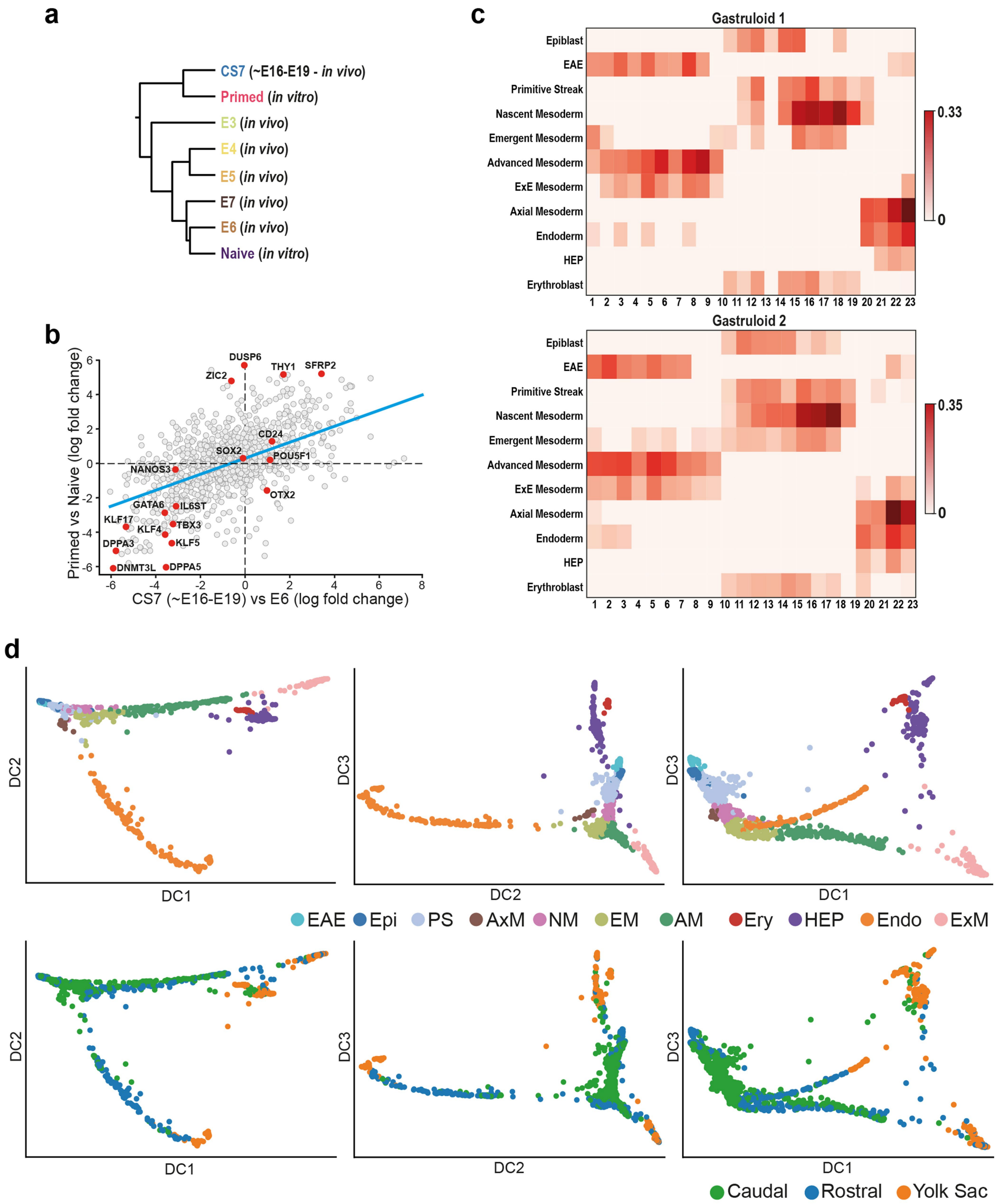
Extended Data Fig. 1 | See next page for caption.

Extended Data Fig. 1 | Quality control of scRNA-seq dataset. a, Dorsal view of the dissected embryonic disk showing the primitive streak and node (Scale bar = 500µm; n = 1). **b**, Brightfield images showing embryo dissection with schematic diagrams highlighting the three anatomical regions collected (yolk sac, rostral and caudal regions of embryonic disk; Scale bar = 500µm; n = 1). **c**, Metrics used to assess the quality of the scRNA-seq libraries. From top to bottom the scatter plots show the number of detected genes, the fraction of reads mapped to the human genome, the fraction of reads mapped to mitochondrial genes and the fraction of reads mapped to ERCC spike-ins, all as a function of the total number of reads. Cells that passed quality control are marked by green circles, while black circles indicate cells that failed the quality control and were excluded from downstream analyses. **d**, The boxplots show the total log expression of normalized counts for XIST and Y-genes across all clusters. While XIST was mostly not detected, Y-chromosome genes had always non-zero counts; this suggests that there is no contamination from maternal tissues in any of the clusters. n = 1195 cells were examined from a single embryo. Horizontal black lines denote median values and boxes cover the 25th and 75th percentiles range; whiskers extend to 1.5 × IQR. **e**, The stacked barplots indicate the percentages of cells from each cluster in the phase G1, S or G2/M of the cell cycle, as predicted from their transcriptomic profiles. **f**, Insertion-deletion length and size distribution of gastrula and fetal liver data. Y axis represents total number of indels on merged cells, while x axis represents indel length in base pairs. Hemato-Endothelial Progenitors (HEP), Endoderm (End), Advanced Mesoderm (AM), Primitive Streak (PS), Extraembryonic Mesoderm (ExM), Axial Mesoderm (AxM), Erythroblasts (Ery), Emergent Mesoderm (EM), Epiblast (Epi), Nascent Mesoderm (NM), Ectoderm (Amniotic/Embryonic (EAE)).



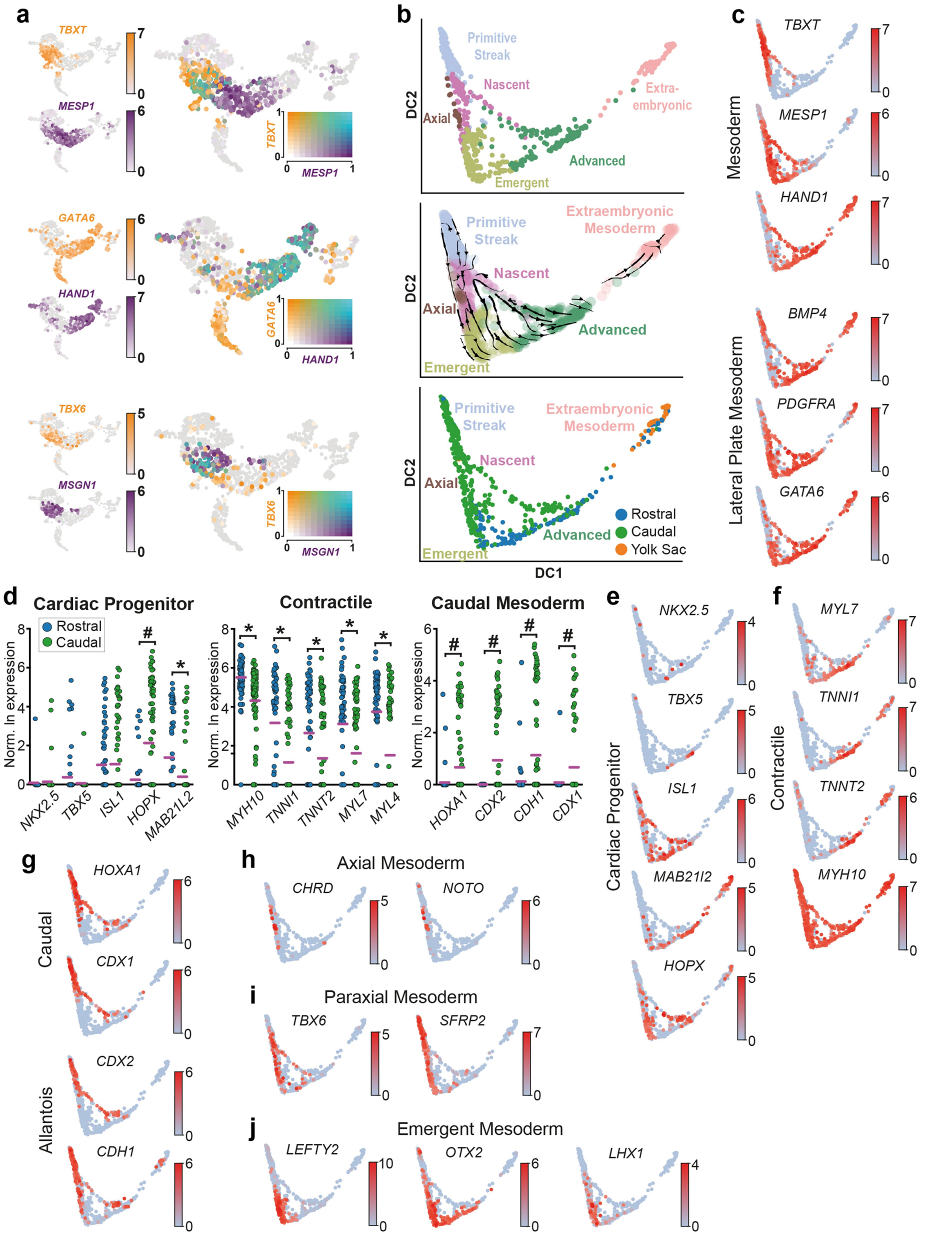
Extended Data Fig. 2 | See next page for caption.

Extended Data Fig. 2 | Characterisation and comparison of a CS7 human gastrula with Non-human primate and Mouse. a, Heatmap with the normalized log expression of well characterized marker genes for the identified cell types: Epiblast (Epi), Ectoderm (Amniotic/Embryonic (EAE)), Primitive Streak (PS), Nascent Mesoderm (NM), Emergent Mesoderm (EM), Advanced Mesoderm (AM), Extraembryonic Mesoderm (ExM), Axial Mesoderm (AxM), Endoderm (Endo), Hemato-Endothelial Progenitors (HEP), Erythroblasts (Ery). **b**, Stacked bar plots highlighting the anatomical region that cells were collected from and the percentage breakdown of each cluster. Numbers in brackets represent the total number of cells per cluster. **c**, Heatmap showing the fraction of human gastrula cells allocated to mouse cell types at E7.25 (data from[18]). **d**, Dendrogram showing hierarchical clustering of the transcriptomes of cell types from human gastrula and cultured cynomolgus macaque embryos at 16-day post-fertilization (from[19]). **e**, Top, UMAP plots showing the log expression of *MEST* and *GCNT2*. Bottom, violin plots showing the log expression of total transcripts (top row) and selected isoforms scaled by the maximum value in different cell types. Isoform names refer to Ensembl nomenclature.

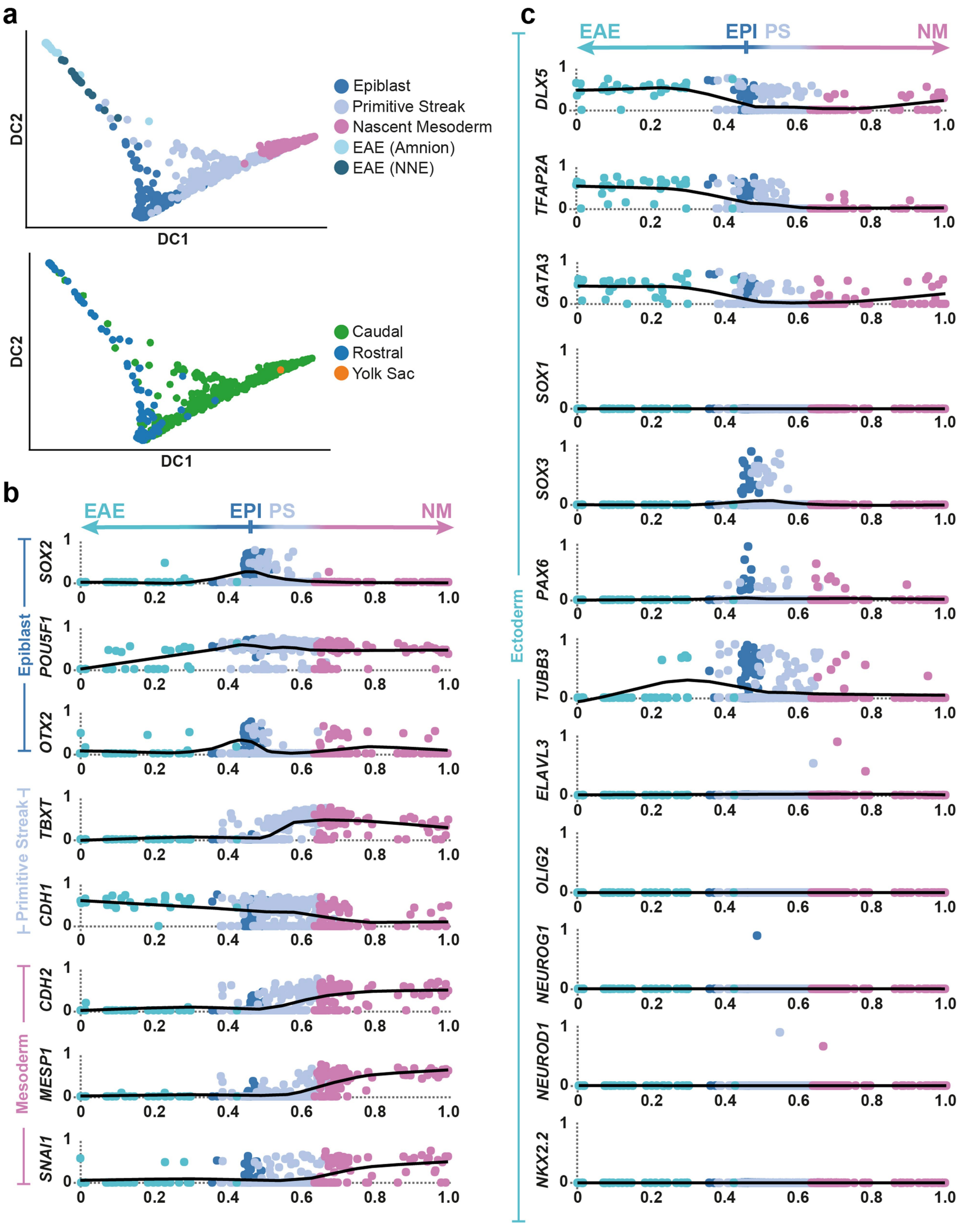


Extended Data Fig. 3 | See next page for caption.

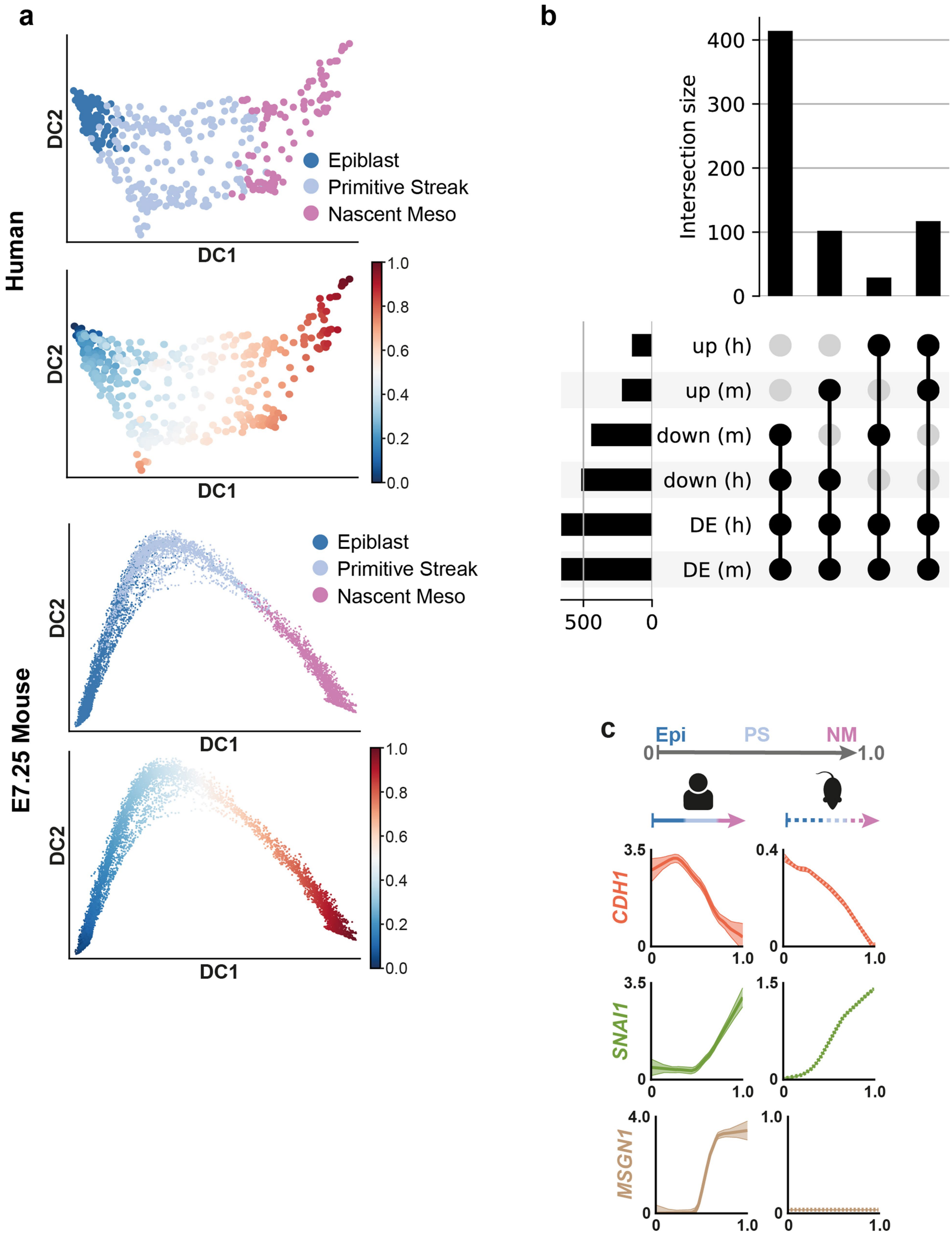
Extended Data Fig. 3 | *In Vitro* vs *In Vivo* comparisons. a, Dendrogram representation built on corrected expression values obtained with Seurat showing comparison of an *in vitro* model of pluripotency with *in vivo* data. **b**, Log-fold changes of expression levels of the genes between primed vs naïve hESC (y axis) and CS7 epiblast vs E6 data (x axis). Selected genes are highlighted in red; the blue line is obtained through a linear regression. A statistically significant positive correlation is found (Pearson's correlation coefficient ~0.63, p-value = 3e-107), indicating that the hESC resemble the *in vivo* primed and naïve states at the transcriptome-wide level. **c**, Heatmaps showing the correlations between the transcriptomic profiles of the human gastrula cell types (rows) and sections of human gastruloids taken at different positions along the rostral-caudal axis (columns) in two different replicates (Gastruloid 1 and Gastruloid 2). Only the values of the statistically significant correlations (p-value < 0.01; 2-tailed Pearson's correlation, see Methods) are reported, while all the non-significant correlations were set to 0. **d**, UMAP representation of the human gastrula data with the PGCs highlighted. d, Diffusion map of cells from all 11 clusters. The first three diffusion components (DC1, 2, 3) are plotted in different combinations. In the top panels, cells are coloured by the clusters they belong to, while in the bottom panels the colours indicate the region each cell was dissected from. Ectoderm (amniotic/embryonic) (EAE), Epiblast (Epi), Primitive Streak (PS), Axial Mesoderm (AxM), Nascent Mesoderm (NM), Emergent Mesoderm (EM), Advanced Mesoderm (AM), Erythroblasts (Ery), Hemato-Endothelial Progenitors (HEP), Endoderm (Endo), Extraembryonic Mesoderm (ExM).



Extended Data Fig. 4 | Rostral and Caudal differences in diversification of mesodermal subtypes. **a**, UMAP highlighting combinatorial gene expression. Individual gene expression (left) is reported as the log expression whilst combinatorial plots (right) show scaled log expression values. **b**, Diffusion map of cells from the 6 mesoderm related clusters (Primitive Streak, PS; Nascent Mesoderm, NM; Emergent Mesoderm, EM; Mesoderm, Meso; Axial Mesoderm, AxM; Extraembryonic Mesoderm, ExM), with the first and the second diffusion components plotted. **c**, Diffusion map of mesodermal showing the log expression levels of mesodermal markers genes. **d**, Differential gene expression between rostral and caudal advanced mesoderm cells. Significantly upregulated in rostral (*) or caudal (#) cells. **e**-**j**, Diffusion map of mesodermal clusters showing log expression levels of mesoderm subtype markers.

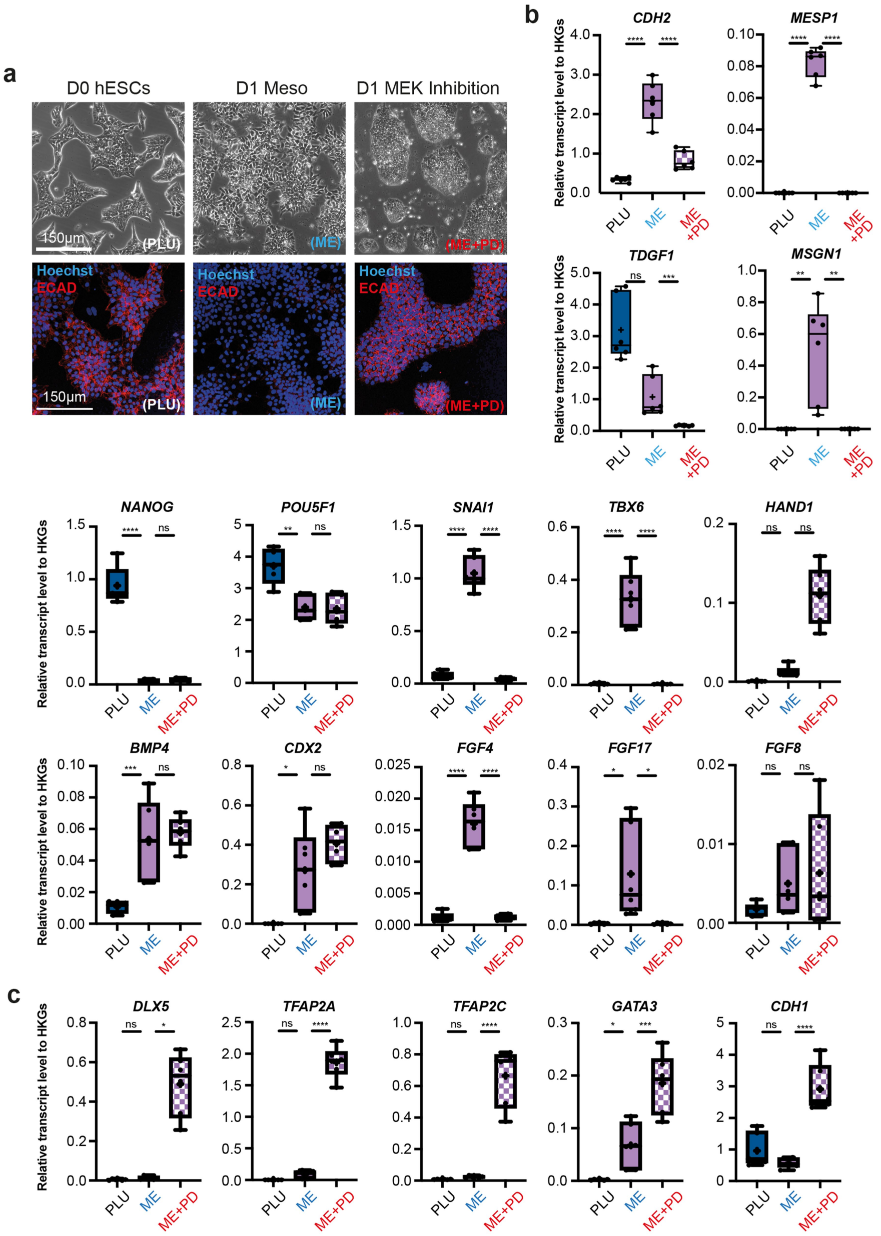


Extended Data Fig. 5 | Differentiation of the epiblast. a, Diffusion map of cells from the Epiblast, Primitive Streak, Nascent Mesoderm and Ectoderm (amniotic/embryonic). The first two diffusion components are plotted (DC1 and DC2) and cells are colored by their cluster (top) or the anatomical region they were isolated from (bottom). **b** and **c**, Normalized log gene expression changes along a pseudotime coordinate (see Fig. 4a) running from 0 to 1 and spanning the Ectoderm (amniotic/embryonic) (EAE), the Epiblast (EPI), the Primitive Streak (PS) and the Nascent Mesoderm (NM), as depicted by the arrow on top. The selected genes highlight Primitive Streak and mesoderm formation (panel b) as well as ectoderm differentiation (panel c).

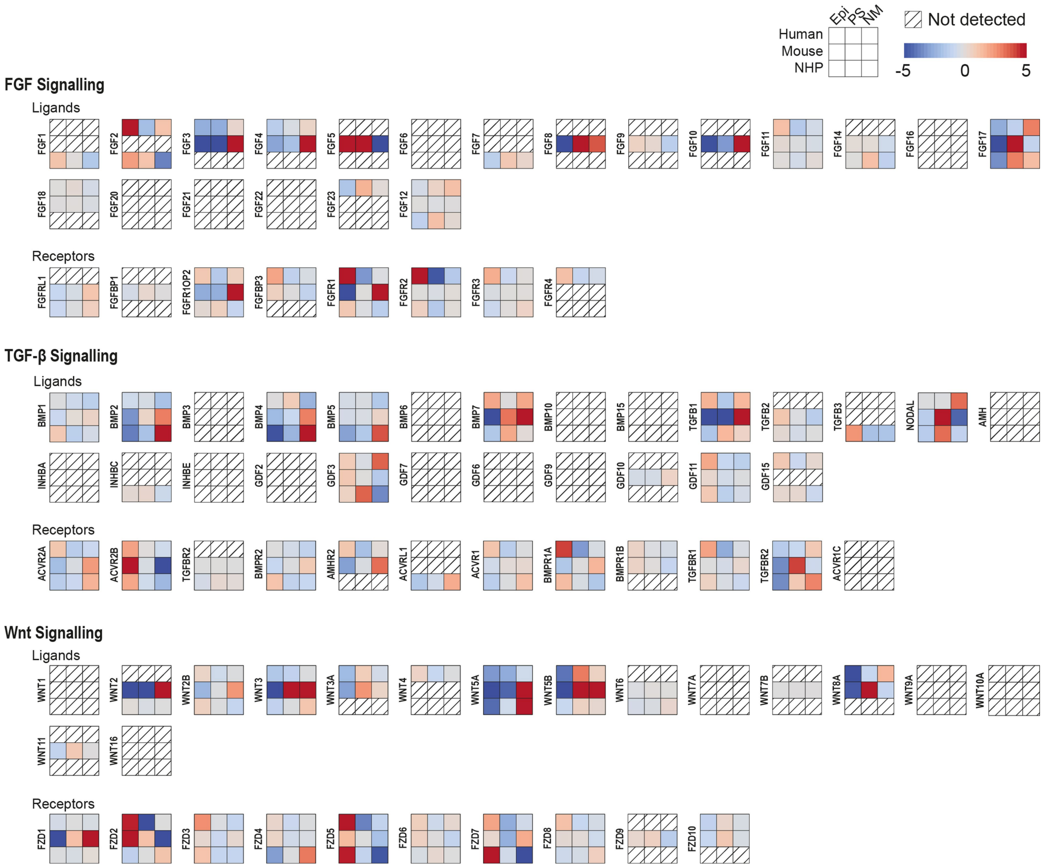


Extended Data Fig. 6 | See next page for caption.

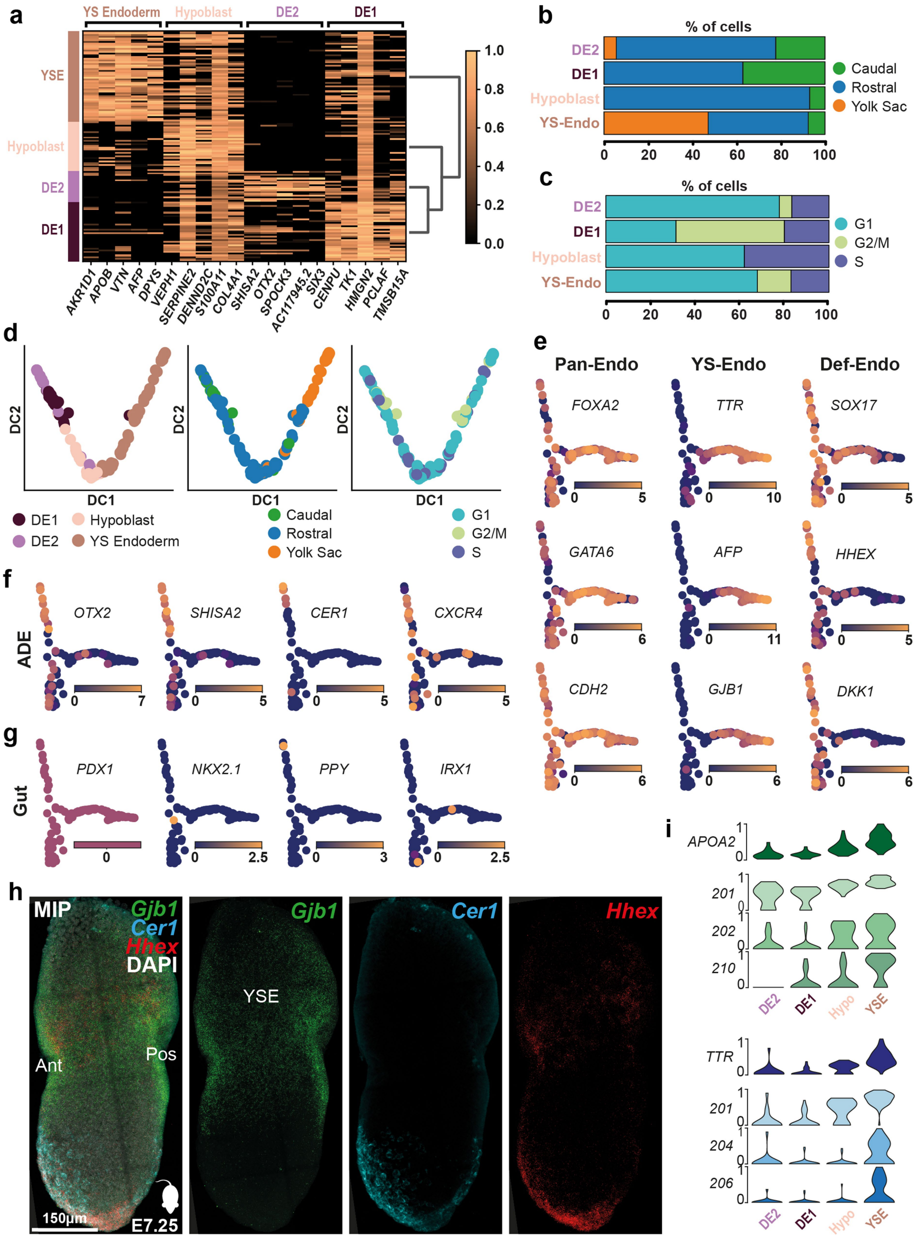
Extended Data Fig. 6 | Mesoderm formation in human and mouse. **a**, Diffusion map with cells from the human (top two plots) or mouse (bottom two plots) Epiblast, Primitive Streak and Nascent Mesoderm clusters. Cells are colored based on their cluster of origin or on their diffusion pseudotime coordinate. **b**, Upset plot for the number of differentially expressed (DE) genes as a function of the diffusion pseudotime (dpt) shown in panel a in mouse (m) or human (h). Here, only genes that are differentially expressed in both species and with a log-fold change > 1 along the trajectory are included. Genes are split according to their increasing (up) or decreasing (down) trend as a function of dpt. **c**, Comparison of pseudotime analysis during primitive streak and nascent mesoderm formation in human and mouse (data from[18]). Cells in epiblast (Epi), Primitive Streak (PS) and Nascent Mesoderm (NM) clusters from human and mouse embryos at matching stages (see Methods) were independently aligned along a differentiation trajectory and a diffusion pseudotime coordinate (dpt) was calculated for each (top). The expression pattern and standard error of the mean of selected genes along pseudotime is plotted for human (left, continuous lines) and mouse (right, dashed lines). Both *SNAI1* and *CDH1* showed comparable expression profiles during mesoderm formation in mouse and human whilst *MSGN1* was differently expressed between species.



Extended Data Fig. 7 | Characterization of EMT during hESC mesoderm formation. a, Bright-field microscopy images of D0 hESC (left), D1 Meso (center) and D1 MEK Inhibition (right) ESC colonies (top panels). Fluorescence microscopy images of E-Cadherin staining (bottom panels). **b**, Quantification of transcript levels for selected pluripotent, EMT and mesendoderm genes across the three conditions PLU, ME, ME+PD. **c**, Quantification of transcript levels for selected non-neural ectoderm genes across the three conditions PLU, ME, ME+PD. (n = 6 from three different experiments. Center line, median; box limits, upper and lower quartiles; whiskers, minimum and maximum; dots, mean value per experiement. ns = p-value ≥ 0.05; *** = p-value < 0.001; **** = p-value < 0.0001 (Ordinary one-way ANOVA after passing a Shapiro-Wilk normality test. Kruskal-Wallis multiple comparison test used if Shapiro-Wilk normality test failed (*MSGN1*, *TDGF1*, *HAND1*, *DLX5*). House-keeping genes, HKGs. See SI Table 17 for source data and exact p-values.

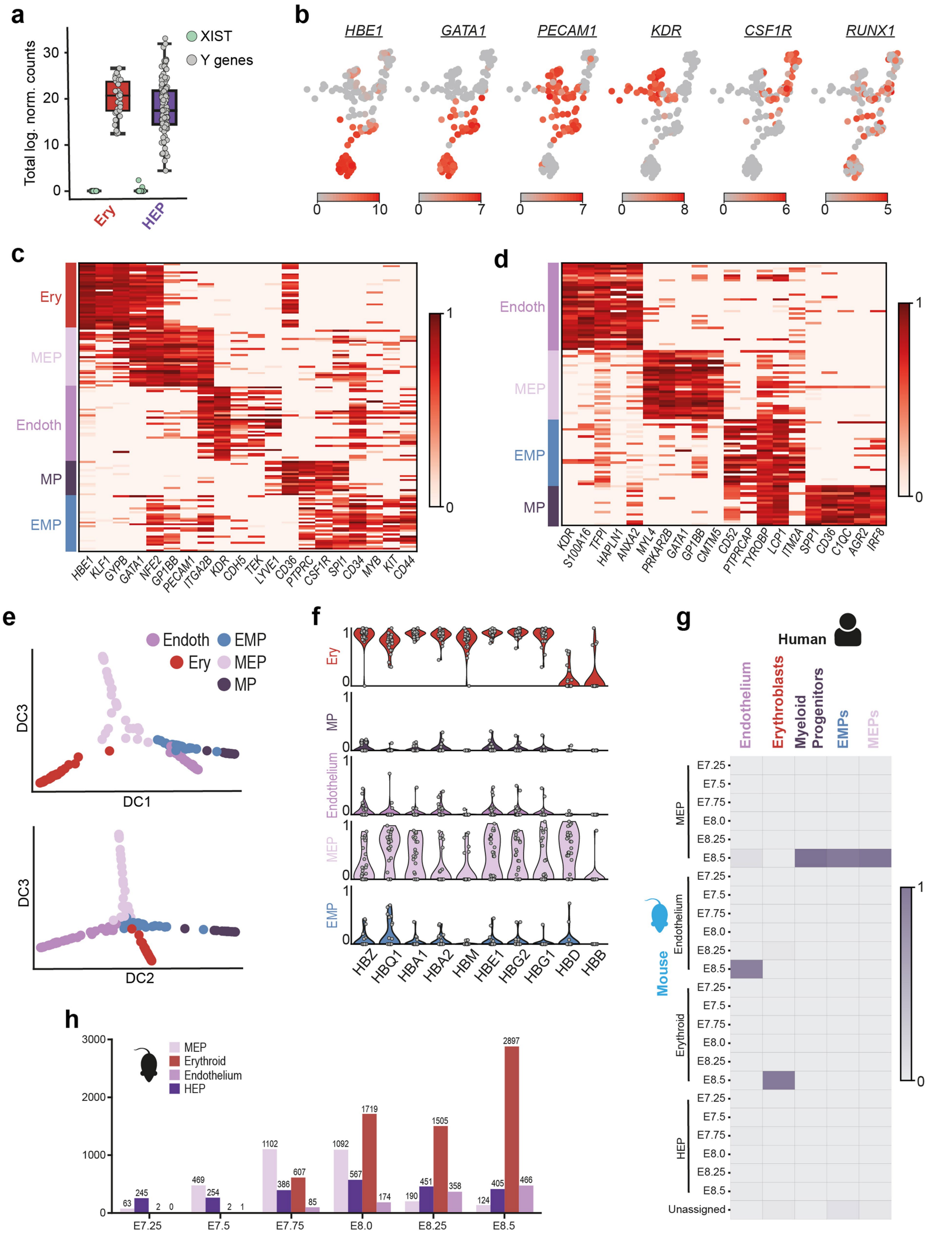


Extended Data Fig. 8 | Comparison of signaling during mesoderm formation in the human and mouse. Heatmap comparison of the z-score-normalized log expression values of components of FGF, TGF-β and Wnt signaling pathways in the human gastrula, mouse embryos (E7.25 stage) and cultured cynomolgus macaque embryos (16 d.p.f stage). From human and mouse we considered the Epiblast (Epi), Primitive Streak (PS) and Nascent Mesoderm (NM) clusters; in the macaque, we used the clusters annotated as postL-Epi, L-Gast1 and L-Gast2.



Extended Data Fig. 9 | See next page for caption.

Extended Data Fig. 9 | Endoderm subcluster identification. a, Heatmap showing the scaled log expression levels of marker genes of the four endodermal subclusters. **b**, Percentage of cells dissected from the Caudal, Rostral or Yolk Sac portion of the embryo in the four endodermal subclusters. **c**, Percentage of cells based on their predicted cell-cycle phase of the four endodermal subclusters. **d**, Diffusion map of cells from the Endoderm cluster. The first two diffusion components (DC1 and DC2) are plotted and cells are coloured by the sub clusters (left), anatomical origin (central) or the predicted cell-cycle phase (right). Yolk Sac, YS; Definitive Endoderm (DE) 1 and 2. **e**, Diffusion map of cells from the Endoderm cluster with DC1 and DC3 plotted, showing log expression levels of Pan-endoderm, Yolk-sac endoderm and definitive endoderm markers. **f**, Log expression levels of Anterior Definitive Endoderm markers. These genes are more highly expressed in DE2. **g**, Log expression levels of Gut Endoderm markers, showing limited expression. **h**, Maximum intensity projection and mid-sagittal section (h') of an E7.0 mouse embryo showing expression of *Gjb1* (yolk sac endoderm marker) as well as *Cer1* and *Hhex* (anterior definitive endoderm markers) using Hybridization Chain Reaction (n = 4). *Cer1* and *Hhex* show greater expression in the anterior embryonic endoderm. Anterior, Ant; Posterior, Pos; Yolk-sac Endoderm, YSE. **i**, Violin plots showing the scaled log expression of total transcripts (top row) and individual isoforms in different endodermal subclusters. Isoform lables refer to Ensembl transcript numbers.



Extended Data Fig. 10 | See next page for caption.

Extended Data Fig. 10 | Hemato-Endothelial Progenitors subclusters. **a**, Boxplots showing the total log expression of normalized counts for XIST and Y-genes in Erythroblasts (Ery) and Hemato-Endothelial Progenitors (HEP), indicating no contamination from maternal tissue. n = 143 cells were examined from a single embryo. Horizontal black lines denote median values and boxes cover the 25th and 75th percentiles range; whiskers extend to 1.5 × IQR. **b**, UMAP of HEP and Erythroblast clusters showing log expression of blood related marker genes. **c**, Heatmap showing the scaled log expression of well-characterized marker genes for both the Hemato-Endothelial Progenitors subclusters and Erythroblast cluster. **d**, Heatmap showing the normalized log expression levels of the top 5 marker genes of the four Hemato-Endothelial Progenitors subclusters. **e**, Diffusion maps of HEP subclusters and Erythroblasts showing diffusion components (DC) 1, 2 and 3. **f**, Violin plots showing the scaled log expression of Globin genes in the five blood related clusters: Erythroblasts (Ery), Myeloid Progenitors (MP), Endothelium, Megakaryocyte-Erythroid Progenitors (MEP) and Erythro-Myeloid progenitors (EMP). Each grey dot represents a single cell. **g**, Heatmap showing the estimated mapping of human Erythroid and HEP subclusters to mouse blood-related clusters. Scalebar represents the fraction of human cells mapped to each category. **h**, Bar graph showing the number of cells present in the mouse scRNA-seq dataset[18] at different development timepoints.

nature research

Corresponding author(s): Shankar Srinivas, Antonio Scialdone

Last updated by author(s): Sep 1, 2021

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	No software was used for data collection
Data analysis	We used existing published sequence analysis packages, as detailed in the methods. Including biomaRt R, velocyto v0.17.17, RaceID package v0.1.5, pypairs' v3.1.1, 'transIndel' v0.1, DESeq2 v3.11, Seurat v3.0 and Salmon v0.17.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The raw data from our study can be downloaded from ArrayExpress under accession code: E-MTAB-9388. The processed data may be downloaded from www.human-gastrula.net. Datasets used as references include; Mouse gastrula data: E-MTAB-6967; Pre-implantation embryo data: E-MTAB-3929; ENSEMBL database (GRCh38.p13); MSigDB database; Signaling Database CellPhoneDB.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	The sample size of this study was dictated by the availability of human CS7 embryos. Embryos at this stage of human development are extremely rare thereby significantly impacting sample size.
Data exclusions	Disaggregated cells from the embryo were excluded on the basis of a Live/Dead stain (described in Methods), to select for live cells. Post sequencing quality control (detailed in manuscript) was used to exclude poor quality sequenced cells.
Replication	Due to the rarity and difficulty in collecting human embryos at this stage of development replication was not possible.
Randomization	Given this study characterises a single human embryo randomisation was not necessary.
Blinding	Given this study characterises a single human embryo blinding was not required.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study	
☐	☒	Antibodies
☐	☒	Eukaryotic cell lines
☒	☐	Palaeontology and archaeology
☐	☒	Animals and other organisms
☐	☒	Human research participants
☒	☐	Clinical data
☒	☐	Dual use research of concern

Methods

n/a	Involved in the study	
☒	☐	ChIP-seq
☒	☐	Flow cytometry
☒	☐	MRI-based neuroimaging

Antibodies

Antibodies used	Anti-Ecadherin antibody (3195, Cell Signaling Technology, 1:200) and anti-Rabbit IgG Alexa Fluoro 568 antibody (A10042, Invitrogen, 1:1000) were used in this study
Validation	Anti-Ecadherin antibody has been validated by western blot analysis and cited 1686 times (https://www.cellsignal.co.uk/products/primary-antibodies/e-cadherin-24e10-rabbit-mab/3195).

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)	Human ESCs (H9/WA09 line; WiCell), Vallier Laboratory
Authentication	Validated using fingerprinting
Mycoplasma contamination	Confirmed negative
Commonly misidentified lines (See ICLAC register)	No commonly misidentified cell lines were used in the study

Animals and other organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research

Laboratory animals	Wild-type Mus musculus embryos were obtained using C57BL/6 males crossed with 8-16 week old CD1 females.
Wild animals	This study did not use any wild animals.
Field-collected samples	This study did not use any field-collected samples.
Ethics oversight	All animal experiments complied with the UK Animals (Scientific Procedures) Act 1986, approved by the local Biological Services Ethical Review Process and were performed under UK Home Office project licenses PPL 30/3420 and PCB8EF1B4.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Human research participants

Policy information about studies involving human research participants

Population characteristics	A single CS7 human embryo.
Recruitment	The sample was collected by the Human Developmental Biology Resource (HDBR - https://www.hdbr.org/general-information). The material was collected after appropriate informed written consent from the donor, by medical termination.
Ethics oversight	HDBR has approval from the UK National Research Ethics Service (London Fulham Research Ethics Committee (18/LO/0822) and the Newcastle and North Tyneside NHS Health Authority Joint Ethics Committee (08/H0906/21+5)) to function as a Research Tissue Bank for registered projects. The HDBR is monitored by The Human Tissue Authority (HTA) for compliance with the Human Tissue Act (HTA; 2004). This work was done as part of project #200295 registered with the HDBR.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

6.2 Publication 2

nature metabolism
ARTICLES
https://doi.org/10.1038/s42255-021-00422-7

Check for updates

Cell competition acts as a purifying selection to eliminate cells with mitochondrial defects during early mouse development

Ana Lima[1,2,14], Gabriele Lubatti[3,4,5,14], Jörg Burgstaller[6], Di Hu[7], Alistair P. Green[8], Aida Di Gregorio[1], Tamzin Zawadzki[1], Barbara Pernaute[1,9], Elmir Mahammadov[3,4,5], Salvador Perez-Montero[1], Marian Dore[2], Juan Miguel Sanchez[1,10], Sarah Bowling[1], Margarida Sancho[1], Thomas Kolbe[11,12], Mohammad M. Karimi[2,13], David Carling[2], Nick Jones[8], Shankar Srinivas[7], Antonio Scialdone[3,4,5,15] ✉ and Tristan A. Rodriguez[1,15] ✉

Cell competition is emerging as a quality-control mechanism that eliminates unfit cells in a wide range of settings from development to the adult. However, the nature of the cells normally eliminated by cell competition and what triggers their elimination remains poorly understood. In mice, 35% of epiblast cells are eliminated before gastrulation. Here we show that cells with mitochondrial defects are eliminated by cell competition during early mouse development. Using single-cell transcriptional profiling of eliminated mouse epiblast cells, we identify hallmarks of cell competition and mitochondrial defects. We demonstrate that mitochondrial defects are common to a range of different loser cell types and that manipulating mitochondrial function triggers cell competition. Moreover, we show that in the mouse embryo, cell competition eliminates cells with sequence changes in mt-Rnr1 and mt-Rnr2, and that even non-pathological changes in mitochondrial DNA sequences can induce cell competition. Our results suggest that cell competition is a purifying selection that optimizes mitochondrial performance before gastrulation.

Cell competition is a fitness-sensing mechanism that eliminates cells that, although viable, are less fit than their neighbours. The cells that are eliminated are generically termed losers, while the fitter cells that survive are referred to as winners. Cell competition has been shown to act in a broad range of settings, from the developing embryo to the ageing organisms[1–3]. It has been primarily studied in *Drosophila*, where it was first described in the imaginal wing disc[4]. Since then, it has also been found to be conserved in mammals. In the mouse embryo, 35% of embryonic cells are eliminated between embryonic day (E) 5.5 and E6.5, and strong evidence suggests that this elimination is through cell competition[5–7]. These and other studies identified a number of read-outs of cell competition in the mouse embryo, such as relative low c-MYC expression, a loss of mTOR (mammalian target of rapamycin) signalling, low TEAD transcription factor activity, high P53 expression or elevated levels of ERK phosphorylation[5–9]. Importantly, there is a substantial overlap with the markers of cell competition originally identified in *Drosophila* as well as those found in other cell competition models, such as Madin–Darby canine kidney cells[1–3]. Despite the advance that having these cell competition markers signifies, given that they were primarily identified by using genetic models that rely on over-expression or mutation, we still have little insight into the overarching features of the cells that are eliminated in the physiological context.

Mitochondria, with their diverse cellular functions ranging from determining the bioenergetic output of the cell to regulating its apoptotic response, are strong candidates for determining competitive cell fitness. During early mouse development, mitochondria undergo profound changes in their shape and activity[10]. In the pre-implantation embryo, mitochondria are rounded, fragmented and contain sparse cristae, but after implantation they fuse to form complex networks with mature cristae[11]. The mode of replication of mitochondrial DNA (mtDNA), which encodes vital components of the bioenergetic machinery, also changes during early mouse development. After fertilization, mtDNA replication ceases and its copy number per cell decreases with every division until the post-implantation stages, when mtDNA replication resumes[10]. As the mutation rate of mtDNA is much higher than that of nuclear DNA[12,13], this increased replication most likely leads to an increased mutation load. In fact, inheritable mtDNA-based diseases are reported with a prevalence of 5–15 cases per 100,000 individuals[14,15]. A number of mechanisms have been proposed to reduce this mutation load, such as the bottleneck effect, purifying selection or biased segregation of mtDNA haplotypes[16–21]. However, how these

---

[1]National Heart and Lung Institute, Imperial College London, London, UK. [2]MRC London Institute of Medical Sciences (LMS), Institute of Clinical Sciences, Imperial College London, London, UK. [3]Institute of Epigenetics and Stem Cells, Helmholtz Zentrum München, Munich, Germany. [4]Institute of Functional Epigenetics, Helmholtz Zentrum München, Neuherberg, Germany. [5]Institute of Computational Biology, Helmholtz Zentrum München, Neuherberg, Germany. [6]Institute of Animal Breeding and Genetics, University of Veterinary Medicine, Vienna, Austria. [7]Department of Physiology, Anatomy and Genetics, University of Oxford, Oxford, UK. [8]EPSRC Centre for the Mathematics of Precision Healthcare, Department of Mathematics, Imperial College London, London, UK. [9]Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology, Barcelona, Spain. [10]Orchard Therapeutics, London, UK. [11]Biomodels Austria (Biat), University of Veterinary Medicine Vienna, Vienna, Austria. [12]Department IFA-Tulln, University of Natural Resources and Life Sciences, Vienna, Austria. [13]Comprehensive Cancer Centre, School of Cancer & Pharmaceutical Sciences, Faculty of Life Sciences & Medicine, King's College London, London, UK. [14]These authors contributed equally: Ana Lima, Gabriele Lubatti. [15]These authors jointly supervised this work: Antonio Scialdone, Tristan A. Rodriguez. ✉e-mail: antonio.scialdone@helmholtz-muenchen.de; tristan.rodriguez@imperial.ac.uk

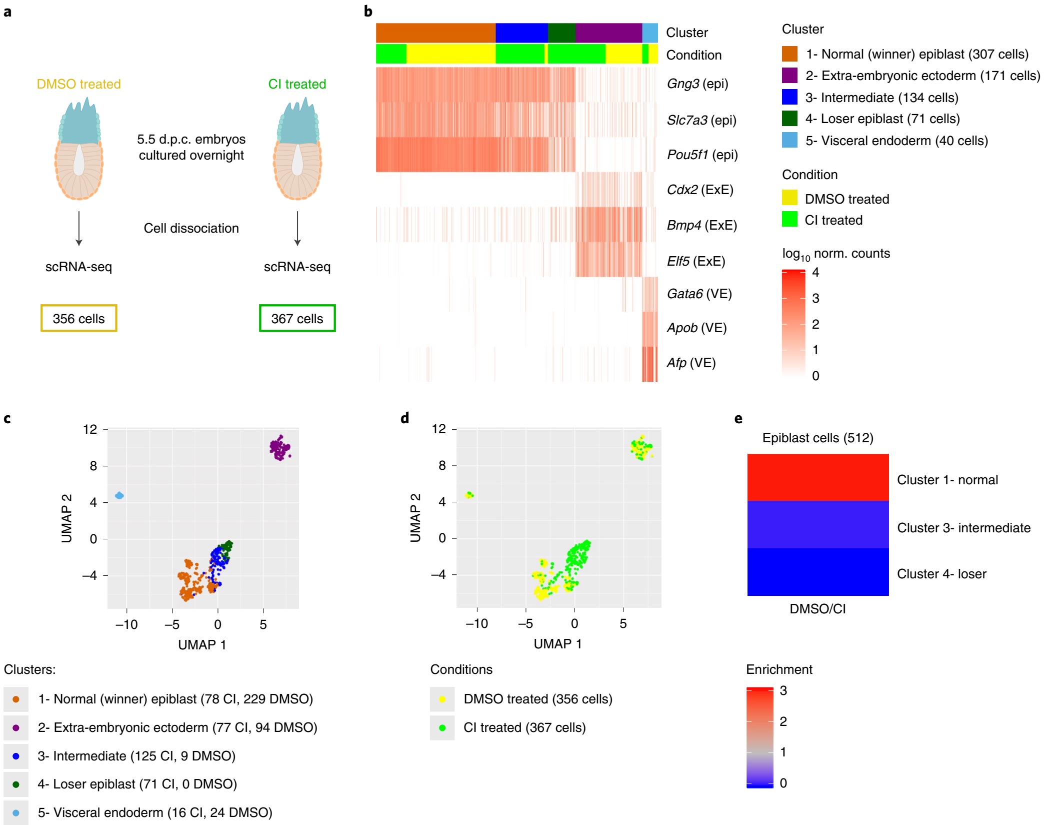


Fig. 1 | Cells eliminated during early mouse embryogenesis have a distinct transcriptional profile. a, Experimental design. The number of cells in the two conditions (DMSO treated and CI treated) refers to the cells that passed the quality control. d.p.c., days post-coitum. **b**, Identification of the clusters according to known gene markers from the different embryonic regions[73]. Three clusters (clusters 1, 3 and 4) show marker genes of the epiblast (Epi), while the remaining clusters correspond to the extra-embryonic visceral endoderm (VE; cluster 5) and extra-embryonic ectoderm (ExE; cluster 2). The epiblast clusters were named 'winner', 'intermediate' and 'loser' on the basis of the relative fraction of cells from CI-treated embryos they include (**e**). **c**,**d**, Uniform manifold approximation projection (UMAP) visualization of the single-cell RNA-seq data, with cells coloured according to cluster (**c**) or condition (**d**). A region made up exclusively by cells from CI-treated embryos emerged. **e**, Ratio between the fraction of cells from DMSO-treated and CI-treated embryos in the three epiblast clusters. While the 'winner' epiblast cluster shows an enrichment of cells from DMSO-treated embryos, the 'intermediate' and the 'loser' epiblast clusters are strongly enriched for cells from CI-treated embryos.

mechanisms act at the molecular and cellular level is still poorly understood.

To understand the nature of the cells eliminated during early mouse post-implantation development, we have analysed their transcriptional profile by single-cell RNA sequencing (scRNA-seq) and found that these cells share a cell competition signature. Analysis of the mis-regulated pathways identified mitochondrial dysfunction as a common feature. Importantly, our studies also found evidence of mtDNA mutations in the eliminated cells. Furthermore, we demonstrate that manipulating mitochondrial activity by either disrupting mitochondrial dynamics or introducing non-pathological mtDNA changes is sufficient to trigger cell competition. Therefore, these results pinpoint mitochondrial performance as a key cellular feature that determines the competitive ability of embryonic cells and suggest that cell competition is acting as a purifying selection during early mammalian development.

Results

Loser cells have a distinct transcriptional profile. We have previously shown that in the early post-implantation mouse embryo about 35% of epiblast cells are eliminated and that these cells are marked by low mTOR signalling[7]. However, we currently do not understand the characteristics of these cells or what triggers their elimination. To answer these questions, we have analysed their transcriptional profile with scRNA-seq. To ensure the eliminated cells can be captured, as we have done before[7], we isolated embryos at E5.5 and cultured them for 16 h in the presence of a caspase inhibitors (CIs) or vehicle (DMSO) (Fig. 1a). Unsupervised clustering of the scRNA-seq data revealed five clusters: two corresponding to extra-embryonic tissues (visceral endoderm and extra-embryonic ectoderm) and three that expressed epiblast marker genes (Fig. 1b,c, Extended Data Fig. 1a–f and Methods). Interestingly, cells from CI-treated and DMSO-treated embryos were unequally distributed across the three epiblast clusters. In particular, one of these clusters (cluster 4) was only composed of cells from CI-treated embryos (Fig. 1d,e). Also notable is that all epiblast clusters contained cells in the G2/M and S phases of the cell cycle, suggesting they are all cycling (Extended Data Fig. 2a).

The three epiblast clusters are highly connected, as highlighted by a connectivity analysis carried out with PAGA[22] (Extended Data Fig. 2b). Hence, to establish the relationship between these epiblast clusters, we computed a diffusion map[23]. For this, we selected only cells captured from CI-treated embryos, to eliminate possible confounding effects due to the CI (Fig. 2a). However, when all epiblast cells were considered, the results remain unchanged (Extended Data Fig. 2c–e). This analysis identified a trajectory between the three epiblast clusters, with those cells unique to CI-treated embryos falling at one extreme end of the trajectory (corresponding to cluster 4; Fig. 2a) and with those cells present in both DMSO-treated and CI-treated embryos at the other (corresponding to cluster 1; Fig. 2a and Extended Data Fig. 2d).

To further define the identity of the epiblast cells of CI-treated embryos, we analysed the genes differentially expressed along the trajectory (Methods and Extended Data Fig. 3a) using ingenuity pathway analysis (IPA) to characterize gene signatures[24]. Importantly, we found that these differentially expressed genes fell under molecular and cellular function categories associated with cell death and survival, protein synthesis and nucleic acids (Fig. 2b). Analysis of the factors with enriched targets within the genes differentially expressed along the trajectory revealed RICTOR (an mTOR component), TLE3, MYC, MYCN, P53 and IGFR (that is, upstream of mTOR) as the top upstream regulators (Fig. 2c). Breaking down the differentially expressed genes into those downregulated or upregulated along the winner-to-loser trajectory revealed that the targets of RICTOR, MYC, MYCN and IGFR primarily fell within the downregulated genes (Supplementary Tables 1 and 2). P53-activated targets were preferentially upregulated and P53-repressed targets were preferentially downregulated (Extended Data Fig. 3b,c). Moreover, genes related to protein synthesis were primarily found to be downregulated.

The observation that the genes differentially expressed along the trajectory fall into cell death categories, as well as being mTOR, MYC and P53 targets, strongly suggests that cells at each end of the trajectory are the winners and losers of cell competition[5–7]. For this reason, we hereafter refer to those epiblast cells unique to CI-treated embryos as 'loser' epiblast cells and to those at the opposite end of the trajectory as the 'winner' epiblast cells. Those cells lying between these two populations on the trajectory are considered 'intermediate'. Using this knowledge, we can define a diffusion pseudotime (dpt) coordinate[25] originating in the 'winner' cluster that tracks the position of cells along the trajectory and that can be interpreted as a 'losing score'; that is, it quantifies how strong the signature of the 'losing' state is in the transcriptome of a cell (Fig. 2d,e).

In accordance with previous studies[6,8,9], we also found evidence for miss-patterning in the eliminated epiblast cells, as a proportion of these cells co-expressed naïve pluripotency and differentiation markers (Fig. 2f and Extended Data Fig. 3d). To test if loser cells are developmentally delayed or advanced compared to control cells, we projected our data onto a previously published diffusion map that includes epiblast cells from E5.5, E6.25 and E6.5 embryos[26]. We found that all epiblast cells, irrespective of the condition in which the embryos were cultured (that is, treated with DMSO or CI) and of their losing state (that is, that they belonged to the winner, intermediate or loser cluster), mostly overlapped with E6.5 epiblast cells (Extended Data Fig. 3e–g). Cells from the loser cluster were slightly closer to the E6.25 stage than the winner and intermediate cells, as shown by their pseudotime coordinate, but they remain far from the earlier E5.5 stage. This result, combined with the higher expression of some differentiation markers observed in loser cells, suggests that these cells are miss-patterned rather than developmentally delayed.

Loser cells have defects in mitochondrial function. Using IPA, we next analysed the cellular pathways mis-regulated in loser epiblast cells and found that the top two pathways (mitochondrial dysfunction and oxidative phosphorylation (OXPHOS)) were related to mitochondrial function (Fig. 3a,b and Supplementary Tables 1 and 2). For example, we found a downregulation along the winner-to-loser trajectory of the mtDNA-encoded subunits *mt-Nd3* and *mt-Atp6*, of regulators of mitochondrial dynamics such as *Opa1* (optic atrophy 1), as well as of genes involved in mitochondrial membrane and cristae organization such as *Samm50* (Fig. 3c), suggesting that mitochondrial function is impaired in loser cells.

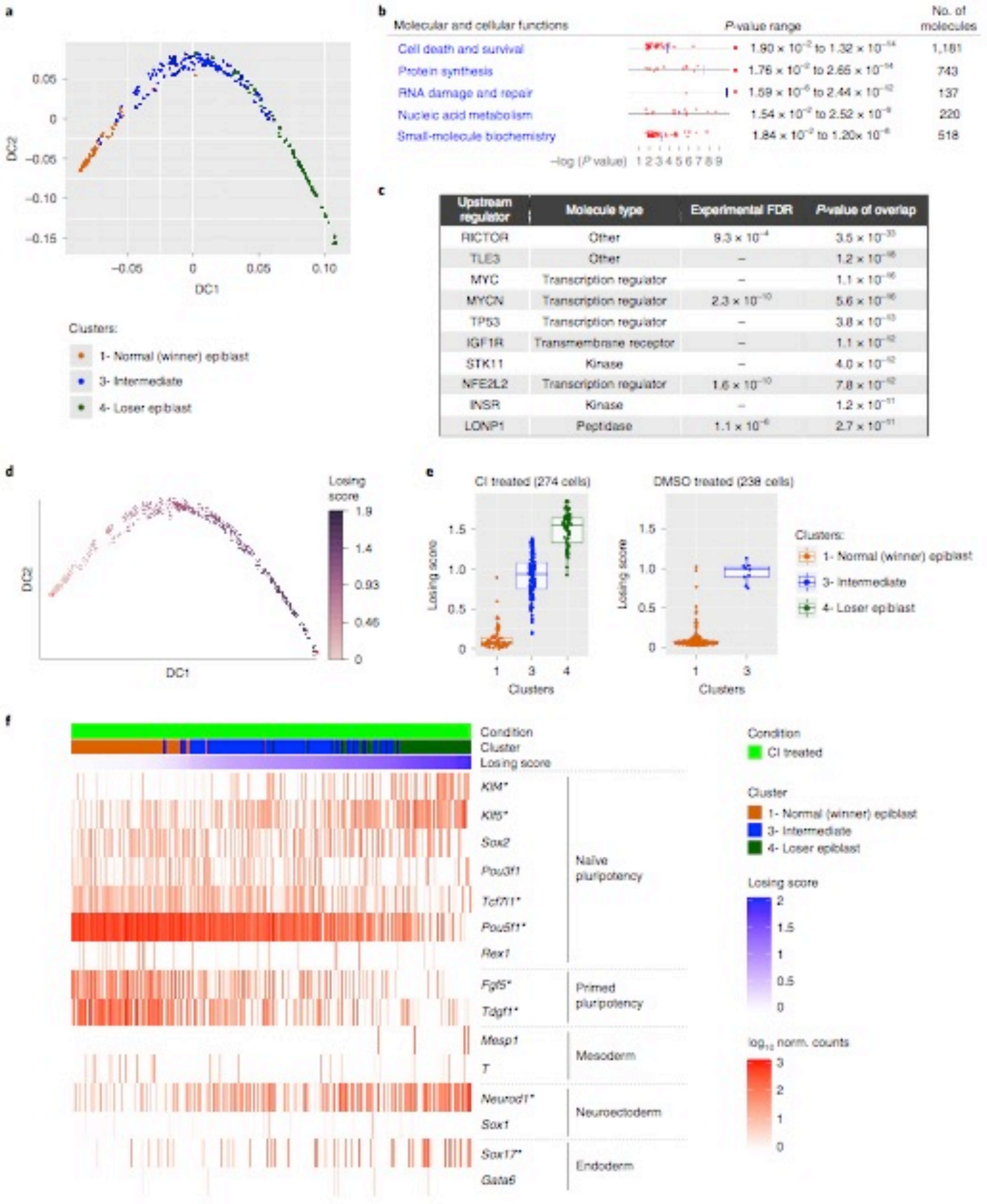
A recent body of evidence has revealed that stress responses, such as the integrated stress response (ISR) or the closely related unfolded protein response (UPR), when triggered in cells with impaired mitochondrial function prompt a transcriptional programme to restore cellular homeostasis[27–29]. We observed that loser epiblast cells displayed a characteristic UPR/ISR signature[30–33] and key regulators of this response, such as *Atf4*, *Ddit3*, *Nfe2l2* (*Nrf2)* and *Foxo3* were all upregulated in these cells (Extended Data Fig. 4a–d). Similarly, *Sesn2*, a target of p53 that controls mTOR activity[34], was also upregulated in loser cells (Extended Data Fig. 4d). These findings support that loser epiblast cells present mitochondrial defects, leading to the activation of a stress response in an attempt to restore cellular homeostasis[35].

To validate the significance of the observed mitochondrial defects, we did two things: First, we asked if the changes in expression of mitochondrial regulators at the mRNA level are also reflected at the protein level. We observed that in CI-treated embryos, loser cells that persist and are marked by low mTOR activity[7] also show significantly lower OPA1 levels (Fig. 3d–f). We also found that DMSO-treated embryos showed strong DDIT3 staining (an UPR/ISR marker also known as CHOP) in the dying cells that accumulate in the pro-amniotic cavity, and that in CI-treated embryos, DDIT3 expression was upregulated in a proportion of epiblast cells (Extended Data Fig. 4e–g). Second, we studied the mitochondrial membrane potential (Δψm), an indication of mitochondrial health, in loser epiblast cells. We observed that while the cells of DMSO-treated embryos showed a high Δψm that fell within a narrow range, in CI-treated embryos the proportion of cells with a low Δψm significantly increased (Fig. 3d,g,h). Together, these results

Fig. 2 | A cell competition transcriptional signature is identified in cells eliminated during mouse embryonic development. a, Diffusion map of epiblast cells (only from CI-treated embryos), coloured by cluster. **b**,**c**, IPA of genes differentially expressed along the diffusion trajectory (Extended Data Fig. 3a) generated lists of the top five molecular and cellular functions (**b**) and upstream regulators (**c**) found to be differentially activated in epiblast cells along the diffusion trajectory from winner (cluster 1) to loser status (cluster 4). **d**, Diffusion map of epiblast cells (only from CI-treated embryos) coloured by the dpt coordinate. The winner and the loser clusters were found at the two extremities of the trajectory, hence the dpt coordinate can be interpreted as a 'losing score'. **e**, Losing score of the cells in the three epiblast clusters in CI-treated (left) or DMSO-treated (right) embryos. The losing score of the cells from DMSO-treated embryos was obtained by projecting them on the diffusion map shown in **d** (Methods). **f**, Expression levels in epiblast cells from CI-treated embryos of genes (in rows) that are markers for naïve pluripotency (*Klf4*, *Klf5*, *Sox2*, *Pou3f1*, *Tcf7l1*, *Pou5f1* and *Zfp42* (*Rex1*)), primed pluripotency (*Fgf5* and *Tdgf1*), mesoderm (*Mesp1* and *T*), neuroectoderm (*Neurod1* and *Sox1*) and endoderm (*Sox17 and Gata6*). Cells (columns) were sorted by their losing scores. The genes marked with an asterisk were differentially expressed along the trajectory. See Methods for details on statistical analysis.

suggest that loser epiblast cells have impaired mitochondrial activity that triggers a stress response.

Mitochondrial dysfunction is common to different loser cells. To address if mitochondrial defects are a common feature of loser cells eliminated by cell competition, we analysed embryonic stem cells (ESCs) that are defective for bone morphogenetic protein (BMP) signalling (*Bmpr1a*$^{-/-}$) and tetraploid cells (4n)[8]. We first carried out a mass spectrometry analysis using the Metabolon platform and found that metabolites and intermediates of the



Molecular and cellular functions	P-value range	No. of molecules
Cell death and survival	1.90×10^{-2} to 1.32×10^{-14}	1,181
Protein synthesis	1.76×10^{-2} to 2.65×10^{-14}	743
RNA damage and repair	1.59×10^{-6} to 2.44×10^{-12}	137
Nucleic acid metabolism	1.54×10^{-2} to 2.52×10^{-9}	220
Small-molecule biochemistry	1.84×10^{-2} to 1.20×10^{-8}	518

Upstream regulator	Molecule type	Experimental FDR	P-value of overlap
RICTOR	Other	9.3×10^{-4}	3.5×10^{-33}
TLE3	Other	–	1.2×10^{-18}
MYC	Transcription regulator	–	1.1×10^{-16}
MYCN	Transcription regulator	2.3×10^{-10}	5.6×10^{-16}
TP53	Transcription regulator	–	3.8×10^{-13}
IGF1R	Transmembrane receptor	–	1.1×10^{-12}
STK11	Kinase	–	4.0×10^{-12}
NFE2L2	Transcription regulator	1.6×10^{-10}	7.8×10^{-12}
INSR	Kinase	–	1.2×10^{-11}
LONP1	Peptidase	1.1×10^{-6}	2.7×10^{-11}

tricarboxylic acid (TCA) cycle, such as malate, fumarate, glutamate and α-ketoglutarate are depleted in both *Bmpr1a*$^{-/-}$ and 4n ESCs in differentiation culture conditions (Fig. 4a). Next, we performed an extracellular flux Seahorse analysis of *Bmpr1a*$^{-/-}$ ESCs to measure their glycolytic and OXPHOS rates. We observed that when these cells are maintained in pluripotency culture conditions that are not permissive for cell competition[6], they exhibit a higher OXPHOS rate than control cells (Extended Data Fig. 5a,b). In contrast, when *Bmpr1a*$^{-/-}$ cells are induced to differentiate, this phenotype is reversed, with mutant cells showing lower ATP generated through OXPHOS and a higher glycolytic capacity than controls (Fig. 4b–e and Extended Data Fig. 5c,d). This suggests that after differentiation *Bmpr1a*$^{-/-}$ cells are unable to sustain proper OXPHOS activity.

To further test the possibility that defective mouse ESCs (mESCs) have impaired mitochondrial function, we assessed their Δψm. We found that whilst *Bmpr1a*$^{-/-}$ and 4n cells had a similar Δψm to control cells in pluripotency conditions (Extended Data Fig. 5e,f), following differentiation both these cell types presented a loss of Δψm, irrespective of whether they were separate or co-cultured with wild-type cells (Fig. 4f,g). This reduction in Δψm is not due to excessive mitochondrial reactive oxygen species (ROS) production or to a lower mitochondrial mass within mutant cells because, as for example, *Bmpr1a*$^{-/-}$ cells had lower ROS levels and similar TOMM20 and mt-CO1 expression to control cells (Fig. 4h–j and Extended Data Fig. 5g). The fact that the loss of Δψm and lower OXPHOS activity can be observed even when loser cells are cultured separately suggests that the mitochondrial dysfunction phenotype is an inherent property of loser cells and not a response to them being out-competed. These results also indicate that the mitochondrial defects are directly linked to the emergence of the loser status: In conditions that are not permissive for cell competition (pluripotency), mutant cells do not show defective mitochondrial function, but when they are switched to differentiation conditions that allow for cell competition, they display impaired mitochondrial function.

To further explore the relationship between mitochondrial activity and the competitive ability of the cell, we analysed the Δψm of BMP-defective cells that are null for p53 (*Bmpr1a*$^{-/-}$;*p53*$^{-/-}$ ESCs), as these are not eliminated by wild-type cells[7]. Remarkably, we observed that mutating *p53* in *Bmpr1a*$^{-/-}$ cells not only rescues the loss of Δψm of these cells, but also causes hyperpolarization of their mitochondria (Fig. 4k). These results suggest a role for P53 in regulating mitochondrial activity of ESCs and strongly support a pivotal role for mitochondrial activity in cell competition.

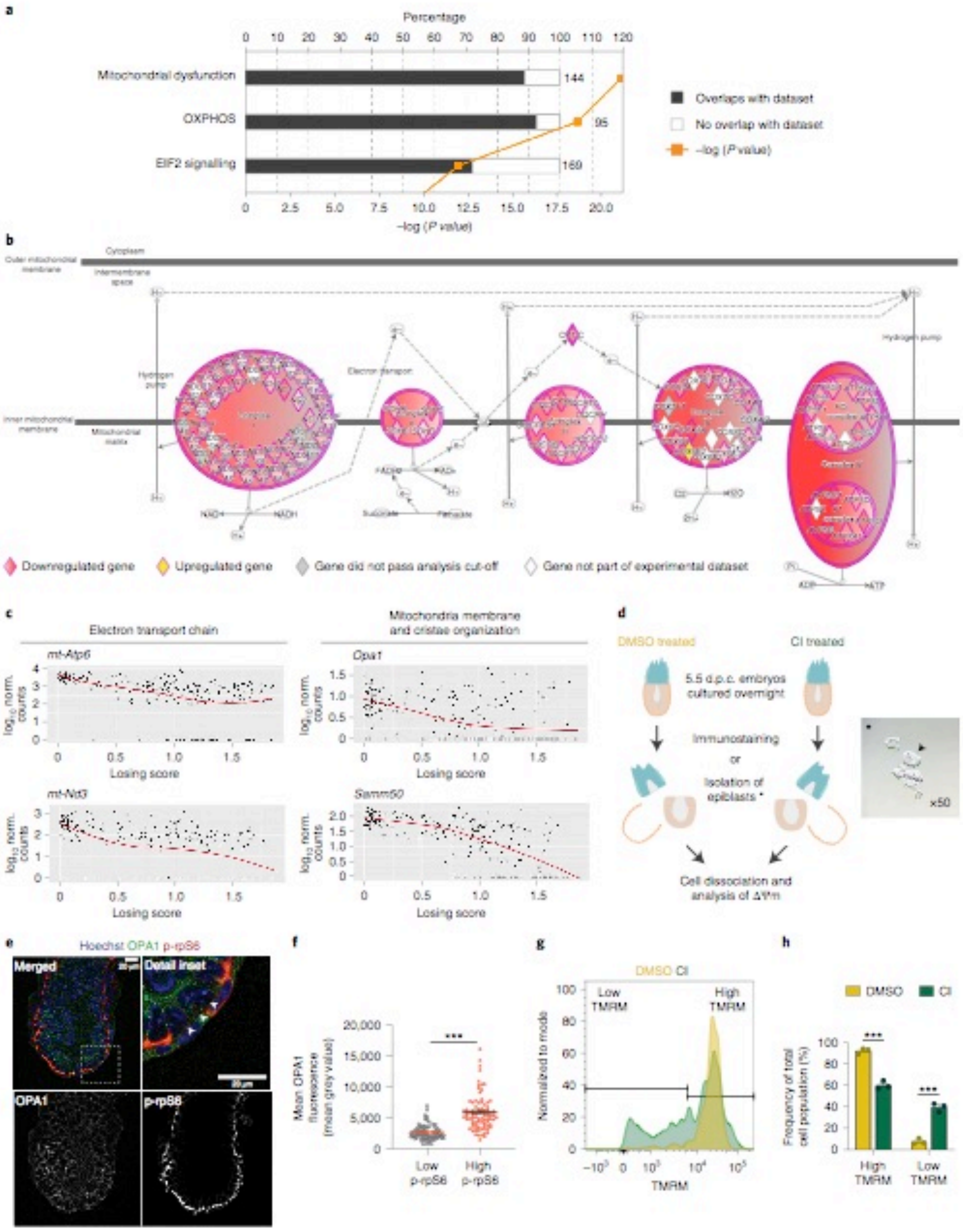
Impaired mitochondrial function triggers cell competition. The mitochondrial defects observed in loser cells led us to ask if disrupting mitochondrial activity alone is sufficient to trigger cell competition. During the onset of differentiation, mitochondrial shape changes substantially. In pluripotent cells, mitochondria have a round and fragmented shape, but after differentiation they fuse and become elongated, forming complex networks[10]. Given that this change in shape correlates with when cell competition occurs, we tested if disrupting mitochondrial dynamics is sufficient to induce cell competition. MFN1 and MFN2 regulate mitochondrial fusion and DRP1/DNM1L controls their fission[36–38]. We generated ESCs null for mitofusin 2 (*Mfn2*$^{-/-}$), which have enlarged globular mitochondria, and ESCs null for dynamin-related protein 1 (*Drp1*$^{-/-}$), which show hyper-elongated mitochondria (Fig. 5a). We first tested the competitive ability of *Mfn2*$^{-/-}$ ESCs in pluripotency conditions, which we have previously found not to induce out-competing in *Bmpr1a*$^{-/-}$ or 4n cells[6]. Interestingly, we found that although *Mfn2*$^{-/-}$ cells grow similarly to wild-type cells in separate cultures, they were out-competed in co-culture (Fig. 5b). Analogously, the *Drp1* mutant cells did not grow significantly slower than wild-type cells when cultured separately in differentiation-inducing conditions, but they were out-competed by wild-type cells in co-culture (Fig. 5c). The observation that disrupting mitochondrial dynamics can induce cell competition even in pluripotency culture conditions, suggests that mitochondrial activity is a dominant parameter determining the competitive ability of the cell.

To establish how disruption of mitochondrial fusion and fission affects mitochondrial performance, we compared the Δψm, respiration rates and mitochondrial ATP production of *Mfn2*$^{-/-}$ and *Drp1*$^{-/-}$ ESCs to those of wild-type cells (Fig. 5d–g). We found that whilst *Mfn2*$^{-/-}$ and *Drp1*$^{-/-}$ ESCs had lower Δψm than control cells (Fig. 5d,f), *Mfn2*$^{-/-}$ ESCs had lower maximal respiration rates but similar basal respiration and ATP production to controls, and *Drp1*$^{-/-}$ ESCs showed similar respiration and ATP production to controls (Fig. 5e,g). This suggests that ATP production or respiration rates alone do not determine the relative competitive ability of ESCs.

Besides mitochondrial dysfunction, another prominent signature of loser cells found in vivo was the UPR/ISR (Extended Data Fig. 4). Because the loss of *Drp1* has been associated with activation of the UPR[39–41], we investigated if the *Drp1*$^{-/-}$ loser cells also showed evidence for the activation of the UPR/ISR. We observed that *Drp1*$^{-/-}$ cells show higher expression of ATF4 and phosphorylated eukaryotic initiation factor 2α (p-eIF2α) than their wild-type counterparts, which is indicative of UPR/ISR activation (Fig. 5h)[39–41]. Another feature previously described following loss of *Drp1* is the proteolytic cleavage of OPA1, where short isoforms (S-OPA1) are accumulated in detriment of the long isoforms (L-OPA1)[39]. When we analysed the expression of OPA1 in wild-type and *Drp1*$^{-/-}$ cells, we observed that while wild-type cells retained L-OPA1 expression, loser cells predominantly expressed the S-OPA1 isoforms and

Fig. 3 | Cells eliminated during early mouse embryogenesis have mitochondrial defects. a, Top canonical pathways, identified by IPA, mis-regulated in loser cells in comparison to normal epiblast cells. The numbers at the end of each bar refer to total amount of genes involved in that pathway. The percentage refers to the number of genes found mis-regulated in loser cells relative to the number total genes within each pathway. **b**, Details of changes in the OXPHOS pathway identified in **a**. Circular and oval shapes represent each of the electron transport chain (ETC) complexes (complexes I to V). Diamond shapes represent subunits of each ETC complex. Downregulated genes in loser cells are coloured in shades of red. Darker shades correspond to lower false discovery rate (FDR) values. *Cox6b2*, in yellow, was upregulated in loser cells. Grey denotes genes that were not differentially expressed between loser and winner cells (FDR > 0.01). White denotes genes from the Knowledge Base that were not tested (for example, because they were not detected in our dataset). **c**, Expression levels of mitochondrial genes as a function of the losing score of cells. **d**, Experimental design adopted to assess mitochondrial function in **e**–**h**. The asterisk indicates a representative micrograph of one of the isolated epiblasts (arrow) used for Δψm analysis after embryo microdissection. **e**, Representative immunohistochemistry of OPA1 in E6.5 embryos where cell death was inhibited (CI treated), quantified in **f**. Loser cells were identified by low mTOR activation (low p-rpS6; arrowheads). Scale bar, 20 µm. **f**, Quantification of OPA1 fluorescence in normal epiblast cells and loser cells. *N* = 6 embryos with a minimum of 8 cells analysed per condition. **g**, Representative histogram of flow cytometry analysis of tetramethylrhodamine methyl ester (TMRM) probe, indicative of Δψm, in epiblast cells from embryos where cell death was allowed (DMSO treated) or inhibited (CI treated), quantified in **h**. **h**, Frequency of epiblast cells with high or low TMRM fluorescence, according to the range defined in **g** from embryos where cell competition was allowed (DMSO treated) or inhibited (CI treated). Data were obtained from three independent experiments and are shown as the mean ± s.e.m. (**g** and **h**). Twelve embryos per condition were pooled for each experiment. See Methods for details on statistical analysis.

displayed almost no expression of L-OPA1 (Fig. 5i). This defect has been associated with mito-ribosomal stalling, a phenotype that can be replicated by treating cells with actinonin (Extended Data Fig. 6)[42]. To test if the shift in isoform expression observed in *Drp1*$^{-/-}$ ESCs is due to aberrant mitochondrial translation, we treated cells with doxycycline, which inhibits translation in mitochondria[43], and



observed that this was sufficient to partially rescue L-OPA1 expression (Fig. 5j). This rescue together with the evidence for UPR/ISR activation suggests that *Drp1*$^{-/-}$ cells display defects in mitochondrial translation.

Loser epiblast cells accumulate mtDNA mutations. There is strong evidence for selection against aberrant mitochondrial function induced by deleterious mtDNA mutations in mammals[21,44–47]. Given our observation that cell competition selects against cells with impaired mitochondrial function, we asked if cell competition could be reducing mtDNA heteroplasmy (frequency of different mtDNA variants) during mouse development. It has been recently shown that scRNA-seq can be used to reliably identify mtDNA variants, although with a lower statistical power compared to more direct approaches, like mtDNA sequencing[48]. We therefore tested if mtDNA heteroplasmy is present in our scRNA-seq data and whether this correlates with the losing score of a cell. Our analysis revealed that the frequency of specific mtDNA polymorphisms increased with the losing score of epiblast cells (Fig. 6a), and such mtDNA changes occurred within *mt-Rnr1* and *mt-Rnr2* (Fig. 6b–h and Extended Data Fig. 7a–e). Moreover, these changes were not dependent on the litter from which the embryo came from (Extended Data Fig. 7f–k). As it was formally possible that these loser-specific sequence changes could originate from contaminating nuclear mitochondrial sequences (NUMTS) or from RNA editing, we performed several controls to confirm that mtDNA polymorphisms are the most likely source of these changes (Methods). For example, we considered only the RNA-seq reads that are uniquely mapped to the mitochondrial genome and not to nuclear DNA, and we confirmed that the variants with highest heteroplasmy found in the ‘loser’ cells were not present in any of the NUMTS that have previously been reported or could be identified using BLAST. Moreover, we verified that the observed sequence changes were not compatible with canonical RNA editing (Methods). It is worth noting that the sequence changes we detected in *mt-Rnr1* and *mt-Rnr2* strongly co-occurred in the same cell, with those closest together having the highest probability of coexisting (Fig. 6i and Extended Data Fig. 7l). This is suggestive of mtDNA replication errors that could be ‘scarring’ the mtDNA, disrupting the function of *mt-Rnr1* (12S rRNA) and *mt-Rnr2* (16S rRNA) and causing the loser phenotype. Importantly, the presence of these specific mtDNA mutations in the loser cells suggests that cell competition could be contributing to the elimination of deleterious mtDNA mutations during early mouse development. Of note, we only report mtDNA variants detected in regions of the genome with high sequencing coverage (Extended Data Fig. 7m); therefore, the presence of other variations in mtDNA sequences between winner and loser cells cannot be excluded.

mtDNA sequence determines the competitive ability of a cell. To explore this possibility further, we analysed if alterations in mtDNA can induce cell competition by testing the competitive ability of ESCs with non-pathological differences in mtDNA sequence. For this we compared the relative competitive ability of ESCs that shared the same nuclear genome background but differed in their mitochondrial genomes by a small number of non-pathological sequence changes. We derived ESCs from hybrid mouse strains that we had previously engineered to have a common nuclear C57BL/6N background, but mtDNAs from different wild-caught mice[16]. Each wild-derived mtDNA variant (or haplotype) contains a specific number of single-nucleotide polymorphisms (SNPs) that lead to a small number of amino acid changes when compared to the C57BL/6N mtDNA haplotype. Furthermore, these haplotypes (BG, HB and ST) can be ranked according to their genetic distance from the C57BL/6N mtDNA (Fig. 7a and Extended Data Fig. 8a). Characterization of the isolated ESCs revealed that they have a range of heteroplasmy (mix of wild-derived and C57BL/6N mtDNAs) that is stable over several passages (Extended Data Fig. 8b). Importantly, these different mtDNA haplotypes and different levels of heteroplasmy do not alter cell size, cell granularity, mitochondrial mass or mitochondrial dynamics, nor do they substantially impact the cell’s $\Delta\psi$m (Extended Data Fig. 8c–f).

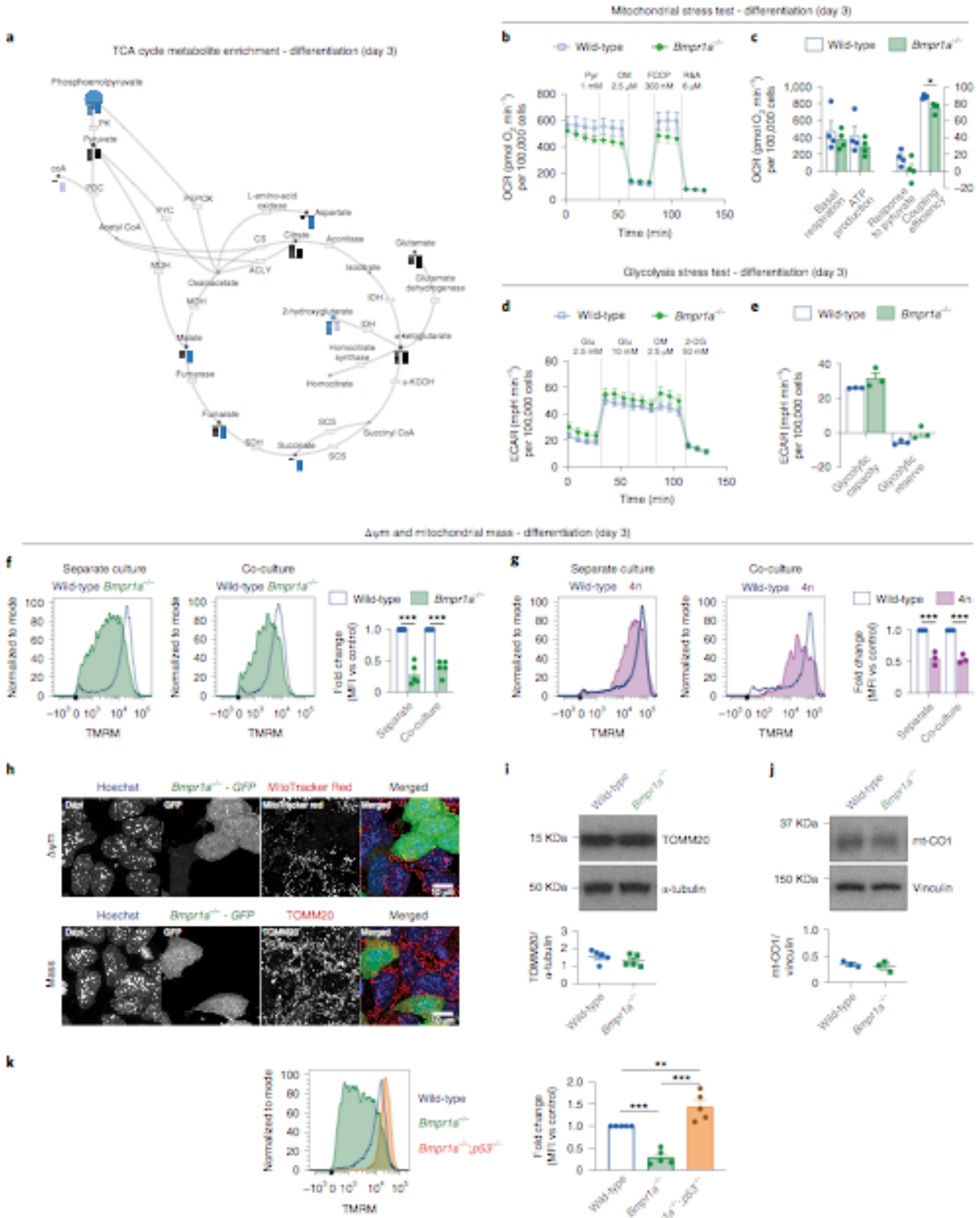
When we tested the competitive ability of these ESCs with different mtDNA content, in pluripotency culture conditions, we observed that cells carrying the mtDNAs that were most distant from the C57BL/6N mtDNA, such as the HB (100%), the HB (24%) and the ST (46%) ESCs could all out-compete the C57BL/6N line (Fig. 7b,c and Extended Data Fig. 8g). Similarly, when we tested the HB (24%) line against the BG (99%) or the BG (95%) lines (which have mtDNAs more closely related to the C57BL/6N mtDNA), we found that cells with the HB haplotype could also out-compete these ESCs (Fig. 7d and Extended Data Fig. 8h). In contrast, we observed that the HB (24%) ESCs were unable to out-compete their homoplasmic counterparts, HB cells (100%) or the ST cells (46%) that carry the most distant mtDNA variant from C57BL/6N (Fig. 7e and Extended Data Fig. 8i). These results tell us three things: First, non-pathological differences in mtDNA sequence can trigger cell competition. Second, a competitive advantage can be conferred by only a small proportion of mtDNA content, as indicated by our finding that HB (24%) behave as winners. Finally, these findings suggest that the phylogenetic proximity between mtDNA variants can potentially determine their competitive cell fitness.

To characterize the mode of competition between cells with different mtDNA, we focused on the HB (24%) and the BG (95%) ESCs. Analysis of these cell lines revealed that specifically when co-cultured, the BG (95%) cells displayed high levels of apoptosis (Fig. 7f), indicating that they are out-competed through their

→

Fig. 4 | Mitochondrial defects are a common feature of cells eliminated by cell competition. **a**, Metabolic enrichment analysis of the TCA cycle and intermediate metabolites obtained using Metabolon platform for defective cells (*Bmpr1a*$^{-/-}$, left bar; 4n, right bar), in comparison to wild-type cells during differentiation. Bars indicate compound levels relative to wild-type cells. Blue bars indicate compounds that were significantly altered ($P < 0.05$), and light-blue bars indicate compounds that were almost significantly altered ($0.05 \leq P \leq 0.1$). Black bars indicate compounds that were altered although not statistically significant in comparison to the levels found in wild-type cells. The enzymes on the pathway are represented as boxes and labelled by their canonical names. **b**–**e**, Metabolic flux analysis of wild-type and BMP-defective cells during differentiating conditions. Analysis of OCR as a measure of mitochondrial function (mitochondrial stress test; **b**). Details of metabolic parameters found changed from the analysis of the mitochondrial stress test (**c**). Analysis of extracellular acidification rate (ECAR) as a measure of glycolytic function (glycolysis stress test; **d**). Details of metabolic parameters found changed from the analysis of the glycolysis stress test (**e**). **f**,**g**, $\Delta\psi$m in defective mESCs undergoing differentiation in separate or co-culture conditions. Representative histograms of TMRM fluorescence and quantification for wild-type and *Bmpr1a*$^{-/-}$ (**f**) and wild-type and 4n (**g**) cells. **h**, Representative micrographs of wild-type and *Bmpr1a*$^{-/-}$ cells co-cultured during differentiation and stained for a reporter of $\Delta\psi$m (MitoTracker Red; top) or mitochondrial mass (TOMM20; bottom). Nuclei were stained with Hoechst. Scale bar, 10 µm. **i**,**j**, Western blot analysis of mitochondrial mass markers TOMM20 (**i**) and mt-CO1 (**j**) for wild-type and *Bmpr1a*$^{-/-}$ cells during differentiation. **k**, Analysis of $\Delta\psi$m for wild-type, *Bmpr1a*$^{-/-}$ and *Bmpr1a*$^{-/-}$;*p53*$^{-/-}$ cells during differentiation. Representative histogram of TMRM fluorescence and quantification. Data are the mean ± s.e.m. Extracellular flux Seahorse data were obtained from three (**d** and **e**) or four (**b** and **c**) independent experiments, with five replicates per cell type in each assay. The remaining data were obtained from three (**g** and **j**) or five (**a**,**f**,**i** and **k**) independent experiments. See Methods for details on statistical analysis. MFI, mean fluorescence intensity.

elimination. To gain further insight, we performed bulk RNA-seq of these cells in separate and co-culture conditions (Extended Data Fig. 8j) and analysed the differentially expressed genes by gene-set enrichment analysis (GSEA). We found that in separate culture the most notable features that distinguished BG (95%) from HB (24%) cells were a downregulation of genes involved in OXPHOS



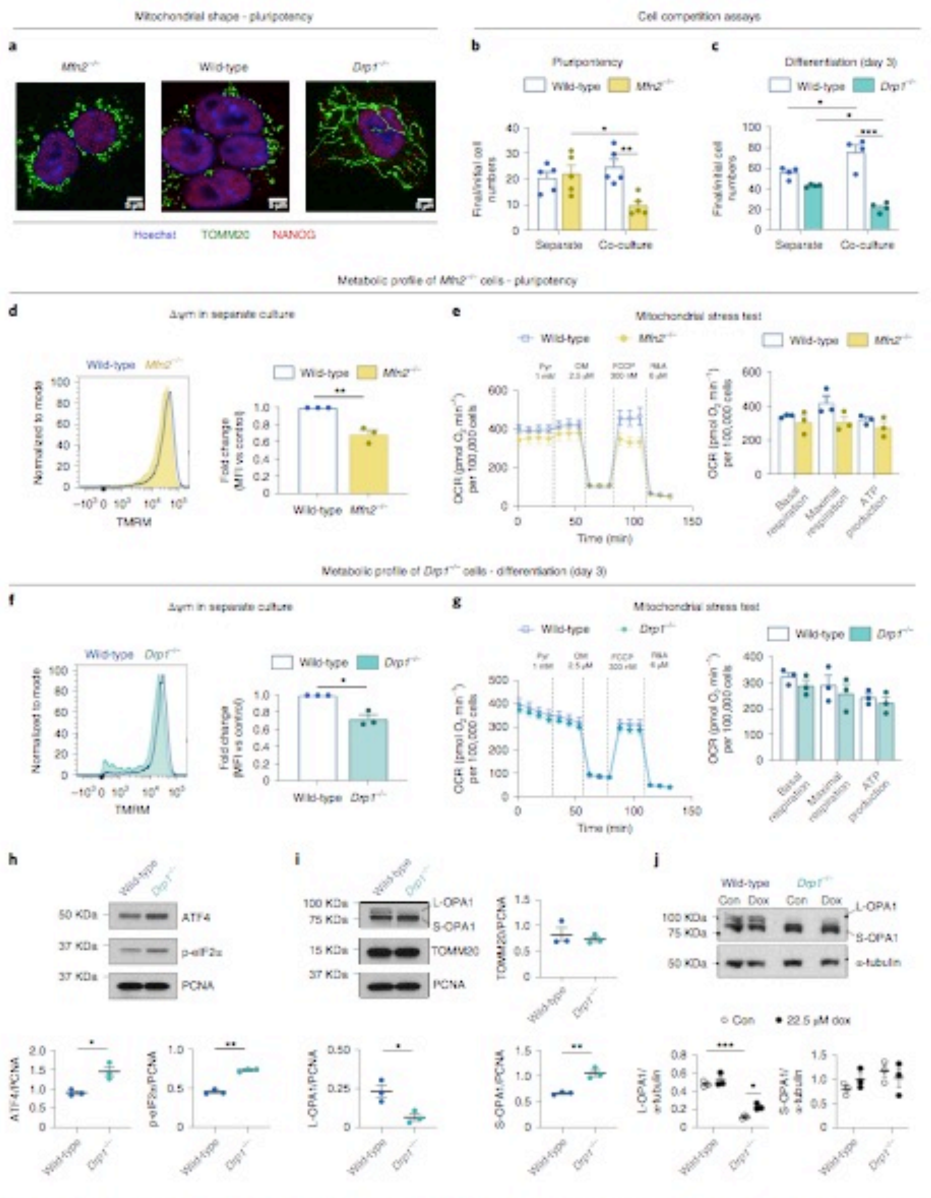


Fig. 5 | Manipulating mitochondria biology is sufficient to trigger cell competition. **a**, Representative micrographs of wild-type, *Mfn2*$^{-/-}$ and *Drp1*$^{-/-}$ mESCs showing alterations in mitochondrial morphology in mutant cells. TOMM20 was used as a mitochondrial marker and NANOG as a pluripotency marker. Nuclei were stained with Hoechst. Scale bar, 5 µm. **b**,**c**, Cell competition assays between wild-type mESCs and cells with altered morphology: *Mfn2*$^{-/-}$ during pluripotency (**b**) and *Drp1*$^{-/-}$ during differentiation (**c**). The ratio of final/initial cell numbers cultured separately or in co-culture is shown. **d**–**j**, Metabolic profile of *Mfn2*$^{-/-}$ and *Drp1*$^{-/-}$ mESCs. Analysis of mitochondrial Δψm for wild-type and *Mfn2*$^{-/-}$ cells cultured separately during pluripotency (**d**) and for wild-type and *Drp1*$^{-/-}$ mESCs $^{-/-}$ during differentiation in a separate culture (**f**). Metabolic flux analysis of wild-type and *Mfn2*$^{-/-}$ mESCs cultured separately during pluripotency (**e**) and for wild-type and *Drp1*$^{-/-}$ undergoing differentiation in separate cultures (**g**). Data were collected from three independent experiments. **h**–**j**, Western blot analysis of markers of UPR and mitochondrial markers in wild-type and *Drp1*$^{-/-}$ during differentiation in separate culture. Cells were treated with doxycycline (Dox, 22.5 µM) or vehicle (Con) from day 1 of differentiation and samples were collected on day 3 (**j**). Data are the mean ± s.e.m. of three (**d**–**j**), four (**c**) or five (**b**) independent experiments. See Methods for details on statistical analysis.

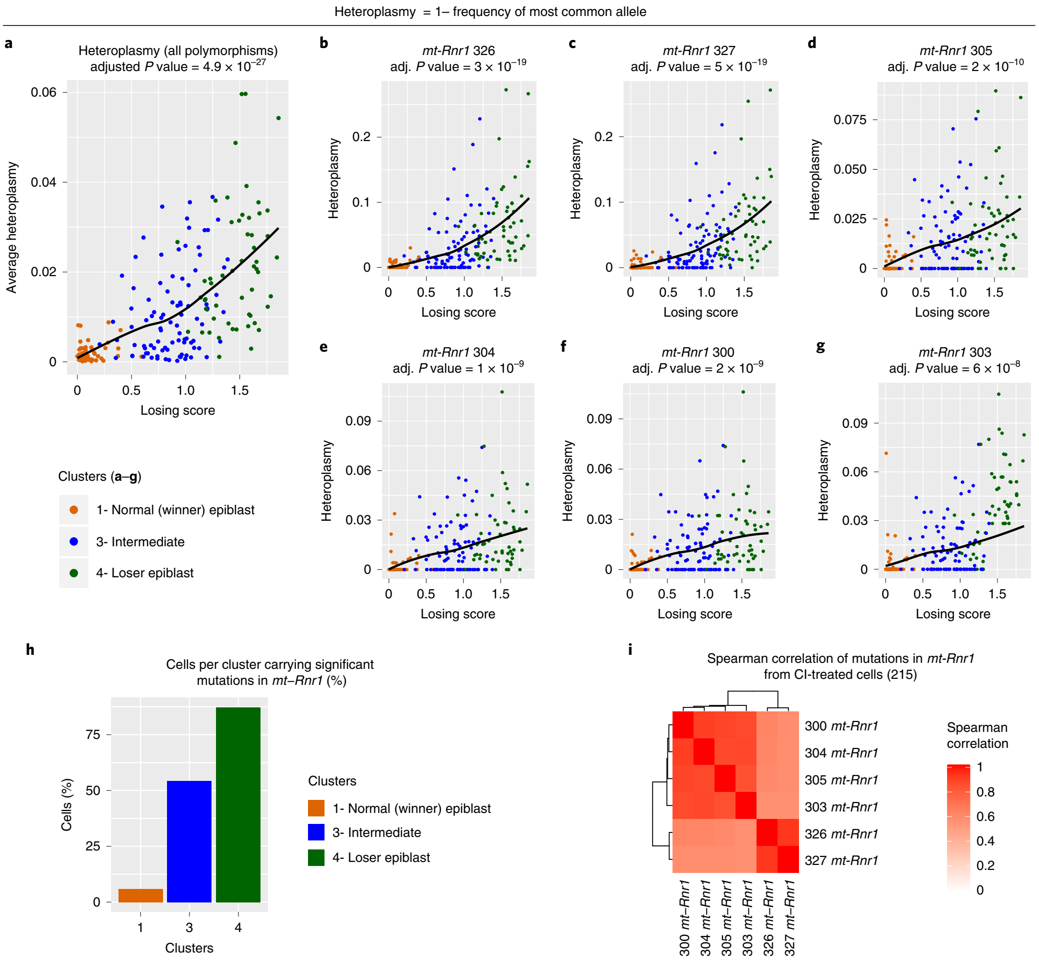


Fig. 6 | Intermediate and loser epiblast cells accumulate polymorphisms in mtDNA sequences. a–g, mtDNA heteroplasmy (plotted as heteroplasmy = 1 minus the frequency of most common allele) in epiblast cells from CI-treated embryos. Average heteroplasmy (considering all 11 polymorphisms that had a statistically significant dependence on the losing score; Methods) as a function of the losing scores of the cells (**a**). mtDNA heteroplasmy for six positions within *mt-Rnr1* (**b–g**). The heteroplasmy at all these positions, as well as the average heteroplasmy, increased with the losing scores of the cells in a statistically significant way (the adjusted *P* value estimated via a generalized linear model is indicated at the top of each plot). **h**, The bar plot indicates the fraction of epiblast cells in each of the clusters indicated on the *x* axis (winner, intermediate and loser) that carries a mean heteroplasmy (computed on the six positions within the *mt-Rnr1* indicated in **b–g**) greater than 0.01. This shows that the level of mtDNA heteroplasmy in *mt-Rnr1* is strongly associated with the loser status of the cells, as ~55% and ~87% of cells in the intermediate and the loser clusters, respectively, had heteroplasmic sequences in this gene compared to only ~5% of cells in the winner cluster. **i**, Spearman's correlation coefficient between the mtDNA heteroplasmy at the six positions shown in **b–g**. See Methods for details on statistical analysis.

and an upregulation of those associated with cytokine activity (Fig. 7g). Interestingly, in the co-culture condition, in addition to these signatures, BG (95%) cells revealed a downregulation in signature markers of MYC activity and mTOR signalling (Fig. 7h), whose downregulation is a known read-out of loser status during cell competition in the embryo[5–7] (Fig. 2c).

To test if the downregulation of genes involved in OXPHOS was also reflected at the functional level, we compared oxygen consumption rates (OCRs) and mitochondrial ATP generation in HB (100%), HB (24%), BG (95%) and C57BL/6N ESCs. We found that the winner cells HB (100%) and HB (24%) had higher basal respiration, higher maximal respiration and higher mitochondrial ATP production than the loser BG (95%) and C57BL/6N ESCs (Extended Data Fig. 9). These data indicate that the mtDNA differences that exist between winner and loser cells are sufficient to affect their mitochondrial performance and this ultimately determines their competitive ability. However, the observation that differentiating *Drp1*$^{-/-}$ ESCs are eliminated by cell competition but do not show differences in respiration rates or

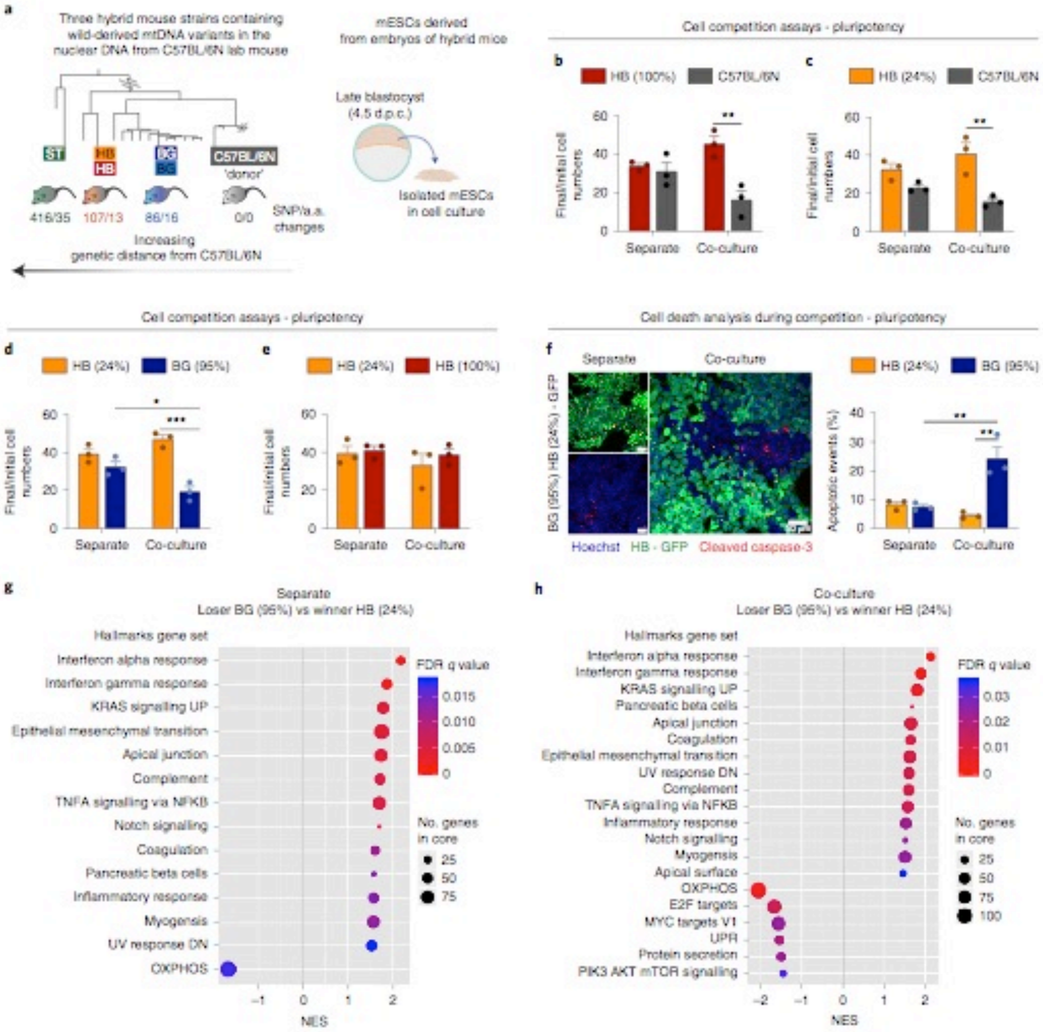


Fig. 7 | Changes in mtDNA sequence can determine the competitive ability of a cell. a, Derivation of mESCs from hybrid mouse strains, generated elsewhere by Burgstaller and colleagues. Neighbour-joining phylogenetic analysis of mtDNA from wild-derived and C57BL/6N mouse strains that were used to generate hybrid mice (adapted from a previous study[illegible]), illustrates the genetic distance of the mtDNA from wild-derived mouse strains to the C57BL/6N laboratory mouse. The number of SNPs and amino acid (a.a.) changes from the wild-derived to laboratory mouse strain is shown. mESCs were derived from embryos of hybrid mice, containing the nuclear background of a C57BL/6N laboratory mouse and mtDNA from three possible wild-derived strains (BG, HB or ST). **b–e**, Cell competition assays between cells derived from the embryos of hybrid mice performed in pluripotency maintenance conditions. The ratio of final/initial cell numbers in a separate culture or co-culture is shown. **f**, Representative micrographs of cleaved caspase-3 staining and quantification of the percentage of apoptotic events in winners HB (24%) and loser BG (95%) mESCs maintained in pluripotency and cultured in separate or co-culture conditions. **g**,**h**, GSEA of differentially expressed genes from bulk RNA-seq in loser BG (95%) compared to winner HB (24%) mESCs maintained in pluripotency and cultured in separate (**g**) or co-culture (**h**) conditions. Gene sets that show positive normalized enrichment scores (NESs) are enriched in loser cells, while gene sets that show negative NESs are depleted in loser cells. Data were obtained from four independent experiments (**g**,**h**). Remaining data are the mean ± s.e.m. of three independent experiments (**b**–**f**). See Methods for details on statistical analysis.

mitochondrial ATP production (Fig. 5c,g), suggests that respiration or ATP production rates alone are unlikely to be the mitochondrial parameters that control competitive cell fitness.

The finding that the genes downregulated in BG (95%) cells when co-cultured with HB (24%) cells fell under functional categories relating to mitochondrial function (Extended Data Fig. 10a)

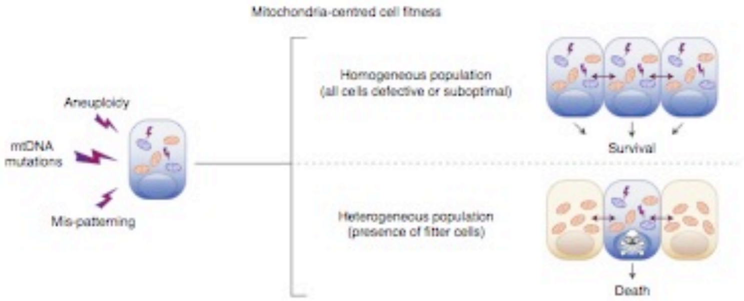

Mitochondria-centred cell fitness
Aneuploidy
mtDNA mutations
Mis-patterning
Homogeneous population (all cells defective or suboptimal)
Survival
Heterogeneous population (presence of fitter cells)
Death

Fig. 8 | Model of cell competition. Summary of the main findings of the study. A range of cellular defects, such as aneuploidy, mis-patterning or mtDNA mutations, cause alterations in mitochondrial function, affecting the relative fitness of cells. The cells with suboptimal mitochondrial activity survive in a homogeneous population but are eliminated by cell competition in the presence of fitter cells.

led us to analyse the degree of overlap between these genes and the genes differentially expressed along the winner-to-loser trajectory in the embryo. We observed a significant overlap in mis-regulated genes (Extended Data Fig. 10b), as well as in the functional components that these genes can be categorized into (Extended Data Fig. 10c). This further highlights the importance of relative mitochondrial activity for determining the competitive ability of embryonic cells.

Discussion

The emerging role of cell competition as a regulator of cell fitness in a wide range of cellular contexts, from the developing embryo to the ageing tissue[1–3], has highlighted the importance of understanding which cell types are normally eliminated by this process. With the aim of understanding this question, we analysed the transcriptional identity of the cells eliminated in the early mouse embryo. We found that they not only present a cell competition signature but also have impaired mitochondrial function and are marked by sequence changes in mt-Rnr1 and mt-Rnr2. Starting from these results, we leveraged in vitro models of cell competition to show that: (1) mitochondrial function is impaired in loser cells eliminated by cell competition, and (2) differences in mitochondrial activity are sufficient to trigger cell competition in ESCs. Overall, this points to mitochondrial performance as a key determinant of the competitive ability of cells during early mammalian embryonic development. One implication of our findings is that a range of different types of defects, such as mis-patterning, karyotypic abnormalities or mtDNA mutations, lead to dysfunctional mitochondria at the onset of differentiation and that ultimately it is their impaired mitochondrial function that triggers cell competition, inducing their elimination (Fig. 8).

Embryos are exposed to different microenvironments in vivo and when cultured ex vivo. Similarly, ESCs also experience a different microenvironment to epiblast cells in the embryo. These different microenvironments could potentially affect the selective pressure and hence the transcriptional signature of loser cells. However, there are two reasons why we think that the loser cell signatures identified here are conserved across systems. First, the transcriptional profile of our epiblast cells from cultured embryos is very similar to that of epiblast cells from freshly isolated embryos (Extended Data Fig. 3e–g). Second, the loser signature identified here is enriched for targets of P53 and depleted for mTOR and c-MYC targets. Given that these are regulators of cell competition identified by us and others in the embryo and in ESCs[5–7], it suggests that the same pathways are inducing loser cell elimination in in vivo, ex vivo and in ESC models of cell competition.

It is well known that the successful development of the embryo can be influenced by the quality of its mitochondrial pool[48]. Moreover, divergence from normal mitochondrial function during embryogenesis can either be lethal or lead to the development of mitochondrial disorders[49]. Deleterious mtDNA mutations are a common cause of mitochondrial diseases and, during development, selection against mutant mtDNA has been described to occur through at least two mechanisms: the bottleneck effect and the intracellular purifying selection. The bottleneck effect is associated specifically with the unequal segregation of mtDNAs during primordial germ cell specification, for example, as seen in the human embryo[50]. In contrast to this, purifying selection, as the name implies, allows for selection against deleterious mtDNAs and has been proposed to take place during both development and postnatal life[51]. Importantly, purifying selection has been found to occur at the molecule and organelle levels, as well as at the cellular level[52]. Our findings indicate that purifying selection can occur not only at the intracellular level but also at the intercellular level (cell non-autonomously). We show that epiblast cells can sense their relative mitochondrial activity and that those cells with mtDNA mutations or with lower or aberrant mitochondrial function are eliminated. By selecting those cells with the most favourable mitochondrial performance, cell competition would not only prevent cells with mitochondrial defects from contributing to the germline or future embryo, but also ensure optimization of the bioenergetic performance of the epiblast, therefore contributing to the synchronization of growth during early development.

Cell competition has been studied in a variety of organisms, from *Drosophila* to mammals, and it is likely that multiple different mechanisms fall under its broad umbrella[1–3]. Despite this, there is considerable interest in understanding if there could be any common feature in at least some of the contexts where cell competition has been described. The first demonstration of cell competition in *Drosophila* was made by inducing clones carrying mutations in the ribosomal gene *Minute*[4] and this has become one of the primary models to study this process. Our finding that during normal early mouse development cell competition eliminates cells carrying mutations in *mt-Rnr1* and *mt-Rnr2* transcripts, demonstrates that in the physiological context mutations in ribosomal genes also trigger cell competition. While we identified 11 mutations specific to loser cells, we cannot exclude the presence of additional variants differentiating winners from losers in those positions that did not have sufficient coverage in our RNA-seq data. Our observation that mis-patterned and karyotypically abnormal cells show impaired mitochondrial activity indicates that during early mouse development different types of defects impair mitochondrial function and trigger cell competition. Interestingly, mtDNA genes are amongst the top mis-regulated factors identified during cell competition in mouse skin[53]. In the *Drosophila* wing disc oxidative stress, a general consequence of dysfunctional mitochondria, underlies the out-competing of *Minute* and *Mahj* mutant cells[54]. Similarly, in Madin–Darby canine kidney cells, a loss of $\Delta\psi$m occurs during the out-competing of RasV12 mutant cells and is key for their extrusion[55]. These observations raise the possibility that differences in mitochondrial activity may be a key determinant of competitive cell fitness in a wide range of systems. Unravelling which mitochondrial features lead to cellular differences that can be sensed between cells during cell competition and if these are conserved in human systems will be key not only for understanding this process, but also to open up the possibility for future therapeutic avenues in the diagnosis or prevention of mitochondrial diseases.

Methods

Animals. Mice were maintained and treated in accordance with the Home Office's Animals (Scientific Procedures) Act 1986 and covered by the Home Office project licence PBBEBDCDA. All mice were housed on a 10–14-h light–dark cycle with access to water and food ad libitum. All mice were housed within individually ventilated cages. Temperature was maintained between 21–24 °C and humidity between 45–65%. Mattings were generally set up in the afternoon. Noon on the day of finding a vaginal plug was designated as E0.5. Embryo dissection was performed at appropriate time points in M2 medium (Sigma), using Dumont no.5 forceps (11251-10, FST). No distinction was made between male and female embryos during the analysis.

Cell lines, cell culture routine and drug treatments. E14 mESCs (RRID: CVCL_C320), kindly provided by A. Smith from Cambridge University, were used as wild-type control tdTomato-labelled or unlabelled cells. GFP-labelled or unlabelled cells defective for BMP signalling (*Bmpr1a*$^{-/-}$), tetraploid cells (4n) and *Bmp1a*$^{-/-}$ null for *p53* (*Bmpr1a*$^{-/-}$;*p53*$^{-/-}$) are described elsewhere[6,7]. *Drp1*$^{-/-}$ or *Mfn2*$^{-/-}$ cells were generated by CRISPR mutagenesis. Cells with different mtDNA content in the same nuclear background were derived from embryos of hybrid mice, generated elsewhere[16].

Cells were maintained at pluripotency and cultured at 37 °C in 5% CO_2 in 25-cm^2 flasks (Nunc) coated with 0.1% gelatin (Sigma) in DPBS. Growth medium (ES medium) consisted of GMEM supplemented with 10% FCS, 1 mM sodium pyruvate, 2 mM L-glutamine, 1× minimum essential medium non-essential amino acids and 0.1 mM β-mercaptoethanol (all from Gibco) and 0.1% leukaemia inhibitory factor (LIF, produced and tested in the laboratory). Cells derived from hybrid mice (C57BL/6N nuclear background) were maintained on 0.2% LIF. The growth medium was changed daily, and cells were split every 3 d.

To manipulate mitochondrial translation during differentiation, wild-type and *Drp1*$^{-/-}$ mESCs were treated with doxycycline (22.5 µM), from day 1 to day 3 of culture, or with actinonin (150 µM), for 6 h on day 3 of culture in N2B27 medium ('Differentiation and cell competition assays'). As the control condition, cells were treated with vehicle. Samples were collected on day 3 of differentiation for western blot analysis.

CRISPR mutagenesis. *Drp1* and *Mfn2* knockout ESCs were generated by CRISPR–Cas9-mediated deletion of *Drp1* exon 2 and *Mfn2* exon 3, respectively. sgRNA guides flanking *Drp1* exon 2 or *Mfn2* exon 3 were cloned into the PX459 vector (Addgene)[56]: *Drp1* exon 2 upstream sgRNA: 5′ TGGAACGGTCACAGCTGCAC 3′; *Drp1* exon 2 downstream sgRNA: 5′ TGGTCGCTGAGTTTGAGGCC 3′; *Mfn2* upstream sgRNA: 5′ GTGGTATGACCAATCCCAGA 3′; *Mfn2* downstream sgRNA: 5′ GGCCGGCCACTCTGCACCTT 3′. E14 ESCs were co-transfected with 1 µg of each sgRNA expression using Lipofectamine 2000 (Invitrogen) according to the manufacturer's instructions. As the control, E14 ESCs were transfected in parallel with an equal amount of empty PX459 plasmid. Following 6 d of puromycin selection, single colonies were picked from both *Drp1* sgRNA and ESCs transfected with empty vector and screened for mutations. *Drp1* exon 2 deletion was confirmed by PCR genotyping using the following primers: Drp1_genot F: 5′ GGATACCCCAAGATTTCTGGA 3′; Drp1_genot R: 5′ AGTCAGGTAATCGGGAGGAAA 3′, followed by Sanger sequencing. *Mfn2* exon 3 deletion was confirmed by PCR genotyping using the following primers: Mfn2_genot F: 5′ CAGCCCAGACATTGTTGCTTA 3′; Mfn2_genot R: 5′ AGCTGCCTCTCAGGAAATGAG 3′, followed by Sanger sequencing.

Derivation of mouse embryonic stem cells from hybrid mouse strains. The derivation of new mESC lines was adapted from work by Czechanski et al.[57]. Cells were derived from embryos of hybrid mouse strains BG, HB and ST. These contain the mtDNA of C57BL/6N (BL6) laboratory mouse and mtDNA variants from wild-caught mice[16].

Embryos were isolated at E2.5 (morula stage) and cultured in four-well plates (Nunc, Thermo Scientific) containing KSOM medium (Millipore) plus two inhibitors (KSOM + 2i): 1 µM MEK inhibitor PDO325901 (Sigma-Aldrich) and 3 µM GSK-3 inhibitor CHIR9902 (Cayman Chemicals) for 2 d at 37 °C in a 5% CO_2 incubator. To reduce evaporation, the area surrounding the wells was filled with DPBS. Embryos were further cultured in fresh 4-well plates containing N2B27 + 2i + LIF medium: N2B27 medium supplemented with 1 µM MEK inhibitor PDO325901 and 3 µM GSK-3 inhibitor and 0.1% LIF for up to 3 d until reaching the blastocyst stage. Each embryo was then transferred to a well of a 96-well plate coated with 0.1% gelatin in DPBS and containing 150 µl of N2B27 + 2i + LIF medium per well. In these conditions, the embryos should attach to the wells allowing the epiblast to form an outgrowth. This plate was then incubated at 37 °C in a 5% CO_2 incubator for 3 to 7 d until ESC-like colonies start to develop from the epiblast outgrowth. Cells were passaged by dissociation with Accutase (Sigma) and seeded in gradually increasing growth surface areas (48-well, 24-well and 12-well plates; T12.5 and T25 flasks), until new cell lines were established. At this stage, cells were weaned from N2B27 + 2i + LIF medium and then routinely cultured in ESC medium.

These new cell lines were then subjected to characterization by flow cytometry (cell size, granularity and mitochondrial $\Delta\psi$m) and amplification refractory mutation system (ARMS)–qPCR assay[16] to determine heteroplasmy.

Heteroplasmy quantification by ARMS–qPCR assay. Every qPCR run consisted of the consensus and an ARMS assay.

Consensus assay. CO2-F: TCTTATATGGCCTACCCATTCCAA, CO2-R: GGAAAACAATTATTAGTGTGTGATCATG, CO2-FAM: 6FAM-TTGGTCTACAAGACGCCACATCCCCT-BHQ-1

(amplicon length: 103 bp)

ARMS assays. 16SrRNA2340/Staudach-f: AAACCAACATATCTCATTGACCgAA (haplotype ST), *16SrRNA*2340(3)G-f: AATCAACATATCTTATTGACCaAG (haplotype C57BL/6N), *16SrRNA*2340(3)A-f: AATCAACATATCTTATTGACCgAA (haplotypes BG and HB), *16SrRNA*2458-r: CAC CAT TGG GAT GTC CTG ATC, 16SrRNA-FAM: FAM-CAA TTA GGG TTT ACG ACC TCG ATG TT-BHQ-1.

Lower-case letters indicate the intentional mismatch (ARMS), underlined letters indicate SNP-specific bases (amplicon length: 142 bp for BG and HB; 143 bp for ST).

Master-mixes for triplicate qPCR reactions contained 1× buffer B2 (Solis BioDyne), 4.5 mM $MgCl_2$, 200 µM of the four deoxynucleotides (dNTPs, Solis BioDyne), 0.7 units HOT FIREPol DNA polymerase (Solis BioDyne), 300 nM of each primer and 100 nM hydrolysis probe. For each reaction, 12 µl of master-mix and 3 µl DNA were transferred in triplicates to 384-well PCR plates (Life Technologies) using the automated pipetting system epMotion 5075TMX (Eppendorf). Amplification was performed on the ViiA 7 Real-Time PCR System using the ViiA 7 software v1.1 (Life Technologies). DNA denaturation and enzyme activation were performed for 15 min at 95 °C. DNA was amplified over 40 cycles consisting of 95 °C for 20 s, 58 °C for 20 s and 72 °C for 40 s for all assays.

The standard curve method was applied. Amplification efficiencies were determined for each run separately by DNA dilution series consisting of DNA from mice harbouring the respective analysed mtDNA. Typical results were: slope = −3.462, −3.461, −3.576 and −3.668; mean efficiency = 0.95, 0.94, 0.90 and 0.87; and *y* intercept = 32.4, 33.8, 34.5 and 31.9; for the consensus, C57BL/6N, HB and BG, and ST assays, respectively (Supplementary Figs. 1–4). Coefficient of correlation was ≥0.99 in all assays in all runs. All target samples were within the linear interval of the standard curves. To test for specificity, in each run, a negative control sample, that is, a DNA sample of a mouse harbouring the mtDNA of the non-analysed type in the heteroplasmic mouse (C57BL/6N or the respective wild-derived mtDNA) was measured. All assays could discriminate between C57BL/6N and wild-derived mouse mtDNA at a minimum level of >1%. Target sample DNA was tested for inhibition by dilution in Tris-EDTA buffer (pH 8.0).

For the calculation of mtDNA heteroplasmy, the assay detecting the minor allele (C57BL/6N or wild-derived mice, <50%) was always used. If both specific assays gave values >50% (that can happen at around 50% heteroplasmy), the mean value of both assays was taken. All qPCR runs contained no template controls for all assays; these were negative in 100% of analyses.

ARMS–qPCR standard curves and detection limit. mtDNA heteroplasmy was quantified by ARMS–qPCR, an established method in the field[16,19,58–63]. Calibration curves were created with a dilution series of DNA that showed a 100% match with the respective assay. Therefore, for all assays, necessarily divergent dilution series had to be used. The amount of DNA between the dilution series can diverge and thus values were plotted as arbitrary units. Supplementary Fig. 1 shows the standard curve produced for the consensus assay (detecting *mt-Co2* as a measure of total mtDNA) and Supplementary Figs. 2–4 show standard curves produced for specific mtDNA variants (laboratory mouse mtDNA, C57BL/6N; wild-derived mice mtDNAs BG, HB and ST).

SNP-specific quantification of mtDNA. To test the SNP-specific quantification of mtDNA, mixtures of match and mismatch DNA were analysed in triplicates. All assays could discriminate between C57BL/6N and wild-derived mouse mtDNA (and vice versa) at a minimum level of 1%, as shown by the ARMS–qPCR typical false-positive signal with the 100% mismatch DNA (detection limit, in all assays below 0.3%). The results and amplification plots for the specific quantification of HB and BG wild-derived mouse mtDNA from C57BL/6N mtDNA are shown in Supplementary Fig. 5 and Supplementary Table 9. The results and amplification plots for the specific quantification of ST wild-derived mouse mtDNA from C57BL/6N mtDNA are available in Supplementary Fig. 6 and Supplementary Table 10. Average values of the triplicate values are shown.

Embryo experiments. Early mouse embryos were isolated at E5.5 (from pregnant CD1 females, purchased from Charles River). Following dissection from the decidua, embryos were cultured overnight in poor N2B27 medium (same formulation as N2B27 medium but supplemented with 0.5× B27 supplement and 0.5× N2 supplement) with pan-CIs (100 µM, Z-VAD-FMK, FMK001, R&D Systems) or an equal volume of vehicle (DMSO) as the control. On the next morning, embryos were processed for scRNA-seq or functional validation ($\Delta\psi$m analysis and immunohistochemistry for markers of loser cells).

For the scRNA-seq and $\Delta\psi$m analysis, embryos were dissociated into a single-cell suspension. Briefly, up to 12 embryos were dissociated in 600 µl Accutase (A6964, Sigma) over 12 min at 37 °C, with tapping of the tube at 2-min intervals. Accutase was then neutralized with an equal volume of FCS, cells were spun down and stained with TMRM (for $\Delta\psi$m analysis) or directly resuspended in 300 µl DPBS with 1% FCS (for single-cell sorting and RNA-seq). Sytox Blue (1:1,000 dilution, S34857, Thermo Fisher Scientific) was used for viability staining.

Differentiation and cell competition assays. Cell competition assays between wild-type cells and *Bmpr1a*$^{-/-}$, 4n or *Drp1*$^{-/-}$ cells were performed in differentiating conditions. Cells were seeded onto fibronectin-coated plates (1:100, Merck) in DPBS for 1 h at 37 °C and grown in N2B27 medium to promote the differentiation of mESCs into a stage resembling the post-implantation epiblast, as cell competition was previously shown to occur in these conditions[6]. N2B27 medium consisted of 1:1 DMEM/F12 nutrient mixture and Neurobasal medium supplemented with N2 (1×) and B27 (1×) supplements, 2 mM L-glutamine and 0.1 mM β-mercaptoethanol (all from Gibco). Cell competition assays between wild-type and *Mfn2*$^{-/-}$ cells and between mESCs with different mtDNA content were performed in conditions of pluripotency maintenance (ESC medium).

Cells were either seeded separately or mixed for co-cultures at a 50:50 ratio, onto 12-well plates, at a density of 8×10^4 cells per well, except for assays between wild-type and *Mfn2*$^{-/-}$ mESCs, where 3.2×10^5 cells were seeded per well. The growth of cells was followed daily and compared between separate cultures or co-cultures, to control for cell-intrinsic growth differences, until the fourth day of culture. Viable cells were counted daily using a Vi-CELL XR Analyser (Beckman Coulter), and proportions of each cell type in co-cultures were determined using an LSR II Flow Cytometer (BD Bioscience), based on the fluorescent tag of the ubiquitously expressed GFP or TdTomato in one of the cell populations.

Metabolomic analysis. The metabolic profile was obtained using the Metabolon Platform (Metabolon). Each sample consisted of five biological replicates. For each replicate, 1×10^7 cells were spun down and snap frozen in liquid nitrogen. Pellets from five independent experiments for each condition were analysed by Metabolon using a combination of ultra-high performance liquid chromatography–tandem mass spectroscopy (UHPLC–MS/MS) and gas chromatography–mass spectroscopy (GC–MS). Compounds were identified by comparison to library entries of purified standards based on the retention time/index, mass-to-charge ratio (*m/z*) and chromatographic data (including MS/MS spectral data) on all molecules present in the library. Samples were normalized to protein content measured by Bradford assay. Statistical analysis was performed using Welch's two-sample *t*-test and statistical significance was defined as a P value ≤ 0.05.

Seahorse analysis. The metabolic function of cells was assessed by extracellular flux analysis using Seahorse XF24 (Agilent Technologies). For assays ran during pluripotency, cells were seeded, on the day before the assay, onto 0.1% gelatin-coated plates (Sigma) in 300 µl of ESC medium. All cell types were seeded at 5×10^4 cells per well, except for *Bmpr1a*$^{-/-}$ cells, which were seeded at 5×10^4 cells per well. For assays ran during differentiation, cells were seeded, 3 d before the assay, onto fibronectin-coated plates (1:100 dilution; Merck) in 300 µl of N2B27 medium. All cell types were seeded at 2.4×10^4 cells per well, except for *Bmpr1a*$^{-/-}$ cells, which were seeded at 3.2×10^4 cells per well.

On the day of the assay, cells were carefully washed twice with assay medium and then left with a final volume of 600 µl per well. The plate was then equilibrated on a non-CO_2 incubator at 37 °C for 30 min. The assay medium consisted of unbuffered DMEM (D5030, Sigma) that was supplemented on the day of the assay according to the test performed. For the OCR measurements, the assay medium was supplemented with 0.5 g l^{-1} glucose (Sigma) and 2 mM L-glutamine (Life Technologies), while for the ECAR measurements, the medium was supplemented with 1 mM sodium pyruvate and 2 mM L-glutamine (both from Life Technologies) at pH 7.4 and 37 °C.

The protocol for the assay consisted of four baseline measurements and three measurements after each compound addition. Compounds (all from Sigma) used in OCR and ECAR assays were prepared in the supplemented assay medium. For the OCR assay, the following compounds were added: 1 mM pyruvate, 2.5 µM oligomycin, 300 nM carbonyl cyanide-4-(trifluoromethoxy) phenylhydrazone and a mixture of rotenone and antimycin A at 6 µM each (R&A). For the ECAR assay, the following compounds were added: 2.5 mM and 10 mM of glucose, 2.5 µM of oligomycin and 50 mM of 2-deoxyglucose.

Each of the experiments was performed three times, with five biological replicates of each cell type. For background correction measurements, four wells were left without cells (A1, B4, C3 and D6). ECAR and OCR measurements were performed on the same plate. The assay parameters for both tests were calculated following the Seahorse assay report generator (Agilent Technologies).

At the end of the assay, cells were fixed and stained with Hoechst. Both OCR and ECAR were normalized to cell number, determined by manual cell counts using Fiji software. The normalization of the data was processed on Wave Desktop software (Agilent Technologies) and data were exported to Prism 8 (GraphPad) for statistical analysis.

Analysis of mitochondrial membrane potential and reactive oxygen species. For TMRM staining in single cells from early mouse epiblasts, embryos were dissected at E5.5 and cultured overnight in the presence or absence of CIs. On the following morning, to avoid misleading readings, epiblasts were isolated initially by an enzymatic treatment with 2.5% pancreatin, 0.5% trypsin and 0.5% polyvinylpyrrolidone (PVP40), all from Sigma-Aldrich, to remove the visceral endoderm. Embryos were treated for 8 min at 4 °C, followed by 2 min at room temperature (RT). The visceral endoderm was then peeled with the forceps and the extra-embryonic ectoderm was removed to isolate the epiblasts. Twelve epiblasts were pooled per 600 µl of Accutase (Sigma-Aldrich) for dissociation into single cells before staining. The reaction was stopped with an equal volume of FCS and cells were subjected to TMRM staining. Cells were incubated in 200 µl of 10 nM Nernstian probe TMRM perchlorate (T5428, Sigma), prepared in N2B27 medium. After incubation for 15 min at 37 °C, cells were pelleted again and resuspended in flow cytometry (FC) buffer (3% FCS in DPBS). Sytox Blue (1:1,000 dilution; S34857, Thermo Fisher Scientific) was used as viability staining.

Quantitative analysis of $\Delta\psi$m and mitochondrial ROS was performed by flow cytometry. Cells were grown in pluripotency or differentiating conditions. Cells were dissociated and pelleted to obtain 2×10^5 cells per sample for the staining procedure. For TMRM staining in mESCs, 2×10^5 cells of each cell line were resuspended in 200 µl of 10 nM TMRM (T5428, Sigma), prepared in N2B27 medium. Cells were incubated at 37 °C for 15 min, and then resuspended in FC buffer (3% FCS in DPBS). For the analysis of mitochondrial ROS, cells were grown in differentiating conditions and stained on the third day of culture. Briefly, 2×10^5 cells of each cell line were resuspended in 200 µl of a 5-µM solution of MitoSOX (M36008, Invitrogen) prepared in N2B27 medium. Cells were incubated at 37 °C for 15 min, and then resuspended in FC buffer. Sytox Blue was used for viability staining.

Cell suspensions stained with TMRM or MitoSOX were analysed in a BD LSR II flow cytometer operated through FACSDiva software (Becton Dickinson Biosciences). For TMRM fluorescence detection, the yellow laser was adjusted for excitation at $\lambda=562$ nm, capturing the emission light at $\lambda=585$ nm for TMRM. MitoSOX fluorescence was analysed with the violet laser adjusted for excitation at $\lambda=405$ nm, capturing the emission light at $\lambda=610$ nm. In the case of GFP-labelled cell lines, for GFP fluorescence detection, the blue laser was adjusted for excitation at $\lambda=488$ nm, capturing the emission light at $\lambda=525$ nm. Results were analysed in FlowJo Software v9 or v10.0.7r2. See the FACS gating strategy in Supplementary Fig. 7.

Qualitative analysis of $\Delta\psi$m was performed by confocal microscopy. Wild-type and *Bmpr1a*$^{-/-}$ cells were grown in fibronectin-coated glass coverslips. On the third day of differentiation, cells were loaded with a 200-nM MitoTracker Red probe (Life Technologies), prepared in N2B27 medium, for 15 min at 37 °C. Cells were then washed with DPBS and fixed with 3.7% formaldehyde for subsequent immunocytochemical staining of total mitochondrial mass, with TOMM20 antibody.

Immunofluorescence. Cells were washed with DPBS and fixed with 3.7% formaldehyde (Sigma) in N2B27, for 15 min at 37 °C. Permeabilization of the cell membranes was performed with 0.4% Triton X-100 in DPBS (DPBS-Tx), at RT with agitation. The blocking step with 5% BSA in DPBS-Tx 0.1% was performed for 30 min, at RT with agitation. Mitochondria were labelled with TOMM20 antibody (1:100 dilution; Santa Cruz Biotechnologies). Dead cells were labelled with cleaved caspase-3 antibody (1:400 dilution; CST-9664), and NANOG antibody was used to mark pluripotent cells (1:100 dilution; eBioscience). Secondary antibodies were Alexa Fluor 488 and 568 (1:600 dilution; Invitrogen). Primary antibody incubation was performed overnight at 4 °C and secondary antibody incubation was done for 45 min, together with Hoechst to stain nuclei (1:1,000 dilution; Thermo Scientific) at RT and protected from light. In both cases, antibodies were diluted in blocking solution. Three 10-min washes with DPBS-Tx 0.1% were performed between each critical step and before mounting with Vectashield medium (Vector Laboratories).

Samples were imaged with a Zeiss LSM 780 confocal microscope and processed with Fiji[64]. Mitochondrial stainings were imaged with a ×63/1.4 oil objective. For samples stained with TOMM20 antibody and MitoTracker Red, *z*-stacks were acquired and processed for deconvolution using Huygens software (Scientific Volume Imaging; https://svi.nl/). Samples stained with cleaved caspase-3 were imaged with a ×20/0.8 air objective. Imaging and deconvolution analysis were performed with support and advice from S. Rothery from the Facility for Imaging by Light Microscopy (FILM) at Imperial College London.

Embryo immunofluorescence staining for p-rpS6, OPA1 and DDIT3 (CHOP) markers was performed as follows. Cultured embryos were fixed in 4% paraformaldehyde in DPBS containing 0.01% Triton and 0.1% Tween 20 for 20 min at RT. Permeabilization of the membranes was performed for 10 min in DPBS with 0.5% Triton. Embryos were blocked in 5% BSA in DPBS with 0.25% Triton during 45 min. Incubation with primary antibodies (CHOP (1:500 dilution; CST, 2895), OPA1 (1:100 dilution; BD Biosciences, 612606) and p-rpS6 (1:200 dilution; CST, 5364)) was completed overnight at 4 °C in 2.5% BSA in DPBS with 0.125% Triton. The following morning, hybridization with secondary antibodies Alexa Fluor 568 and Alexa Fluor 488 (Invitrogen, diluted at 1:600 in DPBS with 0.125% Triton) was performed next for 1 h at RT. Hoechst was also added to this mixture to stain nuclei (1:1,000 dilution; Invitrogen). Three 10-min washes with filtered DPBS-Tx 0.1% were performed between each critical step. All steps included gentle agitation.

Embryos were imaged in embryo dishes (Nunc) in a drop of Vectashield using a Zeiss LSM 780 confocal microscope at ×40/1.3 oil objective.

Further details about image acquisition and processing are specified in Supplementary Table 8.

Western blotting. Cells were washed in DPBS and lysed with Laemmli lysis buffer (0.05 M Tris-HCl at pH 6.8, 1% SDS, 10% glycerol and 0.1% β-mercaptoethanol in distilled water). Total protein quantification was done using BCA assay (Thermo Scientific) and samples (15 µg of protein per lane) were loaded into 12% Bis-Tris protein gels (Bio-Rad). Resolved proteins were transferred into nitrocellulose membranes (GE Healthcare). The following primary antibodies, prepared in TBS-0.1% Tween containing 5% BSA were incubated overnight at 4 °C with gentle agitation: rabbit anti-TOMM20 (1:1,000 dilution; CST, 42406), mouse anti-ATPB (1:1,000 dilution; Abcam, ab14730), rabbit anti-α-tubulin (1:1,000 dilution; CST, 2144), mouse anti-mt-CO1 (1:2,000 dilution; Abcam, ab14705), rabbit anti-DRP1 (1:1,000 dilution; CST, 8570), mouse anti-MFN1 (1:1,000 dilution; Abcam, ab57602), mouse anti-MFN2 (1:500 dilution; Abcam, ab56889), mouse anti-vinculin (1:1,000 dilution; Sigma, V9131), mouse anti-OPA1 (1:1,000 dilution; BD Biosciences, 612606), rabbit anti-ATF4 (1:1,000 dilution; CST, 11815), rabbit anti-PCNA (1:5,000 dilution; Abcam, ab18197) and rabbit anti-p-eIF2α (Ser51; 1:1,000 dilution; CST, 9721). On the following morning, HRP-conjugated secondary antibodies (1:5,000 dilution; sc-2004 and sc-2005, Santa Cruz), prepared in TBS-0.1% Tween containing 5% milk (Sigma) were incubated for 1 h at RT under gentle agitation. Membranes were developed with ECL reagents (Promega) and mounted in cassettes for time-controlled exposure to film (GE Healthcare).

Bulk and single-cell RNA sequencing. For bulk RNA-seq in the competitive scenario between cells with different mtDNA, HB (24%) and BG (95%) mESCs were grown separately or in co-culture. On the third day of culture, cells were dissociated and subjected to FACS to separate the cell populations in co-culture according to their GFP label. Propidium iodine (1:1,000 dilution; 81845, Sigma) was used for viability staining. See the FACS gating strategy in Supplementary Fig. 8. To control for eventual transcriptional changes due to the FACS process, a mixture of the two separate populations was subjected to the same procedure as the co-cultured samples. Total RNA isolation was then carried out using RNA extraction Kit (RNeasy Mini Kit, QIAGEN). PolyA selection/enrichment was the method adopted for library preparation, using the NEB Ultra II RNA Prep Kit. Single-end 50-bp libraries were sequenced on an Illumina HiSeq 2500. Raw base call files were converted to fastq files using Illumina's bcl2fastq (v2.1.7). Reads were aligned to the mouse genome (mm9) using Tophat2 (v2.0.11)[65] with default parameters. Mapped reads that fell on genes were counted using featureCounts from Rsubread package[66]. Generated count data were then used to identify differentially expressed genes using DESeq2 (ref. [67]). Genes with very low read counts were excluded. Finally, GSEA was performed using GSEA software[68,69] on a pre-ranked list generated by DESeq2.

To investigate the nature of cells eliminated by cell competition during early mouse embryogenesis by means of scRNA-seq, early mouse embryos were dissected at E5.5 and cultured overnight in the presence or absence of CIs. The next morning, embryos were dissociated with Accutase and subjected to single-cell sorting into 384-well plates. Total RNA isolation was then carried out using an RNA extraction Kit (RNeasy Mini Kit, QIAGEN). scRNA-seq was performed using the Smart-seq2 protocol[70]. PolyA selection/enrichment with Ultra II Kit (NEB) was the method adopted for library preparation.

Data processing, quality control and normalization. We performed transcript quantification in our scRNA-seq data by running Salmon (v0.8.2)[71] in the quasi-mapping-based mode. First, a transcriptome index was created from the mouse reference (version GRCm38.p4) and ERCC spike-in sequences. Then, the quantification step was carried out with the 'quant' function, correcting for the sequence-specific biases ('--seqBias' flag) and the fragment-level GC biases ('--gcBias' flag). Finally, the transcript-level abundances were aggregated to gene-level counts. On the resulting raw count matrix including 1,495 cells, we applied a quality-control check to exclude poor quality cells from downstream analyses.

For quality control, we applied the following criteria: identification of cells that had a $\log_{10}$ total number of reads equal to or greater than 4, a fraction of mapped reads equal to or greater than 0.8, a number of genes with an expression level above ten reads per million equal to or greater than 3,000 and a fraction of reads mapped to endogenous genes equal to or greater than 0.5. This resulted in the selection of 723 cells, which were kept for downstream analyses. Transcripts per million (TPM) normalization (as estimated by Salmon) was used.

Highly variable genes and dimensionality reduction. To identify highly variable genes (HVGs), we first fitted a mean and total variance trend using the R function 'trendVar' and then the variance was decomposed into biological and technical components with the R function 'decomposeVar'; both functions are included in the package 'scran' (v1.6.9)[72].

We considered HVGs those with a biological component that was significantly greater than zero at an FDR (Benjamini–Hochberg method) of 0.05. Then, we applied further filtering steps by keeping only genes that had an average expression greater to or equal than 10 TPM and were significantly correlated with one another (function 'correlatePairs' in 'scran' package, FDR < 0.05). This yielded 1,921 genes, which were used to calculate a distance matrix between cells defined as $\sqrt{(1-\rho)/2}$, where ρ is the Spearman's correlation coefficient between cells. A two-dimensional representation of the data was obtained with the UMAP package (v0.2.0.0; https://cran.r-project.org/web/packages/umap/index.html) using the distance matrix as input.

Cell clustering and connectivity analysis. To classify cells into different clusters, we ran hierarchical clustering on the distance matrix (see above; 'hclust' function in R with ward.D2 aggregation method) followed by the dynamic hybrid cut algorithm ('cutreeDynamic' function in R package 'dynamicTreeCut' (https://CRAN.R-project.org/package=dynamicTreeCut) v1.63.1, with the hybrid method, using a minimum cluster size of 35 cells and a 'deepSplit' parameter equal to 0), which identified five clusters. Cells from different batches were well mixed across these five clusters (Extended Data Fig. 1), suggesting that the batch effect was negligible. The identity of the five clusters was established based on the expression of known marker genes of epiblasts, visceral endoderm and extra-embryonic ectoderm, which were identified in a previous study[73]. The expression levels of some of the top markers are plotted in Fig. 1b.

We performed a robustness analysis on the clustering by exploring in detail how the choices of genes, clustering parameters and algorithms affect the identity and the number of clusters. First, we quantified the cluster robustness by calculating Pearson's gamma and the average silhouette width obtained with 100 random subsets of 60% of the HVGs and different values of the deepSplit parameter. While the robustness at a deepSplit value of 0 and 1 was similar, for greater values of deepSplit (corresponding to less conservative clustering), the robustness rapidly declined (Extended Data Fig. 1e). The clustering with deepSplit value of 0 and 1 (the more robust choices) yielded very similar results, the only difference being the splitting of the intermediate cluster in two subclusters (Extended Data Fig. 1f).

In addition to this, we also used Louvain clustering on the HVGs (resolution = 0.3, $k = 20$ with 20 principal components), which again produced very similar clusters.

We quantified the connectivity between the clusters (using only CI-treated cells) with PAGA[22] implemented in the Python library scanpy (v1.4.7)[74]. The analysis revealed that the three epiblast clusters were connected with each other, whereas the two extra-embryonic tissues (visceral endoderm and extra-embryonic ectoderm) were isolated (Extended Data Fig. 2b).

Identification of a single-cell trajectory in the epiblast. We calculated a diffusion map ('DiffusionMap' function in the R package 'destiny' (v2.6.2)[23]) on the distance defined above on the epiblast cells from CI-treated embryos. The pseudotime coordinate was computed with the 'DPT' function with the root cell in the winner epiblast cluster (identified by the function 'tips' in the 'destiny' package). Such pseudotime coordinates can be interpreted as a 'losing score' for all the epiblast cells from the CI-treated embryos.

We estimated the losing scores of the epiblast cells from DMSO-treated embryos by projecting such data onto the diffusion map previously calculated (function 'dm_predict' in the destiny package). Finally, for each of the projected cells, we assigned the losing score as the average of the losing scores of the ten closest neighbours in the original diffusion map (detected with the function 'projection-dist' in the destiny package).

While for the clustering and the trajectory analysis we used the HVGs computed from the whole dataset, we verified that all results concerning the separation between winner and loser epiblast cells (for example, clusters and losing score) remain unaffected if the HVGs are calculated using only the epiblast cells.

Mapping of data from epiblast cells onto published datasets. We compared the transcriptional profile of epiblasts from embryos cultured in DMSO and CI with that of epiblasts collected from freshly isolated embryos at different stages.

To do this, we considered a dataset published previously[26], which includes epiblast cells from embryos at the stages E5.5 (102 cells), E6.25 (130 cells) and E6.5 (288 cells). A diffusion map and a diffusion pseudotime coordinate were computed with these cells following the same procedure described above (Extended Data Fig. 3e,f). Then, we projected epiblast cells from CI-treated and DMSO-treated embryos and we assigned to them a diffusion pseudotime coordinate as described above (Extended Data Fig. 3g).

Differential gene expression analysis along the trajectory. To identify the genes that were differentially expressed along the trajectory, we first kept only genes that had more than 15 TPM in more than ten cells (this list of genes is provided in Supplementary Table 4); then, we obtained the log-transformed expression levels of these genes (adding 1 as a pseudo-count to avoid infinities) as a function of the losing score and we fitted a generalized additive model (GAM) to them (R function 'gam' from 'GAM' package version 1.16.). We used the ANOVA test for parametric effects provided by the 'gam' function to estimate a *P* value for each tested gene. This yielded a list of 5,311 differentially expressed genes ($\text{FDR} < 0.01$).

Next, we looked for groups of differentially expressed genes that shared similar expression patterns along the trajectory. To this aim, similarly to what we did when clustering cells, we calculated a correlation-based distance matrix between genes, defined as $\sqrt{(1-\rho)/2}$, where ρ is the Spearman's correlation coefficient between genes. Hierarchical clustering was then applied to this matrix ('hclust' function in R, with the 'ward.D2' method) followed by the dynamic hybrid cut algorithm ('dynamicTreeCut' package) to define clusters ('cutreeDynamic' function in R with the hybrid method and a minimum cluster size of 100 genes and a deepSplit parameter equal to 0). This resulted in the definition of four clusters, including three clusters of genes that decreased along the trajectory (merged together for the Gene Ontology enrichment and the IPA analysis) and one cluster of increasing genes (Extended Data Fig. 3a). IPA (QIAGEN; https://www.qiagenbioinformatics.com/products/ingenuity-pathway-analysis/), was run on all genes differentially expressed ($\text{FDR} < 0.01$) along the trajectory from winner to loser cells (Figs. 2a–d and 3a–c), using all the tested genes as a background (Supplementary Table 4). This software generated networks, canonical pathways and functional analysis. The list of decreasing/increasing genes is provided in Supplementary Tables 1 and 2. The pathways found as mis-regulated in Fig. 3 were: mitochondrial dysfunction, $-\log_{10}(P\ \text{value}) = 21.1$; OXPHOS, $-\log_{10}(P\ \text{value}) = 18.6$; EIF2 signalling, $-\log_{10}(P\ \text{value}) = 11.9$. FDRs for the genes shown in Fig. 3b range from 1.25×10^{-51} (for *Atp5b*) to 5.42×10^{-3} (for *Ndufa11*). *Cox6b2* was found to be upregulated in loser cells ($\text{FDR} = 2.69 \times 10^{-13}$).

Analysis of heteroplasmy in a single-cell RNA-seq dataset. We used STAR (v2.7)[75] to align the transcriptome of the epiblast cells from CI-treated embryos (274) to the mouse reference genome (mm10). Only reads that uniquely mapped to the mtDNA were considered. From these, we obtained allele counts at each mtDNA position with a Phred quality score greater than 33 using the samtools 'mpileup' function.

Next, we applied filters to remove cells and mtDNA positions with a low coverage. First, we removed cells with fewer than 2,000 mtDNA positions covered by more than 50 reads. Second, we removed positions having less than 50 reads in more than 50% of cells in each of the three epiblast clusters (winner, intermediate and loser). These two filters resulted in 259 cells and 5,192 mtDNA positions (covered by ~700 reads per cell on average) being considered for further analyses.

Starting from these cells and positions, we applied an additional filter to keep only positions with a sufficiently high level of heteroplasmy. To this aim, for each position with more than 50 reads in a cell, we estimated the heteroplasmy according to equation (1):

$$H = 1 - f_{max} \tag{1}$$

where $f_{\max}$ is the frequency of the most common allele. We kept only positions with $H > 0.01$ in at least ten cells.

Finally, using GAMs (see above), we identified the positions whose heteroplasmy *H* changes as a function of the cells' losing score in a statistically significant way. We found a total of eleven significant positions ($\text{FDR} < 0.001$), six of them in *mt-Rnr1* and five in *mt-Rnr2*. All of these positions had a higher level of heteroplasmy in loser cells (Fig. 6b–g and Extended Data Fig. 7a-e). The results remain substantially unaltered if the Spearman's rank correlation test (as opposed to the GAMs) is used.

The most common substitutions observed in each position were: *mt-Rnr1* 300 A-to-C; *mt-Rnr1* 303 T-to-G; *mt-Rnr1* 304 T-to-G; *mt-Rnr1* 305 C-to-G; *mt-Rnr1* 326 A-to-G; *mt-Rnr1* 327 C-to-G; *mt-Rnr2* 2,031 T-to-G; *mt-Rnr2* 2,074 C-to-G; *mt-Rnr2* 2,077 A-to-C; *mt-Rnr2* 2,079 C-to-T; *mt-Rnr2* 2,081 A-to-G.

For the bar plot shown in Fig. 6h and the correlation heat maps in Fig. 6i and Extended Data Fig. 7l, we took into account only cells that covered with more than 50 reads all the significant positions in the *mt-Rnr1* gene (215 cells; Fig. 6h,i) or in both the *mt-Rnr1* and *mt-Rnr2* genes (214 cells; Extended Data Fig. 7l).

As a negative control, we repeated the analysis described above using the ERCC spike-ins added to each cell. As expected, none of the positions were statistically significant, which suggests that our procedure is robust against sequence errors introduced during PCR amplification.

We also performed the mtDNA heteroplasmy analysis in cells from the visceral endoderm and the extra-embryonic ectoderm in both DMSO and CI conditions; none of these cells had a mtDNA heteroplasmy higher than 0.01 in the 11 significant positions identified within *mt-Rnr1* and *mt-Rnr2* in loser epiblast cells, and the reference allele was always the most common. This reinforces the hypothesis that such variants are specific to loser epiblast cells and are not resulting from contamination.

To test the reliability of our heteroplasmy estimations, we used RNA-seq data from two of the mtDNA cell lines (BG and HB; Fig. 7) for which the heteroplasmy was measured also by ARMS–qPCR. To do so, first we downloaded the fasta files of the two mtDNA cell lines from https://www.ncbi.nlm.nih.gov/nuccore/KC663619.1/ and https://www.ncbi.nlm.nih.gov/nuccore/KC663620.1/, then we identified the mtDNA positions that differed from the BL6 reference genome. Finally, on these different positions, the heteroplasmy, *H*, was computed as explained above. The values of heteroplasmy we found with our computational analysis were very close to those estimated by ARMS–qPCR: for HB (24%), ~17% from RNA-seq data versus ~24% measured by ARMS–qPCR; and for BG (95%), ~93% from RNA-seq data versus ~95% measured by ARMS–qPCR.

Because we are inferring mtDNA changes from RNA-seq data, we also considered additional potential sources for the sequence changes we observed. Specifically, one possible source is contamination from NUMTs. However, a NUMT contamination is very unlikely for the following reasons: (1) we considered only reads that uniquely mapped to the mitochondrial genome; (2) the variants with the highest heteroplasmy identified in 'loser' cells (*mt-Rnr1* 326 and 327) were not present in any of the NUMTs previously reported[76] or those that we identified using *blastn* (also taking into account the SNPs of the mouse strain we used); (3) the variants detected were exclusively found in 'loser' epiblast cells, and they were not detected in any other cell type from the same embryos, that is, neither in 'winner' epiblast cells nor in cells from extra-embryonic tissues; (4) we estimated that if the variants with the strongest heteroplasmy (that is, *mtRnr-1* 326 and 327) were present on a NUMT, in order for them to reach an heteroplasmy of ~20% (Fig. 6b,c), the NUMT would have to be expressed at high levels, comparable to or even higher than many mitochondrial genes.

Another possible cause of the sequence changes is RNA editing. However, the majority of the changes that we found (see above) are not compatible with the canonical RNA editing in Metazoans, which consists of A-to-I (which would be read as A-to-G in RNA-seq) and C-to-U[77].

Common features of scRNA-seq and bulk RNA-seq datasets. Differential expression analysis between the co-cultured winner HB (24%) and loser BG (95%) cell lines was performed using the package EdgeR (v3.20.9)[78].

Batches were specified in the argument of the function 'model.matrix'. We fitted a quasi-likelihood negative binomial generalized log-linear model (with the function 'glmQLFit') to the genes that were filtered by the function filterByExpr (with default parameters). These genes were used as background for the gene enrichment analysis.

We set an FDR of 0.001 as a threshold for significance. The enrichment analysis for both the scRNA-seq and bulk RNA-seq datasets were performed using the tool g:Profiler[79]. The list of upregulated, downregulated and background genes related to the differential expression analysis for the bulk RNA-seq dataset is provided in the Supplementary Tables 5–7.

Quantification, statistical analysis and reproducibility. The quantification of the DDIT3 and OPA1 expression in embryos was performed using two distinct methods. DDIT3 expression was quantified by counting the number of epiblast cells with positive staining in the embryos of each group. The expression of OPA1 was quantified on Fiji software as the mean fluorescence across a ten-pixel-width line drawn on the basal cytoplasm of each cell with high or low p-rpS6 fluorescence intensity, as specified in a previous study[7]. A minimum of eight cells were quantified per condition (high versus low mTOR activity) in each embryo. Six embryos treated with CI were analysed. Mean values of OPA1 fluorescence for each epiblast cell were pooled on the same graph.

Flow cytometry data were analysed with FlowJo Software v9 or v10.0.7r2.

Western blot quantification was performed using Image Studio Lite v5.2.5 (LI-COR). Protein expression levels were normalized to loading controls vinculin, α-tubulin or PCNA.

Normalization of data from metabolic flux analysis with Seahorse was performed using Wave Desktop software (Agilent Technologies) and data were exported to Prism v8 (GraphPad) for statistical analysis.

All box plots show the lower quartile (Q1, 25th percentile), the median (Q2, 50th percentile) and the upper quartile (Q3, 75th percentile). Box length refers to interquartile range (IQR, $Q3 - Q1$). The upper whisker marks the minimum between the maximum value in the dataset and 1.5 times the IQR from Q3 ($Q3 + 1.5 \times \text{IQR}$), while the lower whisker marks the maximum between the minimum value in the dataset and the IQR times 1.5 from Q1 ($Q1 - 1.5 \times \text{IQR}$). Outliers are shown outside the interval defined by box and whiskers as individual points.

The micrographs shown on Fig. 3d represents one of the micro-dissected embryo epiblasts used for the experiment presented in Fig. 3g,h. The representative confocal microscopy images shown in Fig. 4h are from confocal imaging deconvolution performed from one experiment following reproducibility of observations from previous independent experiments.

The statistical analysis of the results generated in wet-lab experiments was performed using GraphPad Prism v8.0.0 for Mac (GraphPad Software). Data were tested for normality using the Shapiro–Wilk normality test. Two-tailed parametric or non-parametric statistical tests were applied accordingly. Statistical significance was considered with a confidence interval of 0.05%; $*P<0.05$; $**P<0.01$; $***P<0.001$.

Here we specify details about the statistical test and multiple-comparisons test (when applicable) used for each experiment. The statistical significance of IPA analysis shown in Figs. 2b,c and 3a,b was calculated with A right-tailed Fisher's exact test ($P<0.05$). Data presented in Figs. 3f and 4i and Extended Data Fig. 4g were analysed by Mann–Whitney test. Data shown in Fig. 4b–e and Extended Data Fig. 5a–d were analysed by an unpaired *t*-test or Mann–Whitney *U* test. Data shown in Figs. 4j and 5e,g–i were analysed with an unpaired *t*-test. A one-sample *t*-test was used to analyse data presented in Fig. 5d,f. Figure 4k and Extended Data Fig. 8b–f show data analysed by one-way ANOVA, followed by Holm–Sidak's multiple-comparisons test. Data presented in Figs. 3h, 4f,g, 5b,c,j and 7b–f and Extended Data Figs. 5e–g, 6b,c and 8g–i were analysed by two-way ANOVA, followed by Holm–Sidak's multiple-comparisons test. The statistical analysis of data from Extended Data Fig. 9 was carried out with one-way ANOVA or Kruskal–Wallis test, followed by Holm–Sidak's or Dunn's multiple-comparison test, respectively. ANOVA for parametric effects on a GAM fit was used test statistical significance for data presented in Fig. 6a–g and Extended Data Fig. 7a–e. The adjusted *P* values (indicated at the top of each plot) were computed using the Benjamini–Hochberg method. The correlation coefficients shown in Fig. 6i and Extended Data Figs. 2e, 3d and 7l were calculated with Spearman's rank correlation rho test (two-sided test, 0.95 confidence level). Data shown in Extended Data Figs. 3b,c, 4a and 10b,c were analysed with Fisher's exact test (two-sided). Gene enrichment analysis shown in Extended Data Fig. 10 was tested for statistical significance with a cumulative hypergeometric test. *P* values were adjusted for multiple comparisons using the g:Profiler algorithm g:SCS (https://doi.org/10.1093/nar/gkm226). Finally, the statistical analysis on data presented in Supplementary Tables 5 and 6 was performed using empirical Bayes quasi-likelihood *F* tests. *P* values were adjusted for multiple comparisons using the Benjamini–Hochberg method.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

Data were analysed with standard programmes and packages, as detailed above. All relevant data are included in the paper and/or its Supplementary Information files. RNA-seq raw data as well as processed data are available through ArrayExpress, under accession numbers E-MTAB-8640, for scRNA-seq data, and E-MTAB-8692, for bulk RNA-seq data. Source data are provided with this paper.

Code availability

Code used to generate the figures in the paper is available at https://github.com/ScialdoneLab/Cell-Competition-Paper-Figures/.

Received: 24 February 2020; Accepted: 2 June 2021;
Published online: 12 July 2021

References

1. Bowling, S., Lawlor, K. & Rodriguez, T. A. Cell competition: the winners and losers of fitness selection. *Development* **146**, dev167486 (2019).
2. Diaz-Diaz, C. & Torres, M. Insights into the quantitative and dynamic aspects of cell competition. *Curr. Opin. Cell Biol.* **60**, 68–74 (2019).
3. Madan, E., Gogna, R. & Moreno, E. Cell competition in development: information from flies and vertebrates. *Curr. Opin. Cell Biol.* **55**, 150–157 (2018).
4. Morata, G. & Ripoll, P. Minutes: mutants of *Drosophila* autonomously affecting cell division rate. *Dev. Biol.* **42**, 211–221 (1975).
5. Claveria, C., Giovinazzo, G., Sierra, R. & Torres, M. Myc-driven endogenous cell competition in the early mammalian embryo. *Nature* **500**, 39–44 (2013).
6. Sancho, M. et al. Competitive interactions eliminate unfit embryonic stem cells at the onset of differentiation. *Dev. Cell* **26**, 19–30 (2013).
7. Bowling, S. et al. P53 and mTOR signalling determine fitness selection through cell competition during early mouse embryonic development. *Nat. Commun.* **9**, 1763 (2018).
8. Diaz-Diaz, C. et al. Pluripotency surveillance by myc-driven competitive elimination of differentiating cells. *Dev. Cell* **42**, 585–599 (2017).
9. Hashimoto, M. & Sasaki, H. Epiblast formation by TEAD-YAP-dependent expression of pluripotency factors and competitive elimination of unspecified cells. *Dev. Cell* **50**, 139–154 (2019).
10. Lima, A., Burgstaller, J., Sanchez-Nieto, J. M. & Rodriguez, T. A. The mitochondria and the regulation of cell fitness during early mammalian development. *Curr. Top. Dev. Biol.* **128**, 339–363 (2018).
11. Zhou, W. et al. HIF1α induced switch from bivalent to exclusively glycolytic metabolism during ESC-to-EpiSC/hESC transition. *EMBO J.* **31**, 2103–2116 (2012).
12. Khrapko, K. et al. Mitochondrial mutational spectra in human cells and tissues. *Proc. Natl Acad. Sci. USA* **94**, 13798–13803 (1997).
13. Allio, R., Donega, S., Galtier, N. & Nabholz, B. Large variation in the ratio of mitochondrial to nuclear mutation rate across animals: implications for genetic diversity and the use of mitochondrial DNA as a molecular marker. *Mol. Biol. Evol.* **34**, 2762–2772 (2017).
14. Burgstaller, J. P., Johnston, I. G. & Poulton, J. Mitochondrial DNA disease and developmental implications for reproductive strategies. *Mol. Hum. Reprod.* **21**, 11–22 (2015).
15. Gorman, G. S. et al. Mitochondrial diseases. *Nat. Rev. Dis. Prim.* **2**, 16080 (2016).
16. Burgstaller, J. P. et al. MtDNA segregation in heteroplasmic tissues is common in vivo and modulated by haplotype differences and developmental stage. *Cell Rep.* **7**, 2031–2041 (2014).
17. Johnston, I. G. et al. Stochastic modelling, Bayesian inference, and new in vivo measurements elucidate the debated mtDNA bottleneck mechanism. *eLife* **4**, e07464 (2015).
18. Latorre-Pellicer, A. et al. Regulation of mother-to-offspring transmission of mtDNA heteroplasmy. *Cell Metab.* **30**, 1120–1130 (2019).
19. Lee, H. S. et al. Rapid mitochondrial DNA segregation in primate preimplantation embryos precedes somatic and germline bottleneck. *Cell Rep.* **1**, 506–515 (2012).
20. Zhang, H., Burr, S. P. & Chinnery, P. F. The mitochondrial DNA genetic bottleneck: inheritance and beyond. *Essays Biochem.* **62**, 225–234 (2018).
21. Sharpley, M. S. et al. Heteroplasmy of mouse mtDNA is genetically unstable and results in altered behavior and cognition. *Cell* **151**, 333–343 (2012).
22. Wolf, F. A. et al. PAGA: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. *Genome Biol.* **20**, 59 (2019).
23. Angerer, P. et al. destiny: diffusion maps for large-scale single-cell data in R. *Bioinformatics* **32**, 1241–1243 (2016).
24. Kramer, A., Green, J., Pollard, J. Jr. & Tugendreich, S. Causal analysis approaches in ingenuity pathway analysis. *Bioinformatics* **30**, 523–530 (2014).
25. Haghverdi, L., Buttner, M., Wolf, F. A., Buettner, F. & Theis, F. J. Diffusion pseudotime robustly reconstructs lineage branching. *Nat. Methods* **13**, 845–848 (2016).
26. Cheng, S. et al. Single-cell RNA-seq reveals cellular heterogeneity of pluripotency transition and X chromosome dynamics during early mouse development. *Cell Rep.* **26**, 2593–2607 (2019).
27. Topf, U., Wrobel, L. & Chacinska, A. Chatty mitochondria: keeping balance in cellular protein homeostasis. *Trends Cell Biol.* **26**, 577–586 (2016).
28. Melber, A. & Haynes, C. M. UPR^{mt} regulation and output: a stress response mediated by mitochondrial–nuclear communication. *Cell Res.* **28**, 281–295 (2018).
29. Munch, C. The different axes of the mammalian mitochondrial unfolded protein response. *BMC Biol.* **16**, 81 (2018).
30. Zhao, Q. et al. A mitochondrial specific stress response in mammalian cells. *EMBO J.* **21**, 4411–4419 (2002).
31. Nargund, A. M., Pellegrino, M. W., Fiorese, C. J., Baker, B. M. & Haynes, C. M. Mitochondrial import efficiency of ATFS-1 regulates mitochondrial UPR activation. *Science* **337**, 587–590 (2012).
32. Quiros, P. M., Mottis, A. & Auwerx, J. Mitonuclear communication in homeostasis and stress. *Nat. Rev. Mol. Cell Biol.* **17**, 213–226 (2016).
33. Mouchiroud, L. et al. The NAD$^+$/sirtuin pathway modulates longevity through activation of mitochondrial UPR and FOXO signaling. *Cell* **154**, 430–441 (2013).
34. Saveljeva, S. et al. Endoplasmic reticulum stress-mediated induction of SESTRIN 2 potentiates cell survival. *Oncotarget* **7**, 12254–12266 (2016).
35. Yun, J. & Finkel, T. Mitohormesis. *Cell Metab.* **19**, 757–766 (2014).
36. Chen, H. et al. Mitofusins Mfn1 and Mfn2 coordinately regulate mitochondrial fusion and are essential for embryonic development. *J. Cell Biol.* **160**, 189–200 (2003).
37. Prudent, J. & McBride, H. M. The mitochondria–endoplasmic reticulum contact sites: a signalling platform for cell death. *Curr. Opin. Cell Biol.* **47**, 52–63 (2017).
38. Smirnova, E., Griparic, L., Shurland, D. L. & van der Bliek, A. M. Dynamin-related protein Drp1 is required for mitochondrial division in mammalian cells. *Mol. Biol. Cell* **12**, 2245–2256 (2001).
39. Favaro, G. et al. DRP1-mediated mitochondrial shape controls calcium homeostasis and muscle mass. *Nat. Commun.* **10**, 2576 (2019).
40. Quiros, P. M. et al. Multi-omics analysis identifies ATF4 as a key regulator of the mitochondrial stress response in mammals. *J. Cell Biol.* **216**, 2027–2045 (2017).
41. Restelli, L. M. et al. Neuronal mitochondrial dysfunction activates the integrated stress response to induce fibroblast growth factor 21. *Cell Rep.* **24**, 1407–1414 (2018).
42. Richter, U. et al. A mitochondrial ribosomal and RNA decay pathway blocks cell proliferation. *Curr. Biol.* **23**, 535–541 (2013).

43. Moullan, N. et al. Tetracyclines disturb mitochondrial function across eukaryotic models: a call for caution in biomedical research. *Cell Rep.* **10**, 1681–1691 (2015).
44. Kauppila, J. H. K. et al. A phenotype-driven approach to generate mouse models with pathogenic mtDNA mutations causing mitochondrial disease. *Cell Rep.* **16**, 2980–2990 (2016).
45. Fan, W. et al. A mouse model of mitochondrial disease reveals germline selection against severe mtDNA mutations. *Science* **319**, 958–962 (2008).
46. Stewart, J. B. et al. Strong purifying selection in transmission of mammalian mitochondrial DNA. *PLoS Biol.* **6**, e10 (2008).
47. Freyer, C. et al. Variation in germline mtDNA heteroplasmy is determined prenatally but modified during subsequent transmission. *Nat. Genet.* **44**, 1282–1285 (2012).
48. Ludwig, L. S. et al. Lineage tracing in humans enabled by mitochondrial mutations and single-cell genomics. *Cell* **176**, 1325–1339 (2019).
49. Chinnery, P. F. & Hudson, G. Mitochondrial genetics. *Br. Med. Bull.* **106**, 135–159 (2013).
50. Floros, V. I. et al. Segregation of mitochondrial DNA heteroplasmy through a developmental genetic bottleneck in human embryos. *Nat. Cell Biol.* **20**, 144–151 (2018).
51. Burr, S. P., Pezet, M. & Chinnery, P. F. Mitochondrial DNA heteroplasmy and purifying selection in the mammalian female germline. *Dev. Growth Differ.* **60**, 21–32 (2018).
52. Rajasimha, H. K., Chinnery, P. F. & Samuels, D. C. Selection against pathogenic mtDNA mutations in a stem cell population leads to the loss of the 3243 A > G mutation in blood. *Am. J. Hum. Genet.* **82**, 333–343 (2008).
53. Ellis, S. J. et al. Distinct modes of cell competition shape mammalian tissue morphogenesis. *Nature* **569**, 497–502 (2019).
54. Kucinski, I., Dinan, M., Kolahgar, G. & Piddini, E. Chronic activation of JNK JAK/STAT and oxidative stress signalling causes the loser cell status. *Nat. Commun.* **8**, 136 (2017).
55. Kon, S. et al. Cell competition with normal epithelial cells promotes apical extrusion of transformed cells through metabolic changes. *Nat. Cell Biol.* **19**, 530–541 (2017).
56. Ran, F. A. et al. Genome engineering using the CRISPR–Cas9 system. *Nat. Protoc.* **8**, 2281–2308 (2013).
57. Czechanski, A. et al. Derivation and characterization of mouse embryonic stem cells from permissive and nonpermissive strains. *Nat. Protoc.* **9**, 559–574 (2014).
58. Burgstaller, J. P. et al. Large-scale genetic analysis reveals mammalian mtDNA heteroplasmy dynamics and variance increase through lifetimes and generations. *Nat. Commun.* **9**, 2488 (2018).
59. Burgstaller, J. P., Schinogl, P., Dinnyes, A., Muller, M. & Steinborn, R. Mitochondrial DNA heteroplasmy in ovine fetuses and sheep cloned by somatic cell nuclear transfer. *BMC Dev. Biol.* **7**, 141 (2007).
60. Kang, E. et al. Mitochondrial replacement in human oocytes carrying pathogenic mitochondrial DNA mutations. *Nature* **540**, 270–275 (2016).
61. Yahata, N., Boda, H. & Hata, R. Elimination of mutant mtDNA by an optimized mpTALEN restores differentiation capacities of heteroplasmic MELAS-iPSCs. *Mol. Ther. Methods Clin. Dev.* **20**, 54–68 (2021).
62. Venegas, V. & Halberg, M.C. Quantification of mtDNA mutation heteroplasmy (ARMS–qPCR). *Methods Mol. Biol.* **837**, 313–326 (2012).
63. Machado, T. S. et al. Real-time PCR quantification of heteroplasmy in a mouse model with mitochondrial DNA of C57BL/6 and NZB/BINJ strains. *PLoS ONE* **10**, e0133650 (2015).
64. Schindelin, J. et al. Fiji: an open-source platform for biological-image analysis. *Nat. Methods* **9**, 676–682 (2012).
65. Kim, D. et al. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* **14**, R36 (2013).
66. Liao, Y., Smyth, G. K. & Shi, W. The R package Rsubread is easier, faster, cheaper and better for alignment and quantification of RNA sequencing reads. *Nucleic Acids Res.* **47**, e47 (2019).
67. 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).
68. Mootha, V. K. et al. PGC-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat. Genet.* **34**, 267–273 (2003).
69. Subramanian, A. et al. Gene-set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA* **102**, 15545–15550 (2005).
70. Picelli, S. et al. Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nat. Methods* **10**, 1096–1098 (2013).
71. Patro, R., Duggal, G., Love, M. I., Irizarry, R. A. & Kingsford, C. Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods* **14**, 417–419 (2017).
72. Lun, A. T., McCarthy, D. J. & Marioni, J. C. A step-by-step workflow for low-level analysis of single-cell RNA-seq data with Bioconductor. *F1000Res* **5**, 2122 (2016).
73. Scialdone, A. et al. Resolving early mesoderm diversification through single-cell expression profiling. *Nature* **535**, 289–293 (2016).
74. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.* **19**, 15 (2018).
75. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
76. Calabrese, F. M., Simone, D. & Attimonelli, M. Primates and mouse NumtS in the UCSC Genome Browser. *BMC Bioinformatics* **13**, S15 (2012).
77. Lukes, J., Kaur, B. & Speijer, D. RNA editing in mitochondria and plastids: weird and widespread. *Trends Genet.* **37**, 99–102 (2021).
78. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).
79. Reimand, J., Arak, T. & Vilo, J. g:Profiler—a web server for functional interpretation of gene lists (2011 update). *Nucleic Acids Res.* **39**, W307–W315 (2011).
80. Zappia, L. & Oshlack, A. Clustering trees: a visualization for evaluating clusterings at multiple resolutions. *GigaScience* **7**, giy083 (2018).
81. Scialdone, A. et al. Computational assignment of cell-cycle stage from single-cell transcriptome data. *Methods* **85**, 54–61 (2015).

Acknowledgements

We thank S. Rothery for guidance and advice with confocal microscopy. The FILM at Imperial College London is supported in part by funding from the Wellcome Trust (grant no. 104931/Z/14/Z) and BBSRC (grant no. BB/L015129/1). We thank J. Elliot and B. Patel from the LMS/NIHR Imperial Biomedical Research Centre Flow Cytometry Facility for support. We are thankful to G. Chennell and A. Sardini for guidance and support with Seahorse experiments. Research in the laboratory of T.A.R. was supported by the MRC project grant (MR/P018467/1) and the BBSRC project grant (BB/S008284/1) and by the British Heart Foundation (BHF) PhD studentships (FS/14/62/31288 and FS/17/64/33476). Work in the laboratory of A.S. is funded by the Helmholtz Association. A.L. was funded by a BHF centre of excellence PhD studentship. S.S. was funded through Wellcome awards 103788/Z/14/Z and 108438/Z/15/Z.

Author contributions

A.L. performed most of the experimental wet-lab work. J.B. and A.L. derived heteroplasmic mESC lines. J.B. performed heteroplasmy measurements in heteroplasmic mESCs. B.P. generated $Mfn2^{-/-}$ and $Drp1^{-/-}$ mESCs, and J.M.S. conducted characterization of mitochondria shape and pluripotency status. S.P.-M. participated in the metabolic characterization of $Drp1^{-/-}$ cells. D.H. performed embryo dissections, treatments and cell dissociation before scRNA-seq experiments. G.L. did the bioinformatic analysis of scRNA-seq data. E.M., N.J. and A.P.G. participated in the analysis of mtDNA heteroplasmy. A.D.G. performed the metabolomic studies using the Metabolon platform and participated in embryo dissections and immunohistochemistry stainings for validation of results obtained by scRNA-seq. T.K. collected the embryos for derivation of the mESCs with different mtDNA content. M.D. and M.K. performed the bioinformatic analysis of bulk RNA-seq experiments. N.J., S.S. and D.C. participated in the design of experimental work and analysis of results. A.L., G.L., A.S and T.A.R. interpreted results and wrote the paper. T.A.R. and A.S. directed and designed the research.

Competing interests

The authors declare no competing interests.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s42255-021-00422-7.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s42255-021-00422-7.

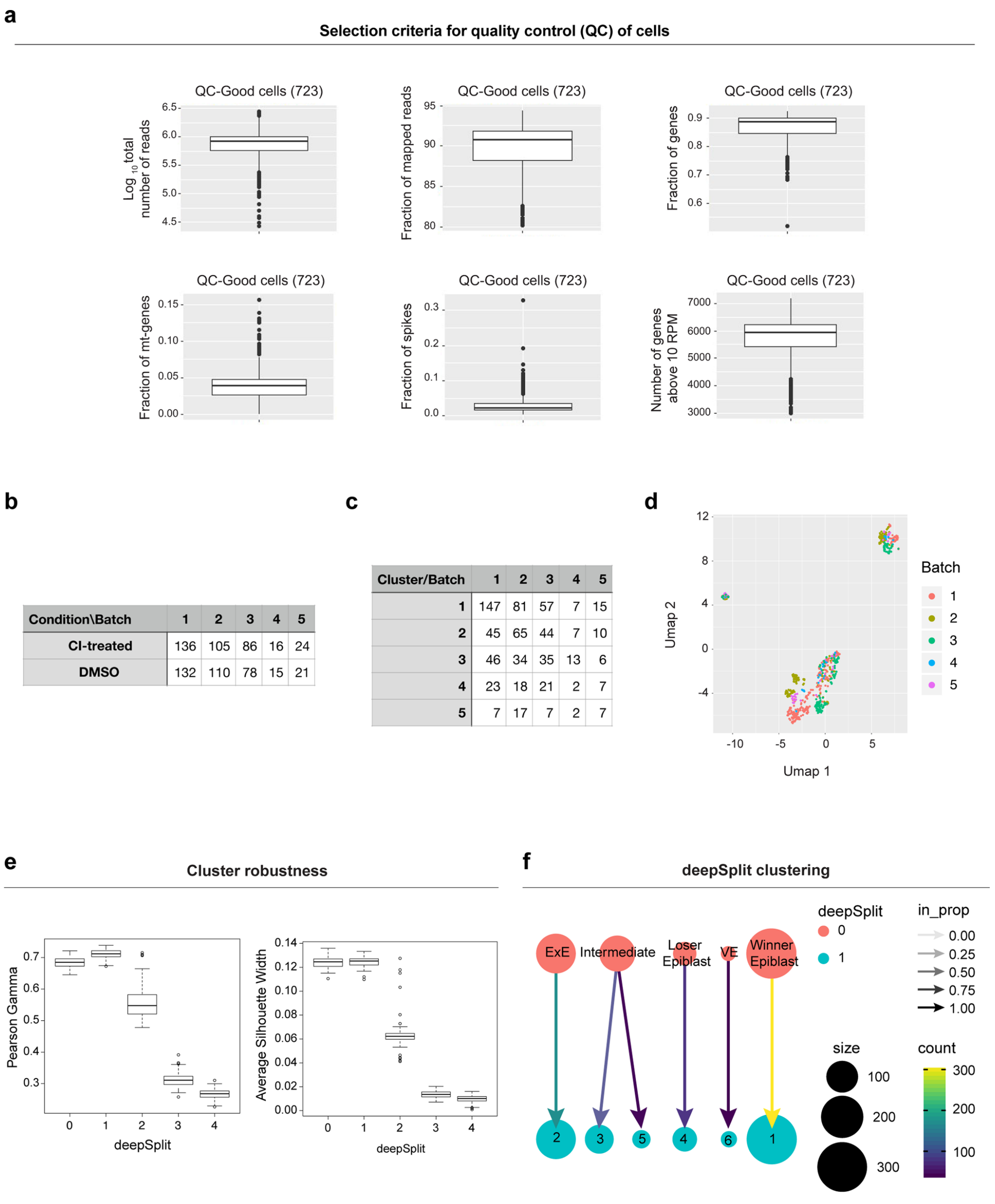
Correspondence and requests for materials should be addressed to A.S. or T.A.R.

Peer review information *Nature Metabolism* thanks Anna-Katerina Hadjantonakis and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Primary Handling Editors: Christoph Schmitt; Elena Bellafante.

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.

© The Author(s), under exclusive licence to Springer Nature Limited 2021

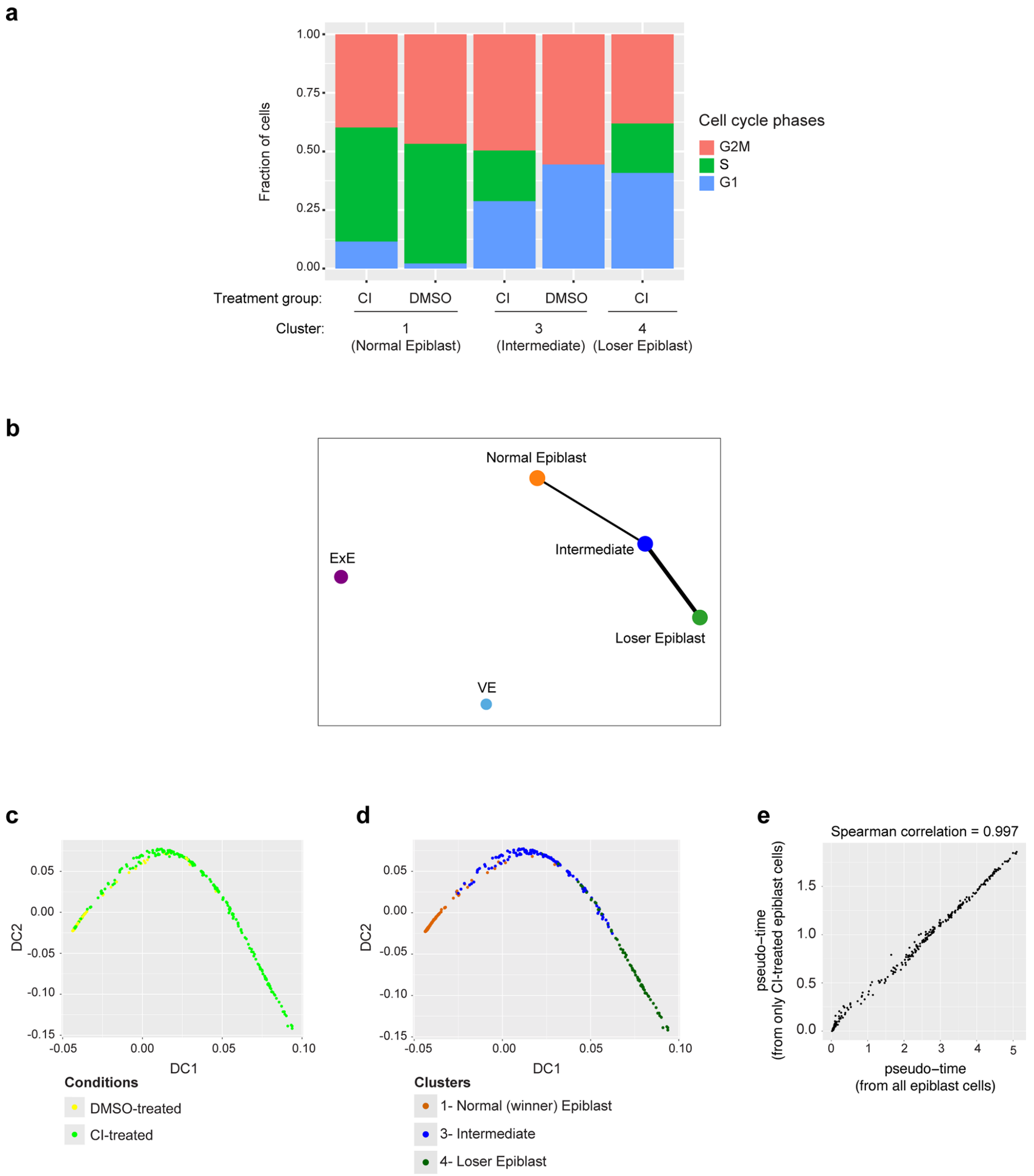


Condition\Batch	1	2	3	4	5
CI-treated	136	105	86	16	24
DMSO	132	110	78	15	21

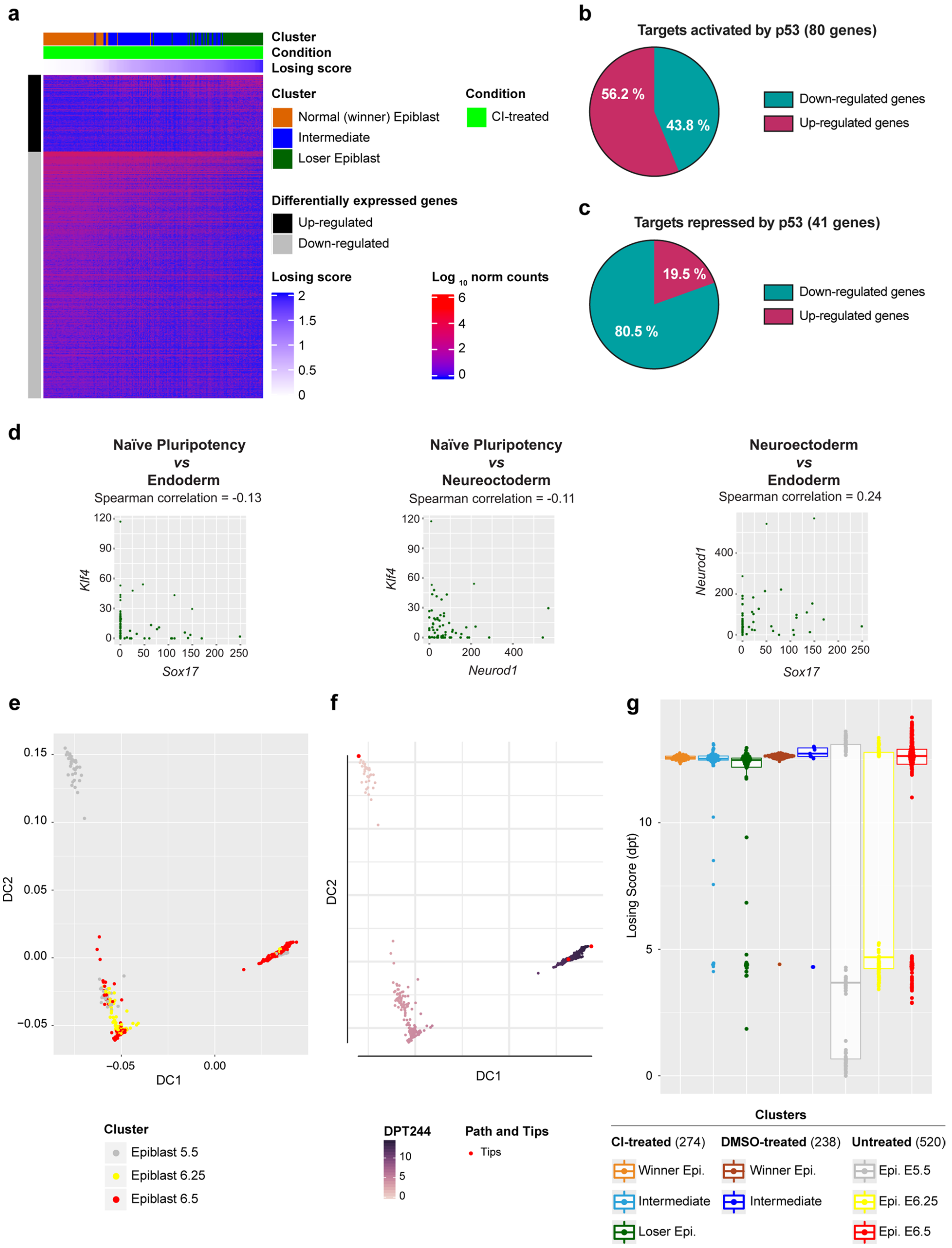
Cluster/Batch	1	2	3	4	5
1	147	81	57	7	15
2	45	65	44	7	10
3	46	34	35	13	6
4	23	18	21	2	7
5	7	17	7	2	7

Extended Data Fig. 1 | See next page for caption.

Extended Data Fig. 1 | Quality controls of scRNA-seq and clustering robustness analysis. a, Selection criteria for quality control (QC) of all cells. A total of 723 passed the quality control (723 good quality cells) and were considered for downstream analysis. All these parameters were computed for each cell. Log_{10} total number of reads (top left): $\log_{10}$ of the sum of the number of reads that were processed in every cell; Fraction of mapped reads (top central): number of reads that are confidentially mapped to the reference genome divided by total number of reads that were processed for each cell. This number is automatically provided by Salmon v0.8.2; Fraction of genes (top right): number of reads mapped to endogenous genes divided by the total sum of reads that were processed; Fraction of mt-genes (bottom left): number of reads mapped to mitochondrial genes divided by the total sum of reads that were processed; Fraction of spikes (bottom central): number of reads mapped to ERCC spike-ins divided by the total sum of reads that were processed; Number of genes above 10 RPM (bottom right): number of genes with expression level above 10 reads per million. **b**, Number of good quality cells in each condition (rows) and batch (columns). **c**, Number of good quality cells per cluster (rows) and batch (columns). **d**, UMAP plot of the data with cells coloured by batch. In each batch there is a balanced distribution of cells in the two conditions and across the five clusters. **e**, The Pearson's gamma (left panel) and the Average Silhouette Width (right panel) was calculated for each set of clusters obtained with 100 random subsamples of 60% of highly variable genes and different values of the deepSplit parameter (see Methods). The most robust clusters correspond to deepSplit values of 0 and 1. **f**, The changes in composition and number of clusters between the clustering obtained with deepSplit 0 (top) and 1 (bottom) are shown using the library 'clustree'[80]. See methods for details on statistical analysis.

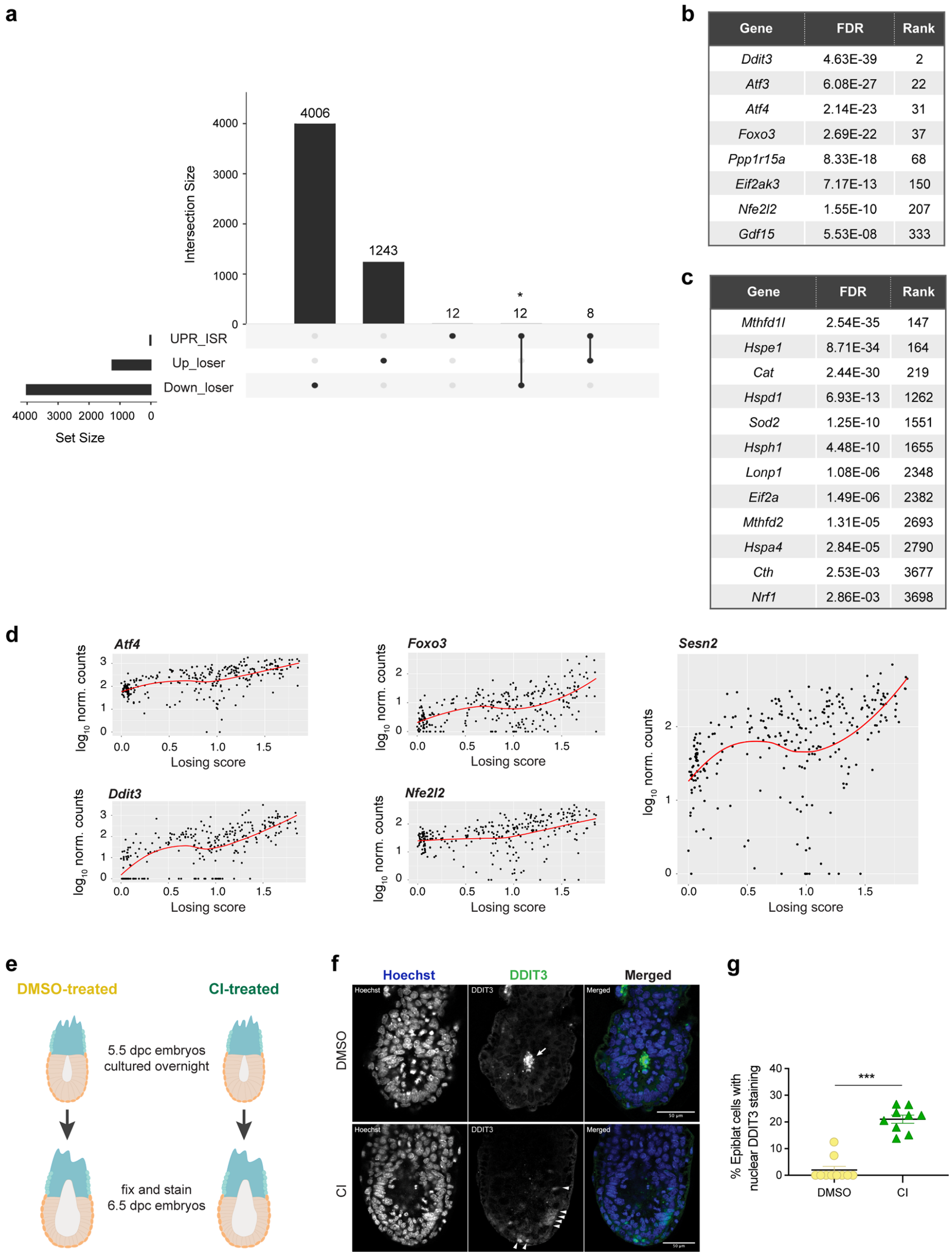


Extended Data Fig. 2 | Cell cycle analysis and cluster connectivity. **a**, Cell cycle analysis of epiblast cells from clusters 1, 3 and 4. Cell cycle phase was predicted with cyclone algorithm[81] and shows that there are cells in S and G2M phase also in the loser and intermediate clusters. **b**, PAGA plot showing the connectivity of the five clusters of cells from CI-treated embryos. **c-d**, Diffusion map analysis in all epiblast cells (from DMSO and CI-treated embryos): cells are coloured according to the condition (**c**) and to the cluster (**d**). **e**, The pseudotime coordinate of the CI-treated epiblast cells obtained from the diffusion map including all epiblast cells correlates extremely well (with the pseudo-time coordinate obtained in the diffusion map calculated only from CI-treated epiblast cells (Fig. 2a). See methods for details on statistical analysis.



Extended Data Fig. 3 | See next page for caption.

Extended Data Fig. 3 | Analysis on epiblast cells from DMSO and CI-treated embryos. a, Heatmap showing the expression pattern of all genes differentially expressed along the trajectory from winning to losing cells in Fig. 2d. **b-c**, Overlap of genes differentially expressed along the trajectory joining winning and losing epiblast cells in CI-treated embryos (Fig. 2a and panel d) and genes targeted by p53. Pie charts show the percentage of genes up- or down-regulated in loser cells within the group of target genes that are activated (**b**) or repressed (**c**) by p53. There is an enrichment of activated/repressed targets among genes upregulated/downregulated in losing cells respectively (p-value=1E-4). The list of p53 targets is taken from[58]. **d**, Scatter plots of the expression levels of different marker genes plotted against each other in loser epiblast cells (cluster 4). Loser cells have higher expression of pluripotency markers as well as higher expression of some lineage-specific markers and the co-expression of these markers is only weakly correlated - the Spearman's correlation coefficient is shown. **e-g** Our scRNA-seq data from epiblast cells is projected on top of previously published data from epiblast collected from freshly isolated embryos at different stages (E5.5, E6.25 and E6.5; data from[26]). First, a diffusion map (**e**) and a pseudotime coordinate (**f**) is computed for the epiblast cells from freshly isolated embryos. Then, a pseudotime coordinate is estimated for our data after projecting it onto the diffusion map. Panel **g** shows the pseudotime coordinates for both datasets, split by stage, treatment and cluster. See methods for details on statistical analysis.



b

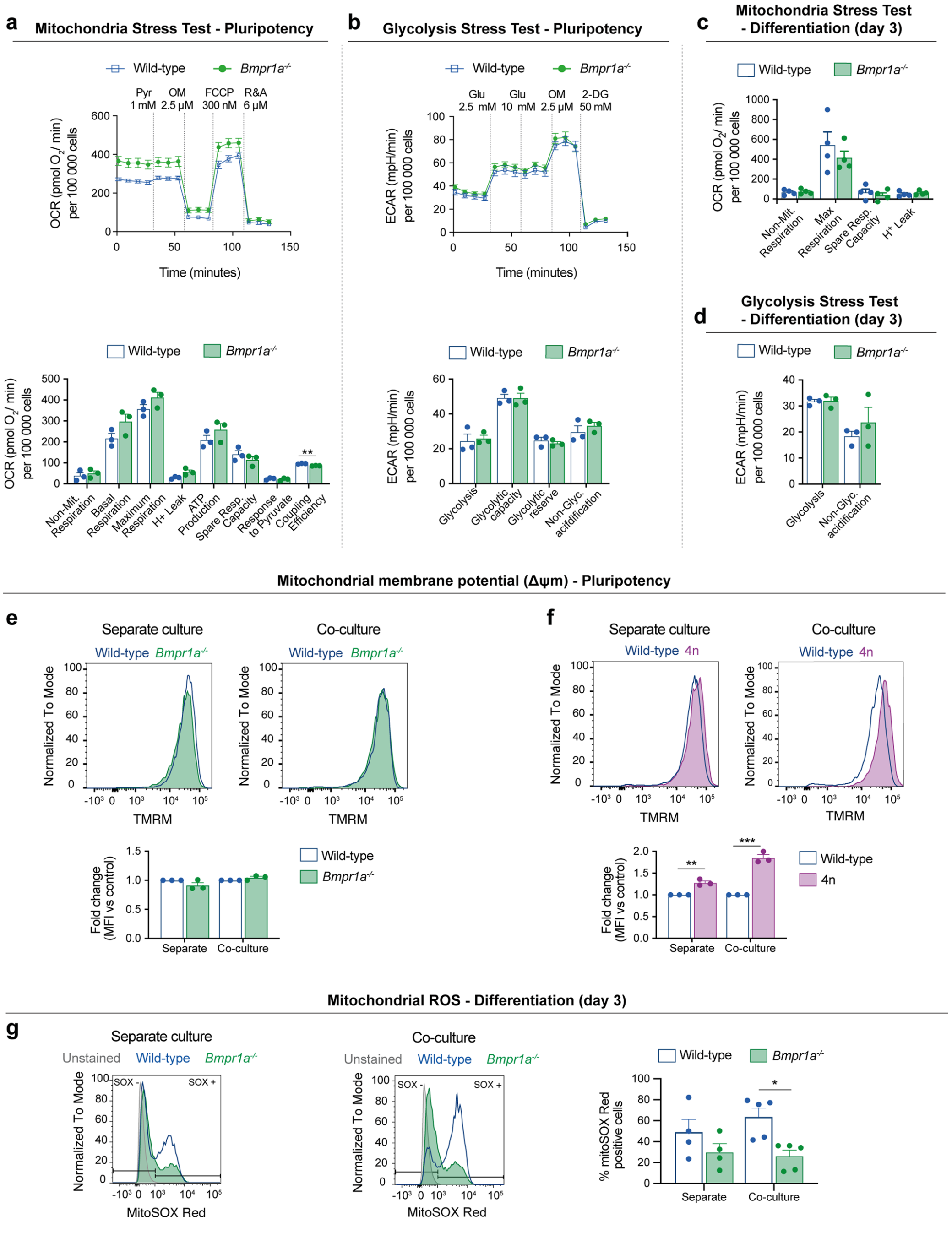
Gene	FDR	Rank
Ddit3	4.63E-39	2
Atf3	6.08E-27	22
Atf4	2.14E-23	31
Foxo3	2.69E-22	37
Ppp1r15a	8.33E-18	68
Eif2ak3	7.17E-13	150
Nfe2l2	1.55E-10	207
Gdf15	5.53E-08	333

c

Gene	FDR	Rank
Mthfd1l	2.54E-35	147
Hspe1	8.71E-34	164
Cat	2.44E-30	219
Hspd1	6.93E-13	1262
Sod2	1.25E-10	1551
Hsph1	4.48E-10	1655
Lonp1	1.08E-06	2348
Eif2a	1.49E-06	2382
Mthfd2	1.31E-05	2693
Hspa4	2.84E-05	2790
Cth	2.53E-03	3677
Nrf1	2.86E-03	3698

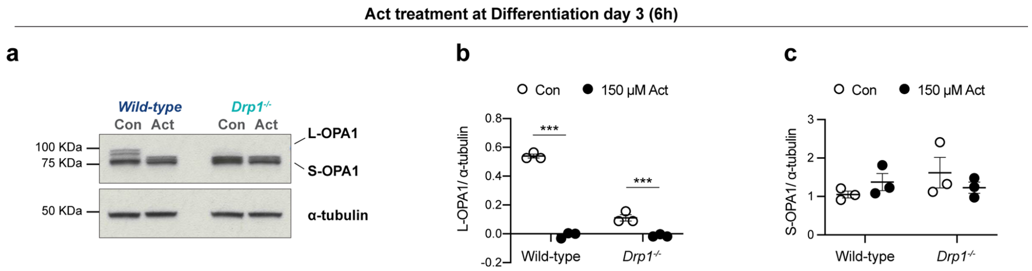
Extended Data Fig. 4 | See next page for caption.

Extended Data Fig. 4 | Cells eliminated during early mouse embryogenesis have activated stress responses. a, Overlap of genes differentially expressed along the trajectory joining winning and losing epiblast cells in CI-treated embryos (Fig. 2a and Extended Data Fig. 3a) and genes related to the unfolded protein response and integrated protein response pathways (UPR_ISR, see Supplementary Table 3). From the 32 genes related to the UPR & ISR pathways, 12 are down-regulated in loser cells, 8 genes are up-regulated in loser cells, and 12 genes are not differentially expressed between loser and winner cells. There is a statistically significant enrichment of UPR&ISR genes among the up-regulated genes in loser cells (odds ratio=3.0, p-value=0.012). The intersection between UPR-ISR genes and the down regulated genes is not significant (odds ratio=1.2, p value=0.69). **b-c**, List of genes from UPR-ISR pathways that are statistically significantly up-regulated (**b**) or down-regulated (**c**) in loser cells. **d**, Scatterplots with the expression levels of genes involved in stress responses in epiblast cells from CI-treated embryos as a function of cells' losing score. **e**, Experimental design with the approach taken to validate the expression of the stress response marker DDIT3 in epiblast cells from DMSO or CI-treated embryos. **f**, Representative micrographs of DMSO (upper panel) or CI-treated embryos (100 µM, lower panel) stained for DDIT3, quantified in (**g**). Nuclei are labelled with Hoechst. In control embryos (DMSO-treated), dying cells in the cavity show very high DDIT3 expression (arrow), while live cells in the epiblast of the CI-treated embryos show more modest levels of DDIT3 expression (arrowheads). Scale bar = 20 µm. **g**, Quantification of the percentage of epiblast cells with nuclear DDIT3 expression. N = 10 DMSO and N = 9 CI-treated embryos. Data shown as mean ± SEM. See methods for details on statistical analysis.

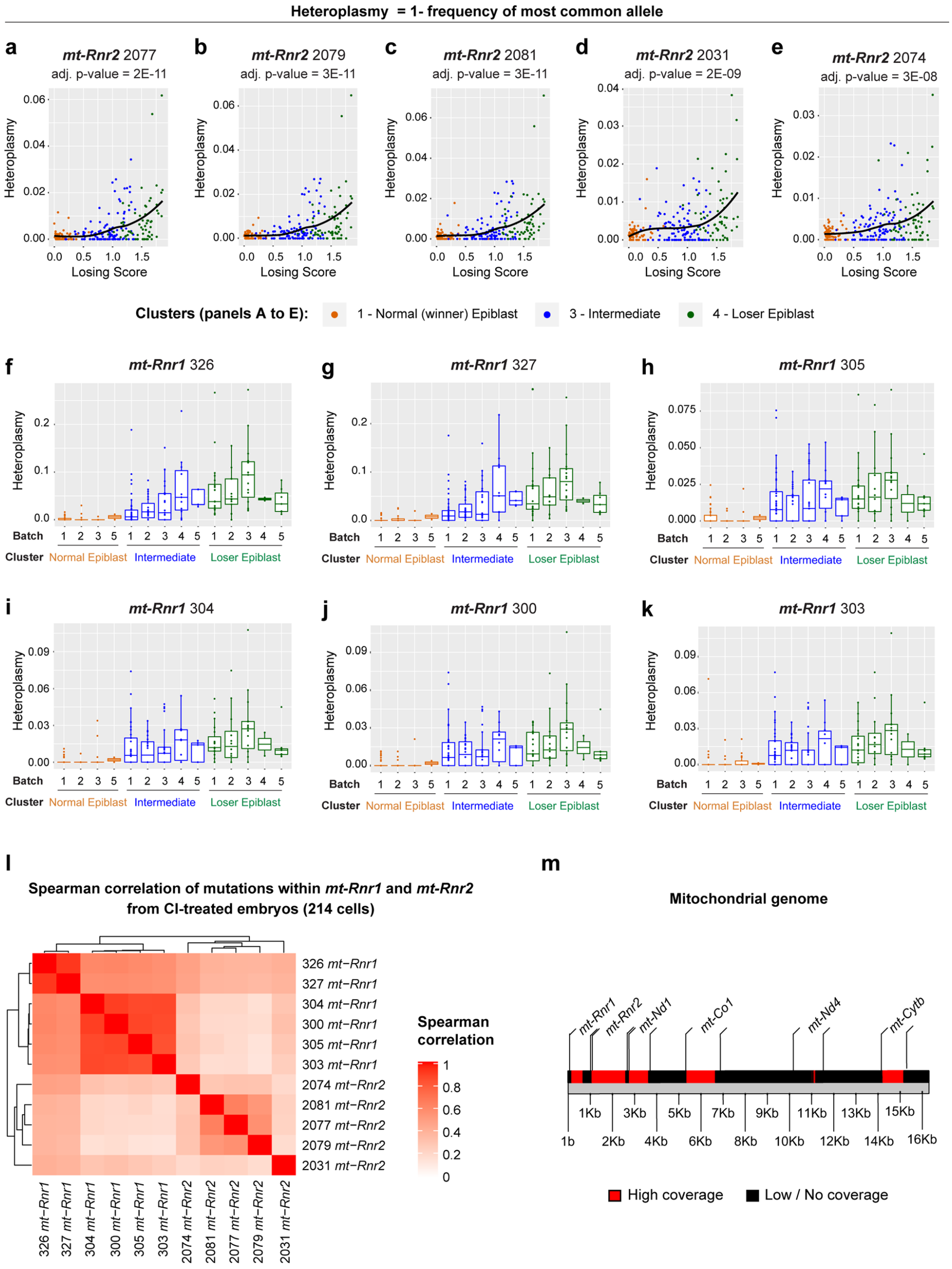


Extended Data Fig. 5 | See next page for caption.

Extended Data Fig. 5 | Mitochondrial function in wild-type, *Bmpr1a*$^{-/-}$ and 4n mESCs. a-d, Metabolic flux analysis of wild-type and *Bmpr1a*$^{-/-}$ mESCs. OCR profile and metabolic parameters assessed during the mitochondria stress test performed in pluripotency conditions (**a**). ECAR profile and metabolic parameters assessed during the glycolysis stress test performed in pluripotency conditions (**b**). Metabolic parameters from the mitochondria stress test found to be similar between wild-type and *Bmpr1a*$^{-/-}$ mESCs during differentiation – day 3 (**c**). Metabolic parameters from the glycolysis stress test found to be similar between wild-type and *Bmpr1a*$^{-/-}$ mESCs during differentiation – day 3 (**d**). Data obtained from 3 (**a,b**) or 5 (**c,d**) independent experiments, with 5 replicates per cell type in each assay. **e-f**, Analysis of mitochondrial membrane potential ($\Delta\psi$m) in defective mESCs maintained in pluripotency conditions, in separate or co-culture. Representative histograms of TMRM fluorescence and quantification for wild-type and *Bmpr1a*$^{-/-}$ (**e**) and wild-type and 4n (**f**). **g**, Analysis of mitochondrial ROS in wild-type and *Bmpr1a*$^{-/-}$ mESCs undergoing differentiation in separate or co-culture: representative histograms of mitoSOX Red fluorescence and quantification of the percentage of mitoSOX positive cells. Data shown as mean ± SEM from 3 (**e-f**) or 5 (**g**) independent experiments. See methods for details on statistical analysis.

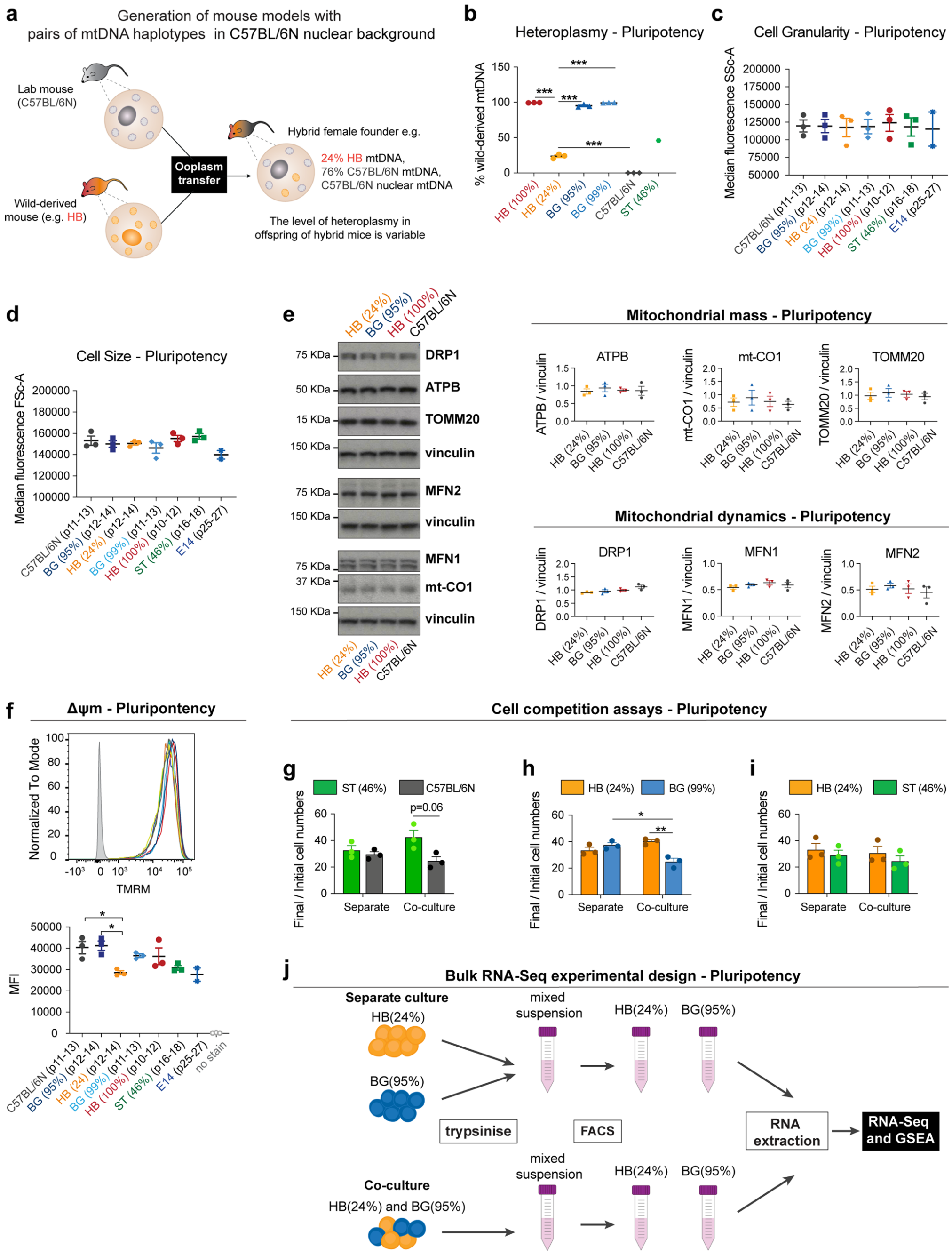


Extended Data Fig. 6 | Effect of actinonin in OPA1 expression in wild-type and *Drp1*$^{-/-}$ cells. a, Western blot analysis of OPA1 expression in wild-type and *Drp1*$^{-/-}$ cells treated with actinonin (Act, 150 μM) during 6 hours on the third day of differentiation, quantified in (**b-c**). **b-c**, Expression levels of L-OPA1 (**b**) and S-OPA1 (**c**) relative to α-tubulin. Data shown as mean ± SEM of 3 independent experiments.



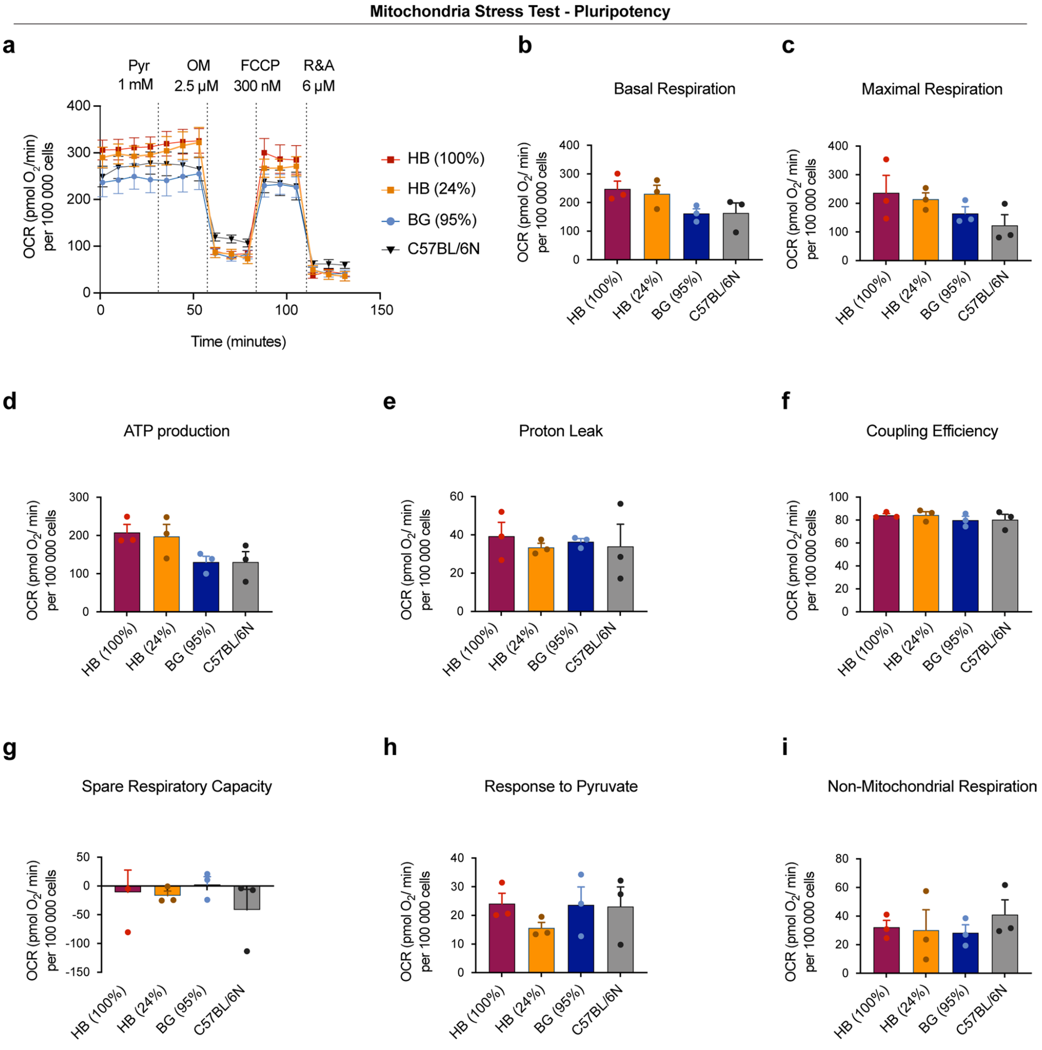
Extended Data Fig. 7 | See next page for caption.

Extended Data Fig. 7 | Analysis of SNPs in mtDNA in epiblast cells. a-e, mtDNA heteroplasmy (plotted as Heteroplasmy = 1- frequency of most common allele) in epiblast cells from CI-treated embryos for five positions within the *mt-Rnr2* gene. All these positions have an heteroplasmy that increases with the cells' losing scores in a statistically significant way - the adjusted p-values are indicated at the top of each plot. **f-k**, The variation in the heteroplasmy across the CI-treated cells is not due to a batch effect for the 6 significant positions within the *mt-Rnr1* gene. The number of cells analysed per cluster (and batch) is as follows: number of cells in Normal Epiblast :42 (1),16 (2),18 (3),0 (4),2 (5); number of cells in Intermediate: 42 (1), 28 (2), 28(3), 12 (4), 5 (5); number of cells in Loser Epiblast: 22 (1), 15(2), 20 (3), 2 (4), 7 (5). **l**, Correlation between the mtDNA heteroplasmy at all the statistically significant positions, six within the gene mt-*Rnr1* and five within the gene mt-*Rnr2*. **m**, Schematic representation of the mitochondrial genome showing in red the positions that passed our filtering based on coverage and were considered for the heteroplasmy analysis. Only the genes that include these positions are indicated. See methods for details on statistical analysis.



Extended Data Fig. 8 | See next page for caption.

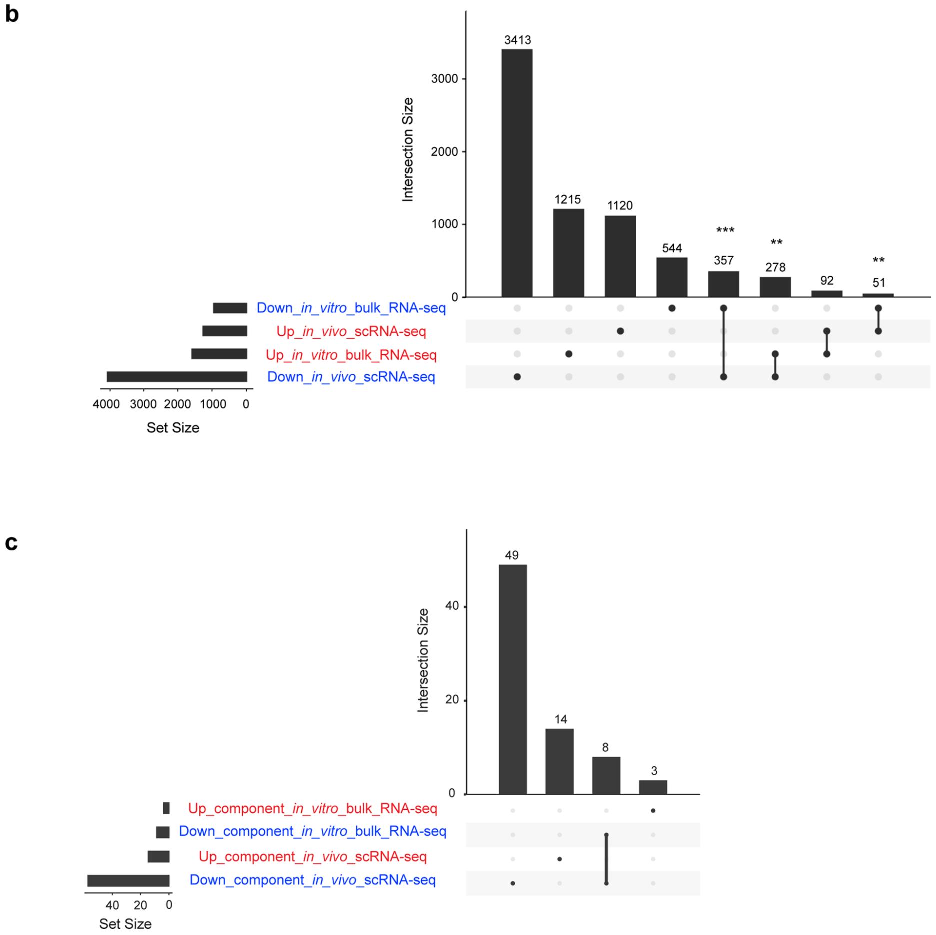
Extended Data Fig. 8 | Changes in mtDNA sequence are enough to trigger cell competition. a, Illustration of the process of derivation of the mESCs lines from mice that are hybrid between the wild-caught strains (BG, HB or ST) and the lab mouse (C57BL/6N). These hybrid mice were generated elsewhere[16] by ooplasmic transfer: the zygote of a C57BL/6N mouse was injected with ooplasm from a wild-caught mouse (orange, HB pictured). Therefore, these hybrid mice contain the nuclear background of the C57BL/6N strain and the mtDNA of wild-caught strain and potentially C57BL/6N mtDNA (heteroplasmic mice strains). mESCs lines were derived from the hybrid mice and characterised. **b-f**, Characterisation of the derived cell lines by flow cytometry, during pluripotency, in comparison to the wild-type cell line used in previous experiments (E14, 129/Ola background). Heteroplasmy analysis of the derived mESC lines from the hybrid mice, indicating the percentage of wild-derived mtDNA (**b**). Cell granularity (internal complexity) given as median fluorescence intensity of SSc-A laser (**c**). Cell size given as median fluorescence intensity of FSc-A laser (**d**). Analysis of the expression of mitochondrial markers: representative western blot and quantification of markers of mitochondrial mass (ATPB, mt-CO1 and TOMM20) and mitochondrial dynamics (DRP1, MFN1and MFN2), relative to vinculin, in cells derived from hybrid mice (**e**). **f**, Representative histograms and quantification of median TMRM fluorescence, indicative of $\Delta\psi$m, for the hybrid cell lines derived, in comparison to the wild-type cell line used in previous experiments (E14, 129/Ola background). **g-i**, Cell competition assays between hybrid cell lines maintained in pluripotency culture conditions. The ratio of final/initial cell numbers in separate or co-culture is shown. **j**, Experimental design for RNA-Seq and gene set enrichment analysis (GSEA). The isolation of RNA from winner HB(24%) and loser BG(95%) cells was performed after three days in separate or co-culture conditions, once cells have been subjected to FACS to isolate the two populations form mixed cultures. Data shown as mean ± SEM of 3 independent experiments. See methods for details on statistical analysis.



Extended Data Fig. 9 | Metabolic flux analysis of the cells with different mtDNA variants: HB(100%), HB(24%), BG(95%) and C57BL/6N. a, OCR profile during mitochondria stress test performed in pluripotency maintenance conditions. **b-i**, Metabolic parameters assessed during the during the mitochondria stress test performed in pluripotency conditions. Data obtained from 3 independent experiments, with 5 replicates per cell type in each assay. Error bars represent SEM. See methods for details on statistical analysis.

a

Source	Term	Adjusted p-value
GO:CC	mitochondrial protein complex	5.91E-05
GO:CC	inner mitochondrial membrane protein complex	8.84E-04
GO:CC	mitochondrial inner membrane	8.93E-04
GO:CC	mitochondrial respirasome	2.44E-03
GO:CC	respiratory chain complex	3.89E-03
GO:CC	respirasome	6.50E-03
GO:CC	mitochondrial part	1.06E-02
GO:CC	organelle inner membrane	4.65E-02
KEGG	oxidative phosphorylation	7.71E-04
KEGG	Huntington disease	2.35E-03
WP	electron transport chain	1.26E-03



Extended Data Fig. 10 | See next page for caption.

Extended Data Fig. 10 | Common features of scRNA-seq and bulk RNA-seq datasets. a, Terms significantly enriched among genes downregulated in BG(95%) (loser) ESCs *in vitro* when co-cultured with HB(24%) cells. The loss of mitochondrial activity emerges as a common feature between loser cells *in vivo* and *in vitro*. The gene enrichment analysis was performed using g-profiler tool (see Methods) and p-values were adjusted for multiple comparisons using the g:Profiler algorithm g:SCS (10.1093/nar/gkm226). **b**, Intersection between differentially expressed genes along the trajectory from winning to losing epiblast cells ('in_vivo_scRNA-seq'; Fig. 2a and Extended Data Fig. 3a, and genes differentially expressed between co-cultured HB(24%) (winner) and BG(95%) (loser) ESCs ('in_vitro_bulk_RNA-seq'). 'Up' and 'Down' here refer to genes up- or down-regulated in loser cells. For the intersection between down-regulated genes from scRNA-seq (*in vivo*) and down-regulated genes from bulk RNA-seq (*in vitro*): p-value, 1.71E-12; odds ratio 1.80. For the intersection between down-regulated genes from scRNA-seq (*in vivo*) and up-regulated genes from bulk RNA-seq (*in vitro*): p-value, 5.20E-3; odds ratio 0.67. For the intersection between up-regulated genes from scRNA-seq (*in vivo*) and down-regulated genes from bulk RNA-seq (*in vitro*): p-value, 4.87E-3; odds ratio 0.80. The intersection between up-regulated genes from sc-RNA-seq (*in vivo*) and up-regulated genes from bulk RNA-Seq (*in vitro*) is not statistically significant: p-value: 0.30, odds ratio 1.14. **c**, Intersection between the significantly enriched terms in genes upregulated or downregulated in loser cells in the epiblast of CI-treated embryos ('*in_vivo*_scRNA-Seq') or in our *in vitro* model of competition between co-cultured HB(24%) (winner) and BG(95%) (loser) ESCs ('*in_vitro*_bulk_RNA-seq'). All the terms enriched among downregulated genes *in vitro* are also enriched *in vivo*. See methods for details on statistical analysis.

nature research

Corresponding author(s): Antonio Scialdone & Tristan Rodriguez

Last updated by author(s): May 24, 2021

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Vi-CELL XR Software (XR v2.04, Beckman Coulter) was used for automated viable cell counts and FACSDiva software for Windows was used for flow cytometry data collection. Excel v16 (currently v16.49) for Mac OS v10 was used for general data compilation prior to statistical analysis.
Data analysis	The software used for RNA-seq data analysis was Salmon v0.8.2, STAR v2.7, samtools v1.11, R studio (https://rstudio.com), Python v3.8.5 and IPA v01-13 (Qiagen). Specific packages used for RNA-seq data analysis are scran v1.6.9, UMAP v0.2.0.0, dynamicTreeCut v1.63.1, Seurat v4.0.1, destiny v2.6.2, GAM v1.16, EdgeR v3.20.9 and scanpy v1.4.7. Microscopy imaging analysis was performed with Fiji. Statistical analysis was performed with GraphPad Prism v8. Western Blot band intensity determined with Image Studio Lite v5.0 (LI-COR). All these were run on Mac OS v10. Flow cytometry analysis data performed with FlowJo v9 & v10.0.7r2 and confocal imaging deconvolution on Huygens (Scientific Volume Imaging) for Windows. Seahorse data normalization was done with Wave Desktop v2.6 (Agilent) run on Windows 10. Data was then plotted subjected to statistical analysis with GraphPad Prism v8.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Authors can confirm that all relevant data are included in the paper and/ or its supplementary information files. Source data for Figures 2-5,7 and for Extended Data

Figures 4-6, 8-9 are provided as Excel files with the paper. Due to big size of files, source data for Figure 6 and Extended Data Figure 7 are available from https://drive.google.com/drive/folders/1hSQ_otFYUtxT1t8rpN2sMCDMIH6FIcnp. RNA-seq raw as well as processed data are available through ArrayExpress, accession numbers E-MTAB-8640, for scRNA-seq data, and E-MTAB-8692, for bulk RNA-seq data.

All the code used for generating the figures in the paper is available at https://github.com/ScialdoneLab/Cell-Competition-Paper-Figures.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Sample size was calculated based on our experience with similar previous experiments.
Data exclusions	No data were excluded.
Replication	Experimental observations were reproducible, as indicated by individual data points plotted and sample size number disclosed in figure legends.
Randomization	For cell culture experiments, plate wells were randomly assigned between treatment and control groups within each cell type. For experiments performed in mouse embryos, embryos from different litters were pooled and randomly assigned to control or treatment groups.
Blinding	If n> 4 investigators were not blinded to group allocation but instead groups were blinded during data collection and analysis analysis. With smaller sample numbers blinding was not feasible as the order of sample collection was determined by the well the cells came from and this order was easy to remember. Additionally we relied on unbiased measurements of quantitative parameters.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study	
☐	☒	Antibodies
☐	☒	Eukaryotic cell lines
☒	☐	Palaeontology
☐	☒	Animals and other organisms
☒	☐	Human research participants
☒	☐	Clinical data

Methods

n/a	Involved in the study	
☒	☐	ChIP-seq
☐	☒	Flow cytometry
☒	☐	MRI-based neuroimaging

Antibodies

Antibodies used	Rabbit anti-ATF-4 (CST-11815, Cell Signaling Technology; RRID:AB_2616025);
	Goat anti-rabbit Alexa Fluor 568 (A-11011, Invitrogen; RRID: AB_143157);
	Goat anti-mouse Alexa Fluor 488(A-11001, Invitrogen; RRID: AB_2534069).
	Mouse anti-ATPB (ab14730, Abcam RRID:AB_301438);
	Mouse anti-CHOP (CST-2895, Cell Signaling Technology; RRID:AB_2089254);
	Rabbit anti-Cleaved caspase-3 (CST-9664, Cell Signaling Technology; RRID:AB_2070042);
	Mouse anti-mt-CO1 (ab14705, Abcam; RRID: AB_2084810);
	Rabbit anti-DRP1 (CST-8570, Cell Signaling Technology; RRID: AB_10950498);
	Rabbit anti-peIF2alpha(Ser51) (CST-9721, Cell Signaling Technology; RRID:AB_330951);
	Mouse anti-MFN1 (ab57602, Abcam; RRID: AB_2142624);
	Mouse anti-MFN2 (ab56889, Abcam; RRID: AB_2142629);
	Goat anti-mouse HRP conjugated (sc-2005, Santa Cruz Biotechnology; RRID: AB_631736);
	Rat anti-NANOG (14-5761-80, eBioScience; RRID: AB_763613);
	Mouse anti-OPA1 (612606, BD Biosciences; RRID: AB_399888);
	Rabbit anti-PCNA (ab18197, Abcam; RRID:AB_444313);
	Rabbit anti-prpS6 (CST-5364, Cell Signaling Technology; RRID: AB_10694233);

	Goat anti-rabbit HRP conjugated (sc-2004, Santa Cruz Biotechnology; RRID: AB_631746);
	Rabbit anti-TOMM20 (sc-11415, Santa Cruz Biotechnology; RRID: AB_2207533);
	Rabbit anti-TOMM20 (CST-42406, Cell Signaling Technology; RRID: AB_2687663);
	Rabbit anti-α-Tubulin (CST- 2144, Cell Signaling Technology; RRID: AB_2210548)
	Mouse anti-Vinculin (V9131, Sigma-Aldrich; RRID: AB_477629).

Validation	All antibodies used were commercially available and therefore validated by the companies. We have performed our own validation in the lab by including samples that were not probed with primary antibody (incubated with secondary antibody only) alongside the complete staining or blotting protocol. Additionally, we compared target protein expression patterns with the ones seen in relevant published literature.

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)	All cell lines used were mouse embryonic stem cells (mESCs). E14 cells, the wild-type mESCs (RRID: CVCL_C320), were a gift from Prof A. Smith (Cambridge). Tetraploid cells (4n), mESCs null for Bmpr1a , mESCs null for both Bmpr1a and p53 are described elsewhere: Di-Gregorio et al., 2007, Sancho et al., 2013; and Bowling et al., 2018, respectively. Cells null for Mfn2 and cells null for Drp1 are described in this manuscript.
Authentication	Species authentication was performed by PCR or RNA seq.
Mycoplasma contamination	All cell lines tested negative for mycoplasma contamination.
Commonly misidentified lines (See ICLAC register)	No cell lines used in this study were found listed in the database of known misidentified cell lines.

Animals and other organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research

Laboratory animals	For scRNA-seq and validation experiments, pregnant CD1 mice purchased from Charles River were 6-10 weeks old. Embryos were dissected at embryonic day 5.5 (E5.5). For the derivation of hybrid mESC lines, embryos at morula stage (E2.5) were isolated from hybrid mouse strains generated elsewhere (Burgstaller et al., 2014). These contain the mtDNA of C57BL/6N lab mouse and mtDNA variants from wild-caught mice.
Wild animals	No wild animals were used in this study.
Field-collected samples	This study did not involve field-collected samples.
Ethics oversight	All animal work was done in accordance with the Home Office's Animals (Scientific Procedures) Act 1986 and covered by the Home Office project license PBBEBDCDA.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Flow Cytometry

Plots

Confirm that:

☒ The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).

☒ The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).

☐ All plots are contour plots with outliers or pseudocolor plots.

☒ A numerical value for number of cells or percentage (with statistics) is provided.

Methodology

Sample preparation	For a flow cytometry performed in epiblast cells isolated from embryos, embryos were dissected ar E5.5 from pregnant CD1 mouse females and cultures overnight in N2B27 poor media with pan-caspase inhibitors (100 µM, Z-VAD-FMK, FMK001, R&D Systems, USA) or equal volume of vehicle (DMSO) as control. On the following morning, to avoid misleading readings, epiblasts were isolated initially by an enzymatic treatment with of 2.5% pancreatin, 0.5% trypsin and 0.5% polyvinylpyrrolidone (PVP40) - all from Sigma-Aldrich - to remove the visceral endoderm (VE). Embryos were treated during 8 min at 4ºC, followed by 2 min at RT. The VE was then pealed with the forceps and the extraembryonic ectoderm removed to isolate the epiblasts. Twelve embryo epiblasts were pooled per treatment condition and dissociated into single cells with 600 µL Acccutase (A6964, Sigma, UK) during 12 min at 37ºC, tapping the tube every two minutes. Accutase was then neutralised with equal volume of FCS, cells span down and stained with 10 nM of the TMRM (T5428, Sigma, UK) prepared in N2B27 media. After incubating for 15 min at 37ºC, cells were pelleted again and re-suspended in 3% FCS in DPBS. Sytox blue (1:1000, S34857, ThermoFisher Scientific, UK), was used as viability staining.

	Quantitative analysis of mitochondrial membrane potential (Δψm) and mitochondrial ROS was performed by flow cytometry. Cells were grown in pluripotency or differentiating conditions, dissociated and pelleted to obtain 2E05 cells per sample for the staining procedure. For TMRM staining in mESCs, 2E05 cells of each cell line were resuspended in 200 µL of 10 nM TMRM (T5428, Sigma, UK), prepared in N2B27 media. Cells were incubated at 37ºC for 15 min, and then resuspended in FC buffer (3% FCS in DPBS). For the analysis of mitochondrial ROS, cells were grown in differentiating conditions and stained on the third day of culture. Briefly, 2E05 cells of each cell line were resuspended in 200 µL of 5 µM solution of MitoSOX (M36008, Invitrogen, UK) prepared in N2B27 media. Cells were incubated at 37ºC for 15 min, and then resuspended in FC buffer. Sytox blue was used as viability staining. Stained cell suspensions with TMRM or MitoSOX were analysed in BD LSRII flow cytometer operated through FACSDiva software (Becton Dickinson Biosciences, UK). For TMRM fluorescence detection the yellow laser was adjusted for excitation at λ=562 nm, capturing the emission light at λ=585 nm for TMRM. MitoSOX fluorescence was analysed with the violet laser adjusted for excitation at λ=405 nm, capturing the emission light at λ=610 nm. In the case of GFP-labelled cell lines, for GFP fluorescence detection the blue laser was adjusted for excitation at λ=488 nm, capturing the emission light at λ=525 nm.
Instrument	BD LSRII cell analyser (Becton Dickinson Biosciences, UK) for cell competition assays, TMRM and MitoSOX staining analysis or BD FACSAria III cell sorter (Becton Dickinson Biosciences, UK) for sorting experiments.
Software	FACSDiva software (Becton Dickinson Biosciences, UK)
Cell population abundance	For cell TMRM and MitoSOX staining analysis, cell population abundance is described in the manuscript figures and corresponding source data. This was determined with FlowJo software. The proportion of cells sorted prior to bulk RNA-seq was determined with FACSDiva software. For scRNA-seq samples, post-sort fractions were determined by computational analysis as described in the Methods.
Gating strategy	The gating strategy is exemplified in the Supplementary Information file, both for experiments using mitochondrial dyes (TMRM or MitoSOX) and sorting of cells within a mixed population based on GFP label prior to bulk RNA collection and sequencing. Briefly, cell debris were excluded on FSC-A vs SCC-A plots. Consequentially, single cells were isolated both with FSC and SSC laser plots and, from the single cell populations, only live cells were considered, based on viability staining applied (Sytox Blue of propidium iodine). Positive staining signal was considered to be equal or above the magnitude of 10^3 on the logarithmic scale of fluorescence intensity for the relevant fluorophore (Sytox Blue shown in example: live cells remain unstained). Flow cytometry data for TMRM and MitoSOX is presented in the form of univariate histogram plots, with x-axis re-labeled with fluorophore name. As plots show more than one sample, data is presented normalised to mode. When distinction between high and low TMRM or MitoSOX positive and negative levels was made, the threshold cut-off value was defined at the magnitude of 5.3^3 and 10^3, respectively, on the logarithmic scale of fluorescence intensity.

☒ Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.

6.3 Publication 3

The Journal of Open Source Software

MitoHEAR: an R package for the estimation and downstream statistical analysis of the mitochondrial DNA heteroplasmy calculated from single-cell datasets

Gabriele Lubatti[*][1,2,3], **Elmir Mahammadov**[†][1,2,3], **and Antonio Scialdone**[1,2,3¶]

1 Institute of Epigenetics and Stem Cells, Helmholtz Zentrum München, Munich, Germany **2** Institute of Functional Epigenetics, Helmholtz Zentrum München, Neuherberg, Germany **3** Institute of Computational Biology, Helmholtz Zentrum München, Neuherberg, Germany ¶ Corresponding author

DOI: 10.21105/joss.04265

Software

- Review
- Repository
- Archive

Editor: Charlotte Soneson

Reviewers:

- @BatoolMM
- @juanvillada

Submitted: 21 February 2022
Published: 01 June 2022

License
Authors of papers retain copyright and release the work under a Creative Commons Attribution 4.0 International License (CC BY 4.0).

Summary

Eukaryotic cells rely on mitochondria: organelles that are equipped with their own DNA (mtDNA) to produce the energy they need. Each cell includes multiple mtDNA copies that are not perfectly identical but have differences in their sequence; such sequence variability is called heteroplasmy. mtDNA heteroplasmy has been associated with diseases (Nissanka & Moraes, 2020), which can affect cellular fitness and have an impact on cellular competition (Lima et al., 2021). Several single-cell sequencing protocols provide the data to estimate mtDNA heteroplasmy, including single-cell DNA-seq, RNA-seq, and ATAC-seq, in addition to dedicated protocols like MAESTER (Miller et al., 2022). Here, we provide MitoHEAR (Mitochondrial HEteroplasmy AnalyzeR), a user-friendly software package written in R that allows this estimation as well as downstream statistical analysis of the mtDNA heteroplasmy calculated from single-cell datasets. MitoHEAR takes as input BAM files, computes the frequency of each allele and, starting from these, estimates the mtDNA heteroplasmy at each covered position for each cell.

The analysis parameters (e.g., the filtering of the mtDNA positions based on read quality and coverage) are easily tuneable. Moreover, statistical tests are available to explore the dependency of the mtDNA heteroplasmy on continuous or discrete cell covariates (e.g., culture conditions, differentiation states, etc.), as extensively shown in the included detailed tutorials.

Statement of need

Although mtDNA heteroplasmy has important consequences on human health (Stewart & Chinnery, 2015) and embryonic development (Floros et al., 2019), there are still many open questions on how heteroplasmy affects cells' ability to function and how cells keep it under control. With the increasing availability of single-cell data, many questions can begin to be answered. Still, it is essential to have efficient and streamlined computational tools that enable researchers to estimate and analyse mtDNA heteroplasmy. Existing packages (Calabrese et al., 2014; Huang & Huang, 2021; Prashant et al., 2021) focus only on the first step of quantifying heteroplasmy from BAM files, and do not provide any specific tools for further statistical analyses or plotting. MitoHEAR covers all steps of the analysis in a unique user-friendly package, with highly customisable functions. Starting from BAM files, MitoHEAR estimates heteroplasmy and offers several options for downstream analyses. For example, statistical tests are provided to investigate the relationship of the mtDNA heteroplasmy with continuous or

---

[*] first author
[†] co-author

discrete cell covariates. Moreover, it includes plotting functions to visualise heteroplasmy and allele frequencies and to perform hierarchical clustering of cells based on heteroplasmy values.

Key functions

The two main functions of `MitoHEAR` are:

1. `get_raw_counts_allele`: A parallelised function that relies on Rsamtools and generates the raw counts matrix starting from BAM files, with cells as rows and bases with the four possible alleles as columns.
2. `get_heteroplasmy`: Starting from the output of `get_raw_counts_allele`, this function computes the matrix with heteroplasmy values (defined as 1 minus the frequency of the most common allele) and the matrix with allele frequency values, for all the cells and bases that pass a filtering procedure.

Among the downstream analyses implemented in the package are:

- Several statistical tests (e.g., Wilcoxon rank-sum test) for the identification of the mtDNA positions with the most different levels of heteroplasmy between discrete groups of cells or along a trajectory of cells (i.e., cells sorted according to a diffusion pseudo-time) (**Figure 1** and **Figure 2**).
- Plotting functions for the visualisation of heteroplasmy and the corresponding allele frequency values among cells.
- Unsupervised hierarchical clustering of cells based on a distance matrix defined from the angular distance of allele frequencies that could be relevant for lineage tracing analysis (Ludwig et al., 2019) (**Figure 3**).

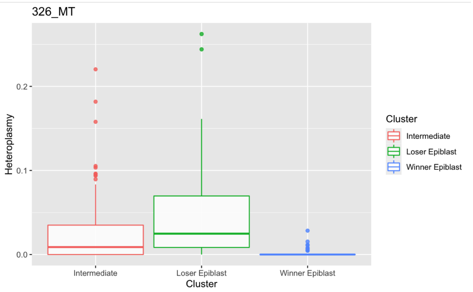


Figure 1: Example of an output plot generated by MitoHEAR showing heteroplasmy values at a given position estimated from single cells in three clusters indicated on the x-axis. Data from Lima et al. (2021).

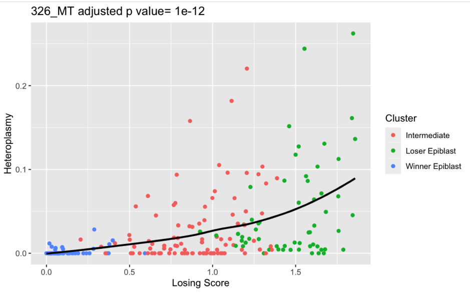


Figure 2: Example of an output figure generated by MitoHEAR where the heteroplasmy is plotted as a function of the pseudo-time coordinate of each cell. Cells are classified into three clusters. The heteroplasmy shows a statistically significant change along the pseudo-time, as indicated by the adjusted p-value reported at the top, which is computed by a generalised additive model fit. Data from Lima et al. (2021).

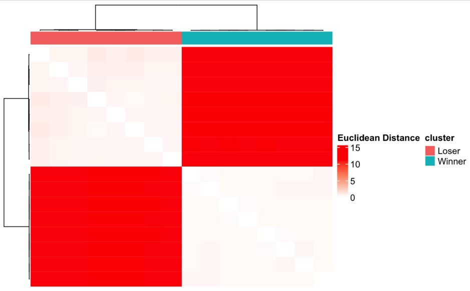


Figure 3: Unsupervised hierarchical clustering of cells based on a distance matrix defined from the angular distance of allele frequencies. The data shown is bulk RNA-seq mouse data from two mtDNA cell lines labelled *Loser* and *Winner*. Data from Lima et al. (2021).

The package has been used in a recently published paper (Lima et al., 2021), where we revealed that cells with higher levels of heteroplasmy are eliminated by cell competition in mouse embryos and are characterised by specific gene expression patterns.

References

Calabrese, C., Simone, D., Diroma, M. A., Santorsola, M., Guttà, C., Gasparre, G., Picardi, E., Pesole, G., & Attimonelli, M. (2014). MToolBox: a highly automated pipeline for heteroplasmy annotation and prioritization analysis of human mitochondrial variants in high-throughput sequencing. *Bioinformatics*, *30*(21), 3115–3117. https://doi.org/10.1093/bioinformatics/btu483

Floros, V., Pyle, A., Dietmann, S., Wei, W., Tang, W., Irie, N., Payne, B., Capalbo, A., Noli, L., Coxhead, J., Hudson, G., Crosier, M., Strahl, H., Khalaf, Y., Saitou, M., Ilic, D., Surani, M., & Chinnery, P. (2019). Segregation of mitochondrial DNA heteroplasmy through a developmental genetic bottleneck in human embryos. *Nature Cell Biology*. https://doi.org/10.1038/s41556-017-0017-8

Huang, X., & Huang, Y. (2021). Cellsnp-lite: an efficient tool for genotyping single cells. *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btab358

Lima, A., Lubatti, G., Burgstaller, J., Hu, D., Green, A., Gregorio, A. D., Zawadzki, T., Pernaute, B., Mahammadov, E., Dore, M., Sanchez, J. M., Bowling, S., Sancho, M., Karimi, M., Carling, D., Jones, N., Srinivas, S., Scialdone, A., & Rodriguez, T. A. (2021). Cell competition acts as a purifying selection to eliminate cells with mitochondrial defects during early mouse development. *Nature Metabolism*. https://doi.org/10.1038/s42255-021-00422-7

Ludwig, L. S., Lareau, C. A., Ulirsch, J. C., Christian, E., Muus, C., Li, L. H., Pelka, K., Ge, W., Oren, Y., Brack, A., Law, T., Rodman, C., Chen, J. H., Boland, G. M., Hacohen, N., Rozenblatt-Rosen, O., Aryee, M. J., Buenrostro, J. D., Regev, A., & Sankaran, V. G. (2019). Lineage tracing in humans enabled by mitochondrial mutations and single-cell genomics. *Cell*, *176*(6), 1325–1339.e22. https://doi.org/10.1016/j.cell.2019.01.022

Miller, T. E., Lareau, C. A., Verga, J. A., Ssozi, D., Ludwig, L. S., Farran, C. E., Griffin, G. K., Lane, A. A., Bernstein, B. E., Sankaran, V. G., & van Galen, P. (2022). Mitochondrial variant enrichment from high-throughput single-cell RNA-seq resolves clonal populations. *Nature Biotechnology*. https://doi.org/10.1038/s41587-022-01210-8

Nissanka, N., & Moraes, C. T. (2020). Mitochondrial DNA heteroplasmy in disease and targeted nuclease-based therapeutic approaches. *EMBO Reports*, *21*(3), e49612. https://doi.org/10.15252/embr.201949612

Prashant, N., Alomran, N., Chen, Y., Liu, H., Bousounis, P., Movassagh, M., Edwards, N., & Horvath, A. (2021). SCReadCounts: Estimation of cell-level SNVs expression from scRNA-seq data. *BMC Genomics*. https://doi.org/10.1186/s12864-021-07974-8

Stewart, J., & Chinnery, P. (2015). The dynamics of mitochondrial DNA heteroplasmy: Implications for human health and disease. *Nature Reviews Genetics*. https://doi.org/10.1038/nrg3966

6.4 Publication 4

ARTICLES
https://doi.org/10.1038/s41588-022-01023-0

nature genetics

Check for updates

OPEN

DNA replication fork speed underlies cell fate changes and promotes reprogramming

Tsunetoshi Nakatani[1], Jiangwei Lin[1], Fei Ji[2,3], Andreas Ettinger[1], Julien Pontabry[1], Mikiko Tokoro[4], Luis Altamirano-Pacheco[1], Jonathan Fiorentino[1,5,6], Elmir Mahammadov[1,5,6], Yu Hatano[4], Capucine Van Rechem[7], Damayanti Chakraborty[7], Elias R. Ruiz-Morales[1], Paola Y. Arguello Pascualli[1], Antonio Scialdone[1,5,6], Kazuo Yamagata[4], Johnathan R. Whetstine[7,8,9], Ruslan I. Sadreyev[2,10] and Maria-Elena Torres-Padilla[1,11] ✉

Totipotency emerges in early embryogenesis, but its molecular underpinnings remain poorly characterized. In the present study, we employed DNA fiber analysis to investigate how pluripotent stem cells are reprogrammed into totipotent-like 2-cell-like cells (2CLCs). We show that totipotent cells of the early mouse embryo have slow DNA replication fork speed and that 2CLCs recapitulate this feature, suggesting that fork speed underlies the transition to a totipotent-like state. 2CLCs emerge concomitant with DNA replication and display changes in replication timing (RT), particularly during the early S-phase. RT changes occur prior to 2CLC emergence, suggesting that RT may predispose to gene expression changes and consequent reprogramming of cell fate. Slowing down replication fork speed experimentally induces 2CLCs. In vivo, slowing fork speed improves the reprogramming efficiency of somatic cell nuclear transfer. Our data suggest that fork speed regulates cellular plasticity and that remodeling of replication features leads to changes in cell fate and reprogramming.

Cellular plasticity is an essential requirement for multicellular organisms. Cells in the early mammalian embryo are most plastic because they can generate every cell type in the body. In particular, the mouse zygote and each of the blastomeres in 2-cell-stage embryos are totipotent[1,2], because they can generate a new organism on their own without the need for carrier cells. This contrasts with pluripotent cells, which can generate all the cells in the body, but not extraembryonic tissues[3,4]. Thus, totipotent cells have greater cellular plasticity. However, the mechanisms that sustain totipotency are poorly understood.

DNA replication is a fundamental process for genetic and epigenetic inheritance. However, how the early mammalian embryo replicates its DNA and whether the acquisition of totipotency is regulated through DNA-replication-dependent mechanisms is unknown. As the molecular properties of the replication fork are central to the regulation of replication[5], we set out to investigate replication fork dynamics in totipotent cells in vivo and totipotent-like cells in culture.

Results

2CLCs and totipotent embryos have a slow replication fork speed. Totipotent-like cells resembling 2-cell-stage mouse embryos arise spontaneously in embryonic stem cell (ESC) cultures, but only in very low proportions of around 0.5%[6]. 2CLCs recapitulate several molecular features of the totipotent cells in mouse embryos and display expanded potency, including higher ability to be reprogrammed upon nuclear transfer[6–8]. Similar to 2-cell-stage embryos, 2CLCs express specific repeats such as MERVL[6,9] and thus can be identified by a fluorescent reporter under the control of the MERVL long-terminal repeat[6,10], enabling their characterization and isolation (Fig. 1a). We used DNA fiber analysis to study DNA replication and measure replication fork speed[11,12]. Analysis of replication fork speed in 2CLCs revealed a significantly slower fork speed compared with ESCs (Fig. 1b). Although ESCs displayed an expected rate of 1.34 kb min^{-1} (ref. [13]), 2CLCs had approximately half this speed (0.56 kb min^{-1}) (Fig. 1c). This suggested that totipotent-like cells in culture replicate DNA much more slowly than pluripotent stem cells. Importantly, the length of the S-phase did not change (see also below), suggesting that 2CLCs may use more origins than ESCs, to compensate for a slower fork progression. Indeed, analysis of the DNA fibers[14] indicated an increase in DNA fibers in which replication stopped after the first label, implying more termination or blockage events (Fig. 1d), consistent with increased origin usage. In agreement, visualization of replication by 5-ethynyl-2′-deoxyuridine (EdU) incorporation revealed that 2CLCs displayed a more dispersed EdU pattern and higher number of replication clusters compared with ESCs (Extended Data Fig. 1a,b).

To address whether slow replication dynamics is a feature of genuine totipotent cells, we measured replication fork speed in 2-cell-stage embryos in vivo (Fig. 1e). Notably, 2-cell-stage embryos displayed a low fork speed during their complete S-phase (median 0.33 kb min^{-1} in early, mid and late S-phase; Fig. 1f). This was in

---

[1]Institute of Epigenetics and Stem Cells, Helmholtz Zentrum München, München, Germany. [2]Department of Molecular Biology, Massachusetts General Hospital, Boston, MA, USA. [3]Department of Genetics, Harvard Medical School, Boston, MA, USA. [4]Faculty of Biology-Oriented Science and Technology, Kindai University, Wakayama, Japan. [5]Institute of Functional Epigenetics, Helmholtz Zentrum München, Neuherberg, Germany. [6]Institute of Computational Biology, Helmholtz Zentrum München, Neuherberg, Germany. [7]Massachusetts General Hospital Cancer Center and Department of Medicine, Harvard Medical School, Charlestown, MA, USA. [8]Cancer Signaling and Epigenetics Program, Fox Chase Cancer Center, Philadelphia, PA, USA. [9]Cancer Epigenetics Institute, Fox Chase Cancer Center, Philadelphia, PA, USA. [10]Department of Pathology, Massachusetts General Hospital and Harvard Medical School, Boston, MA, USA. [11]Faculty of Biology, Ludwig-Maximilians Universität, München, Germany. ✉e-mail: torres-padilla@helmholtz-muenchen.de

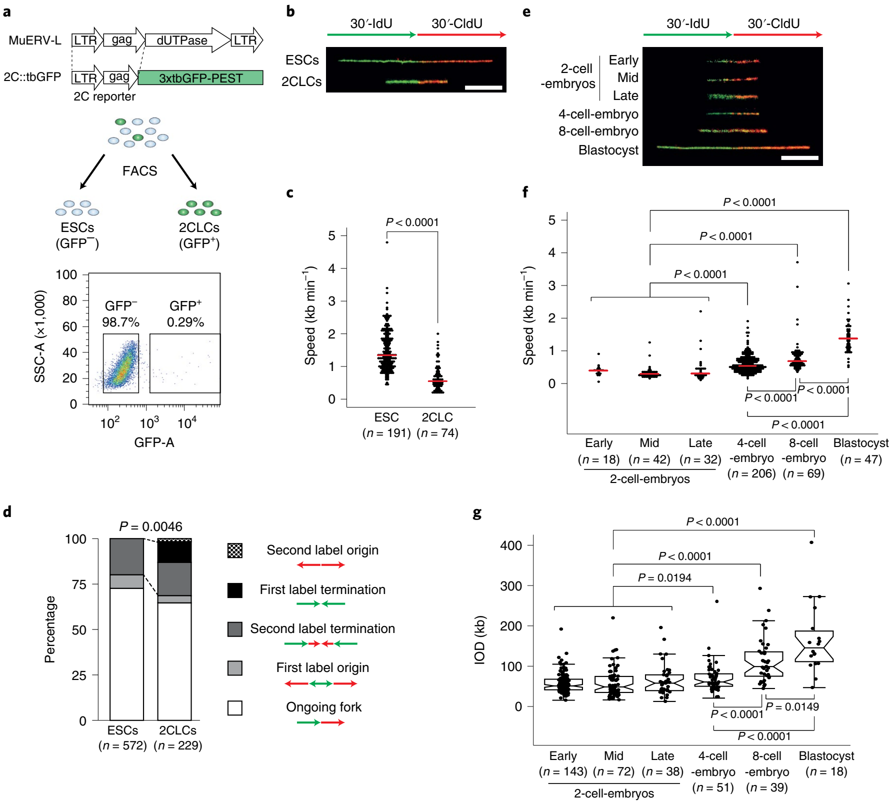


Fig. 1 | 2CLCs and totipotent embryos display slow replication fork speed. **a**, Experimental setup for isolation of ESCs (GFP⁻) and 2CLCs (GFP⁺) based on FACS. **b**–**g**, DNA fiber analysis of pluripotent and totipotent cells by sequential labeling of nascent DNA. Representative fiber images (**b**) and quantification results of fork speed (**c**) from ESCs and 2CLCs are shown. **d**, Distribution of patterns of replication derived from fiber analyses from ESCs and 2CLCs. **e**,**f**, Representative fiber images (**e**) and quantification results of fork speed (**f**) from mouse embryos at the indicated developmental stages. **g**, Quantification of the IOD at the indicated stages of mouse preimplantation embryos. In **c** and **f**, the red line indicates the median. In **g**, the boxplots show the median and the IQR and whiskers depict the smallest and largest values within 1.5 × IQR. In **c**, **f** and **g**, statistical analyses were performed with a two-sided Wilcoxon's rank-sum test. In **d**, statistical analyses were performed using a two-sided binomial test. In **b** and **e**, scale bars, 5 μm.

contrast to 4- and 8-cell-stage embryos, which displayed faster replication dynamics that increased further at the blastocyst stage (0.53 kb min^{-1}, 0.68 kb min^{-1} and 1.37 kb min^{-1}, respectively; Fig. 1e,f). The slow fork speed in 2-cell-stage embryos was accompanied by an increase in the number of replication clusters compared with ESCs, considering the difference in nuclear volume (Extended Data Fig. 1c–e), suggestive of an increase in the number of replication foci and potentially also of origins used. To explore this possibility, we quantified the proportion of 'origin label' events as well as inter-origin distance (IOD). The 2-cell-stage embryos displayed a higher ratio of origins to forks (first label origin) on average confined to the early S-phase and compared with 8-cell-stage embryos and blastocysts (Extended Data Fig. 1f). In addition, the IODs—known to inversely correlate with the number of active origins[15]—were significantly shorter in 2-cell-stage embryos, compared with 8-cell embryos and blastocysts (Fig. 1g). Thus, totipotent cells in the early embryo replicate DNA with slow fork dynamics, which increases as development proceeds. These data underscore fundamental properties of DNA replication dynamics in the early mouse embryo.

Emergence of 2CLCs requires DNA replication. Next, we reasoned that, if replication fork dynamics is relevant for 2CLC reprogramming, the S-phase may play a critical role for 2CLC emergence. To address this, we synchronized cells at G1/S using a double thymidine block, after which we removed pre-existing 2CLCs from the culture using fluorescence-activated cell sorting (FACS), and measured the number of newly emerging 2CLCs every hour after releasing the culture from the block, using the 2C MERVL-driven reporter as readout (Fig. 2a). This analysis revealed 2CLC emergence along with cell cycle progression, which reached the same proportion as the synchronized population on completion of the S-phase within ~6 h after release (Fig. 2b and Extended Data Fig. 2a). Inhibition of DNA synthesis, upon addition of aphidicolin or thymidine after release from G1/S, led to a reduction in the proportion of 2CLCs. This suggests that, although DNA synthesis partially contributes to

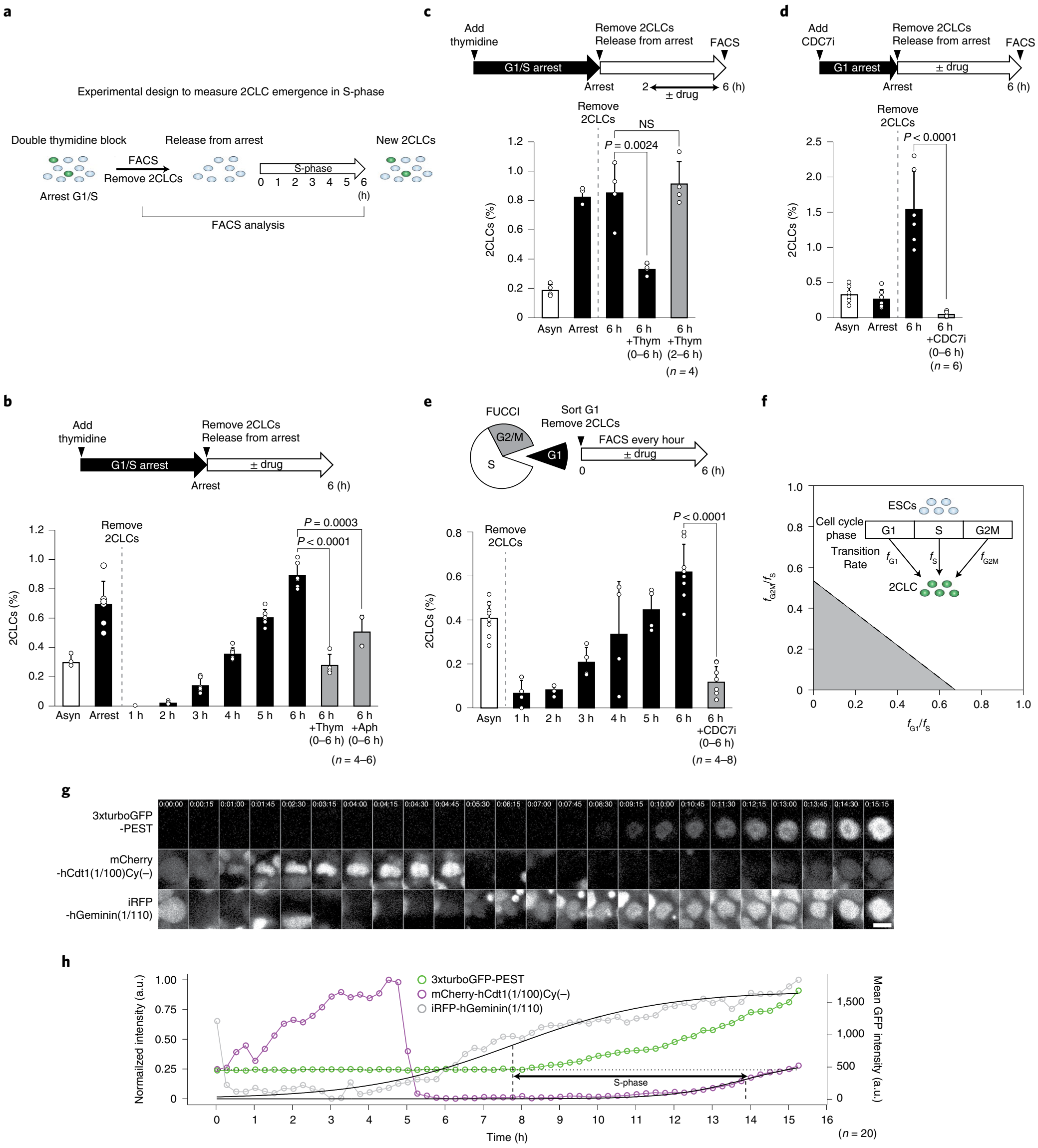
2CLC emergence, it is entry into the S-phase, which neither thymidine nor aphidicolin blocks, that is important for 2CLC reprogramming (Fig. 2b). We obtained similar results using another ESC line[6] (Extended Data Fig. 2b,c). We also investigated whether 2CLC induction is related to checkpoint activation, but obtained no evidence for the requirement of checkpoint activation in 2CLC induction or increased γH2A.X levels in 2CLCs compared with ESCs[10] (Extended Data Fig. 2d,e). To address whether completion of the S-phase is important for 2CLC induction, we added thymidine 2 h after release from the G1/S block. This resulted in full 2CLC induction (Fig. 2c), in concordance with cells progressing through the S-phase but accumulating before the G2/M peak (Extended Data Fig. 2f), suggesting that entry into the S-phase, but not necessarily completion, is relevant for 2CLC reprogramming. We then asked whether preventing origin firing affects 2CLC emergence. G1 synchronization and sustained treatment with a CDC7 kinase inhibitor (Extended Data Fig. 2g,h), which blocks MCM phosphorylation and thereby origin firing[16], resulted in an almost complete suppression of 2CLC emergence (Fig. 2d). Importantly, synchronization at the G2/M transition did not increase the proportion of 2CLCs (Extended Data Fig. 3a), suggesting that our results do not reflect cell cycle inhibition in general, but rather reflect 2CLC emergence together with DNA replication. In agreement, irreversible cell cycle arrest prevented 2CLC emergence (Extended Data Fig. 3b,c). Of note, we observed an increase in 2CLCs after G2/M release, which paralleled progression into the next S-phase and was prevented on CDK1 inhibition, which blocks origin firing[17] (Extended Data Fig. 3a,d). As cell cycle arrest using chemical inhibitors may have indirect effects, we looked at whether 2CLCs emerge during the S-phase in normal, cycling ESCs. Sorting G1 cells using the FUCCI (fluorescence ubiquitination cell cycle indicator) system[18] in the absence of any chemical arrest confirmed de novo 2CLC emergence coincident with S-phase progression (Fig. 2e and Extended Data Fig. 3e). We also performed mathematical modeling using our cell cycle data (Extended Data Fig. 3f and Methods), which indicated that 2CLCs emerge primarily during the S-phase, because the transition rates (*f*) in other phases of the cell cycle are negligible and smaller than the transition rate in the S-phase (that is, $f_{G1}, f_{G2M} < f_S$; Fig. 2f). Accordingly, direct observation with live microscopy using the FUCCI system indicated that most 2CLCs emerge together with S-phase progression (Fig. 2g,h). We conclude that 2CLC emergence occurs concomitant with DNA replication and that entry into the S-phase is key for this reprogramming.

Slowing replication fork speed induces 2CLCs. To address how S-phase enables 2CLC reprogramming and given our observations above (Fig. 1), we focused on replication fork speed. We asked whether modulating replication fork speed can regulate reprogramming toward 2CLCs. For this, we sought to reduce fork speed experimentally. The USP7 deubiquitinase modulates small ubiquitin-like modifier (SUMO) levels at sites of DNA replication, thereby regulating replication fork progression. Inhibiting USP7 decreases fork speed in human cells and fibroblasts[19]. We thus asked whether ubiquitin-specific-processing protease 7 (USP7) depletion can induce 2CLCs. *Usp7* downregulation in ESCs led to reduced fork speed (Fig. 3a,b and Extended Data Fig. 4a) without significantly affecting cell proliferation (Extended Data Fig. 4b). Strikingly, *Usp7* RNA interference (RNAi) led to more than approximately sixfold induction of 2CLCs (Fig. 3c) and a concomitant increase in the transcription of the MERVL retrotransposon (Extended Data Fig. 4c), a marker of 2CLCs and 2-cell-stage embryos. The 2CLCs induced upon *Usp7* knockdown displayed typical 2CLC features, such as ZSCAN4 expression, downregulation of OCT4 (POU5F1), chromocenter dispersion (Fig. 3d) and a high gene expression profile overlap with endogenous 2CLCs[7] (Fig. 3e and Extended Data Fig. 4d), including upregulation of MERVL and MT2_Mm and an enrichment of '2C' genes (Supplementary Tables 1 and 2 and Extended Data Fig. 4e–g). Unsupervised clustering of transcriptomes from early embryos[20], ESCs and several 2CLC datasets[6,7,21,22] confirmed that USP7 knockdown-induced 2CLCs are transcriptionally more similar to 2CLCs and 2-cell-stage embryos (Fig. 3f). In line with their 2CLC identity[8,23], they express the transcription factor *Dux* and MERVL activation—as determined using the 2C reporter—that was dependent on *Dux* (Fig. 3e and Extended Data Fig. 4h). As USP7 can have multiple functions throughout the cell cycle[24,25], we next asked whether USP7 functions to regulate 2CLC emergence during or outside the S-phase. For this, we first depleted USP7 using small interfering (si)RNA, then synchronized cells at the G2/M transition using a PLK1 (polo-like kinase 1) inhibitor (PLKi) and cultured them for another 6 h (Extended Data Fig. 4i,j), after which we determined the number of 2CLCs. Addition of the PLK1i prevented induction of 2CLCs after synchronization at G2/M (Extended Data Fig. 4k), suggesting that the effect of USP7 depletion in inducing 2CLCs occurs before G2. To address this directly, we engineered a knock-in ESC homozygous *Usp7* allele with an auxin-induced degron (AID) (Extended Data Fig. 4l), which enables precise temporal control of USP7 protein using auxin (Extended Data Fig. 4m). With this approach, we were able to deplete USP7 specifically from the early, mid or late S-phase (Fig. 3g). Using these conditions, we determined the impact of the temporal depletion of USP7 on 2CLC emergence in the S-phase immediately after release from double thymidine block as above (Fig. 3h). The steady-state population of 2CLCs was higher in the USP7–AID cell line, presumably because our transgene causes slightly lower USP7 expression

Fig. 2 | Emergence of 2CLCs occurs together with or after DNA replication. a, Strategy to evaluate 2CLC emergence during the S-phase. **b**, After synchronization of ESCs at G1/S by double thymidine block, existing 2CLCs were removed. The remaining cells were released from the block and cultured with or without the indicated inhibitors. Emerging 2CLCs were quantified by FACS. Asyn, asynchronized. **c**, After synchronization as in **b**, 2CLCs were removed by FACS and thymidine added 2–6 h after release to prevent S-phase completion. Emerging 2CLCs were quantified 6 h after release. NS, not significant. **d**, ESCs synchronized in G1 using a CDC7 inhibitor, after which existing 2CLCs removed. Cells were subsequently grown with or without CDC7 inhibitor and newly emerging 2CLCs were quantified 6 h after release. Barplots show mean ± s.d. Statistical analyses are by two-sided Student's *t*-test. **e**, ESCs in G1 sorted based on their FUCCI (mCherry-hCdt1(1/100)Cy(−) and iRFP-hGeminin (1/110)) fluorescence and new 2CLCs quantified hourly by FACS. The means ± s.d. of at least four independent biological replicates are shown. Statistical analyses are by two-sided Student's *t*-test. **f**. Mathematical modeling showing the quantitative relationships between the transition rates (*f*) of ESCs into 2CLCs during cell cycle phases (that is, f_{G1}, f_S and f_{G2M}). The transition rate is the probability that an ESC changes its fate to 2CLC during a given unit of time. The gray area demarcates all possible values compatible with the data: all the values of transition rates falling within the gray area fit the experimental data. As the dashed line cuts the *y* and *x* axes at values <1 for both G2/M over S (f_{G2M}/f_S, *y* axis) and G1 over S (f_{G1}/f_S, *x* axis), transitions from ESCs to 2CLCs must occur most frequently in the S-phase. **g**,**h**, Live-cell microscopy indicating that 2CLCs emerge concomitantly with S-phase progression. a.u., arbitrary units. Live-imaging stills representative of 20 time-lapse recordings of emerging 2CLCs using FUCCI (mCherry-hCdt1(1/100)Cy(−) and iRFP-hGeminin (1/110)). **h**, Quantification of the representative emerging 2CLC in **g** depicting normalized mean fluorescence intensities (mCherry, iRFP, left axis) and mean raw fluorescence (GFP, right axis) over time. The S-phase duration is indicated. The majority of cells analyzed displayed similar results, with onset of 2C::tbGFP fluorescence during the S-phase or S/G2 transition. Scale bar, 10 µm. Barplots, mean ± s.d.; dots, values of each replicate; *n*, number of biological replicates.

compared with the parental clone (Extended Data Fig. 4n). USP7 depletion resulted in a 2CLC increase, compared with basal levels, exclusively when depleted from early S-phase onward, but not from mid or late S-phase (Fig. 3h). These experiments demonstrate that entry into the S-phase and/or early S-phase is critical for 2CLC emergence.

As an orthologous approach to slow down replication fork speed, we employed low doses of hydroxyurea (HU)[26]. We verified that HU treatment led to a reduction in fork speed (Fig. 3i). HU treatment resulted in a striking, approximately tenfold increase in 2CLCs (Fig. 3j), which displayed typical 2CLC features (Fig. 3k and Extended Data Fig. 4h). As a third approach, we used RNAi to achieve partial downregulation for the ribonucleotide reductase subunits RRM1 and RRM2, known to result in reduction of fork speed[26,27](Extended Data Fig. 4o,p). Downregulation of both RRM1 and RRM2 led to a robust 2CLC increase of ~20- and 10-fold, respectively (Extended Data Fig. 4q), suggesting that slowing the replication fork speed regulates changes in cell fate and highlighting the relevance of replication fork dynamics for 2CLC reprogramming. We also addressed whether our findings may be applicable to other reprogramming

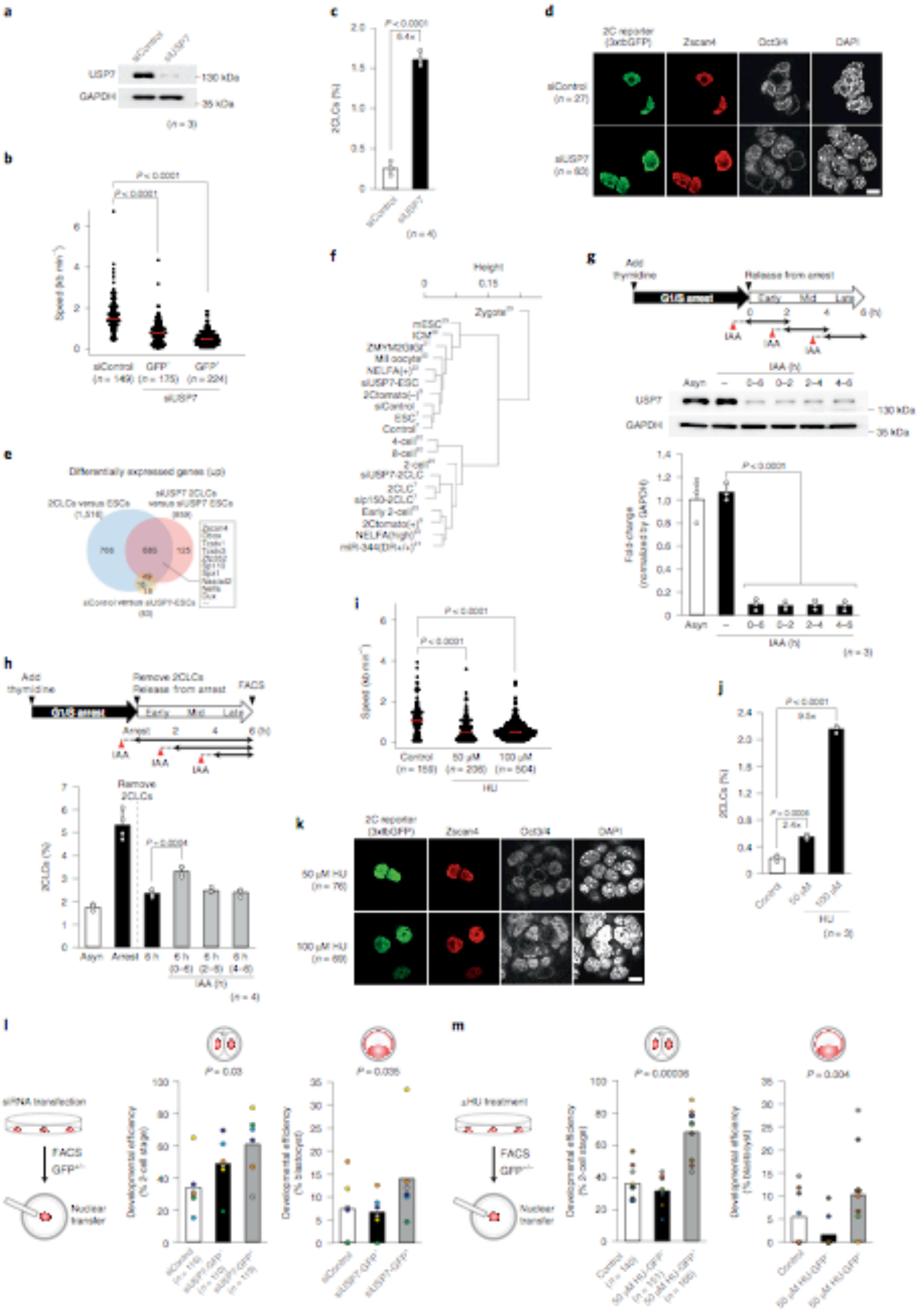


systems. Namely, we addressed whether induced pluripotent cell (iPSC) generation can be improved upon incubation with low doses of HU. Our results (Extended Data Fig. 5a,b) indicate an increase in the number of iPSC colonies after exposure to HU and may suggest a more general role for fork speed in cell reprogramming.

To further characterize the 2CLCs induced by USP7 depletion or HU, we examined their developmental potential compared with ESCs using two approaches. First, we performed morula aggregation with ESCs and 2CLCs produced after USP7 downregulation or by HU treatment, and analyzed their lineage contribution in blastocysts reconstructed in three dimensions, based on confocal microscopy. In each experiment we aggregated an equivalent number of cells and scored the number of cells in the inner cell mass (ICM) or the trophectoderm (TE) to account for variability between embryos. Although we found ESCs contributing to both the ICM and the TE, with a strong bias toward the ICM, in agreement with previous reports under these conditions[28–30], 2CLCs more frequently contributed to both (Extended Data Fig. 5c,d), in line with the suggested bipotentiality of 2CLCs. Single-cell chimera injections confirmed that 2CLCs can contribute to cells that express OCT4 and CDX2 (Extended Data Fig. 5e). Second, we asked whether depletion of USP7 or HU treatment can improve developmental efficiency after nuclear transfer (NT), as a readout for expanded cell potency as previously described for 2CLCs[7,21,31]. We performed NT into enucleated mouse oocytes using 2CLCs induced after siRNA for USP7 or upon HU treatment as donor. Remarkably, the number of embryos that cleaved to the 2-cell stage and formed hatching blastocysts was greatly increased when USP7-depleted or HU-treated green fluorescent protein-positive (GFP$^+$) cells were used as donors, compared with controls (Fig. 3l,m, Extended Data Fig. 5f and Supplementary Table 3). These findings are in line with the known increased reprogrammability of control 2CLCs[7]. These experiments using USP7-depleted and HU-induced 2CLCs as donors suggest that they correspond to endogenous 2CLCs[6,7] in terms of cellular potency. Thus, we conclude that reducing replication fork speed generates cells with a higher propensity to be reprogrammed upon NT.

2CLCs display distinctive changes in RT. Next, we explored the possible consequences of the differences in fork speed between ESCs and 2CLCs. We hypothesized that a slower fork speed, known to entail an increase in active origins to maintain the duration of the S-phase[32], may result in changes in RT. Mammalian cells display an orderly program for replicating their genome in units of around 400–800 kb, which are coordinately replicated at determined times during the S-phase[33–35]. Early replication often correlates with the transcriptional potential of a gene[36], although a causal relationship between RT and gene expression has not been firmly established. We first investigated whether 2CLC reprogramming entails a change in RT. We generated genome-wide RT maps from sorted ESCs and 2CLCs in early, mid and late S-phase (Extended Data Fig. 6a–c). A survey over the genome browser revealed specific gene regions shifting to earlier RT in 2CLCs. These included '2C'-specific genes such as *Zscan4*, *Obox2/3* and *Dux* (Fig. 4a and Extended Data Fig. 6d). Inquiry into the genomic regions shifting RT between the two cell types[37] revealed changes across the genome in the replication timing of 2CLCs, compared with ESCs (Fig. 4b). These changes represented approximately 3% of the genome, and most occurred by shifting at early S-phase (Fig. 4c), in line with our observations above suggesting that the early S-phase is critical for 2CLC emergence. These changes corresponded primarily to enlarged early replication domains in 2CLCs, leading to larger replication domains in the early S-phase in 2CLCs, compared with ESCs (Fig. 4d). In addition, domains shifting to earlier RT were enriched for MERVL sequences, in particular the MERVL promoter (LTR, MT2_Mm) and internal sequences (MERVL), but not for other endogenous retroviruses, LINE-1 or SINE-B2 elements (Fig. 4e). This shift to earlier RT matches a higher expression of MERVL elements in 2CLCs compared with those that change RT or shifted to a later pattern of replication (Fig. 4f and Extended Data Fig. 6e). We next examined the genes that change RT in 2CLCs. We identified 440 genes that shifted in their RT profile, most of which changed to an earlier phase (76%; $n=333$ genes) (Supplementary Table 4). Among them, most changed from mid- and late RT in ESCs to earlier replication in 2CLCs (98%; $n=328$ genes) (Fig. 4g). These genes included genes from the '2C' program, such as *Zscan4* and *Dux* (Fig. 4a and Extended Data Fig. 6d). Approximately a quarter of the RT-changing genes shifted to a later pattern of replication ($n=107$ genes). Among the genes that changed RT, only 30% ($n=136$) were differentially expressed in 2CLCs compared with ESCs and most of these shifted to an earlier RT (Fig. 4h). This suggests that only a fraction of the changes in RT of 2CLCs is concordant with changes in gene expression. To address the chromatin status of the genes that shift RT, we analyzed ESC and 2-cell-stage embryo chromatin immunoprecipitation followed by sequencing (ChIP-seq) datasets. In general, RT genes displayed enrichment of H3K4me3 at promoters or had bivalent signatures (Extended Data Fig. 6f–h), in agreement with their expression state in ESCs and 2-cell-stage embryos[38]. Some were enriched with H3K9me3 (Extended Data Fig. 6f–h) and the ENCODE term 'heterochromatin' was significantly over-represented in the RT regions that shift to earlier RT in 2CLCs (Extended Data Fig. 6i). This is in line with our observation that MERVL shifts to earlier RT in 2CLCs. RT profiles in

Fig. 3 | Slowing replication fork speed induces 2CLCs. a, USP7 expression 48 h after siRNA transfection. **b**, Fork speed in ESCs, GFP$^+$ (Usp7KD-induced 2CLCs) and GFP$^-$ cells after USP7 depletion. Statistical analysis was by two-sided Wilcoxon's rank-sum test. **c**, FACS quantification of 2CLCs 48 h after USP7 siRNA transfection. Statistical analysis was by two-sided Student's *t*-test. **d**, ZSCAN4 and OCT4 immunofluorescence in 2CLCs induced upon USP7 knockdown. **e**, Venn diagram of upregulated genes in control, USP7-depleted ESCs and USP7-depleted 2CLCs. **f**, Dendrogram of transcriptomes from various 2CLCs, early embryos, siControl-transfected ESCs, siUSP7-transfected ESCs and siUSP7-transfected 2CLCs. **g**,**h**, Early S-phase is critical for 2CLC induction on USP7 depletion. **g**, Western blot in an AID–USP7 knock-in cell line at indicated hours of auxin (indole-3-acetic acid (IAA)) treatment. IAA was added 30 min before early, mid or late S-phase (red arrowhead). GAPDH, glyceraldehyde 3-phosphate dehydrogenase. **h**, ESCs synchronized with double thymidine block, existing 2CLCs removed by FACS and IAA added as indicated. Emerging 2CLCs were quantified 6 h after release. Statistical analyses for pairwise comparison with control group were with a two-sided Student's *t*-test. **i**, Fork speed in HU-treated ESCs. Statistical analyses were by Wilcoxon's rank-sum test. **j**, 2CLCs induced by HU. The apparent higher increase in 2CLC percentage in 100 μM HU compared with 50 μM HU may be due to selective increase in ESC death and an increase in the number of cells in the S-phase with 100 μM HU (Extended Data Fig. 7j). Statistical analyses for pairwise comparison with control group used a two-sided Student's *t*-test. **k**, ZSCAN4 and OCT4 immunofluorescence in 2CLCs induced by HU. **l**,**m**, Greater reprogrammability of 2CLCs, induced by slowing fork speed. Nuclei of sorted GFP$^+$ and GFP$^-$ cells after USP7 siRNA (**l**) or HU (**m**) treatment were transferred into enucleated oocytes. Reprogramming efficiency is indicated by development of NT-derived embryos to 2-cell (left) and blastocyst (right). Barplots show average percentage of developmental efficiency across 6 (**l**) and 10 (**m**) independent experiments; each dot indicates percentages obtained in each experiment and color depicts side-by-side experiments; *n*, number of embryos analyzed. Statistical analyses were by two-sided Welch's test for unequal variances. **b**,**i**, Red line: median; barplots: mean ± s.d.; dots, values of each replicate; *n*, number of independent biological replicates. In **d** and **k**, scale bars, 10 μm.


a
siControl
siUSP7
USP7
GAPDH
130 kDa
35 kDa
(n = 3)
b
Speed (kb min⁻¹)
P < 0.0001
P < 0.0001
siControl (n = 149)
GFP⁻ (n = 175)
GFP⁺ (n = 224)
siUSP7
c
2CLCs (%)
P < 0.0001
6.4×
siControl
siUSP7
(n = 4)
d
2C reporter (3xtbGFP)
Zscan4
Oct3/4
DAPI
siControl (n = 27)
siUSP7 (n = 93)
e
Differentially expressed genes (up)
2CLCs versus ESCs (1,518)
siUSP7 2CLCs versus siUSP7 ESCs (858)
siControl versus siUSP7-ESCs (83)
766
685
125
49
18
Zscan4
Dux
f
Height
0
0.15
Zygote
mESC
ICM
ZMYM2OKO
MII oocyte
NELFA(+)
siUSP7-ESC
2Ctomato(−)
siControl
ESC
Control
4-cell
8-cell
2-cell
siUSP7-2CLC
2CLC
sip150-2CLC
Early 2-cell
2Ctomato(+)
NELFA(high)
miR-344(DR+/+)
g
Add thymidine
G1/S arrest
Release from arrest
Early
Mid
Late
0
2
4
6 (h)
IAA
IAA (h)
Asyn
−
0–6
0–2
2–4
4–6
USP7
GAPDH
130 kDa
35 kDa
Fold-change (normalized by GAPDH)
P < 0.0001
(n = 3)
h
Add thymidine
Remove 2CLCs
Release from arrest
FACS
G1/S arrest
Early
Mid
Late
Arrest
2
4
6 (h)
IAA
Remove 2CLCs
2CLCs (%)
P = 0.0004
Asyn
Arrest
6 h
6 h (0–6)
6 h (2–6)
6 h (4–6)
IAA (h)
(n = 4)
i
Speed (kb min⁻¹)
P < 0.0001
P < 0.0001
Control (n = 158)
50 μM (n = 206)
100 μM (n = 504)
HU
j
2CLCs (%)
P < 0.0001
9.5×
P = 0.0006
2.4×
Control
50 μM
100 μM
HU
(n = 3)
k
2C reporter (3xtbGFP)
Zscan4
Oct3/4
DAPI
50 μM HU (n = 76)
100 μM HU (n = 69)
l
siRNA transfection
FACS
GFP⁺/⁻
Nuclear transfer
P = 0.03
Developmental efficiency (% 2-cell stage)
siControl (n = 116)
siUSP7-GFP⁻ (n = 110)
siUSP7-GFP⁺ (n = 115)
P = 0.036
Developmental efficiency (% blastocyst)
siControl
siUSP7-GFP⁻
siUSP7-GFP⁺
m
±HU treatment
FACS
GFP⁺/⁻
Nuclear transfer
P = 0.00006
Developmental efficiency (% 2-cell stage)
Control (n = 140)
50 μM HU-GFP⁻ (n = 151)
50 μM HU-GFP⁺ (n = 166)
P = 0.004
Developmental efficiency (% blastocyst)
Control
50 μM HU-GFP⁻
50 μM HU-GFP⁺

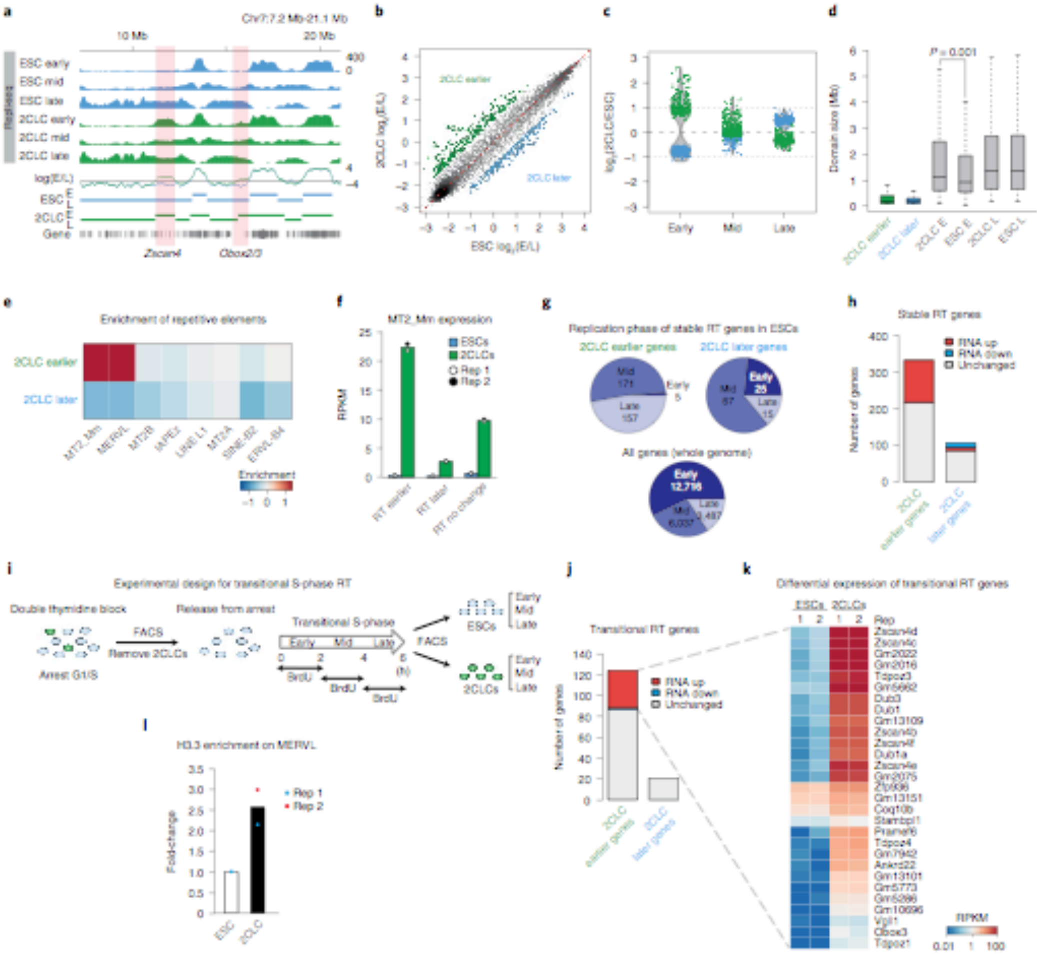


Fig. 4 | 2CLCs display changes in RT and slowing replication fork speed promotes reprogramming to totipotency during SCNT. a, Repli-seq tracks indicating early:late ratio as $\log_2$(E/L) for ESCs and 2CLCs. Horizontal lines indicate early (E) and late (L) replicating domains. Orange highlights regions of differential RT. **b**, Comparison of $\log_2$(E/L) between 2CLCs and ESCs at 100-kb bins across the genome. Green and blue points are regions of differential RT (twofold cutoff). **c**, Violin plot of log(fold differences) between 2CLSs and ESCs at early, mid and late S-phase over twofold differential $\log_2$(E/L) 100-kb bins. **d**, Early replicating domains are larger in 2CLCs. Boxplot of domain sizes for differential RT shows early or late domains from two biological replicates. Boxes show IQR between first and third quartiles, horizontal line shows the median and whiskers show Q3 + 1.5 × IQR and Q1 − 1.5 × IQR. Statistical significance comparing domain sizes was by two-sided Student's *t*-test. **e**, Genomic regions replicating earlier in 2CLCs are enriched in MT2_Mm and MERVL. Heatmap showing $\log_2$(fold enrichment) of repeats within earlier and later RT regions. **f**, Barplot of average expression (RPKM (reads per kilobase per million mapped reads)) in ESCs and 2CLCs of MT2_Mm repeats, the replication of which showed earlier, later or no shift in 2CLCs. Points indicate values for biological RNA-seq replicates. **g**, Shift of most genes to earlier replication in 2CLCs replicating in mid or late S-phase in ESCs. Proportions are shown by the numbers of genes according to their replication profile in ESCs, for the gene sets that shifted to earlier or later timing in 2CLCs. **h**, Genes replicating earlier in 2CLCs tend to be upregulated in 2CLCs. The barplot shows number of genes shifting to earlier or later RT in 2CLCs according to their expression changes (115 upregulated, 0 downregulated and 218 unchanged). **i**, Strategy to map replication timing in emerging 2CLCs during their transitional S-phase. **j**, Barplot depicting number of genes shifting to earlier or later RT during the transitional S-phase, during which 2CLCs emerge, according to their expression changes (29 upregulated, 0 downregulated and 95 unchanged) in 2CLCs. **k**, Most upregulated genes showing replication shift to earlier in transitional S-phase are repressed in ESCs. Heatmap depicts differential gene expression of genes shifting to earlier RT during the transitional S-phase. RPKM derive from two biological replicates. **l**, Relative H3.3 enrichment in ESCs and 2CLCs expressing SNAP-tagged-H3.3 analyzed by qPCR CUT&RUN. Dots represent biological replicates.

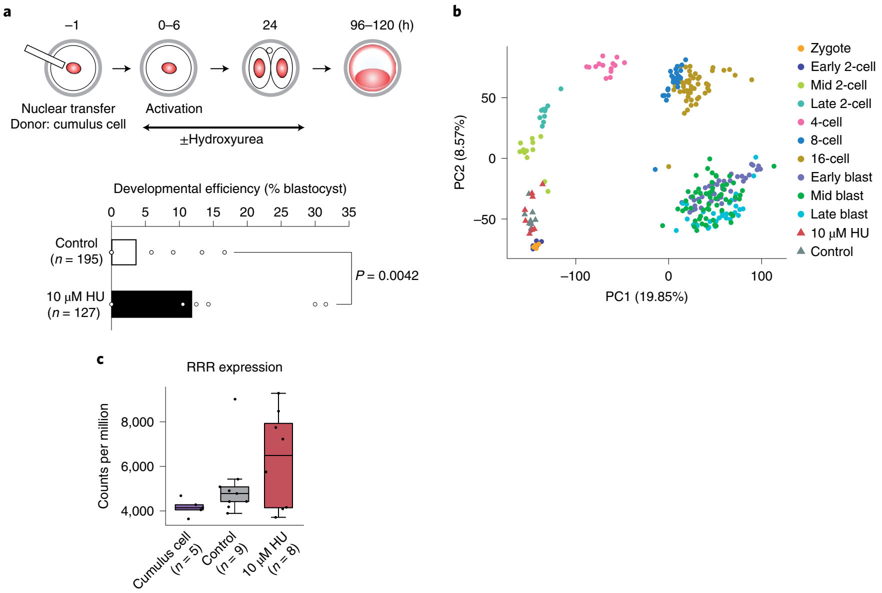


Fig. 5 | Improvement of the developmental potential of SCNT-derived embryos. a, SCNT embryos from cumulus cells treated with HU for 24 h after NT. Reprogramming efficiency was estimated by calculating the developmental rate of NT-derived embryos to the blastocyst stage. Barplots indicate the percentage of developmental efficiency of ten (control) and seven (10 μM HU) independent experiments. Each dot indicates the percentage obtained in each of these experiments and *n* indicates the total number of activated oocytes analyzed. Statistical analyses were performed using the *z*-score test for two population proportions (two tailed). **b**, Principal component (PC) analysis depicting the transcriptional profile of all NT embryos 28 h after activation, analyzed by single embryo RNA-seq, in comparison with wild-type embryos[38]. Note that NT-derived embryos cluster at the corresponding developmental time at which they were collected, indicating transcriptional reprogramming. **c**, Boxplot depicting the expression levels across RRRs in cumulus cells and control and HU-treated NT embryos. Each dot represents individual embryos (biological replicates). The boxplots indicate the first and third quartiles as the lower and upper hinges and the whiskers extend to the lowest and highest value no further than 1.5 × IQR.

2CLCs induced by *Usp7* knockdown (Extended Data Fig. 6j,k and Supplementary Table 5) displayed overall a similar RT profile compared with endogenous 2CLCs (Extended Data Fig. 6l), suggesting that the changes in replication fork speed during the S-phase, elicited by USP7 depletion, lead to a similar change in the RT profile in 2CLCs. Thus, we conclude that 2CLCs display a distinctive RT profile, characterized by changes to early replication of MERVLs and part of the 2C program. Importantly, as an excess number of origins are licensed in G1 than are used during the S-phase[39,40], these data are consistent with our observations indicating that entry into early S-phase is important, and suggest that additional origins may fire during early S-phase to promote 2CLC emergence.

MERVL shift to earlier replication during 2CLC reprogramming. Next, to address whether changes in RT temporally precede changes in cell fate, we devised an approach to map RT of emerging 2CLCs in the S-phase during which they transition toward 2CLCs, which we referred to as the 'transitional' S-phase (Fig. 4i). Our experimental design enabled us to analyze all cells that would undergo reprogramming in a synchronized fashion during the S-phase. Notably, the length of the S-phase of the transitioning cells, albeit variable, did not differ significantly in either the 'mothers' of the 2CLCs or the emerging 2CLCs themselves, compared with ESCs (Extended Data Fig. 7a), which enabled direct comparison of the RT profiles in both cell types. Our transitional RT datasets showed good correlation among replicates (Extended Data Fig. 7b) and revealed minor changes in RT compared with ESCs (Extended Data Fig. 7c), similar to the RT datasets of 'stable' 2CLCs. Analysis of the genes, which shift RT during the transitional S-phase, revealed 6 genes that shifted earlier with at least a 2-fold difference between ESCs and 2CLCs, and 145 genes with at least a 1.5-fold difference (Methods; Extended Data Fig. 7d and Supplementary Table 6). The fact that RT analysis in stable 2CLCs displays a higher number of genes that change in RT compared with the transitional RT dataset could indicate that part of the RT program of 2CLCs changes during the transitional S-phase, during which 2CLCs emerge, but another portion is achieved and consolidated once 2CLCs have been reprogrammed. Notably, most genes that shift to earlier RT during this transitional S-phase are not expressed in ESCs and become highly upregulated in 2CLCs (Fig. 4j,k)[7]. These genes belong to both the Zscan4-signature and the 2C signature[6,10,41,42]. Likewise, MERVL elements were enriched in domains shifting to earlier RT before 2CLC emergence (Extended Data Fig. 7e). As we detected differences in RT already during the transitional S-phase before 2CLC emergence, these data suggest that changes in RT of a subset of 2C genes and MERVL elements occur before changes in cell fate and transcriptional profile.

To address how a change in RT could potentially affect MERVL expression, we investigated their chromatin status because alteration of RT can disrupt chromatin modifications[43]. To restore the chromatin template after replication and preserve the corresponding epigenetic information, the replication machinery interacts with and recruits chromatin modifiers and remodelers[44]. Distinct chromatin proteins associate with the replication machinery in early versus late S-phase[45,46]. For example, 'new' histone H3.3 is known to be enriched at nascent chromatin specifically in the early S-phase[47]. H3.3 is associated with transcriptionally active chromatin and is incorporated throughout the cell cycle[48,49]. Thus, we investigated the distribution of H3.3 at MERVL. CUT&RUN for H3.3 indeed revealed that H3.3 is enriched at MERVL in 2CLCs, compared with ESCs (Fig. 4l). H3.3 is also enriched at MERVL in 2-cell-stage embryos (Extended Data Fig. 7f), coincident with the onset of MERVL expression[50]. Thus, a change in RT is associated with H3.3 enrichment at MERVL upon 2CLC emergence.

Slowing replication promotes reprogramming during SCNT. Finally, we sought to address the functional relevance of the replication dynamics and fork remodeling for reprogramming to totipotency. Terminally differentiated somatic cells can be reprogrammed to totipotency upon transplantation into enucleated oocytes[51,52]. However, this process is inefficient and often development beyond the 2-cell stage is considered to be a bottleneck[31]. Considering the slower fork speed that we observed in 2-cell-stage embryos, we addressed whether reducing fork speed improves somatic cell nuclear transfer (SCNT) efficiency using cumulus cells as donors. In normal fertilized embryos, HU treatment did not affect developmental progression (Extended Data Fig. 7g). Remarkably, HU treatment greatly increased SCNT efficiency, leading to significantly higher developmental rates compared with the controls (3.5-fold, $P = 0042$; Fig. 5a). RNA-seq analysis of NT embryos indicated that cloned embryos have effectively reset their transcriptional landscape, including activation of zygotic genome activation genes and importantly, also, of 'reprogramming resistant regions' (RRRs)[31] (Fig. 5b,c and Extended Data Fig. 7h,i). Thus, these results suggest that manipulating replication fork speed can improve cloning and facilitate reprogramming to totipotency.

Discussion

The overall rate of DNA synthesis is controlled by altering the rate at which individual replication forks synthesize DNA and/or changing the total number of active forks in the S-phase. In other vertebrates, such as *Xenopus*, embryonic cells divide extremely fast when the embryo goes from 50 to >5,000 cells, with S-phase lasting ~14 min at the earliest measured stage[53]. Although fork speed has not been determined before the midblastula transition, work with egg extracts supports a model whereby a high density of randomly positioned origins ensures genome duplication within this very short time[54,55]. DNA combing at the ribosomal DNA locus also revealed that frequency of initiation decreases from the early blastula onward[56–58]. However, similar analyses have not been done in mammals. Our data in the mouse indicate that the mammalian embryo replicates its DNA with low speed in the first three cell cycles after fertilization.

Our data suggest a working model whereby slower fork speed and the concomitant higher ratio of origins to forks enable a shift of RT of specific genomic regions, which are enriched in MERVL, toward early S-phase. Early replication may enable the recruitment of factors preferentially associated with replicative chromatin in early S-phase compared with late S-phase[46,47]. We propose that a change in RT provides a window of opportunity to alter the chromatin template toward transcriptionally permissive chromatin, for example, through the incorporation of the histone variant H3.3 (ref. [47]). Indeed, H3.3 can be deposited during the S-phase[59,60] and therefore changes in the distribution of H3.3 can potentially occur as a consequence of earlier replication. This is consistent with our data showing that MERVLs, which shift toward earlier RT, become highly expressed in 2CLCs and with data indicating that H3.3 enrichment at MERVL in the 2-cell-stage embryos is dependent on DNA replication[50]. This, in turn, may facilitate the expression of 2C genes driven by MERVL[6,9,21,61]. Indeed, H3.3 is required for de novo global transcription and embryonic development[62]. Molecular studies to determine the position and the number of origins used are currently impossible in embryos or 2CLCs, primarily because techniques to identify origins require amounts in the millions of cells. Identifying the mechanisms for origin firing during reprogramming and early development will demand further study and the development of low-input protocols. Our work contributes to the molecular characterization of 2CLCs, for which similarities to and differences from the 2-cell-stage embryo have started to emerge[61,63–65].

DNA damage induces *Zscan4* expression[66] and has recently been shown to promote expression of *Dux* through direct transactivation by p53 (ref. [67]). It is interesting that, upon DNA damage, the DNA-damage response (DDR) kinases ATR (ataxia telangiectasia and Rad3 related) and ATM (ataxia telangiectasia mutated) are required for DNA-damage-induced 2CLCs[67]. Earlier work documented that aphidicolin treatment, leading to increased phosphorylation of CHK1 in ESCs, induces *Zscan4* and *MERVL* expression[68]. However, although chemical inhibition of ATR partly reduced the extent of ZSCAN4 activation, this was not the case in ATR-deficient ESCs[68]. Although checkpoint activation and DNA damage can induce 2CLCs[67], 2CLC emergence can also occur without checkpoint activation[69]. It is noteworthy that most studies on the role of checkpoint activation in 2CLC induction are based on experimental induction of DNA damage, but only few have been performed in unperturbed conditions. Our work in naturally cycling 2CLCs, demonstrating the lack of detectable increase in γH2A.X in 2CLCs and that depletion of several checkpoint mediators does not impact the number of 2CLCs[10], suggests that DDR is not necessarily always involved in this process. This is in line with recent findings by Grow et al., which support both p53-dependent and p53-independent mechanisms for regulating DUX[67].

Overall, we suggest that regulation of fork speed can act as a fate determinant factor. Thus, our work highlights fundamental features of DNA replication in reprogramming cell fate.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41588-022-01023-0.

Received: 26 October 2021; Accepted: 27 January 2022;
Published online: 7 March 2022

References

1. Casser, E. et al. Totipotency segregates between the sister blastomeres of two-cell stage mouse embryos. *Sci. Rep.* **7**, 8299 (2017).
2. Tarkowski, A. K. Experiments on the development of isolated blastomeres of mouse eggs. *Nature* **184**, 1286–1287 (1959).
3. Ishiuchi, T. & Torres-Padilla, M. E. Towards an understanding of the regulatory mechanisms of totipotency. *Curr. Opin. Genet. Dev.* **23**, 512–518 (2013).
4. Baker, C. L. & Pera, M. F. Capturing totipotent stem cells. *Cell Stem Cell* **22**, 25–34 (2018).
5. Merchut-Maya, J. M., Bartek, J. & Maya-Mendoza, A. Regulation of replication fork speed: mechanisms and impact on genomic stability. *DNA Repair* **81**, 102654 (2019).
6. Macfarlan, T. S. et al. Embryonic stem cell potency fluctuates with endogenous retrovirus activity. *Nature* **487**, 57–63 (2012).
7. Ishiuchi, T. et al. Early embryonic-like cells are induced by downregulating replication-dependent chromatin assembly. *Nat. Struct. Mol. Biol.* **22**, 662–671 (2015).
8. 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).
9. Peaston, A. E. et al. Retrotransposons regulate host genes in mouse oocytes and preimplantation embryos. *Dev. Cell* **7**, 597–606 (2004).
10. Rodriguez-Terrones, D. et al. A molecular roadmap for the emergence of early-embryonic-like cells in culture. *Nat. Genet.* **50**, 106–119 (2018).

11. Michalet, X. et al. Dynamic molecular combing: stretching the whole human genome for high-resolution studies. *Science* **277**, 1518–1523 (1997).
12. Techer, H. et al. Replication dynamics: biases and robustness of DNA fiber analysis. *J. Mol. Biol.* **425**, 4845–4855 (2013).
13. Ahuja, A. K. et al. A short G1 phase imposes constitutive replication stress and fork remodelling in mouse embryonic stem cells. *Nat. Commun.* **7**, 10660 (2016).
14. Nieminuszczy, J., Schwab, R. A. & Niedzwiedz, W. The DNA fibre technique—tracking helicases at work. *Methods* **108**, 92–98 (2016).
15. Anglana, M., Apiou, F., Bensimon, A. & Debatisse, M. Dynamics of DNA replication in mammalian somatic cells: nucleotide pool modulates origin choice and interorigin spacing. *Cell* **114**, 385–394 (2003).
16. Montagnoli, A. et al. A Cdc7 kinase inhibitor restricts initiation of DNA replication and has antitumor activity. *Nat. Chem. Biol.* **4**, 357–365 (2008).
17. Katsuno, Y. et al. Cyclin A-Cdk1 regulates the origin firing program in mammalian cells. *Proc. Natl Acad. Sci. USA* **106**, 3184–3189 (2009).
18. Sakaue-Sawano, A. et al. Visualizing spatiotemporal dynamics of multicellular cell-cycle progression. *Cell* **132**, 487–498 (2008).
19. Lecona, E. et al. USP7 is a SUMO deubiquitinase essential for DNA replication. *Nat. Struct. Mol. Biol.* **23**, 270–277 (2016).
20. Wu, J. et al. The landscape of accessible chromatin in mammalian preimplantation embryos. *Nature* **534**, 652–657 (2016).
21. Yang, F. et al. DUX-miR-344-ZMYM2-mediated activation of MERVL LTRs induces a totipotent 2C-like state. *Cell Stem Cell* **26**, 234–250.e7 (2020).
22. Hu, Z. et al. Maternal factor NELFA drives a 2C-like state in mouse embryonic stem cells. *Nat. Cell Biol.* **22**, 175–186 (2020).
23. De Iaco, A. et al. DUX-family transcription factors regulate zygotic genome activation in placental mammals. *Nat. Genet.* **49**, 941–945 (2017).
24. Hernandez-Perez, S. et al. DUB3 and USP7 de-ubiquitinating enzymes control replication inhibitor Geminin: molecular characterization and associations with breast cancer. *Oncogene* **36**, 4817 (2017).
25. Alonso-de Vega, I., Martin, Y. & Smits, V. A. USP7 controls Chk1 protein stability by direct deubiquitination. *Cell Cycle* **13**, 3921–3926 (2014).
26. Poli, J. et al. dNTP pools determine fork progression and origin usage under replication stress. *EMBO J.* **31**, 883–894 (2012).
27. Somyajit, K. et al. Redox-sensitive alteration of replisome architecture safeguards genome integrity. *Science* **358**, 797–802 (2017).
28. Martin Gonzalez, J. et al. Embryonic stem cell culture conditions support distinct states associated with different developmental stages and potency. *Stem Cell Rep.* **7**, 177–191 (2016).
29. Wood, S. A. et al. Simple and efficient production of embryonic stem cell-embryo chimeras by coculture. *Proc. Natl Acad. Sci. USA* **90**, 4582–4585 (1993).
30. Beddington, R. S. & Robertson, E. J. An assessment of the developmental potential of embryonic stem cells in the midgestation mouse embryo. *Development* **105**, 733–737 (1989).
31. Matoba, S. et al. Embryonic development following somatic cell nuclear transfer impeded by persisting histone methylation. *Cell* **159**, 884–895 (2014).
32. Zhong, Y. et al. The level of origin firing inversely affects the rate of replication fork progression. *J. Cell Biol.* **201**, 373–383 (2013).
33. Rivera-Mulia, J. C. & Gilbert, D. M. Replicating large genomes: divide and conquer. *Mol. Cell* **62**, 756–765 (2016).
34. Goren, A. & Cedar, H. Replicating by the clock. *Nat. Rev. Mol. Cell Biol.* **4**, 25–32 (2003).
35. MacAlpine, D. M., Rodriguez, H. K. & Bell, S. P. Coordination of replication and transcription along a *Drosophila* chromosome. *Genes Dev.* **18**, 3094–3105 (2004).
36. Farkash-Amar, S. et al. Global organization of replication time zones of the mouse genome. *Genome Res.* **18**, 1562–1570 (2008).
37. Marchal, C. et al. Genome-wide analysis of replication timing by next-generation sequencing with E/L Repli-seq. *Nat. Protoc.* **13**, 819–839 (2018).
38. Deng, Q., Ramskold, D., Reinius, B. & Sandberg, R. Single-cell RNA-seq reveals dynamic, random monoallelic gene expression in mammalian cells. *Science* **343**, 193–196 (2014).
39. Dimitrova, D. S. & Gilbert, D. M. The spatial position and replication timing of chromosomal domains are both established in early G1 phase. *Mol. Cell* **4**, 983–993 (1999).
40. Fragkos, M., Ganier, O., Coulombe, P. & Mechali, M. DNA replication origin activation in space and time. *Nat. Rev. Mol. Cell Biol.* **16**, 360–374 (2015).
41. Eckersley-Maslin, M. A. et al. MERVL/Zscan4 network activation results in transient genome-wide DNA demethylation of mESCs. *Cell Rep.* **17**, 179–192 (2016).
42. Cerulo, L. et al. Identification of a novel gene signature of ES cells self-renewal fluctuation through system-wide analysis. *PLoS ONE* **9**, e83235 (2014).
43. Klein, K. N. et al. Replication timing maintains the global epigenetic state in human cells. *Science* **372**, 371–378 (2021).
44. Probst, A. V., Dunleavy, E. & Almouzni, G. Epigenetic inheritance during the cell cycle. *Nat. Rev. Mol. Cell Biol.* **10**, 192–206 (2009).
45. Miller, A. M. & Nasmyth, K. A. Role of DNA replication in the repression of silent mating type loci in yeast. *Nature* **312**, 247–251 (1984).
46. Stewart-Morgan, K. R., Petryk, N. & Groth, A. Chromatin replication and epigenetic cell memory. *Nat. Cell Biol.* **22**, 361–371 (2020).
47. Alabert, C. et al. Two distinct modes for propagation of histone PTMs across the cell cycle. *Genes Dev.* **29**, 585–590 (2015).
48. Ahmad, K. & Henikoff, S. The histone variant H3.3 marks active chromatin by replication-independent nucleosome assembly. *Mol. Cell* **9**, 1191–1200 (2002).
49. Clement, C. et al. High-resolution visualization of H3 variants during replication reveals their controlled recycling. *Nat. Commun.* **9**, 3181 (2018).
50. Ishiuchi, T. et al. Reprogramming of the histone H3.3 landscape in the early mouse embryo. *Nat. Struct. Mol. Biol.* **28**, 38–49 (2021).
51. Wakayama, T., Perry, A. C., Zuccotti, M., Johnson, K. R. & Yanagimachi, R. Full-term development of mice from enucleated oocytes injected with cumulus cell nuclei. *Nature* **394**, 369–374 (1998).
52. Wilmut, I., Schnieke, A. E., McWhir, J., Kind, A. J. & Campbell, K. H. Viable offspring derived from fetal and adult mammalian cells. *Nature* **385**, 810–813 (1997).
53. Graham, C. F. & Morgan, R. W. Changes in the cell cycle during early amphibian development. *Dev. Biol.* **14**, 439–460 (1966).
54. Kermi, C., Lo Furno, E. & Maiorano, D. Regulation of DNA replication in early embryonic cleavages. *Genes* **8**, 42 (2017).
55. Herrick, J., Stanislawski, P., Hyrien, O. & Bensimon, A. Replication fork density increases during DNA synthesis in *X. laevis* egg extracts. *J. Mol. Biol.* **300**, 1133–1142 (2000).
56. Hyrien, O., Maric, C. & Mechali, M. Transition in specification of embryonic metazoan DNA replication origins. *Science* **270**, 994–997 (1995).
57. Walter, J. & Newport, J. W. Regulation of replicon size in *Xenopus* egg extracts. *Science* **275**, 993–995 (1997).
58. Collart, C., Allen, G. E., Bradshaw, C. R., Smith, J. C. & Zegerman, P. Titration of four replication factors is essential for the *Xenopus laevis* midblastula transition. *Science* **341**, 893–896 (2013).
59. Santenard, A. et al. Heterochromatin formation in the mouse embryo requires critical residues of the histone variant H3.3. *Nat. Cell Biol.* **12**, 853–862 (2010).
60. Dunleavy, E. M., Almouzni, G. & Karpen, G. H. H3.3 is deposited at centromeres in S phase as a placeholder for newly assembled CENP-A in G(1) phase. *Nucleus* **2**, 146–157 (2011).
61. Kruse, K. et al. Transposable elements drive reorganisation of 3D chromatin during early embryogenesis. Preprint at *bioRxiv* https://doi.org/10.1101/523712 (2019).
62. Kong, Q. et al. Histone variant H3.3-mediated chromatin remodeling is essential for paternal genome activation in mouse preimplantation embryos. *J. Biol. Chem.* **293**, 3829–3838 (2018).
63. Genet, M. & Torres-Padilla, M. E. The molecular and cellular features of 2-cell-like cells: a reference guide. *Development* **147**, dev189688 (2020).
64. Zhang, Y. et al. Unique patterns of H3K4me3 and H3K27me3 in 2-cell-like embryonic stem cells. *Stem Cell Rep.* **16**, 458–469 (2021).
65. Yu, J. et al. Relaxed 3D genome conformation facilitates the pluripotent to totipotent-like state transition in embryonic stem cells. *Nucleic Acids Res.* **49**, 12167–12177 (2021).
66. Storm, M. P. et al. Zscan4 is regulated by PI3-kinase and DNA-damaging agents and directly interacts with the transcriptional repressors LSD1 and CtBP2 in mouse embryonic stem cells. *PLoS ONE* **9**, e89821 (2014).
67. Grow, E. J. et al. p53 convergently activates Dux/DUX4 in embryonic stem cells and in facioscapulohumeral muscular dystrophy cell models. *Nat. Genet.* **53**, 1207–1220 (2021).
68. Atashpaz, S. et al. ATR expands embryonic stem cell fate potential in response to replication stress. *eLife* **9**, e54756 (2020).
69. Zhu, Y. et al. Cell cycle heterogeneity directs spontaneous 2C state entry and exit in mouse embryonic stem cells. *Stem Cell Rep.* **16**, 2659–2673 (2021).

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

Methods

Embryo collection and culture. All mouse experiments were approved by the Ethics Committee of the Université de Strasbourg (Com'eth Institute of Genetics, Molecular and Cellular Biology) and performed under the compliance of either French legislation or the government of Upper Bavaria. F1 female mice (C57Bl/6J × CBA) aged <10 weeks were superovulated by intraperitoneal injection of 10 U of human chorionic gonadotropin (hCG) followed by 10 U of pregnant mare serum gonadotropin 48 h later, and then mated with F1 male (C57Bl/6J × CBA) mice. Zygotes were collected from the oviduct, placed in drops of KSOM (potassium-supplemented SOM) and cultured at 37 °C with 5% CO_2 as previously described[70].

ESC culture. Mouse E14 ESC lines were cultured in Dulbecco's modified Eagle's medium (DMEM) with GlutaMAX (Invitrogen) containing 15% fetal calf serum, 2× leukemia inhibitory factor, penicillin–streptomycin, 0.1 mM 2-mercaptoethanol, 3 µM CHIR99021 (GSK3β inhibitor) and 1 µM PD0325901 (MEK inhibitor) on gelatin-coated plates unless otherwise stated.

FACS. For isolation and quantification of 2CLCs, cells were washed twice with phosphate-buffered saline (PBS) and treated with 0.25% trypsin. After neutralization with ESC medium, cells were collected by centrifugation and the dissociated single cells were resuspended in ESC medium. To calculate the population of 2CLCs, we counted turbo GFP^+ ESCs after exclusion of dead and doublet cells based on the forward and side-scatter profiles. After sorting, cells were collected in normal culture medium and kept at 4 °C. For collection of cells in G1-phase in Fig. 2e and Extended Data Fig. 3e, we sorted the mCherry-hCdt1(1/100)Cy(−)-positive, iRFP-hGeminin(1/110)-negative subpopulation based on their fluorescence. For cell cycle analysis, the dissociated single cells were fixed with 70% ethanol for 30 min. After treatment with 250 µg ml^{-1} of RNase A (Thermo Fisher Scientific) for 5 min, cells were treated with 50 µg ml^{-1} of propidium iodide (PI) to stain DNA. For the cell death analysis in Extended Data Fig. 7j, harvested cells were incubated with Annexin-V, APC conjugate (A35110) for 15 min at room temperature in binding buffer (10 mM Hepes, pH 7.4, 140 mM NaCl, 2.5 mM $CaCl_2$), according to the manufacturer's protocol. Cells were subsequently washed with binding buffer and stained with 0.5 µg ml^{-1} of PI for 15 min on ice. Sorting was performed on a BD Biosciences FACSAria III and FACSMelody. Percentage of 2CLCs was calculated using FACSDiva and FACSChorus and the analysis of other FACS data was performed using FlowJo software.

DNA fibers in embryos and ESCs. DNA fibers were prepared as described[12,71], which we applied to low cell numbers. Embryos and ESCs transfected with siRNA or treated with HU were sequentially pulse labeled with 25 µM 5-iodo-2′-deoxyuridine (IdU; Sigma-Aldrich) and 50 µM 5-chloro-2′-deoxyuridine (CldU; Sigma-Aldrich) for 30 min each and collected. Labeled cells were lysed and DNA fibers were stretched on to the slide glass by tilting. The fibers were fixed in methanol:acetic acid (3:1), then denatured with 2.5 M HCl for 1 h, neutralized with PBS and blocked with 1% bovine serum albumin/0.1% Tween-20 in PBS. CldU and IdU tracks were detected with anti-bromodeoxyuridine (anti-BrdU) antibodies (described in Supplementary Table 7) recognizing CldU and IdU, respectively, and appropriate secondary antibodies. After the detection of IdU and CldU tracks, DNA was detected using an antibody against single-stranded DNA and the corresponding secondary antibody. 2-cell embryos in early S-phase, mid S-phase and late S-phase, and 4-cell embryos, 8-cell embryos and blastocysts were collected at 35, 37, 39, 53, 70 and 96 h post-hCG injection, respectively. Images were acquired on a Leica SP8 confocal microscope using a ×40 Plan/Apo NA1.3 oil immersion objective (Leica) at 2,048 × 2,048 pixels^2 at an effective pixel size of 142 nm. To calculate fork speed, we used the established conversion 1 µm = 2 kb (ref. [72]). Analysis of DNA fibers was performed by two different researchers using a customized image analysis pipeline that consisted of three steps: (1) localization of fibers in confocal images, (2) detection of branch modes in each fiber and (3) statistical analysis of different fiber parameters (for example, pattern proportion, branch length). As a prerequisite step, we employed masks to select regions of interest in the images, which contained a sufficient number of fibers to be analyzed. Briefly, for the fiber localization, we used a vessel detection algorithm, using a space-scale local variational approach, followed by a morphological reconstruction to extract the median line by B-spline fitting. To overcome issues of noise and signal heterogeneity, we implemented a structure reconstruction with a spatially variant morphological closing[73]. The process uses a small segment (at least the size of disconnection, for example, 20 pixels) as a structuring element. The map is then thresholded to a certain value (typically 0.5) and single fibers are identified separately by a connected component algorithm. Then, the skeletons of the fibers were identified by a morphological thinning and fitted to achieve subpixel accuracy. To detect patterns in the extracted fibers, we used a branch detection strategy. Briefly, intensity profiles from both channels were sampled along the median line. As the channels were not directly comparable in absolute intensity value, the logarithm of their point-wise intensity ratio was used instead. We used regression tree structures in combination with the CART algorithm[74], which uses a partitioning algorithm to detect the patterns of the DNA fibers. Subsequently, a semi-automated step to verify fiber detection and features was implemented manually. The fiber analysis software is written in Python and is available at https://github.com/IES-HelmholtzZentrumMunchen/dna-fibers-analysis. To calculate the IOD, we manually selected sufficiently long fiber stretches from our DNA fiber dataset in the DNA channel, which encompassed several IdU/CldU boundaries. To facilitate the analysis, we generated a Fiji (ImageJ) macro to open the regions of interest in the images and applied the ImageJ 'Straighten' function with a width of 19 pixels to convert bent fibers into approximately two-dimensional images, where the channel intensities were interpolated along the *x* axis. In the stretched fiber images, we then manually selected all identifiable IdU/CldU boundaries. The remaining analysis was performed in R. We first calculated from the *x* coordinates of the boundaries all origin positions by averaging between two adjacent boundary points. We then determined the pairwise difference between origins to obtain the IOD. IOD and boxplots were created using the ggplot2 library in R.

Cell cycle synchronization and drug treatment. For all G1/S synchronization with thymidine, a double thymidine block was used as follows: cells were incubated for 12 h with 2.5 mM thymidine, released for 9 h after washing out the thymidine, and then blocked again with 2.5 mM thymidine for 14 h to arrest all cells at the beginning of S-phase. For release experiments (Figs. 2a–c, 3g,h and 4i–k and Extended Data Fig. 2a–c,f), cell cycle arrest was subsequently released with two washes of thymidine-free medium. After release, cells were harvested at 1-h intervals or treated with 1 µM aphidicolin or 2.5 mM thymidine for 6 h. For other drug treatments (Fig. 2d and Extended Data Figs. 2g,h, 3a–d and 4i–k), the following inhibitors and concentrations were used: CDC7 inhibitor (PHA-767491; 10 µM), CDK1 inhibitor (RO-3306; 10 µM), PLK1 inhibitor (BI-6727; 500 nM) were used to synchronize cells for 8, 10 and 4 h, respectively. In Fig. 2e and Extended Data Fig. 3e, cells in G1-phase were sorted by FACS based on their FUCCI reporter system as described in FACS. After sorting, cells were plated under normal culture conditions or with medium supplemented with 10 µM CDC7 inhibitor. After culturing for 6 h, cells were analyzed by FACS to calculate the number of 2CLCs.

RNA-seq. Forty-eight hours after transfection of siRNA for control and USP7, cells were FACS sorted into ESCs and 2CLCs based on the GFP fluorescence, reflecting the 2C::tbGFP reporter activity. Total RNA was extracted using PicoPure RNA Isolation Kit (Thermo Fisher Scientific) and treated with turbo DNase (Life Technologies). Two biological replicates were prepared for each sample and their quality was checked using the 2100 Bioanalyzer with the RNA 6000 Nano Kit (Agilent). Libraries for strand-specific sequencing were created with a TruSeq Stranded Total RNA Library Prep Human/Mouse/Rat (Illumina) and IDT for Illumina-TruSeq RNA UD Indexes (Illumina) according to the manufacturer's protocol. Excess primers were removed through a purification step using AMPure XP beads (Agencourt Biosciences Corporation). The quality and quantity of the complementary DNA libraries were verified with the 2100 Bioanalyzer using the High Sensitivity DNA Kit (Agilent). Sequencing was carried out on an Illumina HiSeq 4000 (Illumina) with a 150-bp paired-end protocol according to Illumina's instructions.

NT with 2CLCs and ESCs. NT was performed as described[51] with slight modifications[75,76]. Metaphase II-arrested oocytes were collected from superovulated F1 female mice (C57Bl/6J × CBA) aged <10 weeks and cumulus cells were removed using hyaluronidase. Oocytes were enucleated in a droplet of M2 medium containing 5 µg ml^{-1} of cytochalasin B (CB) using a blunt Piezo-driven pipette. After enucleation, the spindle-free oocytes were washed extensively and maintained in CZB medium up to 2 h before nucleus injection. Nuclei of ESCs and 2CLCs (E14 background, originally derived from 129/Ola mouse strain) cultured in serum/leukemia inhibitory factor (nontreated, siControl-transfected, siUSP7-transfected or HU-treated cells) were collected by FACS based on their GFP fluorescence and size, and were aspirated in and out of the injection pipette to remove the cytoplasmic material and then injected into enucleated oocytes. The reconstructed oocytes were cultured in CZB medium for 1 h and activated for 6 h in Ca^{2+}-free CZB medium containing 10 mM Sr^{2+} and 5 µg ml^{-1} of CB. After activation, the reconstructed embryos were cultured in KSOM at 37 °C under 5% CO_2:air for 5 d and subsequently checked for their developmental efficiency. Note that, although most NT protocols employ Trichostatin A, we purposely refrained from using Trichostatin A to avoid confounding effects due to potential alterations to chromatin structure.

SCNT. SCNT was performed using cumulus cells as donors. For these experiments, we used two different F1 mouse strains to provide robustness and validation: C57BL/6J × DBA/2J and C57Bl/6J × CBA. The same protocol as for 2CLCs and ESCs was used, with slight modifications. Briefly, MII oocytes were collected and enucleated in CZB medium and then allowed to recover in KSOM until they were used for NT. The nuclei of donor cumulus cells were injected into the enucleated oocytes using a Piezo-driven micromanipulator. After reconstruction, oocytes were cultured for 1 h in KSOM and activated for 6 h in KSOM containing 10 mM Sr^{2+} and 5 µg ml^{-1} of CB supplemented with 2 mM (ethylenebis(oxonitrilo))tetra-acetate[77]. Embryos were then randomly distributed in medium with or without HU (10 µM), which was replaced by fresh medium without HU after 24 h. Experimental design and scoring were double blinded. The SCNT data derived from the two mouse strains were verified for consistency and the sum of the compiled data is shown in Fig. 5a.

Replication timing. For the stable RT and USP7 RT, synchronously cycling cells were pulse labeled with the nucleotide analog BrdU for 2 h, respectively. Cells were sorted into early, mid and late S-phase fractions, 20,000 cells each, on the basis of DNA content using FACS. For the transitional RT, existing 2CLCs were removed after double thymidine block. After release from G1/S arrest, ESCs were treated with BrdU for 2 h during the specific time windows indicated in Extended Data Figure 4i (0–2 h for early S-phase, 2–4 h for mid S-phase or 4–6 h for late S-phase). ESCs and newly emerged 2CLCs were sorted by FACS based on the 2C::tbGFP fluorescence 6 h after release from G1/S block, and genomic (g)DNA was isolated from each condition (that is, early, mid or late S-phase for ESCs and 2CLCs) using sodium dodecylsulfate–proteinase K buffer and purified by phenol–chloroform extraction. The gDNA was fragmented using the Covaris sonicator to obtain fragments of 700 bp on average. The sheared, BrdU-labeled DNA from each fraction was immunoprecipitated using 0.5 μg of mouse anti-BrdU antibody followed by addition of 50 μl of precleared Dynabeads coupled to sheep anti-mouse immunoglobulin G (Invitrogen). The immunoprecipitated pellet was digested overnight with proteinase K and purified by phenol–chloroform extraction. RT libraries were prepared based on Accel-NGS methyl seq library kit (Swift Biosciences) according to the manufacturer's instructions. The BrdU-immunoprecipitated DNA was denatured and subjected to Adaptase reaction. This step was followed by an extension reaction with two cleanup steps utilizing Agencourt Ampure XP beads (Beckman Coulter). The eluate was subjected to a ligation step, followed by Ampure bead-mediated purification. Indexing PCR was performed at 98 °C for 30 s, 9 cycles at 98 °C for 10 s, 60 °C for 30 s and 68 °C for 60 s, followed by a 4 °C hold cycle. The PCR product was further purified by Ampure beads and eluted in a 20-μl volume using Tris–EDTA buffer provided by the manufacturer. The libraries were verified using Agilent 2200 Tape Station (Agilent) utilizing DNA high-sensitivity tape (Agilent). Up to 12 libraries were pooled together after Qubit quantification with Qubit DNA HS assay kit (Thermo Fisher Scientific) and loaded into Nextseq 500/550 high-output cartridge (Illumina) for 75 cycles of single-end sequencing.

RT analysis. Repli-seq reads from early, mid and late time points of S-phase were mapped to the reference mm9 genome using BWA[78] and counted over 100-kb genomic bins across the genome, followed by the Loess smoothing of bin counts as previously described[37]. The E/L was calculated from the read counts in early and late S-phase. Regions with differential RT between ESCs (GFP$^-$) and 2CLCs (GFP$^+$) cells were determined based on 2-fold (or 1.5-fold for the transitional S-phase) cutoff of change in E/L ratio. Domains of early and late replication were identified using the DNAcopy package[79]. Genes were classified as early, mid or late replicating based on the stage of S-phase with the highest read density over the gene body. This three-stage classification was highly consistent with the traditional E/L based only on the reads from early and late stages.

Single embryo RNA-seq and library preparation. Control and HU-treated (10 μM) nuclear transferred embryos were cultured until 28 h after activation, at which point a representative proportion of embryos was collected, washed with PBS, placed in tubes with 1× Clontech lysis buffer (Z5013N) containing ERCC RNA Spike-In Mix (Invitrogen) and flash-frozen in liquid nitrogen. RNA-seq was carried out using the SMART-seq2 protocol[80] and subjected to paired-end sequencing on a Nextseq 500 (Illumina) platform. A total of nine control and eight HU-treated embryos derived from two independent experiments were sequenced. In parallel, we collected 12 single cumulus cells used as donors and processed them for RNA-seq under identical conditions.

Statistical analyses. To assess whether the data were normally distributed, we performed a Shapiro–Wilk test or *F*-test. For normally distributed data, we applied the Student's *t*-test to perform pairwise comparisons between groups, as indicated throughout the figure legends; otherwise we applied the nonparametric Mann–Whitney (Wilcoxon's rank-sum) test. The proportions of patterns from the DNA fiber data were analyzed by a binomial test in R (two-sample test for equality of proportions with continuity correction). Where data are shown as box-and-whisker plots, we followed the convention for boxplots[81] (thick bar, median; boxes, IQR; whiskers, range without outliers; dots outside whiskers, outliers beyond 3× or 2× IQR). For datasets with unequal variance (Fig. 3l,m), we applied Welch's test for unequal or unknown variances.

Antibodies. Antibodies used in this work are described in Supplementary Table 7.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The Repli-seq and RNA-seq data from the present study are available from the Gene Expression Omnibus database, accession nos. GSE136228 and GSE166338. Previously published RNA-seq datasets reanalyzed in the present study are available under accession nos. GSM1933935, GSM1625860, GSM1933937, GSM1625862, GSM1625864, GSM1625867, GSM1625868, GSM838739, GSM838738, GSM1625873, E-MTAB-2684 and GSM1933935. ChIP-seq datasets reanalyzed in the present study are available under accession nos. GSE73952, GSE97778, GSE73952, GSE23943 and GSE139527. Source data are provided with this paper. All other data supporting the findings of the present study are available from the corresponding author upon reasonable request.

Code availability

All next-generation sequencing data were analyzed using standard programs and packages, as detailed in Methods. Code for DNA fiber analysis is available at: https://github.com/IES-HelmholtzZentrumMunchen/dna-fibres-analysis.

References

70. Hogan, B., Beddington, R., Costantini, F. & Lacy, E. *Manipulating the Mouse Embryo: A Laboratory Manual* 4th edn (Cold Spring Harbor Laboratory Press, 1994).
71. Miotto, B. et al. The RBBP6/ZBTB38/MCM10 axis regulates DNA replication and common fragile site stability. *Cell Rep.* **7**, 575–587 (2014).
72. Conti, C. et al. Replication fork velocities at adjacent replication origins are coordinately modified during DNA replication in human cells. *Mol. Biol. Cell* **18**, 3059–3067 (2007).
73. Tankyevych, O., Talbot, H. & Dokladal, P. Curvilinear morpho-Hessian filter. In *Proc. Internal Symposium on Biomedical Imaging: From Nano to Macro (ISBI)* 1011–1014 (IEEE, 2008).
74. Breiman, L., Friedman, J., Stone, C. & Olshen, R. *Classification and Regression Trees* (CRC Press, 1984).
75. Kishigami, S. et al. Significant improvement of mouse cloning technique by treatment with trichostatin A after somatic nuclear transfer. *Biochem. Biophys. Res. Commun.* **340**, 183–189 (2006).
76. Li, J., Ishii, T., Feinstein, P. & Mombaerts, P. Odorant receptor gene choice is reset by nuclear transfer from mouse olfactory sensory neurons. *Nature* **428**, 393–399 (2004).
77. Kishigami, S. & Wakayama, T. Efficient strontium-induced activation of mouse oocytes in standard culture media by chelating calcium. *J. Reprod. Dev.* **53**, 1207–1215 (2007).
78. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics* **26**, 589–595 (2010).
79. Hiratani, I. et al. Global reorganization of replication domains during embryonic stem cell differentiation. *PLoS Biol.* **6**, e245 (2008).
80. Picelli, S. et al. Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nat. Methods* **10**, 1096–1098 (2013).
81. Tukey, J. W. *Exploratory Data Analysis* (Addison-Wesley, 1977).

Acknowledgements

We thank P. Pasero and P.A. Defossez for critical advice on DNA fiber protocol and analysis, B. Pfander and E. Lecona for helpful discussions and advice, D. Pich and W. Hammerschmidt for access to FACS, M. Serrano, J. Cossec and A. Dejean for the iMEFs and S. Hamperl and A. Burton for critical reading of the manuscript. Work in the Torres-Padilla laboratory is funded by the Helmholtz Association, the German Research Council (CRC 1064), an ET124/2-1 grant awarded to A.E. and H2020 Marie-Curie Actions ITN EpiSystem and ChromDesign. R.I.S. acknowledges NIH's National Institute of Diabetes and Digestive and Kidney Diseases (grant no. P30 DK040561). J.R.W. acknowledges NIH's National Institute of General Medical Sciences (grant nos. R01GM097360 and R35GM144131) and NIH/National Cancer Institute's Cancer Center Support (grant no. P30 CA006927) and K.Y. acknowledges JSPS KAKENHI (grant no. JP18H05528). J.R.W. is a recipient of an ALA Lung Cancer Discovery Award.

Author contributions

T.N. designed, performed and analyzed most of the experiments. J.L., K.Y. and M.T. performed the NT. A.E. and J.P. performed image analyses. J.F. and M.E. established a mathematical model under A.S.'s supervision. F.J. performed most bioinformatic analyses under R.S.'s supervision. L.A.-P. analyzed single embryo RNA-seq. E.R.R.-M. and P.Y.A.P. analyzed histone modification profiles. C.V.R. and D.C. performed library preparation for RT analysis under J.R.W.'s supervision. M.E.T.-P. conceived, designed and supervised the study. All authors contributed to manuscript preparation, and read, commented on and approved the manuscript.

Funding

Open access funding provided by Helmholtz Zentrum München - Deutsches Forschungszentrum für Gesundheit und Umwelt (GmbH).

Competing interests

J.R.W. is/was serving as consultant and advisor for Qsonica, Salarius Pharmaceuticals, Daiichi Sankyo, Inc. and Vyne Therapeutics, and currently receives research sponsoring from Salarius Pharmaceuticals. The other authors declare that they have no competing interests.

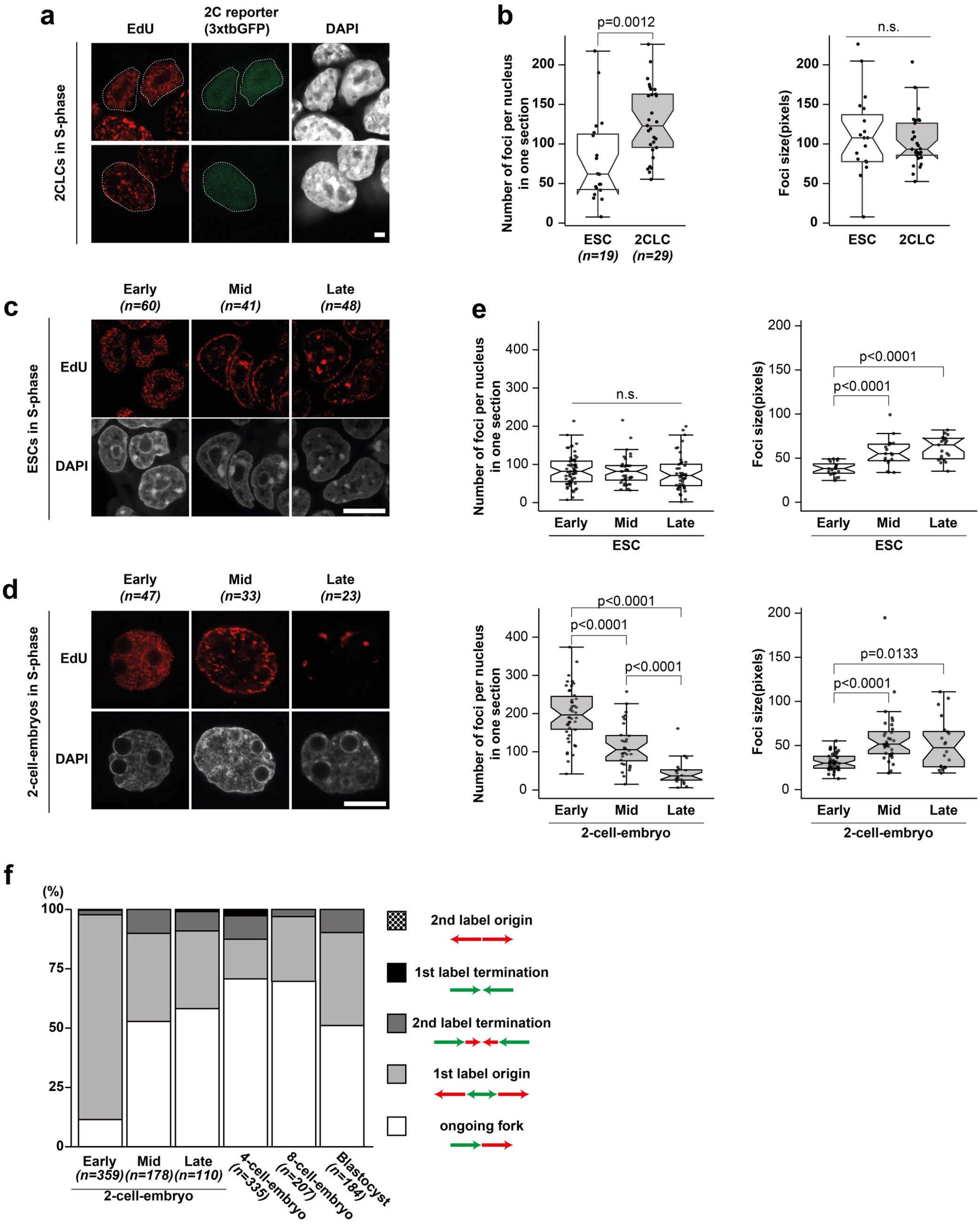
Additional information

Extended data are available for this paper at https://doi.org/10.1038/s41588-022-01023-0.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41588-022-01023-0.

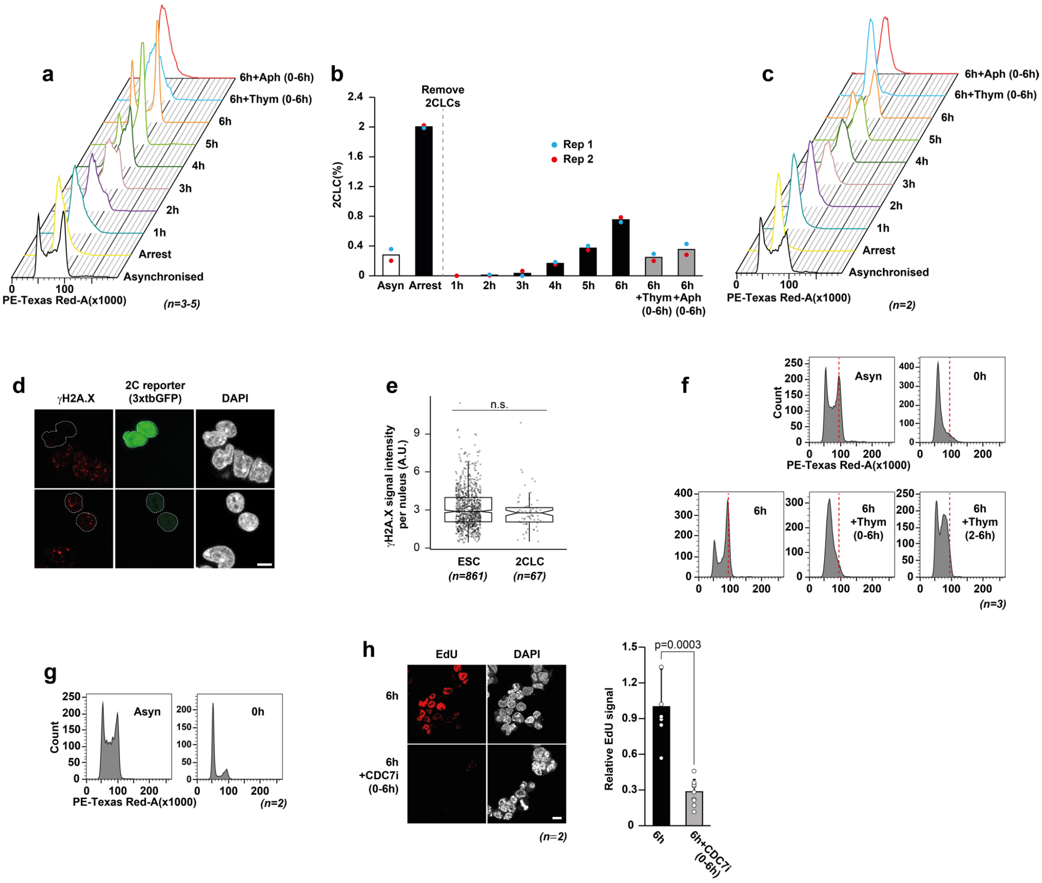
Correspondence and requests for materials should be addressed to Maria-Elena Torres-Padilla.

Reprints and permissions information is available at www.nature.com/reprints.

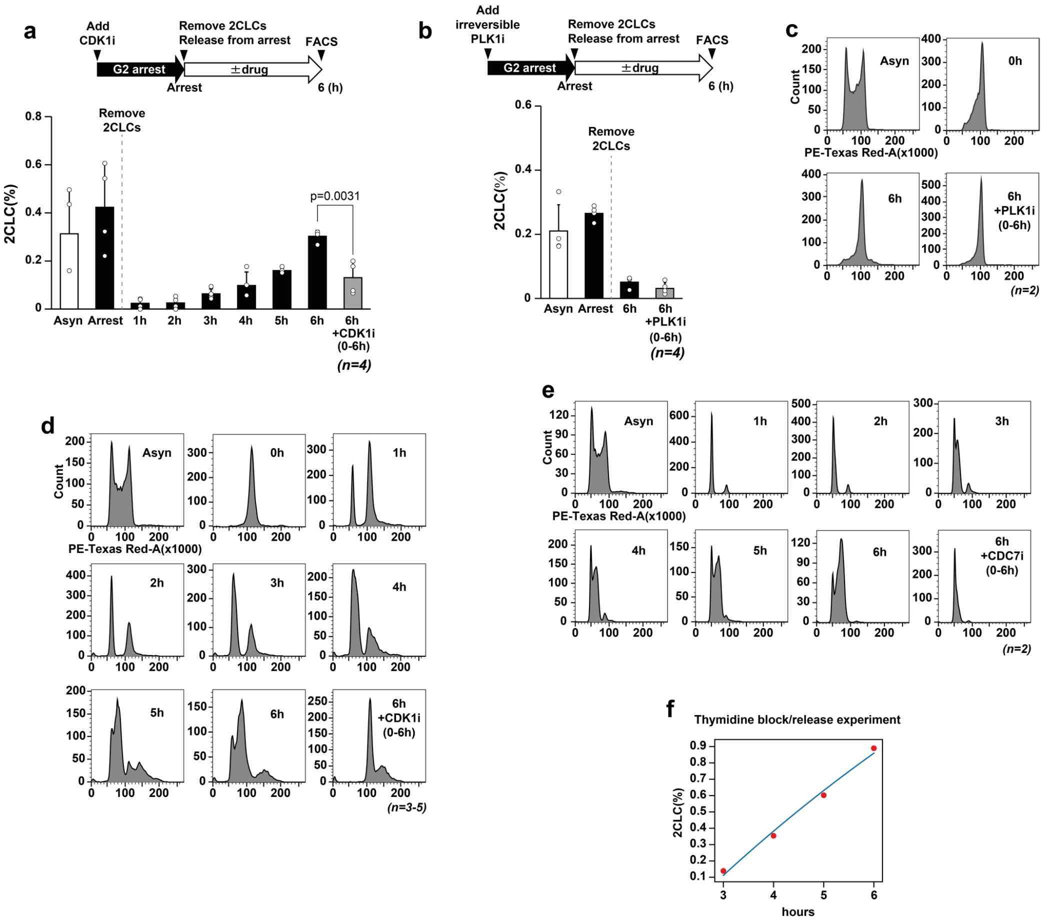


Extended Data Fig. 1 | See next page for caption.

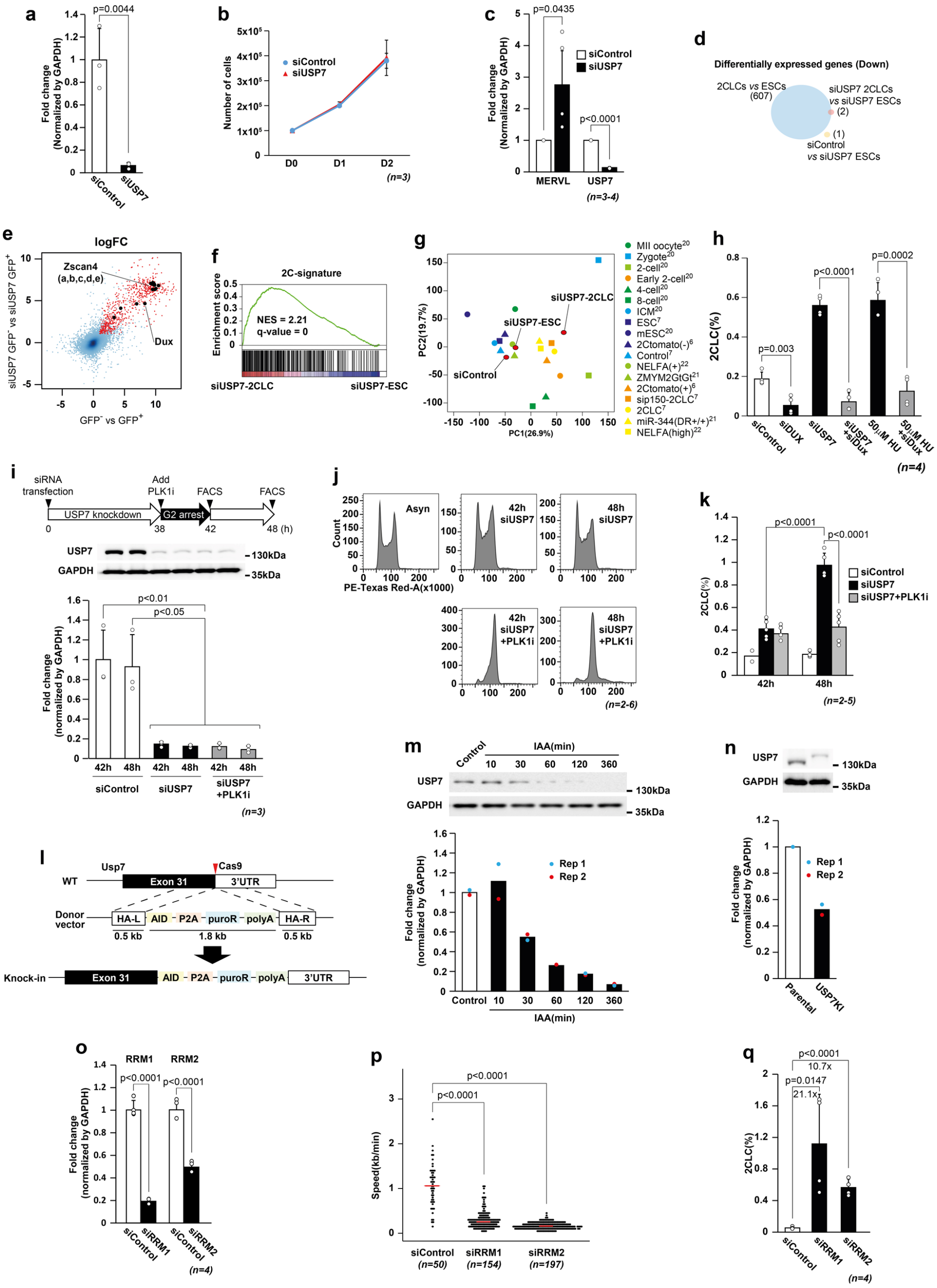
Extended Data Fig. 1 | Increasing the number of active origins in 2-cell-embryos. a. DNA replication in asynchronous ESCs and 2CLCs visualized with STED microscopy. Cells were pulse-labeled with EdU for 20 min. White dotted line indicates 2CLCs. Scale bar, 2.5 µm. Images were acquired side by side and analyzed using identical parameters and are therefore comparable. **b**. Number and size distribution of EdU foci of representative, randomly selected ESCs and 2CLCs. Each dot indicates the number of EdU foci (left) and their size (right) in one single STED section in one nucleus. Boxplots: median and interquartile range (IQR); whiskers: smallest and largest values within 1.5×IQR. 1 pixel equals 20.6 nm. Statistical analyses: two-sided Wilcoxon rank-sum test. **c**. DNA replication patterns in ESCs under STED microscopy. EdU was added 0, 2, 4 h after double thymidine block release for early, mid, and late S-phase, respectively. **d**. DNA replication patterns in 2-cell-embryos using STED microscopy. EdU incubation was from 34, 36, and 38 h post-hCG injection for early, mid, and late S-phase, respectively. Images in c and d were acquired side by side, analyzed using identical parameters and are comparable (but not with panels in a and b). **c, d**, Scale bar, 10 µm; n, number of nuclei analyzed. **e**. Number and size distribution of replication foci of early, mid, and late S-phase in ESCs and 2-cell-embryos. Each dot indicates the number of EdU foci (left) and their size (right) in each nucleus in one STED section. Note that because the nuclear volume of 2-cell embryos is approximately 20 times bigger than ESCs, the total number of DNA foci in 2-cell embryos is much higher than ESCs. Boxplots: median and interquartile range (IQR); whiskers: smallest and largest values within 1.5×IQR. 1 pixel equals 20.6 nm. Statistical analyses: two-sided Wilcoxon rank-sum test. **f**. Replication patterns from fiber analyses at the indicated embryonic stages. N: number of fibers analyzed.



Extended Data Fig. 2 | Effect of cell cycle progression on the emergence of 2 CLCs. a. Cell cycle profiles determined by FACS based on propidium iodide staining of ESCs after release from double thymidine block, which corresponds to Fig. 2b. **b**. Population of 2CLCs (in %) after release from double thymidine block detected with 2C::tdTomato reporter[6]. Following synchronization of ESCs at G1/S by double thymidine block, existing 2CLCs were removed and the remaining cells were released from the block and cultured with or without the indicated DNA replication inhibitor. Newly emerging 2CLCs were quantified by FACS at indicated time points after release. **c**. Cell cycle profiles determined by FACS based on propidium iodide staining of the 2C::tdTomato ESCs reporter line after release from double thymidine block, which corresponds to Extended Data Fig. 2b. **d, e**. γH2A.X immunostaining in asynchronous ESCs and 2CLCs. Representative images (d) and the corresponding quantification of global γH2A.X levels (e). 2CLCs are outlined with white dotted lines. Boxplots show median and interquartile range (IQR), whiskers depict the smallest and largest values within 1.5×IQR. Scale bar, 10 μm. Statistical analyzes: two-sided Wilcoxon rank-sum test. **f**. Cell cycle profiles determined by FACS based on propidium iodide staining of ESCs under thymidine treatment from 2 h after release from double thymidine block, which corresponds to Fig. 2c. The dashed, red line indicates the G2/M peak. **g, h**. S-phase progression after removal of the CDC7 inhibitor. The effects of CDC7 inhibitor on the cell cycle arrest (g) and restoration of the DNA replication after releasing from block (h, left and right panels) were verified by FACS and EdU incorporation, respectively. In **b** mean values are shown, dots indicate values for each replicate. In **h**, bar plots show mean±S.D and dots indicate values of each image. n indicates number of independent biological replicates. Statistical comparisons: two-sided Student's *t*-test.

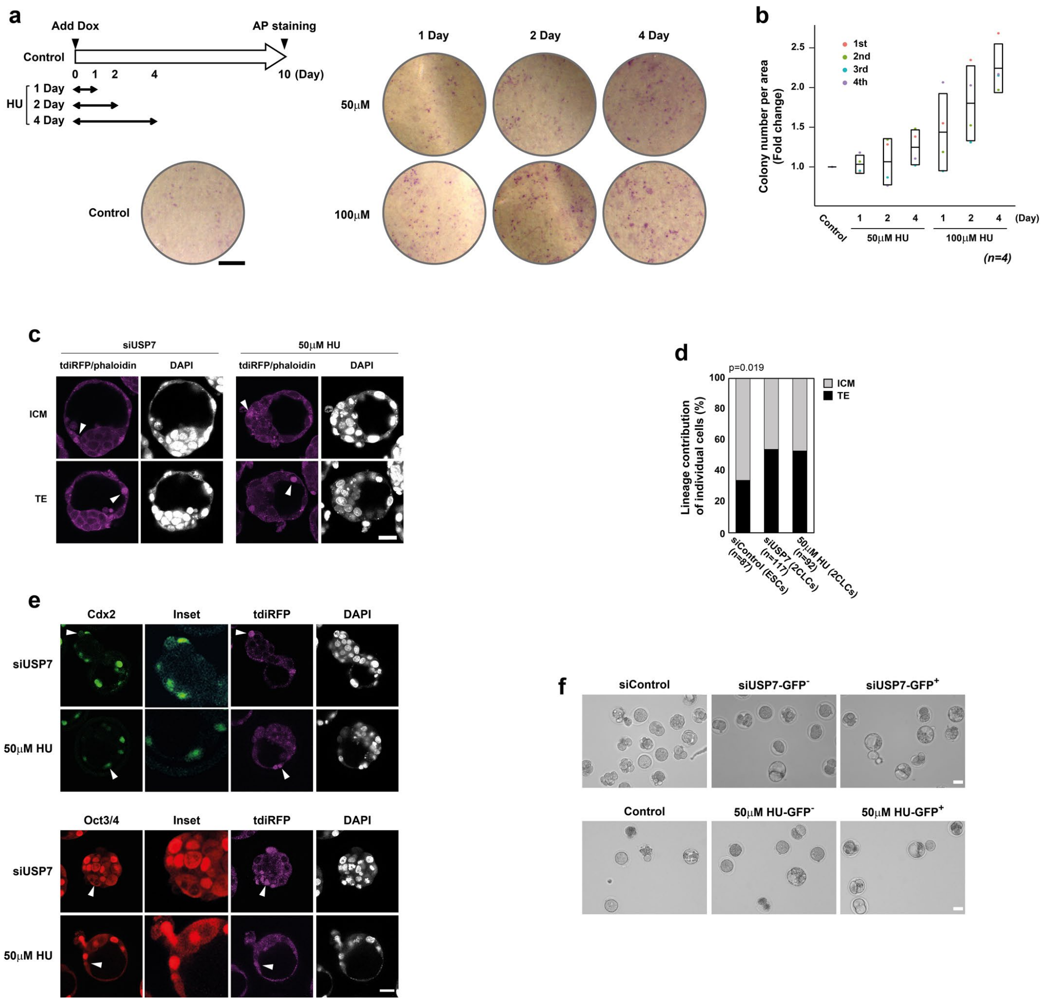


Extended Data Fig. 3 | Effect of entry into S-phase on 2CLC reprogramming. a. ESCs were synchronized using a CDK1 inhibitor, after which existing 2CLCs were removed. Cells were subsequently grown with or without CDK1 inhibitor and newly emerging 2CLCs were quantified hourly by FACS until 6 h after block release. **b**. ESCs were synchronized using a PLK1 inhibitor, after which existing 2CLCs were removed. Cells were subsequently grown with or without PLK1 inhibitor and newly emerging 2CLCs were quantified by FACS 6 h after block release. **c**. Cell cycle profiles determined by FACS for propidium iodide staining of ESCs after release from PLK1 inhibitor, which corresponds to Extended Data Fig. 2b. **d**. Cell cycle profiles based on propidium iodide content of ESCs after release from CDK1 inhibitor, which corresponds to Extended Data Fig. 2a. **e**. Cell cycle profiles based on propidium iodide content of ESCs after collection in G1-phase using FACS without the addition of drugs (for example without cell cycle synchronization) based on the FUCCI reporter. These data correspond to Fig. 2e. **f**. The panel shows the data from the double thymidine block and release experiment described in Fig. 2A and the corresponding fit. The estimation of f_s obtained from this fit was then used to identify the values of $\frac{f_{G1}}{f_S}$ and $\frac{f_{G2M}}{f_S}$ compatible with the data, shown in Fig. 2f. These values lie on a line, shown in Fig. 2f, computed from eq. (1) in the Methods. The values of the parameters used are: $\frac{N_{2c}}{N_E} = 0.01$, $T_c = 8\ h$, $\frac{1}{\omega - \varphi_{2c}} = 12\ h$. In **a** and **b**, the bar plots show mean±S.D. and dots indicate the values of each replicate. n indicates the number of independent biological replicates. Statistical comparisons were performed by two-sided Student's *t*-test.

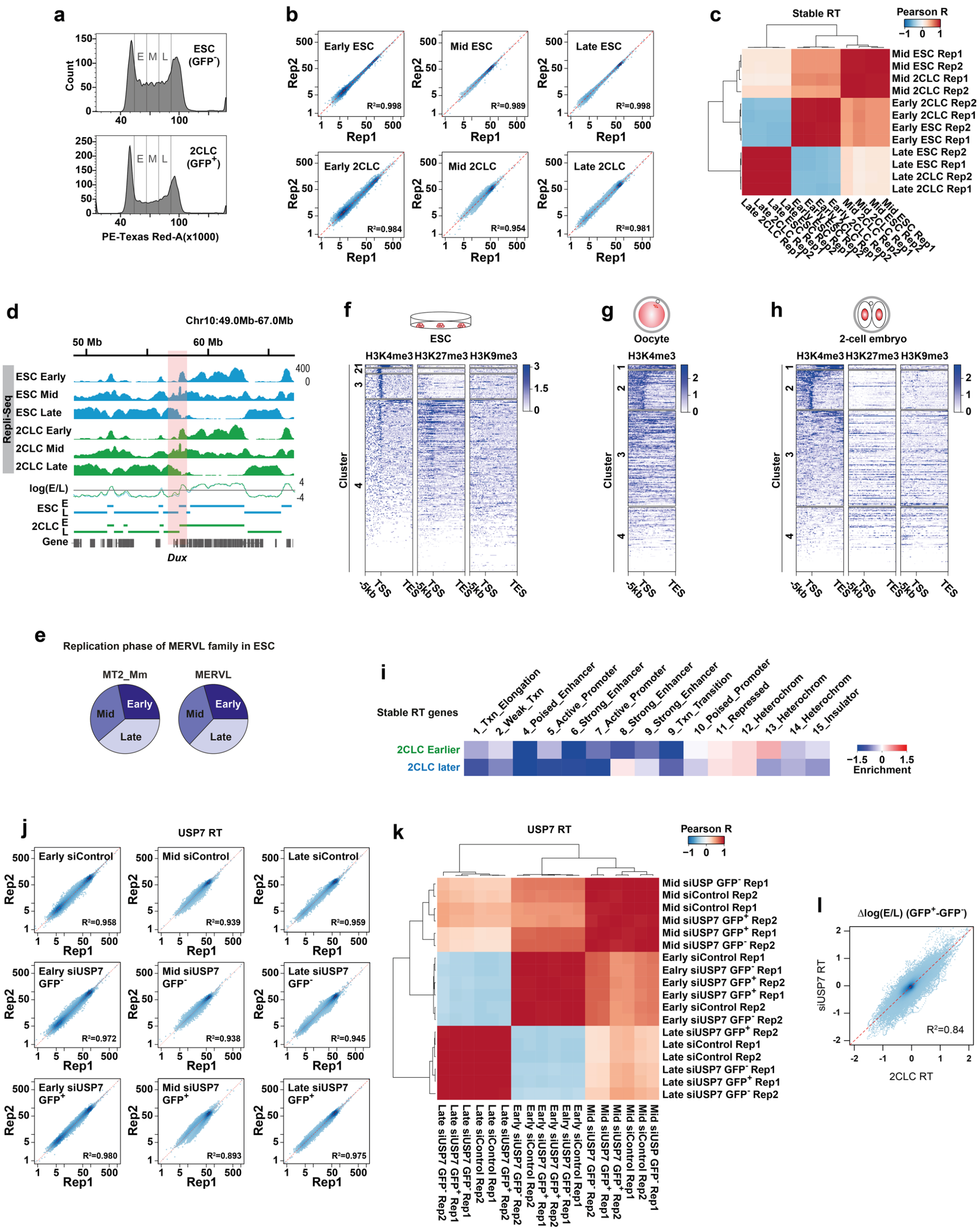


Extended Data Fig. 4 | See next page for caption.

Extended Data Fig. 4 | S-phase-dependent effect of USP7 on the emergence of 2CLCs. a. USP7 protein quantification upon *Usp7* siRNA (corresponds to Fig.3a). Statistical comparisons: two-sided Welch's *t*-test. **b**. Growth curve after *Usp7* RNAi at day 1(D1) 2(D2) after seeding(D0). **c**. MERVL and *Usp7* qRT-QPCR upon *Usp7* RNAi. Statistical comparisons: two-sided Welch's *t*-test. **d**. Venn diagram comparing downregulated genes in control, USP7-depleted ESCs, and USP7-depleted 2CLCs. **e**. logFC scatter plots between siUSP7-2CLCs and endogenous 2CLCs showing high overlap of upregulated genes (red) between both 2CLC samples. **f**. Gene-set enrichment analysis of siUSP7-induced 2CLCs against a '2C' signature. **g**. PCA of siControl, siUSP7 ESCs (GFP^{-} cells) and siUSP7-induced 2CLCs transcriptomes compared with embryos and other 2CLC datasets. **h**. 2CLC percentage upon siRNA for *Usp7* or HU treatment combined with siRNA for *Dux*. Statistical comparisons: two-sided Student's *t*-test for pairwise comparison only between the indicated groups. **i**. Western Blot after *Usp7* siRNA or control siRNA transfection and treatment of PLK1 inhibitor at indicated times. Statistical comparisons: two-sided Student's *t*-test between each sample. Highest p-values are shown. **j-k**. Cell cycle profiles of ESCs (j) and 2CLC percentage (k) after transfection of control or *Usp7* siRNA, followed by treatment with PLK1 inhibitor at indicated times. In (k) n for 42 h control samples is 2 biological replicates; for 48 h control is 4 and for all other samples is 5. Statistical comparisons between the indicated groups: two-sided Student's *t*-test. **l**. *Usp7* gene locus and knock-in strategy to insert Auxin-Inducible Degron (AID) at the C-terminus of USP7. **m**. Western Blot of USP7 after auxin (IAA) treatment. **n**. Western Blot in parental and knock-in USP7-AID line. The USP7-AID transgene causes lower USP7 expression compared to the parental clone, presumably leading to higher steady-state 2CLC population. **o**. *Rrm1* and *Rrm2* RT-qPCR 48 h after transfection with their respective siRNAs compared to control. Statistical comparisons: two-sided Welch's *t*-test. **p**. Fork speed upon RNAi for RRM1, RRM2 or control. Statistical analysis: two-sided Wilcoxon rank-sum test. **q**. 2CLCs quantification 48 h after siRNA transfection. Statistical analysis: two-sided Student's *t*-test. In bar graphs plots are mean±S.D and dots are individual replicate values. n: number of independent biological replicates. In m and n, dots are values from biological replicates.

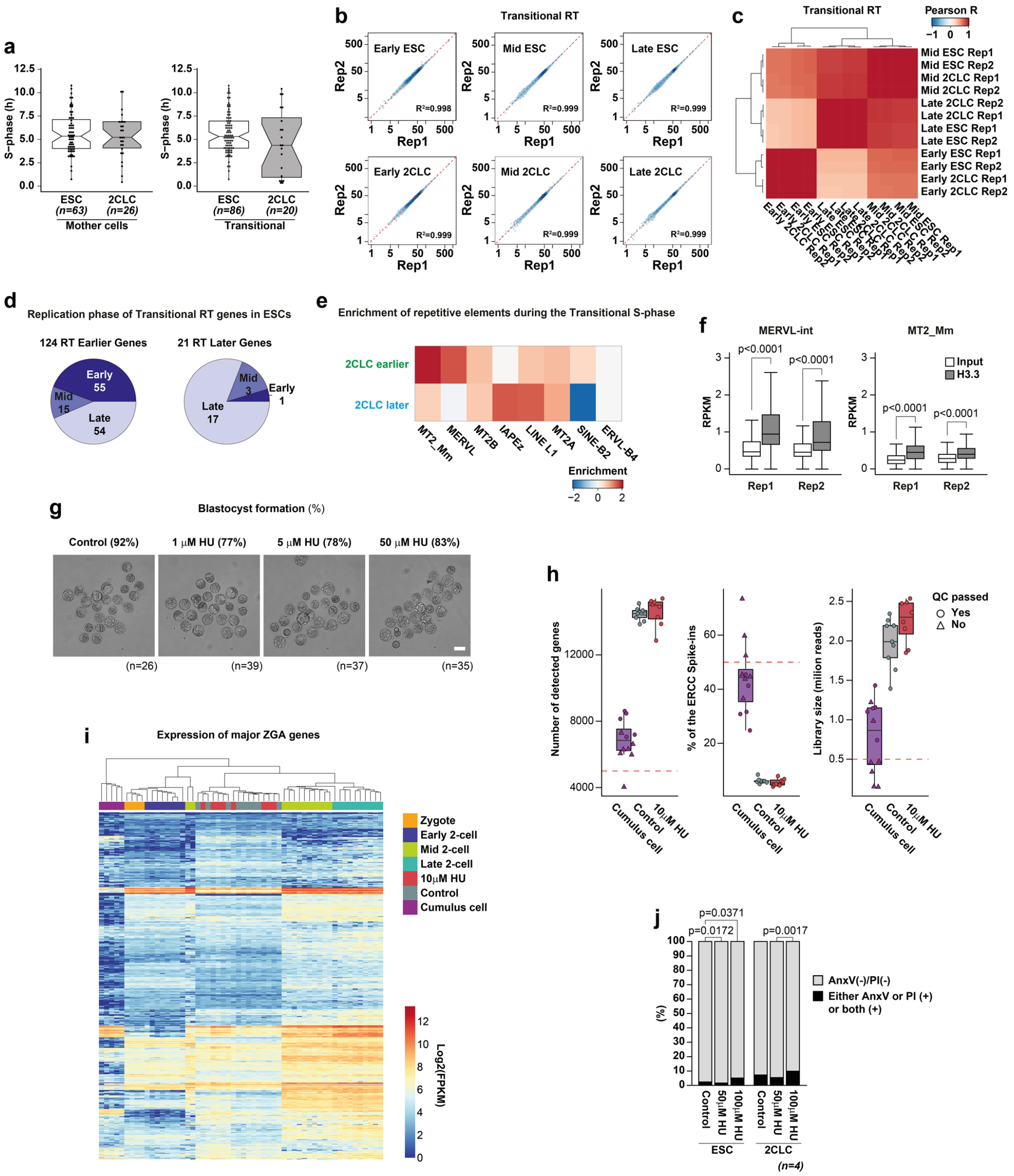


Extended Data Fig. 5 | Impact of modulating replication fork speed on reprogramming. a, b. HU treatment facilitates iPS reprogramming. Alkaline phosphatase-positive iPSC colonies (a) and their quantification results (b) after OKSM induction by Dox in reprogrammable MEFs treated during the indicated time windows of HU. In **b**, the dots indicate the values from individual experiments compared to the control, the middle line is the mean and the boxes depict mean ± SD. Statistical analyses were performed with a generalized linear model using a Poisson distribution. Both the concentration and the days as well as the combination have a significant effect ($p < 0.0001$) on reprogramming efficiency. Scale bar, 1 mm. **c**. **d**. Lineage contribution of ESCs and 2CLCs in chimeric blastocysts. siControl-transfected ESCs, siUSP7-induced 2CLCs, and HU-induced 2CLCs were aggregated with 4–8 cell stage embryos and cultured for 2 days. Blastocysts were analyzed by confocal microscopy, and reconstructed in 3D to determine the position of individual cells in each lineage using phalloidin as cell membrane label. Representative images of cells within ICM and TE (c) and the quantification results (d) of 26, 38 and 32 embryos analyzed per group, respectively, are shown. Data are displayed as the percentage of cells, which upon aggregation, display inner (ICM) or outer (TE) position. Statistical analyses were performed with a Kruskal-Wallis test. Scale bar, 25 μm. **e**. Representative images showing immunostaining of chimeric blastocysts injected with H2B-tdiRFP expressing siUSP7-2CLCs or 50 μM HU induced 2CLCs. H2B-tdiRFP-positive cells expressing Cdx2 and Oct3/4 are indicated by arrowheads and the corresponding Insets at higher magnification are shown. A total of 17 embryos were injected with siUSP7-2CLCs and 21 embryos were injected with HU induced 2CLCs. Note that the Oct3/4 positive cell in the siUSP7 panel depicts a cell in mitosis. Scale bars, 25 μm. **f**. Representative images of nuclear transferred embryos derived from the indicated donor cells 4 days after activation, corresponding to a representative experiment related to Figs. 3l and m. Scale bar, 50 μm.



Extended Data Fig. 6 | See next page for caption.

Extended Data Fig. 6 | Genome-wide analysis of replication timing (RT) in 2CLCs. a. Collection of early, mid, and late S-phase of ESCs and 2CLCs. Each subpopulation of S-phase in ESCs (top) and 2CLCs (bottom) was sorted based on their DNA content upon propidium iodide staining. **b**. Scatter plots of read density at 100 Kbp bins across the whole genome between Repli-seq replicates of ESCs (GFP$^-$ cells) and 2CLCs (GFP$^+$ cells). Pearson R^2 is indicated. **c**. Heatmap depicting Pearson R correlation based on read density at 100 Kbp bins across the genome for the indicated samples. **d**. Repli-seq tracks around the Dux locus at early, mid, and late S-phase in ESCs and 2CLCs indicating early to late ratio as log2(E/L). **e**. Distribution of MERVL elements (MT2_Mm and MERVL_int) according to early, mid, and late replication regions in ESCs. **f**. Heatmaps of histone modification densities in ESCs in the 5-Kbp vicinity of gene bodies for 333 genes with differential replication timing. **g**. Heatmaps of histone modification densities in oocyte in the 5-Kbp vicinity of gene bodies for 333 genes with differential replication timing. **h**. Heatmaps of histone modification densities in 2-cell embryos in the 5-Kbp vicinity of gene bodies for 333 genes with differential replication timing. **i**. Log2-fold enrichment of ENCODE chromatin states among genomic regions shifting to earlier (top) and later (bottom) replication in 2CLCs. **j**. Scatter plots of read density at 100 Kbp bins across the genome between Repli-seq replicates of control siRNA-transfected ESCs, and GFP$^+$ (USP7KD-induced 2CLC) and GFP$^-$ cells following USP7 depletion. Pearson R^2 is indicated. **k**. Heatmap depicting Pearson R correlation based on read density at 100 Kbp bins across the genome in each S-phase of control siRNA transfected ESCs, and GFP$^+$ (USP7KD-induced 2CLC) and GFP$^-$ cells following USP7 depletion. **l**. Usp7KD-induced 2CLCs display changes in replication timing similar to 2CLCs, when compared to ESCs. Scatter plot shows a high degree of correlation (R^2=0.84) between the differences in Early to Late ratio for 2CLCs vs ESCs (x-axis) and for Usp7KD-induced 2CLC vs GFP$^-$ cells following USP7 depletion, among 100 Kbp bins across the genome.



Extended Data Fig. 7 | See next page for caption.

Extended Data Fig. 7 | Genome-wide analysis of replication timing (RT) in the transitional S-phase during which 2CLC emerge. a. S-phase length in mother cells of 2CLCs and ESCs and in ESCs and 2CLCs during the transitional S-phase. Boxplots: median (middle line) IQR (boxes) and extent of data without outliers (whiskers,>1.5x IQR). Notches extend to +/−1.58xIQR/sqrt(n), indicating confidence intervals. Dots are individual measurements arranged in 0.2 h bins. **b**. Scatter plots of read density at 100 Kbp bins across the genome between Repli-seq replicates of the transitional S-phase of cells transitioning from ESC to 2CLC. Pearson R^2 is indicated. **c**. Pearson R correlation heatmap based on read density at 100 Kbp bins across the genome in each S-phase of cells transitioning from ESC to 2CLC. **d**. Pie charts of numbers of genes that replicate in early, mid and late S-phase in ESCs, for gene-sets whose replication shifted to earlier and later timing during the transitional S-phase in emerging 2CLCs. **e**. Enrichment of repeat elements across genomic regions changing to an earlier and later replication timing during the transitional S-phase at which 2CLCs emerge. **f**. H3.3 enrichment at MERVL-int and MT2_Mm repeats in 2-cell embryos. Reads were normalized by sequencing depth and length, data from two biological replicates shown separately as 25th and 75th percentiles (box), median (line) and smallest and largest values within 1.5×IQR of the hinge (whiskers). Statistical analyses against the input were with two-sided Wilcoxon-signed-rank test. **g**. Developmental progression of fertilized embryos upon HU treatment. Zygotes collected at 17-18 h post-hCG were treated with HU until 48 h posthCG. Embryos reaching the blastocyst stage (%) are indicated; n: number of embryos analyzed. Scale bar, 100 µm. **h**. RNAseq quality control (QC) metrics for nuclear transferred embryos (control and 10µM HU-treated) and single cumulus cells. QC thresholds (red dotted lines) are indicated; samples failing QC (triangles) were discarded. Boxplots show median and IQR; whiskers depict the smallest and largest values within 1.5×IQR. **i**. Heatmap with expression of ZGA genes upon nuclear transfer compared to *in vivo* derived embryos. **j**. Cell death analysis by dual Annexin-V and propidium iodide (PI) staining following HU treatment. Cells positive for either or both Annexin-V and PI were considered dead. Statistical analyses: two-sided Student's *t*-test.

nature portfolio

Corresponding author(s): Maria-Elena Torres-Padilla

Last updated by author(s): Nov 19, 2021

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	LAS X (version 3) and Prairie View (version 5) were used for imaging data acquisition and basic image manipulation. BD FACSDiva software (version 8) was used for cell sorting data acquisition.
Data analysis	Microscoft excel (version 15) were used for data analysis. Adobe Illustrator CS6 (version 16) and Adobe Photoshop CS6 (version 13) were used for Figure preparation. FlowJo (version 10), BD FACSDiva software (version 8), and BD FACSChorus (version 2.0) were used for analysis of cell cycle distribution and cell population. Illumina TruSeq adapters and the overrepresented sequences in FastQC were trimmed using the palindrome mode of trimmomatic v0.38 under the parameters ILLUMINACLIP:Adapters:3:30:8:1:true LEADING:10 TRAILING:10 SLIDINGWINDOW:4:15 MINLEN:10. Bowtie2 was run for aligning the trimmed reads to the mm10 mouse genome vM19 (GRCm38.p6) downloaded from GENCODE. Reads were fixed using fixmate; unmapped and multimapped reads were removed. Peak calling was carried out using the callpeak function of MACS2 v2.1.2.20181002, by setting a threshold of q=0.01. Deeptools toolkit v3.1.3, was used to compute the peak scores and plot the heatmap using the functions computeMatrix and plotHeatmap. For RNA sequencing analysis and sample clustering, STAR aligner was used to map sequencing reads to transcripts in the mouse mm9 reference genome. Read counts for individual transcripts were produced with HTSeq-count, followed by the estimation of expression values and detection of differentially expressed transcripts using EdgeR. For Analysis of H3.3 enrichment on MERVL, reads overlapping MERVL elements (MT2_Mm, MERVL-int) were quantified for each locus using bedtools (v2.26.0) and normalized by the sequencing depth and length of the fragment. The GTF annotation used was from the TEtranscripts. For Single embryo RNA sequencing analysis, analyses were carried out on R (version 4.0.2). Reads were aligned with STAR (2.7.3a) to the mm10 genome with the default settings and counting the reads for every gene using the option “--quantMode GeneCounts”.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

Sequencing data generated during this study have been deposited in the Gene Expression Omnibus (GEO) under accession codes GSE136228 (Repli-seq and RNA-seq data).
Previously published RNA-seq datasets re-analysed here are available under accession codes GSM1933935 (MII oocyte); GSM1625860 (Zygote); GSM1933937 (Early 2-cell); GSM1625862 (2-cell); GSM1625864 (4-cell); GSM1625867 (8-cell); GSM1625868 (ICM); GSM838739 (2Ctomato negative ESCs); GSM838738 (2Ctomato positive 2CLCs); GSM1625873 (mESC); E-MTAB-2684 (Control ES cells without treatment); E-MTAB-2684 (ES cells, untreated GFP minus); E-MTAB-2684 (2CLCs, untreated GFP plus); E-MTAB-2684 (CAF-1 KD induced 2CLCs, si-p150 GFPplus); GSM 1933935 (ZMYM2-depleted ESC); GSM3110926 (Dox-induced NELFA positive cells) ; GSM3110919 (NELFA(high) GFP positive); GSM4224405 (miR-344(DR+/+)).
Previously published ChIP-seq datasets re-analysed here are available under accession codes GSE73952 (H3K4me3 and H3K27me3, 2-cell-embryo); GSE97778 (H3K9me3, 2-cell-embryo); GSE23943 (H3K4me3, H3K9me3, and H3K27me3, ESC); GSE74952 (H3K4me3, oocyte); GSE139527 (H3.3, 2-cell-embryo).
Figures with associated raw data are as follows: Fig. 3e, 3f, 4a-k, 5b, 5c, Extended Data Fig. 4d-g, 6b-l, 7b-f, 7h, and 7i.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	To do the statistical test, at least 3 biological replicates were included (unless otherwise stated) based on previously published work and preliminary studies as standard for this field of research. See Figures legends for each experiment.
Data exclusions	No data were excluded.
Replication	All attempts at replication were successful as reported in the manuscript
Randomization	Cells were allocated at random to experimental groups as stated in the Methods
Blinding	Double-blind counting was carried out for Extended Data Fig. 5d in which relatively subjective counting was performed. All other analysis was objectively performed using automated approaches.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study	
☐	☒	Antibodies
☐	☒	Eukaryotic cell lines
☒	☐	Palaeontology and archaeology
☐	☒	Animals and other organisms
☒	☐	Human research participants
☒	☐	Clinical data
☒	☐	Dual use research of concern

Methods

n/a	Involved in the study	
☒	☐	ChIP-seq
☐	☒	Flow cytometry
☒	☐	MRI-based neuroimaging

Antibodies

Antibodies used	Antibodies used were as follows: anti-BrdU(IdU) (BD Biosciences 347580), anti-BrdU(CldU) (Novus NB500-169, Abcam ab6326), anti-

Antibodies used	single stranded DNA (Milipore MAB3034), anti-Zscan4 (Milipore AB3430), anti-Oct3/4 (BD Biosciences 611203, MBL PM048), anti-Cdx2 (BioGenex AM392-5M), anti-Usp7 (BETHYL A300-034), anti-Gapdh (Milipore MAB374), anti-BrdU (Sigma B8434), anti-gH2AX (Abcam ab22551), anti-SNAP (NEB P9310), Secondary antibodies used were A11001, A121429, A11077, A21236, A16078, and A16110. More detailed information of antibodies are written in Supplementary Table S7.
Validation	Manufacturer validated that three anti-BrdU (BD Biosciences 347580), (Novus NB500-169), and (Abcam ab6326) recognize IdU (https://www.bdbiosciences.com/us/applications/research/apoptosis/purified-antibodies/purified-mouse-anti-brdu-b44/p/347580), CldU (https://www.novusbio.com/products/bromodeoxyuridine-brdu-antibody-bu1-75-icr1-_nb500-169), and CldU(https://www.abcam.com/brdu-antibody-bu175-icr1-proliferation-marker-ab6326.html), respectively. Anti-BrdU (Sigma B8434) were reported to be applicanle for IP (Shibata, Elife, 2016). Anti-single stranded DNA (Milipore MAB3034) was validated by the manufacturer (https://www.merckmillipore.com/DE/de/product/Anti-DNA-Antibody-single-stranded-clone-16-19,MM_NF-MAB3034?ReferrerURL=https%3A%2F%2Fwww.google.com%2F&bd=1). Anti-Usp7 (BETHYL A300-034) was validated by Western blot combination with siRNA experiment. Anti-Zscan4 (Milipore AB3430), anti-Oct3/4 (BD Biosciences 611203), and anti-Cdx2 (BioGenex AM392-5M) were validated our previous studies (Ishiuchi, NSCB, 2015, Rodriguez-Terrones, Nat. Genet., 2018, Burton, NCB, 2020). Anti-gH2AX (https://www.abcam.com/gamma-h2ax-phospho-s139-antibody-3f2-ab22551.html), anti-Oct3/4 (https://ruo.mbl.co.jp/bio/dtl/A/?pcd=PM048), and anti-SNAP (https://international.neb.com/products/p9310-anti-snap-tag-antibody-polyclonal#Product%20Information) were validated by manufacturerer.

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)	All of ES cells originating from our previous study (Miyanari, Nature, 2012 and Ishiuchi, NSCB, 2015).
Authentication	We have validated that our 2C-reporter cell line reflect endogenous expression of MERV-L by IF and also performed side by side comparisons based on RNAseq with the reporter cell lines that we and others have validated before (Ishiuchi, NSCB, 2015, Macfarlan, Nature, 2012, De Iaco, Nat. genet., 2017, Hendrickson, Nat. genet., 2017, Rodriguez-Terrones, Nat. Genet., 2018). Validation of FUCCI cells was done by sorting mCherry-hCdt1(1/100)Cy(-)-positive, iRFP-hGeminin(1/110)-negative cells and confirmed that subpopulation corresponded to G1 peak which was obtained by PI staining. Knock-in of the AID cassette was validated by genomic PCR with specific primer sets after the drug selection and homoallelic mutant were used for the experiments.
Mycoplasma contamination	All cell lines tested negative for mycoplasma contamination.
Commonly misidentified lines (See ICLAC register)	No commercially misidentified cell lines were used.

Animals and other organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research

Laboratory animals	F1 (C57BL6 X CBA/H) mice were used to provide oocytes and crossed with F1 males to provide zygotes.
Wild animals	This study did not use wild animals.
Field-collected samples	This study did not involve field-collected samples.
Ethics oversight	All experiments were performed under the authorization of French legislation or the Upper Bavarian authorities.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Flow Cytometry

Plots

Confirm that:

☒ The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).

☒ The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).

☒ All plots are contour plots with outliers or pseudocolor plots.

☒ A numerical value for number of cells or percentage (with statistics) is provided.

Methodology

Sample preparation	For isolation and quantification of 2CLCs, cells were washed with twice with PBS and treated with 0.25% trypsin. After neutralization with ESC medium, cells were collected by centrifugation and the dissociated single cells were resuspended in ESC medium. To calculate the population of 2CLCs, we counted turboGFP-positive ESCs after extrusion of dead and doublet cells based on the forward and side scatter profiles. After sorting, cells were collected in normal culture medium and kept at 4°C. For collection of cells in G1-phase in Fig 2e and Extended Data Fig. 2e, we sorted mCherry-hCdt1(1/100)Cy(-)-positive, iRFP-hGeminin(1/110)-negative subpopulation based on their fluorescence. For cell cycle analysis, the dissociated single cells

	were fixed with 70 % ethanol for 30 min. After treatment with 250 µg/mL RNaseA (Thermofisher Scientific) for 5 min, cells were treated with 50 µg/mL propidium iodide (PI) to stain DNA.
Instrument	Sorting was performed on a BD Biosciences FACSAria III.
Software	FlowJo (version 10) and BD FACSDiva software (version 8) were used for analysis of cell cycle distribution and cell population, respectively.
Cell population abundance	Whenever cell numbers were not an issue, fluorescence was verified after sorting and was usually 95-100%. Downstream experiments always confirmed a very high degree of sorting purity.
Gating strategy	Stringent gatings were always used, leaving a significant gap in between negative/positive or low/high populations.

☒ Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.

6.5 Manuscript 1

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.12.476097; this version posted January 12, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Myeloid-biased HSC require Semaphorin 4A from the bone marrow niche for self-renewal under stress and life-long persistence

Dorsa Toghani[1,13], Sharon Zeng[1,13], Elmir Mahammadov[2,14], Edie I. Crosse[1,14], Negar Seyedhassantehrani[3], Christian Burns[3], David Gravano[3], Stefan Radtke[1], Hans-Peter Kiem[1], Sonia Rodriguez[4], Nadia Carlesso[4], Amogh Pradeep[1], Nicola K. Wilson[5], Sarah J. Kinston[5], Berthold Göttgens[5], Claus Nerlov[6], Eric Pietras[7], Marion Mesnieres[8], Christa Maes[8], Atsushi Kumanogoh[9], Thomas Worzfeld[10,11], Peter Kharchenko[12], David T. Scadden[12], Antonio Scialdone[2], Joel A Spencer[3], and Lev Silberstein[1]*

[1]Fred Hutchinson Cancer Research Center, Seattle, USA,[2]Helmholtz Zentrum Muenchen, Munich, Germany, [3]University of California Merced, USA, [4]Beckman Research Institute, City of Hope, USA, [5]. Department of Haematology, Jeffrey Cheah Biomedical Centre, Wellcome - MRC Cambridge Stem Cell Institute, University of Cambridge, UK, [6]University of Oxford, UK, [7]University of Colorado, USA, [8]KU Leuven, Belgium, [9]University of Osaka, Japan, [10]University of Marburg, Germany, [11]Max-Planck-Institute for Heart and Lung Research, Germany, [12]Harvard University, Boston, USA

[13]DT and SZ contributed equally

[14]EM and EIC contributed equally

*Corresponding author, email lsilbers@fredhutch.org

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.12.476097; this version posted January 12, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

SUMMARY

Tissue stem cells are hierarchically organized. Those that are most primitive serve as key drivers of regenerative response but the signals that selectively preserve their functional integrity are largely unknown. Here, we identify a secreted factor, Semaphorin 4A (Sema4A), as a specific regulator of myeloid-biased hematopoietic stem cells (myHSC), which are positioned at the top of the HSC hierarchy. Lack of Sema4A leads to exaggerated myHSC (but not downstream “balanced” HSC) proliferation after acute inflammatory stress, indicating that Sema4A enforces myHSC quiescence. Strikingly, aged Sema4A knock-out myHSC expand but almost completely lose reconstitution capacity. The effect of Sema4A is non cell-autonomous, since upon transplantation into Sema4A-deficient environment, wild-type myHSC excessively proliferate but fail to engraft long-term. Sema4A constrains inflammatory signaling in myHSC and acts via a surface receptor Plexin-D1. Our data support a model whereby the most primitive tissue stem cells critically rely on a dedicated signal from the niche for self-renewal and life-long persistence.

Hematopoietic Stem Cells, Inflammation, Aging, Niche

INTRODUCTION

In multiple tissues, stem cells are hierarchically organized and contain distinct, functionally specialized subsets. For example, in the brain (Sachewsky et al., 2019), skeletal muscle (Scaramozza et al., 2019), cornea (Altshuler et al., 2021; Farrelly et al., 2021) and skin (Hsu et al., 2011; Rompolas et al., 2013), the most primitive stem cells (which are also largely quiescent) become activated by injury or stress, whereas their “downstream”, more differentiated counterparts are more proliferative and mainly engaged in on-going tissue repair. Although this two-tiered organization of the stem cell compartment is essential for life-long

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.12.476097; this version posted January 12, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

tissue maintenance, the mechanisms that ensure functional preservation and persistence of individual stem cell subsets within a tissue stem cell hierarchy remain largely unknown.

In the bone marrow, the stem cell hierarchy is exemplified by the myeloid-biased and “balanced” (HSC) subsets, in which the former is considered the most primitive (Challen et al., 2010; Morita et al., 2010; Sanjuan-Pla et al., 2013). Compared to balanced HSC (balHSC), myeloid-biased HSC (myHSC) are inherently skewed towards myeloid differentiation, endowed with a higher self-renewal potential and possess a superior ability to enter cell cycle in response to inflammatory stress (Mann et al., 2018; Matatall et al., 2014; Mitroulis et al., 2018). Although these properties are beneficial for powerful and timely host defense response, they likely account for myHSC expansion during inflammation and aging, which is associated with profound and irreversible functional loss (Beerman et al., 2010; Esplin et al., 2011; Grover et al., 2016; Pang et al., 2011). Thus, despite being positioned at the top of the HSC hierarchy, myHSC appear most vulnerable to stress-induced damage.

In the current study, we identify a membrane-bound and secreted protein Semaphorin 4A (Sema4A) as a myHSC-protective factor. We show that under stress conditions, such as aging and transplantation, the absence of Sema4A results in excessive expansion and functional attrition of myHSC. Surprisingly, balHSC are only minimally affected, suggesting that the effect of Sema4A is myHSC-specific. We further demonstrate that Sema4A from the bone marrow niche is essential for myHSC self-renewal and identify Plexin-D1 as a functional receptor. Our results reveal that by selectively preserving a functional myHSC pool, Sema4A plays a key role in long-term maintenance of the HSC hierarchy.

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.12.476097; this version posted January 12, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

RESULTS

Sema4A regulates quiescence of mouse and human hematopoietic stem/progenitor cells

We have previously established proximity-based analysis as a platform for niche factor discovery (Silberstein et al., 2016). We compared single cell RNA-Seq signatures of osteolineage cells (OLC) that were located in close proximity to a single transplanted HSC (proximal OLC) and further away (distal OLC), and functionally validated several secreted factors with a higher expression in proximal OLC (Angiogenin, IL18, Embigin, VEGF-C) as non cell-autonomous regulators of hematopoietic stem cell/progenitor quiescence (Fang et al., 2020; Goncalves et al., 2016; Silberstein *et al.*, 2016). Because Sema4A displayed a similar expression difference by being significantly more abundant in proximal OLC, as shown in Fig. 1A, we hypothesized that it could also act as a niche-derived HSC quiescence regulator.

Semaphorins are a large family of membrane-bound and/or secreted proteins which mediate cell-cell communications in neural development, angiogenesis, immune response and cancer (Fard and Tamagnone, 2021; Kolodkin et al., 1993). In keeping with our hypothesis for a possible non cell-autonomous regulatory role of Sema4A in hematopoiesis, we found that Sema4A transcripts were detectable in niche cell subsets, such as CD31^{+} endothelial cells and niche factor-enriched VCAM1^{high}Embigin^{+} OLC fraction, as also supported by the published data (Baccin et al., 2020) (Fig. S1A and S1B). Of note, expression of Sema4A in CD45^{-}Ter119^{-}ALCAM^{+} bone-lining cells (Valletta et al., 2020) was significantly increased in aged mice (Fig. S1C). In human bone marrow, Sema4A mRNA was also present in bone-lining and endothelial cells, as demonstrated by single-molecule fluorescent in situ hybridization (Fig. S1D).

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.12.476097; this version posted January 12, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

In order to gain initial functional insights, we tested the effect of Sema4A on proliferation of mouse and human hematopoietic stem/progenitor cells *in vitro*. Consistent with our hypothesis, addition of recombinant Sema4A-Fc protein resulted in fewer lin$^-$c-Kit$^+$Sca-1$^+$ (LKS) cells after 24 hours of liquid culture and suppression of hematopoietic colony formation in a dose-dependent manner (Fig. 1B and S1E). Similarly, human recombinant Sema4A-Fc inhibited *in vitro* proliferation of human bone marrow CD34$^+$ cells from several donors, as measured by the Carboxyfluorescein succinimidyl ester (CFSE) dilution assay (Takizawa et al., 2011) (Fig. 1C and S1F). Collectively, these data indicate that Sema4A non cell-autonomously restricts HSPC proliferation, and that this property is conserved between mice and humans.

Next, we examined the role of Sema4A in steady-state hematopoiesis using a germline knock-out model. Sema4A knock-out (Sema4AKO) mice are viable and have a normal life span (Kumanogoh et al., 2005). However, baseline analysis of young Sema4AKO animals revealed subtle but reproducible anemia and thrombocytosis (Fig. S1G). Moreover, in the bone marrow, while the number and frequency of long-term HSC (LT-HSC, defined as LKS CD48$^-$CD34$^-$Flk2$^-$ CD150$^+$) was similar between the genotypes, the differentiation was skewed towards the myeloid lineage, as evidenced by increased frequency of myeloid-committed progenitors MPP2 and mature myeloid cells (Fig. 1D, see Fig. S1H for gating strategy). Importantly, young Sema4AKO HSC displayed more active cycling, as assessed by Ki-67/DAPI staining and EdU incorporation (Fig. 1E, S1I and S1J). Single-cell RNA-Seq of HSPC from young WT and Sema4AKO mice, while showing no difference in cluster distribution (Fig. S1K), revealed positive enrichment for the terms “Kegg ribosome” and “Electron transport chain oxphos system in mitochondria” in Sema4AKO cells within the HSC cluster, consistent with disruption of quiescence (Fig. 1F, 1G, S1L, and Suppl. File). Collectively, these results indicate that loss of Sema4A leads to increased HSC proliferation, metabolic activation and myeloid-biased differentiation.

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.12.476097; this version posted January 12, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Sema4A/PlxnD1 signaling constrains the response of myeloid-biased HSC to proliferative stress

Given recent evidence suggesting that lineage-restricted HSC subsets are differentially regulated by niche-derived signals, we hypothesized that myeloid bias in the Sema4AKO model is due to the lack of quiescence-inducing effect of Sema4A specifically on myHSC. While baseline analysis demonstrated no appreciable difference in cell cycle status between Sema4AKO myHSC (LKS $CD48^{-}CD34^{-}Flk2^{-}CD150^{high}$) and balHSC (LKS $CD48^{-}CD34^{-}Flk2^{-}$ $CD150^{low}$) (Beerman *et al.*, 2010) [data not shown], exposure to acute inflammatory stress revealed important HSC subset-specific differences.

In particular, twenty-four hours after injection with polyinosinic:polycytidilic acid (Poly (I:C)) (Walter et al., 2015)(Fig 2A), we found a significant increase in the percentage of Sema4AKO myHSC in G2M phase of cell cycle (Fig.2B and S2B) while no cell cycle difference was observed in balHSC subset (Fig. 2C and S2C, see Fig. S2A for gating strategy under inflammatory conditions) (Hirche et al., 2017). Of note, myHSC and balHSC cycling was comparable in PBS-injected WT/Sema4AKO animals (Fig S2D and S2E) suggesting that the above changes in Sema4AKO myHSC were due to exaggerated response to inflammatory stress. Indeed, subsequent RNA-Seq analysis of myHSC and balHSC from Poly (I:C)-injected animals revealed enrichment for the terms “IL6-Jak Stat3 signaling” and “Interferon alpha response” which was unique to Sema4AKO myHSC dataset (Fig. 2D, 2E, S2F, Suppl. File). Thus, our results reveal that the absence of Sema4A promotes myHSC cell cycle entry and lead to enhanced myHSC sensitivity to inflammatory signaling.

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.12.476097; this version posted January 12, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Next, we asked if Sema4A deletion differentially impacts long-term reconstitution capacity of the two HSC subsets. To this end, we isolated myHSC and balHSC from WT and Sema4AKO (CD45.2) donors and transplanted equal number of cells from each subset into lethally irradiated WT (CD45.1) recipients (Fig. 2F). As shown in Fig. 2G, Sema4AKO myHSC displayed a significantly higher level of post-transplant reconstitution as compared to WT myHSC controls, with the difference increasing over time (see Fig. S2G for gating strategy used in chimerism analysis). In contrast, this trend was considerably weaker in the recipients of WT/Sema4AKO balHSC and no longer detectable 24 weeks post-transplant (Fig. 2H). In addition, Sema4AKO myHSC (but not balHSC) graft displayed excessive lymphoid skewing (Fig. S2H and S2I). These data provide further support for the myHSC-specific action of Sema4A, as evidenced by enhanced output and impaired differentiation of transplanted Sema4AKO myHSC.

In order to establish a cellular mechanism for this effect, we sought to identify a functional receptor for Sema4A on myHSC. Analysis of published HSC gene expression datasets (Cabezas-Wallscheid et al., 2017; Cabezas-Wallscheid et al., 2014) revealed that amongst known receptors for Sema4A, Plexin-B2 *(PlxnB2*) and Plexin-D1 *(PlxnD1)* had the highest expression level in HSC (Fig. 2I). However, PlxnB2 has been described as a receptor for Angiogenin (Yu et al., 2017) which has no effect on myeloid differentiation (Goncalves *et al.*, 2016). We therefore considered PlxnD1 the most likely candidate. Interestingly, the ability of Sema4A/PlxnD1 signaling to constrain stress-induced proliferation (as it would be for myHSC) has been already shown for the endothelial cells (Toyofuku et al., 2007). Furthermore, our analysis of PlxnD1-GFP reporter mice (Gong et al., 2003) revealed a significantly higher level of PlxnD1 expression in myHSC as compared to balHSC (Fig. 2J), which was consistent with a predominant functional effect of Sema4A. Of note, a fraction of CD34$^+$CD90$^+$ human HSC also expressed PlxnD1 (Fig. S2J).

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.12.476097; this version posted January 12, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Global deletion of PlxnD1 in mice is embryonic lethal due to structural cardiac and vascular defects (Serini et al., 2003), thus precluding functional analysis of adult HSC from these animals. We therefore conditionally deleted PlxnD1 by crossing PlxnD1 "floxed" (Zhang et al., 2009) mice with the Mx1-Cre strain (Kuhn et al., 1995). We confirmed a complete excision of PlxnD1 by PCR and Q-PCR analysis of sorted LKS cells after Poly (I:C) induction (Fig. S2K and S2L). Baseline analysis of PlxnD1$^{fl/fl}$ Mx1-Cre(+) and PlxnD1$^{fl/fl}$ Mx1-Cre(-) mice revealed no significant differences in blood counts, HSC cell cycle and HSPC and mature cell frequency, except for a slight increase in MPP2 and lin$^-$Sca-1$^-$c-Kit$^+$ myeloid progenitors, suggesting myeloid bias (Fig. S2M and S2N, Suppl. Table 1). However, competitive transplantation of myHSC and balHSC from of PlxnD1$^{fl/fl}$ Mx1-Cre(+) and PlxnD1$^{fl/fl}$ Mx1-Cre(-) mice revealed significantly higher reconstitution by PlxnD1-deficient myHSC while their balHSC counterparts engrafted normally, thus recapitulating the phenotype of transplanted myHSC and balHSC from Sema4AKO mice (Fig. 2K-L and S2O-Q). In sum, our results are consistent with a previously unrecognized role of PlxnD1 as a functional receptor for Sema4A on myHSC.

Sema4A prevents excessive myHSC expansion and functional loss with age

Our observation that Sema4A loss enhances myHSC responsiveness to proliferative challenges, such as acute inflammation and transplantation, prompted us to investigate whether this will lead to impaired myHSC function upon chronic inflammatory stimulation, as occurs during aging (Kovtonyuk et al., 2016). Analysis of peripheral blood in aged Sema4AKO mice revealed progressive anemia, thrombocytosis and neutrophilia (Fig. 3A). In order to rule out systemic inflammation as a cause of the above abnormalities, we examined the plasma levels of 36 proinflammatory cytokines in aged WT and Sema4AKO mice (including thrombopoietin and G-CSF) but detected no significant differences (Suppl. Table 2).

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.12.476097; this version posted January 12, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Immunophenotypic analysis of the bone marrow in aged Sema4AKO mice demonstrated a higher number of primitive hematopoietic cells and marked myeloid expansion, as evidenced by increased frequency and absolute number of myeloid progenitors and mature myeloid cells (Fig. 3B and S3A). Critically, we observed a marked (~2.5-fold) increase in the absolute number of myHSC while the number of balHSC was comparable with that of aged-matched WT controls (Fig. 3C and S3B).

The amplified aged Sema4AKO myHSC population may represent either expanded *bona fide* myHSC or a more differentiated progeny which retained immunophenotypic features of myHSC but lost long-term regenerative potential following expansion (Bernitz et al., 2016). To distinguish between these two possibilities, we competitively transplanted equal numbers of myHSC from aged Sema4AKO and WT animals into lethally irradiated WT recipients (Fig. 3D).

Strikingly, aged Sema4AKO myHSC, while still displaying myeloid bias, generated a markedly lower level of donor chimerism in peripheral blood (range 0-1.43% vs 9.93-77.5% in aged WT myHSC controls) (Fig. 3E and S3C) and failed to produce a detectable long-term graft in the bone marrow (Fig. S3E). In contrast, balHSC from both WT and Sema4AKO mice gave rise to comparable levels of peripheral blood and bone marrow donor chimerism (Fig. 3F and S3D). These data demonstrate that during aging, Sema4A absence leads to profound functional attrition of phenotypic myHSC but is inconsequential for balHSC.

Aiming to understand the molecular events which are responsible for the myHSC-specific effect of Sema4A, we performed single cell RNA-Seq analysis of myHSC and balHSC from aged WT and Sema4AKO mice using the Smart-Seq2 protocol (Picelli et al., 2014). We sorted 192 cells per group (768 total), of which 642 were selected for analysis following quality control (see Methods). As expected, WT myHSC had a higher expression of *Slamf1*, self-renewal/low-

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.12.476097; this version posted January 12, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

output-associated genes (*CD74*, *Ly6a*, *vWF*, *Procr*) and lower expression of cell cycle-related genes (*Cdk6* and *Mki67*) as compared to WT balHSC (Fig. S3F) (Becker-Herman et al., 2021; Kent et al., 2009; Kent et al., 2008; Laurenti et al., 2015; Morcos et al., 2017; Rodriguez-Fraticelli et al., 2020).

A transcriptome-wide analysis revealed that within the myHSC fraction, WT and Sema4AKO cells formed distinct, minimally overlapping clusters (Fig. 3G) while balHSC of both genotypes merged together, indicating that the absence of Sema4A induces transcriptional changes predominantly in myHSC (Fig. 3H). We quantified this HSC subset-specific difference by detecting a greater correlation-based distance between WT/ Sema4AKO myHSC compared to WT/ Sema4AKO balHSC (Fig. 3I); Wilcoxon Rank-Sum test, p-value 3.1e-85, see Methods for further details).

Next, we examined the transcriptional features of the above aged HSC subsets in more detail. Consistent with the results of the clustering analysis, the number of genes which were differentially expressed in aged Sema4AKO vs WT myHSC was much greater (431) compared to aged Sema4AKO vs WT balHSC (30). In the aged Sema4AKO myHSC signature, we noted markedly reduced expression of genes that normally constrain HSC pool and promote HSC self-renewal (*CD74, vWF, Ly6a, Mllt3*)(Becker-Herman *et al.*, 2021; Calvanese et al., 2019; Kent *et al.*, 2009; Kent *et al.*, 2008), consistent with their excessive expansion and loss of stemness (Fig. S3G). Moreover, GSEA demonstrated a significant enrichment for the terms “p53 pathways” (top genes: *Jun, Fos, Sesn1*) and “TNF-alpha/NFkB signaling” (top genes: *Fosb, Egr1, Jun*) and as well as a recently defined “core aging HSC signature” (Svendsen et al, 2021) (Fig. 3J, 3K, S3H, Suppl. File).

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.12.476097; this version posted January 12, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

While “TNF-alpha/NFkB signaling” was also enriched in aged Sema4AKO balHSC (Fig. S3I), no enrichment for “p53 pathway” was observed, and enrichment for the “core aging HSC signature” was much weaker (FDR=0.09, P=0.047 for balHSC vs FDR=0.002, P=0.001 for myHSC, (Fig.S3H and data not shown). These findings suggest that in the absence of Sema4A, aged myHSC sustain a greater degree of stress- and inflammation-induced damage (Walter *et al.*, 2015). Consistent with this notion, *in silico* cell cycle analysis (Scialdone et al., 2015) revealed reduced cycling in aged Sema4AKO myHSC but not in balHSC (Fig. 3L). While loss of proliferative capacity in myHSC occurs during normal aging (Montecino-Rodriguez et al., 2019), it was more prominent in aged Sema4AKO myHSC. In conjunction with other phenotypic (expansion and functional loss) and molecular (aging HSC signature) features of normal aging which were exaggerated in aged Sema4AKO myHSC, this suggests that the absence of Sema4A leads to premature myHSC aging.

Given that amplification of aged Sema4AKO phenotypic myHSC was accompanied by a marked expansion of downstream myeloid progeny, we wondered if accelerated differentiation was another factor which would explain their functional loss. We addressed this question using diffusion pseudotime (DPT) analysis (Haghverdi et al., 2016), which can quantify the differentiation state of each cell going from naive (corresponding to HSC) to more mature (multipotent progenitors, MPP). To this end, we first generated 10x Genomics single cell RNA-Seq profiles of lin$^-$c-Kit$^+$ HSPC from 74-weeks old WT animals, i.e. age-matched with WT/Sema4AKO animals for the Smart-Seq2 single-cell RNA-Seq experiment described above. In this 10x dataset, by mapping expression of previously described markers (Nestorowa et al., 2016) we identified the clusters that correspond to HSC (Cluster 2; *Ly6a, Procr, Hlf*) and MPP (Cluster 0; *Cd34, Cebpa, Ctsg*) (see Methods and Fig. S3J-L). Next, we utilized the transcriptomes of cells within these clusters to estimate a differentiation trajectory (Fig. S3M, see Methods for further details), in which higher DPT values correspond to more mature cells.

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.12.476097; this version posted January 12, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

As expected, analysis of known self-renewing marker genes revealed downregulation of vWF, *Mpl*, *Fdg5*, C*tnnal1*, *Procr*, and upregulation of *Ctsg* and *Cbpa* as cells progressed from HSC to MPP (Fig. S3N).

We then estimated a DPT value along this trajectory for myHSC and balHSC from aged WT and Sema4AKO mice which were profiled in the Smart-Seq2 experiment described above. Consistent with previously reported myHSC/balHSC hierarchy, the DPT values of WT aged myHSC were lower than WT balHSC, indicating that myHSC are more primitive than balHSC (Fig. S3O), (Carrelha et al., 2018; Morita *et al.*, 2010). Importantly, the comparison between aged WT myHSC and aged Sema4AKO myHSC revealed that the DPT values for aged Sema4AKO myHSC were higher, suggesting that they became more differentiated (p-value = 0.0002, Wilcoxon rank-sum test, Fig. 3M). Conversely, no significant difference was found between the DPT distributions of aged WT and aged Sema4AKO balHSC (Fig.3N). Thus, our DPT analysis demonstrates that the absence of Sema4A during aging leads to premature, myHSC-specific activation of the differentiation transcriptional program.

In sum, our immunophenotypic, functional and transcriptional analysis identified Sema4A as a critical regulator of myHSC self-renewal and differentiation. The profound loss of regenerative capacity in aged Sema4AKO myHSC, as observed in the transplant experiments, likely represents a cumulative effect of partially overlapping cellular defects, such as inflammatory injury, premature aging and accelerated differentiation.

Sema4A from the bone marrow niche restrains stress-induced myHSC proliferation and maintains self-renewal

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.12.476097; this version posted January 12, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Since Sema4A is expressed in both non-hematopoietic and hematopoietic cells, including HSC(Baccin *et al.*, 2020; Cabezas-Wallscheid *et al.*, 2017; Cabezas-Wallscheid *et al.*, 2014), we asked which cellular source of Sema4A was functionally indispensable for myHSC function. First, we investigated the role of HSC-derived Sema4A by employing conditional deletion with Mx1-Cre. We confirmed Cre-induced recombination of the “floxed” Sema4A allele in HSPC by PCR and Q-PCR analysis (Fig. S4A and S4B). Analysis of Sema4A$^{fl/fl}$ Mx-1Cre (+) animals at the steady-state revealed no difference in peripheral blood counts, bone marrow cellularity, cell cycle and frequency of HSPC and mature cells, as compared to Sema4A$^{fl/fl}$ Mx1-Cre(-) controls (Suppl. Table 1, Fig. S4C and S4D). In competitive transplantation assay, no difference in long-term reconstitution capacity of myHSC and balHSC from Sema4A$^{fl/fl}$ Mx1-Cre(+) and Sema4A$^{fl/fl}$ Mx1-Cre(-) mice was observed (Fig. 4A-B and Fig. S4E-G). These data indicate that hematopoietic-derived Sema4A is dispensable for myHSC and balHSC function.

In order to elucidate the role of bone marrow microenvironment-derived Sema4A, we non-competitively transplanted lethally irradiated WT and Sema4AKO recipients with a radioprotective dose of myHSC and balHSC from WT mice (Fig. 4C). Strikingly, we observed a ~50% post-transplant mortality in Sema4AKO recipients of myHSC (Fig. 4D, left panel). The surviving animals from this group displayed marked anemia and neutrophilia, which resembled blood count abnormalities in aged Sema4AKO mice (Fig. 4D, right panel and Fig. S4H). In contrast, no survival difference was observed in the recipients of balHSC, which showed only mild blood counts changes (Fig. 4E and Fig. S4I). This experiment revealed that Sema4A from the host hematopoietic niche is critical for myHSC self-renewal under stress but plays no significant role in balHSC regeneration.

Among the subsets which make up the bone marrow niche, Sema4A is expressed by endothelial and osteoprogenitor cells (Fig. S1A and Fig. S1B). In order to refine their

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.12.476097; this version posted January 12, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

physiological relevance as cellular sources of Sema4A in the niche, we conditionally deleted Sema4A from each of the two cell types by crossing “floxed” Sema4AKO mice to either VECad-CreERT2 or Osx-Cre animals. Steady-state analysis of Sema4A$^{fl/fl}$ VE-CadCre ERT2(+) and Sema4A$^{fl/fl}$ Osx-Cre (+) mice revealed no difference in peripheral blood counts, bone marrow cellularity, cell cycle and frequency of HSPC and mature cells, as compared to Sema4A$^{fl/fl}$ VE-CadCre ERT2(-) (Sorensen et al., 2009) and Sema4A$^{fl/wt}$ Osx-Cre (+) (Rodda and McMahon, 2006) controls, respectively (Fig. S4J-M and Suppl. Table 1). Transplantation of WT myHSC and balHSC into lethally irradiated donors of the above genotypes showed that both endothelial- and osteoprogenitor-specific deletion of Sema4A only partially recapitulated the effect of a complete Sema4A absence in the host. Specifically, we observed increased mortality in Sema4A$^{fl/fl}$ Osx-Cre (+) recipients of myHSC, but the difference was not statistically significant (Fig. S4N-P). In Sema4A$^{fl/fl}$ VE-CadCre ERT2(+) recipients of myHSC, we detected a slight reduction in hematocrit and no impact on survival (Fig. S4Q-S). Taken together, these studies suggest that a cumulative production by osteoprogenitors, endothelial cells and likely other cellular source(s) may be responsible for the full functional effect of microenvironment-derived Sema4A on myHSC.

Having demonstrated that a complete absence of Sema4A in the host is critical for myHSC engraftment (Fig. 4D), we asked how it may affect early homing, expansion and motility of transplanted myHSC in real time. To this end, we isolated myHSC and balHSC from WT mice and fluorescently labeled them with DiD. We then transferred equal numbers of these cells into lethally irradiated WT and Sema4AKO recipients and performed intravital time-lapse two-photon microscopy of the calvarial bone marrow (Christodoulou et al., 2020). We recorded 3D z-stacks and time-lapse movies 15-20 hours after the cell injection. Single cells and clusters (defined as two or more cells whose cell-to-cell edge are within 15 μm) were detected throughout the calvarial bone marrow in all mice (Fig. S4T). Notably, we observed a ~3 times

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.12.476097; this version posted January 12, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

higher number of transplanted cells in Sema4AKO recipients of myHSC compared to WT controls (mean ~106 cells vs. 34 cells, [p-value = 0.0295]) whereas cell number in WT/Sema4A recipients of balHSC were not significantly different (mean ~68 cells vs. 41 cells, respectively [p-value = 0.2793]) (Fig. 4F and S4U). These results indicate that the absence of Sema4A in the host leads to excessive myHSC expansion but is inconsequential for balHSC. As further evidence for this, we found a similar 3-fold increase in the number of cell clusters in the Sema4AKO recipients of myHSC as compared to WT controls (mean ~21 vs. 7 clusters [p-value = 0.0219] whereas the trend in the balHSC recipients was much weaker (~15 vs. 8 clusters [p-value = 0.2591]) (Fig. S4V).

Next, we analyzed the effect of Sema4A on myHSC and balHSC localization by measuring the 3D distance to the endosteal surface, an established location of post-transplant HSC niche (Lo Celso et al., 2009). Importantly, we observed that in Sema4AKO recipients of myHSC, transplanted cells were found nearly 2x farther from the endosteum compared to WT controls (mean ~8.2 μm vs. 4.9 μm, respectively [p-value = 0.0024]) whereas this difference was smaller and not statistically significant in the Sema4AKO/WT balHSC recipients (mean ~5.9 μm vs. 4.0 μm, respectively [p-value = 0.0674]) (Fig. 4G). These results suggest the in the absence of host Sema4A, myHSC homed away from the niche, whereas localization of balHSC was only marginally altered.

Recent intravital time-lapse microscopy studies revealed that upon proliferative challenge, some HSC within the bone marrow niche become motile (Christodoulou *et al.*, 2020; Upadhaya et al., 2020), indicating that motility may reflect HSC activation state. In order to investigate if motility of transplanted myHSC and balHSC is altered in the absence of Sema4A, we performed time-lapse microscopy for 1.5 hrs. We found that balHSC displayed limited motility (defined as <5 μm movement of the cell centroid over the imaging period) regardless of the host genotype (data not shown). In contrast, a small fraction (~4.7% of total) of myHSC transplanted into

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.12.476097; this version posted January 12, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Sema4AKO mice exhibited highly motile behavior Sema4A (Fig. 4H, Fig.S4V-X, Suppl. Movie). In sum, our intravital imaging data suggests that host absence of Sema4A leads to myHSC hyperactivation, excessive proliferation and mis-localization, which cumulatively may contribute to the loss of self-renewal and engraftment failure in Sema4AKO recipients of myHSC. Intriguingly, post-transplant behavior of balHSC was relatively unaffected, indicating that the two HSC subsets may have fundamentally different requirements for engraftment, including specific dependence of myHSC on Sema4A.

DISCUSSION

Our study provides substantial experimental support for the concept that functionally diverse subsets of somatic stem cells are controlled by distinct non cell-autonomous signals. Prior studies have indicated that within the HSC pool, myHSC and balHSC display differential sensitivity to soluble factors, such as TGF-beta, RANTES, CXCL2 and histamine (Challen *et al.*, 2010; Chen et al., 2017; Ergen et al., 2012; Pinho et al., 2018). However, the impact of these molecules on myHSC longevity and interaction with the bone marrow microenvironment has not been investigated in detail.

In the current study, we identify Sema4A as an indispensable and specific regulator of myHSC quiescence and self-renewal. Semaphorins and plexins are large protein families (Alto and Terman, 2017) whose role in regulation of adult stem cell quiescence and self-renewal is not known. We demonstrate that the absence of Sema4A leads to myHSC over-proliferation and hyperactivation following acute inflammatory insult, which correlates with a dramatic loss of regenerative function with age, likely due to the loss of protection from detrimental effects of inflammatory signaling over the animal’s lifetime (Kaschutnig et al., 2015). Notably, WT myHSC are preferentially activated by inflammation (Mann *et al.*, 2018; Matatall *et al.*, 2014; Mitroulis *et*

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.12.476097; this version posted January 12, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

al., 2018) and become vulnerable to damage, underscoring a physiological need for a dedicated protective signal, such as Sema4A.

Excessive myeloid expansion, as observed in the aged Sema4AKO model, is the cardinal feature of human hematopoietic aging (Pang *et al.*, 2011) and clonal hematopoiesis of indeterminate significance (CHIP) - a common condition which carries a significant risk of progression to myeloid malignancy over time but lacks effective therapeutic intervention (Jaiswal and Ebert, 2019). Our findings raise a possibility that pharmacological augmentation of Sema4A/PlxnD1 signaling may serve as a potential strategy to constrain proliferation of myHSC at the top of expanding myeloid-biased clones, thereby preventing aging-associated HSC dysfunction and reducing the risk of malignant transformation.

Our results underscore the importance of niche-derived signals in life-long maintenance of tissue stem cell hierarchy, which is topped by myHSC in the hematopoietic system. Given that stem cell hierarchies underlie functional organization of other tissues (Altshuler *et al.*, 2021; Farrelly *et al.*, 2021; Hsu *et al.*, 2011; Rompolas *et al.*, 2013; Sachewsky *et al.*, 2019; Scaramozza *et al.*, 2019), the data presented here provide justification for broader efforts to identify subset-specific stem cell regulators, which may lead to development of more precise and effective pro-regenerative therapies.

LIMITATIONS OF THE STUDY

Recent studies revealed that CD150-expressing phenotypic HSC contain a heterogenous mixture of myeloid-restricted progenitors which vary in their capacity for long-term reconstitution and degree of commitment to myeloid, erythroid and platelet lineage (Yamamoto et al., 2018). We recognize that our experiments are unable to resolve the precise identity of a myeloid-restricted subset(s) which is regulated by Sema4A beyond the CD150^{high} HSC fraction. Future

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.12.476097; this version posted January 12, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

studies, including single cell transplantation experiments, will be required to address this question.

Although our data demonstrates that Sema4A suppresses myHSC proliferation during stress, the downstream mediators of this effect are not known. Therefore, the molecular consequences of Sema4A binding to PlxnD1 (and potentially other Sema4A receptors whose functional relevance has not be ruled out by the current study) will need to be further investigated.

ACKNOWLEDGMENTS

Work in the Silberstein Laboratory is supported by the NIH/NHLBI (R01 HL148189), Translational Research Program grant through Leukemia and Lymphoma Society, Washington State CARE Fund and New Development Fund from Fred Hutchinson Cancer Research Center. The work in the Göttgens laboratory was funded in part by the Wellcome Trust [203151/Z/16/Z] and the UKRI Medical Research Council [MC_PC_17230]. For the purpose of open access, the author has applied a CC BY public copyright license to any Author Accepted Manuscript version arising from this submission. The work in the Maes laboratory is supported by grants from the Research Foundation Flanders (FWO) and KU Leuven (C1). The schematics were created with BioRender.com.

We would like to thank Elena Nefyodova (KU Leuven) for expert technical help, Michael Li, Zoe Derauf, Ben Janoschek and Kora Krumm for mouse colony support, and the staff of Flow Cytometry, Genomics and Comparative Medicine cores facilities at Fred Hutchinson Cancer Research Center. We are also grateful to Peter Klein, University of Pennsylvania, and Adam Wilkinson, University of Oxford, for helpful feedback and advice.

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.12.476097; this version posted January 12, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

AUTHOR CONTRIBUTIONS

DT and SZ designed and performed experiments, analyzed and interpreted data, and wrote the manuscript. EM performed computational analyses of sequencing data and contributed to data interpretation. EIC performed experiments, computational analyses of sequencing data and contributed to experimental design and data interpretation. AS data supervised all the computational analyses of sequencing data and contributed to data interpretation. AP supported mouse breeding and performed experiments. NT, CB and DG performed intravital imaging experiments under the supervision of JAS. SR, HPK designed and performed human HSC experiments. NKW and SJK performed single cell RNA-Seq experiments under the supervision of BG. CN, MM, CM and PK contributed data. SR and NC performed experiments with aged mice. TW and AK contributed reagents. EMP and DTS contributed to experimental design and data discussion. L.S. conceived and supervised the project, performed experiments, analyzed data and wrote the manuscript.

DECLARATION OF INTERESTS

SR: Ensoma Inc.: Consultancy; 47 Inc.: Consultancy. HPK: Ensoma Inc.: Consultancy, Current holder of individual stocks in a privately-held company; Homology Medicines: Consultancy; VOR Biopharma: Consultancy. DTS: Fate Therapeutics: Current holder of individual stocks in a privately-held company; Editas Medicines: Current holder of individual stocks in a privately-held company, Membership on an entity's Board of Directors or advisory committees; Clear Creek Bio: Current holder of individual stocks in a privately-held company, Membership on an entity's Board of Directors or advisory committees; Dainippon Sumitomo Pharma: Other: sponsored research; FOG Pharma: Consultancy; Agios Pharmaceuticals: Current holder of individual stocks in a privately-held company, Membership on an entity's Board of Directors or advisory

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.12.476097; this version posted January 12, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

committees; Garuda Therapeutics: Current holder of individual stocks in a privately-held company, Membership on an entity's Board of Directors or advisory committees; VCanBio: Consultancy; Inzen Therapeutics: Membership on an entity's Board of Directors or advisory committees; LifeVaultBio: Current holder of individual stocks in a privately-held company, Membership on an entity's Board of Directors or advisory committees; Magenta Therapeutics: Current holder of individual stocks in a privately-held company, Membership on an entity's Board of Directors or advisory committees. Other authors have nothing to disclose.

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.12.476097; this version posted January 12, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

REFERENCES

Alto, L.T., and Terman, J.R. (2017). Semaphorins and their Signaling Mechanisms. Methods Mol Biol *1493*, 1-25. 10.1007/978-1-4939-6448-2_1.

Altshuler, A., Amitai-Lange, A., Tarazi, N., Dey, S., Strinkovsky, L., Hadad-Porat, S., Bhattacharya, S., Nasser, W., Imeri, J., Ben-David, G., et al. (2021). Discrete limbal epithelial stem cell populations mediate corneal homeostasis and wound healing. Cell Stem Cell *28*, 1248-1261 e1248. 10.1016/j.stem.2021.04.003.

Baccin, C., Al-Sabah, J., Velten, L., Helbling, P.M., Grunschlager, F., Hernandez-Malmierca, P., Nombela-Arrieta, C., Steinmetz, L.M., Trumpp, A., and Haas, S. (2020). Combined single-cell and spatial transcriptomics reveal the molecular, cellular and spatial bone marrow niche organization. Nat Cell Biol *22*, 38-48. 10.1038/s41556-019-0439-6.

Becker-Herman, S., Rozenberg, M., Hillel-Karniel, C., Gil-Yarom, N., Kramer, M.P., Barak, A., Sever, L., David, K., Radomir, L., Lewinsky, H., et al. (2021). CD74 is a regulator of hematopoietic stem cell maintenance. PLoS Biol *19*, e3001121. 10.1371/journal.pbio.3001121.

Beerman, I., Bhattacharya, D., Zandi, S., Sigvardsson, M., Weissman, I.L., Bryder, D., and Rossi, D.J. (2010). Functionally distinct hematopoietic stem cells modulate hematopoietic lineage potential during aging by a mechanism of clonal expansion. Proc Natl Acad Sci U S A *107*, 5465-5470. 10.1073/pnas.1000834107.

Bernitz, J.M., Kim, H.S., MacArthur, B., Sieburg, H., and Moore, K. (2016). Hematopoietic Stem Cells Count and Remember Self-Renewal Divisions. Cell *167*, 1296-1309 e1210. 10.1016/j.cell.2016.10.022.

Cabezas-Wallscheid, N., Buettner, F., Sommerkamp, P., Klimmeck, D., Ladel, L., Thalheimer, F.B., Pastor-Flores, D., Roma, L.P., Renders, S., Zeisberger, P., et al. (2017). Vitamin A-Retinoic Acid Signaling Regulates Hematopoietic Stem Cell Dormancy. Cell *169*, 807-823 e819. 10.1016/j.cell.2017.04.018.

Cabezas-Wallscheid, N., Klimmeck, D., Hansson, J., Lipka, D.B., Reyes, A., Wang, Q., Weichenhan, D., Lier, A., von Paleske, L., Renders, S., et al. (2014). Identification of regulatory networks in HSCs and their immediate progeny via integrated proteome, transcriptome, and DNA methylome analysis. Cell Stem Cell *15*, 507-522. 10.1016/j.stem.2014.07.005.

Calvanese, V., Nguyen, A.T., Bolan, T.J., Vavilina, A., Su, T., Lee, L.K., Wang, Y., Lay, F.D., Magnusson, M., Crooks, G.M., et al. (2019). MLLT3 governs human haematopoietic stem-cell self-renewal and engraftment. Nature *576*, 281-286. 10.1038/s41586-019-1790-2.

Carrelha, J., Meng, Y., Kettyle, L.M., Luis, T.C., Norfo, R., Alcolea, V., Boukarabila, H., Grasso, F., Gambardella, A., Grover, A., et al. (2018). Hierarchically related lineage-restricted fates of multipotent haematopoietic stem cells. Nature *554*, 106-111. 10.1038/nature25455.

Challen, G.A., Boles, N.C., Chambers, S.M., and Goodell, M.A. (2010). Distinct hematopoietic stem cell subtypes are differentially regulated by TGF-beta1. Cell Stem Cell *6*, 265-278. 10.1016/j.stem.2010.02.002.

Chen, X., Deng, H., Churchill, M.J., Luchsinger, L.L., Du, X., Chu, T.H., Friedman, R.A., Middelhoff, M., Ding, H., Tailor, Y.H., et al. (2017). Bone Marrow Myeloid Cells Regulate Myeloid-Biased Hematopoietic Stem Cells via a Histamine-Dependent Feedback Loop. Cell Stem Cell *21*, 747-760 e747. 10.1016/j.stem.2017.11.003.

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.12.476097; this version posted January 12, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Christodoulou, C., Spencer, J.A., Yeh, S.A., Turcotte, R., Kokkaliaris, K.D., Panero, R., Ramos, A., Guo, G., Seyedhassantehrani, N., Esipova, T.V., et al. (2020). Live-animal imaging of native haematopoietic stem and progenitor cells. Nature *578*, 278-283. 10.1038/s41586-020-1971-z.

Ergen, A.V., Boles, N.C., and Goodell, M.A. (2012). Rantes/Ccl5 influences hematopoietic stem cell subtypes and causes myeloid skewing. Blood *119*, 2500-2509. 10.1182/blood-2011-11-391730.

Esplin, B.L., Shimazu, T., Welner, R.S., Garrett, K.P., Nie, L., Zhang, Q., Humphrey, M.B., Yang, Q., Borghesi, L.A., and Kincade, P.W. (2011). Chronic exposure to a TLR ligand injures hematopoietic stem cells. J Immunol *186*, 5367-5375. 10.4049/jimmunol.1003438.

Fang, S., Chen, S., Nurmi, H., Leppanen, V.M., Jeltsch, M., Scadden, D., Silberstein, L., Mikkola, H., and Alitalo, K. (2020). VEGF-C protects the integrity of the bone marrow perivascular niche in mice. Blood *136*, 1871-1883. 10.1182/blood.2020005699.

Fard, D., and Tamagnone, L. (2021). Semaphorins in health and disease. Cytokine Growth Factor Rev *57*, 55-63. 10.1016/j.cytogfr.2020.05.006.

Farrelly, O., Suzuki-Horiuchi, Y., Brewster, M., Kuri, P., Huang, S., Rice, G., Bae, H., Xu, J., Dentchev, T., Lee, V., and Rompolas, P. (2021). Two-photon live imaging of single corneal stem cells reveals compartmentalized organization of the limbal niche. Cell Stem Cell *28*, 1233-1247 e1234. 10.1016/j.stem.2021.02.022.

Goncalves, K.A., Silberstein, L., Li, S., Severe, N., Hu, M.G., Yang, H., Scadden, D.T., and Hu, G.F. (2016). Angiogenin Promotes Hematopoietic Regeneration by Dichotomously Regulating Quiescence of Stem and Progenitor Cells. Cell *166*, 894-906. 10.1016/j.cell.2016.06.042.

Gong, S., Zheng, C., Doughty, M.L., Losos, K., Didkovsky, N., Schambra, U.B., Nowak, N.J., Joyner, A., Leblanc, G., Hatten, M.E., and Heintz, N. (2003). A gene expression atlas of the central nervous system based on bacterial artificial chromosomes. Nature *425*, 917-925. 10.1038/nature02033.

Grover, A., Sanjuan-Pla, A., Thongjuea, S., Carrelha, J., Giustacchini, A., Gambardella, A., Macaulay, I., Mancini, E., Luis, T.C., Mead, A., et al. (2016). Single-cell RNA sequencing reveals molecular and functional platelet bias of aged haematopoietic stem cells. Nat Commun *7*, 11075. 10.1038/ncomms11075.

Haghverdi, L., Buttner, M., Wolf, F.A., Buettner, F., and Theis, F.J. (2016). Diffusion pseudotime robustly reconstructs lineage branching. Nat Methods *13*, 845-848. 10.1038/nmeth.3971.

Hirche, C., Frenz, T., Haas, S.F., Doring, M., Borst, K., Tegtmeyer, P.K., Brizic, I., Jordan, S., Keyser, K., Chhatbar, C., et al. (2017). Systemic Virus Infections Differentially Modulate Cell Cycle State and Functionality of Long-Term Hematopoietic Stem Cells In Vivo. Cell Rep *19*, 2345-2356. 10.1016/j.celrep.2017.05.063.

Hsu, Y.C., Pasolli, H.A., and Fuchs, E. (2011). Dynamics between stem cells, niche, and progeny in the hair follicle. Cell *144*, 92-105. 10.1016/j.cell.2010.11.049.

Jaiswal, S., and Ebert, B.L. (2019). Clonal hematopoiesis in human aging and disease. Science *366*. 10.1126/science.aan4673.

Kaschutnig, P., Bogeska, R., Walter, D., Lier, A., Huntscha, S., and Milsom, M.D. (2015). The Fanconi anemia pathway is required for efficient repair of stress-induced DNA damage in haematopoietic stem cells. Cell Cycle *14*, 2734-2742. 10.1080/15384101.2015.1068474.

Kent, D.G., Copley, M.R., Benz, C., Wohrer, S., Dykstra, B.J., Ma, E., Cheyne, J., Zhao, Y., Bowie, M.B., Zhao, Y., et al. (2009). Prospective isolation and molecular characterization of hematopoietic stem cells with durable self-renewal potential. Blood *113*, 6342-6350. 10.1182/blood-2008-12-192054.

Kent, D.G., Dykstra, B.J., Cheyne, J., Ma, E., and Eaves, C.J. (2008). Steel factor coordinately regulates the molecular signature and biologic function of hematopoietic stem cells. Blood *112*, 560-567. 10.1182/blood-2007-10-117820.

Kolodkin, A.L., Matthes, D.J., and Goodman, C.S. (1993). The semaphorin genes encode a family of transmembrane and secreted growth cone guidance molecules. Cell *75*, 1389-1399. 10.1016/0092-8674(93)90625-z.

Kovtonyuk, L.V., Fritsch, K., Feng, X., Manz, M.G., and Takizawa, H. (2016). Inflamm-Aging of Hematopoiesis, Hematopoietic Stem Cells, and the Bone Marrow Microenvironment. Front Immunol *7*, 502. 10.3389/fimmu.2016.00502.

Kuhn, R., Schwenk, F., Aguet, M., and Rajewsky, K. (1995). Inducible gene targeting in mice. Science *269*, 1427-1429. 10.1126/science.7660125.

Kumanogoh, A., Shikina, T., Suzuki, K., Uematsu, S., Yukawa, K., Kashiwamura, S., Tsutsui, H., Yamamoto, M., Takamatsu, H., Ko-Mitamura, E.P., et al. (2005). Nonredundant roles of Sema4A in the immune system: defective T cell priming and Th1/Th2 regulation in Sema4A-deficient mice. Immunity *22*, 305-316. 10.1016/j.immuni.2005.01.014.

Laurenti, E., Frelin, C., Xie, S., Ferrari, R., Dunant, C.F., Zandi, S., Neumann, A., Plumb, I., Doulatov, S., Chen, J., et al. (2015). CDK6 levels regulate quiescence exit in human hematopoietic stem cells. Cell Stem Cell *16*, 302-313. 10.1016/j.stem.2015.01.017.

Lo Celso, C., Fleming, H.E., Wu, J.W., Zhao, C.X., Miake-Lye, S., Fujisaki, J., Cote, D., Rowe, D.W., Lin, C.P., and Scadden, D.T. (2009). Live-animal tracking of individual haematopoietic stem/progenitor cells in their niche. Nature *457*, 92-96. 10.1038/nature07434.

Mann, M., Mehta, A., de Boer, C.G., Kowalczyk, M.S., Lee, K., Haldeman, P., Rogel, N., Knecht, A.R., Farouq, D., Regev, A., and Baltimore, D. (2018). Heterogeneous Responses of Hematopoietic Stem Cells to Inflammatory Stimuli Are Altered with Age. Cell Rep *25*, 2992-3005 e2995. 10.1016/j.celrep.2018.11.056.

Matatall, K.A., Shen, C.C., Challen, G.A., and King, K.Y. (2014). Type II interferon promotes differentiation of myeloid-biased hematopoietic stem cells. Stem Cells *32*, 3023-3030. 10.1002/stem.1799.

Mitroulis, I., Ruppova, K., Wang, B., Chen, L.S., Grzybek, M., Grinenko, T., Eugster, A., Troullinaki, M., Palladini, A., Kourtzelis, I., et al. (2018). Modulation of Myelopoiesis Progenitors Is an Integral Component of Trained Immunity. Cell *172*, 147-161 e112. 10.1016/j.cell.2017.11.034.

Montecino-Rodriguez, E., Kong, Y., Casero, D., Rouault, A., Dorshkind, K., and Pioli, P.D. (2019). Lymphoid-Biased Hematopoietic Stem Cells Are Maintained with Age and Efficiently Generate Lymphoid Progeny. Stem Cell Reports *12*, 584-596. 10.1016/j.stemcr.2019.01.016.

Morcos, M.N.F., Schoedel, K.B., Hoppe, A., Behrendt, R., Basak, O., Clevers, H.C., Roers, A., and Gerbaulet, A. (2017). SCA-1 Expression Level Identifies Quiescent Hematopoietic Stem and Progenitor Cells. Stem Cell Reports *8*, 1472-1478. 10.1016/j.stemcr.2017.04.012.

Morita, Y., Ema, H., and Nakauchi, H. (2010). Heterogeneity and hierarchy within the most primitive hematopoietic stem cell compartment. J Exp Med *207*, 1173-1182. 10.1084/jem.20091318.

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.12.476097; this version posted January 12, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Nestorowa, S., Hamey, F.K., Pijuan Sala, B., Diamanti, E., Shepherd, M., Laurenti, E., Wilson, N.K., Kent, D.G., and Gottgens, B. (2016). A single-cell resolution map of mouse hematopoietic stem and progenitor cell differentiation. Blood *128*, e20-31. 10.1182/blood-2016-05-716480.
Pang, W.W., Price, E.A., Sahoo, D., Beerman, I., Maloney, W.J., Rossi, D.J., Schrier, S.L., and Weissman, I.L. (2011). Human bone marrow hematopoietic stem cells are increased in frequency and myeloid-biased with age. Proc Natl Acad Sci U S A *108*, 20012-20017. 10.1073/pnas.1116110108.
Picelli, S., Faridani, O.R., Bjorklund, A.K., Winberg, G., Sagasser, S., and Sandberg, R. (2014). Full-length RNA-seq from single cells using Smart-seq2. Nat Protoc *9*, 171-181. 10.1038/nprot.2014.006.
Pinho, S., Marchand, T., Yang, E., Wei, Q., Nerlov, C., and Frenette, P.S. (2018). Lineage-Biased Hematopoietic Stem Cells Are Regulated by Distinct Niches. Dev Cell *44*, 634-641 e634. 10.1016/j.devcel.2018.01.016.
Rodda, S.J., and McMahon, A.P. (2006). Distinct roles for Hedgehog and canonical Wnt signaling in specification, differentiation and maintenance of osteoblast progenitors. Development *133*, 3231-3244. 10.1242/dev.02480.
Rodriguez-Fraticelli, A.E., Weinreb, C., Wang, S.W., Migueles, R.P., Jankovic, M., Usart, M., Klein, A.M., Lowell, S., and Camargo, F.D. (2020). Single-cell lineage tracing unveils a role for TCF15 in haematopoiesis. Nature *583*, 585-589. 10.1038/s41586-020-2503-6.
Rompolas, P., Mesa, K.R., and Greco, V. (2013). Spatial organization within a niche as a determinant of stem-cell fate. Nature *502*, 513-518. 10.1038/nature12602.
Sachewsky, N., Xu, W., Fuehrmann, T., van der Kooy, D., and Morshead, C.M. (2019). Lineage tracing reveals the hierarchical relationship between neural stem cell populations in the mouse forebrain. Sci Rep *9*, 17730. 10.1038/s41598-019-54143-9.
Sanjuan-Pla, A., Macaulay, I.C., Jensen, C.T., Woll, P.S., Luis, T.C., Mead, A., Moore, S., Carella, C., Matsuoka, S., Bouriez Jones, T., et al. (2013). Platelet-biased stem cells reside at the apex of the haematopoietic stem-cell hierarchy. Nature *502*, 232-236. 10.1038/nature12495.
Scaramozza, A., Park, D., Kollu, S., Beerman, I., Sun, X., Rossi, D.J., Lin, C.P., Scadden, D.T., Crist, C., and Brack, A.S. (2019). Lineage Tracing Reveals a Subset of Reserve Muscle Stem Cells Capable of Clonal Expansion under Stress. Cell Stem Cell *24*, 944-957 e945. 10.1016/j.stem.2019.03.020.
Scialdone, A., Natarajan, K.N., Saraiva, L.R., Proserpio, V., Teichmann, S.A., Stegle, O., Marioni, J.C., and Buettner, F. (2015). Computational assignment of cell-cycle stage from single-cell transcriptome data. Methods *85*, 54-61. 10.1016/j.ymeth.2015.06.021.
Serini, G., Valdembri, D., Zanivan, S., Morterra, G., Burkhardt, C., Caccavari, F., Zammataro, L., Primo, L., Tamagnone, L., Logan, M., et al. (2003). Class 3 semaphorins control vascular morphogenesis by inhibiting integrin function. Nature *424*, 391-397. 10.1038/nature01784.
Silberstein, L., Goncalves, K.A., Kharchenko, P.V., Turcotte, R., Kfoury, Y., Mercier, F., Baryawno, N., Severe, N., Bachand, J., Spencer, J.A., et al. (2016). Proximity-Based Differential Single-Cell Analysis of the Niche to Identify Stem/Progenitor Cell Regulators. Cell Stem Cell *19*, 530-543. 10.1016/j.stem.2016.07.004.
Sorensen, I., Adams, R.H., and Gossler, A. (2009). DLL1-mediated Notch activation regulates endothelial identity in mouse fetal arteries. Blood *113*, 5680-5688. 10.1182/blood-2008-08-174508.

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.12.476097; this version posted January 12, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Takizawa, H., Regoes, R.R., Boddupalli, C.S., Bonhoeffer, S., and Manz, M.G. (2011). Dynamic variation in cycling of hematopoietic stem cells in steady state and inflammation. J Exp Med *208*, 273-284. 10.1084/jem.20101643.

Toyofuku, T., Yabuki, M., Kamei, J., Kamei, M., Makino, N., Kumanogoh, A., and Hori, M. (2007). Semaphorin-4A, an activator for T-cell-mediated immunity, suppresses angiogenesis via Plexin-D1. EMBO J *26*, 1373-1384. 10.1038/sj.emboj.7601589.

Upadhaya, S., Krichevsky, O., Akhmetzyanova, I., Sawai, C.M., Fooksman, D.R., and Reizis, B. (2020). Intravital Imaging Reveals Motility of Adult Hematopoietic Stem Cells in the Bone Marrow Niche. Cell Stem Cell *27*, 336-345 e334. 10.1016/j.stem.2020.06.003.

Valletta, S., Thomas, A., Meng, Y., Ren, X., Drissen, R., Sengul, H., Di Genua, C., and Nerlov, C. (2020). Micro-environmental sensing by bone marrow stroma identifies IL-6 and TGFbeta1 as regulators of hematopoietic ageing. Nat Commun *11*, 4075. 10.1038/s41467-020-17942-7.

Walter, D., Lier, A., Geiselhart, A., Thalheimer, F.B., Huntscha, S., Sobotta, M.C., Moehrle, B., Brocks, D., Bayindir, I., Kaschutnig, P., et al. (2015). Exit from dormancy provokes DNA-damage-induced attrition in haematopoietic stem cells. Nature *520*, 549-552. 10.1038/nature14131.

Yamamoto, R., Wilkinson, A.C., Ooehara, J., Lan, X., Lai, C.Y., Nakauchi, Y., Pritchard, J.K., and Nakauchi, H. (2018). Large-Scale Clonal Analysis Resolves Aging of the Mouse Hematopoietic Stem Cell Compartment. Cell Stem Cell *22*, 600-607 e604. 10.1016/j.stem.2018.03.013.

Yu, W., Goncalves, K.A., Li, S., Kishikawa, H., Sun, G., Yang, H., Vanli, N., Wu, Y., Jiang, Y., Hu, M.G., et al. (2017). Plexin-B2 Mediates Physiologic and Pathologic Functions of Angiogenin. Cell *171*, 849-864 e825. 10.1016/j.cell.2017.10.005.

Zhang, Y., Singh, M.K., Degenhardt, K.R., Lu, M.M., Bennett, J., Yoshida, Y., and Epstein, J.A. (2009). Tie2Cre-mediated inactivation of plexinD1 results in congenital heart, vascular and skeletal defects. Dev Biol *325*, 82-93. 10.1016/j.ydbio.2008.09.031.

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.12.476097; this version posted January 12, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

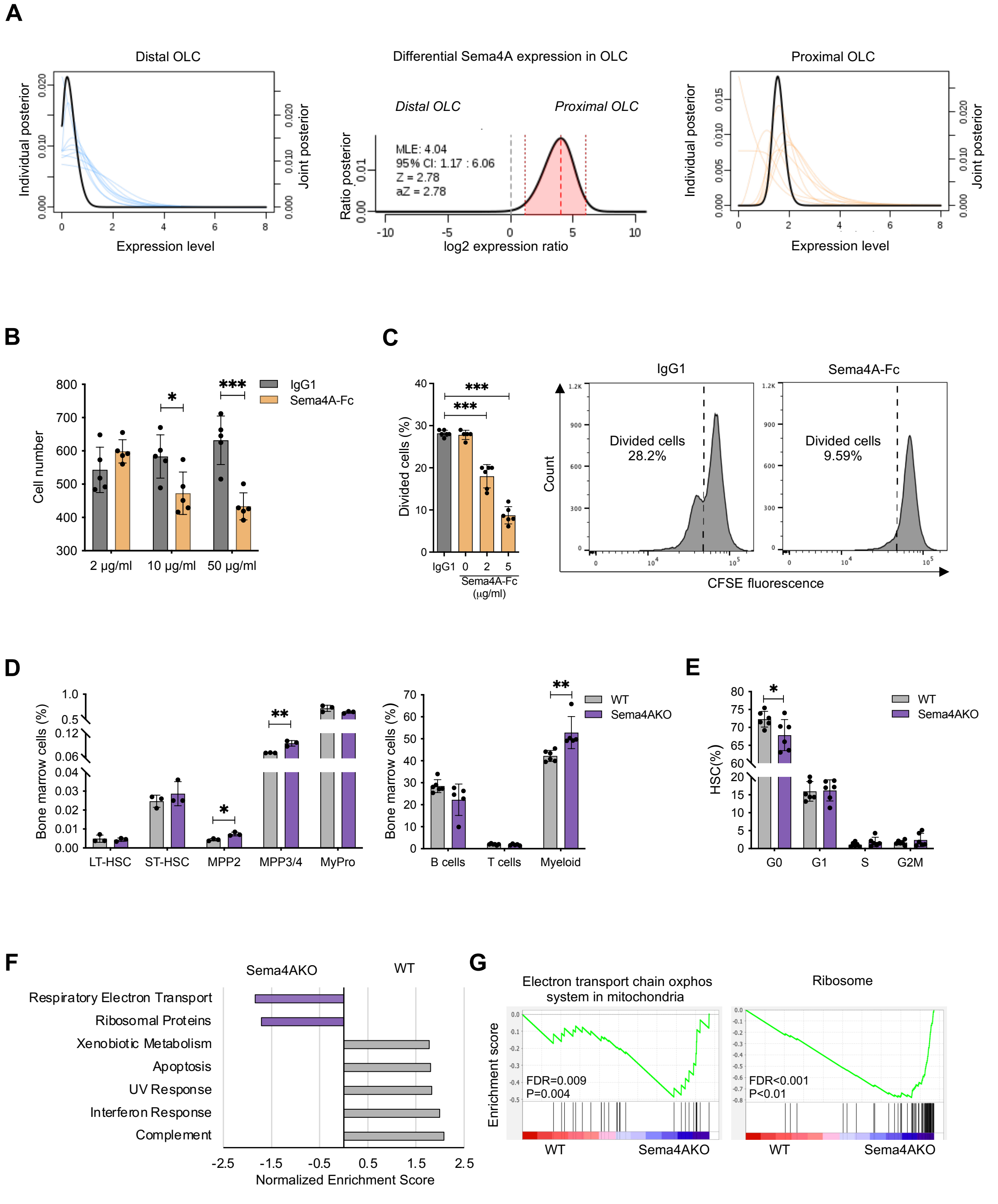


Figure 1

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.12.476097; this version posted January 12, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

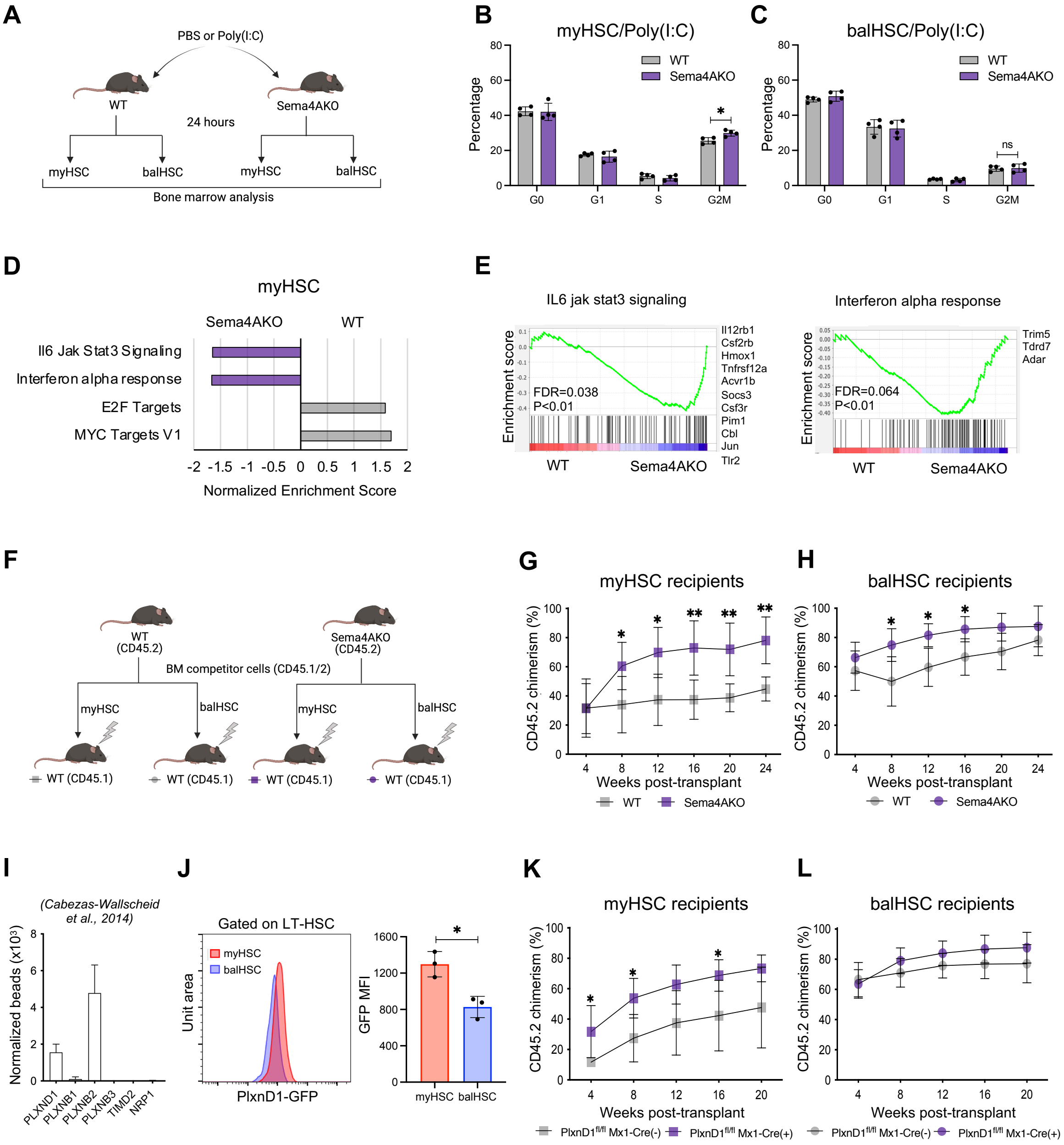


Figure 2

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.12.476097; this version posted January 12, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

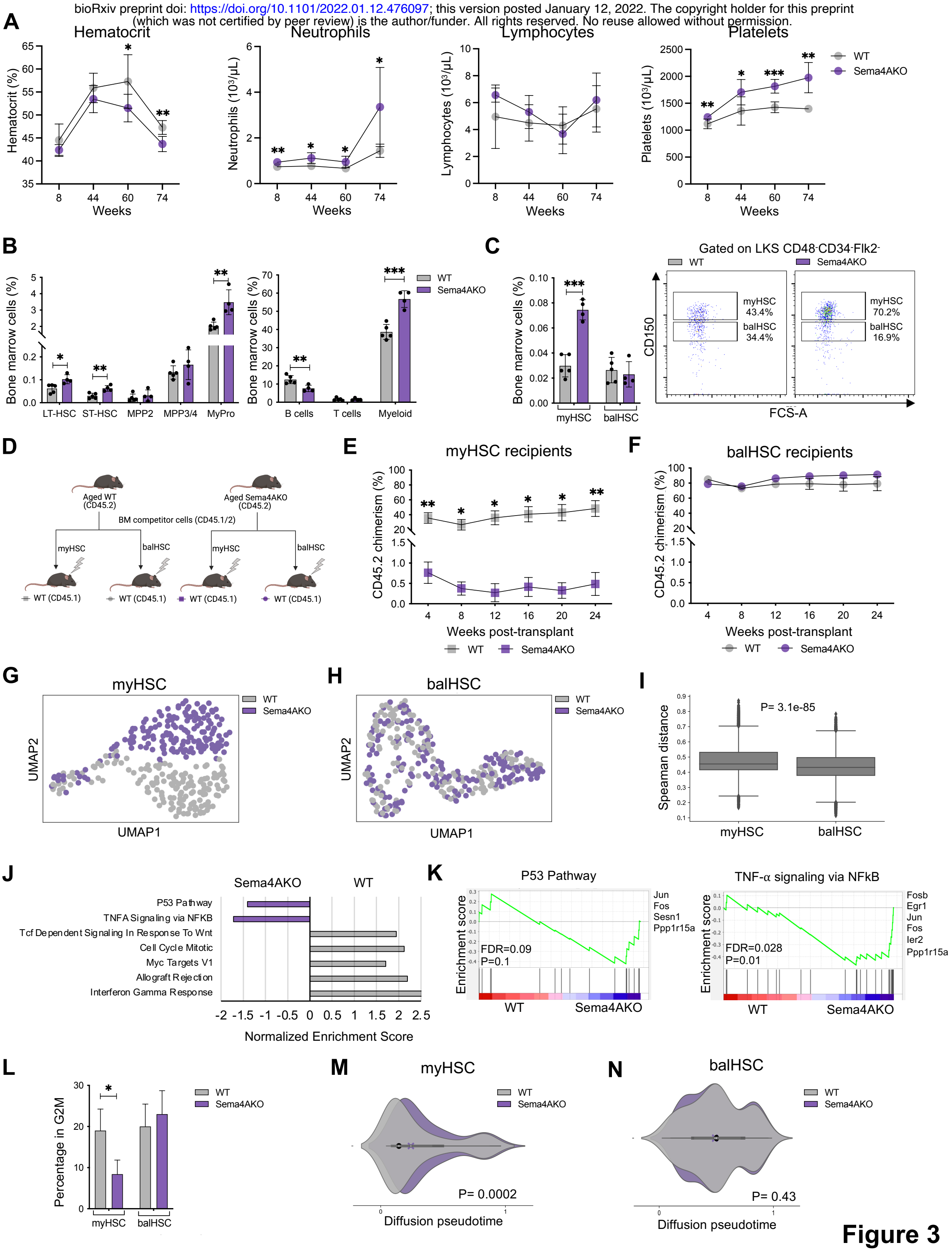


Figure 3

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.12.476097; this version posted January 12, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

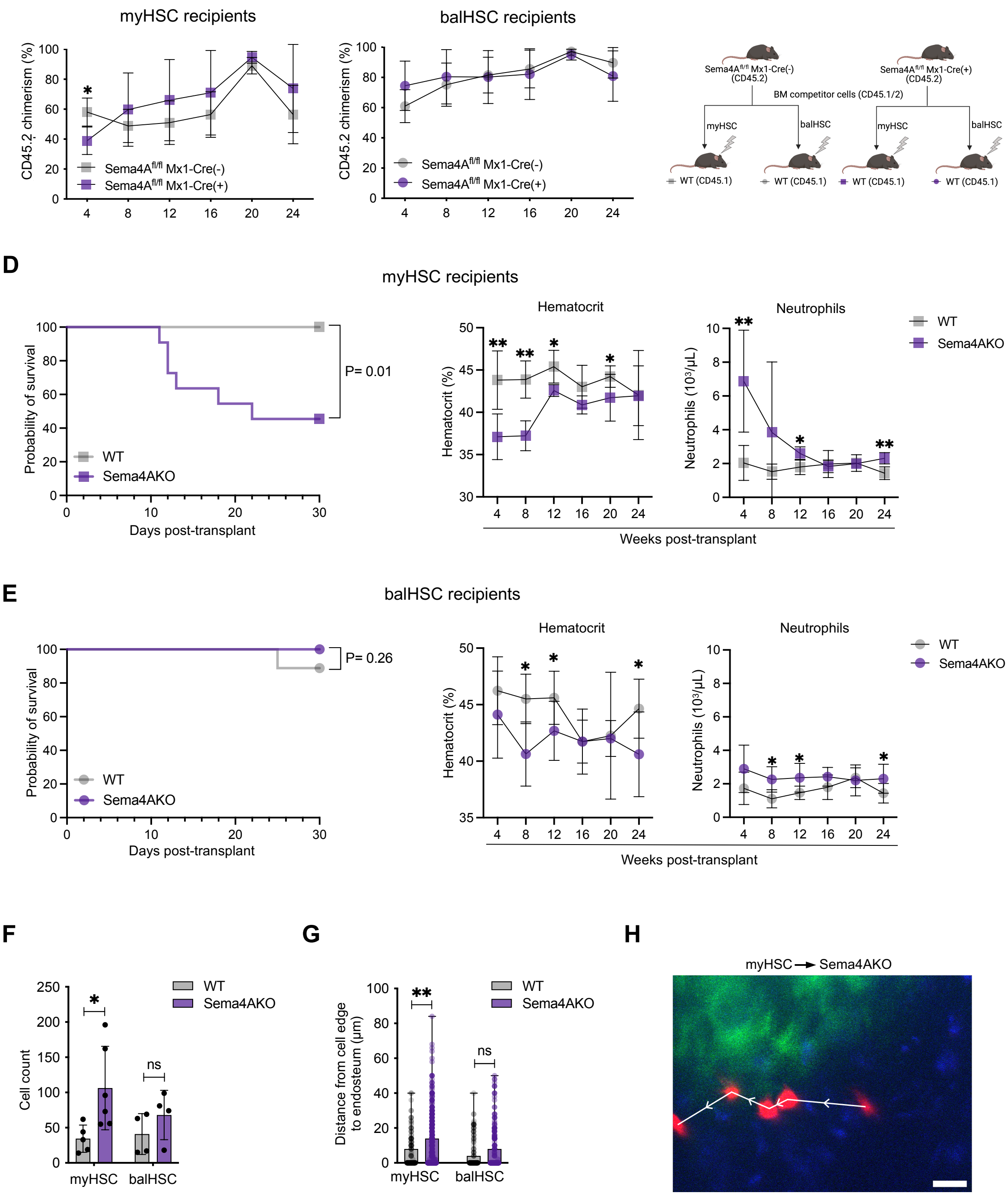


Figure 4

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.12.476097; this version posted January 12, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

MAIN FIGURE LEGENDS

Figure 1. Sema4A regulates quiescence of mouse and human hematopoietic stem/progenitor cells.

(A) Expression of Sema4A in single proximal and distal osteolineage cells. The joint posteriors (black lines) describe the overall estimation of likely expression levels within the proximal (top) and distal (bottom) OLCs and are used to estimate the posterior of the expression fold difference (middle plot). The shaded area under the fold-difference posterior shows 95% confidence region. FPM, fragments per million.

(B) The number of mouse LKS cells 24 hours after addition of mouse Sema4A-Fc/IgG1 control protein (n=5 technical replicates per condition).

(C) CFSE dilution analysis of *ex vivo* proliferation kinetics of human CD34+ cells 24 hours after addition of human Sema4A-Fc/IgG1 protein (n=5 technical replicates per condition, Donor 1). Estimated number of divided cells and representative CSFE fluorescence histograms are shown.

(D) Immunophenotypic analysis of the bone marrow from young WT/Sema4AKO mice (n=3-6 per genotype).

(E) HSC cell cycle analysis using DAPI/Ki-67 staining in young WT/Sema4AKO mice (n=6 per genotype).

(F) Gene set enrichment analysis (GSEA) of single cell RNA-Seq data for the HSC cluster (as defined in Figure S1K) from young WT/Sema4AKO mice. FDR<0.01, top five enriched pathways are shown.

(G) GSEA plots for the pathways as shown in (F).

Data are presented as mean ± SD $^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$ by unpaired T-test.

Figure 2. Sema4A/PlxnD1 signaling constrains the response of myeloid-biased HSC to proliferative stress.

(A) Experimental schema for the acute inflammatory stress model.

(B) Cell cycle analysis of myHSC 24 hours after injection with Poly(I:C) (n=5 mice per genotype).

(C) Cell cycle analysis of balHSC 24 hours after injection with Poly(I:C) (n=5 mice per genotype).

(D) WT vs Sema4AKO GSEA of myHSC from Poly(I:C) injected mice, FDR<0.01.

(E) GSEA plots and top differentially expressed genes for the pathways that were enriched in Sema4AKO myHSC, as shown in (D).

(F) Experimental schema for the transplant studies as shown in (G) and (H).

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.12.476097; this version posted January 12, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

(G) Donor chimerism in the recipients of myHSC from young WT/Sema4AKO mice (n=5 mice per donor genotype).

(H) Donor chimerism in the recipients of balHSC from young WT/Sema4AKO mice (n=4-5 mice per donor genotype).

(I) Published HSC gene expression data (Cabezas-Wallscheid et al., 2014) showing expression of known Sema4A receptors.

(J) Representative histograms (left panel) and quantification of mean fluorescence intensity of GFP expression (right panel) in myHSC (pink) and balHSC (blue) from PlxnD1-GFP reporter mice (n=3 mice).

(K) Donor chimerism in the recipients of myHSC from PlxnD1$^{fl/fl}$ Mx1-Cre(+)and PlxnD1$^{fl/fl}$ Mx1-Cre (-) mice (n=5 recipient mice per genotype).

(L) Donor chimerism in the recipients of balHSC from PlxnD1$^{fl/fl}$ Mx1-Cre(+)and PlxnD1$^{fl/fl}$ Mx1-Cre (-) mice (n=5 recipient mice per genotype).

Data are presented as mean ± SD *p < 0.05; **p < 0.01; ***p < 0.001 by unpaired T-test.

Figure 3. Sema4A prevents excessive myHSC expansion and functional loss with age.

(A) Serial peripheral blood counts of WT and Sema4AKO mice during aging (n=4-9 mice per genotype, age range 8-74 weeks).

(B) Immunophenotypic analysis of the bone marrow from aged (74-weeks old) WT/Sema4AKO mice (n=4-5 mice per genotype).

(C) Frequency of myHSC and balHSC in aged (74-weeks old) WT/Sema4AKO mice (representative flow cytometry plots shown on the right) (n=4-5 mice per genotype).

(D) Experimental schema for competitive myHSC/balHSC transplantation experiments shown in (E) and (F).

(E) Donor chimerism in the recipients of myHSC from aged WT/Sema4AKO mice (n=3-5 per donor genotype).

(F) Donor chimerism in the recipients of balHSC from aged WT/Sema4AKO mice (n=4-5 per donor genotype).

(G) UMAP representation of 162 myHSC from aged WT and Sema4AKO mice (n=2 mice per genotype).

(H) UMAP representation of 165 balHSC cells from WT and Sema4AKO mice (n=2 mice per genotype).

(I) Distribution of pairwise Spearman's correlation distances between aged WT and Sema4AKO myHSC (left) and balHSC (right). The two distributions are statistically significantly different according to a Wilcoxon rank-sum test (p-value = 3.1e-85).

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.12.476097; this version posted January 12, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

(J) GSEA of myHSC from aged WT/Sema4AKO mice, FDR<0.01.

(K) GSEA plots and top differentially expressed genes for pathways enriched in Sema4AKO myHSC.

(L) Percentage of myHSC and balHSC cells from aged WT/Sema4AKO in the G2M phase of the cell cycle as estimated from their transcriptome using *Cyclone*.

(M, N) Distributions of diffusion pseudotime values of myHSC (panel M) and balHSC (panel N) from aged WT/Sema4AKO. The P-values shown at the bottom were computed with a Wilcoxon-rank sum test.

Data are presented as mean ± SD *p < 0.05; **p < 0.01; ***p < 0.001 by unpaired T-test.

Figure 4. Sema4A from the bone marrow niche restrains stress-induced myHSC proliferation and maintains self-renewal.

(A) Donor chimerism in the recipients of myHSC from Sema4A$^{fl/fl}$ Cre(+) and Sema4A$^{fl/fl}$ Cre(-) mice (n=4-5 mice per donor genotype).

(B) Donor chimerism in the recipients of balHSC from Sema4A$^{fl/fl}$ Cre(+) and Sema4A$^{fl/fl}$ Cre(-) mice (n=4-5 mice per donor genotype).

(C) Experimental schema for non-competitive transplant experiments shown in (D) and (E).

(D) Survival curve (left panel) and hematocrit/neutrophil count (right panel) in WT/Sema4AKO recipients of myHSC (data are the summary of 6 independent experiments involving a total of 9-11 recipients per genotype).

(E) Survival curve and peripheral blood counts in WT/Sema4AKO recipients of balHSC (data are the summary of 6 independent experiments involving a total of 9-11 recipients per genotype).

(F) Average number of cells per mouse ~15-20 hours after transplantation of WT myHSC or balHSC into WT/Sema4AKO recipients, as assessed by two-photon intravital imaging of the calvarial bone marrow (Data are the summary of 6 independent experiments involving a total of 4-6 recipients per genotype).

(G) Quantification of 3D distances between individual transplanted cells and the nearest endosteal surface (n = 171, 628, 161, and 266 total cells for WT myHSC, Sema4AKO myHSC, WT balHSC, and Sema4AKO balHSC, respectively; Data are the summary of 6 independent experiments involving a total of 4-6 recipients per group).

(H) Representative time-lapse two-photon intravital image of single motile WT myHSC in the calvaria of Sema4AKO recipient. Each cell and white arrow correspond to a different timepoint at increments of 10 mins. The myHSC (DiD, red), bone (SHG, green), and autofluorescence (blue) are shown. Scale Bar ~ 25 μm.

Data are presented as mean ± SD ∗p < 0.05; ∗∗p < 0.01; ∗∗∗p < 0.001 by unpaired T-test.

7. Discussion

The studies included in the thesis reveal the power of utilizing single-cell transcriptomics in understanding cellular identity and differentiation in different tissues. Several aspects of cell fate decision-making could be explored by applying computational and mathematical methods to the data collected from single cells. The tissues investigated have varying degrees of cellular diversity. However, it was possible to capture the nature of heterogeneity between cells thanks to the high resolution of information that can be obtained with the current single-cell technologies. By employing computational tools on scRNA-seq data, I explored various ways of studying cellular decision-making in mammals.

7.1 Single-cell transcriptomic characterization of a human gastrula

This project entailed the first-ever molecular glimpse into a fundamental stage of mammalian development in humans – gastrulation. Data obtained through scRNA-seq from a rare sample in Carnegie stage 7 (E16-19) was crucial in corroborating previous knowledge, as well as revealing new insights. Some of these analyses include cell type identification, comparison with published data from *in vitro* systems and model organisms, and detection of cellular state transitions through trajectory reconstruction. They are briefly discussed below in more detail.

Quality control of the embryo

Before performing downstream analysis, we wanted to check the viability of the sample. This allowed us to obtain desirable knowledge from this rare sample more confidently. One way was to infer the cell cycle information from the cells. Inferring the cell cycle phases of the cells can help understand features of cellular plasticity. This was demonstrated in the studies included in this thesis. Cell cycle analysis was used to pinpoint different aspects in different systems. For human gastrula, assigning the cell cycle phases to the cells helped to confirm its viability. Because only one embryo was analyzed, such quality control was necessary. By showing that cells show differences in their cell cycle stage, we could conclude that normal cell division was probably going on throughout the embryo. Along with other metrics, cell cycle analysis provided reassurance about the quality of the specimen.

High-throughput sequencing of single cells can also provide an ability to take a glimpse into the DNA of the cells and infer embryo viability. Because of the higher cost associated with direct DNA sequencing, RNA sequencing can be an efficient way of handling questions related to the genome. Thus, without performing an additional whole-genome sequencing experiment, it is possible to extract relevant information from scRNA-seq data. The usefulness of this was demonstrated when checking the viability of the human embryo at the gastrula stage. For this purpose, insertions and deletions (indels) were detected in the genomes of the cells and compared to published scRNA-seq data obtained from fetal tissue (Segal et al., 2019). Like the cell cycle analysis, this provided assurance that the embryo was going through normal development, as no significant difference was detected between the two samples. Particularly, the distribution of indel lengths was quite similar. This demonstrates that scRNA-seq data can be useful to extract information other than gene expression.

Identification of cell-type diversification

Our transcriptomic characterization of the human gastrula confirmed the presence at this stage of many cell types that were expected based on the work done in other model organisms, like mouse. On the other hand, with our analysis, we also observed interesting differences between species, such as differences in the timing of blood cell formation. For example, the major germ layers were detected in the data by using the Leiden algorithm and known marker genes. Additionally, the epiblast that gives rise to these germ layers could also be found. Mesoderm, endoderm, and ectoderm themselves possess heterogeneity, so further clustering them revealed several cell types, like amniotic ectoderm, which was a sub-cluster of the ectoderm cluster. Microscopy images also indicated the presence of red-pigmented cells, which corresponded to one of the two blood clusters (erythroblasts). Characterizing these known cell types allowed for various downstream analyses. However, because a lot of cells are in a transition state towards their fates, pinpointing their exact identities was not straightforward. This could be seen in mesodermal clusters (nascent, emergent, and advanced), which were annotated based on their transcriptional similarity with the primitive streak, spatial origins of the cells, as well as gene expression patterns. While advanced mesodermal cells mostly occupied the rostral region, nascent mesoderm (NM) cells were all in the caudal region of the embryo. NM cells also expressed Brachury (TBXT), which is a PS marker.

Rare cell types usually get lost in the first stage of clustering in scRNA-seq datasets. Therefore, an additional step was needed to explore rare cell populations in the human gastrula data. Because common clustering algorithms are not able to capture small rare populations, a different algorithm was employed. RaceID (Grün et al., 2015) is designed specifically to detect these kinds of populations. It was applied to the epiblast and primitive streak (PS) populations separately. By using RaceID as well as previously known marker genes, we identified primordial germ cells (PGCs), which were not observed and transcriptionally characterized in *in vivo* human embryos before. This shows the power of the scRNA-seq in finding rare cell populations.

We also performed a more in-depth computational analysis that allowed us to detect cell-type specific splicing isoforms, by taking advantage of the fact that the scRNA-seq data was generated with a full-length protocol (Smart-seq2 (Picelli et al., 2014)). While a lot of genes are expressed at comparable levels in all cell populations, specific isoforms of those genes show differences in their expression. For example, mesodermal populations expressed specific isoforms of mesoderm specific transcript (MEST) gene.

Gene expression was not the only information used to characterize the identity of the cells. Because of the challenges mentioned above, having additional prior knowledge can help with cell annotation. In the case of the gastrula dataset, there was additional information on the location of the cells in the embryo. In particular, a sub-dissection of the sample prior to single-cell dissociation provided information about the spatial origin of each cell, i.e., whether each cell came from the rostral or caudal portion of the embryonic disc or from the yolk sac. This helped to confirm the annotation of clusters, as well as appreciate the heterogeneity within them. For instance, it is known that hematopoietic cells arise from the yolk sac. Two blood-related clusters detected in the data, erythroblasts and haemato-endothelial progenitors (HEPs) were mainly found in yolk-sac. The primitive streak and nascent mesoderm cells were only in the caudal region, as expected. As the mesodermal cells differentiate into a more mature state, they occupy the rostral part. In the advanced mesoderm cluster, unsupervised clustering did not indicate region specificity between the sub-clusters. However, comparing rostral and caudal regions directly revealed genes that were differentially expressed. These nuanced differences would not have been possible to detect without the information on spatial regions. Overall these suggest that although gene expression can be enough to identify the cells, additional information can be useful in confirming their identities and establishing ground truth for researchers who want to compare their gene expression data to the human gastrula.

Human gastrula data as a reference for *in vitro* models of embryonic development

Identifying and annotating the cells in human gastrula for the first time provides a unique atlas for future studies. Many studies have already used this dataset as a reference to test *in vitro* models of embryonic development (Jo et al., 2022), or to perform cross-species comparisons, e.g., with non-human primates (Bergmann et al., 2022).

In our paper, we used the human gastrula dataset to carry out two comparisons with *in vitro* models. To test whether human primed stem cells resemble CS7 epiblast, rather than naïve stem cells, two datasets were considered (Messmer et al., 2019; Petropoulos et al., 2016). Using Harmony as data integration method, we observed that human epiblast is closest to primed stem cells obtained from human ES cells (hESCs), while naïve hESCs were transcriptionally closest to embryonic day 6 and 7 of *in vivo* embryos. This result was also confirmed by integrating these datasets employing the algorithm implemented in Seurat v3 (Stuart et al., 2019), followed by hierarchical clustering (based on Spearman's correlation distance) on the corrected gene expression.

We also compared data collected across species (human, mouse and non-human primate) and with different experimental protocols (Smart-seq2 and 10x) when we analyzed the species-specific molecular fingerprints of PGCs. In this case, we computed the intra-sample z-scores in order to identify upregulated and downregulated genes in PGCs, compared to epiblast and PS populations from each species. The results require caution, however, since observing lack of gene expression can be due to drop-outs, especially in the more sparse 10x data.

We observed an overall good match between populations when comparing the human gastrula to mouse and non-human primate (NHP). In addition to the Seurat integration and hierarchical clustering method (see above), scmap (Kiselev et al., 2018) tool was also used for these comparisons. One issue in inter-species comparison arises from choosing orthologous genes. The methods for integrating two datasets usually require these datasets to have the same genes. Because one gene in an organism can have multiple orthologues in another organism, the integration task can provide an additional challenge. However, the number of genes with one-to-one mapping is usually far higher and can be enough for the algorithms to perform well.

Trajectory inference reveals molecular insights into cell differentiation during human gastrulation

Gastrulation is a stage at which many cell differentiation events occur. To explore the increase in cellular heterogeneity, it is important to establish trajectories to link cell states in pseudotime. For example, Epithelial to mesenchymal transition (EMT) is an integral part of gastrulation. To explore gene expression dynamics during EMT, a trajectory from epiblast to nascent mesoderm through the primitive streak cells was defined by using diffusion maps and diffusion pseudotime. More specifically, expression trends of the genes were estimated through computational and statistical tools. These tools are based on fitting gene expression levels as a function of pseudotime through generalized additive models, followed by ANOVA test to detect genes significantly up/down-regulated. One challenge here was to define the trend of the genes. Not all genes have a clear up or down trend because some only get highly expressed in the primitive streak and downregulate in the nascent mesoderm. One way to solve this issue was to set a threshold for log fold-change of the genes along the trajectory. This allowed inter-species comparison of gene expression patterns along EMT from epiblast to nascent mesoderm. Many expected similarities, as well as some intriguing differences, were found between mouse and human gastrulation. Although the same signaling pathways are involved in human and mouse EMT, specific members of the pathways showed differences in their regulation. Some of these bioinformatics-driven results were validated by *in* vitro differentiation model.

Overall, these results show the ability of scRNA-seq data to provide valuable molecular insights into differences in cell differentiation occurring between species. The inter-species analyses only considered genes with similar trends. However, it is also possible to align the timing of the gene regulation in both mouse and human EMT. For instance, one could use a method based on dynamic time warping (e.g. Trajan (Do et al., 2019)) to detect genes that undergo similar regulation between different organisms. However, the interpretation of this analysis requires caution as it is based on the assumption that the expression patterns of most genes are preserved across species.

Analysis of the entire data through diffusion pseudotime allowed the observation of three main cell fate decisions during gastrulation. These correspond to the major germ layers-ectoderm, endoderm, and mesoderm. RNA velocity was used to obtain information about the direction of differentiation. The velocities overlaid on top of the diffusion map highlighted the points at which the fate decisions were made.

7.2 Effect of Semaphorin 4a on mouse hematopoietic stem cell dormancy

In this project (Manuscript I), the aim was to investigate the role of Semaphorin 4a(Sema4a) on HSCs that possess a bias toward myeloid lineage (myHSCs). By generating single-cell libraries from WT and Sema4a KO samples from mouse HSCs, we performed a comparative analysis to test the hypothesis that the Sema4a is required for regulation of myHSC stemness.

Trajectory inference demonstrates the effect of Sema4a on myHSCs

To investigate the effects of Sema4a loss in myHSC differentiation, we analyzed the transcriptional trajectory joining myHSC and the progenitor cells. Using diffusion pseudo-time, we found evidence that the absence of Sema4a might make myeloid biased HSCs lose their stemness.

The ability of scRNA-seq data to provide an opportunity for the analyses mentioned above show the usefulness of transcriptional trajectory identification and analysis, across conditions (e.g., to find the effects of a KO, as discussed here) or species (e.g., as discussed above in the comparative analysis of gastrulation).

Estimation of HSC proliferation through cell cycle analysis

Consistent with the overproliferation of HSCs without Sema4a, cell cycle analysis of the scRNA-seq data showed that in the absence of Sema4a, there is a higher fraction of myHSCs in the G2/M cycle phase compared to WT cells (Appendix 1). The analysis also confirmed the hypothesis that Sema4a only affects the subset of HSC that are more biased towards myeloid lineage, as there was no difference between balanced HSCs in the absence of Sema4a. This shows that the transcriptional analysis of single cells can capture cell cycle information and their relation to defining cellular state. One limitation of the cell cycle detection method was the inability to differentiate between cells in G0 and G1 phases.

Because analyzing dormancy requires this information, checking marker gene expression specific to G0 between the conditions can be necessary. Additionally, RNA velocity can potentially be used to distinguish cells in the G0 phase from the cells G1 phase, where velocity vectors would originate specifically from G1 cells in the direction of cells in the S phase.

7.3 Detecting heteroplasmy from scRNA-seq of mouse epiblast cells

Aside from computationally detecting large structural genomic variations, such as indels and copy number variations (CNV) from scRNA-seq data (Yang et al., 2018; Serin Harmanci et al., 2020), investigating single nucleotide variants (SNV) is also possible. This can be especially useful when constructing a lineage tree with genetic scars or natural barcodes, as described previously in the thesis. An example of this was when scRNA-seq was used to build a lineage tree with mitochondrial heteroplasmy (Ludwig et al., 2019). Although direct sequencing of mtDNA was better at detecting SNVs in that study, scRNA-seq still captured most of the variants. We used this approach in exploring the relationship between mitochondria heteroplasmy and cellular competition in Publication 2 (Lima et al., 2021). Not only did the overall heteroplasmy levels between “loser” and “winning” cells have differences, but it was also possible to detect mtDNA variants specific to cells losing the competition.

It is still important to be cautious to detect SNVs from scRNA-seq data. Modifications specific to transcripts might not be present in the DNA. Additionally, the variable gene expression levels between cells can lead to biases stemming from differing read depth. We took the read depth issue into account while creating a pipeline to quantify mitochondrial heteroplasmy from scRNA-seq data, in order to remove the genomic regions that do not have adequate coverage (Lubatti et al., 2022). Additionally, the nuclear mitochondrial sequences (NUMTs) can lead to possible artifacts which would bias the results. Therefore, in our analyses we also checked whether the sequences with high heteroplasmy correspond to NUMTs.

7.4 Future outlook

The successful use of scRNA-seq to study cellular fate decision in various contexts has been demonstrated in this thesis. We saw how gene expression patterns can reveal insights into cellular heterogeneity in various biological systems by using appropriate computational tools. As both the experimental and computational limitations are gradually solved, the single-cell transcriptomics field constantly offers more exciting research. In combination with different types of molecular profiling, such as single-cell epigenomics and proteomics, new powerful ways of exploring cellular heterogeneity and identity can be established. Additionally, by obtaining spatial transcriptomics on the same sample, cells can be analyzed in their localized context to elucidate their interactions (Longo et al., 2021). Finally, the computational tools used to analyze scRNA-seq data can be complemented with quantitative models to help explain molecular dynamics within cells. Therefore, integration of experiments, data analysis and mechanistic modeling can bring more insight in cellular fate decision, as was the case in Publication 4 (Nakatani et al., 2022).

8. References

Bergen, V., Soldatov, R. A., Kharchenko, P. V, & Theis, F. J. (2021). RNA velocity—current challenges and future perspectives. *Molecular Systems Biology*, *17*(8), e10282. https://doi.org/https://doi.org/10.15252/msb.202110282

Bergmann, S., Penfold, C. A., Slatery, E., Siriwardena, D., Drummer, C., Clark, S., Strawbridge, S. E., Kishimoto, K., Vickers, A., Tewary, M., Kohler, T. N., Hollfelder, F., Reik, W., Sasaki, E., Behr, R., & Boroviak, T. E. (n.d.). Spatial profiling of early primate gastrulation in utero. *Nature*, 1–3. https://doi.org/10.1038/s41586-022-04953-1

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, *2008*(10), P10008. https://doi.org/10.1088/1742-5468/2008/10/p10008

Briggs, J. A., Weinreb, C., Wagner, D. E., Megason, S., Peshkin, L., Kirschner, M. W., & Klein, A. M. (2018). The dynamics of gene expression in vertebrate embryogenesis at single-cell resolution. *Science (New York, N.Y.)*, *360*(6392). https://doi.org/10.1126/science.aar5780

Buettner, F., Natarajan, K. N., Casale, F. P., Proserpio, V., Scialdone, A., Theis, F. J., Teichmann, S. A., Marioni, J. C., & Stegle, O. (2015). Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. *Nature Biotechnology*, *33*(2), 155–160. https://doi.org/10.1038/nbt.3102

Büttner, M., Miao, Z., Wolf, F. A., Teichmann, S. A., & Theis, F. J. (2019). A test metric for assessing single-cell RNA-seq batch correction. *Nature Methods*, *16*(1), 43–49. https://doi.org/10.1038/s41592-018-0254-1

Carnevali, M. D. C. (2006). Regeneration in Echinoderms: repair, regrowth, cloning. *ISJ-Invertebrate Survival Journal*, *3*, 64–76.

Cauffman, G., Liebaers, I., Van Steirteghem, A., & Van de Velde, H. (2006). POU5F1 isoforms show different expression patterns in human embryonic stem cells and preimplantation embryos. *Stem Cells (Dayton, Ohio)*, *24*(12), 2685–2691. https://doi.org/10.1634/stemcells.2005-0611

Conklin, E. G. (1905). Mosaic development in ascidian eggs. *Journal of Experimental Zoology*, *2*(2), 145–223. https://doi.org/https://doi.org/10.1002/jez.1400020202

Coons, A. H., Creech, H. J., & Jones, R. N. (1941). Immunological Properties of an Antibody Containing a Fluorescent Group. *Proceedings of the Society for Experimental Biology and Medicine*, *47*(2), 200–202. https://doi.org/10.3181/00379727-47-13084P

Das, S., Rai, A., Merchant, M. L., Cave, M. C., & Rai, S. N. (2021). A Comprehensive Survey of Statistical Approaches for Differential Expression Analysis in Single-Cell RNA Sequencing Studies. *Genes*, *12*(12). https://doi.org/10.3390/genes12121947

Delaney, C., Schnell, A., Cammarata, L. V, Yao-Smith, A., Regev, A., Kuchroo, V. K., & Singer, M. (2019). Combinatorial prediction of marker panels from single-cell transcriptomic data. *Molecular Systems Biology*, *15*(10), e9005. https://doi.org/https://doi.org/10.15252/msb.20199005

Do, V. H., Blažević, M., Monteagudo, P., Borozan, L., Elbassioni, K., Laue, S., Rojas Ringeling, F., Matijević, D., & Canzar, S. (2019). *Dynamic pseudo-time warping of complex single-cell trajectories*. https://doi.org/10.1101/522672

Finak, G., McDavid, A., Yajima, M., Deng, J., Gersuk, V., Shalek, A. K., Slichter, C. K., Miller, H. W., McElrath, M. J., Prlic, M., Linsley, P. S., & Gottardo, R. (2015). MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biology*, *16*(1), 278. https://doi.org/10.1186/s13059-015-0844-5

Gauthier, J., Vincent, A. T., Charette, S. J., & Derome, N. (2019). A brief history of bioinformatics. *Briefings in Bioinformatics*, *20*(6), 1981–1996. https://doi.org/10.1093/bib/bby063

Goldsby, H. J., Dornhaus, A., Kerr, B., & Ofria, C. (2012). Task-switching costs promote the evolution of division of labor and shifts in individuality. *Proceedings of the National Academy of Sciences*, *109*(34), 13686–13691. https://doi.org/10.1073/pnas.1202233109

Golgi, C. (1883). Sulla fina anatomia degli organi centrali del sistema nervoso. *Rivista Sperimentale Di Freniatria e Di Medicina Legale*, *IX*.

Grün, D., Lyubimova, A., Kester, L., Wiebrands, K., Basak, O., Sasaki, N., Clevers, H., & van Oudenaarden, A. (2015). Single-cell messenger RNA sequencing reveals rare intestinal cell types. *Nature*, *525*(7568), 251–255. https://doi.org/10.1038/nature14966

Gyanchandani, R., Kota, K. J., Jonnalagadda, A. R., Minteer, T., Knapick, B. A., Oesterreich, S., Brufsky, A. M., Lee, A. V, & Puhalla, S. L. (2017). Detection of ESR1 mutations in circulating cell-free DNA from patients with metastatic breast cancer treated with palbociclib and letrozole. *Oncotarget*, *8*(40), 66901–66911. https://doi.org/10.18632/oncotarget.11383

Haghverdi, L., Lun, A. T. L., Morgan, M. D., & Marioni, J. C. (2018). Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nature Biotechnology*, *36*(5), 421–427. https://doi.org/10.1038/nbt.4091

Hie, B., Bryson, B., & Berger, B. (2019). Efficient integration of heterogeneous single-cell transcriptomes using Scanorama. *Nature Biotechnology*, *37*(6), 685–691. https://doi.org/10.1038/s41587-019-0113-3

Jo, K., Teague, S., Chen, B., Khan, H. A., Freeburne, E., Li, H., Li, B., Ran, R., Spence, J. R., & Heemskerk, I. (2022). Efficient differentiation of human primordial germ cells through geometric control reveals a key role for Nodal signaling. *ELife*, *11*, e72811. https://doi.org/10.7554/elife.72811

Kanton, S., Boyle, M. J., He, Z., Santel, M., Weigert, A., Sanchís-Calleja, F., Guijarro, P., Sidow, L., Fleck, J. S., Han, D., Qian, Z., Heide, M., Huttner, W. B., Khaitovich, P., Pääbo, S., Treutlein, B., & Camp, J. G. (2019). Organoid single-cell genomic atlas uncovers human-specific features of brain development. *Nature*, *574*(7778), 418–422. https://doi.org/10.1038/s41586-019-1654-9

Kiselev, V. Y., Andrews, T. S., & Hemberg, M. (2019). Challenges in unsupervised clustering of single-cell RNA-seq data. *Nature Reviews Genetics*, *20*(5), 273–282. https://doi.org/10.1038/s41576-018-0088-9

Kiselev, V. Y., Yiu, A., & Hemberg, M. (2018). scmap: projection of single-cell RNA-seq data across data sets. *Nature Methods*, *15*(5), 359–362. https://doi.org/10.1038/nmeth.4644

Knoll, A. H. (2011). The Multiple Origins of Complex Multicellularity. *Annual Review of Earth and Planetary Sciences*, *39*(1), 217–239. https://doi.org/10.1146/annurev.earth.031208.100209

Korsunsky, I., Millard, N., Fan, J., Slowikowski, K., Zhang, F., Wei, K., Baglaenko, Y., Brenner, M., Loh, P., & Raychaudhuri, S. (2019). Fast, sensitive and accurate integration of single-cell data with Harmony. *Nature Methods*, *16*(12), 1289–1296. https://doi.org/10.1038/s41592-019-0619-0

Korthauer, K. D., Chu, L.-F., Newton, M. A., Li, Y., Thomson, J., Stewart, R., & Kendziorski, C. (2016). A statistical approach for identifying differential distributions in single-cell RNA-seq experiments. *Genome Biology*, *17*(1), 222. https://doi.org/10.1186/s13059-016-1077-y

Kretzschmar, K., & Watt, F. M. (2012). Lineage Tracing. *Cell*, *148*(1), 33–45. https://doi.org/https://doi.org/10.1016/j.cell.2012.01.002

La Manno, G., Soldatov, R., Zeisel, A., Braun, E., Hochgerner, H., Petukhov, V., Lidschreiber, K., Kastriti, M. E., Lönnerberg, P., Furlan, A., Fan, J., Borm, L. E., Liu, Z., van Bruggen, D., Guo, J., He, X., Barker, R., Sundström, E., Castelo-Branco, G., … Kharchenko, P. V. (2018). RNA velocity of single cells. *Nature*, *560*(7719), 494–498. https://doi.org/10.1038/s41586-018-0414-6

Lähnemann, D., Köster, J., Szczurek, E., McCarthy, D. J., Hicks, S. C., Robinson, M. D., Vallejos, C. A., Campbell, K. R., Beerenwinkel, N., Mahfouz, A., Pinello, L., Skums, P., Stamatakis, A., Attolini, C. S.-O., Aparicio, S., Baaijens, J., Balvert, M., Barbanson, B. de, Cappuccio, A., … Schönhuth, A. (2020). Eleven grand challenges in single-cell data science. *Genome Biology*, *21*(1), 31. https://doi.org/10.1186/s13059-020-1926-6

Lima, A., Lubatti, G., Burgstaller, J., Hu, D., Green, A. P., Di Gregorio, A., Zawadzki, T., Pernaute, B., Mahammadov, E., Perez-Montero, S., Dore, M., Sanchez, J. M., Bowling, S., Sancho, M., Kolbe, T., Karimi, M. M., Carling, D., Jones, N., Srinivas, S., … Rodriguez, T. A. (2021). Cell competition acts as a purifying selection to eliminate cells with mitochondrial defects during early mouse development. *Nature Metabolism*, *3*(8), 1091–1108. https://doi.org/10.1038/s42255-021-00422-7

Longo, S. K., Guo, M. G., Ji, A. L., & Khavari, P. A. (2021). Integrating single-cell and spatial transcriptomics to elucidate intercellular tissue dynamics. *Nature Reviews Genetics*, *22*(10), 627–644. https://doi.org/10.1038/s41576-021-00370-8

Lopez, R., Regier, J., Cole, M. B., Jordan, M. I., & Yosef, N. (2018). Deep generative modeling for single-cell transcriptomics. *Nature Methods*, *15*(12), 1053–1058. https://doi.org/10.1038/s41592-018-0229-2

Lorthongpanich, C., Doris, T. P. Y., Limviphuvadh, V., Knowles, B. B., & Solter, D. (2012). Developmental fate and lineage commitment of singled mouse blastomeres. *Development (Cambridge, England)*, *139*(20), 3722–3731. https://doi.org/10.1242/dev.086454

Lotfollahi, M., Naghipourfar, M., Luecken, M. D., Khajavi, M., Büttner, M., Wagenstetter, M., Avsec, Ž., Gayoso, A., Yosef, N., Interlandi, M., Rybakov, S., Misharin, A. V, & Theis, F. J. (2022). Mapping single-cell data to reference atlases by transfer learning. *Nature Biotechnology*, *40*(1), 121–130. https://doi.org/10.1038/s41587-021-01001-7

Lotfollahi, M., Wolf, F. A., & Theis, F. J. (2019). scGen predicts single-cell perturbation responses. *Nature Methods*, *16*(8), 715–721. https://doi.org/10.1038/s41592-019-0494-8

Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, *15*(12), 550. https://doi.org/10.1186/s13059-014-0550-8

Lubatti, G., Mahammadov, E., & Scialdone, A. (2022). MitoHEAR: an R package for the estimation and downstream statistical analysis of the mitochondrial DNA heteroplasmy calculated from single-cell datasets. *Journal of Open Source Software*, *7*, 4265. https://doi.org/10.21105/joss.04265

Ludwig, L. S., Lareau, C. A., Ulirsch, J. C., Christian, E., Muus, C., Li, L. H., Pelka, K., Ge, W., Oren, Y., Brack, A., Law, T., Rodman, C., Chen, J. H., Boland, G. M., Hacohen, N., Rozenblatt-Rosen, O., Aryee, M. J., Buenrostro, J. D., Regev, A., & Sankaran, V. G. (2019). Lineage Tracing in Humans Enabled by Mitochondrial Mutations and Single-Cell Genomics. *Cell*, *176*(6), 1325-1339.e22. https://doi.org/10.1016/j.cell.2019.01.022

Luecken, M. D., Büttner, M., Chaichoompu, K., Danese, A., Interlandi, M., Mueller, M. F., Strobl, D. C., Zappia, L., Dugas, M., Colomé-Tatché, M., & Theis, F. J. (2022). Benchmarking atlas-level data integration in single-cell genomics. *Nature Methods*, *19*(1), 41–50. https://doi.org/10.1038/s41592-021-01336-8

Luecken, M. D., & Theis, F. J. (2019). Current best practices in single-cell RNA-seq analysis: a tutorial. *Molecular Systems Biology*, *15*(6), e8746. https://doi.org/https://doi.org/10.15252/msb.20188746

Ma, H., Zhai, J., Wan, H., Jiang, X., Wang, X., Wang, L., Xiang, Y., He, X., Zhao, Z.-A., Zhao, B., Zheng, P., Li, L., & Wang, H. (2019). In vitro culture of cynomolgus monkey embryos beyond early gastrulation. *Science (New York, N.Y.)*, *366*(6467). https://doi.org/10.1126/science.aax7890

McKenna, A., Findlay, G. M., Gagnon, J. A., Horwitz, M. S., Schier, A. F., & Shendure, J. (2016). Whole-organism lineage tracing by combinatorial and cumulative genome editing. *Science*, *353*(6298), aaf7907. https://doi.org/10.1126/science.aaf7907

Messmer, T., von Meyenn, F., Savino, A., Santos, F., Mohammed, H., Lun, A. T. L., Marioni, J. C., & Reik, W. (2019). Transcriptional Heterogeneity in Naive and Primed Human Pluripotent Stem Cells at Single-Cell Resolution. *Cell Reports*, *26*(4), 815-824.e4. https://doi.org/10.1016/j.celrep.2018.12.099

Moris, N., Anlas, K., van den Brink, S. C., Alemany, A., Schröder, J., Ghimire, S., Balayo, T., van Oudenaarden, A., & Martinez Arias, A. (2020). An in vitro model of early anteroposterior organization during human development. *Nature*, *582*(7812), 410–415. https://doi.org/10.1038/s41586-020-2383-9

Nakatani, T., Lin, J., Ji, F., Ettinger, A., Pontabry, J., Tokoro, M., Altamirano-Pacheco, L., Fiorentino, J., Mahammadov, E., Hatano, Y., Van Rechem, C., Chakraborty, D., Ruiz-Morales, E. R., Arguello Pascualli, P. Y., Scialdone, A., Yamagata, K., Whetstine, J. R., Sadreyev, R. I., & Torres-Padilla, M.-E. (2022). DNA replication fork speed underlies cell fate changes and promotes reprogramming. *Nature Genetics*, *54*(3), 318–327. https://doi.org/10.1038/s41588-022-01023-0

Nguyen, P. K., Nag, D., & Wu, J. C. (2010). Methods to assess stem cell lineage, fate and function. *Advanced Drug Delivery Reviews*, *62*(12), 1175–1186. https://doi.org/https://doi.org/10.1016/j.addr.2010.08.008

Ntege, E. H., Sunami, H., & Shimizu, Y. (2020). Advances in regenerative therapy: A review of the literature and future directions. *Regenerative Therapy*, *14*, 136–153. https://doi.org/10.1016/j.reth.2020.01.004

Pardue, M. L., & Gall, J. G. (1969). Molecular hybridization of radioactive DNA to the DNA of cytological preparations. *Proceedings of the National Academy of Sciences of the United States of America*, *64*(2), 600–604. https://doi.org/10.1073/pnas.64.2.600

Parfrey, L. W., & Lahr, D. J. G. (2013). Multicellularity arose several times in the evolution of eukaryotes (response to DOI 10.1002/bies.201100187). *BioEssays : News and Reviews in Molecular, Cellular and Developmental Biology*, *35*(4), 339–347. https://doi.org/10.1002/bies.201200143

Pei, W., Feyerabend, T. B., Rössler, J., Wang, X., Postrach, D., Busch, K., Rode, I., Klapproth, K., Dietlein, N., Quedenau, C., Chen, W., Sauer, S., Wolf, S., Höfer, T., & Rodewald, H.-R. (2017). Polylox barcoding reveals haematopoietic stem cell fates realized in vivo. *Nature*, *548*(7668), 456–460. https://doi.org/10.1038/nature23653

Petropoulos, S., Edsgärd, D., Reinius, B., Deng, Q., Panula, S. P., Codeluppi, S., Plaza Reyes, A., Linnarsson, S., Sandberg, R., & Lanner, F. (2016). Single-Cell RNA-Seq Reveals Lineage and X Chromosome Dynamics in Human Preimplantation Embryos. *Cell*, *165*(4), 1012–1026. https://doi.org/10.1016/j.cell.2016.03.023

Picelli, S., Faridani, O. R., Björklund, Å. K., Winberg, G., Sagasser, S., & Sandberg, R. (2014). Full-length RNA-seq from single cells using Smart-seq2. *Nature Protocols*, *9*(1), 171–181. https://doi.org/10.1038/nprot.2014.006

Pijuan-Sala, B., Griffiths, J. A., Guibentif, C., Hiscock, T. W., Jawaid, W., Calero-Nieto, F. J., Mulas, C., Ibarra-Soria, X., Tyser, R. C. V, Ho, D. L. L., Reik, W., Srinivas, S., Simons, B. D., Nichols, J., Marioni, J. C., & Göttgens, B. (2019). A single-cell molecular map of mouse gastrulation and early organogenesis. *Nature*, *566*(7745), 490–495. https://doi.org/10.1038/s41586-019-0933-9

Réaumur, R. A. F. de. (1712). *Observations sur les diverses reproductions qui se font dans les Écrevisses, les Omards, les Crabes, etc., et entr'autres sur celles de leurs jambes et de leurs écailles*. Académie royale des sciences.

Regev, A., Teichmann, S. A., Lander, E. S., Amit, I., Benoist, C., Birney, E., Bodenmiller, B., Campbell, P., Carninci, P., Clatworthy, M., Clevers, H., Deplancke, B., Dunham, I., Eberwine, J., Eils, R., Enard, W., Farmer, A., Fugger, L., Göttgens, B., … Yosef, N. (2017). The Human Cell Atlas. *ELife*, *6*. https://doi.org/10.7554/eLife.27041

Saelens, W., Cannoodt, R., Todorov, H., & Saeys, Y. (2019). A comparison of single-cell trajectory inference methods. *Nature Biotechnology*, *37*(5), 547–554. https://doi.org/10.1038/s41587-019-0071-9

Sanjuan-Pla, A., Macaulay, I. C., Jensen, C. T., Woll, P. S., Luis, T. C., Mead, A., Moore, S., Carella, C., Matsuoka, S., Bouriez Jones, T., Chowdhury, O., Stenson, L., Lutteropp, M., Green, J. C. A., Facchini, R., Boukarabila, H., Grover, A., Gambardella, A., Thongjuea, S., … Jacobsen, S. E. W. (2013). Platelet-biased stem cells reside at the apex of the haematopoietic stem-cell hierarchy. *Nature*, *502*(7470), 232–236. https://doi.org/10.1038/nature12495

Scialdone, A., Natarajan, K. N., Saraiva, L. R., Proserpio, V., Teichmann, S. A., Stegle, O., Marioni, J. C., & Buettner, F. (2015). Computational assignment of cell-cycle stage from single-cell transcriptome data. *Methods*, *85*, 54–61. https://doi.org/https://doi.org/10.1016/j.ymeth.2015.06.021

Segal, J. M., Kent, D., Wesche, D. J., Ng, S. S., Serra, M., Oulès, B., Kar, G., Emerton, G., Blackford, S. J. I., Darmanis, S., Miquel, R., Luong, T. V., Yamamoto, R., Bonham, A., Jassem, W., Heaton, N., Vigilante, A., King, A., Sancho, R., … Rashid, S. T. (2019). Single cell analysis of human foetal liver captures the transcriptional profile of hepatobiliary hybrid progenitors. *Nature Communications*, *10*(1), 3350. https://doi.org/10.1038/s41467-019-11266-x

Serin Harmanci, A., Harmanci, A. O., & Zhou, X. (2020). CaSpER identifies and visualizes CNV events by integrative analysis of single-cell or bulk RNA-sequencing data. *Nature Communications*, *11*(1), 89. https://doi.org/10.1038/s41467-019-13779-x

Sogabe, S., Hatleberg, W. L., Kocot, K. M., Say, T. E., Stoupin, D., Roper, K. E., Fernandez-Valverde, S. L., Degnan, S. M., & Degnan, B. M. (2019). Pluripotency and the origin of animal multicellularity. *Nature*, *570*(7762), 519–522. https://doi.org/10.1038/s41586-019-1290-4

Soneson, C., & Robinson, M. D. (2018). Bias, robustness and scalability in single-cell differential expression analysis. *Nature Methods*, *15*(4), 255–261. https://doi.org/10.1038/nmeth.4612

Squair, J. W., Gautier, M., Kathe, C., Anderson, M. A., James, N. D., Hutson, T. H., Hudelle, R., Qaiser, T., Matson, K. J. E., Barraud, Q., Levine, A. J., La Manno, G., Skinnider, M. A., & Courtine, G. (2021). Confronting false discoveries in single-cell differential expression. *Nature Communications*, *12*(1), 5692. https://doi.org/10.1038/s41467-021-25960-2

Stark, R., Grzelak, M., & Hadfield, J. (2019). RNA sequencing: the teenage years. *Nature Reviews. Genetics*, *20*(11), 631–656. https://doi.org/10.1038/s41576-019-0150-2

Stegle, O., Teichmann, S. A., & Marioni, J. C. (2015). Computational and analytical challenges in single-cell transcriptomics. *Nature Reviews. Genetics*, *16*(3), 133–145. https://doi.org/10.1038/nrg3833

Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck III, W. M., Hao, Y., Stoeckius, M., Smibert, P., & Satija, R. (2019). Comprehensive Integration of Single-Cell Data. *Cell*, *177*(7), 1888-1902.e21. https://doi.org/10.1016/j.cell.2019.05.031

Sulston, J. E., Schierenberg, E., White, J. G., & Thomson, J. N. (1983). The embryonic cell lineage of the nematode Caenorhabditis elegans. *Developmental Biology*, *100*(1), 64–119. https://doi.org/10.1016/0012-1606(83)90201-4

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

Tang, F., Barbacioru, C., Wang, Y., Nordman, E., Lee, C., Xu, N., Wang, X., Bodeau, J., Tuch, B. B., Siddiqui, A., Lao, K., & Surani, M. A. (2009). mRNA-Seq whole-transcriptome analysis of a single cell. *Nature Methods*, *6*(5), 377–382. https://doi.org/10.1038/nmeth.1315

Tong, K., Bozdag, G. O., & Ratcliff, W. C. (2022). Selective drivers of simple multicellularity. *Current Opinion in Microbiology*, *67*, 102141. https://doi.org/https://doi.org/10.1016/j.mib.2022.102141

Traag, V. A., Waltman, L., & van Eck, N. J. (2019). From Louvain to Leiden: guaranteeing well-connected communities. *Scientific Reports*, *9*(1), 5233. https://doi.org/10.1038/s41598-019-41695-z

Trapnell, C. (2015). Defining cell types and states with single-cell genomics. *Genome Research*, *25*(10), 1491–1498. https://doi.org/10.1101/gr.190595.115

Trembley, A. (1744). *Mémoires, Pour Servir à l'Histoire d'un Genre de Polypes d'Eau Douce, à Bras en Forme de Cornes.* Verbeek.

Tyser, R. C. V, Mahammadov, E., Nakanoh, S., Vallier, L., Scialdone, A., & Srinivas, S. (2021). Single-cell transcriptomic characterization of a gastrulating human embryo. *Nature*, *600*(7888), 285–289. https://doi.org/10.1038/s41586-021-04158-y

Valentine, J. W. (2003). Architectures of Biological Complexity. *Integrative and Comparative Biology*, *43*(1), 99–103. http://www.jstor.org/stable/3884844

Vandenberg, L. N., Adams, D. S., & Levin, M. (2012). Normalized shape and location of perturbed craniofacial structures in the Xenopus tadpole reveal an innate ability to achieve correct morphology. *Developmental Dynamics : An Official Publication of the American Association of Anatomists*, *241*(5), 863–878. https://doi.org/10.1002/dvdy.23770

VanHorn, S., & Morris, S. A. (2021). Next-Generation Lineage Tracing and Fate Mapping to Interrogate Development. *Developmental Cell*, *56*(1), 7–21. https://doi.org/https://doi.org/10.1016/j.devcel.2020.10.021

Vieira, W. A., Wells, K. M., & McCusker, C. D. (2020). Advancements to the Axolotl Model for Regeneration and Aging. *Gerontology*, *66*(3), 212–222. https://doi.org/10.1159/000504294

Wagner, D. E., Weinreb, C., Collins, Z. M., Briggs, J. A., Megason, S. G., & Klein, A. M. (2018). Single-cell mapping of gene expression landscapes and lineage in the zebrafish embryo. *Science*, *360*(6392), 981–987. https://doi.org/10.1126/science.aar4362

Watcham, S., Kucinski, I., & Gottgens, B. (2019). New insights into hematopoietic differentiation landscapes from single-cell RNA sequencing. *Blood*, *133*(13), 1415–1426. https://doi.org/10.1182/blood-2018-08-835355

Weiss, M. C., Sousa, F. L., Mrnjavac, N., Neukirchen, S., Roettger, M., Nelson-Sathi, S., & Martin, W. F. (2016). The physiology and habitat of the last universal common ancestor. *Nature Microbiology*, *1*(9), 16116. https://doi.org/10.1038/nmicrobiol.2016.116

Wolf, F. A., Angerer, P., & Theis, F. J. (2018). SCANPY: large-scale single-cell gene expression data analysis. *Genome Biology*, *19*(1), 15. https://doi.org/10.1186/s13059-017-1382-0

Yang, R., Van Etten, J. L., & Dehm, S. M. (2018). Indel detection from DNA and RNA sequencing data with transIndel. *BMC Genomics*, *19*(1), 270. https://doi.org/10.1186/s12864-018-4671-4

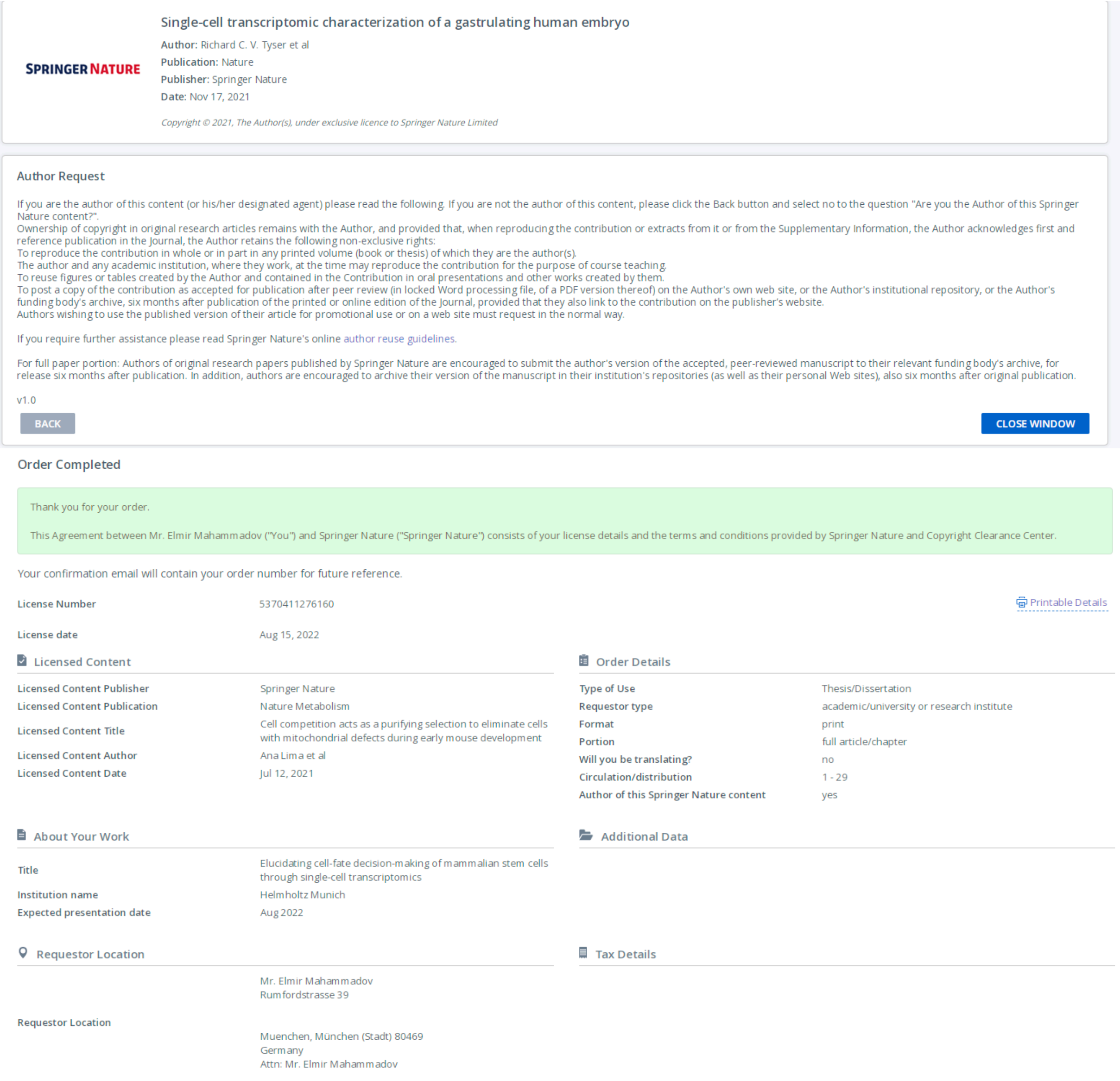
Yao, Z., Liu, H., Xie, F., Fischer, S., Adkins, R. S., Aldridge, A. I., Ament, S. A., Bartlett, A., Behrens, M. M., Van den Berge, K., Bertagnolli, D., de Bézieux, H. R., Biancalani, T., Booeshaghi, A. S., Bravo, H. C., Casper, T., Colantuoni, C., Crabtree, J., Creasy, H., … Mukamel, E. A. (2021). A transcriptomic and epigenomic cell atlas of the mouse primary motor cortex. *Nature*, *598*(7879), 103–110. https://doi.org/10.1038/s41586-021-03500-8

Zappia, L., Phipson, B., & Oshlack, A. (2018). Exploring the single-cell RNA-seq analysis landscape with the scRNA-tools database. *PLOS Computational Biology*, *14*(6), e1006245. https://doi.org/10.1371/journal.pcbi.1006245

Zattara, E. E., Fernández-Álvarez, F. A., Hiebert, T. C., Bely, A. E., & Norenburg, J. L. (2019). A phylum-wide survey reveals multiple independent gains of head regeneration in Nemertea. *Proceedings. Biological Sciences*, *286*(1898), 20182524. https://doi.org/10.1098/rspb.2018.2524

9. Appendices

9.1 Copyright statements



Single-cell transcriptomic characterization of a gastrulating human embryo

Author: Richard C. V. Tyser et al
Publication: Nature
Publisher: Springer Nature
Date: Nov 17, 2021

Copyright © 2021, The Author(s), under exclusive licence to Springer Nature Limited

Author Request

If you are the author of this content (or his/her designated agent) please read the following. If you are not the author of this content, please click the Back button and select no to the question "Are you the Author of this Springer Nature content?".
Ownership of copyright in original research articles remains with the Author, and provided that, when reproducing the contribution or extracts from it or from the Supplementary Information, the Author acknowledges first and reference publication in the Journal, the Author retains the following non-exclusive rights:
To reproduce the contribution in whole or in part in any printed volume (book or thesis) of which they are the author(s).
The author and any academic institution, where they work, at the time may reproduce the contribution for the purpose of course teaching.
To reuse figures or tables created by the Author and contained in the Contribution in oral presentations and other works created by them.
To post a copy of the contribution as accepted for publication after peer review (in locked Word processing file, of a PDF version thereof) on the Author's own web site, or the Author's institutional repository, or the Author's funding body's archive, six months after publication of the printed or online edition of the Journal, provided that they also link to the contribution on the publisher's website.
Authors wishing to use the published version of their article for promotional use or on a web site must request in the normal way.

If you require further assistance please read Springer Nature's online author reuse guidelines.

For full paper portion: Authors of original research papers published by Springer Nature are encouraged to submit the author's version of the accepted, peer-reviewed manuscript to their relevant funding body's archive, for release six months after publication. In addition, authors are encouraged to archive their version of the manuscript in their institution's repositories (as well as their personal Web sites), also six months after original publication.

v1.0

Order Completed

Thank you for your order.

This Agreement between Mr. Elmir Mahammadov ("You") and Springer Nature ("Springer Nature") consists of your license details and the terms and conditions provided by Springer Nature and Copyright Clearance Center.

Your confirmation email will contain your order number for future reference.

License Number	5370411276160
License date	Aug 15, 2022

Licensed Content

Licensed Content Publisher	Springer Nature
Licensed Content Publication	Nature Metabolism
Licensed Content Title	Cell competition acts as a purifying selection to eliminate cells with mitochondrial defects during early mouse development
Licensed Content Author	Ana Lima et al
Licensed Content Date	Jul 12, 2021

Order Details

Type of Use	Thesis/Dissertation
Requestor type	academic/university or research institute
Format	print
Portion	full article/chapter
Will you be translating?	no
Circulation/distribution	1 - 29
Author of this Springer Nature content	yes

About Your Work

Title	Elucidating cell-fate decision-making of mammalian stem cells through single-cell transcriptomics
Institution name	Helmholtz Munich
Expected presentation date	Aug 2022

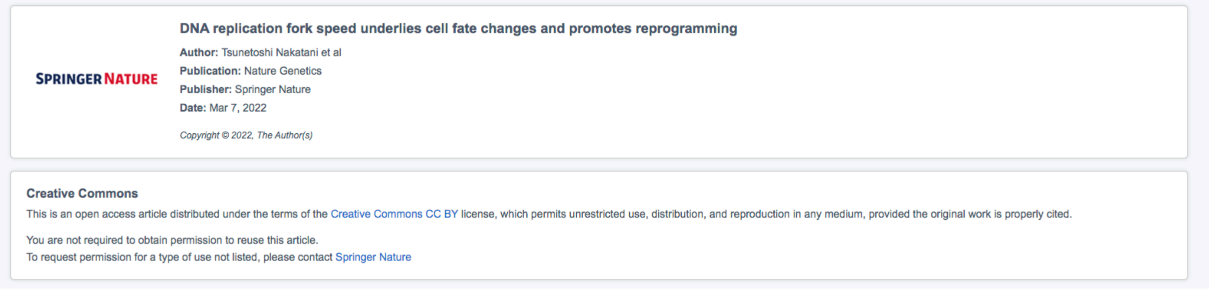
Additional Data

Requestor Location

Requestor Location
Mr. Elmir Mahammadov
Rumfordstrasse 39

Muenchen, München (Stadt) 80469
Germany
Attn: Mr. Elmir Mahammadov

Tax Details



DNA replication fork speed underlies cell fate changes and promotes reprogramming

Author: Tsunetoshi Nakatani et al
Publication: Nature Genetics
Publisher: Springer Nature
Date: Mar 7, 2022

Copyright © 2022, The Author(s)

Creative Commons

This is an open access article distributed under the terms of the Creative Commons CC BY license, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

You are not required to obtain permission to reuse this article.
To request permission for a type of use not listed, please contact Springer Nature

9.2 Curriculum Vitae

Elmir Mahammadov

Rumfordstrasse 39, 80469, Munich • 017682102884 • elmir.mahammadov@helmholtz-muenchen.de

---

Education

Ph.D. candidate in Biology **Dec 2018 - Current**

Ludwig-Maximilians University, Helmholtz Zentrum München (Germany)

Mentor: Dr. Antonio Scialdone

Master of Science in Bioinformatics **Sep 2016 – Sep 2018**

University of Copenhagen (Denmark)

Bachelor of Science in Biological Sciences

Minor in German Language and Literature **Sep 2011 – Jun 2016**

University of Alberta (Canada)

Research Experience

Ph.D. thesis project **Dec 2018 – Current**

Helmholtz Zentrum München – *Dr. Antonio Scialdone group*

Title: Elucidating Cell Fate Decision Making in Early Mammalian Development through Single Cell Transcriptomics

- Analyzed single-cell transcriptomics of human embryo data in early developmental stages
- Developed an interactive data visualization app to explore human gastrula data : human-gastrula.net
- Collaborated with scientists around the world for various projects
- Worked on different biological systems and organisms, such as human embryo and mouse hematopoietic stem cells
- Developed a pipeline to perform mitochondrial heteroplasmy analysis from single-cell data
- Worked on a mathematical model to describe cell cycle transition to 2c like cells

Master thesis project **Feb 2018 – Aug 2018**

Institute of Theoretical Biology, Humboldt University (Berlin) – *Prof. Dr. Hanspeter Herzel group*
Title: Modelling Gene Regulatory Network of the Mammalian Circadian Clock

- Developed a model to mathematically describe the behavior of core clock network in mouse liver
- Parsed and analyzed various omics datasets, such as RNA-seq, ChIP-seq, proteomics to inform the model

Master's internship project **Mar 2017– Jun 2017**

University of Copenhagen –*Prof. Dr. Albin Sandelin group*

Title: Assessment of Different Mapping Tools on CAGE Data Obtained from RRP40 Knock-Down Experiment

- Developed a pipeline to benchmark various mapping tools on CAGE data

Bachelor's internship project **Sep 2015 – Jan 2016**

University of Alberta – *Laboratory of Dr. Oana Caluseriu*

Title: Identification of Disease-Related Candidate Genes for Congenital Anomalies of Kidney and Urinary Tract (CAKUT) by Whole-Exome Sequencing (WES) Analysis

- Investigated WES data from consanguinous family members with a history of CAKUT
- Developed a pipeline to detect potential CAKUT related DNA variants

---

Publications

- Tyser, R.C.V.*, **Mahammadov, E**. *, Nakanoh, S., Vallier L., Scialdone, A., Srivinas S. **Single-cell transcriptomic characterization of a gastrulating human embryo**. *Nature* 600**,** 285–289 (2021).
- Lima, A., Lubatti, G., Burgstaller, J., Hu, D., Green, A., Di Gregorio, A., Zawadzki, T., Pernaute, B., **Mahammadov, E**., Perez-Montero, S., Dore, M., Sanchez, J., Bowling, S., Sancho, M., Kolbe, T., Karimi, M., Carling, D., Jones, N., Srinivas, S., Scialdone, A. and Rodriguez, T. **Cell competition acts as a purifying selection to eliminate cells with mitochondrial defects during early mouse development**. *Nature Metabolism*, 3(8), 1091-1108 (2021).
- Grabe S., **Mahammadov E.**, D. Olmo M., Herzel H., **Synergies of multiple zeitgebers tune entrainment.** *Frontiers in Network Physiology* 1 (2022).
- Nakatani T., Lin K., Ji F., Ettinger A., Pontabry P., Tokoro M., Altamirano. L., Fiorentino J., **Mahammadov E.**, Hatano Y., Van Rechem C., Chakraborty D., Ruiz-Morales E., Arguello Pascualli P., Scialdone A., Yamagata K., Whetstine J., Sadreyev R., Torres-Padilla ME. **Replication fork speed underlies cell fate changes and promotes reprogramming.** *Nature Genetics* 54, 318-327 (2022).

*co-first author

Selected Talks

- Advances in Single Cell Epigenomics (Saarlouis), 04.11.2019. Title: “A spatially resolved single cell atlas of human gastrulation”
- German Stem Cell Network Conference (Virtual), 23.09.2020. Title: “A spatially resolved single cell atlas of human gastrulation”
- Single Cell Biology Conference, Wellcome Genome (Virtual), 10.11.2020. Title: “A spatially resolved single cell atlas of human gastrulation”

Technical Skills

Programming Languages and Tools

Python (pandas, numpy, scikit learn, seaborn)

R (dplyr, ggplot2, Rmarkdown)

Single cell analysis (scanpy, Seurat, scran, etc.)

Interactive data visualisation (Rshiny, D3js, plotly)

Pipeline development

Research model systems

Human embryonic tissue

Mouse hematopoietic stem cells

Mouse embryonic stem cells

Volunteering & Leadership

Co-creator, writer and editor of elmi-spektr.com
Elmi Spektr (Science Spectrum) – Non-profit Popular Science Magazine in Azerbaijan: Nov 2015 - present

Teacher and mentorship in bioinformatics

National Science Academy of Azerbaijan (NSAA) – Genetic Resources Institute: Jul – Aug 2015

Languages

Azerbaijani (native)

Russian (beginner)

English (fluent)

German (intermediate)

Turkish (fluent)

10. Acknowledgments

I would like to extend my deepest gratitude to everyone who supported me throughout these last few years, both during the ups and downs. First, I would like to thank my direct supervisor Dr. Antonio Scialdone for giving me an opportunity to work on exciting projects in his group. His continuous support passion for science and deep knowledge largely shaped me into the scientist I am today. Also, huge props go to all the former and current members of the group, including but not limited to, Dr. Jonathan Fiorentino, Gabriele Lubatti, and Mayra Ruiz for sharing the ride, and providing support and good laughs along the way. The institute where I completed my PhD, Institute of Epigenetics and Stem cells, whose atmosphere was also the biggest deciding factor for me coming here. I met great people here, including Clara, Manuel, Mrinmoy, Adam, Matthias and many others. The other principal investigators in the institute, Dr. Maria Elena Torres-Padilla and Dr. Stefan Hamperl have taught me a lot as well, whose scientific knowledge and passion will be exemplary for me in the future.

I am also grateful to have had great collaborators, from whom I learned a great deal and had the pleasure of working together. Particularly, I will never forget working with Prof. Shankar Srinivas and Dr. Richard Tyser. They set a high standard for me for finding future collaborators.

I want to thank my Azerbaijani friends living in Europe and elsewhere for never leaving me alone, especially during the pandemic. I owe a great deal to Yashar, Nariman, Sadig, Rashad, Mujgan, Artoghrul, Nazakat, Nazim and Agil for listening to my PhD rants and being supportive. Many of them have hosted me to work from their home whenever I needed company.

Finally, I want to dedicate this thesis to my family, my mother Lala, father Mahammad, brother Orkhan, his wife Aytan, and my beloved sister Royala who have supported my academic endeavor from the very beginning and always trusted me.