Principles of gene and regulatory evolution as inferred from cross-species comparisons in primates

Dissertation von Zane Kliesmete

München 2024

Principles of gene and regulatory evolution as inferred from cross-species comparisons in primates

Dissertation an der Fakultät für Biologie

der Ludwig-Maximilians-Universität München

Zane Kliesmete

München, 2024

Diese Dissertation wurde angefertigt
unter der Leitung von PD Dr. Ines Hellmann
an der Fakultät Biologie
der Ludwig-Maximilians-Universität München

Erstgutachter:	PD Dr. Ines Hellmann
Zweitgutachter:	Professor Dr. Jochen Wolf
Tag der Abgabe:	08.07.2024
Tag der mündlichen Prüfung:	05.12.2024

Eidestattliche Versicherung und Erklärung

Eidesstattliche Erklärung

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

München, den 10. Juni 2024

Zane Kliesmete

Erklärung

Hiermit erkläre ich, dass die Dissertation nicht ganz oder in wesentlichen Teilen einer anderen Prüfungskommission vorgelegt worden ist und dass ich mich anderweitig einer Doktorprüfung ohne Erfolg **nicht** unterzogen habe.

München, den 10. Juni 2024

Zane Kliesmete

Contents

Abbreviations

Abbreviation	Definition
AA	amino acid
ATAC-seq	Assay for Transposase-Accessible Chromatin using sequencing
BC	barcode
BM	Brownian Motion
ChIP-seq	Chromatin Immunoprecipitation Sequencing
CRE	cis-regulatory element
CTCF	CCCTC-binding factor
DNA	deoxyribonucleic acid
DNase-seq	DNase I hypersensitive sites sequencing
dN	non-synonymous
dS	synonymous
ESRG	Embryonic stem cell related gene
HERV	human endogenous retrovirus
iPSC	induced pluripotent stem cells
LINE	long interspersed element
lncRNA	long-non-coding RNA
LTR	long terminal repeat
MPRA	Massively parallel reporter assay
MYA	million years ago
NI	neutrality index
NPC	neural progenitor cells
OU	Ornstein-Uhlenbeck
PD	pleiotropic degree
PGLS	phylogenetic generalized least squares
PSC	pluripotent stem cells
PWM	position weight matrix
RNA	ribonucleic acid
RNA-seq	RNA sequencing
sc	single cell
SINE	non-autonomous short interspersed element
SVA	variable-number tandem-repeat Alu elements
STARR-seq	Self Transcribing Active Regulatory Region sequencing
TAD	topologically associated domain
TE	transposable element
TF	transcription factor
TFBS	transcription factor binding site
TRNP1	TMF-regulated nuclear protein 1
N_e	effective population size
UMI	unique molecular identifier
3D	three-dimensional

Chronological List of Publications

I. Takahashi K, Nakamura M, Okubo C, **Kliesmete Z**, Ohnuki M, Narita M, Watanabe A, Ueda M, Takashima Y, Hellmann I, Yamanaka S:
"The pluripotent stem cell-specific transcript ESRG is dispensable for human pluripotency." (2021)
PLoS genetics 17(5):e1009587.
doi: 10.1371/journal.pgen.1009587

II. **Kliesmete Z**, Wange LE, Vieth B, Esgleas M, Radmer J, Hülsmann M, Geuder J, Richter D, Ohnuki M, Götz M, Hellmann I, Enard W:
"Regulatory and coding sequences of TRNP1 co-evolve with brain size and cortical folding in mammals." (2023)
Elife 12:e83593.
doi: 10.7554/eLife.83593

III. Janssen P, **Kliesmete Z**, Vieth B, Adiconis X, Simmons S, Marshall J, McCabe C, Heyn H, Levin JZ, Enard W, Hellmann I:
"The effect of background noise and its removal on the analysis of single-cell expression data." (2023)
Genome Biology 24(1):140.
doi: 10.1186/s13059-023-02978-x

IV. **Kliesmete Z**, Orchard P, Lee VY, Geuder J, Krauss SM, Ohnuki M, Jocher J, Vieth B, Enard W, Hellmann I:
"Evidence for compensatory evolution within pleiotropic regulatory elements." (2024)
bioRxiv
doi: 10.1101/2024.01.10.575014

Other Publications

VI. Gegenfurtner FA, Zisis T, Al Danaf N, Schrimpf W, **Kliesmete Z**, Ziegenhain C, Enard W, Kazmaier U, Lamb DC, Vollmar AM, Zahler S:
"Transcriptional effects of actin-binding compounds: the cytoplasm sets the tone." (2018)
Cellular and Molecular Life Sciences 75:4539-55.
doi: 10.1007/s00018-018-2919-4

VII. Wang S, Gegenfurtner FA, Crevenna AH, Ziegenhain C, **Kliesmete Z**, Enard W, Muller R, Vollmar AM, Schneider S, Zahler S:
"Chivosazole A Modulates Protein–Protein Interactions of Actin." (2019)
Journal of natural products 82(7):1961-70.
doi: 10.1021/acs.jnatprod.9b00335

VIII. Wang S, Crevenna AH, Ugur I, Marion A, Antes I, Kazmaier U, Hoyer M, Lamb DC, Gegenfurtner F, **Kliesmete Z**, Ziegenhain C:
"Actin stabilizing compounds show specific biological effects due to their binding mode." (2019)
Scientific Reports 9(1):9731.
doi: 10.1038/s41598-019-46282-w

IX. Lousada E, **Kliesmete Z**, Janjic A, Burguiere E, Enard W, Schreiweis C: "Expression profiling of the learning striatum." (2023)
bioRxiv
doi: 10.1101/2023.01.03.522560

X. Edenhofer FC, Térmeg A, Ohnuki M, Jocher J, **Kliesmete Z**, Briem E, Hellmann I, Enard W:
"Generation and characterization of inducible KRAB-dCas9 iPSCs from primates for cross-species CRISPRi." (2024)
iScience
doi: 10.1016/j.isci.2024.110090

Declarations of contribution as a first-author

Evidence for compensatory evolution within pleiotropic regulatory elements

This study was conceived by Ines Hellmann and conducted by Ines Hellmann and me. Peter Orchard did the initial peak calling, the scoring of cis-regulatory element pleiotropic degrees and sequence analyses. Beate Vieth helped him. Johanna Geuder, Simon M. Krauß, Mari Ohnuki and Jessica Radmer generated the data under the supervision of Wolfgang Enard. I did all final analyses on sequence, transcription factor binding and regulatory activity evolution. Victor Yan Kin Lee helped in data analyses. Ines Hellmann and I wrote the manuscript.

According to the regulations for the Cumulative Doctoral Thesis at the Faculty of Biology, LMU München, I confirm the substantial contributions of Zane Kliesmete to this publication.

Ines Hellmann

Regulatory and coding sequences of TRNP1 co-evolve with brain size and cortical folding in mammals

This study was proposed by Magdalena Götz and conceived by Wolfgang Enard and Ines Hellmann. Beate Vieth designed all initial sequence acquisitions. MPRA, RNA-seq and coding-sequence data generation was done by Lucas Esteban Wange, supported by Jessica Radmer, Matthias Hülsmann, Daniel Richter. Primate cell culture was done by Johanna Geuder, Jessica Radmer, Mari Ohnuki. Mouse cell manipulation and the quantification of proliferation was done by Miriam Esgleas. Lucas Esteban Wange and I did primary data processing. I analysed and integrated all data. The manuscript was written by me, Ines Hellmann and Wolfgang Enard.

According to the regulations for the Cumulative Doctoral Thesis at the Faculty of Biology, LMU München, we confirm the substantial contributions of Zane Kliesmete to this publication.

Lucas Esteban Wange

Ines Hellmann

Wolfgang Enard

Declarations of contribution as a co-author

The effect of background noise and its removal on the analysis of single-cell expression data
This study was conceived by Philipp Janssen and Ines Hellmann. I helped with data processing, genotype estimation and deconvolution. Ines Hellmann and Philipp Janssen wrote the manuscript.

The pluripotent stem cell-specific transcript ESRG is dispensable for human pluripotency
This study was conceptualized by Kazutoshi Takahashi and Shinya Yamanaka. I did evolutionary analyses of the ESRG promoter and gene sequence under the supervision of Ines Hellmann. The manuscript was written by Kazutoshi Takahashi.

According to the regulations for the Cumulative Doctoral Thesis at the Faculty of Biology, LMU München, I confirm the above contributions of Zane Kliesmete to these publications.

____

Ines Hellmann

Summary

The sequence contained in the 3.2 Gb long haploid stretches of our DNA has been registered but we are still far from having decoded the information that it contains. Among the approaches that facilitate a closer insight into the relevance of individual elements for existent phenotypes is the comparative approach. It extends beyond focusing solely on one species, instead exploiting the knowledge gained from investigating patterns of evolutionary change. Evolutionary comparisons have an advantage over other techniques that rely on genetic change, in that they inform on the types of changes that have evidently occurred in nature. In this thesis, I bridge advances made in gathering genetic information and in generating high-throughput functional assays in a cross-species context to answer fundamental questions in evolutionary genomics.

To be able to rely on recently developed genome-wide functional assays like RNA-seq, we should know the amount of error that these measurements contain. Using genetic variation between species, I contribute to estimating the precision with which we measure expression. We further evaluate and compare computational methods that are designed to remove this noise, using our substitution-based error estimates as the ground truth.

Then, I study multiple aspects of gene and regulatory evolution by leveraging cross-species data on DNA, expression, accessibility and the activity of regulatory and protein sequences. An important current task in genomics is to improve our ability to read and interpret the regulatory code that governs expression. Therefore, I study how constraint is reflected in a range of functional properties of cis-regulatory elements (CREs), using their tissue-specificity as a proxy for functional importance. Based on theoretical considerations and patterns seen in the case of genes, pleiotropic CREs that are utilized in all or the majority of tissues are expected to be under most constraint. This turns out to be true for the conservation patterns of transcription factor binding site repertoires, whereas the exact binding sites as well as the underlying sequences show even lower conservation than that of tissue-specific CREs. Considering the highly conserved accessibility of pleiotropic CREs and the conserved downstream gene expression, these findings suggest pervasive compensatory evolution acting within the sequences of pleiotropic CREs and, likely, across functionally orthologous tissue-specific CREs. This study underlines the importance to evaluate CRE conservation and functionality using metrics beyond simple sequence conservation.

Further, I touch another currently highly debated aspect of genome evolution: The role of newly evolved elements in species-specific rewiring of gene regulatory networks. Transposable element-derived regulatory and gene sequences are gaining increasing attention due to their ability to expand the genome in a clade- or species-specific manner. In addition, some types of TEs, such as long terminal repeat (LTR) elements, carry binding sites for important transcription factors active in pluripotent stem or other cell types. This leads to new regulatory sequences and transcripts. While some of these have been proposed or indeed shown to contribute to the cellular phenotypes, in the current study we revisit one such candidate long non-coding gene, *ESRG*, and find that in spite of its high expression in human pluripotent stem cells, it is dispensable to the function of these cells. We also find no evidence for selection using sequence divergence and polymorphism-based analyses. This study is a reminder to be careful in interpreting expression as a sign of function.

Finally, I combine evolutionary and functional measures to assess the association between genetic and phenotypic evolution. Specifically, I focus on the association between brain evolution and the evolution of a particular brain developmental gene *TRNP1* across over 30 mammalian species. I find that *TRNP1* coding sequence evolution, *TRNP1*-dependent proliferation rates and the activity of a cis-regulatory element of *TRNP1* co-evolve with brain size and the degree of gyrification. These findings advance our evolutionary and neurodevelopmental understanding of how larger and more folded brains evolve. Moreover, with the increasing availability of high-quality genomes and possibilities to assay genetic variants in massively parallel assays, this and similar studies are demonstrations of how evolutionary information can be leveraged by combining phylogenetic approaches with functional assays.

1 | Introduction

The extraordinary precision and far reach of our ability to process, preserve and spread information is among the key capabilities of the human species that has allowed us to build up on the knowledge of many generations [1,2]. Indeed, everything in the human world increasingly relies on and evolves around it. In the era of information, the genetic code still remains among the most relevant, mysterious and surprising sources of information that contains the full instructions to create a whole organism and to generate offsprings. The genetic code can be seen as a likely accidental, self-preserving machine containing sets of instructions to be executed at a particular time in a particular space. It is written using sequences of a simple 4-letter code which is so powerful that it enables the development of such complex structures as our brains, able to process the intense information flow as it does on an every-day-life basis. Therefore, genetics, and biology in general, have inspired many fields beyond medicine [3,4,5], adding to the motivation to study it.

Among the most challenging tasks in biology is to identify causal variants that are responsible for diverged organismal phenotypes within and across species. This is particularly difficult in higher complexity organisms like mammals, and even more challenging across primates where genetic manipulations are out of discussion due to clear ethical issues. Studying natural genetic variation and selectional signatures across regions of the genome can tell us a lot about how to interpret a certain genetic change [6]. Moreover, approaches that rely on natural genetic and phenotypic variation can be combined with molecular assays and known functional features of the elements. In this thesis, I combine evolutionary, molecular and functional measures to study multiple aspects of genome evolution and how these contribute to the evolution of (molecular) phenotypes. Because I focus on the interpretation of the human genome, information about genetic change is mainly derived from within-human variation and species genetically closely related to the human - other primates. In the following sections I provide background information that is important to understand the context of my work. First, I give an overview on the genetic elements that are relevant for this thesis in the sections **Central regulatory mechanisms governing the tissue-specificity of gene expression** and **Emergence of novel elements through the activity of mobile genetic elements**. In the section **Evolutionary forces shaping genome evolution**, I briefly explain basic population genetic and evolutionary concepts relevant for adequately interpreting genetic change. I also outline the evolutionary modes under which primate genomes generally evolve. In the following chapter, **Studying the mode of evolution in different modalities**, I discuss specific molecular assays and evolutionary measures which, if combined properly, can inform on the type of selection acting on a genetic or molecular element. I discuss important measures of protein, gene expression and regulatory evolution and introduce frameworks using which genotype-phenotype co-evolution can be investigated.

1.1 Central regulatory mechanisms governing the tissue-specificity of gene expression

While all cells of the same multicellular organism contain nearly identical genetic information, distinguished sets of information are utilized in different cells yielding distinct cellular phenotypes. This is enabled through gene expression regulation happening on multiple levels. Firstly, sequence of the particular gene and the genetic elements that regulate its expression have to be accessible to the transcription machinery. This is controlled by epigenetic modifications to DNA, such as DNA methylation that affects gene silencing[7], and modifications to histone proteins that regulate the compactness of the DNA[8,9]. Moreover, the required regulatory RNA and proteins including transcription factors (TFs), co-factors and the right type of RNA-polymerase all have to be present in the cell and bind regulatory DNA to initiate transcription of the gene[10,11] (Figure 1.1). The specificity of gene regulation is to a good part attributed to the specificity of the transcription factors present in the cell and their binding to accessible cis-regulatory elements (CREs), commonly classified as promoters, enhancers, silencers and insulators[12]. Notably, further important regulation of the gene product happens also at post-transcriptional and post-translational stages, however those processes are beyond the scope of this thesis. Altogether these mechanisms control cell fate determination during development through complex interplay between external signals, chromatin remodelling and temporal patterns of regulatory protein activity, affecting expression networks of other genes and properties of the cell.

1.1.1 Trans-regulation by transcription factors

Transcription factors are regulatory proteins that bind cis-regulatory elements, often in the vicinity of a gene, and other transcription factors and co-factors, thereby enabling the positioning of transcription machinery and suitable DNA conformation to induce or suppress expression[13]. The estimated total number of human TFs is ∼1,600[13]. One way to detect their presence in a specific cellular context of interest is by quantifying their expression using RNA-seq[14,15]. TFs tend to be expressed at lower levels than other genes, possibly to establish

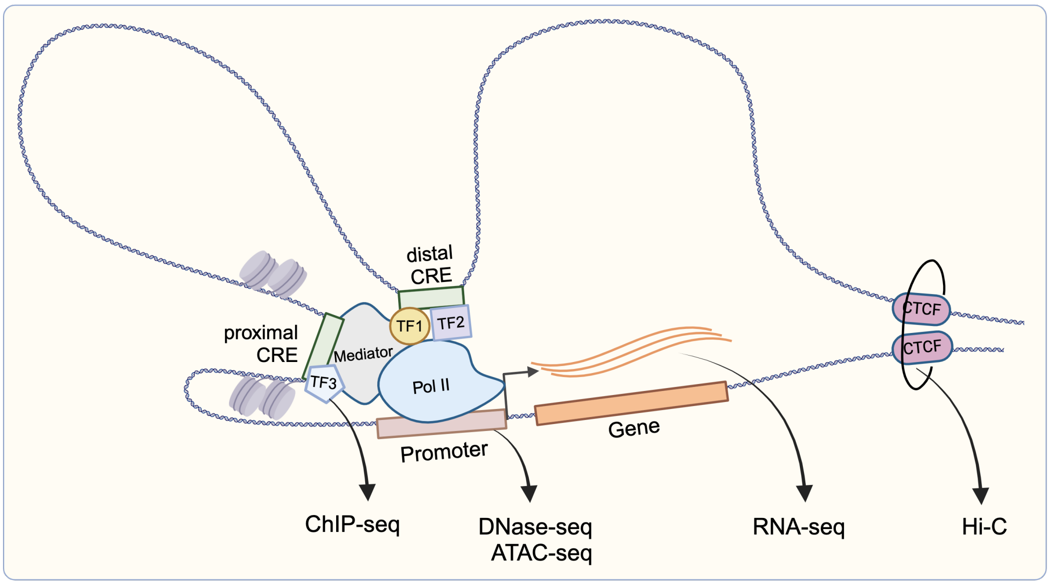


Figure 1.1. Overview of the interplay between chromatin accessibility, transcription factor binding, and transcriptional machinery in regulating gene expression. Proximal and distal cis-regulatory elements (CREs) bind transcription factors (TFs) and co-factors, facilitating the positioning of the transcription machinery near the promoter and transcription start site (TSS) to initiate expression. Pairs of CTCF proteins, in combination with cohesin rings, define the topologically associating domain (TAD) compartments of the DNA, allowing specific regions to interact more frequently within these domains. TF binding can be measured using ChIP-seq targeting the respective TFs. Chromatin accessibility can be measured using DNase-seq and ATAC-seq, while gene expression is commonly quantified using RNA-seq. TAD boundaries and overall 3D genome organization are often measured using Hi-C.

binding specificity[16]. Historically, TFs have been classified as activators or repressors of gene expression[17], however evidence is accumulating that their function is rarely binary as many TFs have been shown to act as repressors or activators in a context-dependent manner[18,19,20]. TFs can bind regulatory sequences as monomers, homo- or hetero-dimers, or multi-mers[21], in the latter cases meaning that they form larger regulatory protein complexes. These complexes can have their own specific effect on gene regulation. Importantly, TFs bind selected, typically ∼6-12 bp long DNA sequences of a particular composition which are embedded in the CREs. In some cases, the preferred binding sequence can have a more strictly defined composition, but empirical experimental findings show frequent binding site degeneracy, i.e. that most TF-DNA interactions are robust to some or even large variation in the binding sequence[22,23]. Moreover, weak binding of TFs to flanking sequences surrounding the target regulatory sequences can also be advantageous for more stable gene expression [24,25].

TFs can be classified into ∼54 families based on their structural and DNA binding site domain similarity[26]. Some classes are associated with particular developmental processes, for example, HOX TF family governs cell fates[27] and the rapidly evolving zinc finger TF family controls the activity of transposable elements[28]. TFs can also be classified by their cellular function. For example, pioneer factors such as MYOD1[29], PAX5[30], FOXA1[31] are known to control or interact with histone modifications and recruit chromatin-remodelling complexes, subsequently initiating chromatin accessibility and the binding of other regulators. Another relevant recently identified group of TFs are the so-called 'Universal stripe factors'[32] that facilitate stable and prolonged CRE accessibility and the binding of other TFs, likely leading to more stable downstream gene expression. A groundbreaking discovery, that also affects the cellular systems used in this thesis, has been the identification of pluripotency factors OCT3/4, SOX2, c-MYC and KLF4 that are sufficient to reprogram differentiated cells into induced pluripotent stem cells (iPSCs)[33]. In summary, the cellular presence and binding of TFs to CREs is central for cell-specific gene regulation. The downstream effects of TF binding does not have a fit-for-all rulebook, instead it depends on the overall combinatorial binding across TFs and the cellular context.

1.1.2 Cis-regulatory elements

Historically, CREs were identified through their proximity to gene sequences and through their sequence conservation, as it is in average higher than that of non-functional sequences but lower than that of protein-coding sequences[34,35]. However, CRE landscapes have proven to be highly dynamic across tissues and developmental stages within and across species[36,37]. The advent of high-throughput assays has enabled more direct ways for active CRE detection in the specific cellular context of interest. DNase I hypersensitive sites sequencing (DNase-seq) [38] and Assay for Transposase-Accessible Chromatin using sequencing (ATAC-seq)[39,40] are among the most popular assays for mapping the location of accessible chromatin, large part of which is accessible because of its regulatory activity. Another, more direct way to identify active or primed CREs is by targeting certain combinations of histone modifications[41] or TF binding using Chromatin immunoprecipitation followed by sequencing (ChIP-Seq)[42,43], however this assay tends to generate larger peaks making it more difficult to identify the exact sequence that is the source of the regulatory activity.

Knowing the complex nature of TF binding, it also comes as no surprise that the regulatory element architectures and sequences that are bound by TFs can be fairly variable. Promoters are the core regulatory units required for trancription initiation, located within few kilobases from or even overlapping the transcription start site(s) of the gene[35]. Promoters that are utilized across a broader range of tissues tend to be large, CpG-rich and often accessible even if the gene is not extensively transcribed[35,44]. A possible reason for the more stable accessibility is to set the baseline requirements ready for the transcription once it is necessary. Tissue-specific promoters tend to be narrower and less CpG-rich, for example TATA-box promoters that regulate tissue-specific, inducible response genes[45]. Enhancers and silencers are more distal regulatory elements that are in average shorter than promoters, often less CpG-rich and accessible in a more cell-type specific fashion[46,47,48]. While enhancers contribute to gene transcription activation, silencers repress transcription initiation. However, quantification of silencer activity is challenging[49,50]. In practice, the distinction between CRE functional classes is often unclear, because the same regulatory sequence can possess the activity of a promoter, enhancer or silencer depending on the cellular context[51,52]. Moreover, CREs often show functional redundancy[53], for example, the activity of some individual enhancers appears to be buffered by equally functional so-called shadow enhancers[54,55].

Experimental evidence suggests that distal enhancers physically contact promoters by chromatin looping during transcription initiation, enabling regulator interactions and better DNA conformation for efficient transcription[56]. The range of possible genomic region interactions are controlled by another important class of CREs called insulators that define the boundaries of a higher-level spatial chromatin organization into context-dependent topologically associated domains (TADs)[57,58,59]. They are bound by a special zinc finger TF called CCCTC-binding factor (CTCF)[60] and thought to contribute to gene expression regulation by limiting the possible interactions between different CREs and genes.

Knowing the boundaries of TADs can also help in narrowing down the possible search space for identifying CRE-to-gene associations. The overall simplest strategy for this task is to associate CREs to genes based on their genomic proximity[61,62], further pruned for being within the same TAD compartment if this information is available. When working with many samples as in the case of single cell data, methods that rely on accessibility and expression co-variation can also be utilized[63,64,65]. Machine learning-based models trained on published data can also be considered for this task[66,67]. Having identified CRE-to-gene associations, various aspects on regulatory principles can be studied.

To summarize, genes and the associated regulatory mechanisms are central to the functionality of the genome. Evolutionary change in their sequence or accessibility can have direct consequences for the phenotype.

1.1.3 Constraint on genetic elements imposed by pleiotropy

The ability of a genetic element to evolve is influenced by the number of molecular contexts in which it is utilized. This phenomenon is widely recognised as pleiotropy, commonly quantified as the number of cell types or tissues in which the element is used. While in some cases better fitness could be achieved by adjusting a particular phenotype, the underlying genetic element(s) may also be essential in another phenotypic context and, hence, lead to detrimental outcome in case of change, thus imposing constraint on its evolution. Therefore, pleiotropic elements are expected to be under more constraint than tissue-specific elements. In recent years, evidence has accumulated that genes with more pleiotropic expression patterns indeed have more conserved protein sequences[68,69]. Moreover, pleiotropic genes also tend to be evolutionary older[70] and show more conserved, i.e. more similar expression patterns across species[71,72,73]. In summary, tissue-specificity is generally a good predictor for constraint[74]. Given this prior information, it is reasonable to assume that also cis-regulatory elements (CREs) that are pleiotropic are on average under more constraint than tissue-specific CREs. However, given that CREs are in many aspects different from genes, it is an open and relevant question how constraint imposed by pleiotropy is reflected in the properties of CREs. Answering this question also brings us closer to understanding the principles that relate CRE sequence evolution to gene expression evolution.

1.2 Emergence of novel elements through the activity of mobile genetic elements

A major contributor to the total content of eukaryotic genomes, and primate genomes in particular, is the activity of mobile genetic elements, also called transposable elements (TEs)[75,76]. TEs can be classified into two main types by their mechanism of activity: DNA-transposons ('cut-and-paste') and RNA-transposons, also called retrotransposons ('copy-and-paste'). It follows that the latter type contributes considerably more to the genome content ($\sim$ 45%)[77]. The main types of retrotransposons of a potential importance for primate evolution due to their abundance are endogenous retroviruses (ERVs), autonomous long interspersed nuclear elements (LINEs), non-autonomous short interspersed nuclear elements (SINEs), that include primate-specific Alu elements (Figure 1.2A), and the primate-specific SINE variable-number tandem-repeat Alu elements (SINE-VNTR-Alu or SVA), which are composites of ERV and Alu elements[78,76,79,80].

In order for these insertions to be heritable, they generally have to happen in the germline or the pre-implantation embryo. Although the identified TE sequences constitute around half of the human genome, by taking into account the likely byproducts of their activity such as pseudogenes and by now unrecognizable products of their ancient activity, their actual contribution might be even around 75%[79]. While most of the fixed TE insertions are thought to be (nearly) neutral, in some cases there can be larger fitness and phenotypic consequences. On one hand, the mutagenic effect of these insertions can have detrimental effects on genome stability and interrupt sequences of functional importance[82]. Additionally, some of these elements are linked to health issues, many of which manifest later in life when stringent regulation does not result in enhanced reproductive success. On the other hand, through the regulatory potential of these sequences, some of the insertions result in novel, potentially functional CREs[83,84] and new genes[85], including regulatory RNAs and expressed chimeric products fused with downstream neighbor sequences called long-non-coding RNAs (lncRNAs)[86] (Figure 1.2C). Indeed, some TE-derived elements have been shown to regulate important processes, including placental development, immune response and the cell-type complexity and signalling in the brain[82]. Some famous functional examples include XIST

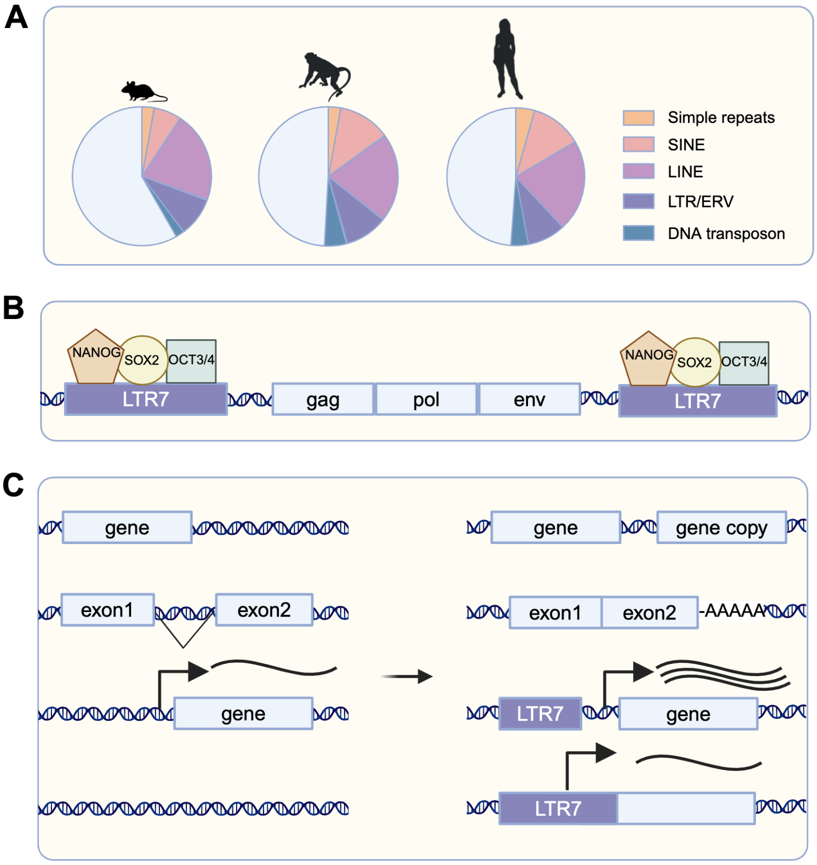


Figure 1.2. **A** Transposable element content in the genomes of a mouse (*Mus musculus*), rhesus macaque (*Macaca mulatta*) and human (*Homo sapiens*)[81,76]. **B** Intact LTR/ERVs consist of *gag*, *pol* and *env* genes, surrounded by two identical LTR elements that function as promoters. The subclass of LTR7 elements have binding sites for important pluripotency TFs including OCT3/4, SOX2 and NANOG. **C** Possible effects of TE jumping and insertions, exemplified for the case of HERVH/LTR7. It can lead to insertions of new copies of a gene, processed pseudogenes with polyA-tails, altered expression of existing genes or the generation of chimeric, expressed products containing parts of the TE and parts of the genomic sequence.

RNA that regulates X chromosome silencing[87,88,89,90], Syncytin genes that allow for nutrient and gas exchange between the mother and the fetus during pregnancy[91,92] and the Arc protein involved in synaptic plasticity[93,94].

Particularly relevant for primate- or human-specific regulatory evolution might be the ERV-derived long terminal repeat (LTR) retrotransposons. They are characterized by identical LTR elements at both ends of the retrotransposal element-specific genes that facilitate transcription of the retroelement and integration of the new copy into the host DNA. ERV/LTR elements are thought to have arisen as a result of ancient and ongoing viral infections of the germline. Estimates based on their age suggest waves of higher activity and fixation around 50 MYA and multiple later waves on the branches leading to Old World monkeys and great apes[79,95]. A relevant subclass are the human ERVs (HERVs), that were found in humans but some of which are also present in other primates. Approximately 8% of the human genome is estimated to stem from HERVs[77], showing great diversity[81]. Different classes carry different sets of transcription factor binding sites in their LTR sequences and show increased regulatory activity in different cell types, such as pluripotent, embryonic endoderm/mesoderm, hematopoetic and immune cells[96]. In the context of pluripotency, particularly relevant are the LTR7 elements that have binding sites for key pluripotency transcription factors (Figure 1.2B). These include OCT3/4 (POU5F1), SOX2, and NANOG, making these elements a potent source of novel CREs[96,97]. A knock-down of LTR7-derived elements was shown to impair cell reprogramming to iPSCs[98]. Some other studies suggest that HERVH-derived enhancers are involved in chromatin opening in human embryonic stem cells (hESCs), followed by the activation of evolutionarily similarly old classes of Krüppel-associated box zinc finger TFs that repress their activity. Later during development, the same enhancers are utilized during cell-type specific differentiation processes[99,100,101,28]. Hence, by some, TE-derived elements are considered a major source of gene regulatory innovations[102,99,103], whereas others have not found evidence for such rewiring[104] or consider it rather an obstacle where a possible rewiring is at most a result of compensation for the disruptions[79]. Whether functional or not, it is clear that TE activity is a major source for the emergence of species-specific elements. Overall, our genome is exposed to factors beyond instantaneous functionality that have and might inevitably happen, all of which can be either removed, coped with, tolerated or utilized[105,100].

1.3 Evolutionary forces shaping genome evolution

If much of the genome is not functional or even slightly disadvantageous, how come it is there and how come the process of genome expansion is happening? The answer becomes obvious through the basic principles of genome evolution. These also elucidate the necessity to interpret the meaning of genetic change cautiously and underlines the importance of carefully designed statistical evolutionary tests of constraint and adaptation that take into account varying population sizes and local mutation rates.

1.3.1 Effective population size

The effective population size (N_e) is among the key factors affecting the relative contributions of selection versus random drift to the genetic material of a species. The types of evolution that can occur within a species depend strongly on it, e.g., if N_e is sufficiently small, the noise associated with random sampling of alleles can influence their fixation probability to the extent as if selection was virtually absent[79]. It is almost always smaller than the total population, partially because 1) some individuals do not contribute to the following generation, 2) sex ratios that deviate from 1:1 reduce N_e, 3) individuals tend to mate with locally related individuals, especially spatial semi-isolated patches decrease the variability. The previous bottlenecks in species variability have large, long lasting effects on its N_e [106,107,108].

Hence, in many cases N_e is considerably smaller than the observed population. For example, the estimated average N_e across vertebrate species is only around 10% of the total number of breeding adults (N)[109,110]. Ignoring this discrepancy would lead to serious errors, while using N_e instead of N has practical advantages as it permits the application of population genetics models that assume the population to behave as an ideal Wright-Fisher population.

1.3.2 Random genetic drift and mutation

The initial spread of all genetic mutations during the first few generations in diploid organisms is mainly determined by drift which is inversely proportional to the population size, i.e. $\frac{1}{2N}$. A certain variant, independently of its long-term neutral, advantageous or disadvantageous effects, has to first survive the stage of being a transient polymorphism. Mutations arise at the frequency of $\mu 2N$ where μ indicates the mutation rate. Hence, in the complete absence of selection, the total rate of fixation of neutral alleles is equal to the product of the number arising per generation and the fixation probability of individual mutations, where both depend on $2N$ in opposing ways[111]. Therefore, the long-term rate of neutral evolution is equal to the genetic mutation rate of the species, e.g.,

$$p_0 = \frac{\mu 2N}{2N} = \mu$$

While the fixation probability of neutral mutations does not depend on N_e, the time that it will take until fixation does, which is in average $4N_e$ generations. This implies that larger N_e results in elevated amount of within-species variation.

To make things more complicated, germline mutation rates vary between different organisms, up to a factor of 40 across vertebrates[112]. Moreover, mutation rate is not uniformly distributed across the genome, instead it depends on multiple local genomic characteristics including structural features such as GC content, chromatin organization, mismatch repair efficiency and recombination rates[113,114,115]. Taking these factors into account is important when looking for selectional signatures between or within species.

1.3.3 Variation in strength and efficacy of selection

Understanding the type of selection acting on genetic elements is a core interest in biology as it informs on their importance and helps connecting genotypes to function and phenotypes. Negative selection, also called purifying selection, constraints the change in functional genetic elements[116]. Given a set of random possible mutations in a functional genetic element, the overall likelihood that many of these will be (weakly) disadvantageous is high, thus decreasing the fixation probability of such changes[117]. In general, disadvantageous mutations are indicated by a negative selection coefficient $s < 0$. The amount of purifying selection tends to be higher in more essential elements, thus serving as an approximation for the importance of the element and as a guidance for prioritising certain genetic regions in case of disease. Overall, non-synonymous sites in protein-coding sequences tend to be under most constraint[118], i.e., most possible changes are associated with relatively large, negative s. For the case of CREs, a few ultra-conserved elements appear to be under strong negative selection[119]. However, most individual CREs might be associated with rather small selection coefficients, possibly due to their redundancy and the potential presence of proto-CREs [120,121], allowing for high turnover rates. Therefore, the majority of protein-coding sequences, including the ones encoding most transcription factors, evolve considerably slower than CRE sequences[118]. The constraint is also generally expected to be higher in pleiotropic elements, functional in multiple cellular or developmental contexts, than in elements that are functional only one or a few contexts[122].

Positive selection, also called directional selection, implies that a genetic change is advantageous for the survival or the reproductive potential of a species and thus have higher fixation probability than a neutral mutation[117], indicated by a positive selection coefficient $s > 0$. A special case of positive selection is the compensatory evolution that can happen because of prior fixation of slightly deleterious alleles[123,124] due to drift or fluctuating selection pressures. Detection of positive selection can help identify the genetic source of species-specific characteristics. Only a few non-synonymous sites of protein-coding sequences appear to evolve under positive selection, but this proportion can vary considerably depending on their structural properties and function[122]. However, even though these events are rare, the selection strength and the resulting adaptive changes can have large phenotypic effects. In comparison, the possibilities for adaptive changes in CREs might be more frequent but each individual change is likely associated with smaller selection coefficients[118]. Hence, overall only a small amount of substitutions is thought to have arisen through positive selection[116]. In addition, it can be challenging to distinguish the actual mutations under positive selection that rapidly increase in frequency through selective sweep from the ones that are dragged along to fixation simply because of being in the genetic vicinity of the selected position and thereby linked, coined as hitchhiking[125].

In general, while the fixation probability of neutral mutations does not depend on N_e, N_e does have an impact on the fixation probability of advantageous and deleterious mutations, i.e., the efficacy of selection after these mutations have survived the first random spread across a few generations. For the scenario where alleles have an additive effect, the selection coefficient is $|s| < 0.1$ (which is mostly the case) and $N_e << N$, the probability of fixation is approximated as follows[126,127]:

$$p_f \approx \frac{(2sN_e/N)}{1 - e^{-4N_e s}}$$

Thus, if the absolute selection coefficient is sufficiently large relative to random drift, e.g., if $4N_e s >> 1$, the fixation probability is different to the neutral expectation by $4N_e s$. This means that the same coefficient s will have a different fixation probability depending on the N_e of the population, in which higher N_e boosts the efficacy of selection[128]. With sufficiently small N_e, the same mutant allele with a certain s can appear nearly identical to a neutral allele, thereby possessing effective neutrality[129,130,131].

1.3.4 Evolutionary modes in primates

Primate, and human evolution in particular, appears to be strongly influenced by non-adaptive evolution, i.e., drift and mutation. The effective population sizes in primates, especially great apes, are estimated to be significantly smaller than the number of breeding individuals. This difference appears to be more drastic than in many other animal clades. These estimates are based on the low amount of polymorphisms within humans[132,133,134,135] and phylogenetic gene comparisons across great apes[136,137]. The reconstructed N_e during human evolution indicates a $\approx$ 10-fold reduction since the common ancestor of humans and chimpanzees [137,138], resulting in the estimated current $N_e \approx 10,000$[139,134,140,138]. In comparison, these numbers are moderately larger for macaques ($N_e \approx 70,000$)[141,138] and much larger for the rodent species ($N_e \approx 450,000 - 820,000$)[142]. This implies that the efficacy of selection in primates is lower than in other species, including model organisms like mice or flies with larger N_e, limiting the fixation of weakly adaptive mutations. Also, neutral mutations get fixed quicker due to drift, therefore the genetic variability is low. Importantly, small N_e also promotes the accumulation of weakly deleterious mutations in short term, altogether leading to an increased fixation of non-adaptive genetic changes including point mutations, gene duplications, TE-derived insertions and even whole chromosome rearrangements as observed for the human. Although potentially disadvantageous at first, in the long run the resultant alterations can give a fertile ground for secondary adaptive or compensatory changes and morphological evolution that is infeasible in large populations[79,143].

1.4 Studying the mode of evolution in different modalities

1.4.1 Quantification of protein evolution rates

Learning to properly read protein-coding sequences and to interpret their change have been important tasks long before any vertebrate genome was fully sequenced[144]. Therefore, these are among the best understood types of elements. Briefly, protein-coding sequences contain codons (triplets of DNA), where each triplet translates to a certain amino acid (AA) in the resulting protein. The beginning, the end and the intron-exon boundaries of the coding sequences are marked by distinct sets of codons or bases[145]. The resulting product can also be detected using the transcribed sequence or the translated protein, altogether alleviating the identification of protein-coding sequences.

It follows that rules for sequence-based quantification of protein evolution rates have also been long studied and certain metrics have been established. The consensus way to quantify the evolutionary rates of a protein is to calculate the ratio of non-synonymous (dN) to synonymous (dS) substitutions, i.e. dN/dS[68,69,146]. Non-synonymous substitutions in the codon sequence change the resulting AA, whereas synonymous substitutions do not affect the AA, thereby serving as a local baseline approximation of the sequence change that is unrelated to the protein evolution. dN/dS values close to 0 indicate slow protein evolution rates (i.e., strong negative selection), while values around 1 suggest nearly-neutral evolution rates and values > 1 suggest the possibility of positive selection[147,148].

Given this useful metric, many protein sequences have been screened across the mammalian or the primate phylogeny[149,150] using maximum likelihood-based phylogenetic frameworks of which the most frequently used is PAML[151,149]. Multiple assumptions regarding the evolution of the protein across the phylogeny need to be made, including but not limited to the presence of a molecular clock, codon frequency and AA distance matrix. In general, most screened proteins appear to evolve under strong negative selection, yielding an overall dN/dS close to zero[152,118]. If positive selection happens, for most proteins it tends to be concentrated in certain functional domains and only stands out when specifically looking at the particular positions using the so-called site model. It can also be limited to certain branches of the phylogenetic tree (branch model, branch-site model). Different tests have been implemented to test the likelihood of these alternative hypotheses relative to a respective null hypothesis[153]. However, these tests are rather conservative and underpowered, detecting adaptive evolution only if dN is higher than dS. Therefore, the case where positive selection has happened in only few lineages and few AAs is difficult to detect[153], particularly between closely related species where the accumulation of synonymous mutations might also be low. Another, less conservative metric for detecting positive or negative selection in protein-coding sequences is the McDonald-Kreitman test[154] (Figure 1.3) which also relies on dN and dS, but adds interpretability and better control over varying N_e by normalizing the cross-species estimates by the within-species variability, i.e. polymorphisms π_N and π_S, at the same sites. The null hypothesis is that the ratio of non-synonynonymous to synonymous variation within species is the same as between species. This can be summarized using a neutrality index:

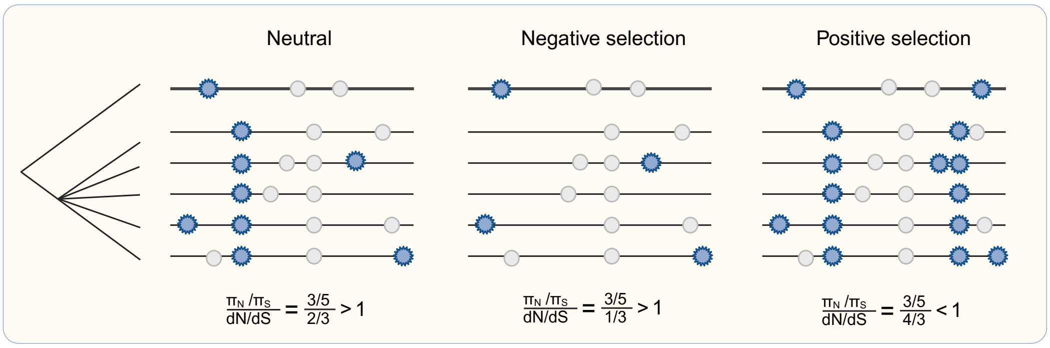

Neutral
Negative selection
Positive selection
$\frac{\pi_N/\pi_S}{dN/dS} = \frac{3/5}{2/3} > 1$
$\frac{\pi_N/\pi_S}{dN/dS} = \frac{3/5}{1/3} > 1$
$\frac{\pi_N/\pi_S}{dN/dS} = \frac{3/5}{4/3} < 1$

Figure 1.3. A visual depiction of McDonald-Kreitman test where an outgroup species is used to estimate the number of fixed substitutions across non-synonymous (dN) and synonymous (dS) sites, while the number of polymorphisms (π_N, π_S) within the same functional categories provide an estimate for the variation within the species.

$$NI = \frac{\pi_N/\pi_S}{dN/dS}$$

An NI>1 indicates negative selection, whereas NI<1 positive selection. Limitations of this approach are related to the fact that the levels of polymorphisms might be influenced by demographic effects or weak negative selection[155,156], thereby breaking the assumption of the approximate neutrality.

Although informative, the evolution rate alone does not tell much about the potential role of the protein changes for phenotypic evolution. If the trait of interest can be quantified as a categorical or continuous characteristic of the species, methods that jointly reconstruct genetic and phenotypic ancestral states across a phylogeny can be used to infer their correlation. Bayesian phylogenetic approaches such as Coevol[157] have an advantage over frequentist methods that the uncertainty in the reconstructed values can be carried along during reconstruction and taken into account when inferring the probability of co-evolution. Clearly, such methods need to be codon- and phylogeny-aware. In theory, different assumptions about the correlation structure between species can be made. The currently implemented (and simplest) correlation structure is the Brownian Motion (Figure 1.4A) as it is aligns with the assumptions of normality[158] and implies that the substitution rates and the continuous traits are evolving with no mean shift and variance σ^2.

While these metrics are well established, they depend heavily on the assumption that the alignments are correct. A small misalignment can cause a frame shift, thereby changing the whole interpretation of the resulting translated protein sequence. Hence, establishing the correct exon-intron boundaries in different species at least for the consensus transcripts is a central and non-trivial task. Currently, many individual studies simply focus on the more easily alignable proteins or subparts of the protein[150,159], which might bias the analysis towards lower dN/dS. Long-read sequencing technologies and recent additional efforts like the mammalian consortium Zoonomia[160] or Vertebrate Genomes Project[161] are important to improve the quality of such analyses.

Beyond comparing protein-coding sequences, functional properties of the orthologous proteins can also be compared using phylogenetic frameworks such as their activity. For this, a thoughtful design of the functional assay is necessary that fits the function of the protein in the relevant cellular system.

1.4.2 Quantification of gene expression divergence

A large part of the diversity observed on the phenotypic level across species is thought to be attributable to differential regulation of genes, particularly during the development[162,10,163]. To investigate a dynamic process in cellular systems or tissues, the compared cell states and the mixture of cell types at hand need to be orthologous between the species[164]. While this is a valid consideration also for comparisons between different human individuals, generation of comparable cross-species systems is even more challenging due to potential systematic differences in differentiation speed, cell type diversity and the fact that most workflows and reagents are generally optimized for human (or mouse) samples. Hence, systematic technical differences that might be associated with the cross-species approach need to be eliminated during the experimental part of the research or later computationally.

Technological advances facilitating high-throughput measurements of gene expression enable quantitative comparisons of the whole transcriptome between different species in tissues or cell types of interest. The most recent and widely used technique is RNA-seq[165,14,15] that involves converting RNA to complementary DNA (cDNA), adding a sample- or cell-specific barcode and amplifying the cDNA molecules to improve their detection. RNA-seq can be used to capture the full sequence of the transcript or, for the sake of sequencing costs, can be targeted to capture only the 5' or 3' end of the transcribed sequence[166].

After gene expression in orthologous cell types has been quantified, statistical approaches that enable unbiased cross-species investigation of gene expression evolution are necessary. Assuming that one-to-one orthologous genes are of interest (and readily identified), simple differential expression analysis[167,168] can be used for comparisons that involve only two or a few species that are evolutionarily similarly distant from each other[169,170]. The simplest type of analysis measures the absolute expression differences between species. This can be used to compare groups of genes, however requires a particularly carefully designed experimental setup as the comparison can be influenced by technical batches that coincide with the species origin of the samples[171]. In addition, the biological interpretation of the absolute expression differences can be difficult even under well controlled conditions. Comparisons of relative expression change across conditions such as a differentiation timeline can be interpreted as differential regulation between species, thus often more informative. Differential regulation can be inferred by specifying an interaction term between species and time or condition. In general, measurements from multiple different individuals within each species should be used to establish a baseline variation, explained by factors other than species divergence. Mixed effects models[172] offer an extension to account for clustered expression data.

To compare expression differences across many species with large divergence times, more sophisticated evolutionary models designed for modelling continuous trait evolution are appropriate. Here, species topology and thereby their non-independence is explicitly taken into account through inclusion of a correlation structure. Popular approaches are phylogenetic ANOVA[173] and phylogenetic regression models[174], including phylogenetic generalized least squares[175,176]. Evolutionary modes can be investigated by comparing the likelihoods or performing an F-test between models that make different assumptions about the expression variance across evolutionary time[158,177,178,72,179]. Random drift is commonly modelled using Brownian motion (BM) where the rate of expression change per unit of time is constant and shows no directionality (Figure 1.4A). As evolutionary time goes to infinity, expression values can be modelled as normally distributed with mean θ and variance σ^2, e.g., $X \sim \mathcal{N}(\theta, \sigma^2)$. Ornstein Uhlenbeck (OU) process can be used to model drift and stabilizing

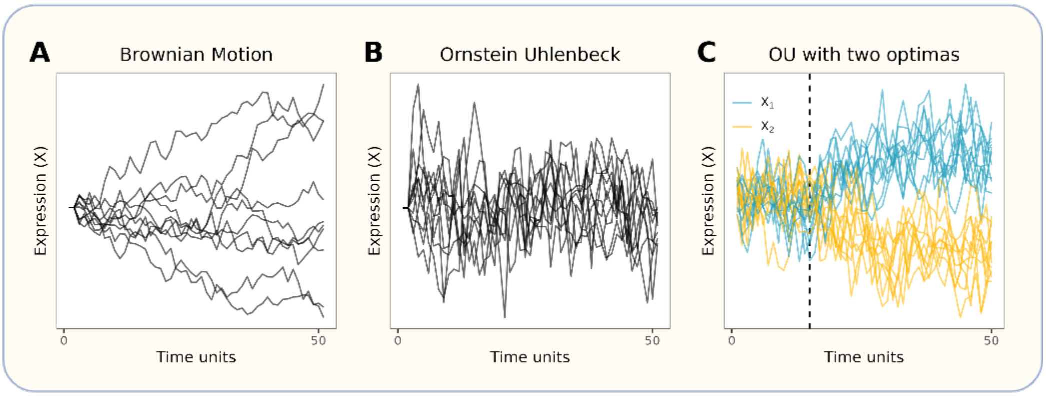


Figure 1.4. Evolutionary models used for modeling continuous trait evolution across time. **A** Brownian Motion is used to model drift, e.g., $X \sim \mathcal{N}(\theta, \sigma^2)$. **B** Ornstein Uhlenbeck process can be used to model negative selection with a 'pull' parameter α describing the selection strength towards the optimum, e.g., $X \sim \mathcal{N}(\theta, \sigma^2/2\alpha)$. **C** Ornstein Uhlenbeck process with two optimas that can be used to model directional selection, e.g., $X_1 \sim \mathcal{N}(\theta_1, \sigma^2/2\alpha)$, $X_2 \sim \mathcal{N}(\theta_2, \sigma^2/2\alpha)$.

(negative) selection where α indicates the selection strength towards the optimum, e.g., $X \sim \mathcal{N}(\theta, \sigma^2/2\alpha)$ (Figure 1.4B). Furthermore, multivariate OU can be applied to test for directional selection on certain branches of the phylogeny by assuming simultaneous existence of different optimal values for different clades, e.g., $X_1 \sim \mathcal{N}(\theta_1, \sigma^2/2\alpha)$, $X_2 \sim \mathcal{N}(\theta_2, \sigma^2/2\alpha)$ [179] (Figure 1.5C). Moreover, because these frameworks are based on regression, they can also be used to measure whether certain covariates can explain some of the remaining variation. Important to note, the more complex the alternative hypothesis, the more species need to be included to have sufficient statistical power. Although overall still simplistic[180,181], these models are useful tools to approximate evolutionary modes of continuous traits like expression.

1.4.3 Measuring cis-regulatory element evolution

To understand the sources of expression patterns across species, we need to be able to read the regulatory code that governs expression in orthologous cell stages, much of which is contained in cis-regulatory elements. The ultimate function of CREs is to enable the binding of the context-relevant TF, co-factors and transcription machinery, contributing to the required DNA conformation that facilitates expression of the associated gene(s)[35,13].

As noted previously, the basic rules for interpreting protein-coding sequence change are rather clear, where discrete changes in the DNA sequence result in known, discrete changes in the protein sequence. In contrast, establishing rules for CRE sequence change appears to be a more complex task due to their inherent flexibility in position and combinatorial usage that can be different for the same gene depending on the cellular context. Moreover, not all nucleotides within a CRE are equally functional - the sequence at some positions of TF binding motifs do not seem to matter[182,183] and TF motif orientation, spacing and composition can be rather flexible - often it is enough with cooperative binding of multiple relevant TFs[184,185,186,187,188,189]. Hence, TF binding potential to a CRE can be seen as a continuous property and certain sequence change does not have discrete nor linear effects on TF binding or expression in most cases. Therefore, CRE evolution rates should be investigated on multiple functional levels beyond sequence, including their transcription factor binding site (TFBS) repertoire, position and regulatory activity conservation (Figure 1.5).

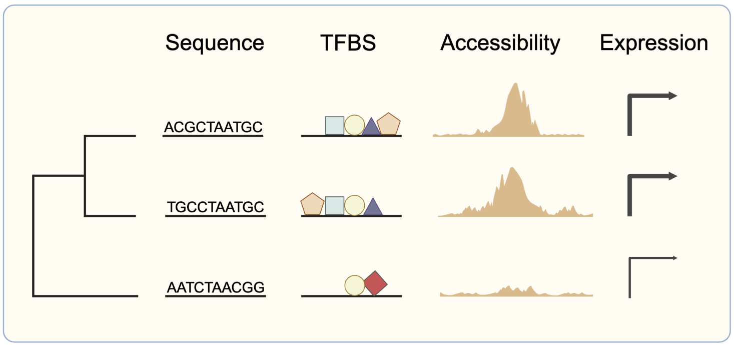


Figure 1.5. Different functional levels that can be used and combined to understand the evolutionary modes of cis-regulatory elements.

Sequence To quantify regulatory sequence evolution and compare it across different regions of the genome, the local mutation rate variation that depends on GC content, chromatin organization, mismatch repair efficiency and recombination rates should be incorporated in the statistical frameworks[113,114,115]. The detection of selectional signatures in the sites of interest can be improved by contrasting them to the evolution rates of nearby putatively-neutrally evolving sites[190,191,192,193]. Moreover, principles similar to McDonald-Kreitman test can be applied also for regulatory sequences where signatures of recent natural selection and effects of random drift are disentangled using patterns of polymorphism and divergence[194,192].

Transcription factor binding sites TFBS repertoires are quantified by scoring the match between position weight matrices (PWMs) of the screened TFs to the CRE sequence of interest. Selecting only expressed TF motifs can increase the interpretability for a specific cellular context. To estimate the specificity of a PWM-CRE match, the base composition of the respective CRE and the nearby flanking sequences can serve as a baseline [195,196,197]. However, this type of contrast is still up to debate, as some studies have indicated functional relevance of surrounding weak binding sites in flanking sequences for attracting TFs and helping stabilizing TF binding to the strongest sites[24,25]. Methods that evaluate the strength of motif clusters, composed of multiple different binding sequences, can be used to estimate the total regulatory potential of the CRE. Another, more specific experimental way to investigate TF binding is through ChIP-seq screens targeting TFs[198,199]. However, the availability of such cross-species assays is currently still limited to only few tissues, species and TFs[200,198,102,199]. For pairwise comparisons across species, simple distance-based metrics can be used for the calculation of binding divergence. In the case of multiple species comparisons, evolutionary models intended for continuous traits that account for species divergence can be applied here similarly to expression[158].

Regulatory activity Finally, a comparison of CRE regulatory activity across species is a powerful approach to establish CRE functional divergence. This level of characterization can also be useful to connect different sequence or TFBS properties to functional evolution and thereby study the rules of the regulatory code. An indirect strategy to quantify regulatory activity is by associating CREs to their putative target genes and modeling the observed expression conservation of the respective gene using the different features of the CRE landscape as predictors[62,201,104,202,52]. Since CREs are investigated in their natural genomic location, allowing for their correct 3D interaction with each other and cellular regulators, this analysis can yield meaningful interpretation of their combinatorial effects on gene expression. On the other hand, it is difficult to distinguish the contributions of individual CREs in modulating gene expression.

A direct way to quantify regulatory change of individual CREs between species or conditions is to assay orthologous CRE activity in the selected cell type(s) of different species. Historically, this was done using reporter assays where the CRE of interest and a fluorescent or luminescent reporter gene, typically luciferase, are cloned into a plasmid[203]. Here, CRE activity is reflected in the amount of the reporter gene product. More recent approaches enable quantification of the activity of thousands of CREs simultaneously. Massively Parallel Reporter Assay (MPRA), which relies on barcode detection to quantify the activity[204,48], was used for the evolutionary analyses in this thesis. It requires *in silico* synthesis of the CRE sequences that is currently limited to a length of around 300 bases. This is below the average size of enhancers (∼420 bp) and especially promoters, that can be up to three times larger[205,206,207].

Thus, tiling of the CREs is necessary. The final activity of a CRE can be calculated by summarizing across the tiles covering it, however such back-calculated additive activity is not necessarily the same as the activity of the full, intact CRE. In addition, the process of DNA synthesis is costly, limiting the number of CREs that can be assayed. However, the sequence synthesis step of an MPRA has the advantage that the effect of a specific sequence change at specific positions, e.g., an *in silico* mutation, can also be measured in a specific cellular context of interest[208].

1.5 Aims of this thesis

In this thesis, by combining evolutionary, molecular and functional measures, I aim to contribute to answering questions related to the following ongoing research:

1. Estimation of the amount of error in expression measurements using RNA-seq
2. Tissue-specificity of regulatory elements and how it relates to functional importance and evolutionary constraint
3. The role of recently evolved elements in species-specific rewiring of gene regulatory networks
4. The association between genotype with a phenotype of interest across a phylogeny.

This work should be informative for domain specialists interested in the specific case studies that are included in this thesis. It could also be interesting to molecular and evolutionary biologists in a broader sense as it touches a range of generally relevant aspects of genome evolution and possible ways to study it.

2 | Results

2.1 The effect of background noise and its removal on the analysis of single-cell expression data

Janssen P, **Kliesmete Z**, Vieth B, Adiconis X, Simmons S, Marshall J, McCabe C, Heyn H, Levin JZ, Enard W, Hellmann I:

"The effect of background noise and its removal on the analysis of single-cell expression data."

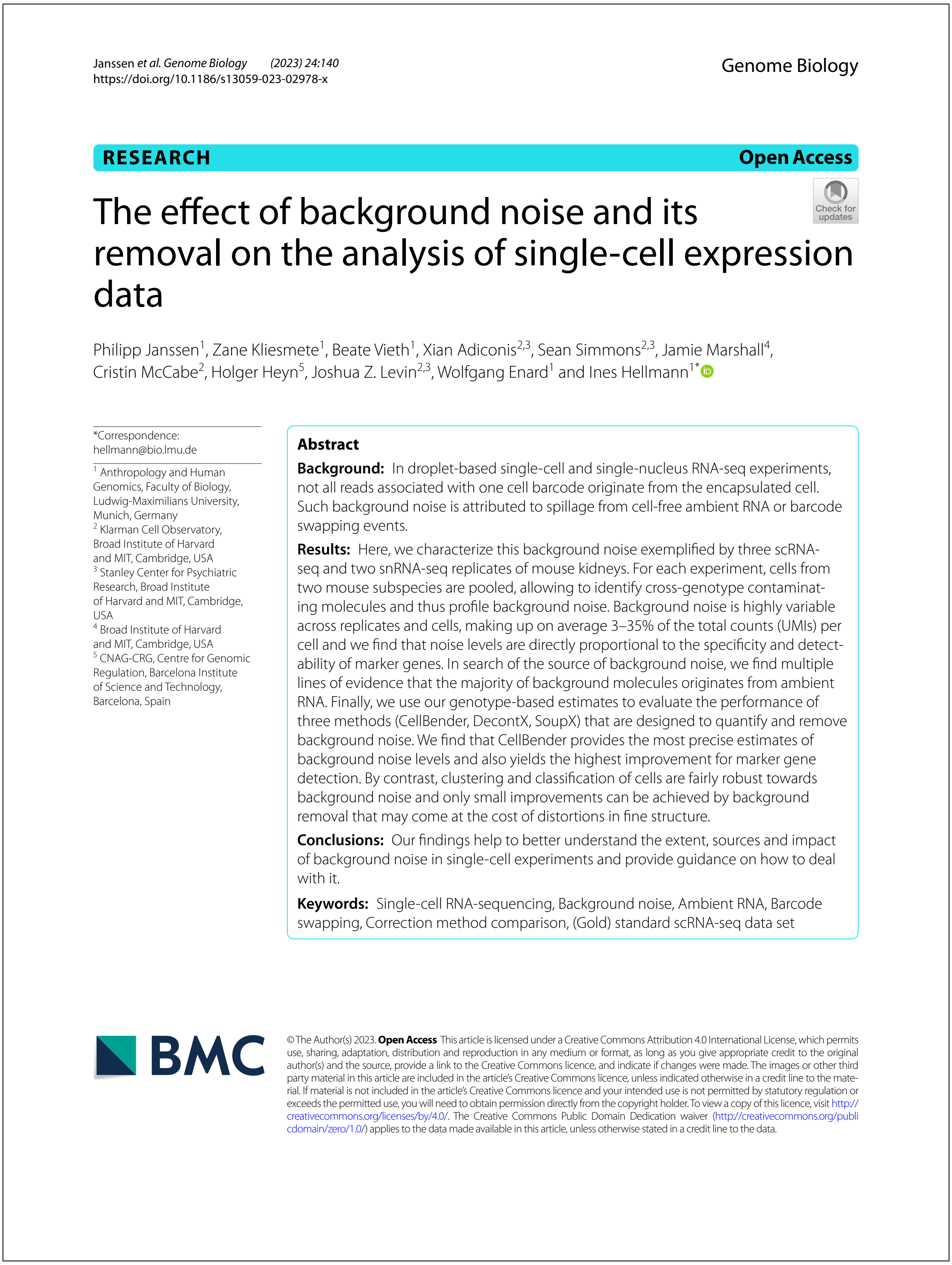
(2023)

Genome Biology 24(1):140.

doi: 10.1186/s13059-023-02978-x

Supplementary Information is freely available at the publisher’s website:

`https://genomebiology.biomedcentral.com/articles/10.1186/s13059-023-02978-x`



Janssen *et al. Genome Biology* (2023) 24:140
https://doi.org/10.1186/s13059-023-02978-x

Genome Biology

RESEARCH **Open Access**

The effect of background noise and its removal on the analysis of single-cell expression data

Philipp Janssen[1], Zane Kliesmete[1], Beate Vieth[1], Xian Adiconis[2,3], Sean Simmons[2,3], Jamie Marshall[4], Cristin McCabe[2], Holger Heyn[5], Joshua Z. Levin[2,3], Wolfgang Enard[1] and Ines Hellmann[1*]

*Correspondence:
hellmann@bio.lmu.de

[1] Anthropology and Human Genomics, Faculty of Biology, Ludwig-Maximilians University, Munich, Germany
[2] Klarman Cell Observatory, Broad Institute of Harvard and MIT, Cambridge, USA
[3] Stanley Center for Psychiatric Research, Broad Institute of Harvard and MIT, Cambridge, USA
[4] Broad Institute of Harvard and MIT, Cambridge, USA
[5] CNAG-CRG, Centre for Genomic Regulation, Barcelona Institute of Science and Technology, Barcelona, Spain

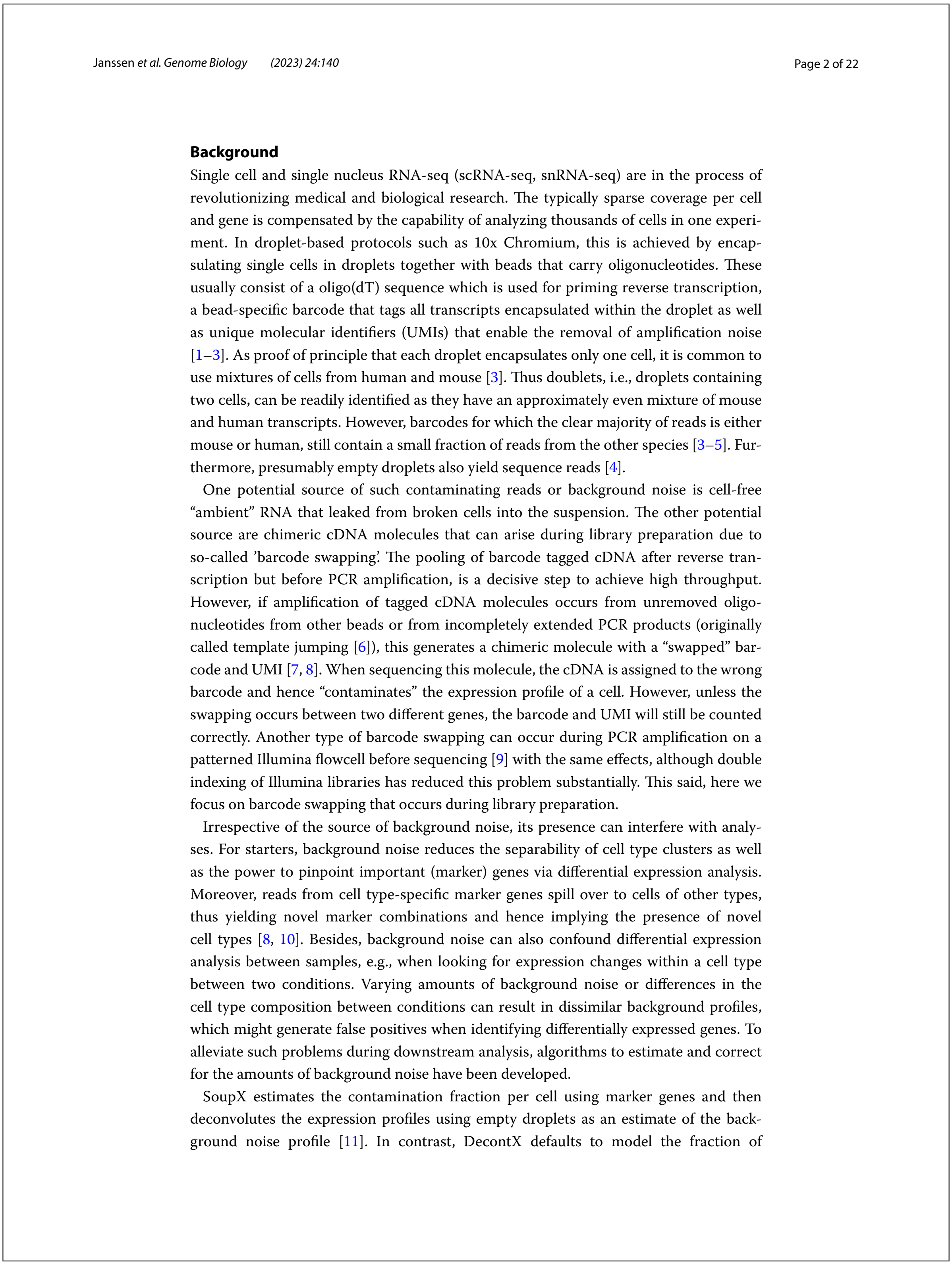
Abstract

Background: In droplet-based single-cell and single-nucleus RNA-seq experiments, not all reads associated with one cell barcode originate from the encapsulated cell. Such background noise is attributed to spillage from cell-free ambient RNA or barcode swapping events.

Results: Here, we characterize this background noise exemplified by three scRNA-seq and two snRNA-seq replicates of mouse kidneys. For each experiment, cells from two mouse subspecies are pooled, allowing to identify cross-genotype contaminating molecules and thus profile background noise. Background noise is highly variable across replicates and cells, making up on average 3–35% of the total counts (UMIs) per cell and we find that noise levels are directly proportional to the specificity and detectability of marker genes. In search of the source of background noise, we find multiple lines of evidence that the majority of background molecules originates from ambient RNA. Finally, we use our genotype-based estimates to evaluate the performance of three methods (CellBender, DecontX, SoupX) that are designed to quantify and remove background noise. We find that CellBender provides the most precise estimates of background noise levels and also yields the highest improvement for marker gene detection. By contrast, clustering and classification of cells are fairly robust towards background noise and only small improvements can be achieved by background removal that may come at the cost of distortions in fine structure.

Conclusions: Our findings help to better understand the extent, sources and impact of background noise in single-cell experiments and provide guidance on how to deal with it.

Keywords: Single-cell RNA-sequencing, Background noise, Ambient RNA, Barcode swapping, Correction method comparison, (Gold) standard scRNA-seq data set



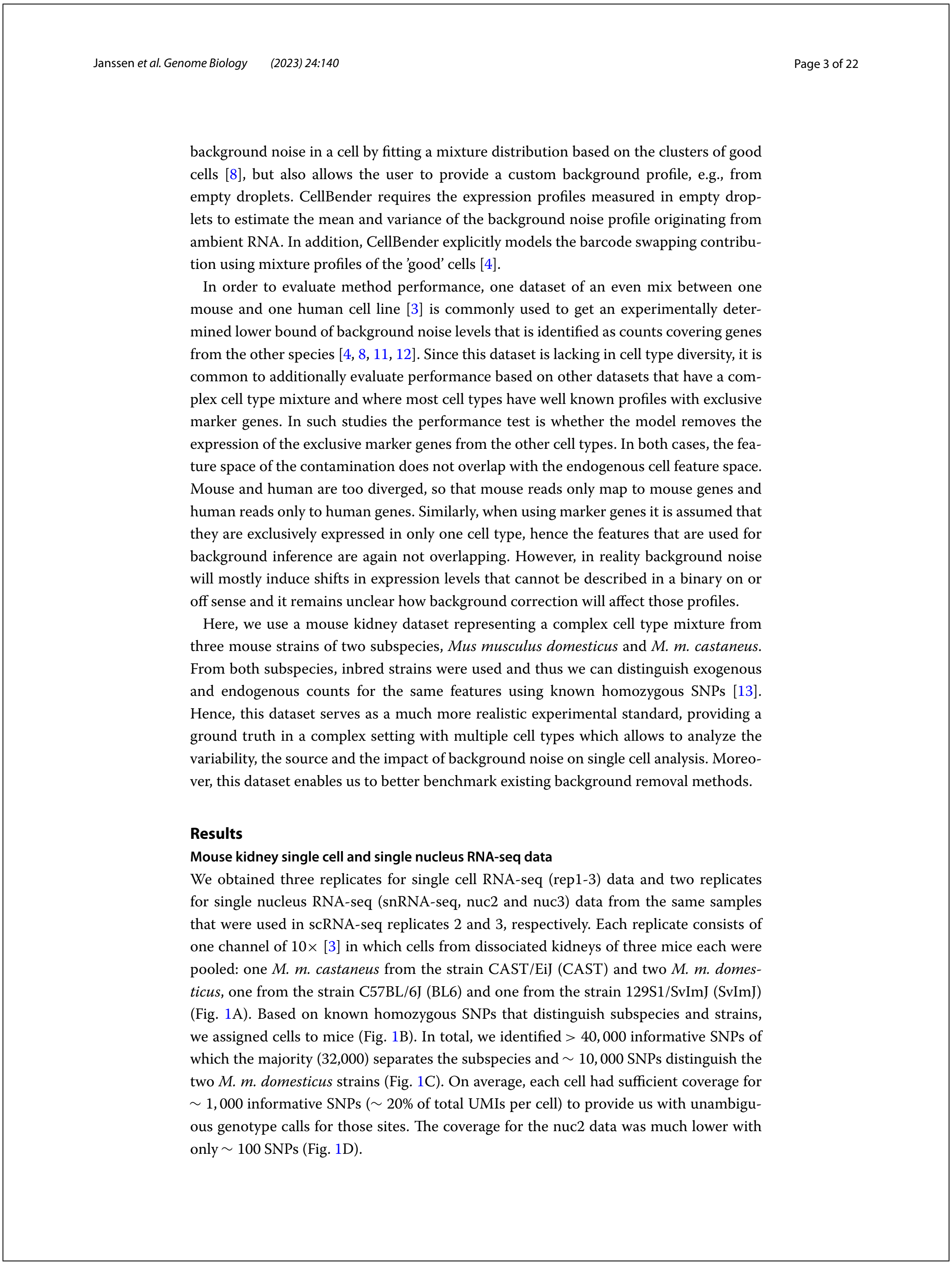
Background

Single cell and single nucleus RNA-seq (scRNA-seq, snRNA-seq) are in the process of revolutionizing medical and biological research. The typically sparse coverage per cell and gene is compensated by the capability of analyzing thousands of cells in one experiment. In droplet-based protocols such as 10x Chromium, this is achieved by encapsulating single cells in droplets together with beads that carry oligonucleotides. These usually consist of a oligo(dT) sequence which is used for priming reverse transcription, a bead-specific barcode that tags all transcripts encapsulated within the droplet as well as unique molecular identifiers (UMIs) that enable the removal of amplification noise [1–3]. As proof of principle that each droplet encapsulates only one cell, it is common to use mixtures of cells from human and mouse [3]. Thus doublets, i.e., droplets containing two cells, can be readily identified as they have an approximately even mixture of mouse and human transcripts. However, barcodes for which the clear majority of reads is either mouse or human, still contain a small fraction of reads from the other species [3–5]. Furthermore, presumably empty droplets also yield sequence reads [4].

One potential source of such contaminating reads or background noise is cell-free “ambient” RNA that leaked from broken cells into the suspension. The other potential source are chimeric cDNA molecules that can arise during library preparation due to so-called ’barcode swapping’. The pooling of barcode tagged cDNA after reverse transcription but before PCR amplification, is a decisive step to achieve high throughput. However, if amplification of tagged cDNA molecules occurs from unremoved oligonucleotides from other beads or from incompletely extended PCR products (originally called template jumping [6]), this generates a chimeric molecule with a “swapped” barcode and UMI [7, 8]. When sequencing this molecule, the cDNA is assigned to the wrong barcode and hence “contaminates” the expression profile of a cell. However, unless the swapping occurs between two different genes, the barcode and UMI will still be counted correctly. Another type of barcode swapping can occur during PCR amplification on a patterned Illumina flowcell before sequencing [9] with the same effects, although double indexing of Illumina libraries has reduced this problem substantially. This said, here we focus on barcode swapping that occurs during library preparation.

Irrespective of the source of background noise, its presence can interfere with analyses. For starters, background noise reduces the separability of cell type clusters as well as the power to pinpoint important (marker) genes via differential expression analysis. Moreover, reads from cell type-specific marker genes spill over to cells of other types, thus yielding novel marker combinations and hence implying the presence of novel cell types [8, 10]. Besides, background noise can also confound differential expression analysis between samples, e.g., when looking for expression changes within a cell type between two conditions. Varying amounts of background noise or differences in the cell type composition between conditions can result in dissimilar background profiles, which might generate false positives when identifying differentially expressed genes. To alleviate such problems during downstream analysis, algorithms to estimate and correct for the amounts of background noise have been developed.

SoupX estimates the contamination fraction per cell using marker genes and then deconvolutes the expression profiles using empty droplets as an estimate of the background noise profile [11]. In contrast, DecontX defaults to model the fraction of



background noise in a cell by fitting a mixture distribution based on the clusters of good cells [8], but also allows the user to provide a custom background profile, e.g., from empty droplets. CellBender requires the expression profiles measured in empty droplets to estimate the mean and variance of the background noise profile originating from ambient RNA. In addition, CellBender explicitly models the barcode swapping contribution using mixture profiles of the 'good' cells [4].

In order to evaluate method performance, one dataset of an even mix between one mouse and one human cell line [3] is commonly used to get an experimentally determined lower bound of background noise levels that is identified as counts covering genes from the other species [4, 8, 11, 12]. Since this dataset is lacking in cell type diversity, it is common to additionally evaluate performance based on other datasets that have a complex cell type mixture and where most cell types have well known profiles with exclusive marker genes. In such studies the performance test is whether the model removes the expression of the exclusive marker genes from the other cell types. In both cases, the feature space of the contamination does not overlap with the endogenous cell feature space. Mouse and human are too diverged, so that mouse reads only map to mouse genes and human reads only to human genes. Similarly, when using marker genes it is assumed that they are exclusively expressed in only one cell type, hence the features that are used for background inference are again not overlapping. However, in reality background noise will mostly induce shifts in expression levels that cannot be described in a binary on or off sense and it remains unclear how background correction will affect those profiles.

Here, we use a mouse kidney dataset representing a complex cell type mixture from three mouse strains of two subspecies, *Mus musculus domesticus* and *M. m. castaneus*. From both subspecies, inbred strains were used and thus we can distinguish exogenous and endogenous counts for the same features using known homozygous SNPs [13]. Hence, this dataset serves as a much more realistic experimental standard, providing a ground truth in a complex setting with multiple cell types which allows to analyze the variability, the source and the impact of background noise on single cell analysis. Moreover, this dataset enables us to better benchmark existing background removal methods.

Results

Mouse kidney single cell and single nucleus RNA-seq data

We obtained three replicates for single cell RNA-seq (rep1-3) data and two replicates for single nucleus RNA-seq (snRNA-seq, nuc2 and nuc3) data from the same samples that were used in scRNA-seq replicates 2 and 3, respectively. Each replicate consists of one channel of 10× [3] in which cells from dissociated kidneys of three mice each were pooled: one *M. m. castaneus* from the strain CAST/EiJ (CAST) and two *M. m. domesticus*, one from the strain C57BL/6J (BL6) and one from the strain 129S1/SvImJ (SvImJ) (Fig. 1A). Based on known homozygous SNPs that distinguish subspecies and strains, we assigned cells to mice (Fig. 1B). In total, we identified $> 40,000$ informative SNPs of which the majority (32,000) separates the subspecies and $\sim 10,000$ SNPs distinguish the two *M. m. domesticus* strains (Fig. 1C). On average, each cell had sufficient coverage for $\sim 1,000$ informative SNPs ($\sim$ 20% of total UMIs per cell) to provide us with unambiguous genotype calls for those sites. The coverage for the nuc2 data was much lower with only ~ 100 SNPs (Fig. 1D).

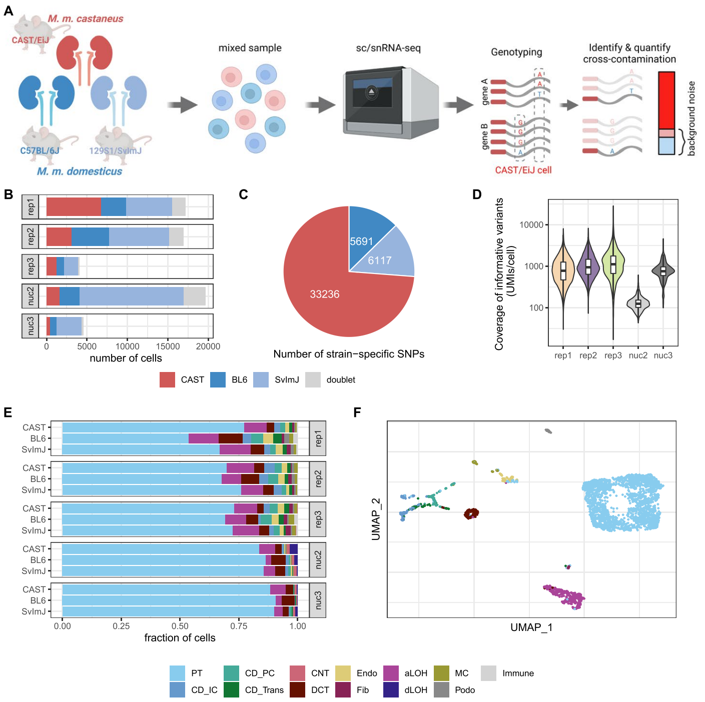


Fig. 1 Generation of mouse strain mixture datasets to quantify background noise. **A** Experimental design (created with BioRender.com). **B** Strain composition in 5 different replicates, subjected to scRNA-seq (rep1-3) or snRNA-seq (nuc2, nuc3). The replicates rep2 and nuc2 and rep3 and nuc3 were generated from the same samples each. CAST: CAST/EiJ strain; BL6: C57BL/6J strain; SvImJ: 129S1/SvImJ. **C** Number of homozygous SNPs with a coverage of more than 100 UMIs that distinguish one strain from the other two. **D** Per cell coverage in *M. m. castaneus* cells of informative variants that distinguish *M. m. castaneus* and *M. m. domesticus*. **E** Cell type composition per replicate and strain; labels were obtained by reference-based classification using mouse kidney data from Denisenko et al. [14] as reference. **F** UMAP visualization of *M. m. castaneus* cells in single-cell replicate 2, colored by assigned cell type. PT, proximal tubule; CD_IC, intercalated cells of collecting duct; CD_PC, principal cells of collecting duct; CD_Trans, transitional cells of collecting duct; CNT, connecting tubule; DCT, distal convoluted tubule; Endo, endothelial; Fib, fibroblasts; aLOH, ascending loop of Henle; dLOH, descending loop of Henle; MC, mesangial cells; Podo, podocytes

Overall, each experiment yielded 5000–20,000 good cells with 9–43% *M. m. castaneus* (Fig. 1B). Thus, the majority of background noise in any *M. m. castaneus* cell is expected to be from *M. m. domesticus* (Additional file 1: Fig. S1B) and therefore we expect that genotype-based estimates of cell-wise amounts of background noise for *M. m. castaneus* to be fairly accurate (Additional file 1: Fig. S2). Hence from here on out we focus on *M. m. castaneus* cells for the analysis of the origins of background noise and also as the ground truth for benchmarking background removal methods.

This dataset has two advantages over the commonly used mouse-human mix [3]. Firstly, the kidney data have a high cell type diversity. Using the data from Denisenko et al. [14] as reference dataset for kidney cell types, we could identify 13 cell types.

Encouragingly, the cell type composition is very similar across mouse strains as well as replicates with proximal tubule cells constituting 66–89% of the cells (Fig. 1E, F; Additional file 1: Fig. S3). Secondly, due to the higher similarity of the mouse subspecies, we can identify contaminating reads for the same features. $\sim 7,000$ genes carry at least one informative SNP about the subspecies. Because so many genes have informative SNPs, the fraction of UMIs that cover an informative SNP is a little higher for PTs, the most frequent cell type, but very comparable across all other cell types, allowing us to quantify contaminating reads (Additional file 1: Fig. S1A).

Background noise fractions differ between replicates and cells

Around 5–20% of the UMI counts are from molecules that contain a SNP that is informative about the subspecies of origin. We quantify in each *M. m. castaneus* cell how often an endogenous *M. m. castaneus* allele or a foreign *M. m. domesticus* allele was covered. Assuming that the count fractions covering the SNPs are representative of the whole cell, we detect a median of 2–27% counts from the foreign genotype over all cells per experiment (Additional file 1: Fig. S1C). This observed cross-genotype contamination fraction represents a lower bound of the overall amounts of background noise. As suggested in Heaton et al. [15], we then integrate over the foreign allele fractions of all informative SNPs to obtain a maximum likelihood estimate of the background noise fraction (ρ_{cell}) of each cell that extrapolates to also include contamination from the same genotype (see the "Methods" section, Additional file 1: Fig. S2). Based on these estimates, we find that background noise levels vary considerably between replicates and do not appear to depend on the overall success of the experiment measured as the cell yield per lane (Fig. 2). For example in scRNA-seq rep3 (3900 cells), we detected overall the fewest good cells, but most of those cells had less than 3% background noise, while the much more successful rep2 (15,000 cells) we estimated the median background noise level at around 11% (Fig. 2A). This said, the snRNA-seq data generated from frozen tissue have much higher background levels than the corresponding scRNA-seq replicates — 35% in nuc2 vs. 11% rep2 and 17% in nuc3 vs. 3% in rep3. How we define good cells based on the UMI counts has little impact on this variability. We still find by far the highest background levels in nuc2 and the lowest in rep3 (Additional file 1: Fig. S4). This high variability is not very surprising. This being a real life experiment and experimental conditions were improved for nuc3 based on the experience with nuc2 (see the "Methods" section). The

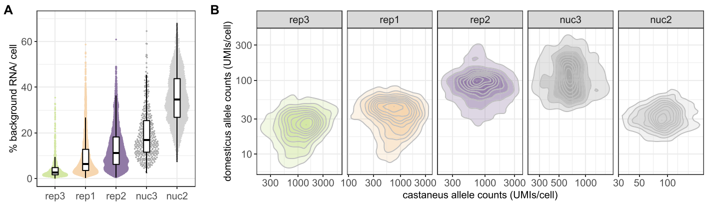


Fig. 2 The level of background noise is variable across replicates and single cells. **A** Estimated fraction of background noise per cell. The replicates on the *x*-axis are ordered by ascending median background noise fraction. **B** In *M. m. castaneus* cells both endogenous *M. m. castaneus* specific alleles (*x*-axis) and *M. m. domesticus* specific alleles (*y*-axis) have coverage in each cell. The detection of *M. m. domesticus* specific alleles can be seen as background noise originating from cells of a different mouse

number of contaminating RNA-molecules (UMIs) depends only weakly on the total UMI counts covering informative variants as a proxy for sequencing depth of the cell (Fig. 2B, Additional file 1: Table S1). Such a weak correlation could be explained by variation in the capture efficiency in each droplet. An alternative, but not mutually exclusive explanation of such a correlation could be that the source of some contaminating molecules is barcode swapping that can occur during library amplification.

However, by and large the absolute amount of background noise is approximately constant across cells and thus the contamination fraction mainly depends on the amount of endogenous RNA: the larger the cell, the smaller the fraction of background noise, pointing towards ambient RNA as the major source of the detected background (Fig. 2B).

Contamination profiles show a high similarity to ambient RNA profiles

In order to better understand the effects of background noise, it is helpful to understand its origins and composition. To this end, we constructed profiles representing endogenous, contaminating and ambient expression profiles by using *M. m. domesticus* allele counts in *M. m. domesticus* cells (endogenous), *M. m. domesticus* allele counts in *M. m. castaneus* cells (contamination) and *M. m. domesticus* allele counts in empty droplets (empty) (Fig. 3A , B; Additional file 1: Fig. S5A-E).

The number of contaminating UMI counts per cell is at a similar level as the UMI counts in empty droplets in all replicates (Fig. 3C, Additional file 1: Fig. S5F). Moreover, if the median UMI count in empty droplets is high for one replicate, we also observe

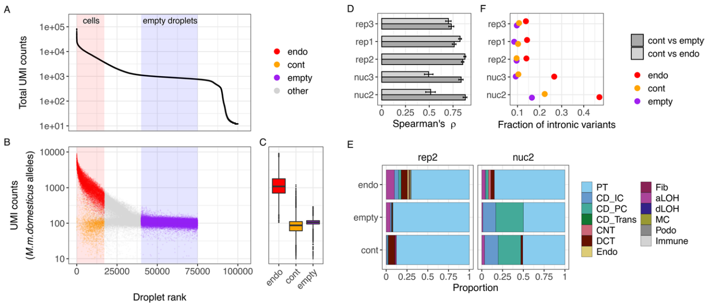
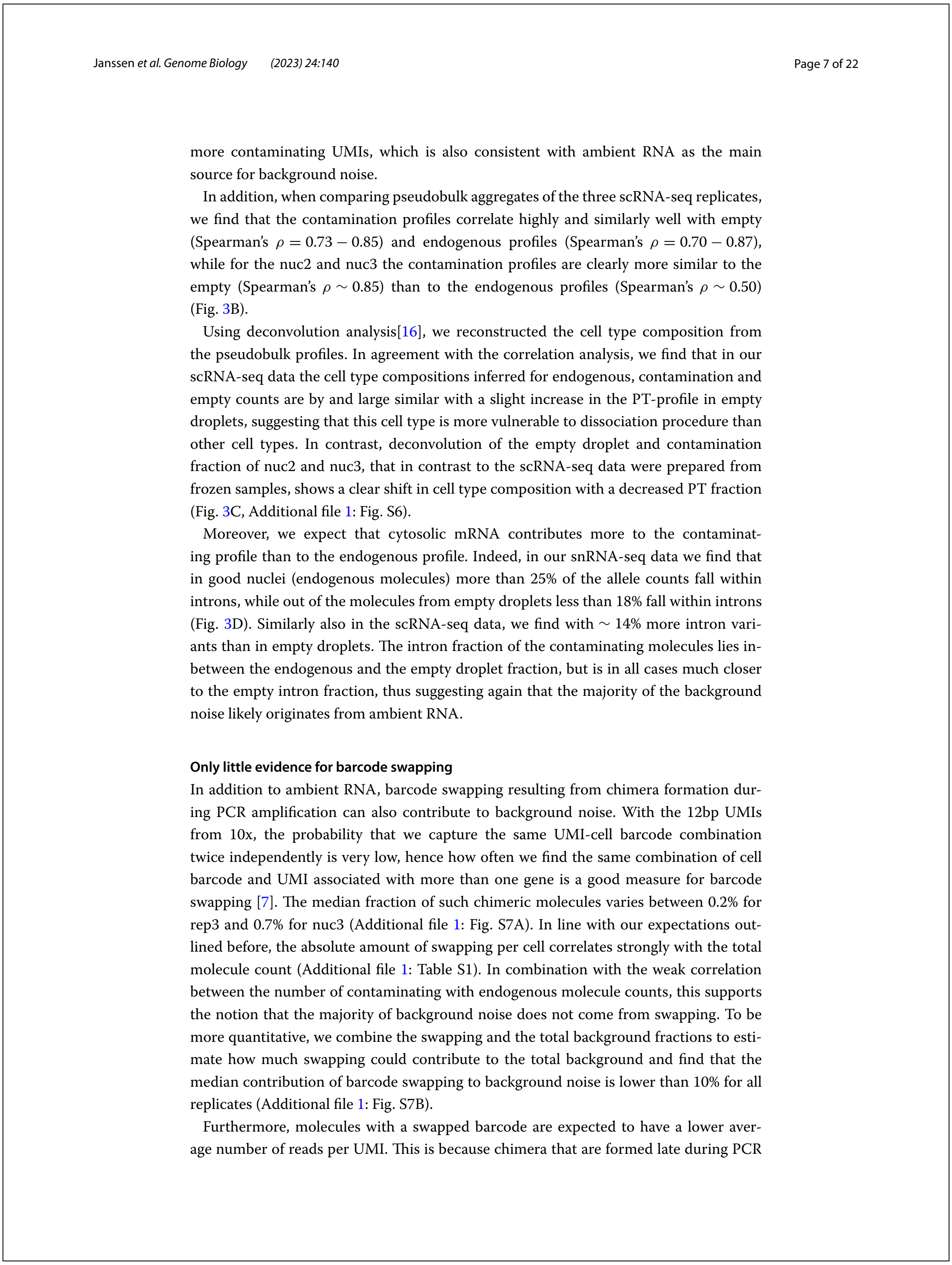


Fig. 3 Characterization of ambient RNA in cells and empty droplets. **A** Ordering droplet barcodes by their total UMI count to distinguish cell-containing droplets with high UMI counts from empty droplets that only contain cell-free ambient RNA and are identifiable as a plateau in the UMI curve, shown here for replicate 2. **B** UMI counts of reads covering *M. m. domesticus* specific alleles were used to construct three profiles depending on whether they were associated with *M. m. domesticus* cell barcodes (endogenous counts, endo), *M. m. castaneus* cell barcodes (contaminating counts, cont) or empty droplet barcodes (empty). Counts from droplets that are not clearly assignable as cell-containing or empty were excluded from further analysis (other). **C** UMI counts per cell for each of the three profiles. **D** Spearman rank correlation between pseudobulk profiles. Error bars indicate 95% confidence intervals obtained by bootstrapping over genes. **E** Deconvolution of cell type contributions to each pseudobulk profile, exemplified by replicates rep2 and nuc2. The stacked barplots depict the estimated fraction of each cell type in the profile as inferred by SCDC using the annotated single cell data of each replicate as reference. PT, proximal tubule; CD_IC, intercalated cells of collecting duct; CD_PC, principal cells of collecting duct; CD_Trans, transitional cells of collecting duct; CNT, connecting tubule; DCT, distal convoluted tubule; Endo, endothelial; Fib, fibroblasts; aLOH, ascending loop of Henle; dLOH, descending loop of Henle; MC, mesangial cells; Podo, podocytes. **F** Fraction of reads covering intronic variants in each of the three profiles



more contaminating UMIs, which is also consistent with ambient RNA as the main source for background noise.

In addition, when comparing pseudobulk aggregates of the three scRNA-seq replicates, we find that the contamination profiles correlate highly and similarly well with empty (Spearman's $\rho = 0.73 - 0.85$) and endogenous profiles (Spearman's $\rho = 0.70 - 0.87$), while for the nuc2 and nuc3 the contamination profiles are clearly more similar to the empty (Spearman's $\rho \sim 0.85$) than to the endogenous profiles (Spearman's $\rho \sim 0.50$) (Fig. 3B).

Using deconvolution analysis[16], we reconstructed the cell type composition from the pseudobulk profiles. In agreement with the correlation analysis, we find that in our scRNA-seq data the cell type compositions inferred for endogenous, contamination and empty counts are by and large similar with a slight increase in the PT-profile in empty droplets, suggesting that this cell type is more vulnerable to dissociation procedure than other cell types. In contrast, deconvolution of the empty droplet and contamination fraction of nuc2 and nuc3, that in contrast to the scRNA-seq data were prepared from frozen samples, shows a clear shift in cell type composition with a decreased PT fraction (Fig. 3C, Additional file 1: Fig. S6).

Moreover, we expect that cytosolic mRNA contributes more to the contaminating profile than to the endogenous profile. Indeed, in our snRNA-seq data we find that in good nuclei (endogenous molecules) more than 25% of the allele counts fall within introns, while out of the molecules from empty droplets less than 18% fall within introns (Fig. 3D). Similarly also in the scRNA-seq data, we find with $\sim 14\%$ more intron variants than in empty droplets. The intron fraction of the contaminating molecules lies in-between the endogenous and the empty droplet fraction, but is in all cases much closer to the empty intron fraction, thus suggesting again that the majority of the background noise likely originates from ambient RNA.

Only little evidence for barcode swapping

In addition to ambient RNA, barcode swapping resulting from chimera formation during PCR amplification can also contribute to background noise. With the 12bp UMIs from 10x, the probability that we capture the same UMI-cell barcode combination twice independently is very low, hence how often we find the same combination of cell barcode and UMI associated with more than one gene is a good measure for barcode swapping [7]. The median fraction of such chimeric molecules varies between 0.2% for rep3 and 0.7% for nuc3 (Additional file 1: Fig. S7A). In line with our expectations outlined before, the absolute amount of swapping per cell correlates strongly with the total molecule count (Additional file 1: Table S1). In combination with the weak correlation between the number of contaminating with endogenous molecule counts, this supports the notion that the majority of background noise does not come from swapping. To be more quantitative, we combine the swapping and the total background fractions to estimate how much swapping could contribute to the total background and find that the median contribution of barcode swapping to background noise is lower than 10% for all replicates (Additional file 1: Fig. S7B).

Furthermore, molecules with a swapped barcode are expected to have a lower average number of reads per UMI. This is because chimera that are formed late during PCR subsequently undergo less amplification [7]. Thus, if the majority of contaminating reads were to originate from barcode swapping, we would expect that the distribution of reads per UMI for cross-genotype contaminating molecules (cont) is similar to that of observed chimeras. This is not what we see (Additional file 1: Fig. S7C): The distribution of reads per UMI for contaminating reads is much more distinct from the distribution for chimeras (Kolmogorov-Smirnov distance, $\Delta_n = 0.381$ (rep3) to 0.595 (nuc3)) than for endogenous reads ($\Delta_n = 0.008$ (rep2) to 0.046 (rep3)). In summary, we find that barcode swapping during library preparation only contributes little to the overall background noise in this data.

The impact of contamination on marker gene analyses

The ability to distinguish hitherto unknown cell types and states is one of the greatest achievements made possible by single cell transcriptome analyses. To this end, marker genes are commonly used to annotate cell clusters for which available classifications appear insufficient. An ideal marker gene would be expressed in all cells of one type but in none of the other present cell types. Thus, when comparing expression levels of one cell type versus all others, we expect high log2-fold changes, the higher the change the more reliable the marker. However, such a reliance on marker genes also makes this type of analysis vulnerable to background noise. Our whole kidney data can illustrate this problem well, because with the very frequent proximal tubular (PT) cells we have a dominant cell type for which rather specific marker genes are known [17]. Slc34a1 encodes a phosphate transporter that is known to be expressed exclusively in PT cells [18, 19]. As expected, it is expressed highly in PT cells, but it is also present in a high fraction of other cells (Fig. 4A, E; Additional file 1: Fig. S8). Moreover, the log2-fold changes of Slc34a1 are smaller in replicates with larger background noise, indicating that

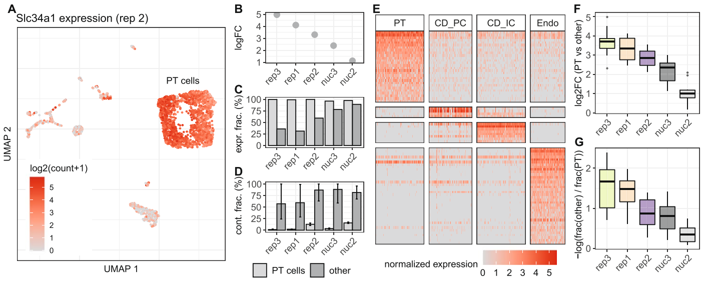


Fig. 4 Background noise affects differential expression and specificity of cell type specific marker genes. **A** UMAP representation of replicate 2 colored by the expression of Slc34a1, a marker gene for cells of the proximal tubule (PT). Besides high counts in a cluster of PT cells, Slc34a1 is also detected in other cell type clusters. Differential expression analysis between PT and all other cells shows a decrease of the detected log fold change of Slc34a1 (**B**) at higher background noise levels, as well as an increase of the fraction of non PT cells in which UMI counts of Slc34a1 were detected (**C**). **D** Estimation of the background noise fraction of Slc34a1 expression indicates that the majority of counts in non PT cells originates from background noise. Error bars indicate 90% profile likelihood confidence intervals. **E** Heatmap of marker gene expression for four cell types in replicate 2, downsampled to a maximum of 100 cells per cell type. **F** Comparison across replicates of log2 fold changes of 10 PT marker genes calculated based on the mean expression in PT cells against mean expression in all other cells. **G** For the same set of genes as in **E**, the log ratio of fraction of cells in which a gene was detected in others and PT cells shows how specific the gene is for PT cells

the detection of Slc34a1 in non-PT cells is likely due to contamination (Fig. 4B–D). We observe the same pattern for other marker genes as well: they are detected across all cell types (Fig. 4E, Additional file 1: Fig. S9) and an increase of background noise levels goes along with decreasing log2-fold changes and increasing detection rates in other cell types (Fig. 4F,G). Thus, the power to accurately detect marker genes decreases in the presence of background noise.

Benchmark of background noise estimation tools

Given that background noise will be present to varying degrees in almost all scRNA-seq and snRNA-seq replicates, the question is whether background removal methods can alleviate the problem without the information from genetic variants. SoupX [11], DecontX [16] and CellBender [4], all provide an estimate of the background noise level per cell. Here, we use our genotype-based background estimates as ground truth to compare it to the estimates of the three background removal methods (Fig. 5A, Additional file 1: Fig. S10). All methods have adjustable parameters, but also provide a set of defaults. For CellBender the user can adjust the nominal false positive rate to put a cap on losing information from true counts. For SoupX and DecontX the resolution of the clustering of cells that is later used to model the endogenous counts can be adjusted. In addition, SoupX can be provided with an expected background level and for DecontX the user can provide a custom background profile rather than using the

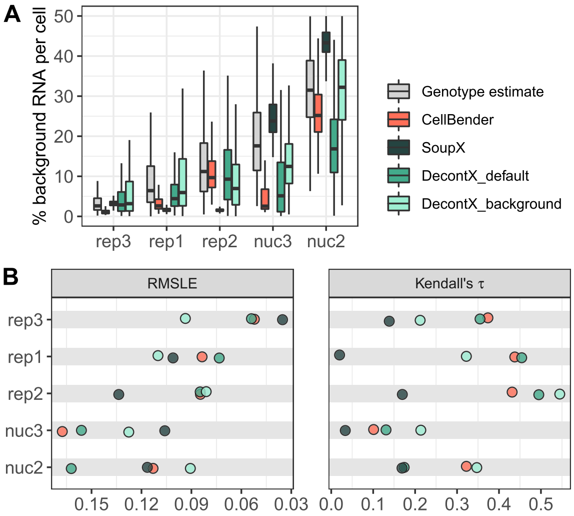
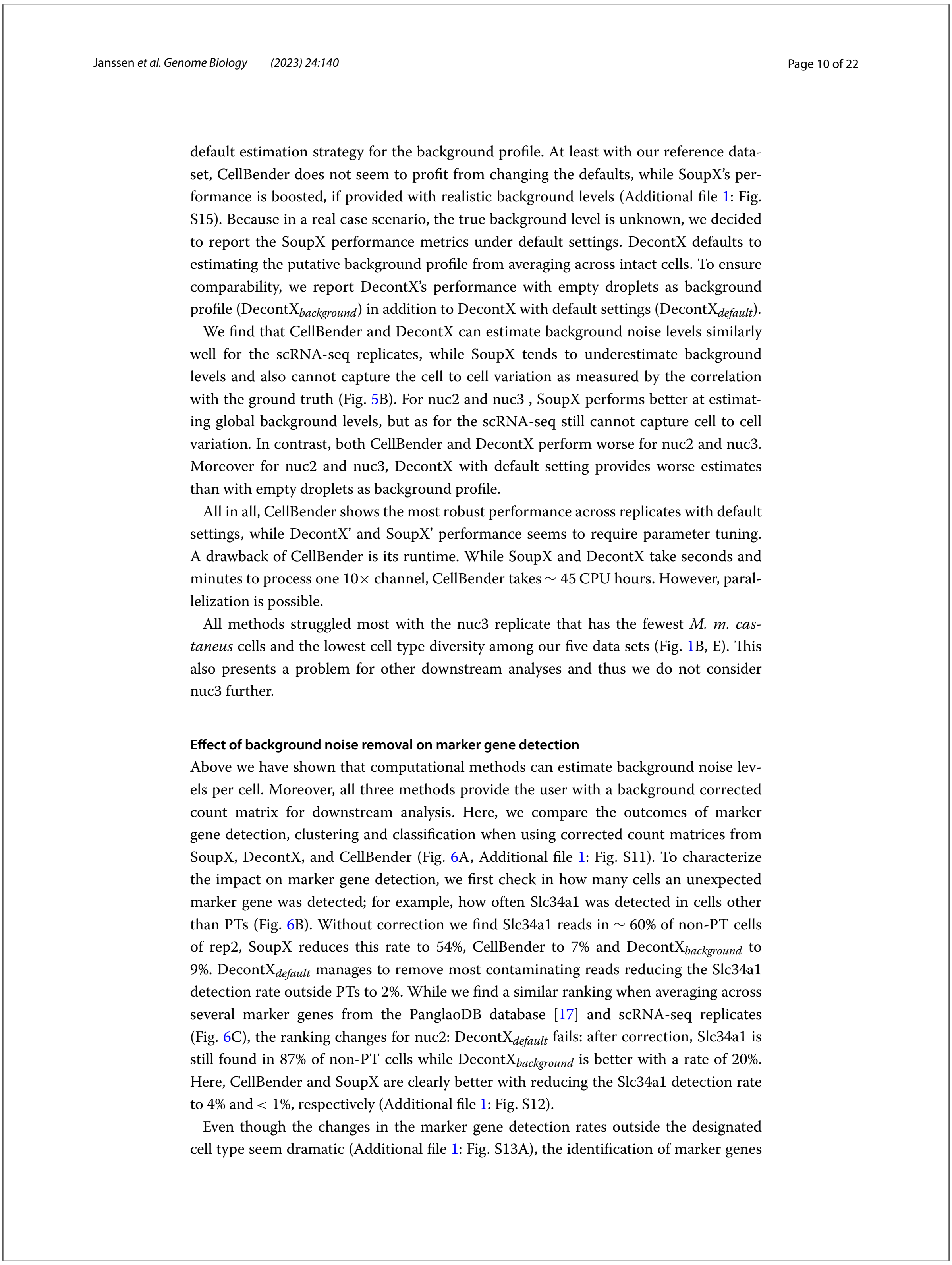


Fig. 5 Accuracy of computational background noise estimation. **A** Estimated background noise levels per cell based on genetic variants (gray) and different computational tools. **B** Taking the genotype-based estimates as ground truth, Root Mean Squared Logarithmic Error (RMSLE) and Kendall rank correlation serve as evaluation metrics for cell-wise background noise estimates of different methods. Low RMSLE values indicate high similarity between estimated values and the assumed ground truth. High values of Kendall's τ correspond to good representation of cell to cell variability in the estimated values



default estimation strategy for the background profile. At least with our reference dataset, CellBender does not seem to profit from changing the defaults, while SoupX's performance is boosted, if provided with realistic background levels (Additional file 1: Fig. S15). Because in a real case scenario, the true background level is unknown, we decided to report the SoupX performance metrics under default settings. DecontX defaults to estimating the putative background profile from averaging across intact cells. To ensure comparability, we report DecontX's performance with empty droplets as background profile (DecontX$_{background}$) in addition to DecontX with default settings (DecontX$_{default}$).

We find that CellBender and DecontX can estimate background noise levels similarly well for the scRNA-seq replicates, while SoupX tends to underestimate background levels and also cannot capture the cell to cell variation as measured by the correlation with the ground truth (Fig. 5B). For nuc2 and nuc3 , SoupX performs better at estimating global background levels, but as for the scRNA-seq still cannot capture cell to cell variation. In contrast, both CellBender and DecontX perform worse for nuc2 and nuc3. Moreover for nuc2 and nuc3, DecontX with default setting provides worse estimates than with empty droplets as background profile.

All in all, CellBender shows the most robust performance across replicates with default settings, while DecontX' and SoupX' performance seems to require parameter tuning. A drawback of CellBender is its runtime. While SoupX and DecontX take seconds and minutes to process one 10× channel, CellBender takes $\sim$ 45 CPU hours. However, parallelization is possible.

All methods struggled most with the nuc3 replicate that has the fewest *M. m. castaneus* cells and the lowest cell type diversity among our five data sets (Fig. 1B, E). This also presents a problem for other downstream analyses and thus we do not consider nuc3 further.

Effect of background noise removal on marker gene detection

Above we have shown that computational methods can estimate background noise levels per cell. Moreover, all three methods provide the user with a background corrected count matrix for downstream analysis. Here, we compare the outcomes of marker gene detection, clustering and classification when using corrected count matrices from SoupX, DecontX, and CellBender (Fig. 6A, Additional file 1: Fig. S11). To characterize the impact on marker gene detection, we first check in how many cells an unexpected marker gene was detected; for example, how often Slc34a1 was detected in cells other than PTs (Fig. 6B). Without correction we find Slc34a1 reads in $\sim$ 60% of non-PT cells of rep2, SoupX reduces this rate to 54%, CellBender to 7% and DecontX$_{background}$ to 9%. DecontX$_{default}$ manages to remove most contaminating reads reducing the Slc34a1 detection rate outside PTs to 2%. While we find a similar ranking when averaging across several marker genes from the PanglaoDB database [17] and scRNA-seq replicates (Fig. 6C), the ranking changes for nuc2: DecontX$_{default}$ fails: after correction, Slc34a1 is still found in 87% of non-PT cells while DecontX$_{background}$ is better with a rate of 20%. Here, CellBender and SoupX are clearly better with reducing the Slc34a1 detection rate to 4% and $< 1\%$, respectively (Additional file 1: Fig. S12).

Even though the changes in the marker gene detection rates outside the designated cell type seem dramatic (Additional file 1: Fig. S13A), the identification of marker genes

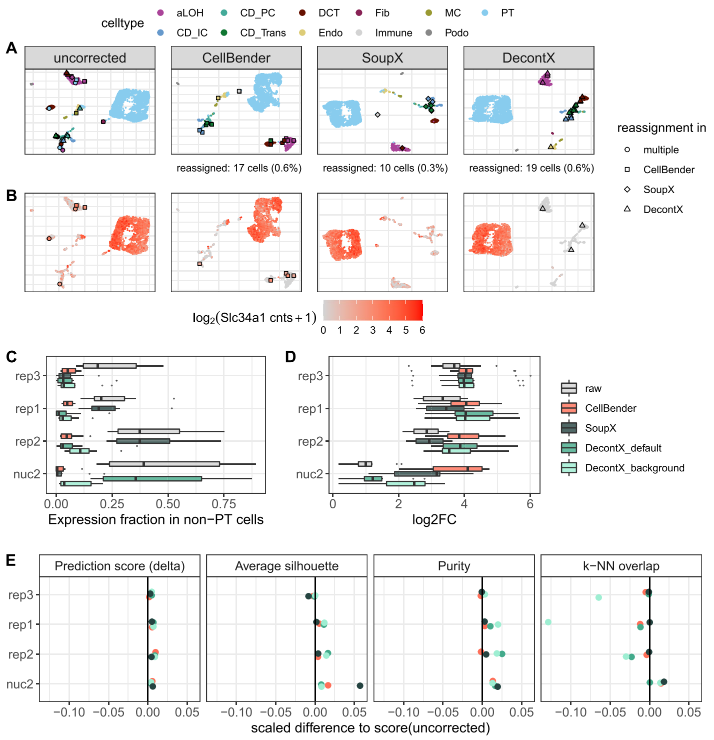
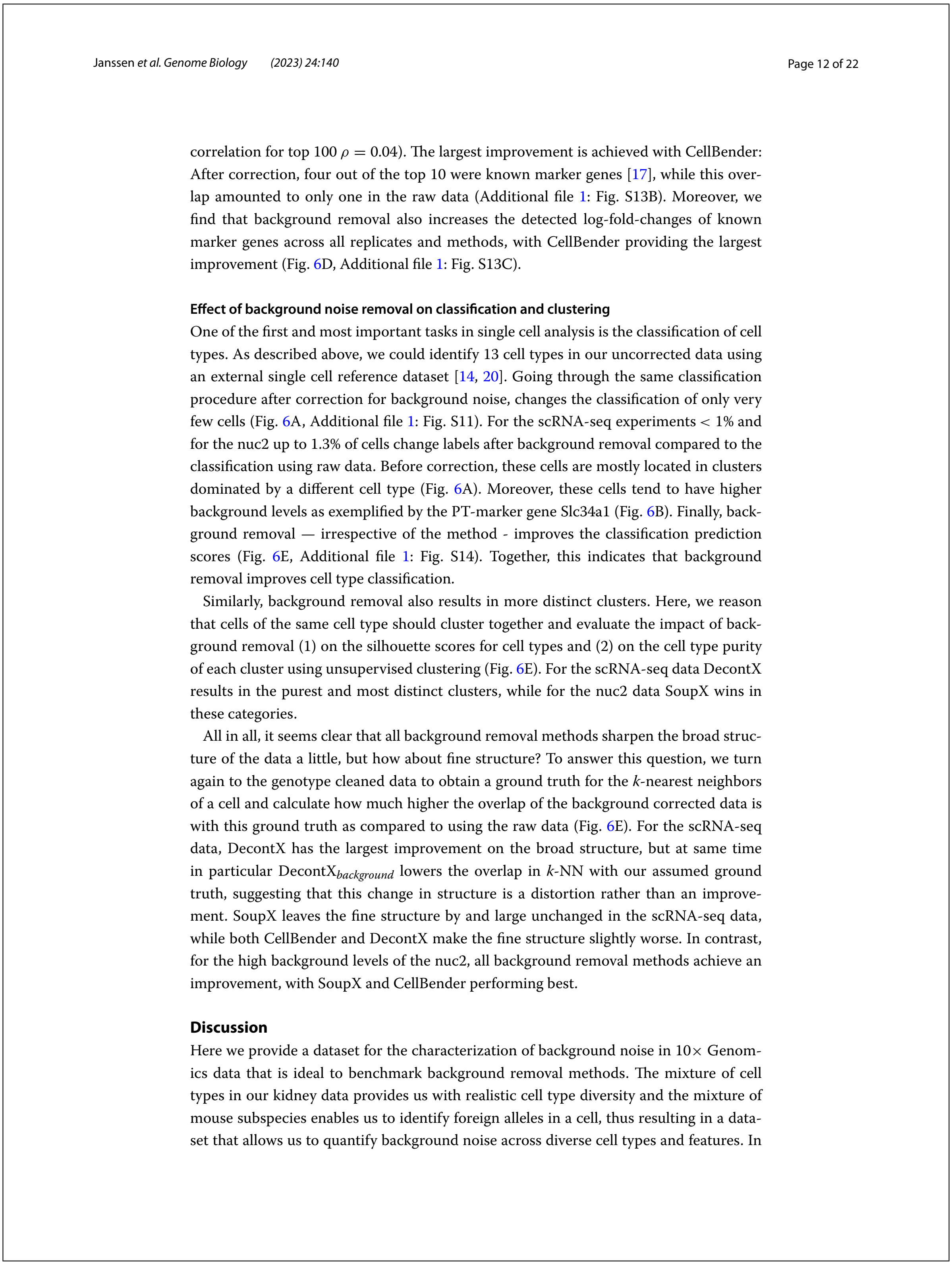


Fig. 6 Effect of background removal on downstream analysis. **A** UMAP representation of replicate 2 single-cell data before and after background noise correction, colored by cell type labels obtained from reference based classification. Individual cells that received a new label after correction are highlighted. PT, proximal tubule; CD_IC, intercalated cells of collecting duct; CD_PC, principal cells of collecting duct; CD_Trans, transitional cells of collecting duct; CNT, connecting tubule; DCT, distal convoluted tubule; Endo, endothelial; Fib, fibroblasts; aLOH, ascending loop of Henle; dLOH, descending loop of Henle; MC, mesangial cells; Podo, podocytes. **B** Expression of the PT cell marker Slc34a1 before and after background noise correction in replicate 2. Cells that were classified as PT cells in the uncorrected data, but got reassigned after correction, are highlighted. **C**, **D** Differential expression analysis of 10 PT markers, evaluating the expression fraction in non-PT cells (**C**) and the log2 fold change between PT and all other cells (**D**). **E** Evaluation metrics for the effect of background noise correction on classification and clustering. For each metric the change relative to the uncorrected data is depicted. The values were scaled by the possible range of each metric. Prediction score: cell-wise score "delta" of reference based classification with SingleR [20]. Average silhouette: Mean of silhouette widths per cell type. Purity: Cluster purity calculated on cell type labels as ground truth and Louvain clusters as test labels. *k*-NN overlap: overlap of the *k*=50 nearest neighbors per cell compared to genotype-cleaned reference *k*-NN graph

[21] is affected only a little. CellBender correction has the largest effect on marker gene detection, yet 8 from the top 10 genes without correction remain marker genes with CellBender correction (Spearman's correlation for top 100 $\rho = 0.84$). In contrast, in the nuc2 data with high background levels, the change in marker gene detection is dramatic. Here, only one of the top 10 marker genes remains after correction (Spearman's



correlation for top 100 $\rho = 0.04$). The largest improvement is achieved with CellBender: After correction, four out of the top 10 were known marker genes [17], while this overlap amounted to only one in the raw data (Additional file 1: Fig. S13B). Moreover, we find that background removal also increases the detected log-fold-changes of known marker genes across all replicates and methods, with CellBender providing the largest improvement (Fig. 6D, Additional file 1: Fig. S13C).

Effect of background noise removal on classification and clustering

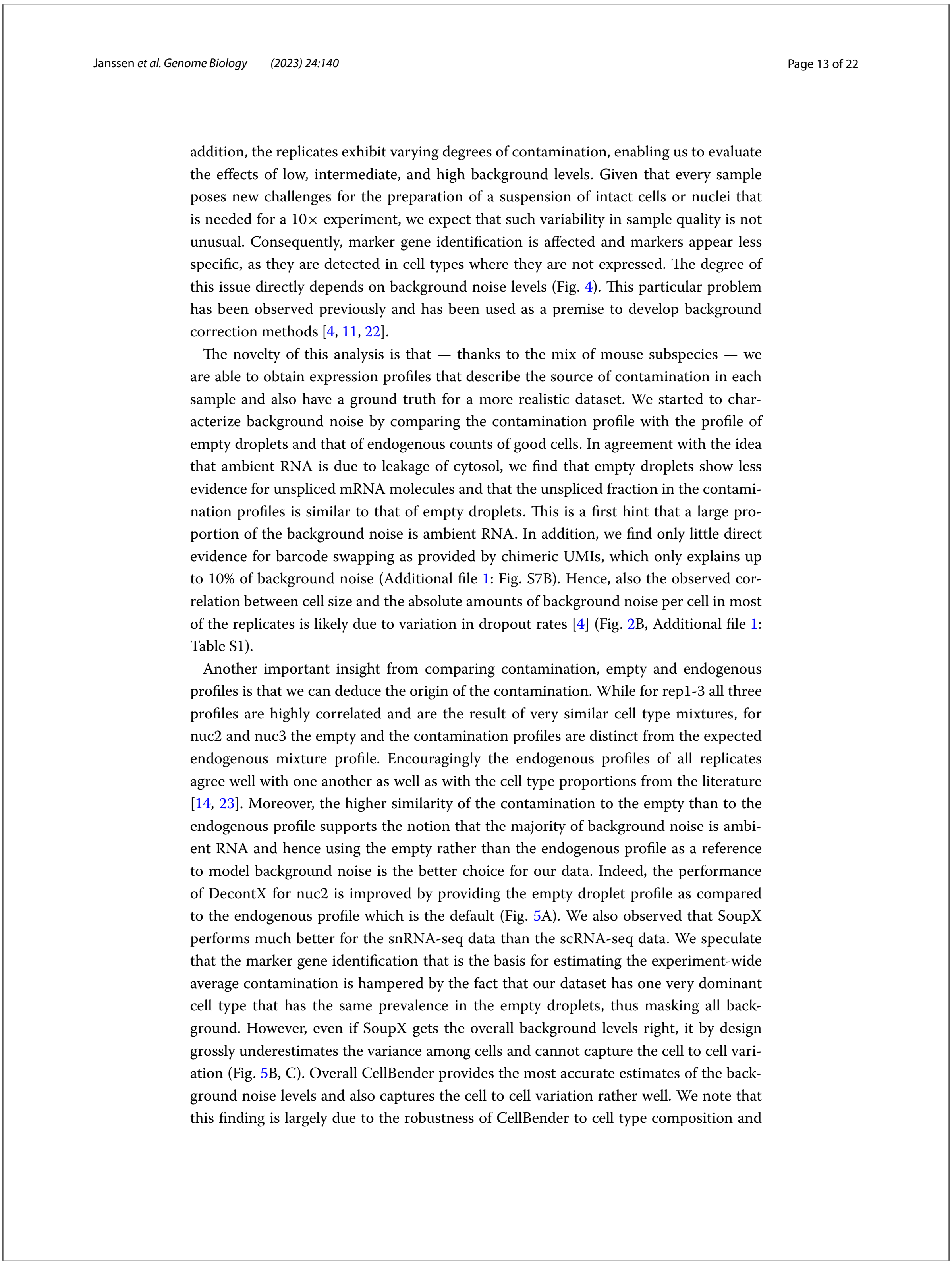
One of the first and most important tasks in single cell analysis is the classification of cell types. As described above, we could identify 13 cell types in our uncorrected data using an external single cell reference dataset [14, 20]. Going through the same classification procedure after correction for background noise, changes the classification of only very few cells (Fig. 6A, Additional file 1: Fig. S11). For the scRNA-seq experiments $<$ 1% and for the nuc2 up to 1.3% of cells change labels after background removal compared to the classification using raw data. Before correction, these cells are mostly located in clusters dominated by a different cell type (Fig. 6A). Moreover, these cells tend to have higher background levels as exemplified by the PT-marker gene Slc34a1 (Fig. 6B). Finally, background removal — irrespective of the method - improves the classification prediction scores (Fig. 6E, Additional file 1: Fig. S14). Together, this indicates that background removal improves cell type classification.

Similarly, background removal also results in more distinct clusters. Here, we reason that cells of the same cell type should cluster together and evaluate the impact of background removal (1) on the silhouette scores for cell types and (2) on the cell type purity of each cluster using unsupervised clustering (Fig. 6E). For the scRNA-seq data DecontX results in the purest and most distinct clusters, while for the nuc2 data SoupX wins in these categories.

All in all, it seems clear that all background removal methods sharpen the broad structure of the data a little, but how about fine structure? To answer this question, we turn again to the genotype cleaned data to obtain a ground truth for the *k*-nearest neighbors of a cell and calculate how much higher the overlap of the background corrected data is with this ground truth as compared to using the raw data (Fig. 6E). For the scRNA-seq data, DecontX has the largest improvement on the broad structure, but at same time in particular DecontX$_{background}$ lowers the overlap in *k*-NN with our assumed ground truth, suggesting that this change in structure is a distortion rather than an improvement. SoupX leaves the fine structure by and large unchanged in the scRNA-seq data, while both CellBender and DecontX make the fine structure slightly worse. In contrast, for the high background levels of the nuc2, all background removal methods achieve an improvement, with SoupX and CellBender performing best.

Discussion

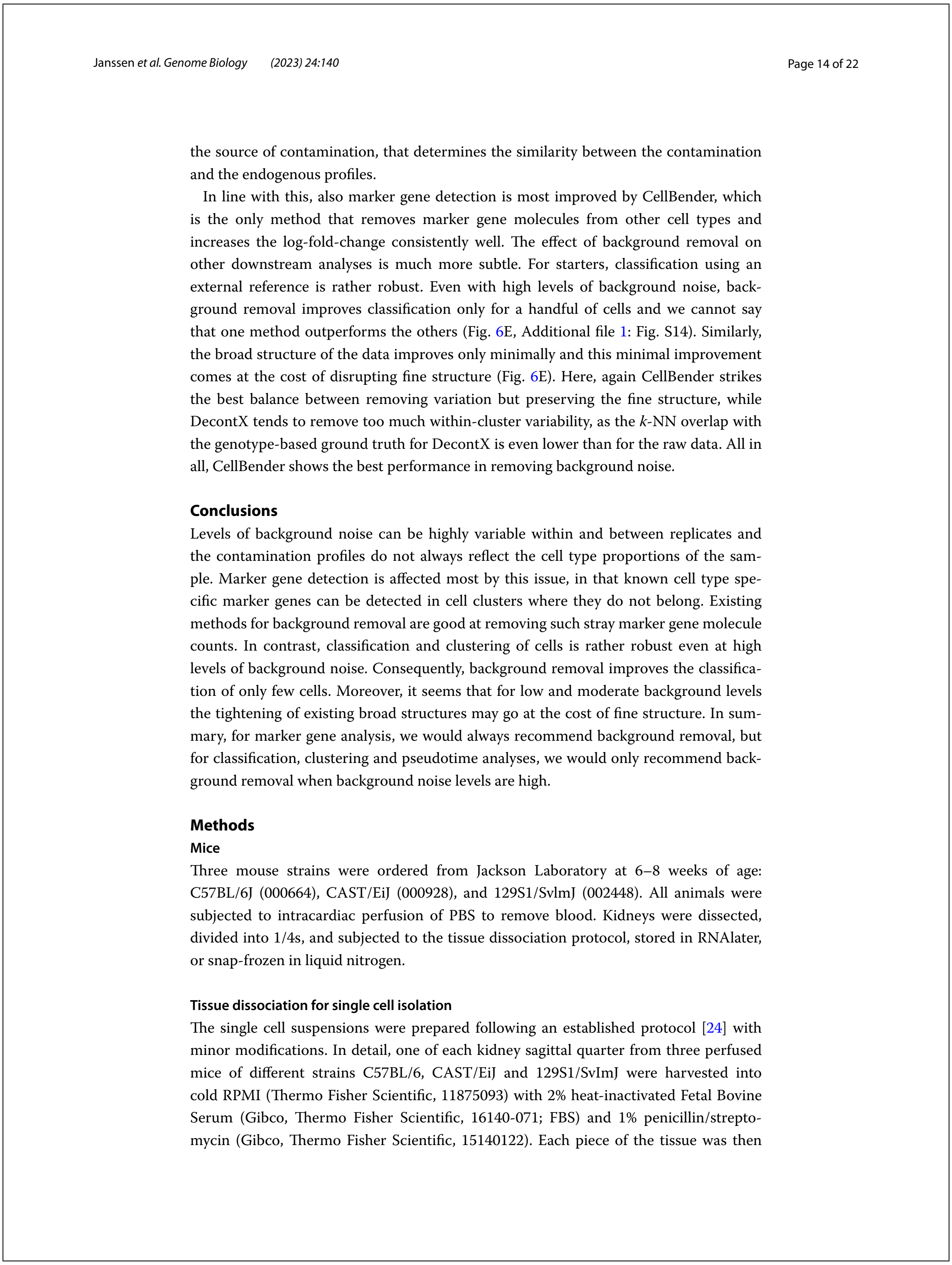
Here we provide a dataset for the characterization of background noise in 10× Genomics data that is ideal to benchmark background removal methods. The mixture of cell types in our kidney data provides us with realistic cell type diversity and the mixture of mouse subspecies enables us to identify foreign alleles in a cell, thus resulting in a dataset that allows us to quantify background noise across diverse cell types and features. In



addition, the replicates exhibit varying degrees of contamination, enabling us to evaluate the effects of low, intermediate, and high background levels. Given that every sample poses new challenges for the preparation of a suspension of intact cells or nuclei that is needed for a 10× experiment, we expect that such variability in sample quality is not unusual. Consequently, marker gene identification is affected and markers appear less specific, as they are detected in cell types where they are not expressed. The degree of this issue directly depends on background noise levels (Fig. 4). This particular problem has been observed previously and has been used as a premise to develop background correction methods [4, 11, 22].

The novelty of this analysis is that — thanks to the mix of mouse subspecies — we are able to obtain expression profiles that describe the source of contamination in each sample and also have a ground truth for a more realistic dataset. We started to characterize background noise by comparing the contamination profile with the profile of empty droplets and that of endogenous counts of good cells. In agreement with the idea that ambient RNA is due to leakage of cytosol, we find that empty droplets show less evidence for unspliced mRNA molecules and that the unspliced fraction in the contamination profiles is similar to that of empty droplets. This is a first hint that a large proportion of the background noise is ambient RNA. In addition, we find only little direct evidence for barcode swapping as provided by chimeric UMIs, which only explains up to 10% of background noise (Additional file 1: Fig. S7B). Hence, also the observed correlation between cell size and the absolute amounts of background noise per cell in most of the replicates is likely due to variation in dropout rates [4] (Fig. 2B, Additional file 1: Table S1).

Another important insight from comparing contamination, empty and endogenous profiles is that we can deduce the origin of the contamination. While for rep1-3 all three profiles are highly correlated and are the result of very similar cell type mixtures, for nuc2 and nuc3 the empty and the contamination profiles are distinct from the expected endogenous mixture profile. Encouragingly the endogenous profiles of all replicates agree well with one another as well as with the cell type proportions from the literature [14, 23]. Moreover, the higher similarity of the contamination to the empty than to the endogenous profile supports the notion that the majority of background noise is ambient RNA and hence using the empty rather than the endogenous profile as a reference to model background noise is the better choice for our data. Indeed, the performance of DecontX for nuc2 is improved by providing the empty droplet profile as compared to the endogenous profile which is the default (Fig. 5A). We also observed that SoupX performs much better for the snRNA-seq data than the scRNA-seq data. We speculate that the marker gene identification that is the basis for estimating the experiment-wide average contamination is hampered by the fact that our dataset has one very dominant cell type that has the same prevalence in the empty droplets, thus masking all background. However, even if SoupX gets the overall background levels right, it by design grossly underestimates the variance among cells and cannot capture the cell to cell variation (Fig. 5B, C). Overall CellBender provides the most accurate estimates of the background noise levels and also captures the cell to cell variation rather well. We note that this finding is largely due to the robustness of CellBender to cell type composition and



the source of contamination, that determines the similarity between the contamination and the endogenous profiles.

In line with this, also marker gene detection is most improved by CellBender, which is the only method that removes marker gene molecules from other cell types and increases the log-fold-change consistently well. The effect of background removal on other downstream analyses is much more subtle. For starters, classification using an external reference is rather robust. Even with high levels of background noise, background removal improves classification only for a handful of cells and we cannot say that one method outperforms the others (Fig. 6E, Additional file 1: Fig. S14). Similarly, the broad structure of the data improves only minimally and this minimal improvement comes at the cost of disrupting fine structure (Fig. 6E). Here, again CellBender strikes the best balance between removing variation but preserving the fine structure, while DecontX tends to remove too much within-cluster variability, as the *k*-NN overlap with the genotype-based ground truth for DecontX is even lower than for the raw data. All in all, CellBender shows the best performance in removing background noise.

Conclusions

Levels of background noise can be highly variable within and between replicates and the contamination profiles do not always reflect the cell type proportions of the sample. Marker gene detection is affected most by this issue, in that known cell type specific marker genes can be detected in cell clusters where they do not belong. Existing methods for background removal are good at removing such stray marker gene molecule counts. In contrast, classification and clustering of cells is rather robust even at high levels of background noise. Consequently, background removal improves the classification of only few cells. Moreover, it seems that for low and moderate background levels the tightening of existing broad structures may go at the cost of fine structure. In summary, for marker gene analysis, we would always recommend background removal, but for classification, clustering and pseudotime analyses, we would only recommend background removal when background noise levels are high.

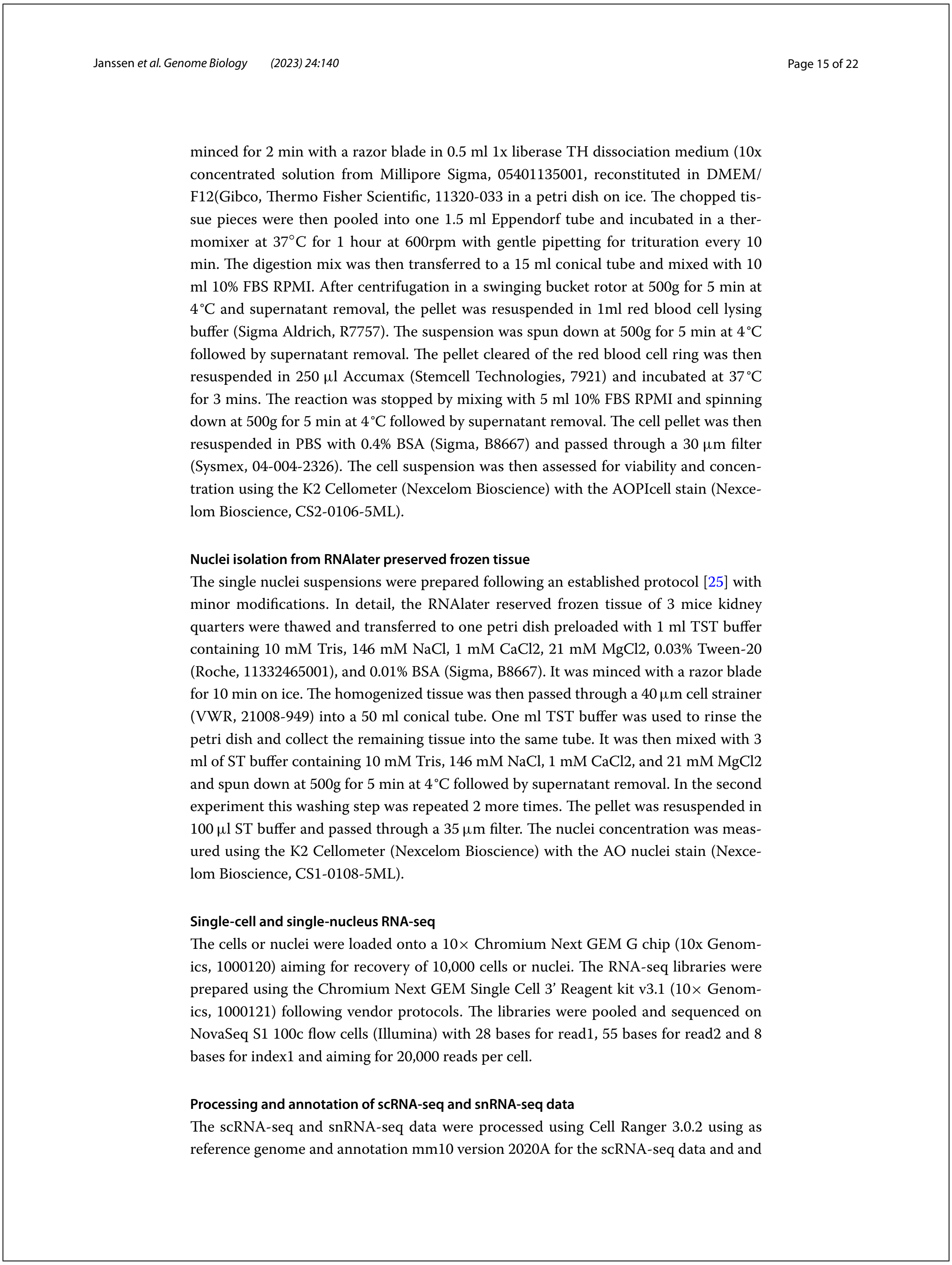
Methods

Mice

Three mouse strains were ordered from Jackson Laboratory at 6–8 weeks of age: C57BL/6J (000664), CAST/EiJ (000928), and 129S1/SvlmJ (002448). All animals were subjected to intracardiac perfusion of PBS to remove blood. Kidneys were dissected, divided into 1/4s, and subjected to the tissue dissociation protocol, stored in RNAlater, or snap-frozen in liquid nitrogen.

Tissue dissociation for single cell isolation

The single cell suspensions were prepared following an established protocol [24] with minor modifications. In detail, one of each kidney sagittal quarter from three perfused mice of different strains C57BL/6, CAST/EiJ and 129S1/SvImJ were harvested into cold RPMI (Thermo Fisher Scientific, 11875093) with 2% heat-inactivated Fetal Bovine Serum (Gibco, Thermo Fisher Scientific, 16140-071; FBS) and 1% penicillin/streptomycin (Gibco, Thermo Fisher Scientific, 15140122). Each piece of the tissue was then



minced for 2 min with a razor blade in 0.5 ml 1x liberase TH dissociation medium (10x concentrated solution from Millipore Sigma, 05401135001, reconstituted in DMEM/F12(Gibco, Thermo Fisher Scientific, 11320-033 in a petri dish on ice. The chopped tissue pieces were then pooled into one 1.5 ml Eppendorf tube and incubated in a thermomixer at 37°C for 1 hour at 600rpm with gentle pipetting for trituration every 10 min. The digestion mix was then transferred to a 15 ml conical tube and mixed with 10 ml 10% FBS RPMI. After centrifugation in a swinging bucket rotor at 500g for 5 min at 4 °C and supernatant removal, the pellet was resuspended in 1ml red blood cell lysing buffer (Sigma Aldrich, R7757). The suspension was spun down at 500g for 5 min at 4 °C followed by supernatant removal. The pellet cleared of the red blood cell ring was then resuspended in 250 μl Accumax (Stemcell Technologies, 7921) and incubated at 37 °C for 3 mins. The reaction was stopped by mixing with 5 ml 10% FBS RPMI and spinning down at 500g for 5 min at 4 °C followed by supernatant removal. The cell pellet was then resuspended in PBS with 0.4% BSA (Sigma, B8667) and passed through a 30 μm filter (Sysmex, 04-004-2326). The cell suspension was then assessed for viability and concentration using the K2 Cellometer (Nexcelom Bioscience) with the AOPIcell stain (Nexcelom Bioscience, CS2-0106-5ML).

Nuclei isolation from RNAlater preserved frozen tissue

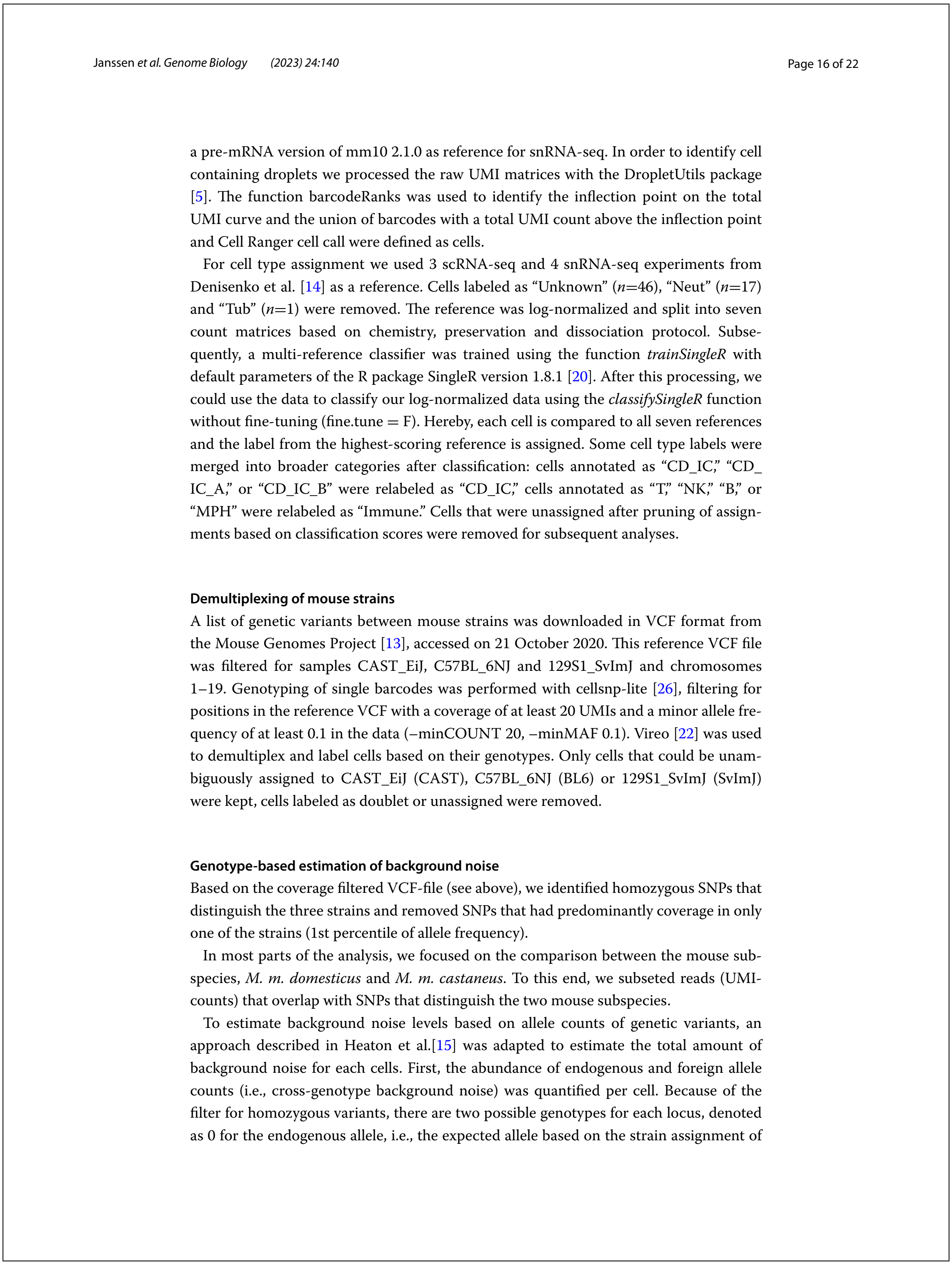
The single nuclei suspensions were prepared following an established protocol [25] with minor modifications. In detail, the RNAlater reserved frozen tissue of 3 mice kidney quarters were thawed and transferred to one petri dish preloaded with 1 ml TST buffer containing 10 mM Tris, 146 mM NaCl, 1 mM CaCl2, 21 mM MgCl2, 0.03% Tween-20 (Roche, 11332465001), and 0.01% BSA (Sigma, B8667). It was minced with a razor blade for 10 min on ice. The homogenized tissue was then passed through a 40 μm cell strainer (VWR, 21008-949) into a 50 ml conical tube. One ml TST buffer was used to rinse the petri dish and collect the remaining tissue into the same tube. It was then mixed with 3 ml of ST buffer containing 10 mM Tris, 146 mM NaCl, 1 mM CaCl2, and 21 mM MgCl2 and spun down at 500g for 5 min at 4 °C followed by supernatant removal. In the second experiment this washing step was repeated 2 more times. The pellet was resuspended in 100 μl ST buffer and passed through a 35 μm filter. The nuclei concentration was measured using the K2 Cellometer (Nexcelom Bioscience) with the AO nuclei stain (Nexcelom Bioscience, CS1-0108-5ML).

Single-cell and single-nucleus RNA-seq

The cells or nuclei were loaded onto a 10× Chromium Next GEM G chip (10x Genomics, 1000120) aiming for recovery of 10,000 cells or nuclei. The RNA-seq libraries were prepared using the Chromium Next GEM Single Cell 3' Reagent kit v3.1 (10× Genomics, 1000121) following vendor protocols. The libraries were pooled and sequenced on NovaSeq S1 100c flow cells (Illumina) with 28 bases for read1, 55 bases for read2 and 8 bases for index1 and aiming for 20,000 reads per cell.

Processing and annotation of scRNA-seq and snRNA-seq data

The scRNA-seq and snRNA-seq data were processed using Cell Ranger 3.0.2 using as reference genome and annotation mm10 version 2020A for the scRNA-seq data and and



a pre-mRNA version of mm10 2.1.0 as reference for snRNA-seq. In order to identify cell containing droplets we processed the raw UMI matrices with the DropletUtils package [5]. The function barcodeRanks was used to identify the inflection point on the total UMI curve and the union of barcodes with a total UMI count above the inflection point and Cell Ranger cell call were defined as cells.

For cell type assignment we used 3 scRNA-seq and 4 snRNA-seq experiments from Denisenko et al. [14] as a reference. Cells labeled as "Unknown" (n=46), "Neut" (n=17) and "Tub" (n=1) were removed. The reference was log-normalized and split into seven count matrices based on chemistry, preservation and dissociation protocol. Subsequently, a multi-reference classifier was trained using the function *trainSingleR* with default parameters of the R package SingleR version 1.8.1 [20]. After this processing, we could use the data to classify our log-normalized data using the *classifySingleR* function without fine-tuning (fine.tune = F). Hereby, each cell is compared to all seven references and the label from the highest-scoring reference is assigned. Some cell type labels were merged into broader categories after classification: cells annotated as "CD_IC," "CD_IC_A," or "CD_IC_B" were relabeled as "CD_IC," cells annotated as "T," "NK," "B," or "MPH" were relabeled as "Immune." Cells that were unassigned after pruning of assignments based on classification scores were removed for subsequent analyses.

Demultiplexing of mouse strains

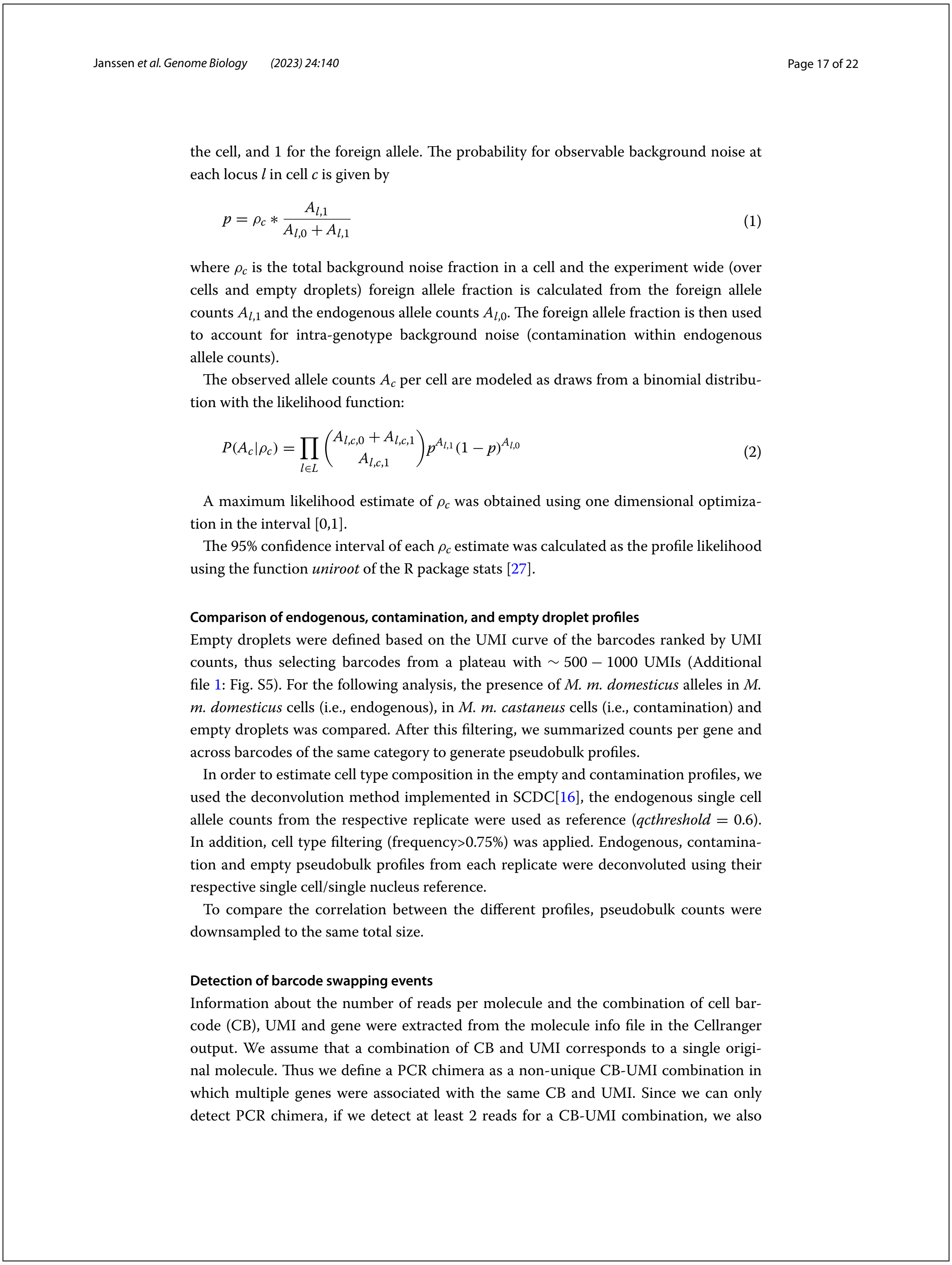
A list of genetic variants between mouse strains was downloaded in VCF format from the Mouse Genomes Project [13], accessed on 21 October 2020. This reference VCF file was filtered for samples CAST_EiJ, C57BL_6NJ and 129S1_SvImJ and chromosomes 1–19. Genotyping of single barcodes was performed with cellsnp-lite [26], filtering for positions in the reference VCF with a coverage of at least 20 UMIs and a minor allele frequency of at least 0.1 in the data (–minCOUNT 20, –minMAF 0.1). Vireo [22] was used to demultiplex and label cells based on their genotypes. Only cells that could be unambiguously assigned to CAST_EiJ (CAST), C57BL_6NJ (BL6) or 129S1_SvImJ (SvImJ) were kept, cells labeled as doublet or unassigned were removed.

Genotype-based estimation of background noise

Based on the coverage filtered VCF-file (see above), we identified homozygous SNPs that distinguish the three strains and removed SNPs that had predominantly coverage in only one of the strains (1st percentile of allele frequency).

In most parts of the analysis, we focused on the comparison between the mouse subspecies, *M. m. domesticus* and *M. m. castaneus*. To this end, we subseted reads (UMI-counts) that overlap with SNPs that distinguish the two mouse subspecies.

To estimate background noise levels based on allele counts of genetic variants, an approach described in Heaton et al.[15] was adapted to estimate the total amount of background noise for each cells. First, the abundance of endogenous and foreign allele counts (i.e., cross-genotype background noise) was quantified per cell. Because of the filter for homozygous variants, there are two possible genotypes for each locus, denoted as 0 for the endogenous allele, i.e., the expected allele based on the strain assignment of



the cell, and 1 for the foreign allele. The probability for observable background noise at each locus l in cell c is given by

$$p = \rho_c * \frac{A_{l,1}}{A_{l,0} + A_{l,1}} \tag{1}$$

where ρ_c is the total background noise fraction in a cell and the experiment wide (over cells and empty droplets) foreign allele fraction is calculated from the foreign allele counts $A_{l,1}$ and the endogenous allele counts $A_{l,0}$. The foreign allele fraction is then used to account for intra-genotype background noise (contamination within endogenous allele counts).

The observed allele counts A_c per cell are modeled as draws from a binomial distribution with the likelihood function:

$$P(A_c|\rho_c) = \prod_{l \in L} \binom{A_{l,c,0} + A_{l,c,1}}{A_{l,c,1}} p^{A_{l,1}} (1-p)^{A_{l,0}} \tag{2}$$

A maximum likelihood estimate of ρ_c was obtained using one dimensional optimization in the interval [0,1].

The 95% confidence interval of each ρ_c estimate was calculated as the profile likelihood using the function *uniroot* of the R package stats [27].

Comparison of endogenous, contamination, and empty droplet profiles

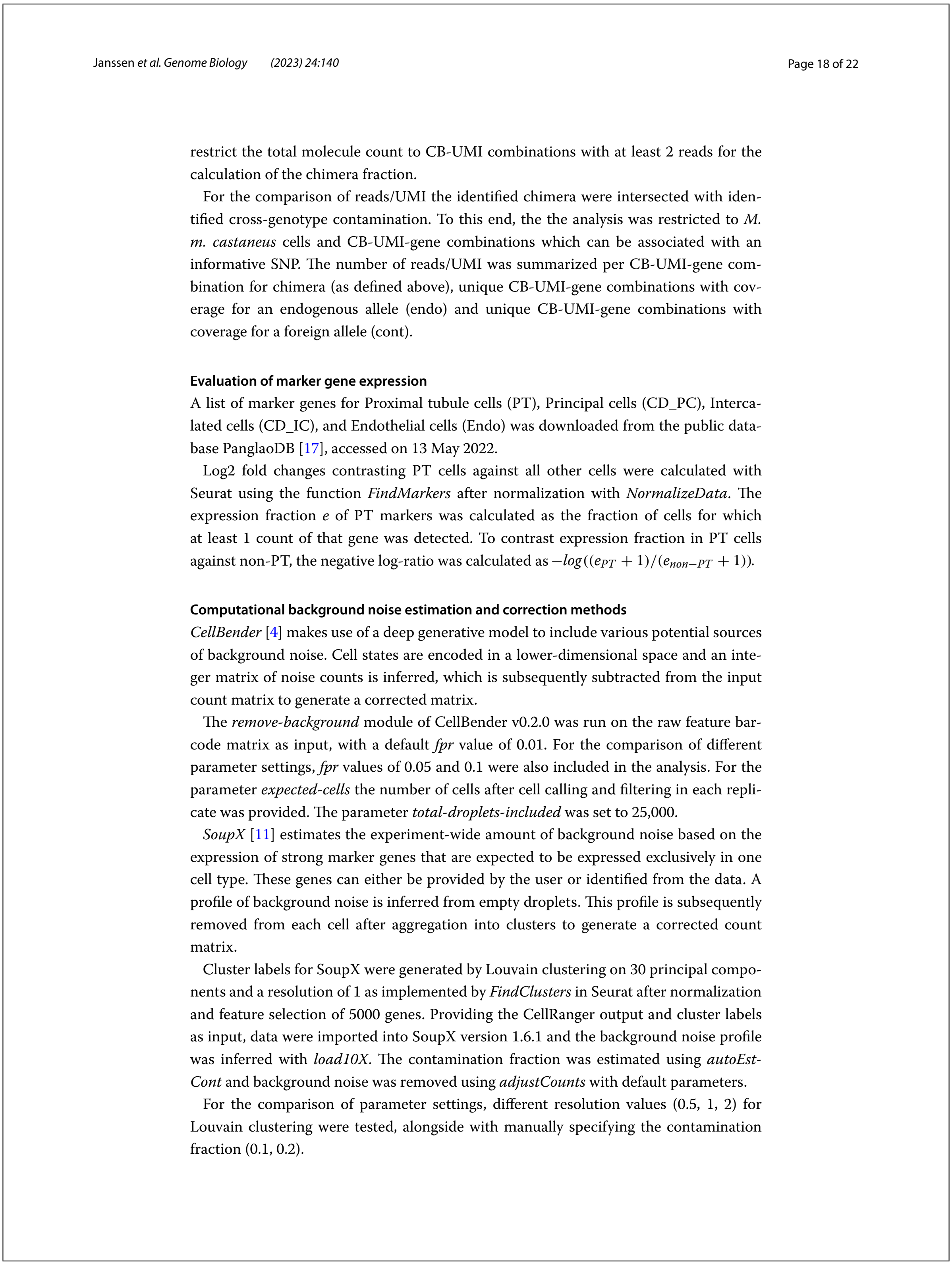
Empty droplets were defined based on the UMI curve of the barcodes ranked by UMI counts, thus selecting barcodes from a plateau with $\sim 500 - 1000$ UMIs (Additional file 1: Fig. S5). For the following analysis, the presence of *M. m. domesticus* alleles in *M. m. domesticus* cells (i.e., endogenous), in *M. m. castaneus* cells (i.e., contamination) and empty droplets was compared. After this filtering, we summarized counts per gene and across barcodes of the same category to generate pseudobulk profiles.

In order to estimate cell type composition in the empty and contamination profiles, we used the deconvolution method implemented in SCDC[16], the endogenous single cell allele counts from the respective replicate were used as reference ($qcthreshold = 0.6$). In addition, cell type filtering (frequency>0.75%) was applied. Endogenous, contamination and empty pseudobulk profiles from each replicate were deconvoluted using their respective single cell/single nucleus reference.

To compare the correlation between the different profiles, pseudobulk counts were downsampled to the same total size.

Detection of barcode swapping events

Information about the number of reads per molecule and the combination of cell barcode (CB), UMI and gene were extracted from the molecule info file in the Cellranger output. We assume that a combination of CB and UMI corresponds to a single original molecule. Thus we define a PCR chimera as a non-unique CB-UMI combination in which multiple genes were associated with the same CB and UMI. Since we can only detect PCR chimera, if we detect at least 2 reads for a CB-UMI combination, we also



restrict the total molecule count to CB-UMI combinations with at least 2 reads for the calculation of the chimera fraction.

For the comparison of reads/UMI the identified chimera were intersected with identified cross-genotype contamination. To this end, the the analysis was restricted to *M. m. castaneus* cells and CB-UMI-gene combinations which can be associated with an informative SNP. The number of reads/UMI was summarized per CB-UMI-gene combination for chimera (as defined above), unique CB-UMI-gene combinations with coverage for an endogenous allele (endo) and unique CB-UMI-gene combinations with coverage for a foreign allele (cont).

Evaluation of marker gene expression

A list of marker genes for Proximal tubule cells (PT), Principal cells (CD_PC), Intercalated cells (CD_IC), and Endothelial cells (Endo) was downloaded from the public database PanglaoDB [17], accessed on 13 May 2022.

Log2 fold changes contrasting PT cells against all other cells were calculated with Seurat using the function *FindMarkers* after normalization with *NormalizeData*. The expression fraction e of PT markers was calculated as the fraction of cells for which at least 1 count of that gene was detected. To contrast expression fraction in PT cells against non-PT, the negative log-ratio was calculated as $-log((e_{PT} + 1)/(e_{non-PT} + 1))$.

Computational background noise estimation and correction methods

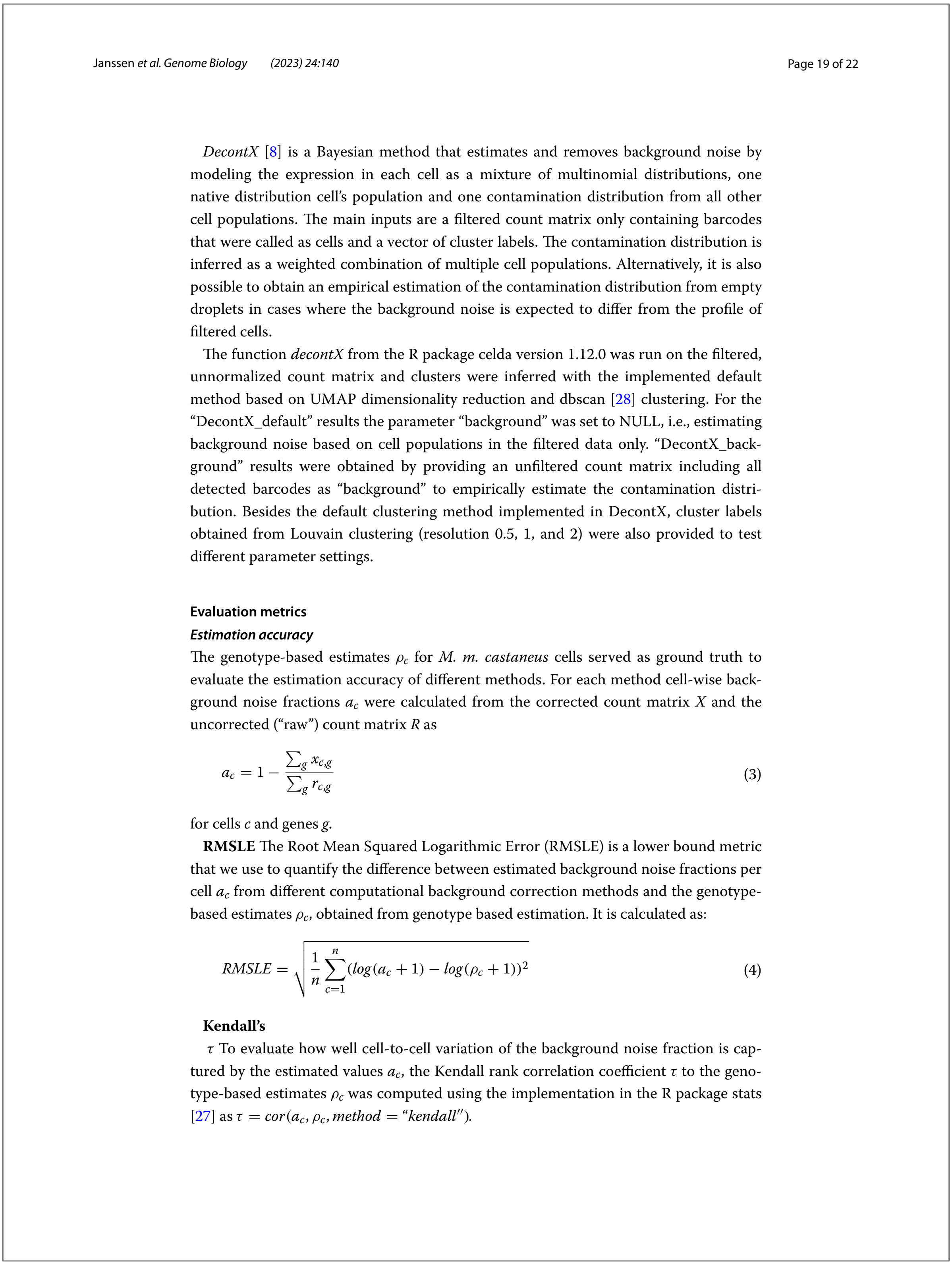
CellBender [4] makes use of a deep generative model to include various potential sources of background noise. Cell states are encoded in a lower-dimensional space and an integer matrix of noise counts is inferred, which is subsequently subtracted from the input count matrix to generate a corrected matrix.

The *remove-background* module of CellBender v0.2.0 was run on the raw feature barcode matrix as input, with a default *fpr* value of 0.01. For the comparison of different parameter settings, *fpr* values of 0.05 and 0.1 were also included in the analysis. For the parameter *expected-cells* the number of cells after cell calling and filtering in each replicate was provided. The parameter *total-droplets-included* was set to 25,000.

SoupX [11] estimates the experiment-wide amount of background noise based on the expression of strong marker genes that are expected to be expressed exclusively in one cell type. These genes can either be provided by the user or identified from the data. A profile of background noise is inferred from empty droplets. This profile is subsequently removed from each cell after aggregation into clusters to generate a corrected count matrix.

Cluster labels for SoupX were generated by Louvain clustering on 30 principal components and a resolution of 1 as implemented by *FindClusters* in Seurat after normalization and feature selection of 5000 genes. Providing the CellRanger output and cluster labels as input, data were imported into SoupX version 1.6.1 and the background noise profile was inferred with *load10X*. The contamination fraction was estimated using *autoEstCont* and background noise was removed using *adjustCounts* with default parameters.

For the comparison of parameter settings, different resolution values (0.5, 1, 2) for Louvain clustering were tested, alongside with manually specifying the contamination fraction (0.1, 0.2).



DecontX [8] is a Bayesian method that estimates and removes background noise by modeling the expression in each cell as a mixture of multinomial distributions, one native distribution cell's population and one contamination distribution from all other cell populations. The main inputs are a filtered count matrix only containing barcodes that were called as cells and a vector of cluster labels. The contamination distribution is inferred as a weighted combination of multiple cell populations. Alternatively, it is also possible to obtain an empirical estimation of the contamination distribution from empty droplets in cases where the background noise is expected to differ from the profile of filtered cells.

The function *decontX* from the R package celda version 1.12.0 was run on the filtered, unnormalized count matrix and clusters were inferred with the implemented default method based on UMAP dimensionality reduction and dbscan [28] clustering. For the "DecontX_default" results the parameter "background" was set to NULL, i.e., estimating background noise based on cell populations in the filtered data only. "DecontX_background" results were obtained by providing an unfiltered count matrix including all detected barcodes as "background" to empirically estimate the contamination distribution. Besides the default clustering method implemented in DecontX, cluster labels obtained from Louvain clustering (resolution 0.5, 1, and 2) were also provided to test different parameter settings.

Evaluation metrics

Estimation accuracy

The genotype-based estimates ρ_c for *M. m. castaneus* cells served as ground truth to evaluate the estimation accuracy of different methods. For each method cell-wise background noise fractions a_c were calculated from the corrected count matrix X and the uncorrected ("raw") count matrix R as

$$a_c = 1 - \frac{\sum_g x_{c,g}}{\sum_g r_{c,g}} \tag{3}$$

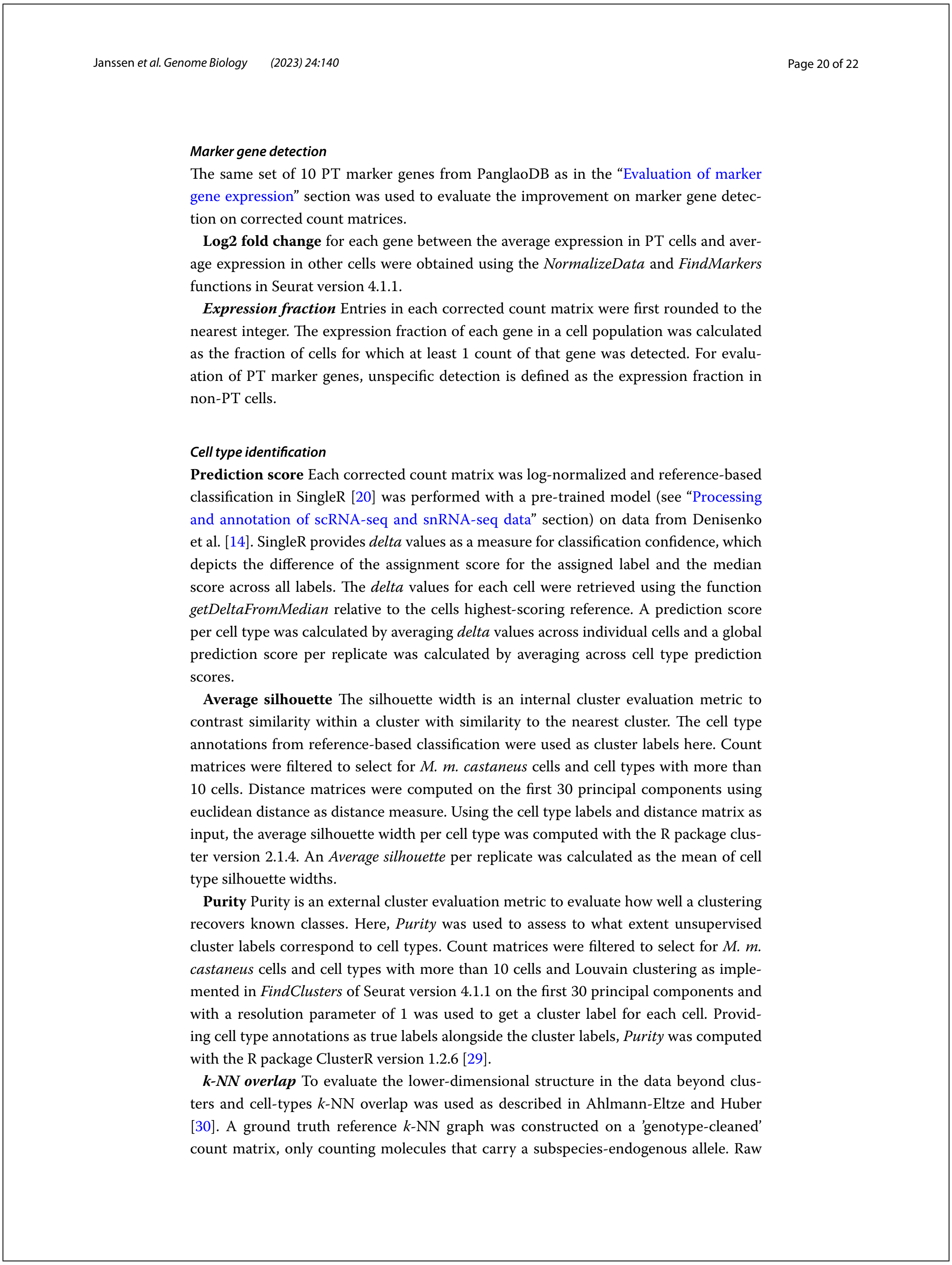
for cells c and genes g.

RMSLE The Root Mean Squared Logarithmic Error (RMSLE) is a lower bound metric that we use to quantify the difference between estimated background noise fractions per cell a_c from different computational background correction methods and the genotype-based estimates ρ_c, obtained from genotype based estimation. It is calculated as:

$$RMSLE = \sqrt{\frac{1}{n}\sum_{c=1}^{n}(log(a_c+1) - log(\rho_c+1))^2} \tag{4}$$

Kendall's

τ To evaluate how well cell-to-cell variation of the background noise fraction is captured by the estimated values a_c, the Kendall rank correlation coefficient τ to the genotype-based estimates ρ_c was computed using the implementation in the R package stats [27] as $\tau = cor(a_c, \rho_c, method = \text{“}kendall''$).



Marker gene detection

The same set of 10 PT marker genes from PanglaoDB as in the "Evaluation of marker gene expression" section was used to evaluate the improvement on marker gene detection on corrected count matrices.

Log2 fold change for each gene between the average expression in PT cells and average expression in other cells were obtained using the *NormalizeData* and *FindMarkers* functions in Seurat version 4.1.1.

Expression fraction Entries in each corrected count matrix were first rounded to the nearest integer. The expression fraction of each gene in a cell population was calculated as the fraction of cells for which at least 1 count of that gene was detected. For evaluation of PT marker genes, unspecific detection is defined as the expression fraction in non-PT cells.

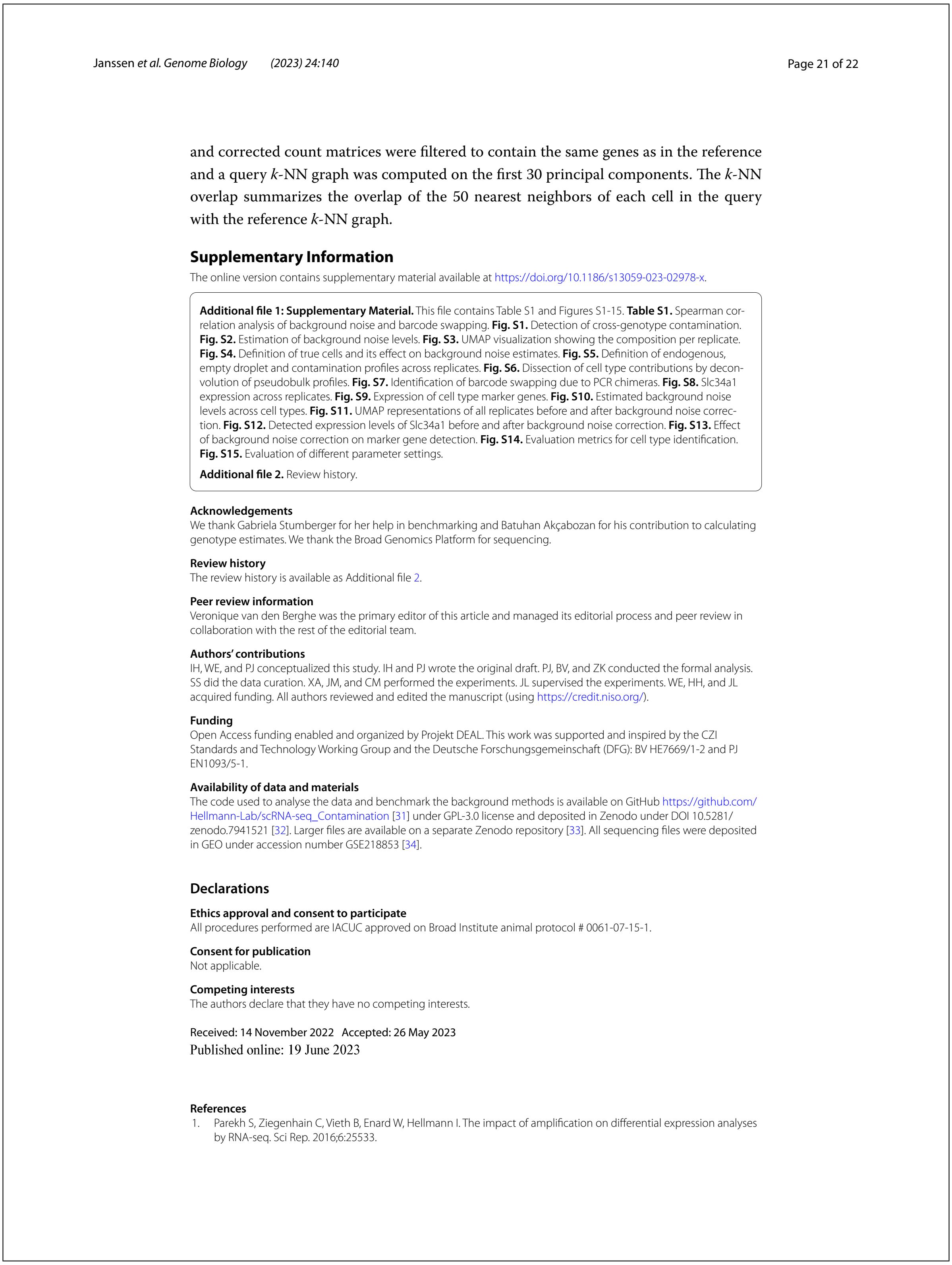
Cell type identification

Prediction score Each corrected count matrix was log-normalized and reference-based classification in SingleR [20] was performed with a pre-trained model (see "Processing and annotation of scRNA-seq and snRNA-seq data" section) on data from Denisenko et al. [14]. SingleR provides *delta* values as a measure for classification confidence, which depicts the difference of the assignment score for the assigned label and the median score across all labels. The *delta* values for each cell were retrieved using the function *getDeltaFromMedian* relative to the cells highest-scoring reference. A prediction score per cell type was calculated by averaging *delta* values across individual cells and a global prediction score per replicate was calculated by averaging across cell type prediction scores.

Average silhouette The silhouette width is an internal cluster evaluation metric to contrast similarity within a cluster with similarity to the nearest cluster. The cell type annotations from reference-based classification were used as cluster labels here. Count matrices were filtered to select for *M. m. castaneus* cells and cell types with more than 10 cells. Distance matrices were computed on the first 30 principal components using euclidean distance as distance measure. Using the cell type labels and distance matrix as input, the average silhouette width per cell type was computed with the R package cluster version 2.1.4. An *Average silhouette* per replicate was calculated as the mean of cell type silhouette widths.

Purity Purity is an external cluster evaluation metric to evaluate how well a clustering recovers known classes. Here, *Purity* was used to assess to what extent unsupervised cluster labels correspond to cell types. Count matrices were filtered to select for *M. m. castaneus* cells and cell types with more than 10 cells and Louvain clustering as implemented in *FindClusters* of Seurat version 4.1.1 on the first 30 principal components and with a resolution parameter of 1 was used to get a cluster label for each cell. Providing cell type annotations as true labels alongside the cluster labels, *Purity* was computed with the R package ClusterR version 1.2.6 [29].

k-NN overlap To evaluate the lower-dimensional structure in the data beyond clusters and cell-types *k*-NN overlap was used as described in Ahlmann-Eltze and Huber [30]. A ground truth reference *k*-NN graph was constructed on a 'genotype-cleaned' count matrix, only counting molecules that carry a subspecies-endogenous allele. Raw



and corrected count matrices were filtered to contain the same genes as in the reference and a query k-NN graph was computed on the first 30 principal components. The k-NN overlap summarizes the overlap of the 50 nearest neighbors of each cell in the query with the reference k-NN graph.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13059-023-02978-x.

Additional file 1: Supplementary Material. This file contains Table S1 and Figures S1-15. **Table S1.** Spearman correlation analysis of background noise and barcode swapping. **Fig. S1.** Detection of cross-genotype contamination. **Fig. S2.** Estimation of background noise levels. **Fig. S3.** UMAP visualization showing the composition per replicate. **Fig. S4.** Definition of true cells and its effect on background noise estimates. **Fig. S5.** Definition of endogenous, empty droplet and contamination profiles across replicates. **Fig. S6.** Dissection of cell type contributions by deconvolution of pseudobulk profiles. **Fig. S7.** Identification of barcode swapping due to PCR chimeras. **Fig. S8.** Slc34a1 expression across replicates. **Fig. S9.** Expression of cell type marker genes. **Fig. S10.** Estimated background noise levels across cell types. **Fig. S11.** UMAP representations of all replicates before and after background noise correction. **Fig. S12.** Detected expression levels of Slc34a1 before and after background noise correction. **Fig. S13.** Effect of background noise correction on marker gene detection. **Fig. S14.** Evaluation metrics for cell type identification. **Fig. S15.** Evaluation of different parameter settings.

Additional file 2. Review history.

Acknowledgements
We thank Gabriela Stumberger for her help in benchmarking and Batuhan Akçabozan for his contribution to calculating genotype estimates. We thank the Broad Genomics Platform for sequencing.

Review history
The review history is available as Additional file 2.

Peer review information
Veronique van den Berghe was the primary editor of this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

Authors' contributions
IH, WE, and PJ conceptualized this study. IH and PJ wrote the original draft. PJ, BV, and ZK conducted the formal analysis. SS did the data curation. XA, JM, and CM performed the experiments. JL supervised the experiments. WE, HH, and JL acquired funding. All authors reviewed and edited the manuscript (using https://credit.niso.org/).

Funding
Open Access funding enabled and organized by Projekt DEAL. This work was supported and inspired by the CZI Standards and Technology Working Group and the Deutsche Forschungsgemeinschaft (DFG): BV HE7669/1-2 and PJ EN1093/5-1.

Availability of data and materials
The code used to analyse the data and benchmark the background methods is available on GitHub https://github.com/Hellmann-Lab/scRNA-seq_Contamination [31] under GPL-3.0 license and deposited in Zenodo under DOI 10.5281/zenodo.7941521 [32]. Larger files are available on a separate Zenodo repository [33]. All sequencing files were deposited in GEO under accession number GSE218853 [34].

Declarations

Ethics approval and consent to participate
All procedures performed are IACUC approved on Broad Institute animal protocol # 0061-07-15-1.

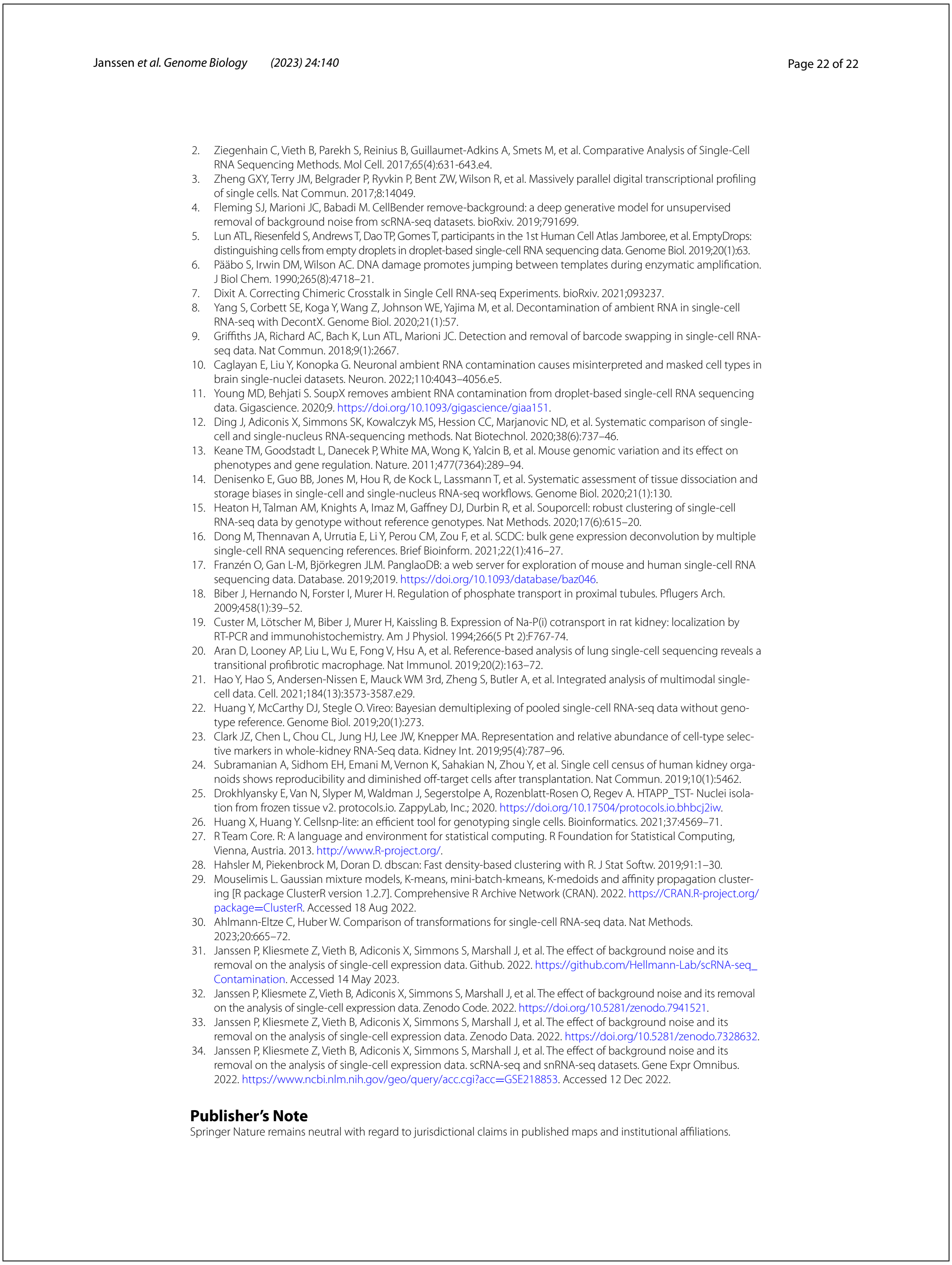
Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Received: 14 November 2022 Accepted: 26 May 2023
Published online: 19 June 2023

References

1. Parekh S, Ziegenhain C, Vieth B, Enard W, Hellmann I. The impact of amplification on differential expression analyses by RNA-seq. Sci Rep. 2016;6:25533.



2. Ziegenhain C, Vieth B, Parekh S, Reinius B, Guillaumet-Adkins A, Smets M, et al. Comparative Analysis of Single-Cell RNA Sequencing Methods. Mol Cell. 2017;65(4):631-643.e4.
3. Zheng GXY, Terry JM, Belgrader P, Ryvkin P, Bent ZW, Wilson R, et al. Massively parallel digital transcriptional profiling of single cells. Nat Commun. 2017;8:14049.
4. Fleming SJ, Marioni JC, Babadi M. CellBender remove-background: a deep generative model for unsupervised removal of background noise from scRNA-seq datasets. bioRxiv. 2019;791699.
5. Lun ATL, Riesenfeld S, Andrews T, Dao TP, Gomes T, participants in the 1st Human Cell Atlas Jamboree, et al. EmptyDrops: distinguishing cells from empty droplets in droplet-based single-cell RNA sequencing data. Genome Biol. 2019;20(1):63.
6. Pääbo S, Irwin DM, Wilson AC. DNA damage promotes jumping between templates during enzymatic amplification. J Biol Chem. 1990;265(8):4718–21.
7. Dixit A. Correcting Chimeric Crosstalk in Single Cell RNA-seq Experiments. bioRxiv. 2021;093237.
8. Yang S, Corbett SE, Koga Y, Wang Z, Johnson WE, Yajima M, et al. Decontamination of ambient RNA in single-cell RNA-seq with DecontX. Genome Biol. 2020;21(1):57.
9. Griffiths JA, Richard AC, Bach K, Lun ATL, Marioni JC. Detection and removal of barcode swapping in single-cell RNA-seq data. Nat Commun. 2018;9(1):2667.
10. Caglayan E, Liu Y, Konopka G. Neuronal ambient RNA contamination causes misinterpreted and masked cell types in brain single-nuclei datasets. Neuron. 2022;110:4043–4056.e5.
11. Young MD, Behjati S. SoupX removes ambient RNA contamination from droplet-based single-cell RNA sequencing data. Gigascience. 2020;9. https://doi.org/10.1093/gigascience/giaa151.
12. Ding J, Adiconis X, Simmons SK, Kowalczyk MS, Hession CC, Marjanovic ND, et al. Systematic comparison of single-cell and single-nucleus RNA-sequencing methods. Nat Biotechnol. 2020;38(6):737–46.
13. Keane TM, Goodstadt L, Danecek P, White MA, Wong K, Yalcin B, et al. Mouse genomic variation and its effect on phenotypes and gene regulation. Nature. 2011;477(7364):289–94.
14. Denisenko E, Guo BB, Jones M, Hou R, de Kock L, Lassmann T, et al. Systematic assessment of tissue dissociation and storage biases in single-cell and single-nucleus RNA-seq workflows. Genome Biol. 2020;21(1):130.
15. Heaton H, Talman AM, Knights A, Imaz M, Gaffney DJ, Durbin R, et al. Souporcell: robust clustering of single-cell RNA-seq data by genotype without reference genotypes. Nat Methods. 2020;17(6):615–20.
16. Dong M, Thennavan A, Urrutia E, Li Y, Perou CM, Zou F, et al. SCDC: bulk gene expression deconvolution by multiple single-cell RNA sequencing references. Brief Bioinform. 2021;22(1):416–27.
17. Franzén O, Gan L-M, Björkegren JLM. PanglaoDB: a web server for exploration of mouse and human single-cell RNA sequencing data. Database. 2019;2019. https://doi.org/10.1093/database/baz046.
18. Biber J, Hernando N, Forster I, Murer H. Regulation of phosphate transport in proximal tubules. Pflugers Arch. 2009;458(1):39–52.
19. Custer M, Lötscher M, Biber J, Murer H, Kaissling B. Expression of Na-P(i) cotransport in rat kidney: localization by RT-PCR and immunohistochemistry. Am J Physiol. 1994;266(5 Pt 2):F767-74.
20. Aran D, Looney AP, Liu L, Wu E, Fong V, Hsu A, et al. Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. Nat Immunol. 2019;20(2):163–72.
21. Hao Y, Hao S, Andersen-Nissen E, Mauck WM 3rd, Zheng S, Butler A, et al. Integrated analysis of multimodal single-cell data. Cell. 2021;184(13):3573-3587.e29.
22. Huang Y, McCarthy DJ, Stegle O. Vireo: Bayesian demultiplexing of pooled single-cell RNA-seq data without genotype reference. Genome Biol. 2019;20(1):273.
23. Clark JZ, Chen L, Chou CL, Jung HJ, Lee JW, Knepper MA. Representation and relative abundance of cell-type selective markers in whole-kidney RNA-Seq data. Kidney Int. 2019;95(4):787–96.
24. Subramanian A, Sidhom EH, Emani M, Vernon K, Sahakian N, Zhou Y, et al. Single cell census of human kidney organoids shows reproducibility and diminished off-target cells after transplantation. Nat Commun. 2019;10(1):5462.
25. Drokhlyansky E, Van N, Slyper M, Waldman J, Segerstolpe A, Rozenblatt-Rosen O, Regev A. HTAPP_TST- Nuclei isolation from frozen tissue v2. protocols.io. ZappyLab, Inc.; 2020. https://doi.org/10.17504/protocols.io.bhbcj2iw.
26. Huang X, Huang Y. Cellsnp-lite: an efficient tool for genotyping single cells. Bioinformatics. 2021;37:4569–71.
27. R Team Core. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2013. http://www.R-project.org/.
28. Hahsler M, Piekenbrock M, Doran D. dbscan: Fast density-based clustering with R. J Stat Softw. 2019;91:1–30.
29. Mouselimis L. Gaussian mixture models, K-means, mini-batch-kmeans, K-medoids and affinity propagation clustering [R package ClusterR version 1.2.7]. Comprehensive R Archive Network (CRAN). 2022. https://CRAN.R-project.org/package=ClusterR. Accessed 18 Aug 2022.
30. Ahlmann-Eltze C, Huber W. Comparison of transformations for single-cell RNA-seq data. Nat Methods. 2023;20:665–72.
31. Janssen P, Kliesmete Z, Vieth B, Adiconis X, Simmons S, Marshall J, et al. The effect of background noise and its removal on the analysis of single-cell expression data. Github. 2022. https://github.com/Hellmann-Lab/scRNA-seq_Contamination. Accessed 14 May 2023.
32. Janssen P, Kliesmete Z, Vieth B, Adiconis X, Simmons S, Marshall J, et al. The effect of background noise and its removal on the analysis of single-cell expression data. Zenodo Code. 2022. https://doi.org/10.5281/zenodo.7941521.
33. Janssen P, Kliesmete Z, Vieth B, Adiconis X, Simmons S, Marshall J, et al. The effect of background noise and its removal on the analysis of single-cell expression data. Zenodo Data. 2022. https://doi.org/10.5281/zenodo.7328632.
34. Janssen P, Kliesmete Z, Vieth B, Adiconis X, Simmons S, Marshall J, et al. The effect of background noise and its removal on the analysis of single-cell expression data. scRNA-seq and snRNA-seq datasets. Gene Expr Omnibus. 2022. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE218853. Accessed 12 Dec 2022.

Publisher's Note

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

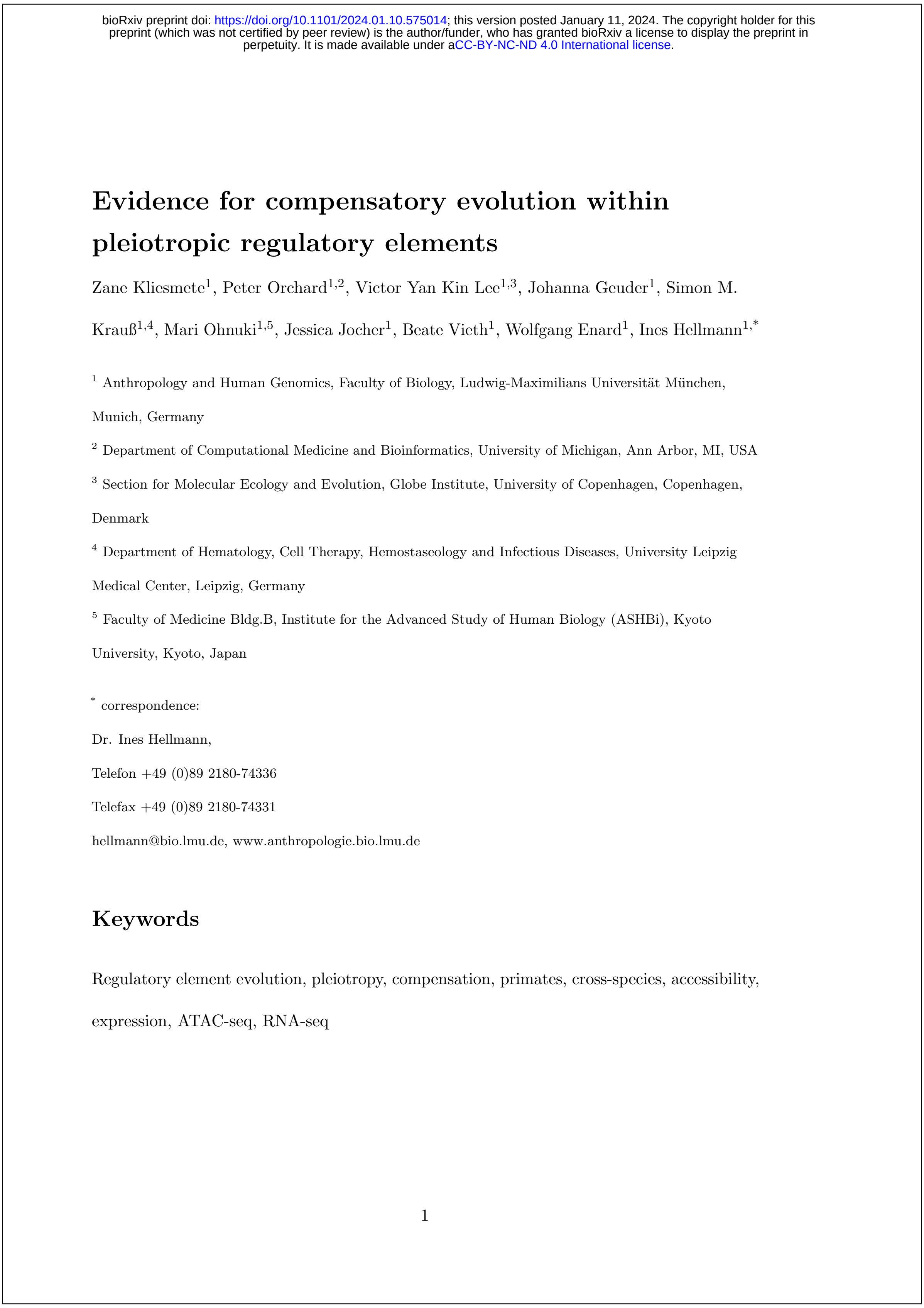
2.2 Evidence for compensatory evolution within pleiotropic regulatory elements

Kliesmete Z, Orchard P, Lee VY, Geuder J, Krauss SM, Ohnuki M, Jocher J, Vieth B, Enard W, Hellmann I:

"Evidence for compensatory evolution within pleiotropic regulatory elements." (2024)

bioRxiv

doi: 10.1101/2024.01.10.575014



Evidence for compensatory evolution within pleiotropic regulatory elements

Zane Kliesmete[1], Peter Orchard[1,2], Victor Yan Kin Lee[1,3], Johanna Geuder[1], Simon M. Krauß[1,4], Mari Ohnuki[1,5], Jessica Jocher[1], Beate Vieth[1], Wolfgang Enard[1], Ines Hellmann[1,*]

[1] Anthropology and Human Genomics, Faculty of Biology, Ludwig-Maximilians Universität München, Munich, Germany

[2] Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI, USA

[3] Section for Molecular Ecology and Evolution, Globe Institute, University of Copenhagen, Copenhagen, Denmark

[4] Department of Hematology, Cell Therapy, Hemostaseology and Infectious Diseases, University Leipzig Medical Center, Leipzig, Germany

[5] Faculty of Medicine Bldg.B, Institute for the Advanced Study of Human Biology (ASHBi), Kyoto University, Kyoto, Japan

[*] correspondence:

Dr. Ines Hellmann,

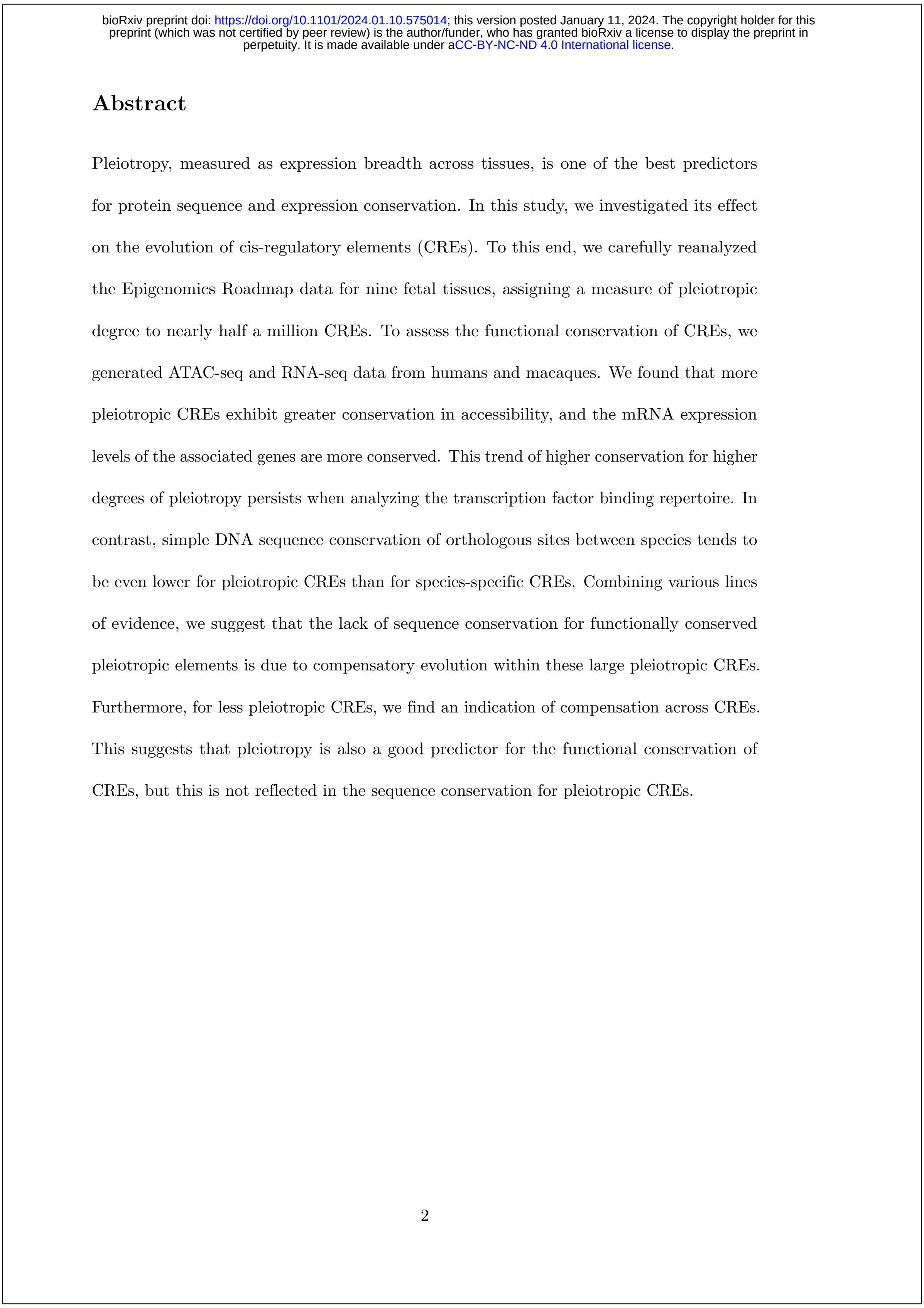
Telefon +49 (0)89 2180-74336

Telefax +49 (0)89 2180-74331

hellmann@bio.lmu.de, www.anthropologie.bio.lmu.de

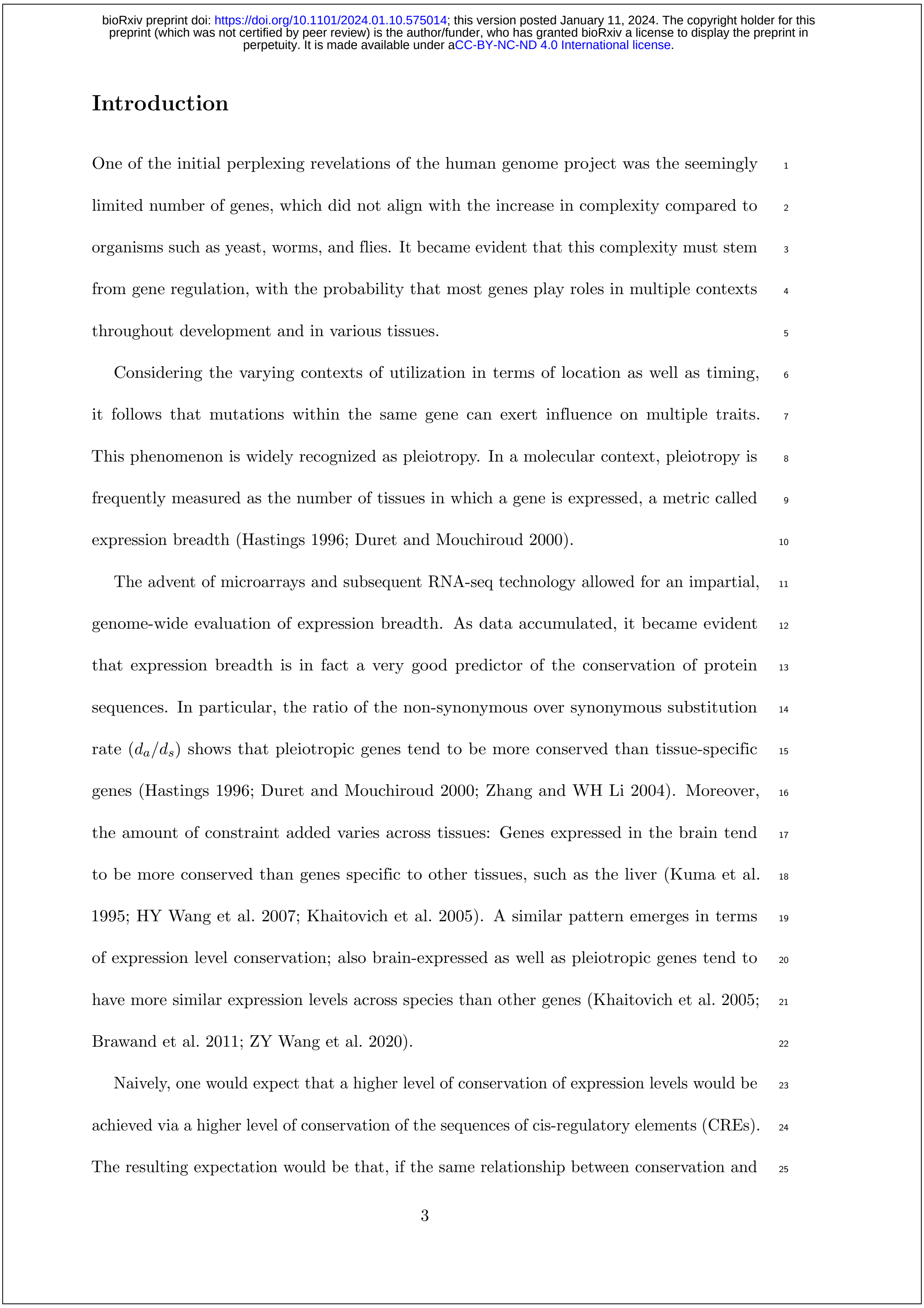
Keywords

Regulatory element evolution, pleiotropy, compensation, primates, cross-species, accessibility, expression, ATAC-seq, RNA-seq



Abstract

Pleiotropy, measured as expression breadth across tissues, is one of the best predictors for protein sequence and expression conservation. In this study, we investigated its effect on the evolution of cis-regulatory elements (CREs). To this end, we carefully reanalyzed the Epigenomics Roadmap data for nine fetal tissues, assigning a measure of pleiotropic degree to nearly half a million CREs. To assess the functional conservation of CREs, we generated ATAC-seq and RNA-seq data from humans and macaques. We found that more pleiotropic CREs exhibit greater conservation in accessibility, and the mRNA expression levels of the associated genes are more conserved. This trend of higher conservation for higher degrees of pleiotropy persists when analyzing the transcription factor binding repertoire. In contrast, simple DNA sequence conservation of orthologous sites between species tends to be even lower for pleiotropic CREs than for species-specific CREs. Combining various lines of evidence, we suggest that the lack of sequence conservation for functionally conserved pleiotropic elements is due to compensatory evolution within these large pleiotropic CREs. Furthermore, for less pleiotropic CREs, we find an indication of compensation across CREs. This suggests that pleiotropy is also a good predictor for the functional conservation of CREs, but this is not reflected in the sequence conservation for pleiotropic CREs.



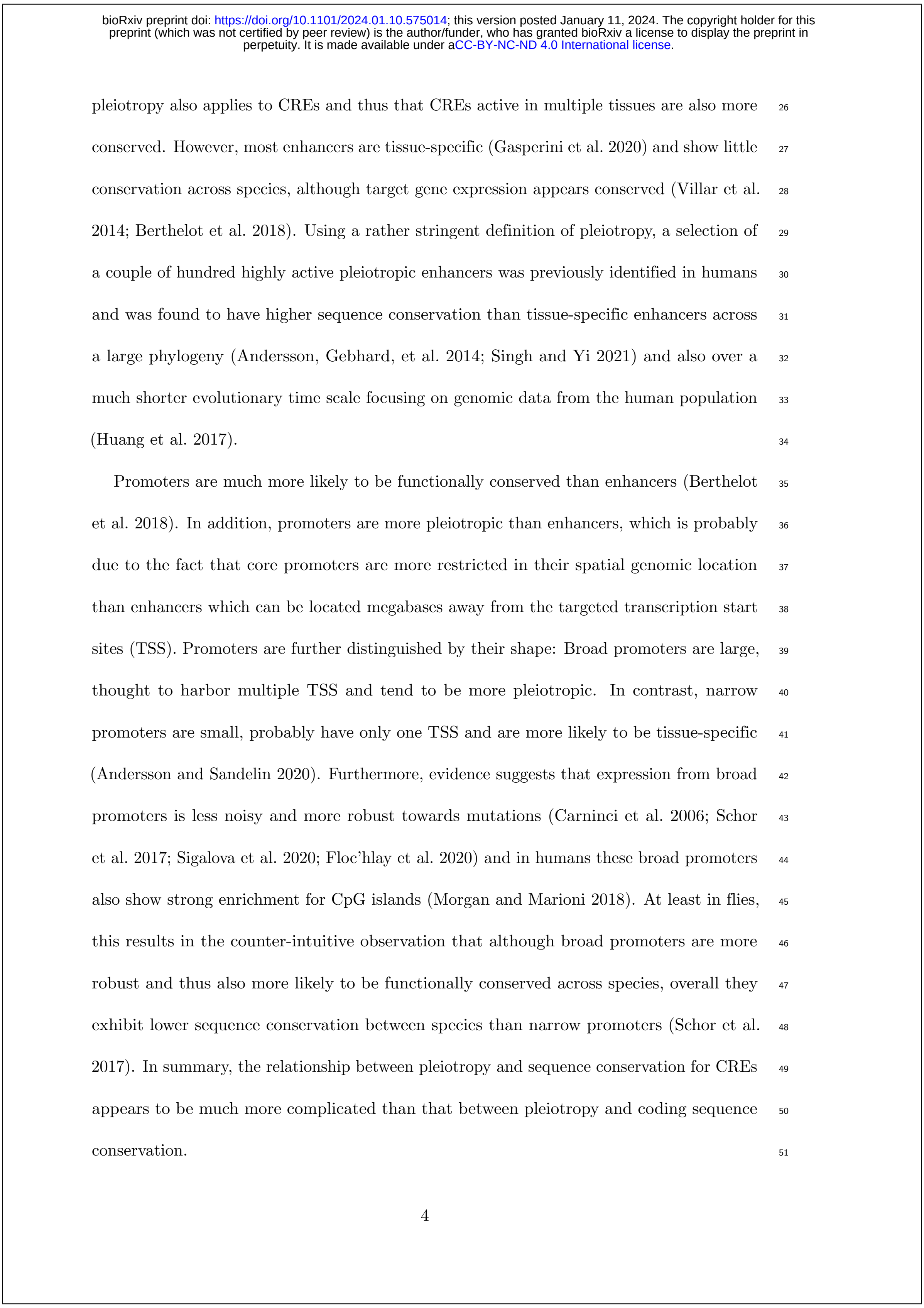
Introduction

One of the initial perplexing revelations of the human genome project was the seemingly limited number of genes, which did not align with the increase in complexity compared to organisms such as yeast, worms, and flies. It became evident that this complexity must stem from gene regulation, with the probability that most genes play roles in multiple contexts throughout development and in various tissues.

Considering the varying contexts of utilization in terms of location as well as timing, it follows that mutations within the same gene can exert influence on multiple traits. This phenomenon is widely recognized as pleiotropy. In a molecular context, pleiotropy is frequently measured as the number of tissues in which a gene is expressed, a metric called expression breadth (Hastings 1996; Duret and Mouchiroud 2000).

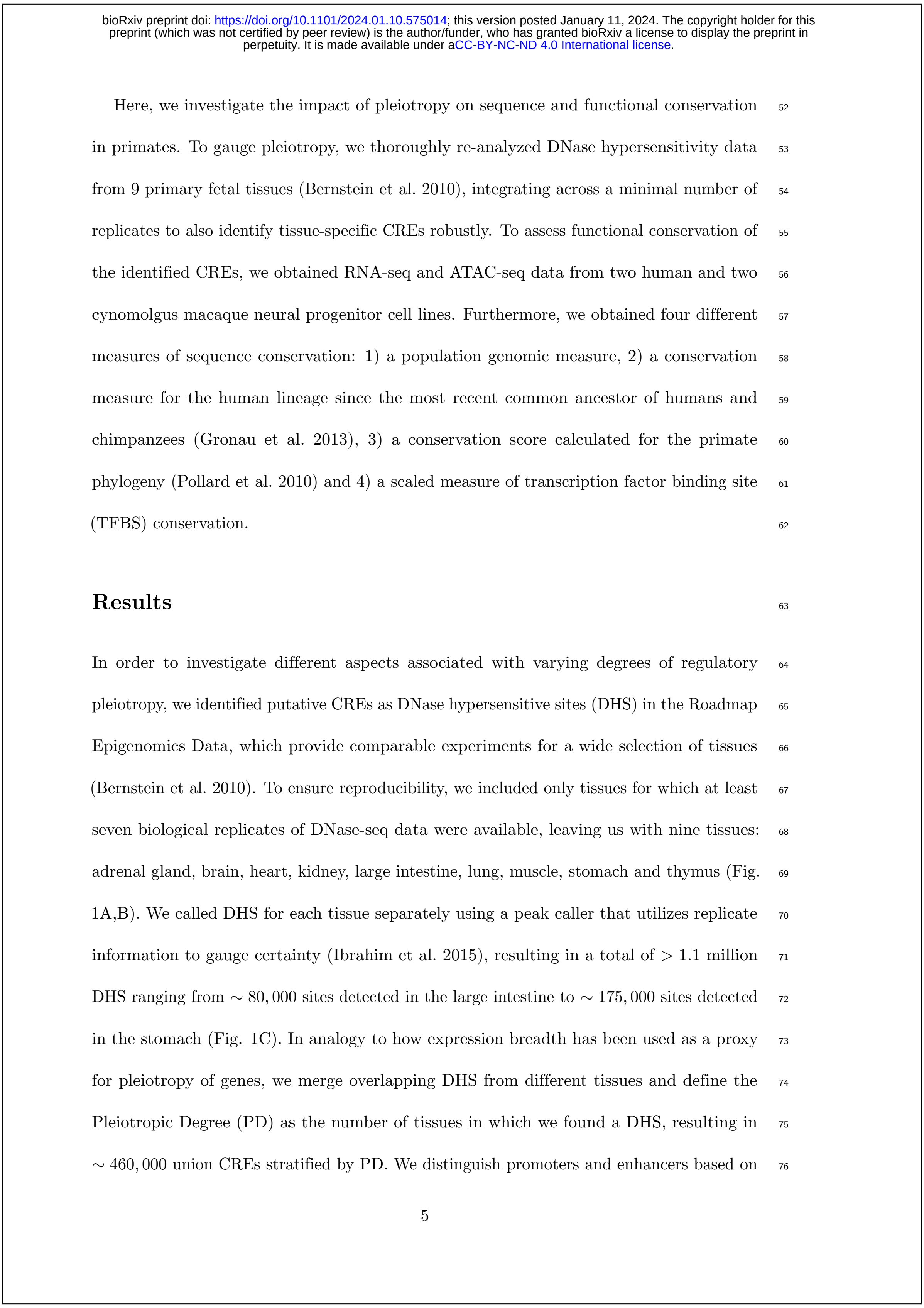
The advent of microarrays and subsequent RNA-seq technology allowed for an impartial, genome-wide evaluation of expression breadth. As data accumulated, it became evident that expression breadth is in fact a very good predictor of the conservation of protein sequences. In particular, the ratio of the non-synonymous over synonymous substitution rate (d_a/d_s) shows that pleiotropic genes tend to be more conserved than tissue-specific genes (Hastings 1996; Duret and Mouchiroud 2000; Zhang and WH Li 2004). Moreover, the amount of constraint added varies across tissues: Genes expressed in the brain tend to be more conserved than genes specific to other tissues, such as the liver (Kuma et al. 1995; HY Wang et al. 2007; Khaitovich et al. 2005). A similar pattern emerges in terms of expression level conservation; also brain-expressed as well as pleiotropic genes tend to have more similar expression levels across species than other genes (Khaitovich et al. 2005; Brawand et al. 2011; ZY Wang et al. 2020).

Naively, one would expect that a higher level of conservation of expression levels would be achieved via a higher level of conservation of the sequences of cis-regulatory elements (CREs). The resulting expectation would be that, if the same relationship between conservation and



pleiotropy also applies to CREs and thus that CREs active in multiple tissues are also more conserved. However, most enhancers are tissue-specific (Gasperini et al. 2020) and show little conservation across species, although target gene expression appears conserved (Villar et al. 2014; Berthelot et al. 2018). Using a rather stringent definition of pleiotropy, a selection of a couple of hundred highly active pleiotropic enhancers was previously identified in humans and was found to have higher sequence conservation than tissue-specific enhancers across a large phylogeny (Andersson, Gebhard, et al. 2014; Singh and Yi 2021) and also over a much shorter evolutionary time scale focusing on genomic data from the human population (Huang et al. 2017).

Promoters are much more likely to be functionally conserved than enhancers (Berthelot et al. 2018). In addition, promoters are more pleiotropic than enhancers, which is probably due to the fact that core promoters are more restricted in their spatial genomic location than enhancers which can be located megabases away from the targeted transcription start sites (TSS). Promoters are further distinguished by their shape: Broad promoters are large, thought to harbor multiple TSS and tend to be more pleiotropic. In contrast, narrow promoters are small, probably have only one TSS and are more likely to be tissue-specific (Andersson and Sandelin 2020). Furthermore, evidence suggests that expression from broad promoters is less noisy and more robust towards mutations (Carninci et al. 2006; Schor et al. 2017; Sigalova et al. 2020; Floc'hlay et al. 2020) and in humans these broad promoters also show strong enrichment for CpG islands (Morgan and Marioni 2018). At least in flies, this results in the counter-intuitive observation that although broad promoters are more robust and thus also more likely to be functionally conserved across species, overall they exhibit lower sequence conservation between species than narrow promoters (Schor et al. 2017). In summary, the relationship between pleiotropy and sequence conservation for CREs appears to be much more complicated than that between pleiotropy and coding sequence conservation.



Here, we investigate the impact of pleiotropy on sequence and functional conservation in primates. To gauge pleiotropy, we thoroughly re-analyzed DNase hypersensitivity data from 9 primary fetal tissues (Bernstein et al. 2010), integrating across a minimal number of replicates to also identify tissue-specific CREs robustly. To assess functional conservation of the identified CREs, we obtained RNA-seq and ATAC-seq data from two human and two cynomolgus macaque neural progenitor cell lines. Furthermore, we obtained four different measures of sequence conservation: 1) a population genomic measure, 2) a conservation measure for the human lineage since the most recent common ancestor of humans and chimpanzees (Gronau et al. 2013), 3) a conservation score calculated for the primate phylogeny (Pollard et al. 2010) and 4) a scaled measure of transcription factor binding site (TFBS) conservation.

Results

In order to investigate different aspects associated with varying degrees of regulatory pleiotropy, we identified putative CREs as DNase hypersensitive sites (DHS) in the Roadmap Epigenomics Data, which provide comparable experiments for a wide selection of tissues (Bernstein et al. 2010). To ensure reproducibility, we included only tissues for which at least seven biological replicates of DNase-seq data were available, leaving us with nine tissues: adrenal gland, brain, heart, kidney, large intestine, lung, muscle, stomach and thymus (Fig. 1A,B). We called DHS for each tissue separately using a peak caller that utilizes replicate information to gauge certainty (Ibrahim et al. 2015), resulting in a total of > 1.1 million DHS ranging from $\sim 80,000$ sites detected in the large intestine to $\sim 175,000$ sites detected in the stomach (Fig. 1C). In analogy to how expression breadth has been used as a proxy for pleiotropy of genes, we merge overlapping DHS from different tissues and define the Pleiotropic Degree (PD) as the number of tissues in which we found a DHS, resulting in $\sim 460,000$ union CREs stratified by PD. We distinguish promoters and enhancers based on genomic distance, while we designate CREs within 2kb of an active annotated TSS (Gencode v.32) as promoters and all other CREs within 1Mb as enhancers (Fishilevich et al. 2017; McLean et al. 2010).

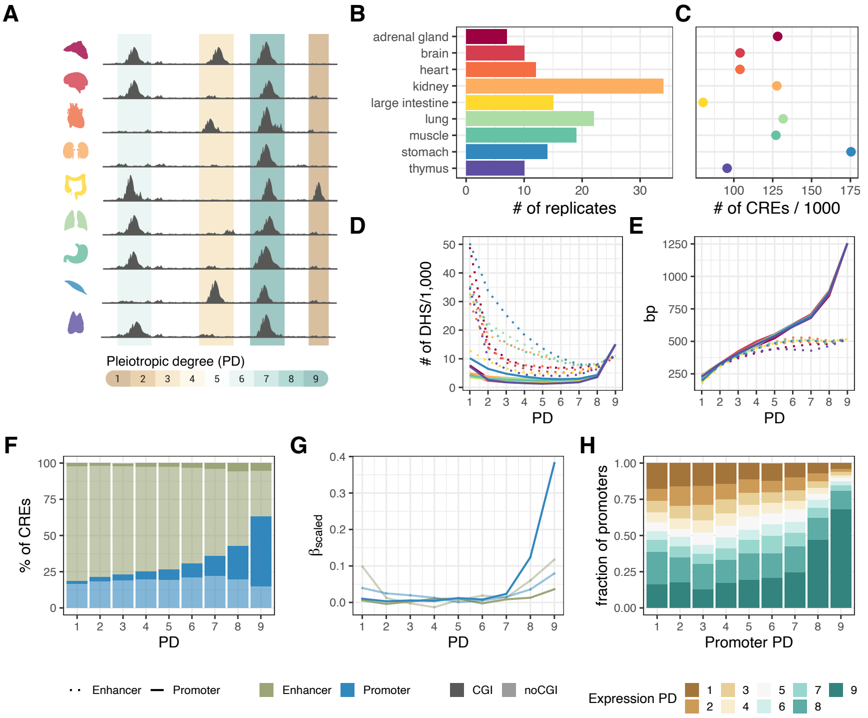
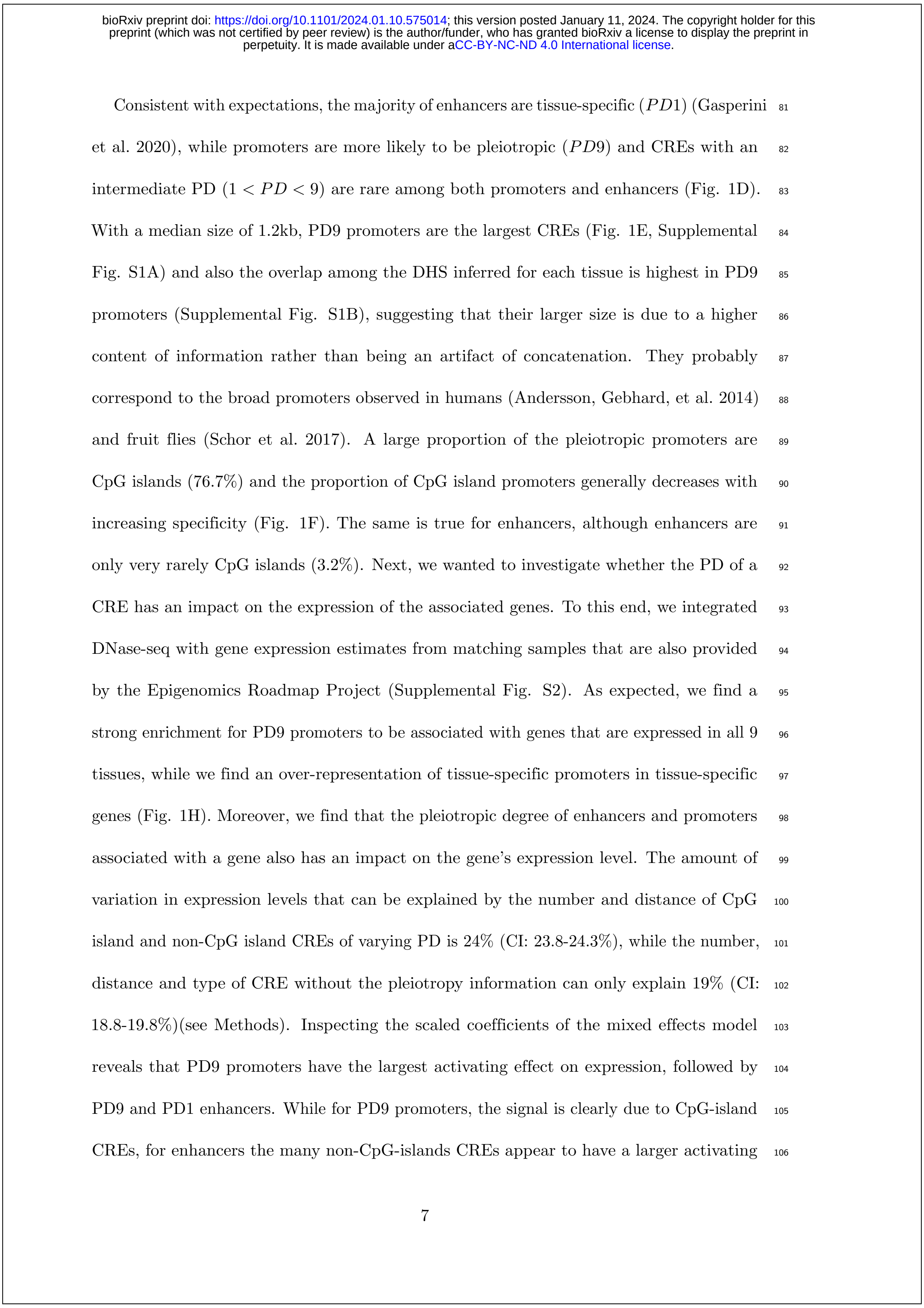
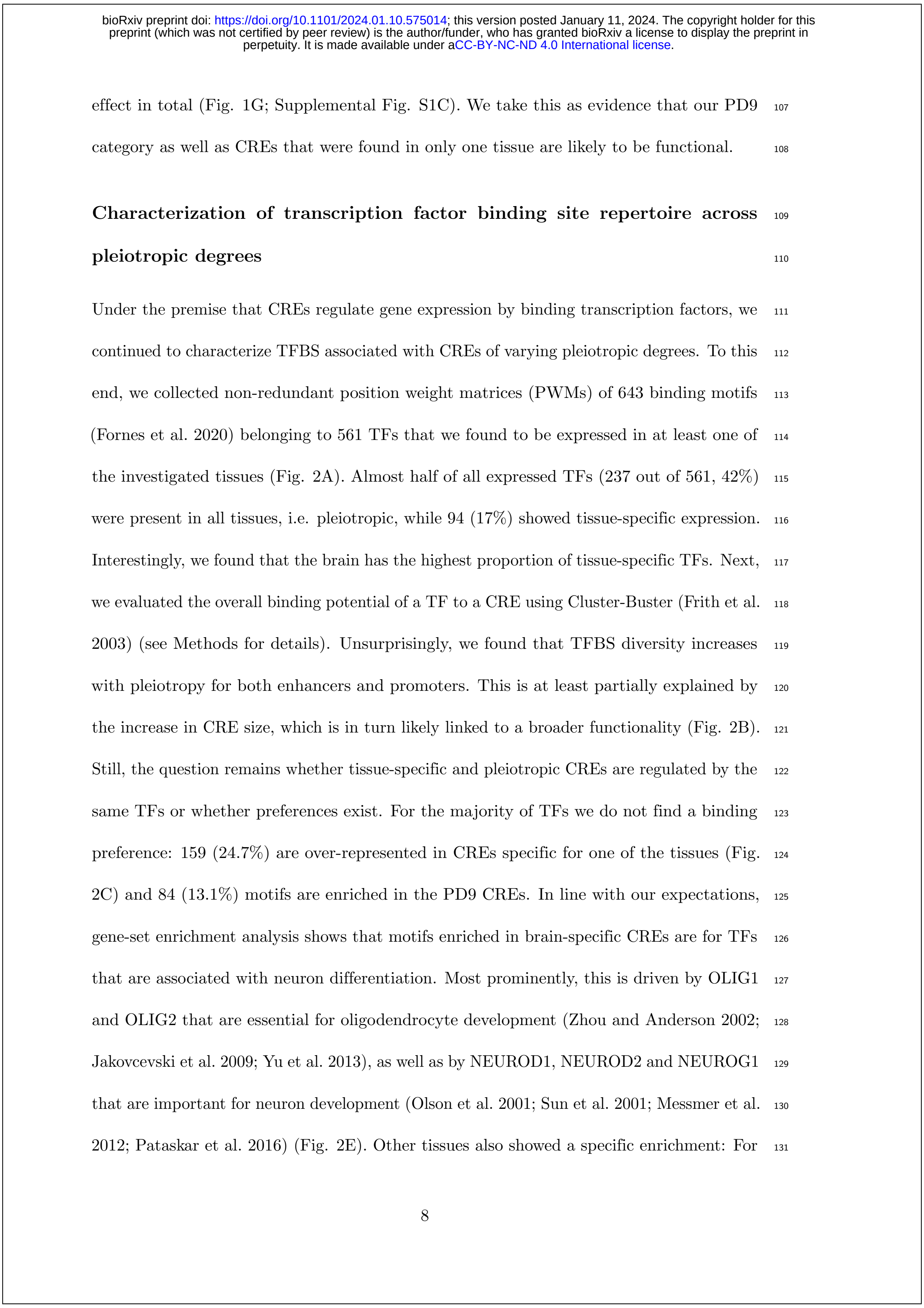


Figure 1. Study overview. (*A*) Open chromatin and expression data from the Roadmap Epigenomics Project (Bernstein et al. 2010) were used to infer the effect of pleiotropy on sequence and TFBS evolution, and associated gene expression in primates. Overlapping DHS peaks between tissues were merged to determine the degree of tissue-specificity per CRE. (*B*) DHS-data from 9 human fetal tissues. The number of biological replicates per tissue varies between 7 and 34. (*C*) The number of CREs per tissue varies 2.3-fold. There is no association between the number of replicates and the number of accessible regions per tissue, suggesting that with > 7 replicates per tissue, sufficient saturation is reached in peak detection. (*D*) Most enhancers (dotted line) are tissue-specific, while promoters (solid line) are mostly pleiotropic. The colors represent the tissues as introduced in (*A*,*B*). (*E*) CRE length increases with the number of tissues, particularly at the promoters. This increase was also observed at the peak level prior to merging (Supplemental Figure S1A). (*F*) The majority of PD9 CREs are CpG-island promoters (solid blue), while tissue-specific elements are rarely CpG-Islands and mainly enhancers (transparent green). (*G*) Scaled coefficients of a linear mixed model to predict gene expression levels using distance scaled CRE counts of different types. (*H*) Pleiotropic promoters are more commonly associated with pleiotropic gene expression patterns. The promoter PD indicates the highest PD of the associated promoters per gene. The y-axis shows the proportions of those x-categories (promoter PD) with associated gene expression pleiotropy ranging from 1 to 9.



Consistent with expectations, the majority of enhancers are tissue-specific ($PD1$) (Gasperini et al. 2020), while promoters are more likely to be pleiotropic ($PD9$) and CREs with an intermediate PD ($1 < PD < 9$) are rare among both promoters and enhancers (Fig. 1D). With a median size of 1.2kb, PD9 promoters are the largest CREs (Fig. 1E, Supplemental Fig. S1A) and also the overlap among the DHS inferred for each tissue is highest in PD9 promoters (Supplemental Fig. S1B), suggesting that their larger size is due to a higher content of information rather than being an artifact of concatenation. They probably correspond to the broad promoters observed in humans (Andersson, Gebhard, et al. 2014) and fruit flies (Schor et al. 2017). A large proportion of the pleiotropic promoters are CpG islands (76.7%) and the proportion of CpG island promoters generally decreases with increasing specificity (Fig. 1F). The same is true for enhancers, although enhancers are only very rarely CpG islands (3.2%). Next, we wanted to investigate whether the PD of a CRE has an impact on the expression of the associated genes. To this end, we integrated DNase-seq with gene expression estimates from matching samples that are also provided by the Epigenomics Roadmap Project (Supplemental Fig. S2). As expected, we find a strong enrichment for PD9 promoters to be associated with genes that are expressed in all 9 tissues, while we find an over-representation of tissue-specific promoters in tissue-specific genes (Fig. 1H). Moreover, we find that the pleiotropic degree of enhancers and promoters associated with a gene also has an impact on the gene's expression level. The amount of variation in expression levels that can be explained by the number and distance of CpG island and non-CpG island CREs of varying PD is 24% (CI: 23.8-24.3%), while the number, distance and type of CRE without the pleiotropy information can only explain 19% (CI: 18.8-19.8%)(see Methods). Inspecting the scaled coefficients of the mixed effects model reveals that PD9 promoters have the largest activating effect on expression, followed by PD9 and PD1 enhancers. While for PD9 promoters, the signal is clearly due to CpG-island CREs, for enhancers the many non-CpG-islands CREs appear to have a larger activating



effect in total (Fig. 1G; Supplemental Fig. S1C). We take this as evidence that our PD9 category as well as CREs that were found in only one tissue are likely to be functional.

Characterization of transcription factor binding site repertoire across pleiotropic degrees

Under the premise that CREs regulate gene expression by binding transcription factors, we continued to characterize TFBS associated with CREs of varying pleiotropic degrees. To this end, we collected non-redundant position weight matrices (PWMs) of 643 binding motifs (Fornes et al. 2020) belonging to 561 TFs that we found to be expressed in at least one of the investigated tissues (Fig. 2A). Almost half of all expressed TFs (237 out of 561, 42%) were present in all tissues, i.e. pleiotropic, while 94 (17%) showed tissue-specific expression. Interestingly, we found that the brain has the highest proportion of tissue-specific TFs. Next, we evaluated the overall binding potential of a TF to a CRE using Cluster-Buster (Frith et al. 2003) (see Methods for details). Unsurprisingly, we found that TFBS diversity increases with pleiotropy for both enhancers and promoters. This is at least partially explained by the increase in CRE size, which is in turn likely linked to a broader functionality (Fig. 2B). Still, the question remains whether tissue-specific and pleiotropic CREs are regulated by the same TFs or whether preferences exist. For the majority of TFs we do not find a binding preference: 159 (24.7%) are over-represented in CREs specific for one of the tissues (Fig. 2C) and 84 (13.1%) motifs are enriched in the PD9 CREs. In line with our expectations, gene-set enrichment analysis shows that motifs enriched in brain-specific CREs are for TFs that are associated with neuron differentiation. Most prominently, this is driven by OLIG1 and OLIG2 that are essential for oligodendrocyte development (Zhou and Anderson 2002; Jakovcevski et al. 2009; Yu et al. 2013), as well as by NEUROD1, NEUROD2 and NEUROG1 that are important for neuron development (Olson et al. 2001; Sun et al. 2001; Messmer et al. 2012; Pataskar et al. 2016) (Fig. 2E). Other tissues also showed a specific enrichment: For example, TFBS that are overrepresented in heart-specific CREs include motifs of MEF2C, TBX20 and NKX2-5 (Fig. 2F), which are essential for cardiac muscle development (He et al. 2011; Schlesinger et al. 2011; Grunert et al. 2016).

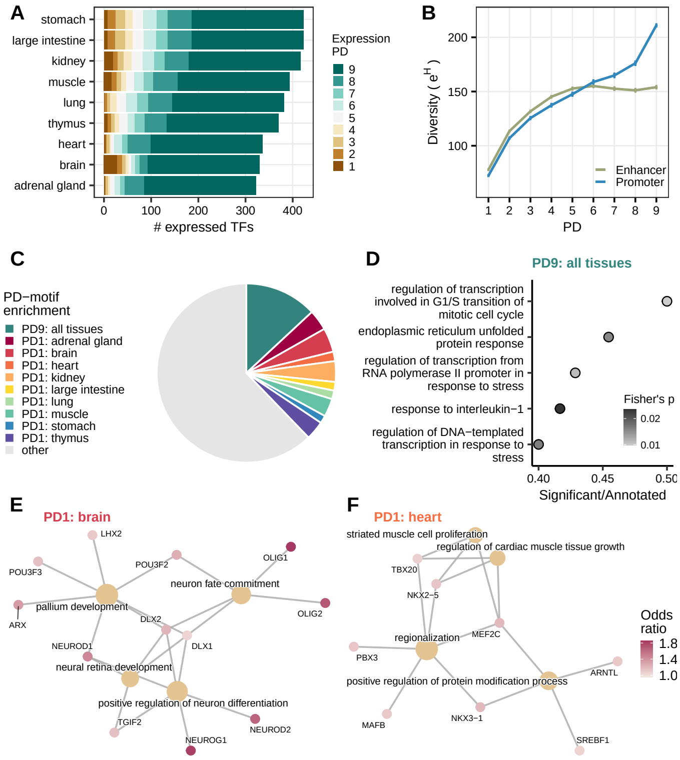
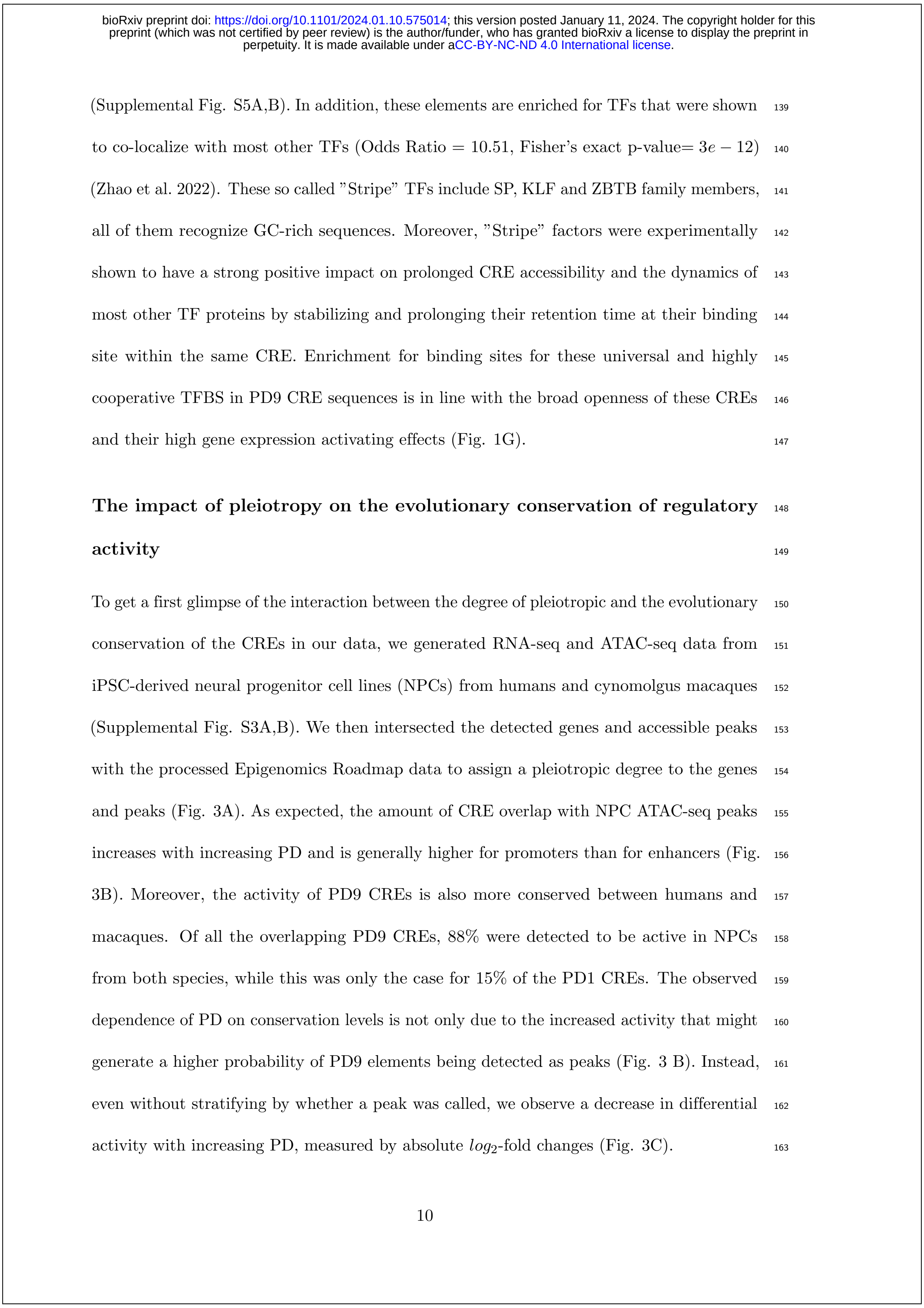


Figure 2. TFBS repertoire diversity and enrichment across tissue-specific and pleiotropic CREs. (*A*) An overview of TF expression across tissues. (*B*) TFBS repertoire diversity increases with PD, particularly across promoters. Depicted are mean +/- SEM. (*C*) Overview of the over-represented motifs in PD9 and PD1 CRE sequences. (*D*) Top 5 categories of gene set enrichment analysis of PD9-enriched motifs using all motifs as background (Gene ontology, Biological Process, Fisher's exact p-value< 0.05). (*E, F*) Top 4 categories of gene set enrichment analysis of tissue-specific PD1 enriched motifs using all motifs as background (Gene ontology, Biological Process, Fisher's exact p-value< 0.05). Fold change depicts the proportion of tissue-specific PD1 CREs with the motif over the global average proportion for that motif. (*E*) Brain-specific PD1 over-represented motifs. (*F*) Heart-specific PD1 over-represented motifs.

In contrast, TFs that show a binding preference for PD9 CREs appear to be associated with more basic cellular processes such as transcription regulation in connection with cell cycle and and stress response (Fig. 2D). These motifs are more GC-rich and tend to have a higher information content than PD1-enriched motifs or motifs without any preference



(Supplemental Fig. S5A,B). In addition, these elements are enriched for TFs that were shown to co-localize with most other TFs (Odds Ratio = 10.51, Fisher's exact p-value= $3e - 12$) (Zhao et al. 2022). These so called "Stripe" TFs include SP, KLF and ZBTB family members, all of them recognize GC-rich sequences. Moreover, "Stripe" factors were experimentally shown to have a strong positive impact on prolonged CRE accessibility and the dynamics of most other TF proteins by stabilizing and prolonging their retention time at their binding site within the same CRE. Enrichment for binding sites for these universal and highly cooperative TFBS in PD9 CRE sequences is in line with the broad openness of these CREs and their high gene expression activating effects (Fig. 1G).

The impact of pleiotropy on the evolutionary conservation of regulatory activity

To get a first glimpse of the interaction between the degree of pleiotropic and the evolutionary conservation of the CREs in our data, we generated RNA-seq and ATAC-seq data from iPSC-derived neural progenitor cell lines (NPCs) from humans and cynomolgus macaques (Supplemental Fig. S3A,B). We then intersected the detected genes and accessible peaks with the processed Epigenomics Roadmap data to assign a pleiotropic degree to the genes and peaks (Fig. 3A). As expected, the amount of CRE overlap with NPC ATAC-seq peaks increases with increasing PD and is generally higher for promoters than for enhancers (Fig. 3B). Moreover, the activity of PD9 CREs is also more conserved between humans and macaques. Of all the overlapping PD9 CREs, 88% were detected to be active in NPCs from both species, while this was only the case for 15% of the PD1 CREs. The observed dependence of PD on conservation levels is not only due to the increased activity that might generate a higher probability of PD9 elements being detected as peaks (Fig. 3 B). Instead, even without stratifying by whether a peak was called, we observe a decrease in differential activity with increasing PD, measured by absolute log_2-fold changes (Fig. 3C).

Next, we wanted to investigate which changes in CRE activity have an impact on the expression of the associated genes. To this end, we tested whether differentially accessible (DA) promoters and enhancers (BH-adjusted Wald test p-value <= 0.1) of a PD category are more likely to be associated with a differentially expressed (DE) gene (BH-adjusted Wald test p-value <= 0.1). Indeed, we find that DA promoters are more likely to be associated with a DE gene (BH adjusted Fisher's exact test). There is a clear enrichment for all promoter PD categories, showing a 2-3 times enrichment (Fig. 3D). Moreover, when we further distinguish CpG island promoters, it turns out that the activity changes there have the greatest potential for downstream effects (Supplemental Fig. S3C,D).

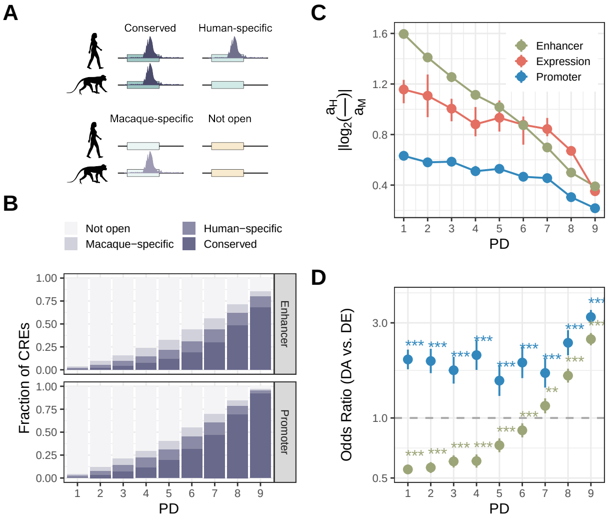
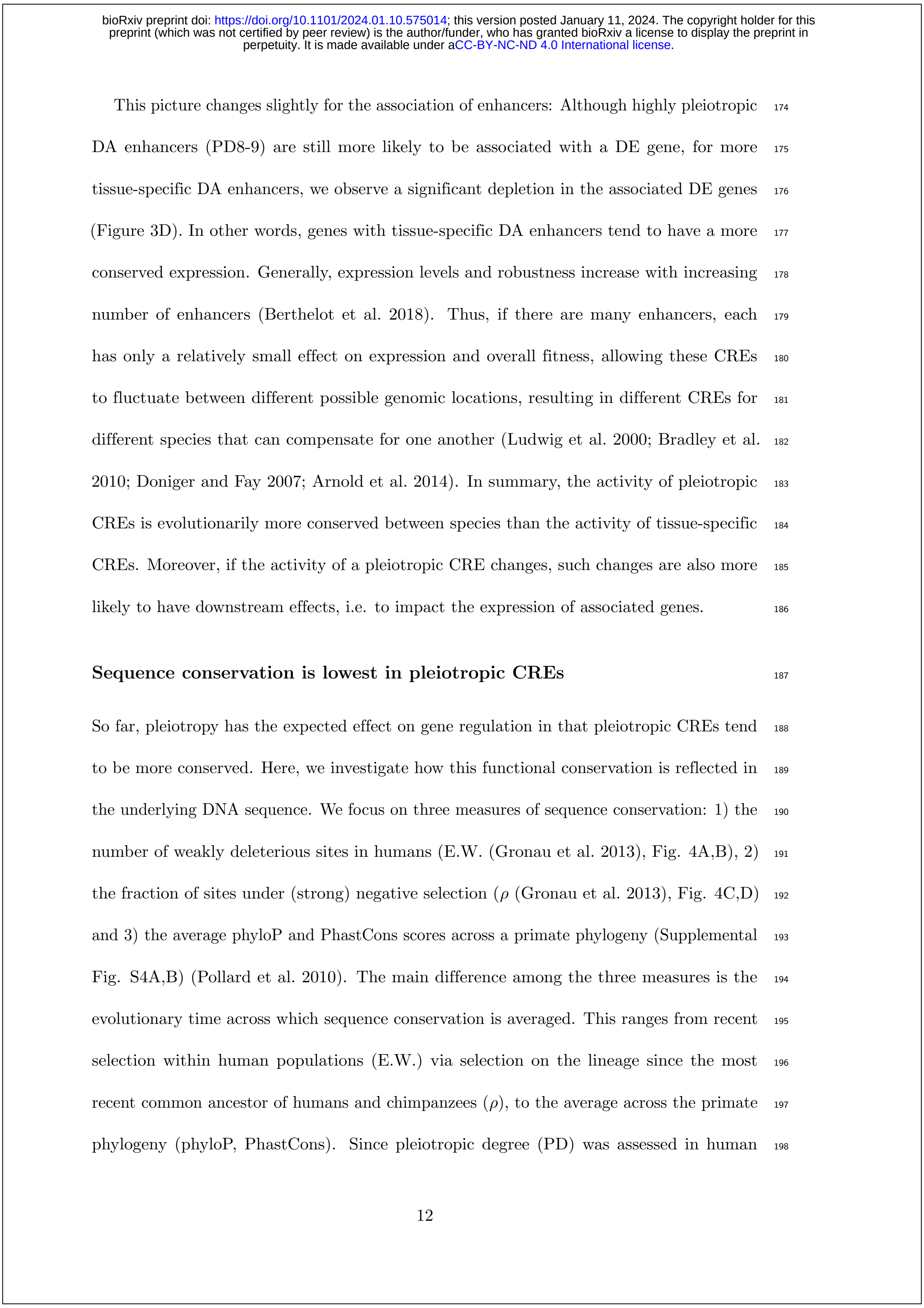


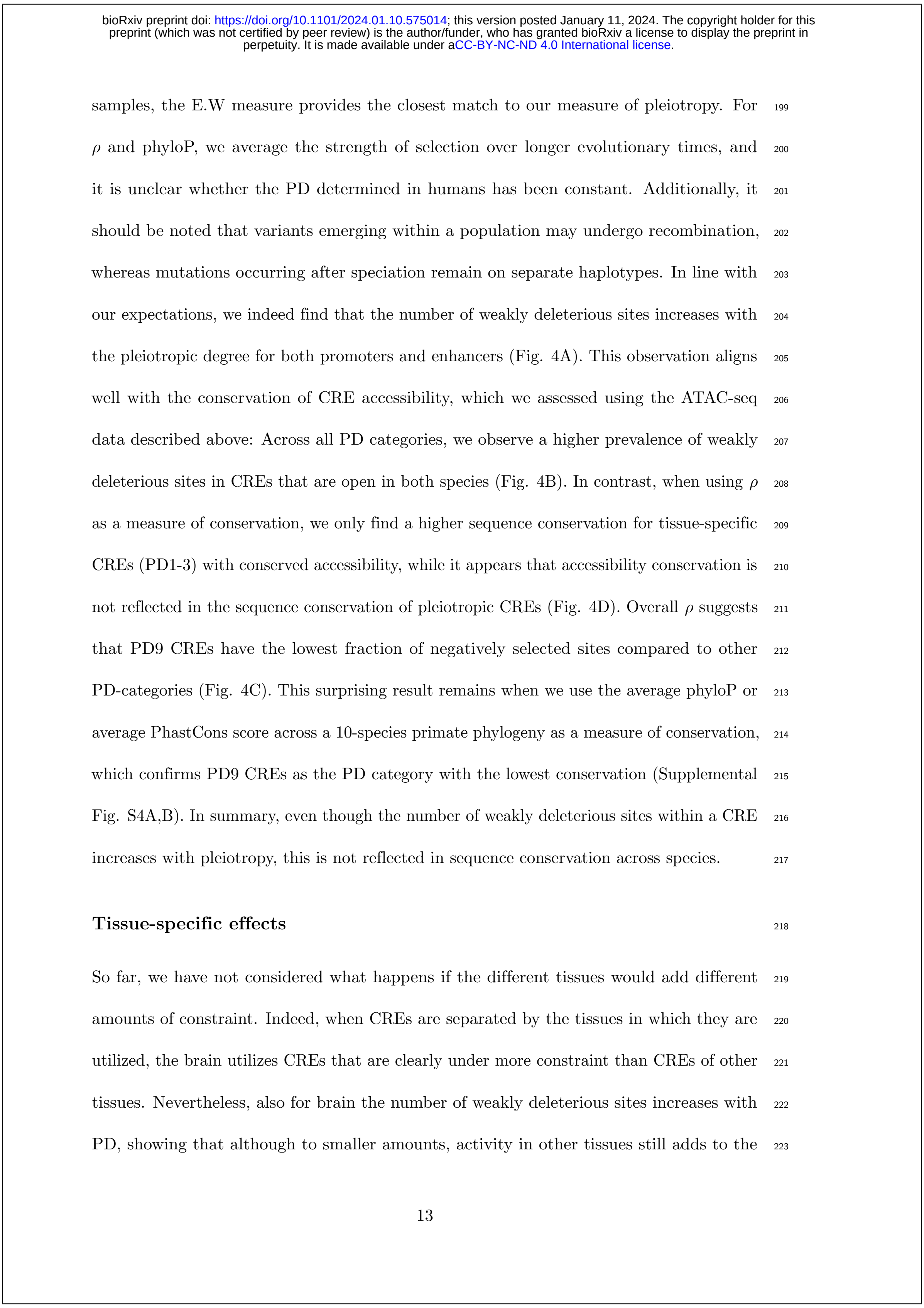
Figure 3. Pleiotropic degree and evolutionary conservation of expression and accessibility between humans and cynomolgus macaques. (*A*,*B*) The fraction of enhancers and promoters of different pleiotropic degrees (PD) as defined using data from 9 tissues from the Epigenomics Roadmap project, which overlapped with ATAC-seq peaks called in neural progenitor cell lines (NPCs) from cynomolgus macaques and humans. The colors indicate whether a human DHS-derived CRE overlapped with a NPC ATAC-seq peak from humans, cynomolgus macaques, or both. (*C*) Mean absolute log$_2$-fold changes of gene expression and activities between humans and cynomolgus macaques. The error bars represent 95% bootstrap confidence intervals. PD9 genes (CREs) have more conserved expression (activity) than more tissue-specific genes. (*D*) We tested for enrichment (odds ratio >1) or depletion (odds ratio <1) of differentially accessible CREs with significantly differentially expressed genes between humans and cynomolgus macaques. Error bars represent the 95% confidence intervals of the odd ratio, and the stars indicate the significance level with Benjamini-Hochberg correction ($\cdot < 0.1$, $^{*} < 0.05$, $^{**} < 0.01$, $^{***} < 0.001$).



This picture changes slightly for the association of enhancers: Although highly pleiotropic DA enhancers (PD8-9) are still more likely to be associated with a DE gene, for more tissue-specific DA enhancers, we observe a significant depletion in the associated DE genes (Figure 3D). In other words, genes with tissue-specific DA enhancers tend to have a more conserved expression. Generally, expression levels and robustness increase with increasing number of enhancers (Berthelot et al. 2018). Thus, if there are many enhancers, each has only a relatively small effect on expression and overall fitness, allowing these CREs to fluctuate between different possible genomic locations, resulting in different CREs for different species that can compensate for one another (Ludwig et al. 2000; Bradley et al. 2010; Doniger and Fay 2007; Arnold et al. 2014). In summary, the activity of pleiotropic CREs is evolutionarily more conserved between species than the activity of tissue-specific CREs. Moreover, if the activity of a pleiotropic CRE changes, such changes are also more likely to have downstream effects, i.e. to impact the expression of associated genes.

Sequence conservation is lowest in pleiotropic CREs

So far, pleiotropy has the expected effect on gene regulation in that pleiotropic CREs tend to be more conserved. Here, we investigate how this functional conservation is reflected in the underlying DNA sequence. We focus on three measures of sequence conservation: 1) the number of weakly deleterious sites in humans (E.W. (Gronau et al. 2013), Fig. 4A,B), 2) the fraction of sites under (strong) negative selection (ρ (Gronau et al. 2013), Fig. 4C,D) and 3) the average phyloP and PhastCons scores across a primate phylogeny (Supplemental Fig. S4A,B) (Pollard et al. 2010). The main difference among the three measures is the evolutionary time across which sequence conservation is averaged. This ranges from recent selection within human populations (E.W.) via selection on the lineage since the most recent common ancestor of humans and chimpanzees (ρ), to the average across the primate phylogeny (phyloP, PhastCons). Since pleiotropic degree (PD) was assessed in human



samples, the E.W measure provides the closest match to our measure of pleiotropy. For ρ and phyloP, we average the strength of selection over longer evolutionary times, and it is unclear whether the PD determined in humans has been constant. Additionally, it should be noted that variants emerging within a population may undergo recombination, whereas mutations occurring after speciation remain on separate haplotypes. In line with our expectations, we indeed find that the number of weakly deleterious sites increases with the pleiotropic degree for both promoters and enhancers (Fig. 4A). This observation aligns well with the conservation of CRE accessibility, which we assessed using the ATAC-seq data described above: Across all PD categories, we observe a higher prevalence of weakly deleterious sites in CREs that are open in both species (Fig. 4B). In contrast, when using ρ as a measure of conservation, we only find a higher sequence conservation for tissue-specific CREs (PD1-3) with conserved accessibility, while it appears that accessibility conservation is not reflected in the sequence conservation of pleiotropic CREs (Fig. 4D). Overall ρ suggests that PD9 CREs have the lowest fraction of negatively selected sites compared to other PD-categories (Fig. 4C). This surprising result remains when we use the average phyloP or average PhastCons score across a 10-species primate phylogeny as a measure of conservation, which confirms PD9 CREs as the PD category with the lowest conservation (Supplemental Fig. S4A,B). In summary, even though the number of weakly deleterious sites within a CRE increases with pleiotropy, this is not reflected in sequence conservation across species.

Tissue-specific effects

So far, we have not considered what happens if the different tissues would add different amounts of constraint. Indeed, when CREs are separated by the tissues in which they are utilized, the brain utilizes CREs that are clearly under more constraint than CREs of other tissues. Nevertheless, also for brain the number of weakly deleterious sites increases with PD, showing that although to smaller amounts, activity in other tissues still adds to the

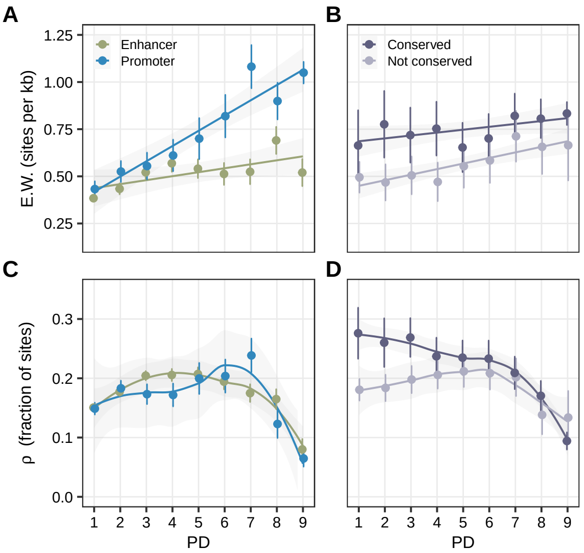


Figure 4. CRE sequence conservation patterns across varying degrees of pleiotropy. (*A*,*B*) Weak negative selection inferred based on human polymorphisms increases with increasing pleiotropic degree (PD). (*A*) Separated by enhancers / promoters. (*B*) Separated by human-macaque accessibility conservation in NPCs. (*C*,*D*) (Strong) negative selection is the highest at the intermediately-specific CREs and lowest in the pleiotropic CRE sequences. (*C*) Separated by enhancers / promoters. (*D*) Separated by human-macaque accessibility conservation in NPCs. (*A*,*B*,*C*,*D*) Depicted are mean estimates per PD category. Error bars indicate SEM.

overall constraint (Fig. 5A). Again, this is not true when considering substitutions on the human lineage as used in the measure ρ (Fig. 5B). Here brain-specific CREs show most constraint on the human lineage, much more than pleiotropic PD9 CREs, which are by definition also utilized in the brain.

To exclude the possibility that the brain effect on the PD9 elements is diluted by the merging of DHS across tissues, we contrast the ρ of the brain peak sequence with adjacent sequences that are part of the same merged CRE but are open in other tissues (Fig. 5C). We find that for PD9 CREs, brain peak sequences show lower sequence conservation on the human lineage than the adjacent sequence utilized only by other tissues, while for less pleiotropic CREs the part that is used in the brain is under much more constraint (Fig. 5D). In summary, even though we find tissue-specific effects, in particular a higher

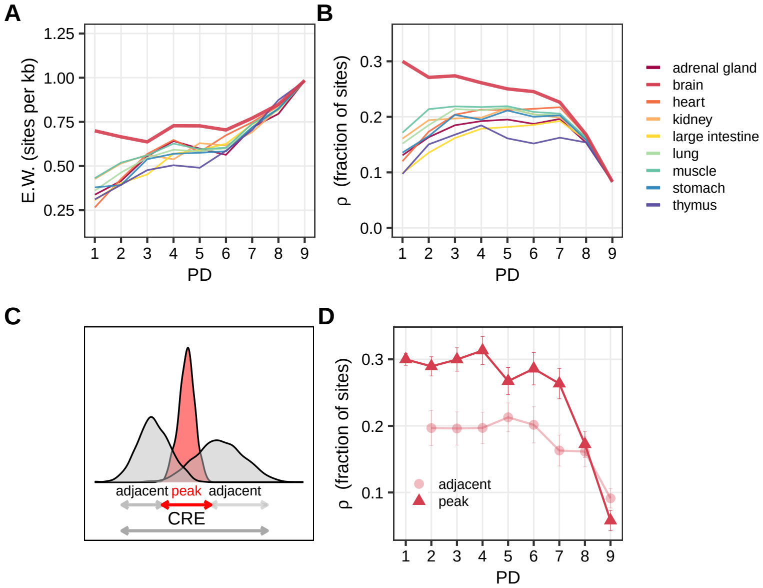
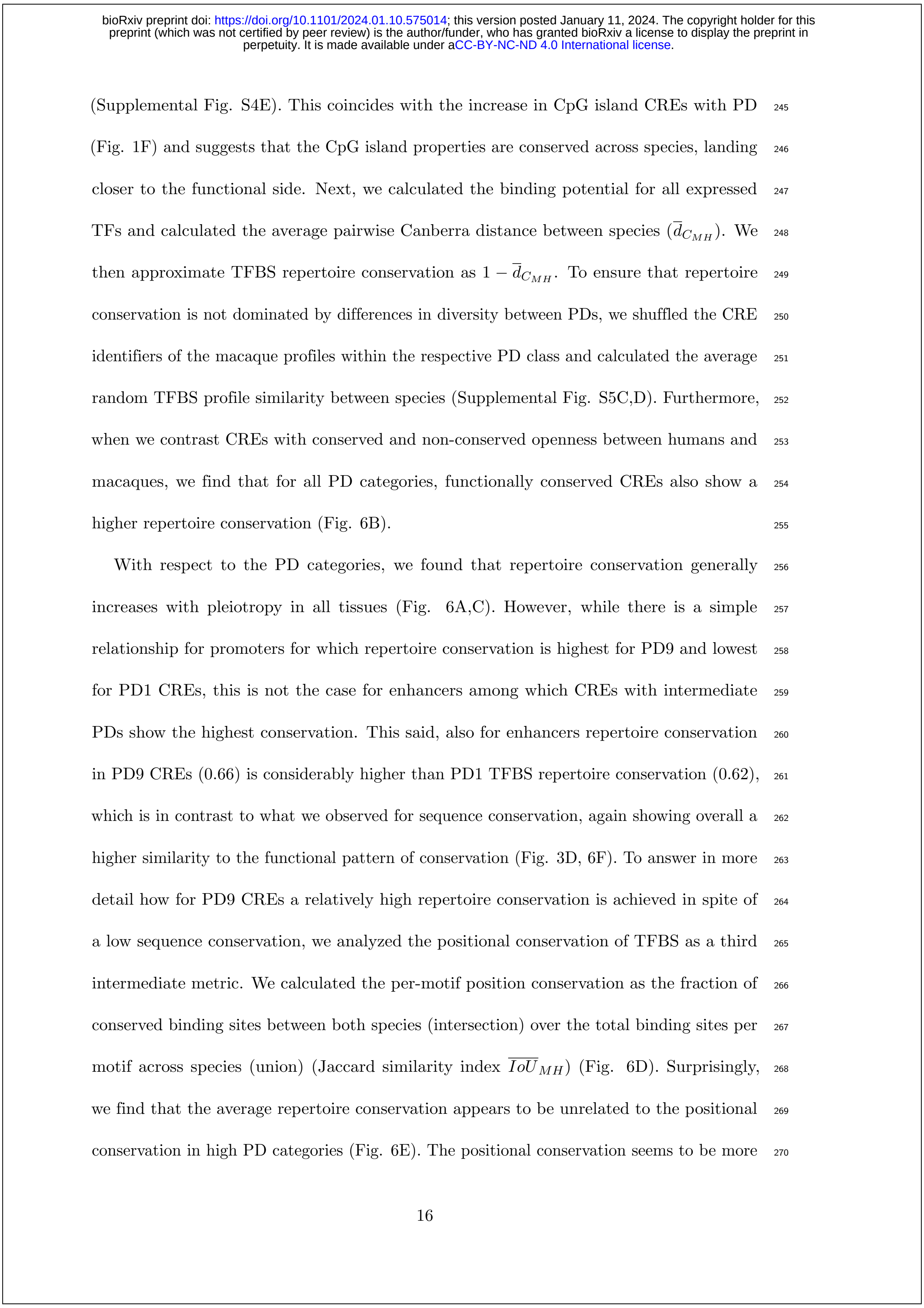


Figure 5. CRE sequence conservation patterns per tissue. (*A*) Weak negative selection inferred based on human polymorphisms separated by the tissue that utilizes the CREs. (*B*) All negative selection separated by the tissue that utilizes the CREs. (*C*) Brain CRE sequences, which showed the highest conservation across tissues, were separated into peak and adjacent sequences. (*D*) The part of the sequence that is used by the brain shows much higher fraction of sites under negative selection than the respective adjacent sequences. (*A*,*B*,*D*) Depicted are mean estimates per PD category. Error bars indicate SEM.

constraint for brain CREs, this cannot explain the overall pattern of the relatively low sequence conservation of pleiotropic CREs. It remains that for pleiotropic CREs there is no simple relationship between sequence and functional conservation between species.

Pleiotropic CRE TF repertoire is conserved, not the binding sites

In order to explain the apparent mismatch between functional and sequence conservation in PD9 CREs across primates, we continued to analyze levels that are intermediate between sequence conservation (less functional) and accessibility conservation (more functional), which are CpG content, TFBS repertoire and position conservation between human CREs and their orthologous sequences in cynomolgus macaques. To begin with, we find that conservation of CpG content increases with PD and is highest for pleiotropic promoters



(Supplemental Fig. S4E). This coincides with the increase in CpG island CREs with PD (Fig. 1F) and suggests that the CpG island properties are conserved across species, landing closer to the functional side. Next, we calculated the binding potential for all expressed TFs and calculated the average pairwise Canberra distance between species ($\overline{d}_{C_{MH}}$). We then approximate TFBS repertoire conservation as $1 - \overline{d}_{C_{MH}}$. To ensure that repertoire conservation is not dominated by differences in diversity between PDs, we shuffled the CRE identifiers of the macaque profiles within the respective PD class and calculated the average random TFBS profile similarity between species (Supplemental Fig. S5C,D). Furthermore, when we contrast CREs with conserved and non-conserved openness between humans and macaques, we find that for all PD categories, functionally conserved CREs also show a higher repertoire conservation (Fig. 6B).

With respect to the PD categories, we found that repertoire conservation generally increases with pleiotropy in all tissues (Fig. 6A,C). However, while there is a simple relationship for promoters for which repertoire conservation is highest for PD9 and lowest for PD1 CREs, this is not the case for enhancers among which CREs with intermediate PDs show the highest conservation. This said, also for enhancers repertoire conservation in PD9 CREs (0.66) is considerably higher than PD1 TFBS repertoire conservation (0.62), which is in contrast to what we observed for sequence conservation, again showing overall a higher similarity to the functional pattern of conservation (Fig. 3D, 6F). To answer in more detail how for PD9 CREs a relatively high repertoire conservation is achieved in spite of a low sequence conservation, we analyzed the positional conservation of TFBS as a third intermediate metric. We calculated the per-motif position conservation as the fraction of conserved binding sites between both species (intersection) over the total binding sites per motif across species (union) (Jaccard similarity index $\overline{IoU}_{MH}$) (Fig. 6D). Surprisingly, we find that the average repertoire conservation appears to be unrelated to the positional conservation in high PD categories (Fig. 6E). The positional conservation seems to be more

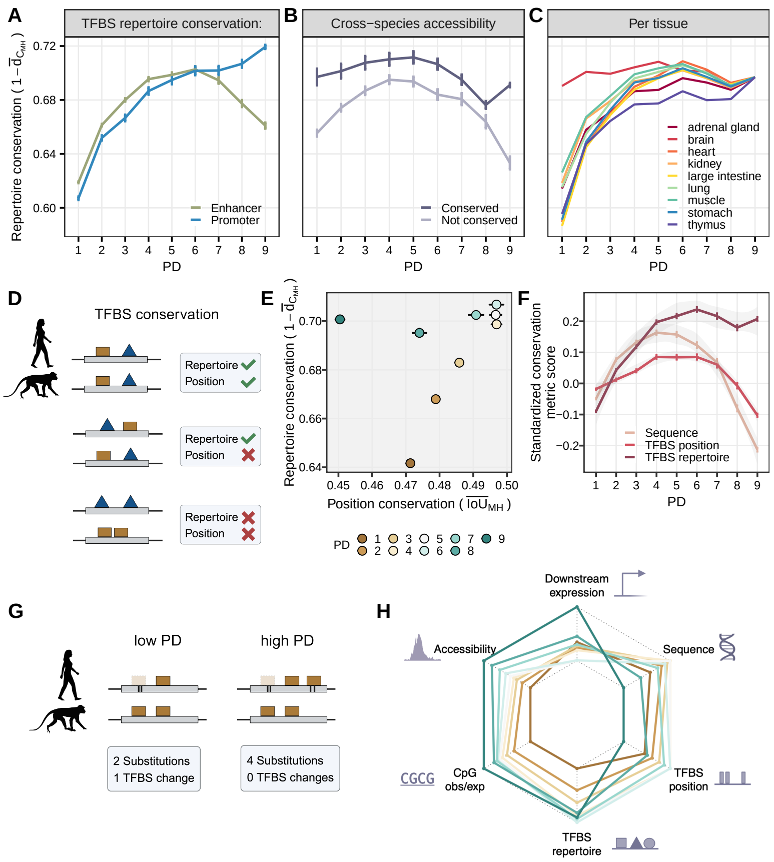
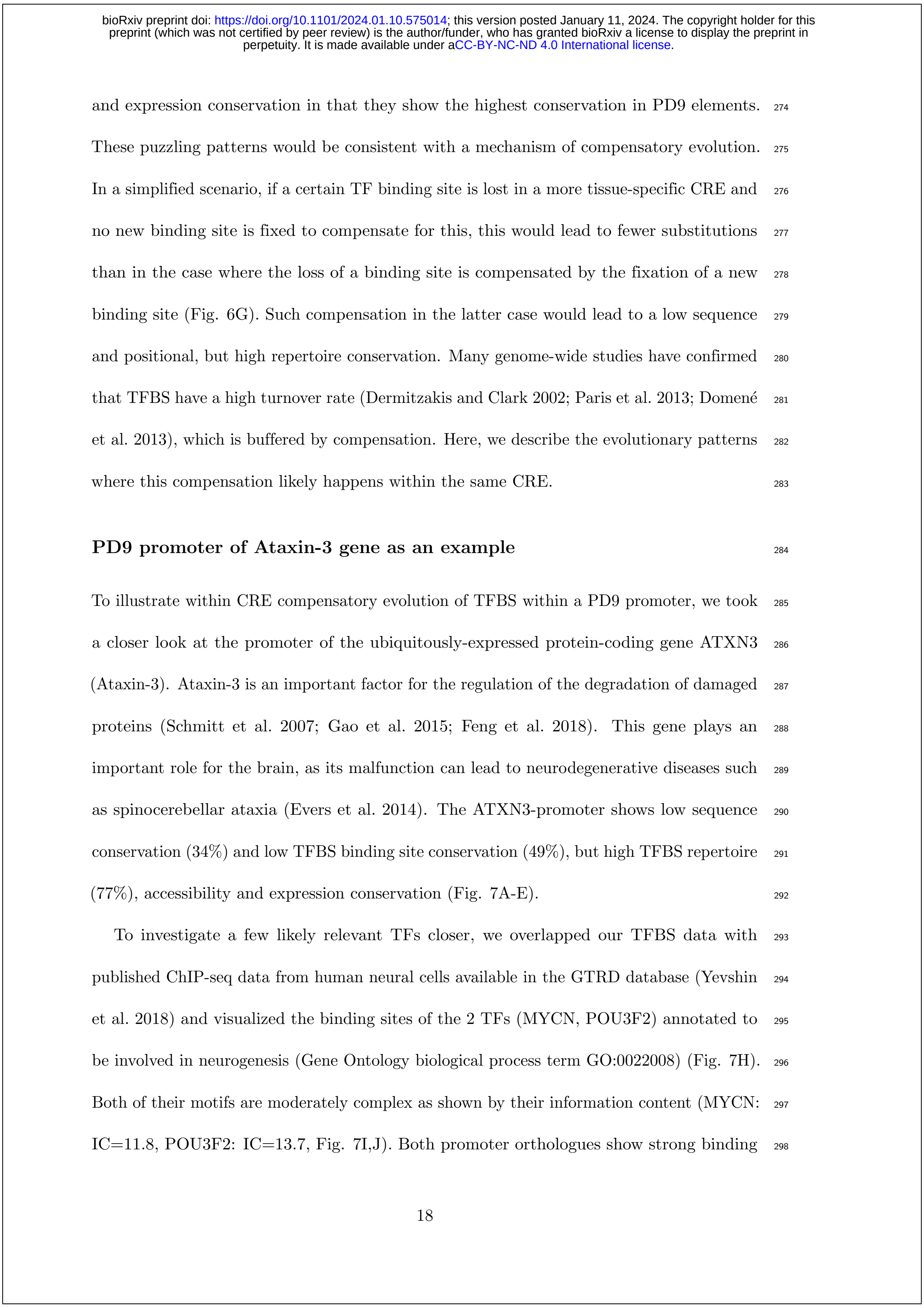


Figure 6. TFBS repertoire and position conservation between orthologous human and macaque CREs. (*A,B,C*) TFBS repertoire conservation across PDs. Depicted are mean +/- SEM. (*A*) TFBS repertoire conservation increases with higher PD among promoters, however, it decreases slightly at high PD-enhancers. (*B*) CREs that overlap NPC peaks with conserved openness show higher TFBS repertoire conservation than species-specific NPC peaks. (*C*) TFBS repertoire conservation differs across tissues, where brain shows the highest conservation at lower PDs. (*D*) Simplified schematic of the measures of repertoire and position conservation. (*E*) TFBS position conservation versus repertoire conservation across PD categories. Depicted are mean values +/- SEM. (*F*) Standardized scores (z-scores) of sequence (primate phyloP), TFBS repertoire and binding site conservation between human and cynomolgus macaque. (*G*) A schematic depicting how lower sequence conservation might lead to higher TFBS repertoire conservation through compensatory mechanisms. (*H*) A summary of the scaled average conservation metric scores across PDs. Sequence: primate phyloP scores, TFBS position: $\overline{IoU}_{MH}$ scores, TFBS repertoire: 1-$\bar{d}_{C_{MH}}$, CpG observed/expected: $|CpG\frac{obs}{exp}_M - CpG\frac{obs}{exp}_H|$, accessibility: —LFC— of NPC-DA results, downstream expression: —LFC— of NPC-DE.

related to sequence conservation, thus landing on the less functional side (Fig. 6F). In summary, while CRE sequence and TFBS positions are least conserved in PD9 elements, CpG content and TFBS-repertoire are in agreement with the more functional metrics accessibility



and expression conservation in that they show the highest conservation in PD9 elements. These puzzling patterns would be consistent with a mechanism of compensatory evolution. In a simplified scenario, if a certain TF binding site is lost in a more tissue-specific CRE and no new binding site is fixed to compensate for this, this would lead to fewer substitutions than in the case where the loss of a binding site is compensated by the fixation of a new binding site (Fig. 6G). Such compensation in the latter case would lead to a low sequence and positional, but high repertoire conservation. Many genome-wide studies have confirmed that TFBS have a high turnover rate (Dermitzakis and Clark 2002; Paris et al. 2013; Domené et al. 2013), which is buffered by compensation. Here, we describe the evolutionary patterns where this compensation likely happens within the same CRE.

PD9 promoter of Ataxin-3 gene as an example

To illustrate within CRE compensatory evolution of TFBS within a PD9 promoter, we took a closer look at the promoter of the ubiquitously-expressed protein-coding gene ATXN3 (Ataxin-3). Ataxin-3 is an important factor for the regulation of the degradation of damaged proteins (Schmitt et al. 2007; Gao et al. 2015; Feng et al. 2018). This gene plays an important role for the brain, as its malfunction can lead to neurodegenerative diseases such as spinocerebellar ataxia (Evers et al. 2014). The ATXN3-promoter shows low sequence conservation (34%) and low TFBS binding site conservation (49%), but high TFBS repertoire (77%), accessibility and expression conservation (Fig. 7A-E).

To investigate a few likely relevant TFs closer, we overlapped our TFBS data with published ChIP-seq data from human neural cells available in the GTRD database (Yevshin et al. 2018) and visualized the binding sites of the 2 TFs (MYCN, POU3F2) annotated to be involved in neurogenesis (Gene Ontology biological process term GO:0022008) (Fig. 7H). Both of their motifs are moderately complex as shown by their information content (MYCN: IC=11.8, POU3F2: IC=13.7, Fig. 7I,J). Both promoter orthologues show strong binding

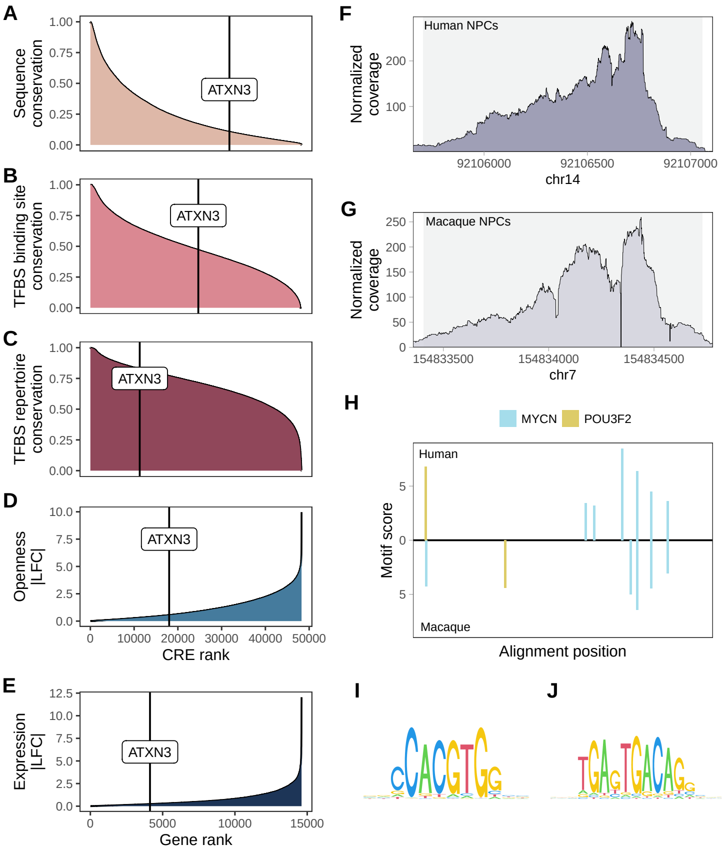
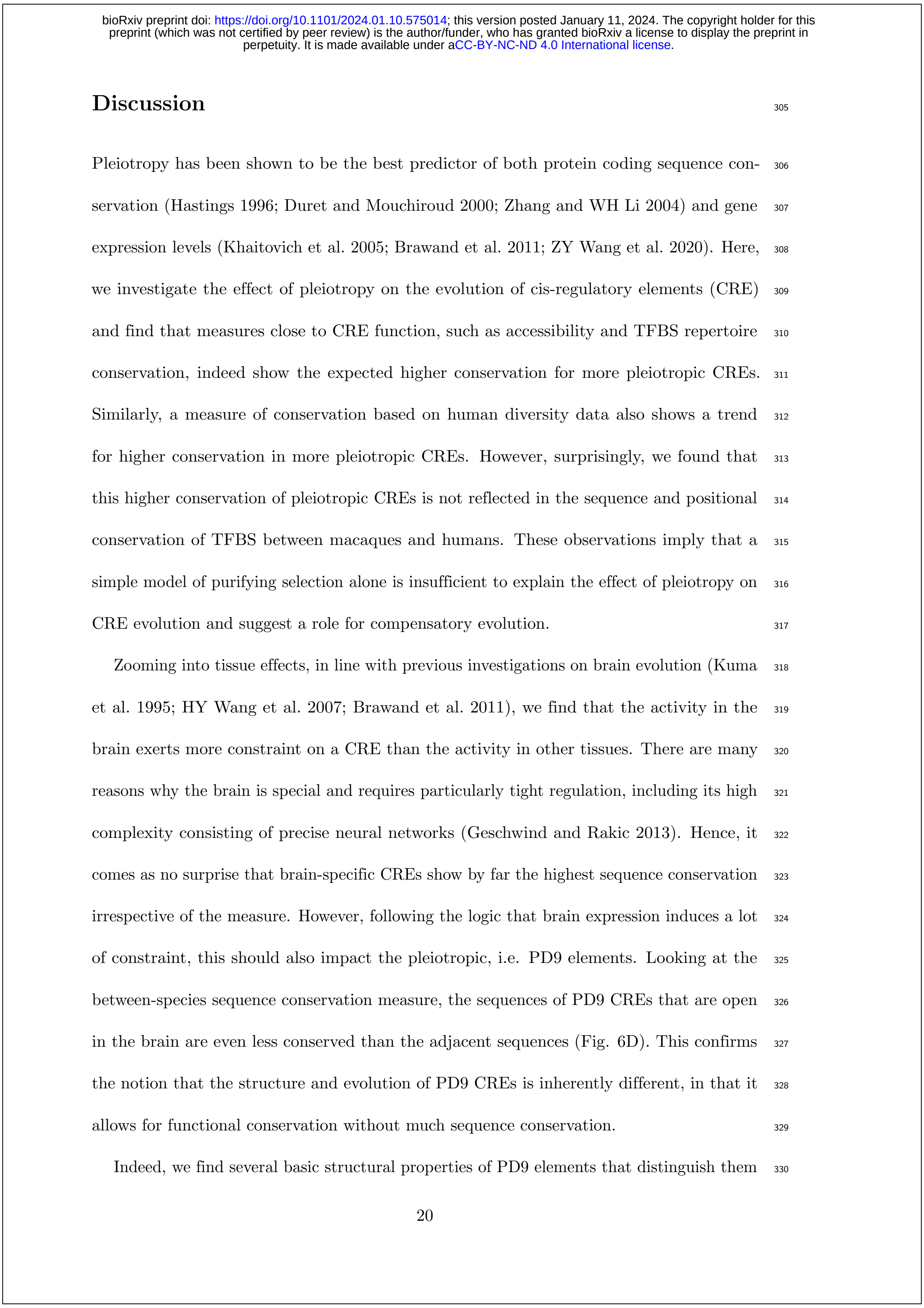


Figure 7. Ranks of ATXN3 PD9 promoter compared to other CREs in terms of (*A*) sequence conservation (mean PhastCons), (*B*) TFBS binding site conservation, (*C*) TFBS repertoire conservation, (*D*) CRE openness conservation between human and cynomolgus macaque in NPCs and (*E*) ATXN3 gene expression conservation between human and cynomolgus macaque in NPCs. (*F, G*) ATXN3 PD9 promoter is accessible in both species. (*H*) ATXN3 promoter shows diverged TFBS positions between species among validated TFs involved in neurogenesis. (*I, J*) PWM logos of the investigated TF motifs with ChIP-seq data available: MYCN (*I*), POU3F2 (*J*).

positions for both TFs. Humans have 6 and macaques 5 MYCN binding sites and both have one POU3F2 binding site, suggesting a rather high repertoire conservation, which is also reflected in similar ATAC-seq peak-shapes (Fig. 7F,G). However, only 3 of the 10 binding sites are positionally conserved between the species. This serves as an example of how the large disagreement between sequence, TF binding site conservation and TFBS repertoire might co-occur.

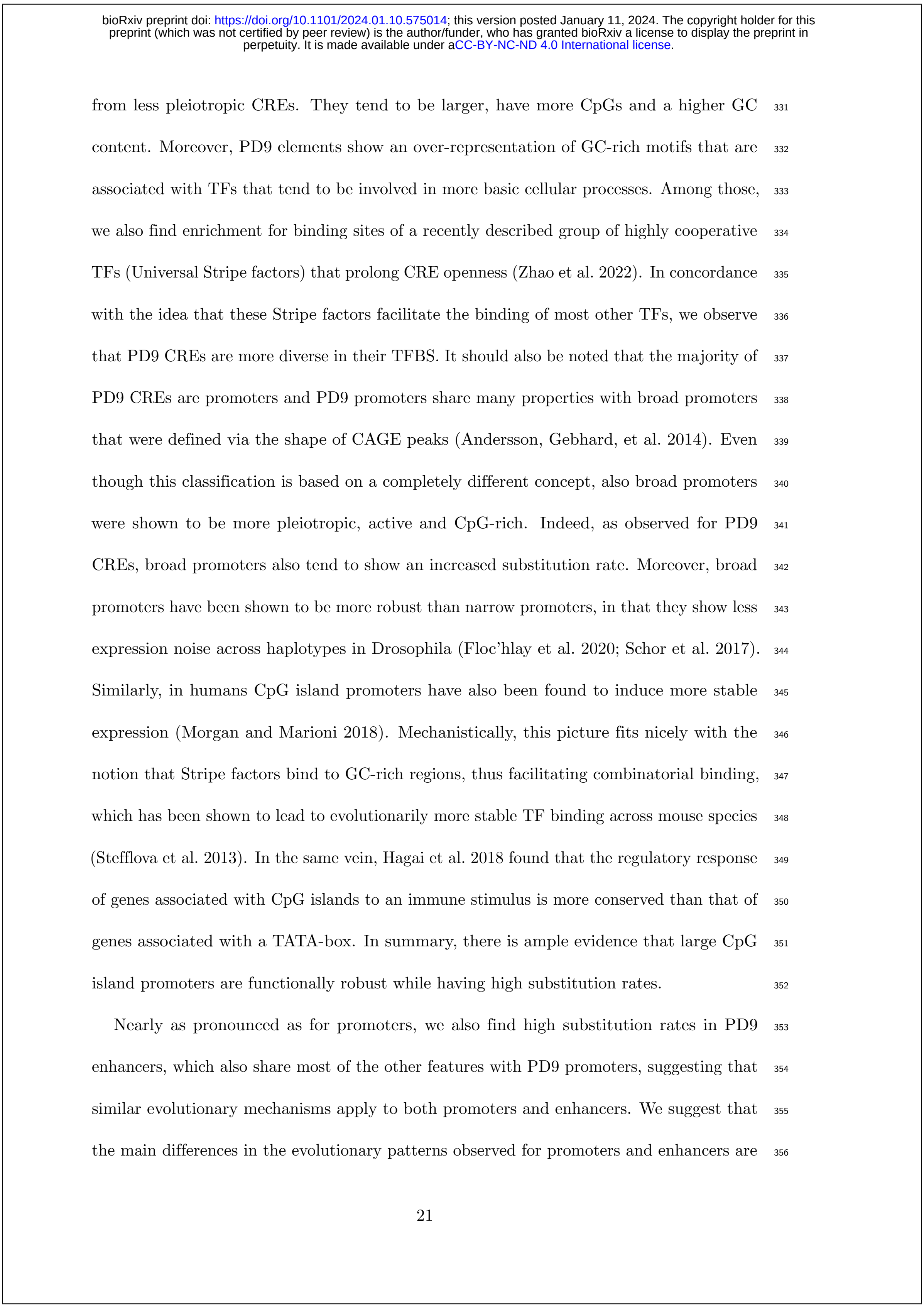


Discussion

Pleiotropy has been shown to be the best predictor of both protein coding sequence conservation (Hastings 1996; Duret and Mouchiroud 2000; Zhang and WH Li 2004) and gene expression levels (Khaitovich et al. 2005; Brawand et al. 2011; ZY Wang et al. 2020). Here, we investigate the effect of pleiotropy on the evolution of cis-regulatory elements (CRE) and find that measures close to CRE function, such as accessibility and TFBS repertoire conservation, indeed show the expected higher conservation for more pleiotropic CREs. Similarly, a measure of conservation based on human diversity data also shows a trend for higher conservation in more pleiotropic CREs. However, surprisingly, we found that this higher conservation of pleiotropic CREs is not reflected in the sequence and positional conservation of TFBS between macaques and humans. These observations imply that a simple model of purifying selection alone is insufficient to explain the effect of pleiotropy on CRE evolution and suggest a role for compensatory evolution.

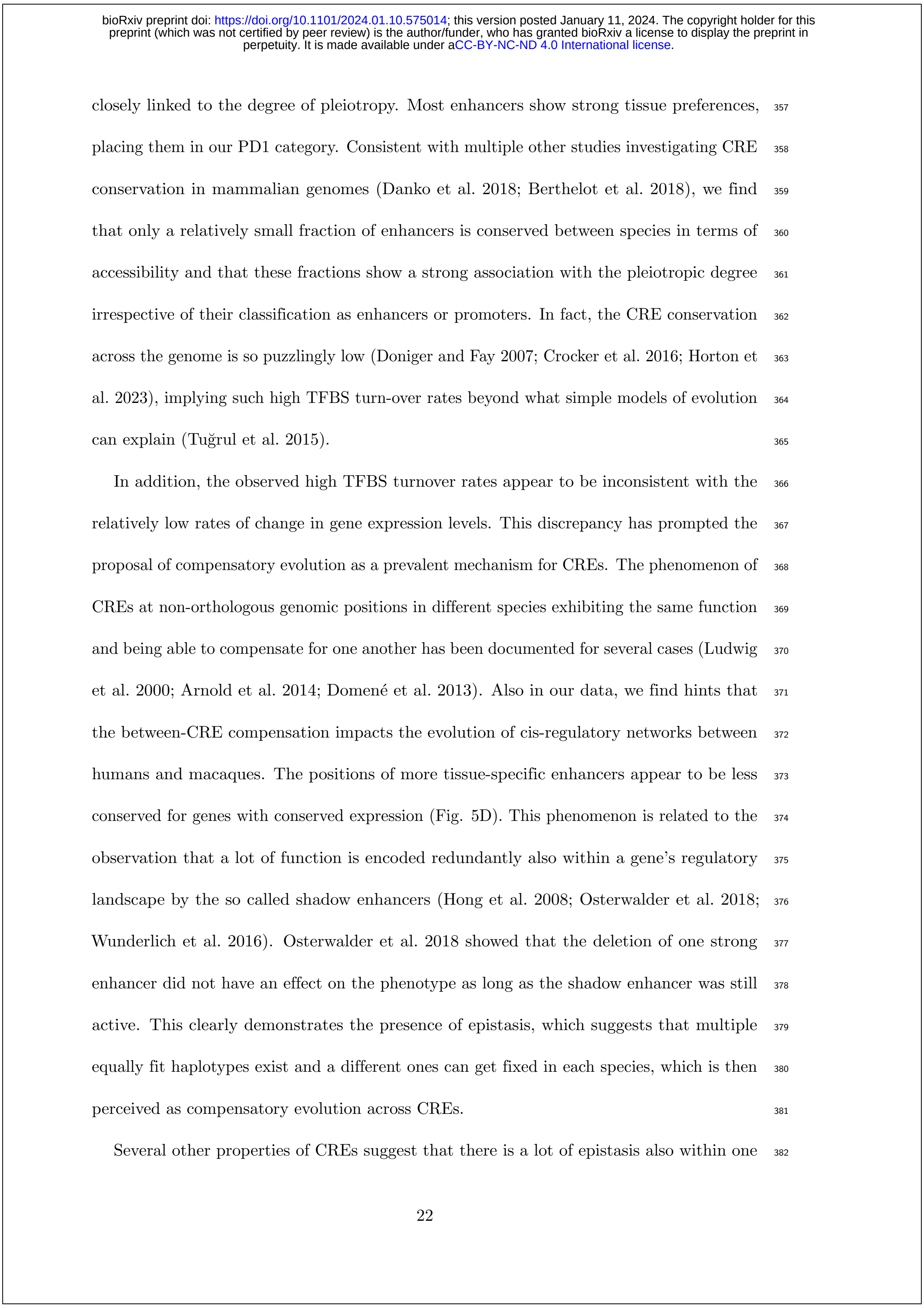
Zooming into tissue effects, in line with previous investigations on brain evolution (Kuma et al. 1995; HY Wang et al. 2007; Brawand et al. 2011), we find that the activity in the brain exerts more constraint on a CRE than the activity in other tissues. There are many reasons why the brain is special and requires particularly tight regulation, including its high complexity consisting of precise neural networks (Geschwind and Rakic 2013). Hence, it comes as no surprise that brain-specific CREs show by far the highest sequence conservation irrespective of the measure. However, following the logic that brain expression induces a lot of constraint, this should also impact the pleiotropic, i.e. PD9 elements. Looking at the between-species sequence conservation measure, the sequences of PD9 CREs that are open in the brain are even less conserved than the adjacent sequences (Fig. 6D). This confirms the notion that the structure and evolution of PD9 CREs is inherently different, in that it allows for functional conservation without much sequence conservation.

Indeed, we find several basic structural properties of PD9 elements that distinguish them



from less pleiotropic CREs. They tend to be larger, have more CpGs and a higher GC content. Moreover, PD9 elements show an over-representation of GC-rich motifs that are associated with TFs that tend to be involved in more basic cellular processes. Among those, we also find enrichment for binding sites of a recently described group of highly cooperative TFs (Universal Stripe factors) that prolong CRE openness (Zhao et al. 2022). In concordance with the idea that these Stripe factors facilitate the binding of most other TFs, we observe that PD9 CREs are more diverse in their TFBS. It should also be noted that the majority of PD9 CREs are promoters and PD9 promoters share many properties with broad promoters that were defined via the shape of CAGE peaks (Andersson, Gebhard, et al. 2014). Even though this classification is based on a completely different concept, also broad promoters were shown to be more pleiotropic, active and CpG-rich. Indeed, as observed for PD9 CREs, broad promoters also tend to show an increased substitution rate. Moreover, broad promoters have been shown to be more robust than narrow promoters, in that they show less expression noise across haplotypes in Drosophila (Floc'hlay et al. 2020; Schor et al. 2017). Similarly, in humans CpG island promoters have also been found to induce more stable expression (Morgan and Marioni 2018). Mechanistically, this picture fits nicely with the notion that Stripe factors bind to GC-rich regions, thus facilitating combinatorial binding, which has been shown to lead to evolutionarily more stable TF binding across mouse species (Stefflova et al. 2013). In the same vein, Hagai et al. 2018 found that the regulatory response of genes associated with CpG islands to an immune stimulus is more conserved than that of genes associated with a TATA-box. In summary, there is ample evidence that large CpG island promoters are functionally robust while having high substitution rates.

Nearly as pronounced as for promoters, we also find high substitution rates in PD9 enhancers, which also share most of the other features with PD9 promoters, suggesting that similar evolutionary mechanisms apply to both promoters and enhancers. We suggest that the main differences in the evolutionary patterns observed for promoters and enhancers are



closely linked to the degree of pleiotropy. Most enhancers show strong tissue preferences, placing them in our PD1 category. Consistent with multiple other studies investigating CRE conservation in mammalian genomes (Danko et al. 2018; Berthelot et al. 2018), we find that only a relatively small fraction of enhancers is conserved between species in terms of accessibility and that these fractions show a strong association with the pleiotropic degree irrespective of their classification as enhancers or promoters. In fact, the CRE conservation across the genome is so puzzlingly low (Doniger and Fay 2007; Crocker et al. 2016; Horton et al. 2023), implying such high TFBS turn-over rates beyond what simple models of evolution can explain (Tuğrul et al. 2015).

In addition, the observed high TFBS turnover rates appear to be inconsistent with the relatively low rates of change in gene expression levels. This discrepancy has prompted the proposal of compensatory evolution as a prevalent mechanism for CREs. The phenomenon of CREs at non-orthologous genomic positions in different species exhibiting the same function and being able to compensate for one another has been documented for several cases (Ludwig et al. 2000; Arnold et al. 2014; Domené et al. 2013). Also in our data, we find hints that the between-CRE compensation impacts the evolution of cis-regulatory networks between humans and macaques. The positions of more tissue-specific enhancers appear to be less conserved for genes with conserved expression (Fig. 5D). This phenomenon is related to the observation that a lot of function is encoded redundantly also within a gene's regulatory landscape by the so called shadow enhancers (Hong et al. 2008; Osterwalder et al. 2018; Wunderlich et al. 2016). Osterwalder et al. 2018 showed that the deletion of one strong enhancer did not have an effect on the phenotype as long as the shadow enhancer was still active. This clearly demonstrates the presence of epistasis, which suggests that multiple equally fit haplotypes exist and a different ones can get fixed in each species, which is then perceived as compensatory evolution across CREs.

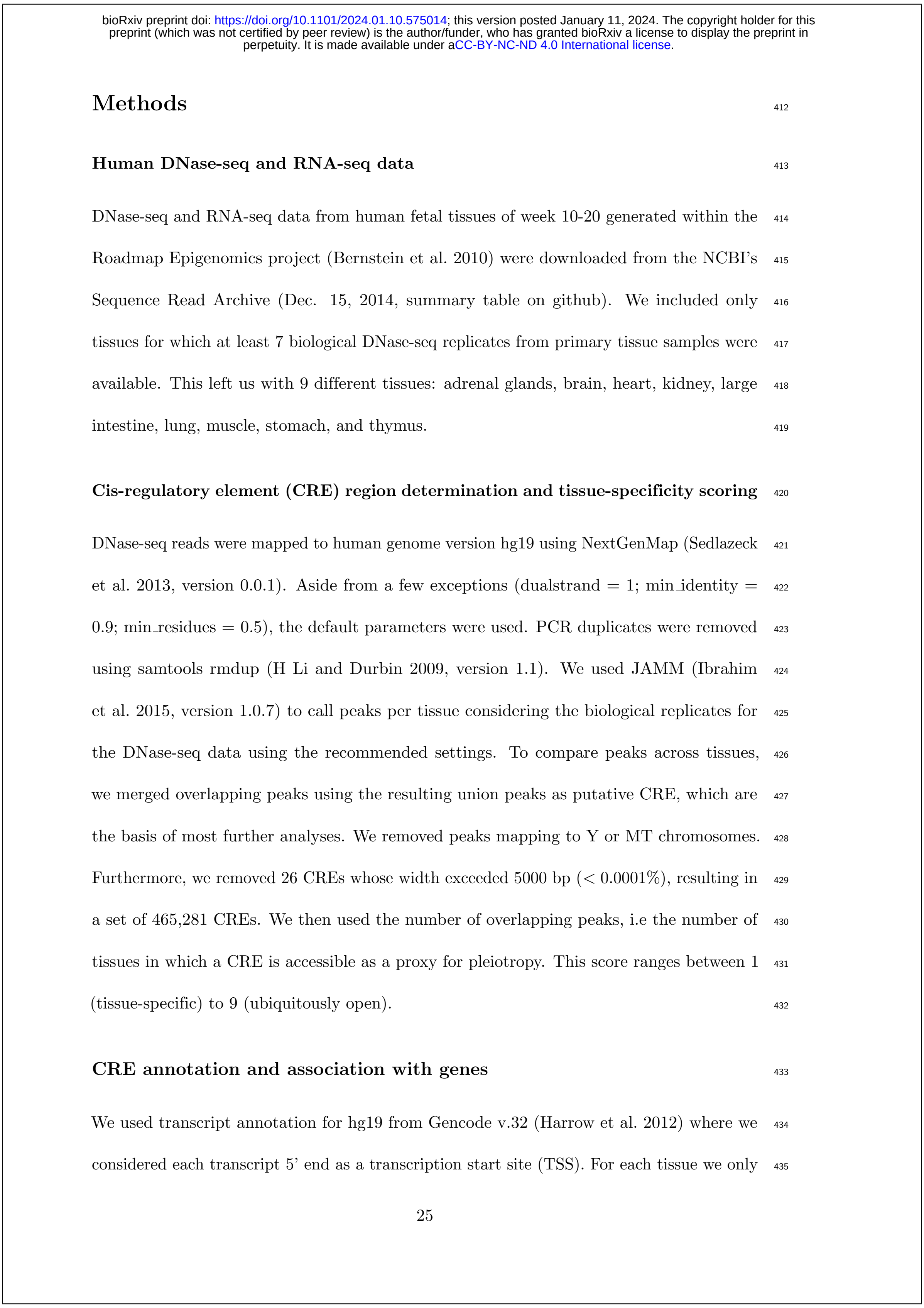
Several other properties of CREs suggest that there is a lot of epistasis also within one

CRE. The billboard model (Kulkarni and Arnosti 2003; Arnosti and Kulkarni 2005) and the TF-collective model (Junion et al. 2012) of enhancer activity suggest that two CRE haplotypes with shifted but similar TFBSs should be functionally equivalent. It follows that the mutations that create these two haplotypes will also have a non-additive effect on fitness. Moreover, some studies showed that binding to a high-affinity site is facilitated by many neighboring low-affinity binding sites, thus providing the raw material for high TFBS turnover rates (Tuğrul et al. 2015).

Thus, we suggest that within-element compensation of TFBS is a common mode of evolution for pleiotropic CREs. This would explain the apparent disparity between the cross-species sequence conservation and the within-species constraint measure E.W. (Fig. 4). Moreover, it would also explain the disparity between the low sequence and the high functional conservation between species as observed in our ATAC-seq and RNA-seq data: If different, functionally equivalent haplotypes got fixed in different species, this would lead to a high sequence divergence while the open chromatin state and downstream gene expression remained conserved (Fig. 6H). Furthermore, we show that even though PD9 TFBS may not have a high positional conservation, the overall binding potential for various TFs across a pleiotropic CRE tends to be conserved.

In summary, we think that compensatory evolution is a prevalent mode for evolution of regulatory elements and goes along with the number of contexts in which the element is utilized. The structure of cis-regulatory networks lends itself to high levels of negative epistasis across more distal CREs, while for the complex, large pleiotropic CREs epistatic interactions are more likely to occur within the same element. The within-element compensation is possibly facilitated by higher spatial restrictions on TFBS locations: Promoters are likely more restricted spatially than enhancers. However, we observe similar patterns for pleiotropic enhancers as well, albeit less pronounced. We speculate that they are also spatially more restricted than less pleiotropic enhancers due to their higher sequence complexity, which is

probably due to highly cooperative binding at the pleiotropic sites. Such complex element structures are less likely to spontaneously occur at distal sites than it is observed for tissue-specific elements.



Methods

Human DNase-seq and RNA-seq data

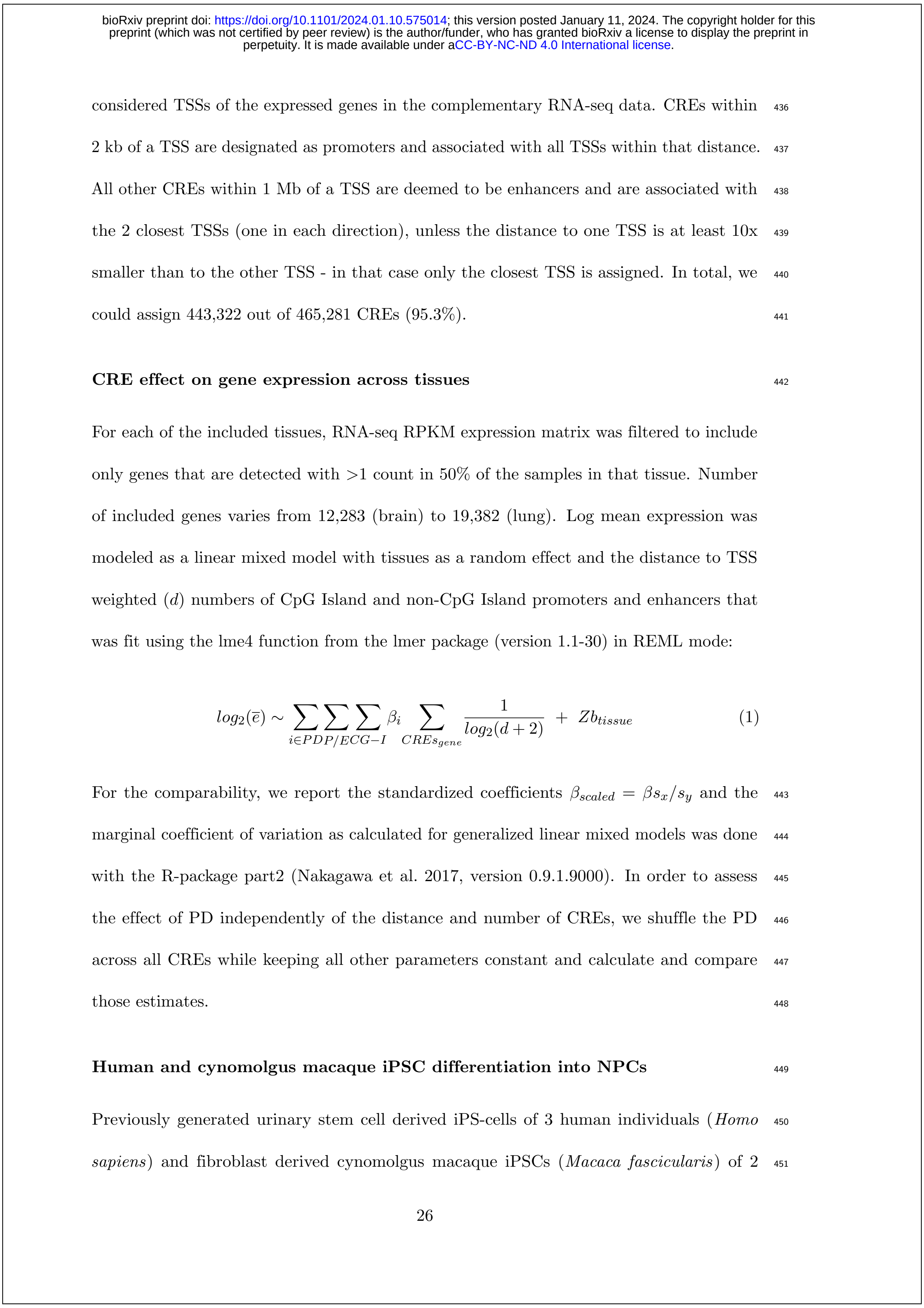
DNase-seq and RNA-seq data from human fetal tissues of week 10-20 generated within the Roadmap Epigenomics project (Bernstein et al. 2010) were downloaded from the NCBI's Sequence Read Archive (Dec. 15, 2014, summary table on github). We included only tissues for which at least 7 biological DNase-seq replicates from primary tissue samples were available. This left us with 9 different tissues: adrenal glands, brain, heart, kidney, large intestine, lung, muscle, stomach, and thymus.

Cis-regulatory element (CRE) region determination and tissue-specificity scoring

DNase-seq reads were mapped to human genome version hg19 using NextGenMap (Sedlazeck et al. 2013, version 0.0.1). Aside from a few exceptions (dualstrand = 1; min_identity = 0.9; min_residues = 0.5), the default parameters were used. PCR duplicates were removed using samtools rmdup (H Li and Durbin 2009, version 1.1). We used JAMM (Ibrahim et al. 2015, version 1.0.7) to call peaks per tissue considering the biological replicates for the DNase-seq data using the recommended settings. To compare peaks across tissues, we merged overlapping peaks using the resulting union peaks as putative CRE, which are the basis of most further analyses. We removed peaks mapping to Y or MT chromosomes. Furthermore, we removed 26 CREs whose width exceeded 5000 bp ($< 0.0001\%$), resulting in a set of 465,281 CREs. We then used the number of overlapping peaks, i.e the number of tissues in which a CRE is accessible as a proxy for pleiotropy. This score ranges between 1 (tissue-specific) to 9 (ubiquitously open).

CRE annotation and association with genes

We used transcript annotation for hg19 from Gencode v.32 (Harrow et al. 2012) where we considered each transcript 5' end as a transcription start site (TSS). For each tissue we only



considered TSSs of the expressed genes in the complementary RNA-seq data. CREs within 2 kb of a TSS are designated as promoters and associated with all TSSs within that distance. All other CREs within 1 Mb of a TSS are deemed to be enhancers and are associated with the 2 closest TSSs (one in each direction), unless the distance to one TSS is at least 10x smaller than to the other TSS - in that case only the closest TSS is assigned. In total, we could assign 443,322 out of 465,281 CREs (95.3%).

CRE effect on gene expression across tissues

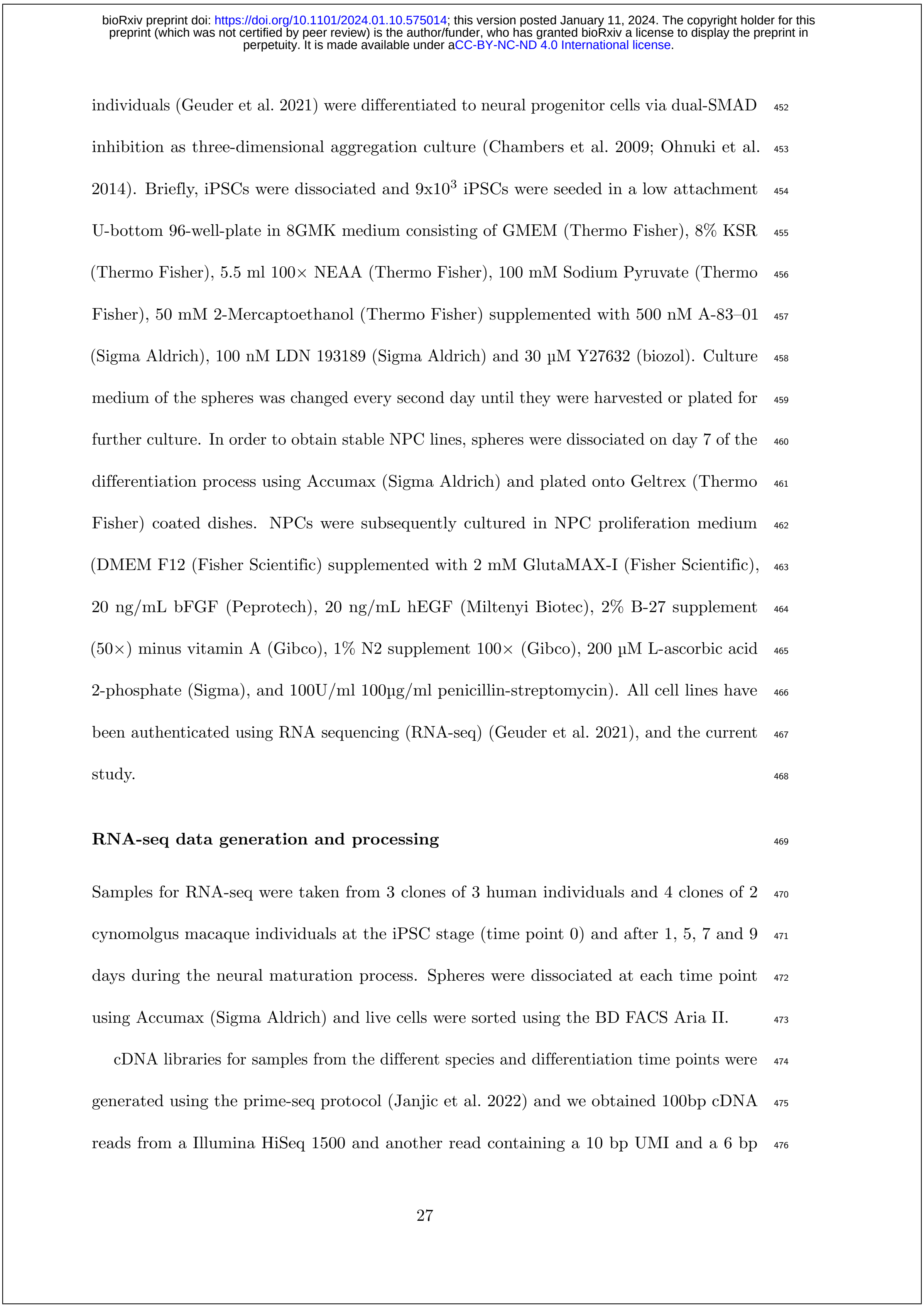
For each of the included tissues, RNA-seq RPKM expression matrix was filtered to include only genes that are detected with >1 count in 50% of the samples in that tissue. Number of included genes varies from 12,283 (brain) to 19,382 (lung). Log mean expression was modeled as a linear mixed model with tissues as a random effect and the distance to TSS weighted (d) numbers of CpG Island and non-CpG Island promoters and enhancers that was fit using the lme4 function from the lmer package (version 1.1-30) in REML mode:

$$log_2(\bar{e}) \sim \sum\sum\sum_{i \in PDP/ECG-I} \beta_i \sum_{CREs_{gene}} \frac{1}{log_2(d+2)} \; + \; Zb_{tissue} \qquad (1)$$

For the comparability, we report the standardized coefficients $\beta_{scaled} = \beta s_x / s_y$ and the marginal coefficient of variation as calculated for generalized linear mixed models was done with the R-package part2 (Nakagawa et al. 2017, version 0.9.1.9000). In order to assess the effect of PD independently of the distance and number of CREs, we shuffle the PD across all CREs while keeping all other parameters constant and calculate and compare those estimates.

Human and cynomolgus macaque iPSC differentiation into NPCs

Previously generated urinary stem cell derived iPS-cells of 3 human individuals (*Homo sapiens*) and fibroblast derived cynomolgus macaque iPSCs (*Macaca fascicularis*) of 2

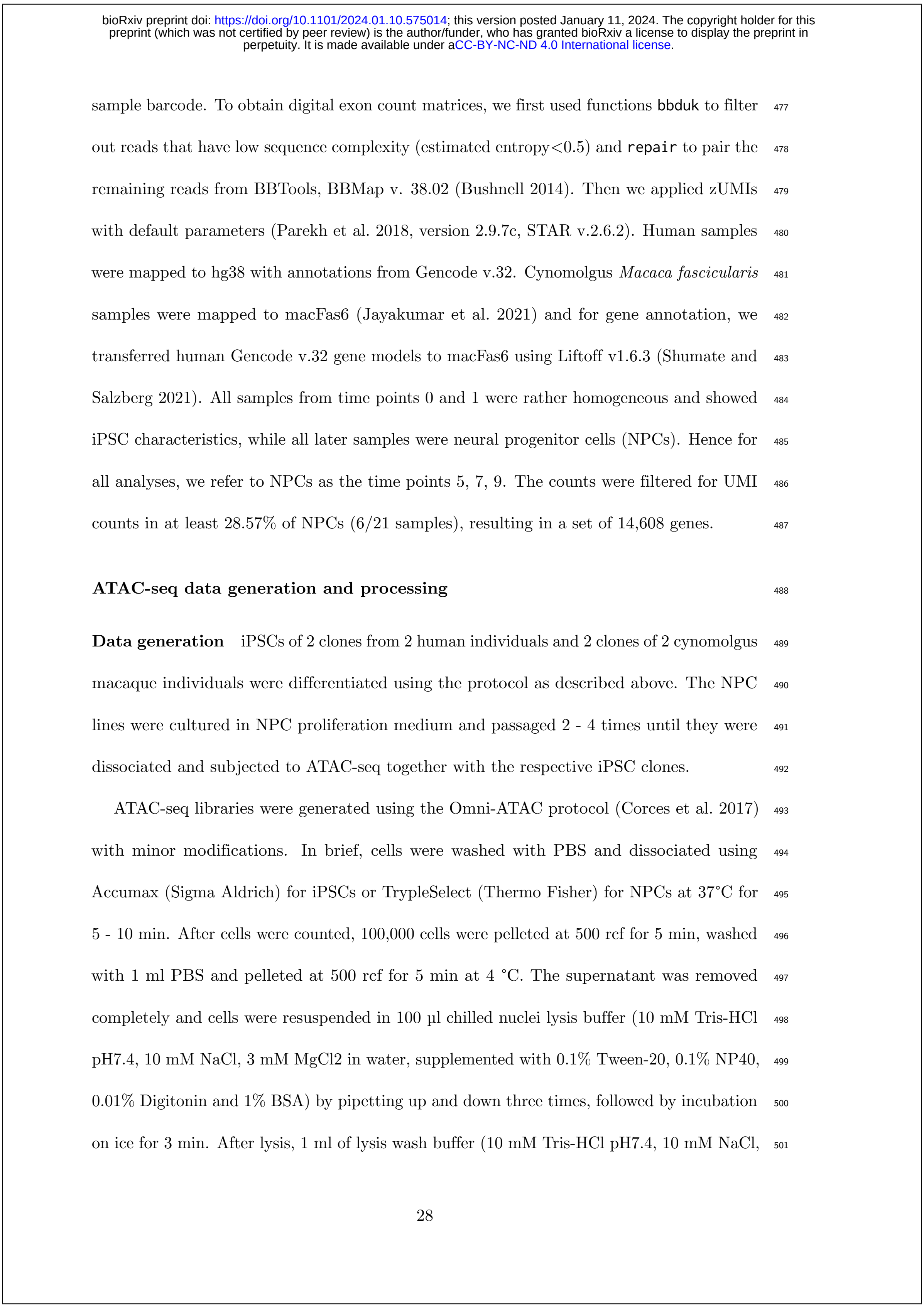


individuals (Geuder et al. 2021) were differentiated to neural progenitor cells via dual-SMAD inhibition as three-dimensional aggregation culture (Chambers et al. 2009; Ohnuki et al. 2014). Briefly, iPSCs were dissociated and 9x10^3 iPSCs were seeded in a low attachment U-bottom 96-well-plate in 8GMK medium consisting of GMEM (Thermo Fisher), 8% KSR (Thermo Fisher), 5.5 ml 100× NEAA (Thermo Fisher), 100 mM Sodium Pyruvate (Thermo Fisher), 50 mM 2-Mercaptoethanol (Thermo Fisher) supplemented with 500 nM A-83–01 (Sigma Aldrich), 100 nM LDN 193189 (Sigma Aldrich) and 30 µM Y27632 (biozol). Culture medium of the spheres was changed every second day until they were harvested or plated for further culture. In order to obtain stable NPC lines, spheres were dissociated on day 7 of the differentiation process using Accumax (Sigma Aldrich) and plated onto Geltrex (Thermo Fisher) coated dishes. NPCs were subsequently cultured in NPC proliferation medium (DMEM F12 (Fisher Scientific) supplemented with 2 mM GlutaMAX-I (Fisher Scientific), 20 ng/mL bFGF (Peprotech), 20 ng/mL hEGF (Miltenyi Biotec), 2% B-27 supplement (50×) minus vitamin A (Gibco), 1% N2 supplement 100× (Gibco), 200 µM L-ascorbic acid 2-phosphate (Sigma), and 100U/ml 100µg/ml penicillin-streptomycin). All cell lines have been authenticated using RNA sequencing (RNA-seq) (Geuder et al. 2021), and the current study.

RNA-seq data generation and processing

Samples for RNA-seq were taken from 3 clones of 3 human individuals and 4 clones of 2 cynomolgus macaque individuals at the iPSC stage (time point 0) and after 1, 5, 7 and 9 days during the neural maturation process. Spheres were dissociated at each time point using Accumax (Sigma Aldrich) and live cells were sorted using the BD FACS Aria II.

cDNA libraries for samples from the different species and differentiation time points were generated using the prime-seq protocol (Janjic et al. 2022) and we obtained 100bp cDNA reads from a Illumina HiSeq 1500 and another read containing a 10 bp UMI and a 6 bp

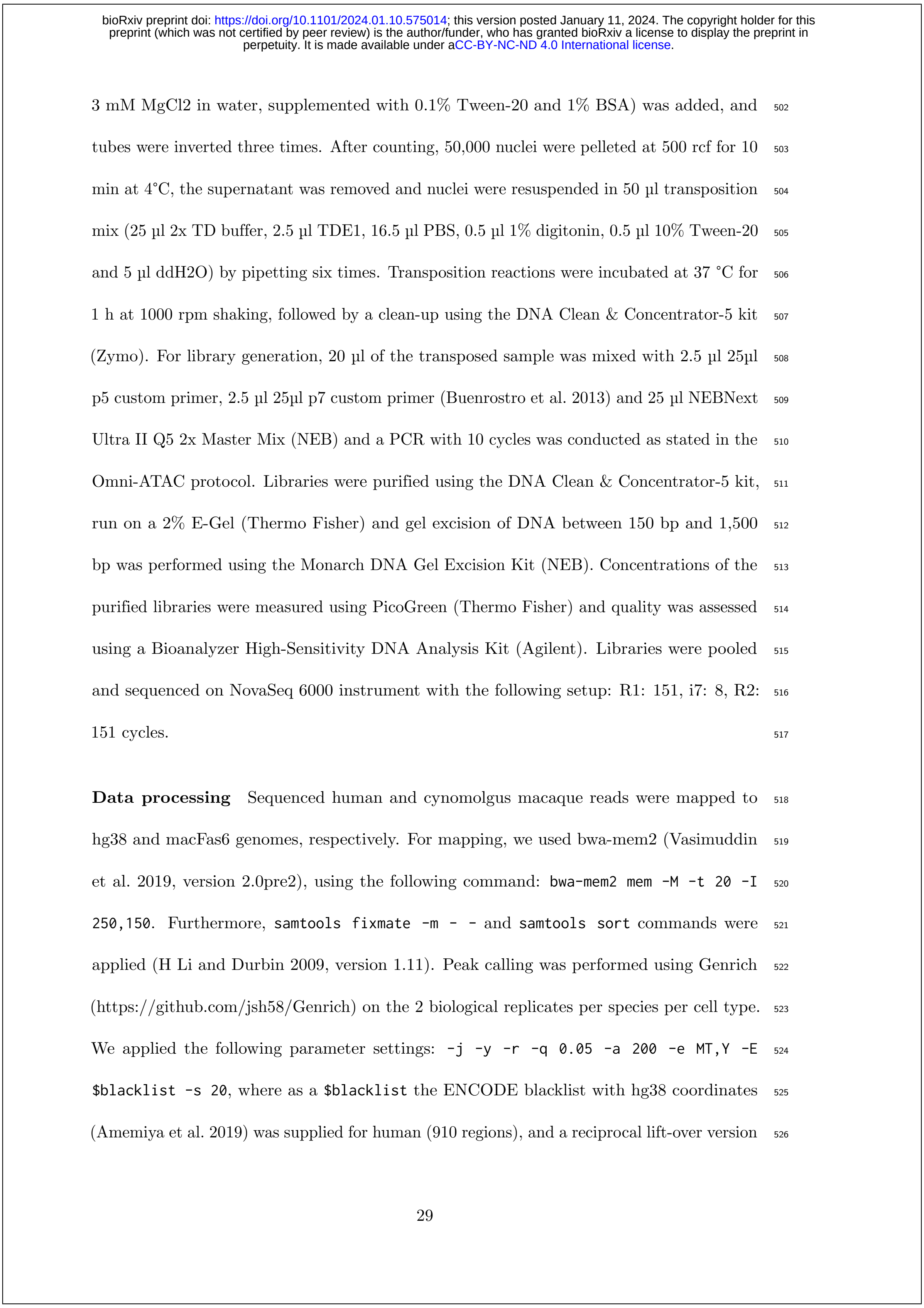


sample barcode. To obtain digital exon count matrices, we first used functions `bbduk` to filter out reads that have low sequence complexity (estimated entropy<0.5) and `repair` to pair the remaining reads from BBTools, BBMap v. 38.02 (Bushnell 2014). Then we applied zUMIs with default parameters (Parekh et al. 2018, version 2.9.7c, STAR v.2.6.2). Human samples were mapped to hg38 with annotations from Gencode v.32. Cynomolgus *Macaca fascicularis* samples were mapped to macFas6 (Jayakumar et al. 2021) and for gene annotation, we transferred human Gencode v.32 gene models to macFas6 using Liftoff v1.6.3 (Shumate and Salzberg 2021). All samples from time points 0 and 1 were rather homogeneous and showed iPSC characteristics, while all later samples were neural progenitor cells (NPCs). Hence for all analyses, we refer to NPCs as the time points 5, 7, 9. The counts were filtered for UMI counts in at least 28.57% of NPCs (6/21 samples), resulting in a set of 14,608 genes.

ATAC-seq data generation and processing

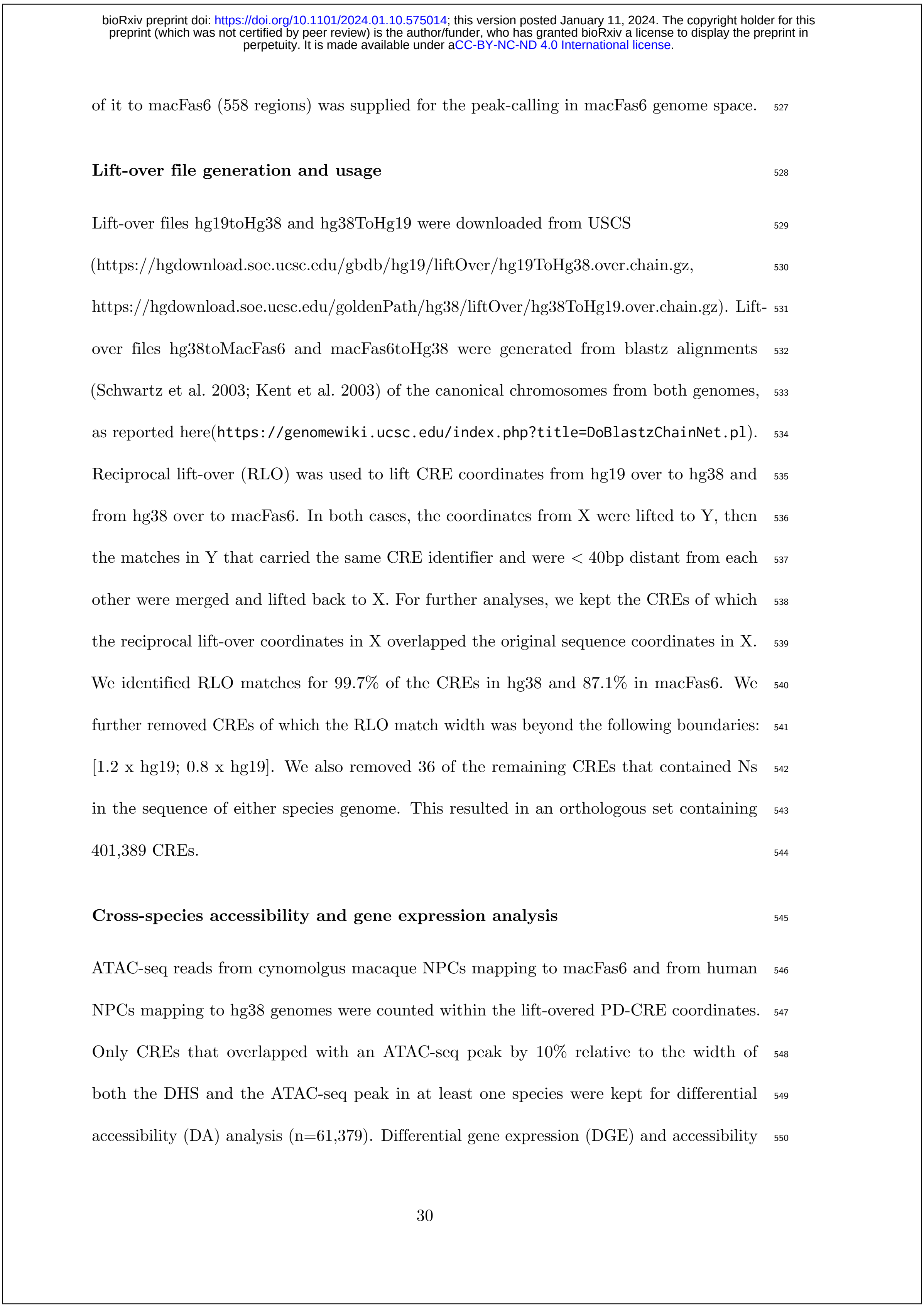
Data generation iPSCs of 2 clones from 2 human individuals and 2 clones of 2 cynomolgus macaque individuals were differentiated using the protocol as described above. The NPC lines were cultured in NPC proliferation medium and passaged 2 - 4 times until they were dissociated and subjected to ATAC-seq together with the respective iPSC clones.

ATAC-seq libraries were generated using the Omni-ATAC protocol (Corces et al. 2017) with minor modifications. In brief, cells were washed with PBS and dissociated using Accumax (Sigma Aldrich) for iPSCs or TrypleSelect (Thermo Fisher) for NPCs at 37°C for 5 - 10 min. After cells were counted, 100,000 cells were pelleted at 500 rcf for 5 min, washed with 1 ml PBS and pelleted at 500 rcf for 5 min at 4 °C. The supernatant was removed completely and cells were resuspended in 100 µl chilled nuclei lysis buffer (10 mM Tris-HCl pH7.4, 10 mM NaCl, 3 mM MgCl2 in water, supplemented with 0.1% Tween-20, 0.1% NP40, 0.01% Digitonin and 1% BSA) by pipetting up and down three times, followed by incubation on ice for 3 min. After lysis, 1 ml of lysis wash buffer (10 mM Tris-HCl pH7.4, 10 mM NaCl,



3 mM MgCl2 in water, supplemented with 0.1% Tween-20 and 1% BSA) was added, and tubes were inverted three times. After counting, 50,000 nuclei were pelleted at 500 rcf for 10 min at 4°C, the supernatant was removed and nuclei were resuspended in 50 µl transposition mix (25 µl 2x TD buffer, 2.5 µl TDE1, 16.5 µl PBS, 0.5 µl 1% digitonin, 0.5 µl 10% Tween-20 and 5 µl ddH2O) by pipetting six times. Transposition reactions were incubated at 37 °C for 1 h at 1000 rpm shaking, followed by a clean-up using the DNA Clean & Concentrator-5 kit (Zymo). For library generation, 20 µl of the transposed sample was mixed with 2.5 µl 25µl p5 custom primer, 2.5 µl 25µl p7 custom primer (Buenrostro et al. 2013) and 25 µl NEBNext Ultra II Q5 2x Master Mix (NEB) and a PCR with 10 cycles was conducted as stated in the Omni-ATAC protocol. Libraries were purified using the DNA Clean & Concentrator-5 kit, run on a 2% E-Gel (Thermo Fisher) and gel excision of DNA between 150 bp and 1,500 bp was performed using the Monarch DNA Gel Excision Kit (NEB). Concentrations of the purified libraries were measured using PicoGreen (Thermo Fisher) and quality was assessed using a Bioanalyzer High-Sensitivity DNA Analysis Kit (Agilent). Libraries were pooled and sequenced on NovaSeq 6000 instrument with the following setup: R1: 151, i7: 8, R2: 151 cycles.

Data processing Sequenced human and cynomolgus macaque reads were mapped to hg38 and macFas6 genomes, respectively. For mapping, we used bwa-mem2 (Vasimuddin et al. 2019, version 2.0pre2), using the following command: `bwa-mem2 mem -M -t 20 -I 250,150`. Furthermore, `samtools fixmate -m - -` and `samtools sort` commands were applied (H Li and Durbin 2009, version 1.11). Peak calling was performed using Genrich (https://github.com/jsh58/Genrich) on the 2 biological replicates per species per cell type. We applied the following parameter settings: `-j -y -r -q 0.05 -a 200 -e MT,Y -E $blacklist -s 20`, where as a `$blacklist` the ENCODE blacklist with hg38 coordinates (Amemiya et al. 2019) was supplied for human (910 regions), and a reciprocal lift-over version



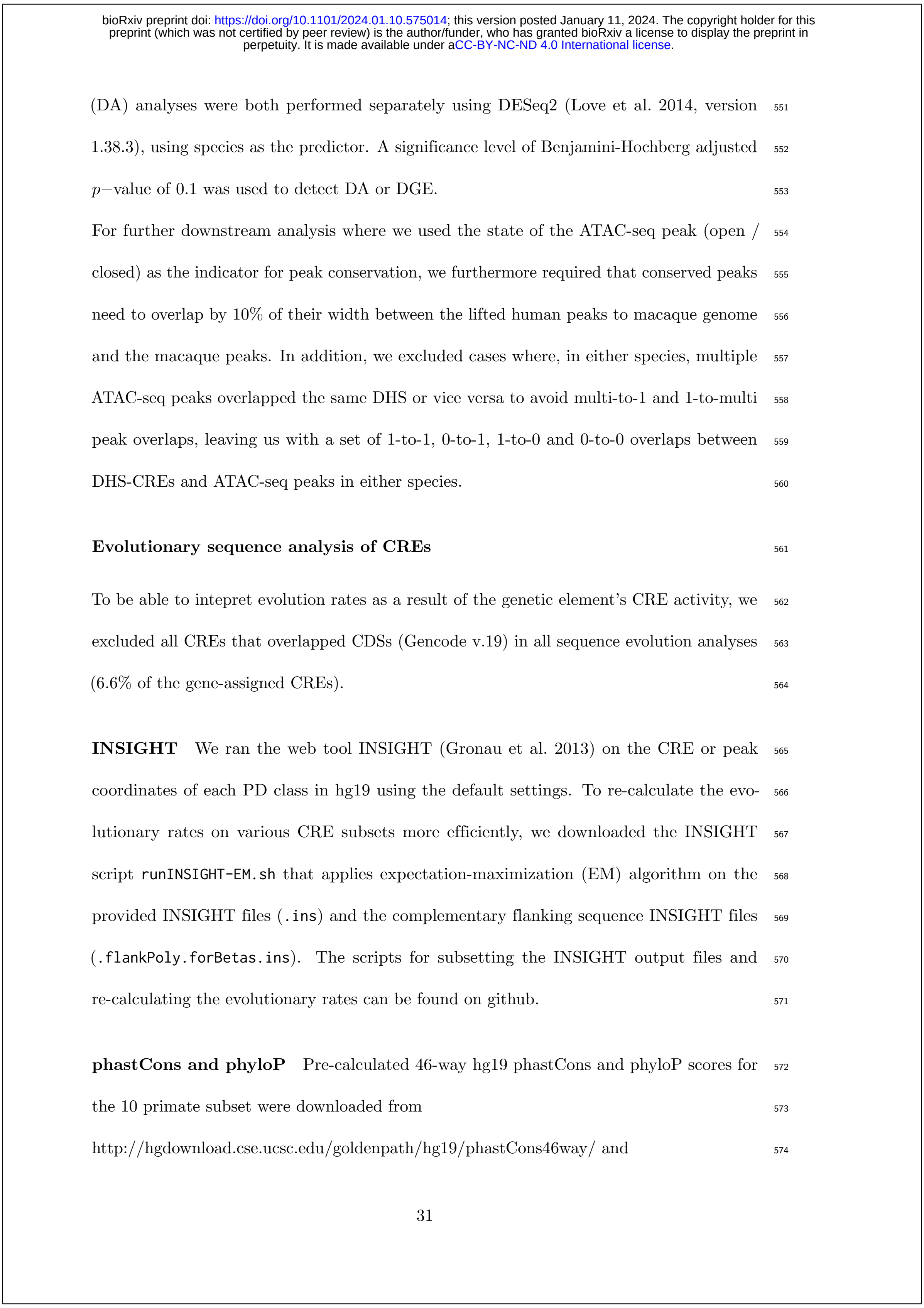
of it to macFas6 (558 regions) was supplied for the peak-calling in macFas6 genome space.

Lift-over file generation and usage

Lift-over files hg19toHg38 and hg38ToHg19 were downloaded from USCS (https://hgdownload.soe.ucsc.edu/gbdb/hg19/liftOver/hg19ToHg38.over.chain.gz, https://hgdownload.soe.ucsc.edu/goldenPath/hg38/liftOver/hg38ToHg19.over.chain.gz). Lift-over files hg38toMacFas6 and macFas6toHg38 were generated from blastz alignments (Schwartz et al. 2003; Kent et al. 2003) of the canonical chromosomes from both genomes, as reported here(`https://genomewiki.ucsc.edu/index.php?title=DoBlastzChainNet.pl`). Reciprocal lift-over (RLO) was used to lift CRE coordinates from hg19 over to hg38 and from hg38 over to macFas6. In both cases, the coordinates from X were lifted to Y, then the matches in Y that carried the same CRE identifier and were $<$ 40bp distant from each other were merged and lifted back to X. For further analyses, we kept the CREs of which the reciprocal lift-over coordinates in X overlapped the original sequence coordinates in X. We identified RLO matches for 99.7% of the CREs in hg38 and 87.1% in macFas6. We further removed CREs of which the RLO match width was beyond the following boundaries: [1.2 x hg19; 0.8 x hg19]. We also removed 36 of the remaining CREs that contained Ns in the sequence of either species genome. This resulted in an orthologous set containing 401,389 CREs.

Cross-species accessibility and gene expression analysis

ATAC-seq reads from cynomolgus macaque NPCs mapping to macFas6 and from human NPCs mapping to hg38 genomes were counted within the lift-overed PD-CRE coordinates. Only CREs that overlapped with an ATAC-seq peak by 10% relative to the width of both the DHS and the ATAC-seq peak in at least one species were kept for differential accessibility (DA) analysis (n=61,379). Differential gene expression (DGE) and accessibility



(DA) analyses were both performed separately using DESeq2 (Love et al. 2014, version 1.38.3), using species as the predictor. A significance level of Benjamini-Hochberg adjusted $p-$value of 0.1 was used to detect DA or DGE.

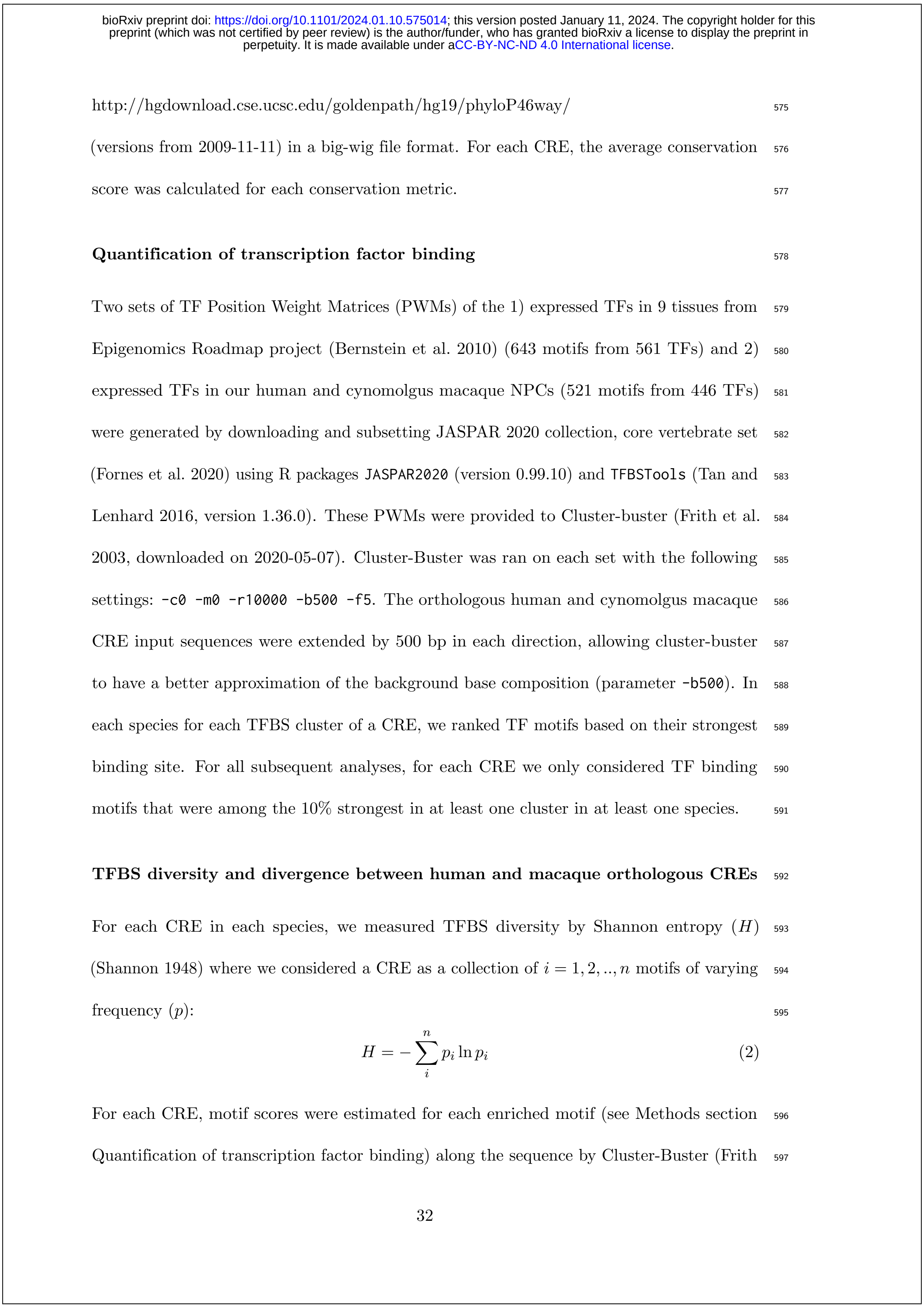
For further downstream analysis where we used the state of the ATAC-seq peak (open / closed) as the indicator for peak conservation, we furthermore required that conserved peaks need to overlap by 10% of their width between the lifted human peaks to macaque genome and the macaque peaks. In addition, we excluded cases where, in either species, multiple ATAC-seq peaks overlapped the same DHS or vice versa to avoid multi-to-1 and 1-to-multi peak overlaps, leaving us with a set of 1-to-1, 0-to-1, 1-to-0 and 0-to-0 overlaps between DHS-CREs and ATAC-seq peaks in either species.

Evolutionary sequence analysis of CREs

To be able to intepret evolution rates as a result of the genetic element's CRE activity, we excluded all CREs that overlapped CDSs (Gencode v.19) in all sequence evolution analyses (6.6% of the gene-assigned CREs).

INSIGHT We ran the web tool INSIGHT (Gronau et al. 2013) on the CRE or peak coordinates of each PD class in hg19 using the default settings. To re-calculate the evolutionary rates on various CRE subsets more efficiently, we downloaded the INSIGHT script `runINSIGHT-EM.sh` that applies expectation-maximization (EM) algorithm on the provided INSIGHT files (`.ins`) and the complementary flanking sequence INSIGHT files (`.flankPoly.forBetas.ins`). The scripts for subsetting the INSIGHT output files and re-calculating the evolutionary rates can be found on github.

phastCons and phyloP Pre-calculated 46-way hg19 phastCons and phyloP scores for the 10 primate subset were downloaded from
http://hgdownload.cse.ucsc.edu/goldenpath/hg19/phastCons46way/ and



http://hgdownload.cse.ucsc.edu/goldenpath/hg19/phyloP46way/ (versions from 2009-11-11) in a big-wig file format. For each CRE, the average conservation score was calculated for each conservation metric.

Quantification of transcription factor binding

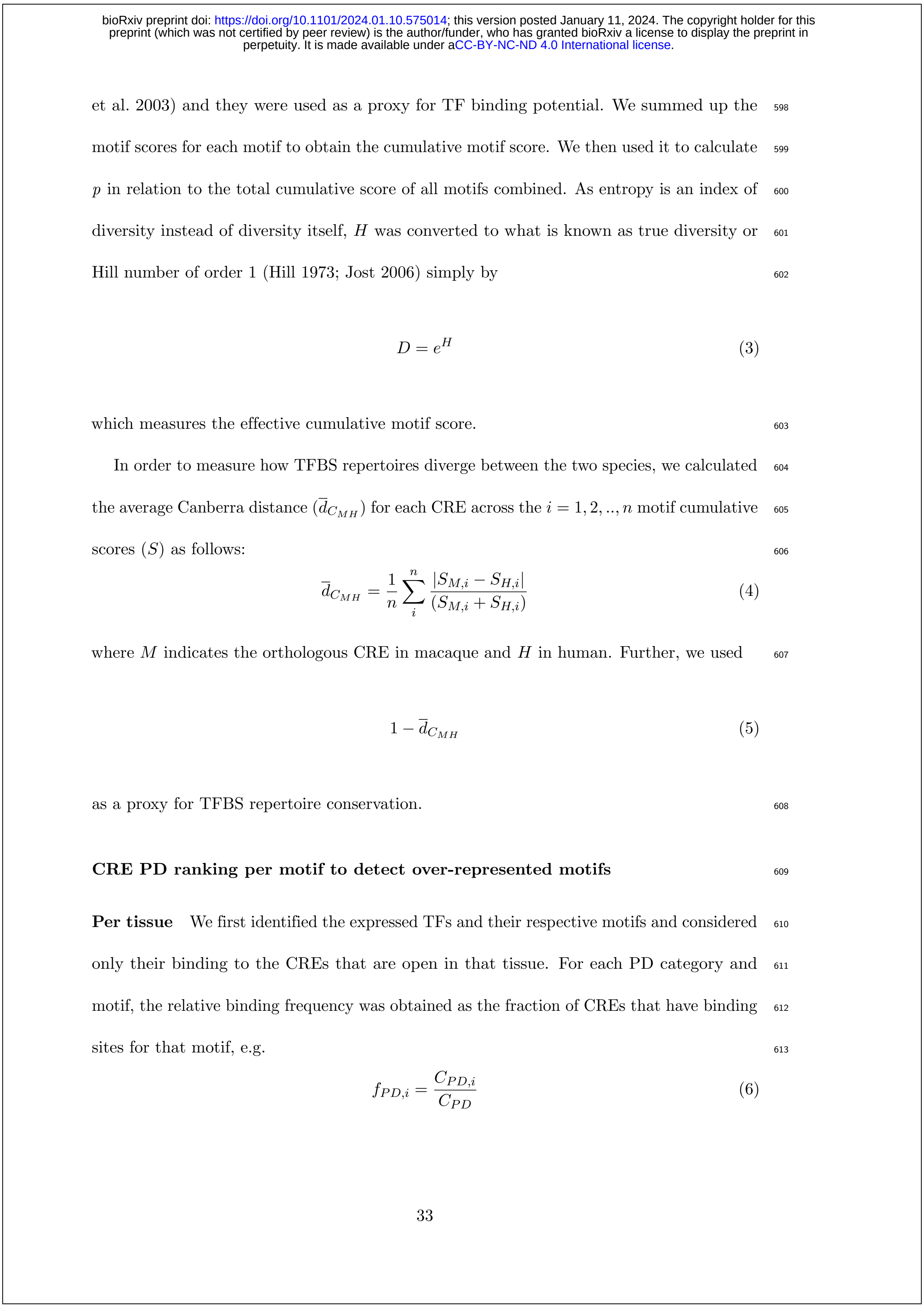
Two sets of TF Position Weight Matrices (PWMs) of the 1) expressed TFs in 9 tissues from Epigenomics Roadmap project (Bernstein et al. 2010) (643 motifs from 561 TFs) and 2) expressed TFs in our human and cynomolgus macaque NPCs (521 motifs from 446 TFs) were generated by downloading and subsetting JASPAR 2020 collection, core vertebrate set (Fornes et al. 2020) using R packages `JASPAR2020` (version 0.99.10) and `TFBSTools` (Tan and Lenhard 2016, version 1.36.0). These PWMs were provided to Cluster-buster (Frith et al. 2003, downloaded on 2020-05-07). Cluster-Buster was ran on each set with the following settings: `-c0 -m0 -r10000 -b500 -f5`. The orthologous human and cynomolgus macaque CRE input sequences were extended by 500 bp in each direction, allowing cluster-buster to have a better approximation of the background base composition (parameter `-b500`). In each species for each TFBS cluster of a CRE, we ranked TF motifs based on their strongest binding site. For all subsequent analyses, for each CRE we only considered TF binding motifs that were among the 10% strongest in at least one cluster in at least one species.

TFBS diversity and divergence between human and macaque orthologous CREs

For each CRE in each species, we measured TFBS diversity by Shannon entropy (H) (Shannon 1948) where we considered a CRE as a collection of $i = 1, 2, .., n$ motifs of varying frequency (p):

$$H = -\sum_{i}^{n} p_i \ln p_i \qquad (2)$$

For each CRE, motif scores were estimated for each enriched motif (see Methods section Quantification of transcription factor binding) along the sequence by Cluster-Buster (Frith



et al. 2003) and they were used as a proxy for TF binding potential. We summed up the motif scores for each motif to obtain the cumulative motif score. We then used it to calculate p in relation to the total cumulative score of all motifs combined. As entropy is an index of diversity instead of diversity itself, H was converted to what is known as true diversity or Hill number of order 1 (Hill 1973; Jost 2006) simply by

$$D = e^{H} \tag{3}$$

which measures the effective cumulative motif score.

In order to measure how TFBS repertoires diverge between the two species, we calculated the average Canberra distance ($\overline{d}_{C_{MH}}$) for each CRE across the $i = 1, 2, .., n$ motif cumulative scores (S) as follows:

$$\overline{d}_{C_{MH}} = \frac{1}{n} \sum_{i}^{n} \frac{|S_{M,i} - S_{H,i}|}{(S_{M,i} + S_{H,i})} \tag{4}$$

where M indicates the orthologous CRE in macaque and H in human. Further, we used

$$1 - \overline{d}_{C_{MH}} \tag{5}$$

as a proxy for TFBS repertoire conservation.

CRE PD ranking per motif to detect over-represented motifs

Per tissue We first identified the expressed TFs and their respective motifs and considered only their binding to the CREs that are open in that tissue. For each PD category and motif, the relative binding frequency was obtained as the fraction of CREs that have binding sites for that motif, e.g.

$$f_{PD,i} = \frac{C_{PD,i}}{C_{PD}} \tag{6}$$

where PD indicates a PD category, i indicates a motif, $C_{PD,i}$ is the count of CREs with motif i binding site(s) present in the particular PD category, C_{PD} is the total CRE count in that PD category. Having obtained these relative frequencies per PD, we then ranked PD categories for each motif. Fold changes of the binding fraction of rank-1 PD relative to the average fraction were calculated for each motif i as:

$$FC_{PD_{(1)},i} = \frac{f_{PD_{(1)},i}}{\frac{1}{9}\sum_{rank=1}^{9} f_{PD_{(rank)},i}} \tag{7}$$

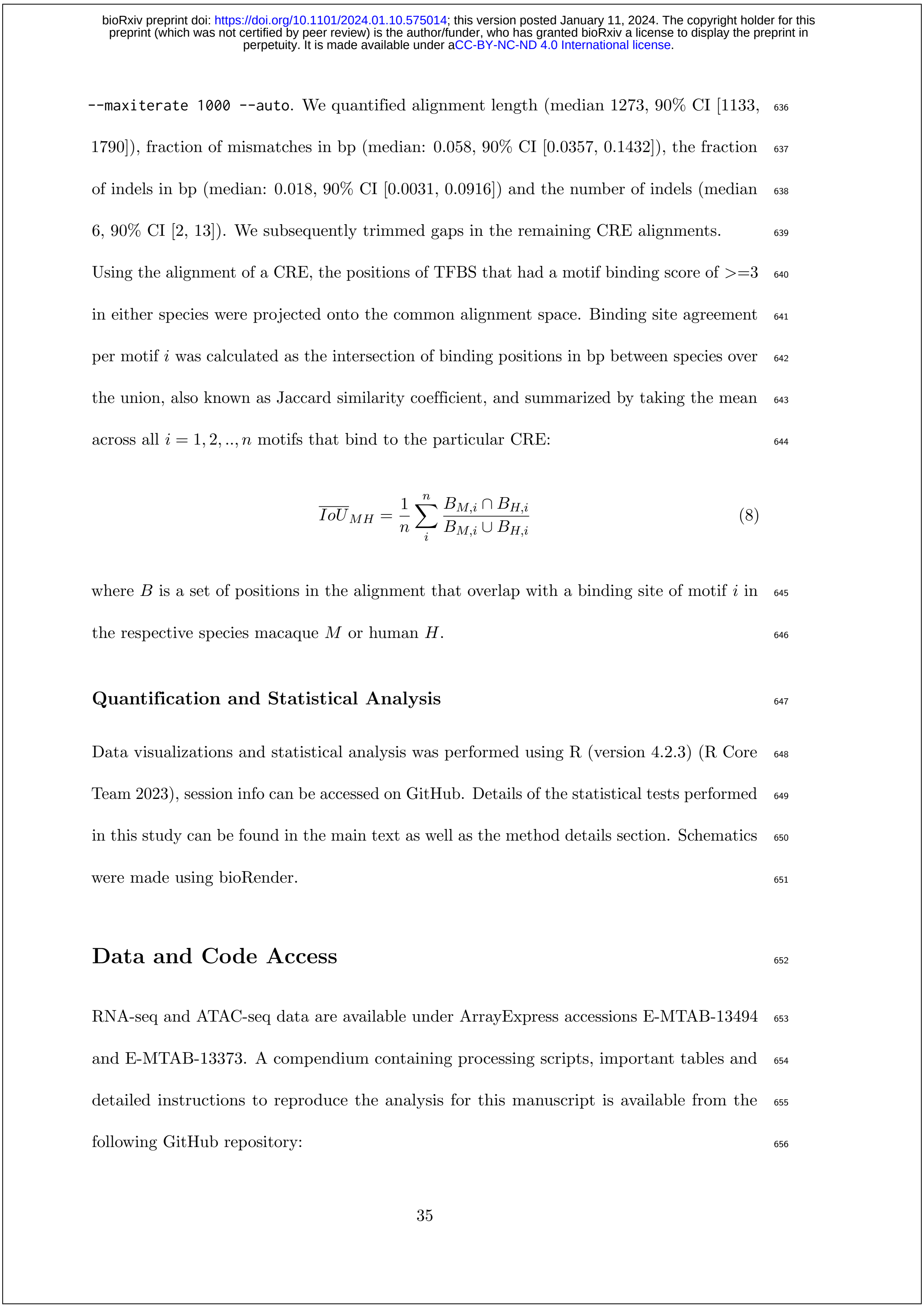
Across tissues To summarise motif-PD enrichment across tissues, we focused on motifs that had the highest binding fractions (rank-1) to either PD9 or PD1. To obtain the PD9-enriched motifs, we identified TF motifs for which PD9 CREs had rank-1 in all tissues. As the PD1-tissue-specific motifs we considered the ones that have PD1 with rank-1 only in that particular tissue, but not in the other tissues. Gene-set enrichment analysis contrasting the respective TF groups with the rest of the expressed TFs was conducted using the Bioconductor package `topGO` (Alexa and Rahnenfuhrer n.d., version 2.50.0), setting the following parameters: `ontology="BP", nodeSize = 10, algorithm = "elim", statistic = "fisher"`.

Stripe factor enrichment analysis

Stripe factor annotation table was obtained from Zhao et al. 2022. We selected the stripe factors detected in human ("Human Stripe Factors") and to subset the universal stripe factors, we used a cutoff of 0.9 for the proportion of total samples in which this TF was detected to be a stripe factor.

TFBS position overlap between human and macaque orthologous CREs

Orthologous human and macaque CRE sequences were pairwise aligned using mafft (Katoh and Standley 2013, version 7) using the following parameters `--adjustdirection`



`--maxiterate 1000 --auto`. We quantified alignment length (median 1273, 90% CI [1133, 1790]), fraction of mismatches in bp (median: 0.058, 90% CI [0.0357, 0.1432]), the fraction of indels in bp (median: 0.018, 90% CI [0.0031, 0.0916]) and the number of indels (median 6, 90% CI [2, 13]). We subsequently trimmed gaps in the remaining CRE alignments. Using the alignment of a CRE, the positions of TFBS that had a motif binding score of >=3 in either species were projected onto the common alignment space. Binding site agreement per motif i was calculated as the intersection of binding positions in bp between species over the union, also known as Jaccard similarity coefficient, and summarized by taking the mean across all $i = 1, 2, .., n$ motifs that bind to the particular CRE:

$$\overline{IoU}_{MH} = \frac{1}{n} \sum_{i}^{n} \frac{B_{M,i} \cap B_{H,i}}{B_{M,i} \cup B_{H,i}} \tag{8}$$

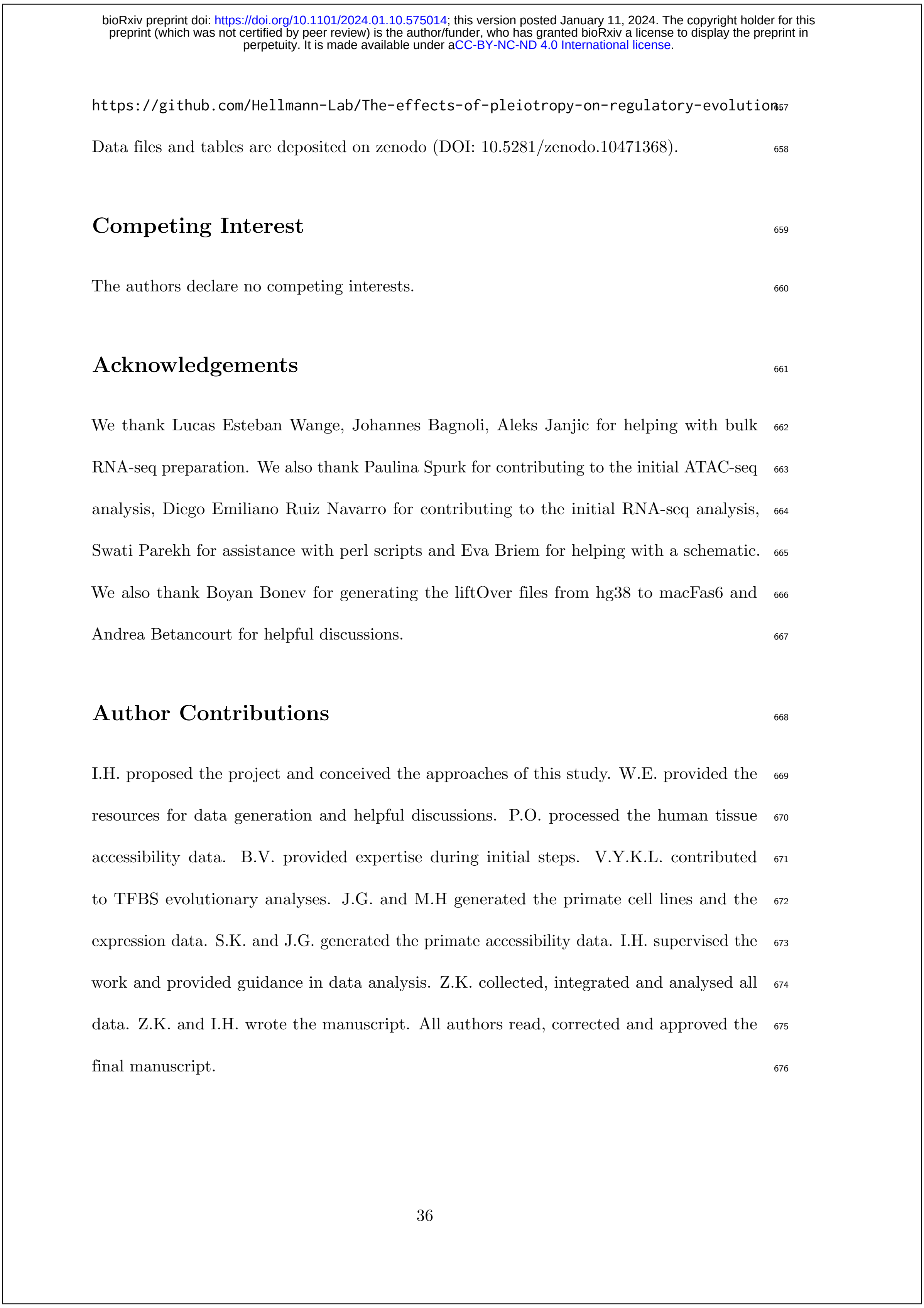
where B is a set of positions in the alignment that overlap with a binding site of motif i in the respective species macaque M or human H.

Quantification and Statistical Analysis

Data visualizations and statistical analysis was performed using R (version 4.2.3) (R Core Team 2023), session info can be accessed on GitHub. Details of the statistical tests performed in this study can be found in the main text as well as the method details section. Schematics were made using bioRender.

Data and Code Access

RNA-seq and ATAC-seq data are available under ArrayExpress accessions E-MTAB-13494 and E-MTAB-13373. A compendium containing processing scripts, important tables and detailed instructions to reproduce the analysis for this manuscript is available from the following GitHub repository:



`https://github.com/Hellmann-Lab/The-effects-of-pleiotropy-on-regulatory-evolution`

Data files and tables are deposited on zenodo (DOI: 10.5281/zenodo.10471368).

Competing Interest

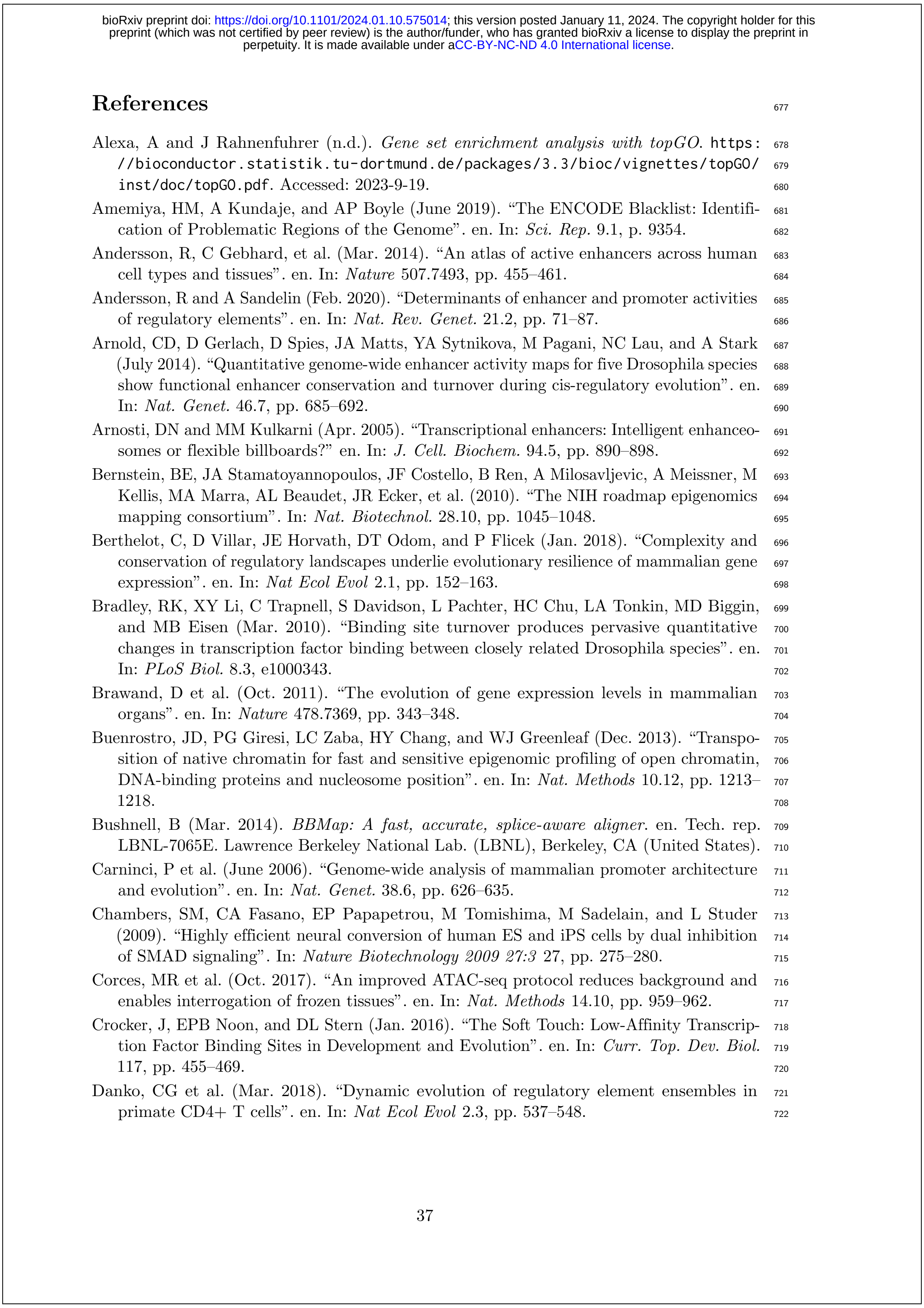
The authors declare no competing interests.

Acknowledgements

We thank Lucas Esteban Wange, Johannes Bagnoli, Aleks Janjic for helping with bulk RNA-seq preparation. We also thank Paulina Spurk for contributing to the initial ATAC-seq analysis, Diego Emiliano Ruiz Navarro for contributing to the initial RNA-seq analysis, Swati Parekh for assistance with perl scripts and Eva Briem for helping with a schematic. We also thank Boyan Bonev for generating the liftOver files from hg38 to macFas6 and Andrea Betancourt for helpful discussions.

Author Contributions

I.H. proposed the project and conceived the approaches of this study. W.E. provided the resources for data generation and helpful discussions. P.O. processed the human tissue accessibility data. B.V. provided expertise during initial steps. V.Y.K.L. contributed to TFBS evolutionary analyses. J.G. and M.H generated the primate cell lines and the expression data. S.K. and J.G. generated the primate accessibility data. I.H. supervised the work and provided guidance in data analysis. Z.K. collected, integrated and analysed all data. Z.K. and I.H. wrote the manuscript. All authors read, corrected and approved the final manuscript.



References

Alexa, A and J Rahnenfuhrer (n.d.). *Gene set enrichment analysis with topGO.* https://bioconductor.statistik.tu-dortmund.de/packages/3.3/bioc/vignettes/topGO/inst/doc/topGO.pdf. Accessed: 2023-9-19.

Amemiya, HM, A Kundaje, and AP Boyle (June 2019). "The ENCODE Blacklist: Identification of Problematic Regions of the Genome". en. In: *Sci. Rep.* 9.1, p. 9354.

Andersson, R, C Gebhard, et al. (Mar. 2014). "An atlas of active enhancers across human cell types and tissues". en. In: *Nature* 507.7493, pp. 455–461.

Andersson, R and A Sandelin (Feb. 2020). "Determinants of enhancer and promoter activities of regulatory elements". en. In: *Nat. Rev. Genet.* 21.2, pp. 71–87.

Arnold, CD, D Gerlach, D Spies, JA Matts, YA Sytnikova, M Pagani, NC Lau, and A Stark (July 2014). "Quantitative genome-wide enhancer activity maps for five Drosophila species show functional enhancer conservation and turnover during cis-regulatory evolution". en. In: *Nat. Genet.* 46.7, pp. 685–692.

Arnosti, DN and MM Kulkarni (Apr. 2005). "Transcriptional enhancers: Intelligent enhanceosomes or flexible billboards?" en. In: *J. Cell. Biochem.* 94.5, pp. 890–898.

Bernstein, BE, JA Stamatoyannopoulos, JF Costello, B Ren, A Milosavljevic, A Meissner, M Kellis, MA Marra, AL Beaudet, JR Ecker, et al. (2010). "The NIH roadmap epigenomics mapping consortium". In: *Nat. Biotechnol.* 28.10, pp. 1045–1048.

Berthelot, C, D Villar, JE Horvath, DT Odom, and P Flicek (Jan. 2018). "Complexity and conservation of regulatory landscapes underlie evolutionary resilience of mammalian gene expression". en. In: *Nat Ecol Evol* 2.1, pp. 152–163.

Bradley, RK, XY Li, C Trapnell, S Davidson, L Pachter, HC Chu, LA Tonkin, MD Biggin, and MB Eisen (Mar. 2010). "Binding site turnover produces pervasive quantitative changes in transcription factor binding between closely related Drosophila species". en. In: *PLoS Biol.* 8.3, e1000343.

Brawand, D et al. (Oct. 2011). "The evolution of gene expression levels in mammalian organs". en. In: *Nature* 478.7369, pp. 343–348.

Buenrostro, JD, PG Giresi, LC Zaba, HY Chang, and WJ Greenleaf (Dec. 2013). "Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position". en. In: *Nat. Methods* 10.12, pp. 1213–1218.

Bushnell, B (Mar. 2014). *BBMap: A fast, accurate, splice-aware aligner.* en. Tech. rep. LBNL-7065E. Lawrence Berkeley National Lab. (LBNL), Berkeley, CA (United States).

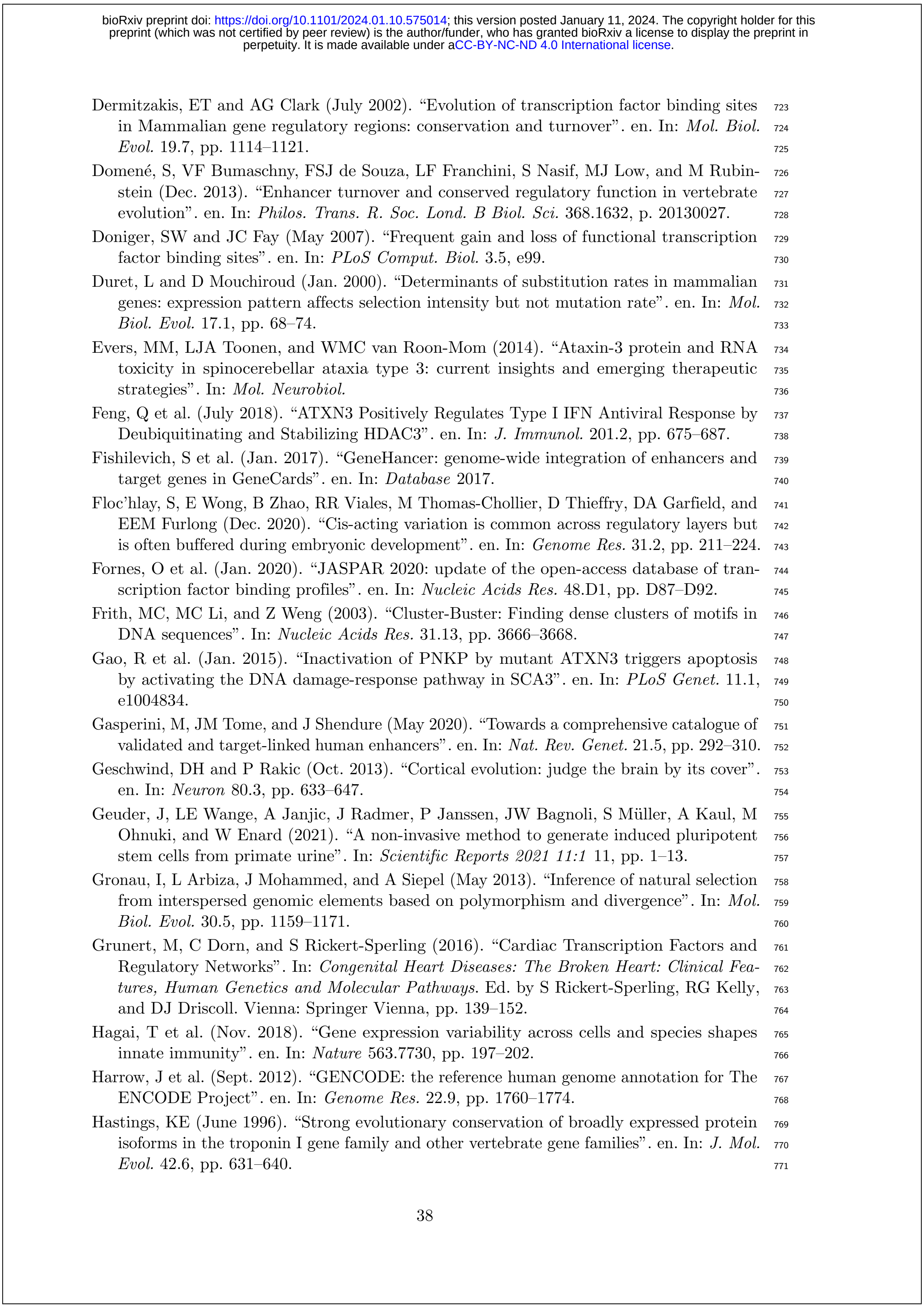
Carninci, P et al. (June 2006). "Genome-wide analysis of mammalian promoter architecture and evolution". en. In: *Nat. Genet.* 38.6, pp. 626–635.

Chambers, SM, CA Fasano, EP Papapetrou, M Tomishima, M Sadelain, and L Studer (2009). "Highly efficient neural conversion of human ES and iPS cells by dual inhibition of SMAD signaling". In: *Nature Biotechnology 2009 27:3* 27, pp. 275–280.

Corces, MR et al. (Oct. 2017). "An improved ATAC-seq protocol reduces background and enables interrogation of frozen tissues". en. In: *Nat. Methods* 14.10, pp. 959–962.

Crocker, J, EPB Noon, and DL Stern (Jan. 2016). "The Soft Touch: Low-Affinity Transcription Factor Binding Sites in Development and Evolution". en. In: *Curr. Top. Dev. Biol.* 117, pp. 455–469.

Danko, CG et al. (Mar. 2018). "Dynamic evolution of regulatory element ensembles in primate CD4+ T cells". en. In: *Nat Ecol Evol* 2.3, pp. 537–548.



Dermitzakis, ET and AG Clark (July 2002). “Evolution of transcription factor binding sites in Mammalian gene regulatory regions: conservation and turnover”. en. In: *Mol. Biol. Evol.* 19.7, pp. 1114–1121.

Domené, S, VF Bumaschny, FSJ de Souza, LF Franchini, S Nasif, MJ Low, and M Rubinstein (Dec. 2013). “Enhancer turnover and conserved regulatory function in vertebrate evolution”. en. In: *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 368.1632, p. 20130027.

Doniger, SW and JC Fay (May 2007). “Frequent gain and loss of functional transcription factor binding sites”. en. In: *PLoS Comput. Biol.* 3.5, e99.

Duret, L and D Mouchiroud (Jan. 2000). “Determinants of substitution rates in mammalian genes: expression pattern affects selection intensity but not mutation rate”. en. In: *Mol. Biol. Evol.* 17.1, pp. 68–74.

Evers, MM, LJA Toonen, and WMC van Roon-Mom (2014). “Ataxin-3 protein and RNA toxicity in spinocerebellar ataxia type 3: current insights and emerging therapeutic strategies”. In: *Mol. Neurobiol.*

Feng, Q et al. (July 2018). “ATXN3 Positively Regulates Type I IFN Antiviral Response by Deubiquitinating and Stabilizing HDAC3”. en. In: *J. Immunol.* 201.2, pp. 675–687.

Fishilevich, S et al. (Jan. 2017). “GeneHancer: genome-wide integration of enhancers and target genes in GeneCards”. en. In: *Database* 2017.

Floc’hlay, S, E Wong, B Zhao, RR Viales, M Thomas-Chollier, D Thieffry, DA Garfield, and EEM Furlong (Dec. 2020). “Cis-acting variation is common across regulatory layers but is often buffered during embryonic development”. en. In: *Genome Res.* 31.2, pp. 211–224.

Fornes, O et al. (Jan. 2020). “JASPAR 2020: update of the open-access database of transcription factor binding profiles”. en. In: *Nucleic Acids Res.* 48.D1, pp. D87–D92.

Frith, MC, MC Li, and Z Weng (2003). “Cluster-Buster: Finding dense clusters of motifs in DNA sequences”. In: *Nucleic Acids Res.* 31.13, pp. 3666–3668.

Gao, R et al. (Jan. 2015). “Inactivation of PNKP by mutant ATXN3 triggers apoptosis by activating the DNA damage-response pathway in SCA3”. en. In: *PLoS Genet.* 11.1, e1004834.

Gasperini, M, JM Tome, and J Shendure (May 2020). “Towards a comprehensive catalogue of validated and target-linked human enhancers”. en. In: *Nat. Rev. Genet.* 21.5, pp. 292–310.

Geschwind, DH and P Rakic (Oct. 2013). “Cortical evolution: judge the brain by its cover”. en. In: *Neuron* 80.3, pp. 633–647.

Geuder, J, LE Wange, A Janjic, J Radmer, P Janssen, JW Bagnoli, S Müller, A Kaul, M Ohnuki, and W Enard (2021). “A non-invasive method to generate induced pluripotent stem cells from primate urine”. In: *Scientific Reports 2021 11:1* 11, pp. 1–13.

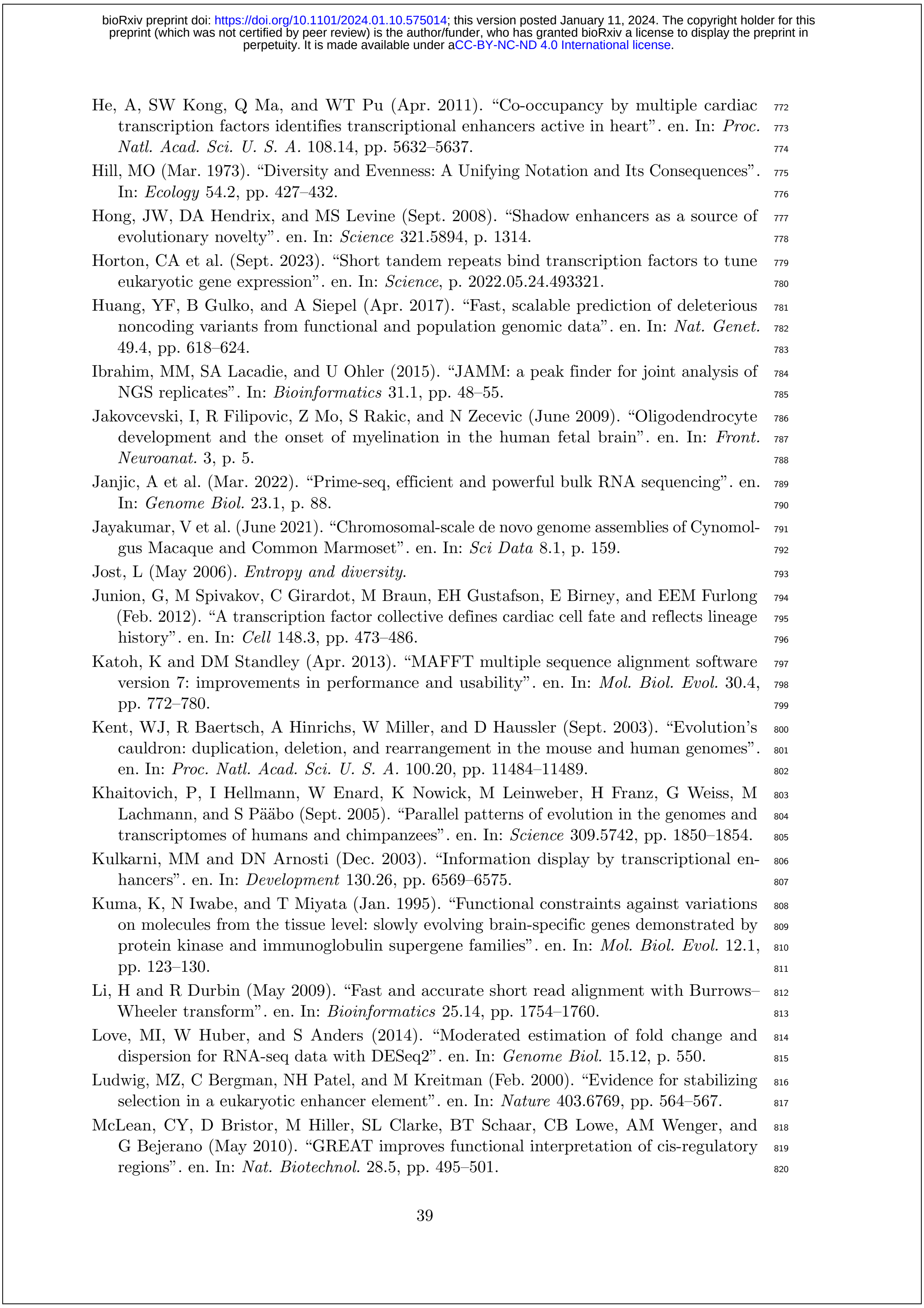
Gronau, I, L Arbiza, J Mohammed, and A Siepel (May 2013). “Inference of natural selection from interspersed genomic elements based on polymorphism and divergence”. In: *Mol. Biol. Evol.* 30.5, pp. 1159–1171.

Grunert, M, C Dorn, and S Rickert-Sperling (2016). “Cardiac Transcription Factors and Regulatory Networks”. In: *Congenital Heart Diseases: The Broken Heart: Clinical Features, Human Genetics and Molecular Pathways.* Ed. by S Rickert-Sperling, RG Kelly, and DJ Driscoll. Vienna: Springer Vienna, pp. 139–152.

Hagai, T et al. (Nov. 2018). “Gene expression variability across cells and species shapes innate immunity”. en. In: *Nature* 563.7730, pp. 197–202.

Harrow, J et al. (Sept. 2012). “GENCODE: the reference human genome annotation for The ENCODE Project”. en. In: *Genome Res.* 22.9, pp. 1760–1774.

Hastings, KE (June 1996). “Strong evolutionary conservation of broadly expressed protein isoforms in the troponin I gene family and other vertebrate gene families”. en. In: *J. Mol. Evol.* 42.6, pp. 631–640.



He, A, SW Kong, Q Ma, and WT Pu (Apr. 2011). "Co-occupancy by multiple cardiac transcription factors identifies transcriptional enhancers active in heart". en. In: *Proc. Natl. Acad. Sci. U. S. A.* 108.14, pp. 5632–5637.

Hill, MO (Mar. 1973). "Diversity and Evenness: A Unifying Notation and Its Consequences". In: *Ecology* 54.2, pp. 427–432.

Hong, JW, DA Hendrix, and MS Levine (Sept. 2008). "Shadow enhancers as a source of evolutionary novelty". en. In: *Science* 321.5894, p. 1314.

Horton, CA et al. (Sept. 2023). "Short tandem repeats bind transcription factors to tune eukaryotic gene expression". en. In: *Science*, p. 2022.05.24.493321.

Huang, YF, B Gulko, and A Siepel (Apr. 2017). "Fast, scalable prediction of deleterious noncoding variants from functional and population genomic data". en. In: *Nat. Genet.* 49.4, pp. 618–624.

Ibrahim, MM, SA Lacadie, and U Ohler (2015). "JAMM: a peak finder for joint analysis of NGS replicates". In: *Bioinformatics* 31.1, pp. 48–55.

Jakovcevski, I, R Filipovic, Z Mo, S Rakic, and N Zecevic (June 2009). "Oligodendrocyte development and the onset of myelination in the human fetal brain". en. In: *Front. Neuroanat.* 3, p. 5.

Janjic, A et al. (Mar. 2022). "Prime-seq, efficient and powerful bulk RNA sequencing". en. In: *Genome Biol.* 23.1, p. 88.

Jayakumar, V et al. (June 2021). "Chromosomal-scale de novo genome assemblies of Cynomolgus Macaque and Common Marmoset". en. In: *Sci Data* 8.1, p. 159.

Jost, L (May 2006). *Entropy and diversity.*

Junion, G, M Spivakov, C Girardot, M Braun, EH Gustafson, E Birney, and EEM Furlong (Feb. 2012). "A transcription factor collective defines cardiac cell fate and reflects lineage history". en. In: *Cell* 148.3, pp. 473–486.

Katoh, K and DM Standley (Apr. 2013). "MAFFT multiple sequence alignment software version 7: improvements in performance and usability". en. In: *Mol. Biol. Evol.* 30.4, pp. 772–780.

Kent, WJ, R Baertsch, A Hinrichs, W Miller, and D Haussler (Sept. 2003). "Evolution's cauldron: duplication, deletion, and rearrangement in the mouse and human genomes". en. In: *Proc. Natl. Acad. Sci. U. S. A.* 100.20, pp. 11484–11489.

Khaitovich, P, I Hellmann, W Enard, K Nowick, M Leinweber, H Franz, G Weiss, M Lachmann, and S Pääbo (Sept. 2005). "Parallel patterns of evolution in the genomes and transcriptomes of humans and chimpanzees". en. In: *Science* 309.5742, pp. 1850–1854.

Kulkarni, MM and DN Arnosti (Dec. 2003). "Information display by transcriptional enhancers". en. In: *Development* 130.26, pp. 6569–6575.

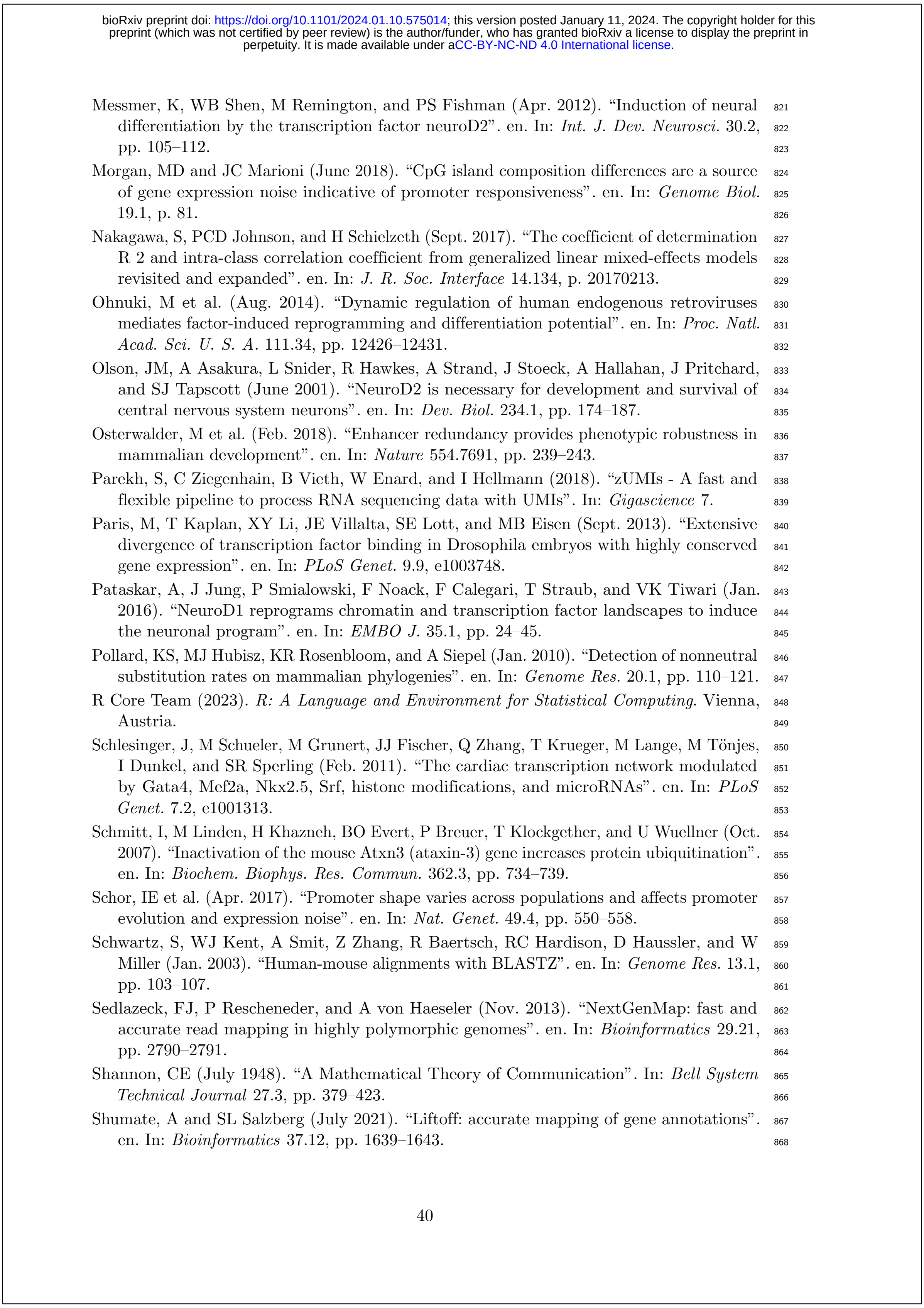
Kuma, K, N Iwabe, and T Miyata (Jan. 1995). "Functional constraints against variations on molecules from the tissue level: slowly evolving brain-specific genes demonstrated by protein kinase and immunoglobulin supergene families". en. In: *Mol. Biol. Evol.* 12.1, pp. 123–130.

Li, H and R Durbin (May 2009). "Fast and accurate short read alignment with Burrows–Wheeler transform". en. In: *Bioinformatics* 25.14, pp. 1754–1760.

Love, MI, W Huber, and S Anders (2014). "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2". en. In: *Genome Biol.* 15.12, p. 550.

Ludwig, MZ, C Bergman, NH Patel, and M Kreitman (Feb. 2000). "Evidence for stabilizing selection in a eukaryotic enhancer element". en. In: *Nature* 403.6769, pp. 564–567.

McLean, CY, D Bristor, M Hiller, SL Clarke, BT Schaar, CB Lowe, AM Wenger, and G Bejerano (May 2010). "GREAT improves functional interpretation of cis-regulatory regions". en. In: *Nat. Biotechnol.* 28.5, pp. 495–501.



Messmer, K, WB Shen, M Remington, and PS Fishman (Apr. 2012). “Induction of neural differentiation by the transcription factor neuroD2”. en. In: *Int. J. Dev. Neurosci.* 30.2, pp. 105–112.

Morgan, MD and JC Marioni (June 2018). “CpG island composition differences are a source of gene expression noise indicative of promoter responsiveness”. en. In: *Genome Biol.* 19.1, p. 81.

Nakagawa, S, PCD Johnson, and H Schielzeth (Sept. 2017). “The coefficient of determination R 2 and intra-class correlation coefficient from generalized linear mixed-effects models revisited and expanded”. en. In: *J. R. Soc. Interface* 14.134, p. 20170213.

Ohnuki, M et al. (Aug. 2014). “Dynamic regulation of human endogenous retroviruses mediates factor-induced reprogramming and differentiation potential”. en. In: *Proc. Natl. Acad. Sci. U. S. A.* 111.34, pp. 12426–12431.

Olson, JM, A Asakura, L Snider, R Hawkes, A Strand, J Stoeck, A Hallahan, J Pritchard, and SJ Tapscott (June 2001). “NeuroD2 is necessary for development and survival of central nervous system neurons”. en. In: *Dev. Biol.* 234.1, pp. 174–187.

Osterwalder, M et al. (Feb. 2018). “Enhancer redundancy provides phenotypic robustness in mammalian development”. en. In: *Nature* 554.7691, pp. 239–243.

Parekh, S, C Ziegenhain, B Vieth, W Enard, and I Hellmann (2018). “zUMIs - A fast and flexible pipeline to process RNA sequencing data with UMIs”. In: *Gigascience* 7.

Paris, M, T Kaplan, XY Li, JE Villalta, SE Lott, and MB Eisen (Sept. 2013). “Extensive divergence of transcription factor binding in Drosophila embryos with highly conserved gene expression”. en. In: *PLoS Genet.* 9.9, e1003748.

Pataskar, A, J Jung, P Smialowski, F Noack, F Calegari, T Straub, and VK Tiwari (Jan. 2016). “NeuroD1 reprograms chromatin and transcription factor landscapes to induce the neuronal program”. en. In: *EMBO J.* 35.1, pp. 24–45.

Pollard, KS, MJ Hubisz, KR Rosenbloom, and A Siepel (Jan. 2010). “Detection of nonneutral substitution rates on mammalian phylogenies”. en. In: *Genome Res.* 20.1, pp. 110–121.

R Core Team (2023). *R: A Language and Environment for Statistical Computing.* Vienna, Austria.

Schlesinger, J, M Schueler, M Grunert, JJ Fischer, Q Zhang, T Krueger, M Lange, M Tönjes, I Dunkel, and SR Sperling (Feb. 2011). “The cardiac transcription network modulated by Gata4, Mef2a, Nkx2.5, Srf, histone modifications, and microRNAs”. en. In: *PLoS Genet.* 7.2, e1001313.

Schmitt, I, M Linden, H Khazneh, BO Evert, P Breuer, T Klockgether, and U Wuellner (Oct. 2007). “Inactivation of the mouse Atxn3 (ataxin-3) gene increases protein ubiquitination”. en. In: *Biochem. Biophys. Res. Commun.* 362.3, pp. 734–739.

Schor, IE et al. (Apr. 2017). “Promoter shape varies across populations and affects promoter evolution and expression noise”. en. In: *Nat. Genet.* 49.4, pp. 550–558.

Schwartz, S, WJ Kent, A Smit, Z Zhang, R Baertsch, RC Hardison, D Haussler, and W Miller (Jan. 2003). “Human-mouse alignments with BLASTZ”. en. In: *Genome Res.* 13.1, pp. 103–107.

Sedlazeck, FJ, P Rescheneder, and A von Haeseler (Nov. 2013). “NextGenMap: fast and accurate read mapping in highly polymorphic genomes”. en. In: *Bioinformatics* 29.21, pp. 2790–2791.

Shannon, CE (July 1948). “A Mathematical Theory of Communication”. In: *Bell System Technical Journal* 27.3, pp. 379–423.

Shumate, A and SL Salzberg (July 2021). “Liftoff: accurate mapping of gene annotations”. en. In: *Bioinformatics* 37.12, pp. 1639–1643.

Sigalova, OM, A Shaeiri, M Forneris, EE Furlong, and JB Zaugg (Aug. 2020). "Predictive features of gene expression variation reveal mechanistic link with differential expression". en. In: *Mol. Syst. Biol.* 16.8, e9539.

Singh, D and SV Yi (Mar. 2021). "Enhancer Pleiotropy, Gene Expression, and the Architecture of Human Enhancer–Gene Interactions". en. In: *Mol. Biol. Evol.* 38.9, pp. 3898–3909.

Stefflova, K et al. (Aug. 2013). "Cooperativity and rapid evolution of cobound transcription factors in closely related mammals". en. In: *Cell* 154.3, pp. 530–540.

Sun, Y, M Nadal-Vicens, S Misono, MZ Lin, A Zubiaga, X Hua, G Fan, and ME Greenberg (Feb. 2001). "Neurogenin promotes neurogenesis and inhibits glial differentiation by independent mechanisms". en. In: *Cell* 104.3, pp. 365–376.

Tan, G and B Lenhard (May 2016). "TFBSTools: an R/bioconductor package for transcription factor binding site analysis". en. In: *Bioinformatics* 32.10, pp. 1555–1556.

Tuğrul, M, T Paixão, NH Barton, and G Tkačik (Nov. 2015). "Dynamics of Transcription Factor Binding Site Evolution". en. In: *PLoS Genet.* 11.11, e1005639.

Vasimuddin, M, S Misra, H Li, and S Aluru (May 2019). "Efficient Architecture-Aware Acceleration of BWA-MEM for Multicore Systems". In: *2019 IEEE International Parallel and Distributed Processing Symposium (IPDPS)*, pp. 314–324.

Villar, D, P Flicek, and DT Odom (Apr. 2014). "Evolution of transcription factor binding in metazoans — mechanisms and functional implications". en. In: *Nat. Rev. Genet.* 15.4, pp. 221–233.

Wang, HY, HC Chien, N Osada, K Hashimoto, S Sugano, T Gojobori, CK Chou, SF Tsai, CI Wu, and CKJ Shen (Feb. 2007). "Rate of evolution in brain-expressed genes in humans and other primates". en. In: *PLoS Biol.* 5.2, e13.

Wang, ZY et al. (Dec. 2020). "Transcriptome and translatome co-evolution in mammals". en. In: *Nature* 588.7839, pp. 642–647.

Wunderlich, Z, MDJ Bragdon, BJ Vincent, JA White, J Estrada, and AH DePace (Mar. 2016). "Krüppel Expression Levels Are Maintained through Compensatory Evolution of Shadow Enhancers". en. In: *Cell Rep.* 14.12, p. 3030.

Yevshin, I, R Sharipov, S Kolmykov, Y Kondrakhin, and F Kolpakov (Nov. 2018). "GTRD: a database on gene transcription regulation—2019 update". en. In: *Nucleic Acids Res.* 47.D1, pp. D100–D105.

Yu, Y et al. (Jan. 2013). "Olig2 targets chromatin remodelers to enhancers to initiate oligodendrocyte differentiation". en. In: *Cell* 152.1-2, pp. 248–261.

Zhang, L and WH Li (Feb. 2004). "Mammalian housekeeping genes evolve more slowly than tissue-specific genes". en. In: *Mol. Biol. Evol.* 21.2, pp. 236–239.

Zhao, Y et al. (Sept. 2022). ""Stripe" transcription factors provide accessibility to co-binding partners in mammalian genomes". en. In: *Mol. Cell* 82.18, 3398–3411.e11.

Zhou, Q and DJ Anderson (Apr. 2002). "The bHLH transcription factors OLIG2 and OLIG1 couple neuronal and glial subtype specification". en. In: *Cell* 109.1, pp. 61–73.

Supplemental Information

Evidence for compensatory evolution within pleiotropic regulatory elements

Zane Kliesmete[1], Peter Orchard[1,2], Victor Yan Kin Lee[1,3], Johanna Geuder[1], Simon M. Krauß[1,4], Mari Ohnuki[1,5], Jessica Radmer[1], Beate Vieth[1], Wolfgang Enard[1], Ines Hellmann[1,*]

[1] Anthropology and Human Genomics, Faculty of Biology, Ludwig-Maximilians Universität München, Munich, Germany

[2] Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI, USA

[3] Section for Molecular Ecology and Evolution, Globe Institute, University of Copenhagen, Copenhagen, Denmark

[4] Department of Hematology, Cell Therapy, Hemostaseology and Infectious Diseases, University Leipzig Medical Center, Leipzig, Germany

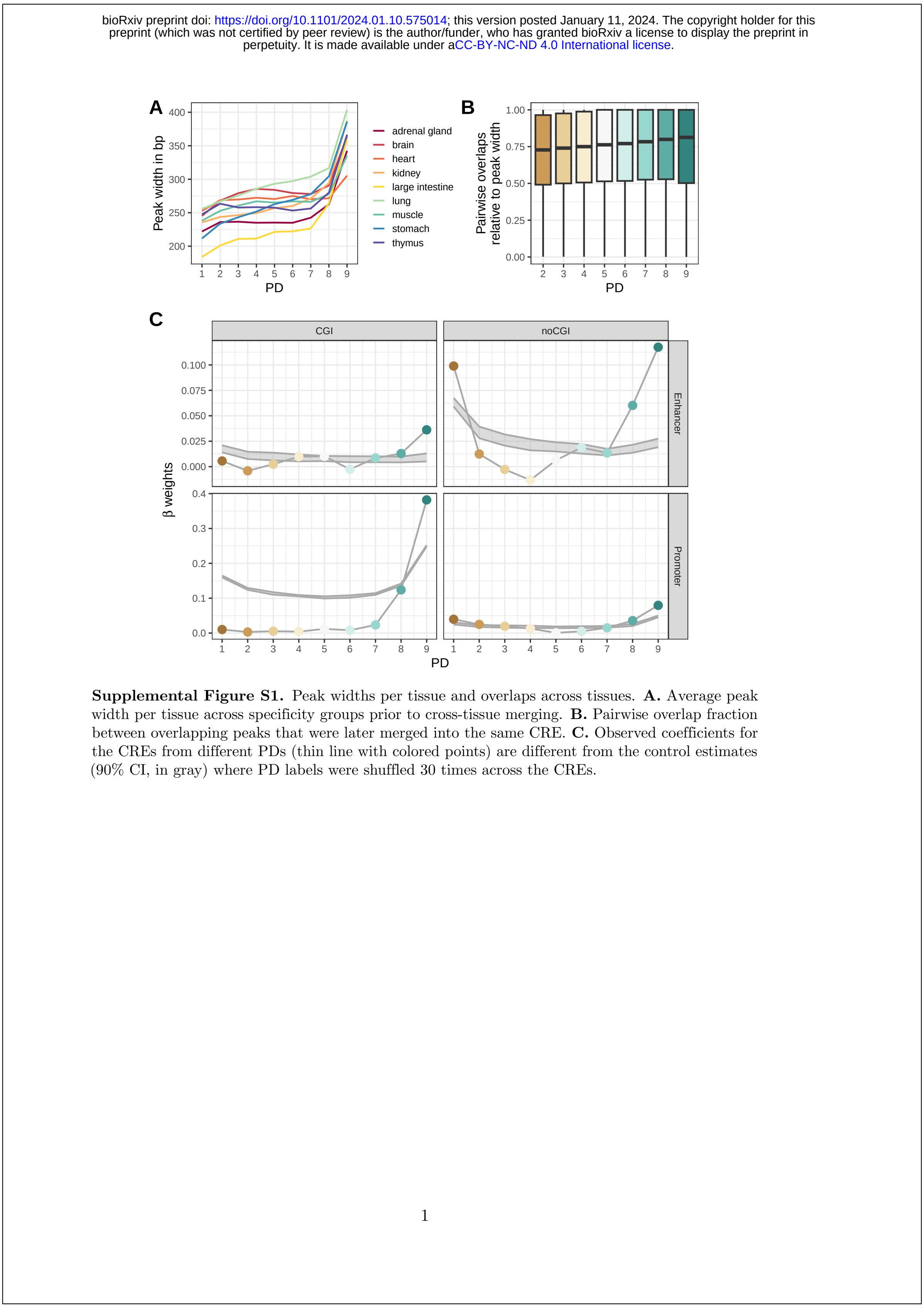
[*] correspondence:

Dr. Ines Hellmann,

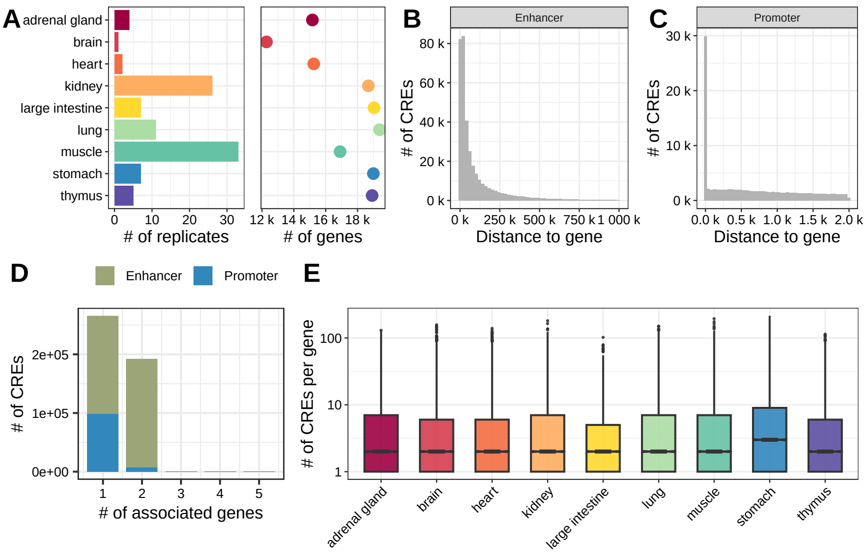
Telefon +49 (0)89 2180-74336

Telefax +49 (0)89 2180-74331

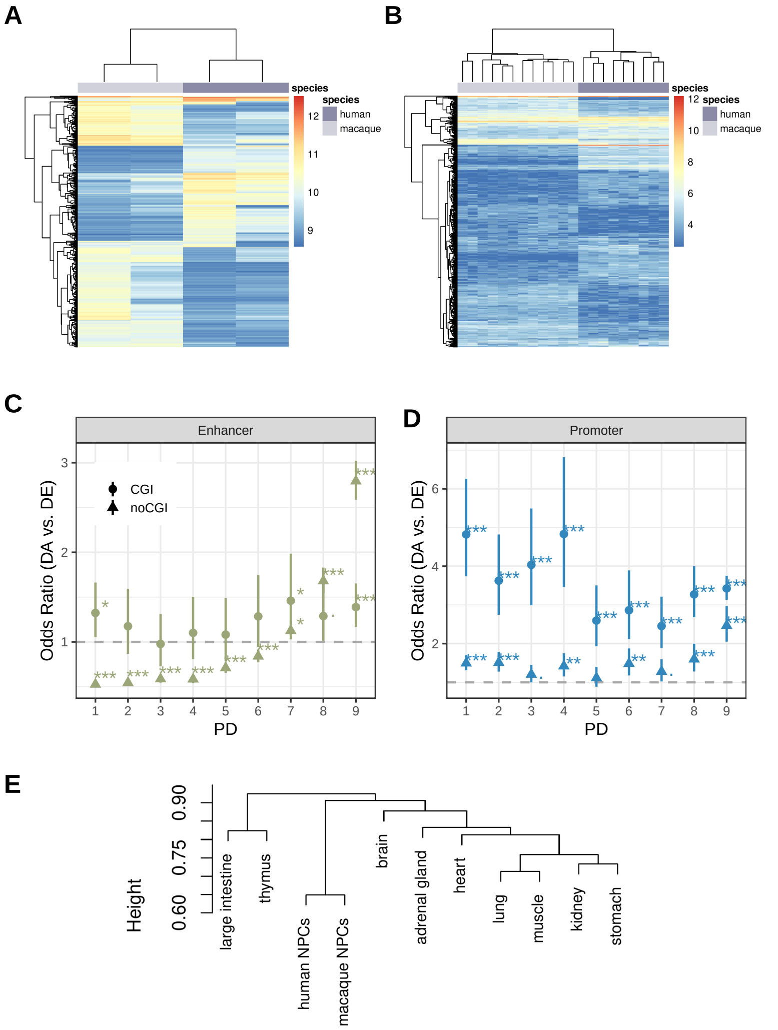
hellmann@bio.lmu.de, www.anthropologie.bio.lmu.de



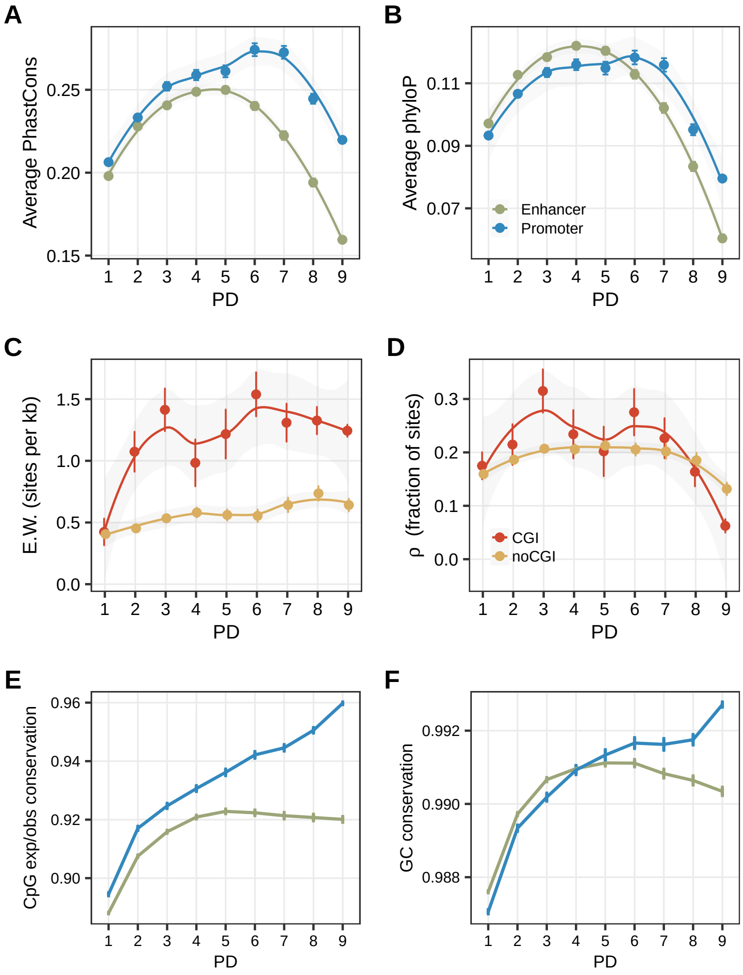
Supplemental Figure S1. Peak widths per tissue and overlaps across tissues. **A.** Average peak width per tissue across specificity groups prior to cross-tissue merging. **B.** Pairwise overlap fraction between overlapping peaks that were later merged into the same CRE. **C.** Observed coefficients for the CREs from different PDs (thin line with colored points) are different from the control estimates (90% CI, in gray) where PD labels were shuffled 30 times across the CREs.



Supplemental Figure S2. CRE to gene association across tissues. **A.** Expression data overview: Number of replicates and number of expressed genes for each tissue. **B-C.** Distance distribution of CREs annotated as enhancers (**B**) and promoters (**C**) to their closest gene. **D.** Number of associated genes per CRE. Enhancers were associated with up to 2 genes, promoters: up to 5. **E.** Associated number of CREs per gene is comparable across tissues.

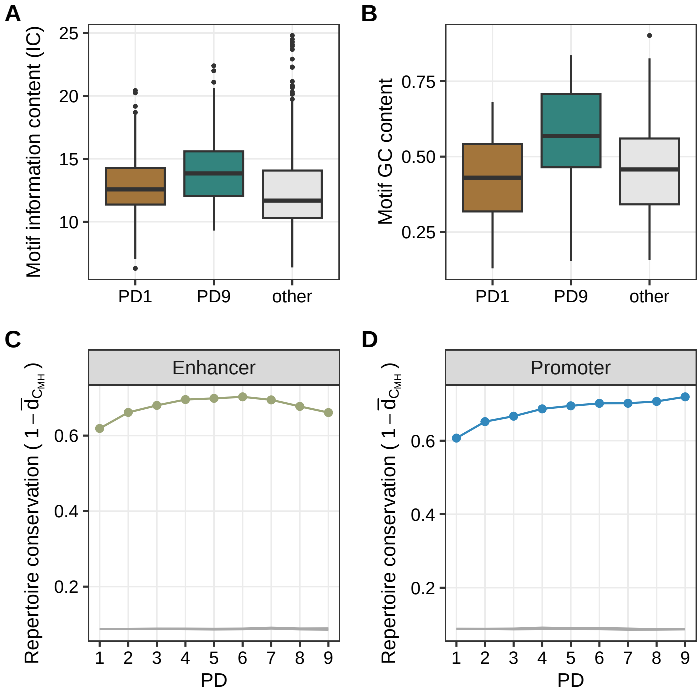


Supplemental Figure S3. Cross-species NPC expression and accessibility. **A.** Heatmap of the top 1000 most variable CREs across the ATAC-seq data in terms of their openness (Euclidean distance, method: complete). **B.** Heatmap of the top 1000 most variable genes across the RNA-seq data (Euclidean distance, method: complete). **C.** Odds ratios of differential accessiblity of enhancers vs. the associated gene differential expression results between humans and cynomolgus macaques, split between CGI and non-CGI CREs. **D.** Odds ratios of differential accessiblity of promoters vs. the associated gene differential expression results between humans and cynomolgus macaques, split between CGI and non-CGI CREs. **C,D.** Error bars represent the 95% confidence intervals of the odd ratio, the stars indicate the significance level after Benjamini-Hochberg correction ($^{\cdot} < 0.1$, $^{*} < 0.05$, $^{**} < 0.01$, $^{***} < 0.001$). **E.** Hierarchical clustering (method = ”binary”, distance = ”complete”) of human and macaque NPC ATAC-seq samples together with human tissue CREs based on the variable CRE binary openness across tissues (PD1-8).



Supplemental Figure S4. Evolutionary sequence analysis of CREs across tissue specificity groups. **A.** PhastCons conservation scores based on a 10-species primate tree. **B.** PhyloP conservation scores based on a 10-species primate tree. **C.** Weak negative selection patterns between CGI and non-CGI CREs suggest an increasing weak selection with higher PD in both groups. **D.** (Strong) negative selection patterns between CGI and non-CGI CREs show similar trend across PDs. **E.** Absolute pairwise distance in CpG expected/observed ratio between human and cynomolgus macaque orthologous CREs. **F.** Absolute pairwise distance in GC content between human and cynomolgus macaque orthologous CREs. **A,B,C,D,E,F.** Depicted are mean estimates per PD category. Error bars indicate SEM.

.



Supplemental Figure S5. TFBS repertoire diversity and conservation. **A.** Motif information content of the PD1-enriched, PD9-enriched and other motifs, classified as in 2D. **B.** Motif GC content of the PD1-enriched, PD9-enriched and other motifs, classified as in 2D. **C,D.** 95% confidence intervals of the average TFBS repertoire conservation when shuffling macaque CRE identifiers within the respective PD class 10 times (grey line). Random CRE similarity is below 10% and does not increase with PD. In comparison, the real observed enhancer and promoter repertoire conservation is depicted in green (**C.**) and blue (**D.**), respectively.

2.3 The pluripotent stem cell-specific transcript ESRG is dispensable for human pluripotency

Takahashi K, Nakamura M, Okubo C, **Kliesmete Z**, Ohnuki M, Narita M, Watanabe A, Ueda M, Takashima Y, Hellmann I, Yamanaka S:

"The pluripotent stem cell-specific transcript ESRG is dispensable for human pluripotency." (2021)

PLoS genetics 17(5):e1009587.

doi: 10.1371/journal.pgen.1009587

Supplementary Information is freely available at the publisher's website:

`https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1009587`

PLOS GENETICS

RESEARCH ARTICLE

The pluripotent stem cell-specific transcript ESRG is dispensable for human pluripotency

Kazutoshi Takahashi[1,2]*, **Michiko Nakamura**[1], **Chikako Okubo**[1], **Zane Kliesmete**[3], **Mari Ohnuki**[3], **Megumi Narita**[1], **Akira Watanabe**[4], **Mai Ueda**[1], **Yasuhiro Takashima**[1], **Ines Hellmann**[3], **Shinya Yamanaka**[1,2,5]

1 Center for iPS Cell Research and Application, Kyoto University, Kyoto, Japan, **2** Gladstone Institute of Cardiovascular Disease, San Francisco, California, United States of America, **3** Anthropology and Human Genomics, Department of Biology II, Ludwig-Maximilians Universitaet, Munich, Germany, **4** Graduate School of Medicine, Kyoto University, Kyoto, Japan, **5** Department of Anatomy, University of California, San Francisco, San Francisco, California, United States of America

* kazu@cira.kyoto-u.ac.jp

Check for updates

OPEN ACCESS

Citation: Takahashi K, Nakamura M, Okubo C, Kliesmete Z, Ohnuki M, Narita M, et al. (2021) The pluripotent stem cell-specific transcript ESRG is dispensable for human pluripotency. PLoS Genet 17(5): e1009587. https://doi.org/10.1371/journal.pgen.1009587

Editor: Marisa S. Bartolomei, University of Pennsylvania, UNITED STATES

Received: November 25, 2020

Accepted: May 6, 2021

Published: May 25, 2021

Peer Review History: PLOS recognizes the benefits of transparency in the peer review process; therefore, we enable the publication of all of the content of peer review and author responses alongside final, published articles. The editorial history of this article is available here: https://doi.org/10.1371/journal.pgen.1009587

Data Availability Statement: The numerical values for the graphs in the manuscript are provided as the Supporting Data. RNA-seq (GSE56568 and GSE171849), ChIP-seq (GSE56567 and

Abstract

Human pluripotent stem cells (PSCs) express human endogenous retrovirus type-H (HERV-H), which exists as more than a thousand copies on the human genome and frequently produces chimeric transcripts as long-non-coding RNAs (lncRNAs) fused with downstream neighbor genes. Previous studies showed that HERV-H expression is required for the maintenance of PSC identity, and aberrant HERV-H expression attenuates neural differentiation potentials, however, little is known about the actual of function of HERV-H. In this study, we focused on ESRG, which is known as a PSC-related HERV-H-driven lncRNA. The global transcriptome data of various tissues and cell lines and quantitative expression analysis of PSCs showed that ESRG expression is much higher than other HERV-Hs and tightly silenced after differentiation. However, the loss of function by the complete excision of the entire ESRG gene body using a CRISPR/Cas9 platform revealed that ESRG is dispensable for the maintenance of the primed and naïve pluripotent states. The loss of ESRG hardly affected the global gene expression of PSCs or the differentiation potential toward tri-lineage. Differentiated cells derived from ESRG-deficient PSCs retained the potential to be reprogrammed into induced PSCs (iPSCs) by the forced expression of OCT3/4, SOX2, and KLF4. In conclusion, ESRG is dispensable for the maintenance and recapturing of human pluripotency.

Author summary

We have been interested in the role of human endogenous retrovirus (HERVs) in human pluripotent stem cells (PSCs). Although we and others have demonstrated that HERV expression is crucial for somatic cell reprogramming to a pluripotent state and the characteristics of PSCs. Little is known which one of more than 1,000 copies of HERVs is important. Thus, in this study, we focused on a HERV-related gene, ESRG which is expressed strongly and specifically in human PSCs but not in differentiated cells. Using a

GSE89976) and Gene expression microarray (GSE54848, GSE156834, GSE159101, and GSE171627) results are accessible in the Gene Expression Omnibus database of the National Center for Biotechnology Information website.

Funding: This work was supported by Grants-in-Aid for Scientific Research (20K20585) to K.T. from the Japanese Society for the Promotion of Science (JSPS); a grant from the Core Center for iPS Cell Research (JP21bm0104001) to S.Y., Research Center Network for Realization of Regenerative Medicine from Japan Agency for Medical Research and Development (AMED) to S. Y.; a grant from the Japan Foundation for Applied Enzymology to K.T.; a grant from the Fujiwara Memorial Foundation to K.T.; a grant from the Takeda Science Foundation to K.T.; and the iPS Cell Research Fund to K.T. from Center for iPS Cell Research and Application, Kyoto University. The study was also supported by funding to S.Y. from Mr. H. Mikitani, Mr. M. Benioff, and the L.K. Whittier Foundation. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: I have read the journal's policy and the authors of this manuscript have the following competing interests: K.T. is on the scientific advisory board of I Peace, Inc. without salary, S.Y. is a scientific advisor (without salary) of iPS Academia Japan, and other authors have declared that no competing interests exist.

CRISPR/Cas9 platform, we generated complete knockout cell lines by deleting the entire gene body of ESRG.

Our results demonstrate that ESRG is dispensable for the PSC characters such as gene expression, self-renewing capacity, and differentiation potential. In addition, ESRG does not contribute to the reprogramming of differentiated cells to a pluripotent state. Altogether, we concluded that ESRG is an excellent marker of pluripotency but dispensable for the PSC identity.

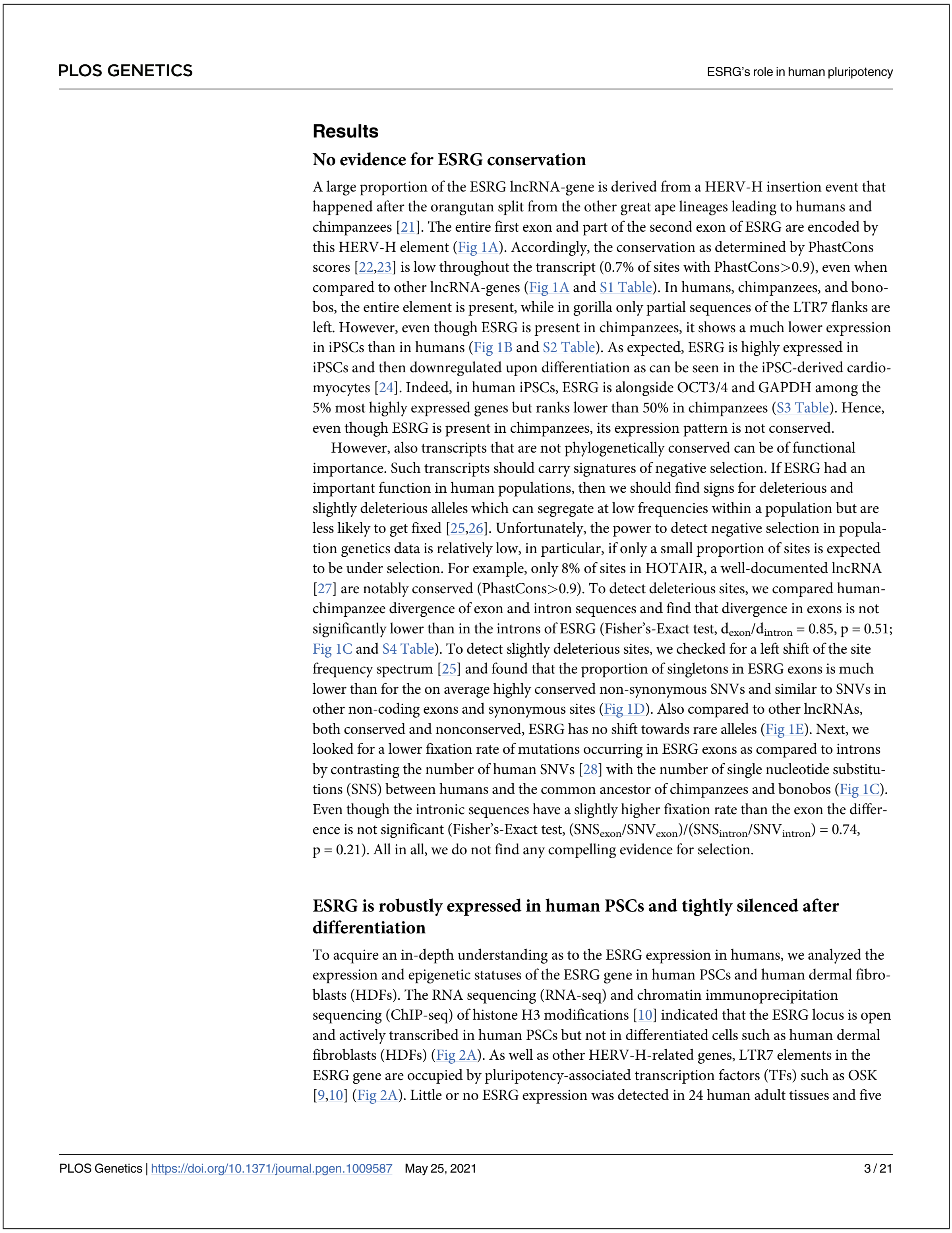
Introduction

Human pluripotent stem cells (PSCs) express several types of human endogenous retroviruses (HERV) [1–3]. The HERV type-H (HERV-H) family is a primate-specific ERV element that was first integrated prior to the New World/Old World divergence. During further primate evolution, this family's major expansion occurred after the branch of Old World monkeys [4]. The typical structure of a HERV-H consists of an interior component, HERV-H-int, flanked by two long terminal repeat 7 (LTR7), which have promoter activity [5,6]. Recent studies have demonstrated that the activity of LTR7 is highly specific in established human PSCs and relatively absent in early human embryos. In contrast, other LTR7 variants such as LTR7B, C, and Y are activated in broad types of early human embryos from the 8-cell to epiblast stages [7].

The importance of HERV-Hs in human PSCs has been shown. The knockdown (KD) of pan HERV-Hs using short hairpin RNAs (shRNAs) against conserved sequences in LTR7 or HERV-H-int regions revealed that HERV-H expression is required for the self-renewal of human PSCs [8,9] and somatic cell reprogramming toward pluripotency [8–14]. In addition to self-renewal, the precise expression of HERV-Hs is crucial for the neural differentiation potential of human PSCs [10,15]. In this way, HERV-H expression contributes to the PSC identity.

The transcription of HERV-H frequently produces a chimeric transcript fused with a downstream neighbor gene, which diversifies HERV-H-driven transcripts. Therefore, many HERV-H-driven RNAs contain unique sequences aside from HERV-H consensus sequences. Indeed, PSC-associated HERV-H-containing long non-coding RNAs (lncRNAs) have been reported [15–17]. One of them, ESRG (embryonic stem cell-related gene; also known as HESRG) was identified as a transcript that is predominantly expressed in undifferentiated human embryonic stem cells (ESCs) [18,19]. ESRG is transcribed from a HERV-H LTR7 promoter [8,20] and is activated in an early stage of somatic cell reprogramming induced by the forced expression of OCT3/4, SOX2, and KLF4 (OSK) [12,13,20]. One previous study showed that the shRNA-mediated KD of ESRG induces the loss of PSC characters such as colony morphology and PSC markers along with the activation of differentiation markers, suggesting the indispensability of ESRG for human pluripotency [8]. However, despite these characterizations, the function of ESRG is still unknown.

In this study, we analyzed the conservation of ESRG to infer its functional importance. Then we completely deleted ESRG alleles to analyze ESRG function in human PSCs with no off-target risk. The loss of ESRG, which is thought to be an essential lncRNA for the PSC identity [8], exhibited no impact on the self-renewal or differentiation potentials of both primed and naïve human PSCs. Neural progenitor cells (NPCs) derived from ESRG-deficient PSCs could be reprogrammed into induced PSC (iPSC) by OSK expression. Altogether, this study revealed that ESRG is dispensable for human pluripotency.



Results

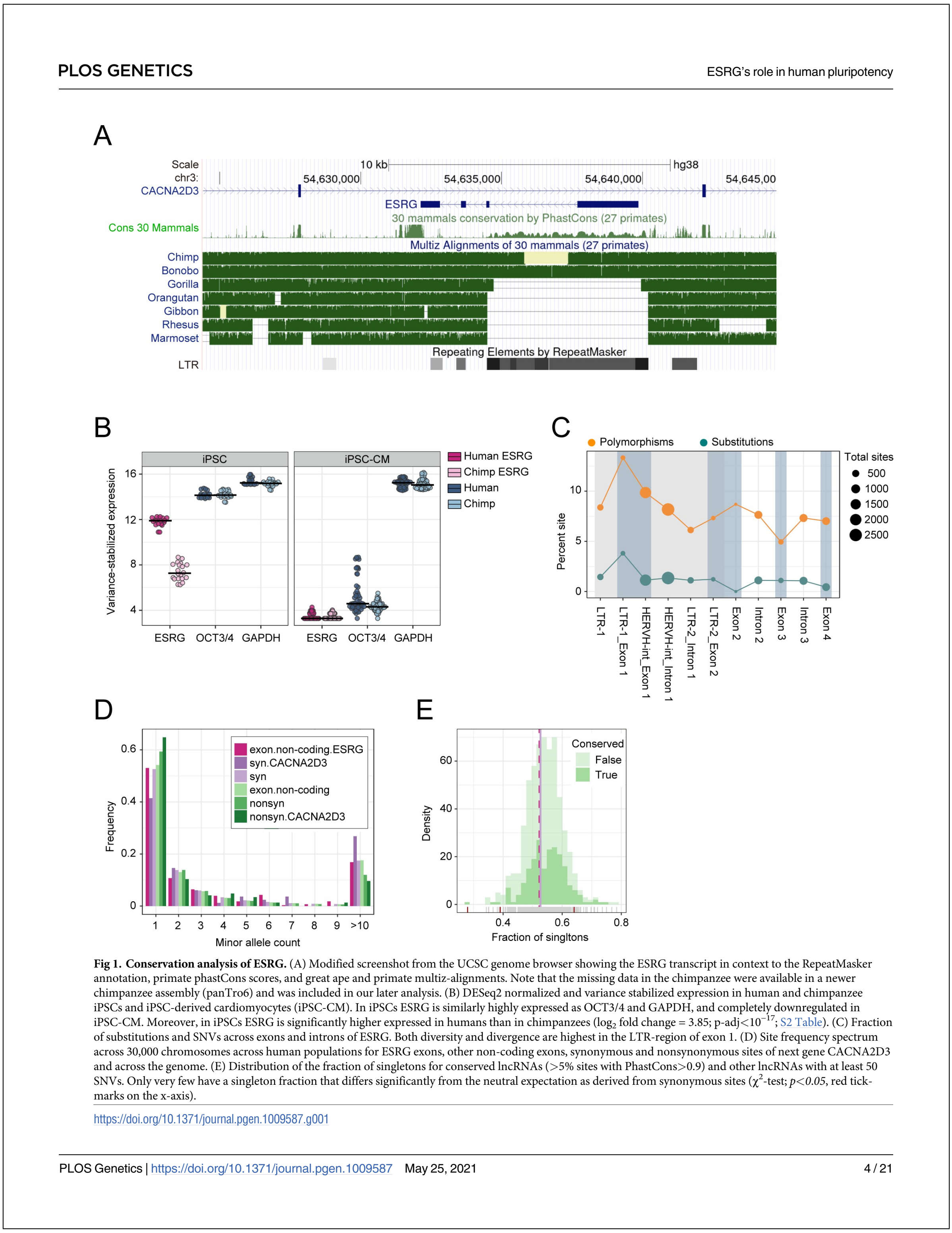
No evidence for ESRG conservation

A large proportion of the ESRG lncRNA-gene is derived from a HERV-H insertion event that happened after the orangutan split from the other great ape lineages leading to humans and chimpanzees [21]. The entire first exon and part of the second exon of ESRG are encoded by this HERV-H element (Fig 1A). Accordingly, the conservation as determined by PhastCons scores [22,23] is low throughout the transcript (0.7% of sites with PhastCons>0.9), even when compared to other lncRNA-genes (Fig 1A and S1 Table). In humans, chimpanzees, and bonobos, the entire element is present, while in gorilla only partial sequences of the LTR7 flanks are left. However, even though ESRG is present in chimpanzees, it shows a much lower expression in iPSCs than in humans (Fig 1B and S2 Table). As expected, ESRG is highly expressed in iPSCs and then downregulated upon differentiation as can be seen in the iPSC-derived cardiomyocytes [24]. Indeed, in human iPSCs, ESRG is alongside OCT3/4 and GAPDH among the 5% most highly expressed genes but ranks lower than 50% in chimpanzees (S3 Table). Hence, even though ESRG is present in chimpanzees, its expression pattern is not conserved.

However, also transcripts that are not phylogenetically conserved can be of functional importance. Such transcripts should carry signatures of negative selection. If ESRG had an important function in human populations, then we should find signs for deleterious and slightly deleterious alleles which can segregate at low frequencies within a population but are less likely to get fixed [25,26]. Unfortunately, the power to detect negative selection in population genetics data is relatively low, in particular, if only a small proportion of sites is expected to be under selection. For example, only 8% of sites in HOTAIR, a well-documented lncRNA [27] are notably conserved (PhastCons>0.9). To detect deleterious sites, we compared human-chimpanzee divergence of exon and intron sequences and find that divergence in exons is not significantly lower than in the introns of ESRG (Fisher's-Exact test, $d_{exon}/d_{intron} = 0.85$, $p = 0.51$; Fig 1C and S4 Table). To detect slightly deleterious sites, we checked for a left shift of the site frequency spectrum [25] and found that the proportion of singletons in ESRG exons is much lower than for the on average highly conserved non-synonymous SNVs and similar to SNVs in other non-coding exons and synonymous sites (Fig 1D). Also compared to other lncRNAs, both conserved and nonconserved, ESRG has no shift towards rare alleles (Fig 1E). Next, we looked for a lower fixation rate of mutations occurring in ESRG exons as compared to introns by contrasting the number of human SNVs [28] with the number of single nucleotide substitutions (SNS) between humans and the common ancestor of chimpanzees and bonobos (Fig 1C). Even though the intronic sequences have a slightly higher fixation rate than the exon the difference is not significant (Fisher's-Exact test, $(SNS_{exon}/SNV_{exon})/(SNS_{intron}/SNV_{intron}) = 0.74$, $p = 0.21$). All in all, we do not find any compelling evidence for selection.

ESRG is robustly expressed in human PSCs and tightly silenced after differentiation

To acquire an in-depth understanding as to the ESRG expression in humans, we analyzed the expression and epigenetic statuses of the ESRG gene in human PSCs and human dermal fibroblasts (HDFs). The RNA sequencing (RNA-seq) and chromatin immunoprecipitation sequencing (ChIP-seq) of histone H3 modifications [10] indicated that the ESRG locus is open and actively transcribed in human PSCs but not in differentiated cells such as human dermal fibroblasts (HDFs) (Fig 2A). As well as other HERV-H-related genes, LTR7 elements in the ESRG gene are occupied by pluripotency-associated transcription factors (TFs) such as OSK [9,10] (Fig 2A). Little or no ESRG expression was detected in 24 human adult tissues and five



PLOS GENETICS

ESRG's role in human pluripotency

Fig 1. Conservation analysis of ESRG. (A) Modified screenshot from the UCSC genome browser showing the ESRG transcript in context to the RepeatMasker annotation, primate phastCons scores, and great ape and primate multiz-alignments. Note that the missing data in the chimpanzee were available in a newer chimpanzee assembly (panTro6) and was included in our later analysis. (B) DESeq2 normalized and variance stabilized expression in human and chimpanzee iPSCs and iPSC-derived cardiomyocytes (iPSC-CM). In iPSCs ESRG is similarly highly expressed as OCT3/4 and GAPDH, and completely downregulated in iPSC-CM. Moreover, in iPSCs ESRG is significantly higher expressed in humans than in chimpanzees ($\log_2$ fold change = 3.85; p-adj$<10^{-17}$; S2 Table). (C) Fraction of substitutions and SNVs across exons and introns of ESRG. Both diversity and divergence are highest in the LTR-region of exon 1. (D) Site frequency spectrum across 30,000 chromosomes across human populations for ESRG exons, other non-coding exons, synonymous and nonsynonymous sites of next gene CACNA2D3 and across the genome. (E) Distribution of the fraction of singletons for conserved lncRNAs (>5% sites with PhastCons>0.9) and other lncRNAs with at least 50 SNVs. Only very few have a singleton fraction that differs significantly from the neutral expectation as derived from synonymous sites (χ^2-test; $p<0.05$, red tick-marks on the x-axis).

https://doi.org/10.1371/journal.pgen.1009587.g001

PLOS Genetics | https://doi.org/10.1371/journal.pgen.1009587 May 25, 2021 4 / 21

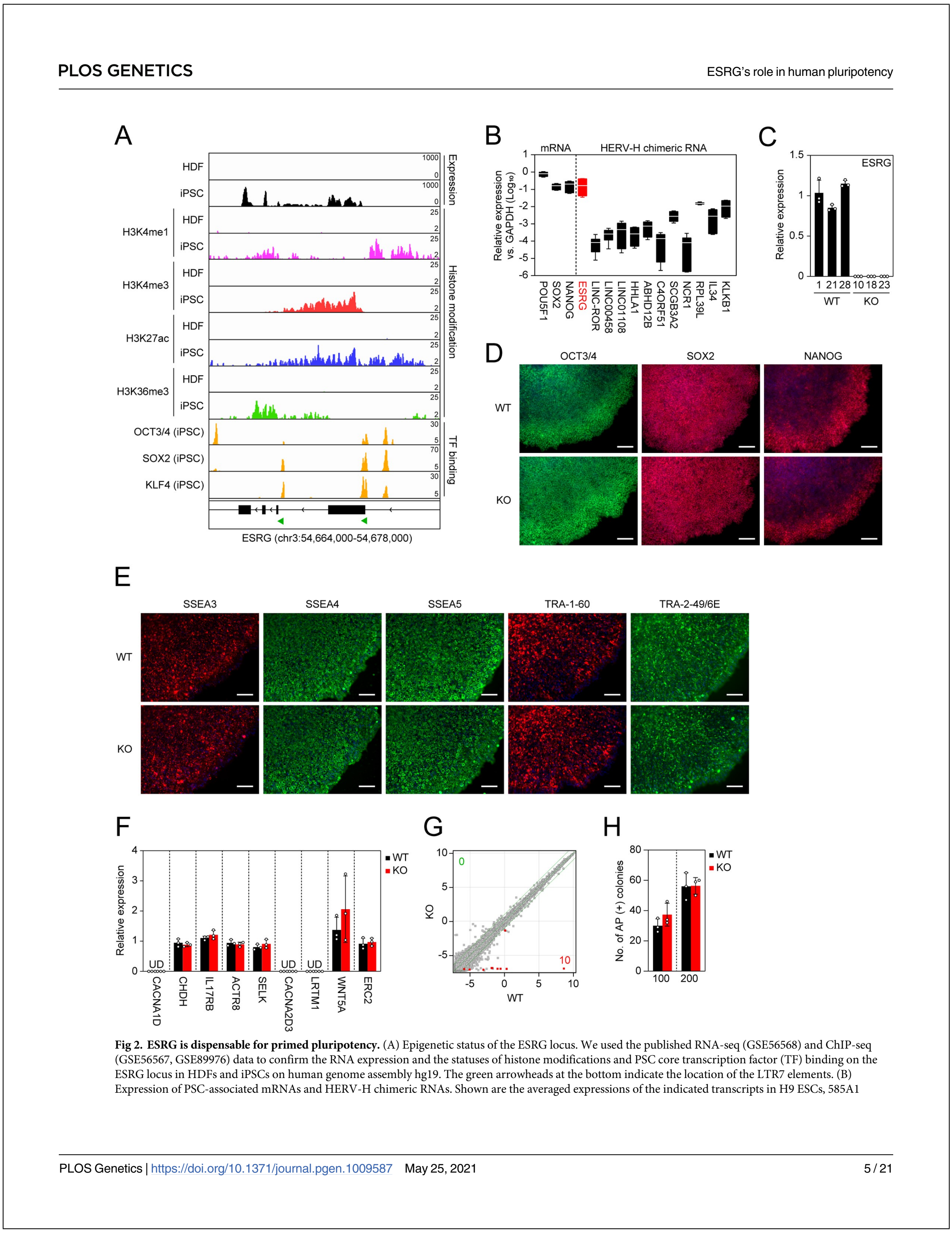
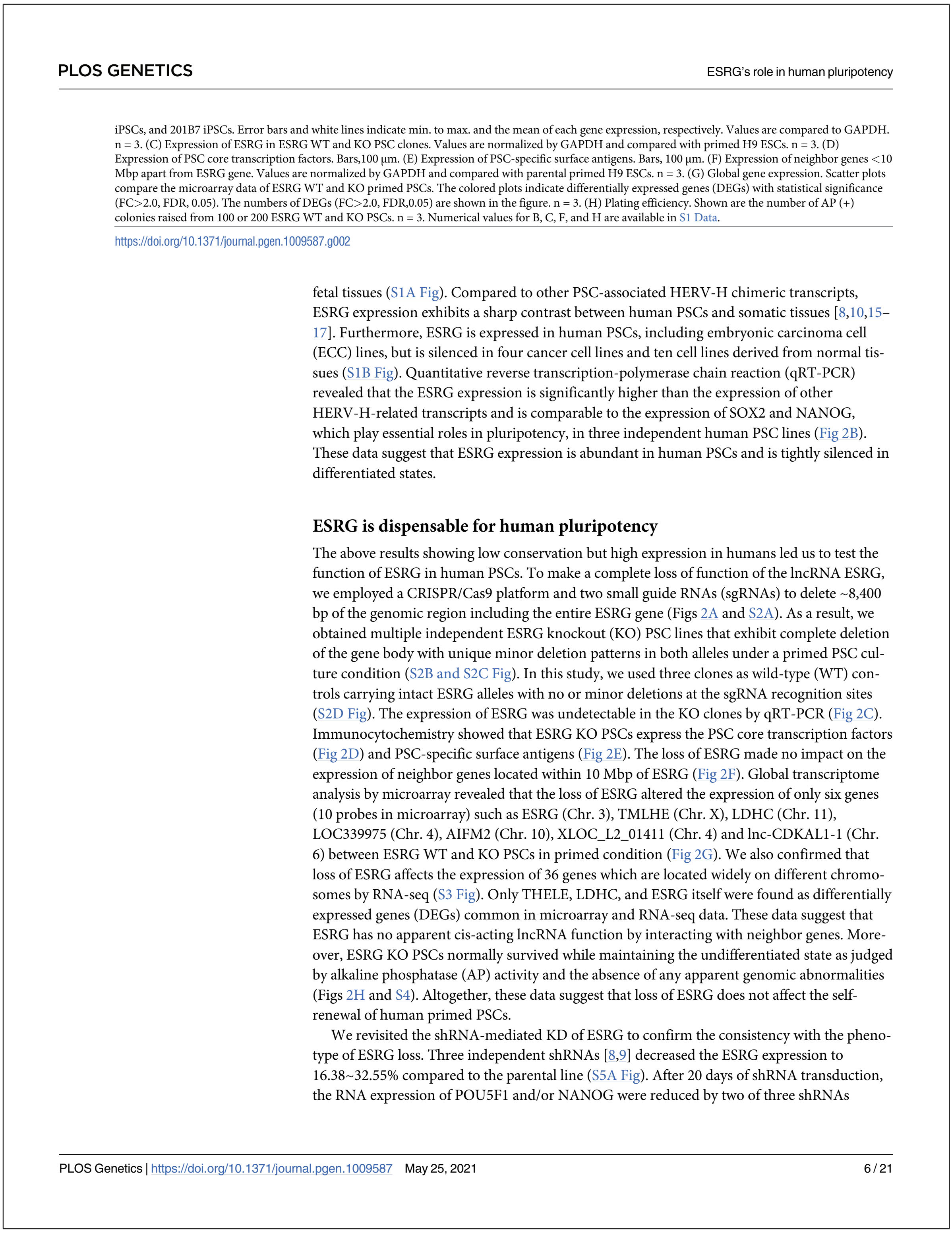


Fig 2. ESRG is dispensable for primed pluripotency. (A) Epigenetic status of the ESRG locus. We used the published RNA-seq (GSE56568) and ChIP-seq (GSE56567, GSE89976) data to confirm the RNA expression and the statuses of histone modifications and PSC core transcription factor (TF) binding on the ESRG locus in HDFs and iPSCs on human genome assembly hg19. The green arrowheads at the bottom indicate the location of the LTR7 elements. (B) Expression of PSC-associated mRNAs and HERV-H chimeric RNAs. Shown are the averaged expressions of the indicated transcripts in H9 ESCs, 585A1



iPSCs, and 201B7 iPSCs. Error bars and white lines indicate min. to max. and the mean of each gene expression, respectively. Values are compared to GAPDH. n = 3. (C) Expression of ESRG in ESRG WT and KO PSC clones. Values are normalized by GAPDH and compared with primed H9 ESCs. n = 3. (D) Expression of PSC core transcription factors. Bars,100 μm. (E) Expression of PSC-specific surface antigens. Bars, 100 μm. (F) Expression of neighbor genes <10 Mbp apart from ESRG gene. Values are normalized by GAPDH and compared with parental primed H9 ESCs. n = 3. (G) Global gene expression. Scatter plots compare the microarray data of ESRG WT and KO primed PSCs. The colored plots indicate differentially expressed genes (DEGs) with statistical significance (FC>2.0, FDR, 0.05). The numbers of DEGs (FC>2.0, FDR,0.05) are shown in the figure. n = 3. (H) Plating efficiency. Shown are the number of AP (+) colonies raised from 100 or 200 ESRG WT and KO PSCs. n = 3. Numerical values for B, C, F, and H are available in S1 Data.

https://doi.org/10.1371/journal.pgen.1009587.g002

fetal tissues (S1A Fig). Compared to other PSC-associated HERV-H chimeric transcripts, ESRG expression exhibits a sharp contrast between human PSCs and somatic tissues [8,10,15–17]. Furthermore, ESRG is expressed in human PSCs, including embryonic carcinoma cell (ECC) lines, but is silenced in four cancer cell lines and ten cell lines derived from normal tissues (S1B Fig). Quantitative reverse transcription-polymerase chain reaction (qRT-PCR) revealed that the ESRG expression is significantly higher than the expression of other HERV-H-related transcripts and is comparable to the expression of SOX2 and NANOG, which play essential roles in pluripotency, in three independent human PSC lines (Fig 2B). These data suggest that ESRG expression is abundant in human PSCs and is tightly silenced in differentiated states.

ESRG is dispensable for human pluripotency

The above results showing low conservation but high expression in humans led us to test the function of ESRG in human PSCs. To make a complete loss of function of the lncRNA ESRG, we employed a CRISPR/Cas9 platform and two small guide RNAs (sgRNAs) to delete ~8,400 bp of the genomic region including the entire ESRG gene (Figs 2A and S2A). As a result, we obtained multiple independent ESRG knockout (KO) PSC lines that exhibit complete deletion of the gene body with unique minor deletion patterns in both alleles under a primed PSC culture condition (S2B and S2C Fig). In this study, we used three clones as wild-type (WT) controls carrying intact ESRG alleles with no or minor deletions at the sgRNA recognition sites (S2D Fig). The expression of ESRG was undetectable in the KO clones by qRT-PCR (Fig 2C). Immunocytochemistry showed that ESRG KO PSCs express the PSC core transcription factors (Fig 2D) and PSC-specific surface antigens (Fig 2E). The loss of ESRG made no impact on the expression of neighbor genes located within 10 Mbp of ESRG (Fig 2F). Global transcriptome analysis by microarray revealed that the loss of ESRG altered the expression of only six genes (10 probes in microarray) such as ESRG (Chr. 3), TMLHE (Chr. X), LDHC (Chr. 11), LOC339975 (Chr. 4), AIFM2 (Chr. 10), XLOC_L2_01411 (Chr. 4) and lnc-CDKAL1-1 (Chr. 6) between ESRG WT and KO PSCs in primed condition (Fig 2G). We also confirmed that loss of ESRG affects the expression of 36 genes which are located widely on different chromosomes by RNA-seq (S3 Fig). Only THELE, LDHC, and ESRG itself were found as differentially expressed genes (DEGs) common in microarray and RNA-seq data. These data suggest that ESRG has no apparent cis-acting lncRNA function by interacting with neighbor genes. Moreover, ESRG KO PSCs normally survived while maintaining the undifferentiated state as judged by alkaline phosphatase (AP) activity and the absence of any apparent genomic abnormalities (Figs 2H and S4). Altogether, these data suggest that loss of ESRG does not affect the self-renewal of human primed PSCs.

We revisited the shRNA-mediated KD of ESRG to confirm the consistency with the phenotype of ESRG loss. Three independent shRNAs [8,9] decreased the ESRG expression to 16.38~32.55% compared to the parental line (S5A Fig). After 20 days of shRNA transduction, the RNA expression of POU5F1 and/or NANOG were reduced by two of three shRNAs (shESRG-4 and 5), although the most effective shRNA (shESRG-2) against ESRG did not alter them (S5A Fig). None of ESRG shRNAs induced the expression of early differentiation markers such as T (mesendoderm) and NES (neuroectoderm) (S5A Fig). The ESRG KD PSCs grew normally with expressing NANOG protein (S5B Fig). These data suggest that ESRG KD by shRNAs does not induce the differentiation of human PSCs in the primed state. We and others previously reported the effects of shRNA-mediated pan HERV-H KD on human PSC characteristics [8–10]. Three shRNAs against the conserved regions of HERV-Hs decreased to 29.06~56.48% compared to the parental line (S6A Fig). One of them (shHERVH-1), as similar efficiency of the ESRG shRNAs, finely knocked down the ESRG expression to 14.55% of the parental line (S5B and S6B Figs). Microarray data suggested that no noticeable changes were detected in the expression of PSC markers and lineage markers (S6B Fig). In addition to the transcriptome data, we confirmed that all three HERV-H KD PSC lines were able to expand with maintaining the stem cell morphologies and NANOG protein expression (S6C Fig). These data support that ESRG is dispensable for the self-renewing of primed PSCs.

In addition to the primed state, we tested if ESRG is required for another state of pluripotency, the so-called naïve state, which also expresses ESRG but at a significantly lower level than the primed state (Fig 3A). Regardless of the ESRG expression, naïve PSCs could be established by switching the media composition and could self-renew while keeping a tightly packed colony formation (Fig 3B) [29–31]. Furthermore, they exhibited a significantly high expression of the naïve pluripotency markers KLF4 and KLF17 and attenuated the expression of the primed PSC marker ZIC2 (Fig 3C) [32,33]. Twenty-nine genes including ESRG and CNCNA2D3 were found as DEGs between ESRG WT and KO PSCs in naïve condition by RNA-seq (S3 Fig), although microarray analysis revealed that ESRG had no effect on the global gene expression of naïve PSCs (Fig 3D). Altogether, these data suggest that ESRG does not contribute to self-renewal and gene expression of human naïve PSCs.

We also differentiated ESRG WT and KO naïve PSCs to the primed pluripotent state. As a result, irrespective of the ESRG genotype, we detected the hallmarks of primed pluripotency such as flatter colony formation, the reactivation of ZIC2 and the suppression of KLF4 and KLF17, suggesting the bidirectional transition between naïve and primed pluripotency does not require ESRG (Fig 3E and 3F). Taken together, these data demonstrate that ESRG is dispensable for the maintenance of human PSCs.

ESRG is not involved in differentiation

Next, we analyzed whether ESRG is required for the differentiation of human primed PSCs by embryoid body (EB) formation. The absence of ESRG had no effect on EB formation by floating culture or differentiation into trilineage such as alpha-fetoprotein (AFP) positive (+) endoderm, smooth muscle actin (SMA) (+) mesoderm, and βIII-TUBULIN (+) ectoderm (Fig 4A and 4B). Other lineage markers such as DCN (endoderm), MSX1 (mesoderm) and MAP2 (ectoderm) were also well induced in EBs derived from either ESRG WT or KO primed PSCs (Fig 4C). Global transcriptome analysis by microarray indicated the loss of ESRG caused no significant gene expression changes during EB differentiation (Fig 4D). These data suggest that ESRG KO PSCs retained the potential to differentiate into all three germ layers.

Previous studies showed that HERV-H expression regulates the neural differentiation potential of human PSCs [10,15,34]. Thus, in addition to the random differentiation by EB formation, we tested whether ESRG contributes to the directed differentiation of human primed PSCs into NPCs by the dual SMAD inhibition method [35,36]. Both ESRG WT and KO PSCs were able to differentiate into expandable NPCs, which expressed the early neural lineage marker PAX6 but not OCT3/4 (Fig 4E). Other NPC markers such as SOX1 and NES were well

PLOS GENETICS

ESRG's role in human pluripotency

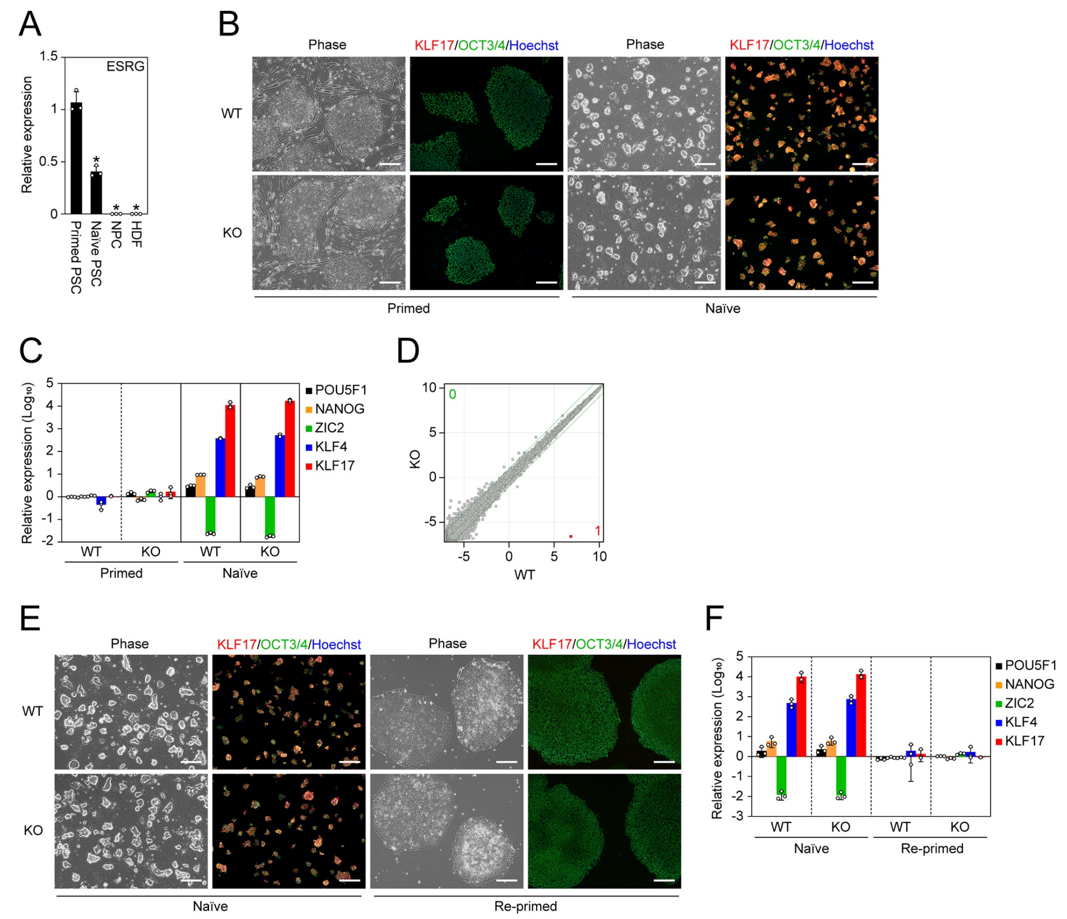


Fig 3. No impact of ESRG on naïve pluripotency. (A) The ESRG expression. Shown are relative expressions of ESRG in primed PSCs, naïve PSCs, NPCs and HDFs. Values are normalized by GAPDH and compared with the primed 585A1 iPSC line. *P<0.05 vs. primed PSCs by unpaired t-test. n = 3. (B) Conversion to naïve pluripotency. Shown are representative images of ESRG WT and KO primed and naïve PSCs under phase contrast and of immunocytochemistry for KLF17 (red) and OCT3/4 (green). Bars, 200 μm. (C) The expression of primed and naïve PSC markers. Shown are the relative expressions of common PSC markers (POU5F1 and NANOG), a primed PSC marker (ZIC2) and naïve PSC markers (KLF4 and KLF17). Values are normalized by GAPDH and compared with primed H9 ESCs. n = 3. (D) Global transcriptome. Scatter plots comparing the microarray data of ESRG WT and KO naïve PSCs. The colored plot indicates DEG with statistical significance (FC>2.0, FDR,0.05). The numbers of DEGs (FC>2.0, FDR,0.05) are shown in the figure. n = 3. (E) Differentiation to primed pluripotency. Representative images of ESRG WT and KO naïve PSCs before and after conversion to the primed pluripotent state are shown. Bars, 200 μm. (F) The expression of primed and naïve PSC markers. Shown are the relative expressions of the marker genes in (C) in ESRG WT and KO naïve PSCs before and after the differentiation to the primed pluripotent state. Values are normalized by GAPDH and compared with primed H9 ESCs. n = 3. Numerical values for A, C, and F are available in S1 Data.

https://doi.org/10.1371/journal.pgen.1009587.g003

induced, whereas the PSC marker NANOG was silenced (Fig 4F). These data suggest that ESRG is not responsible for HERV-H-regulated neural differentiation. Taken together, we concluded that ESRG is not required for the differentiation of human PSCs.

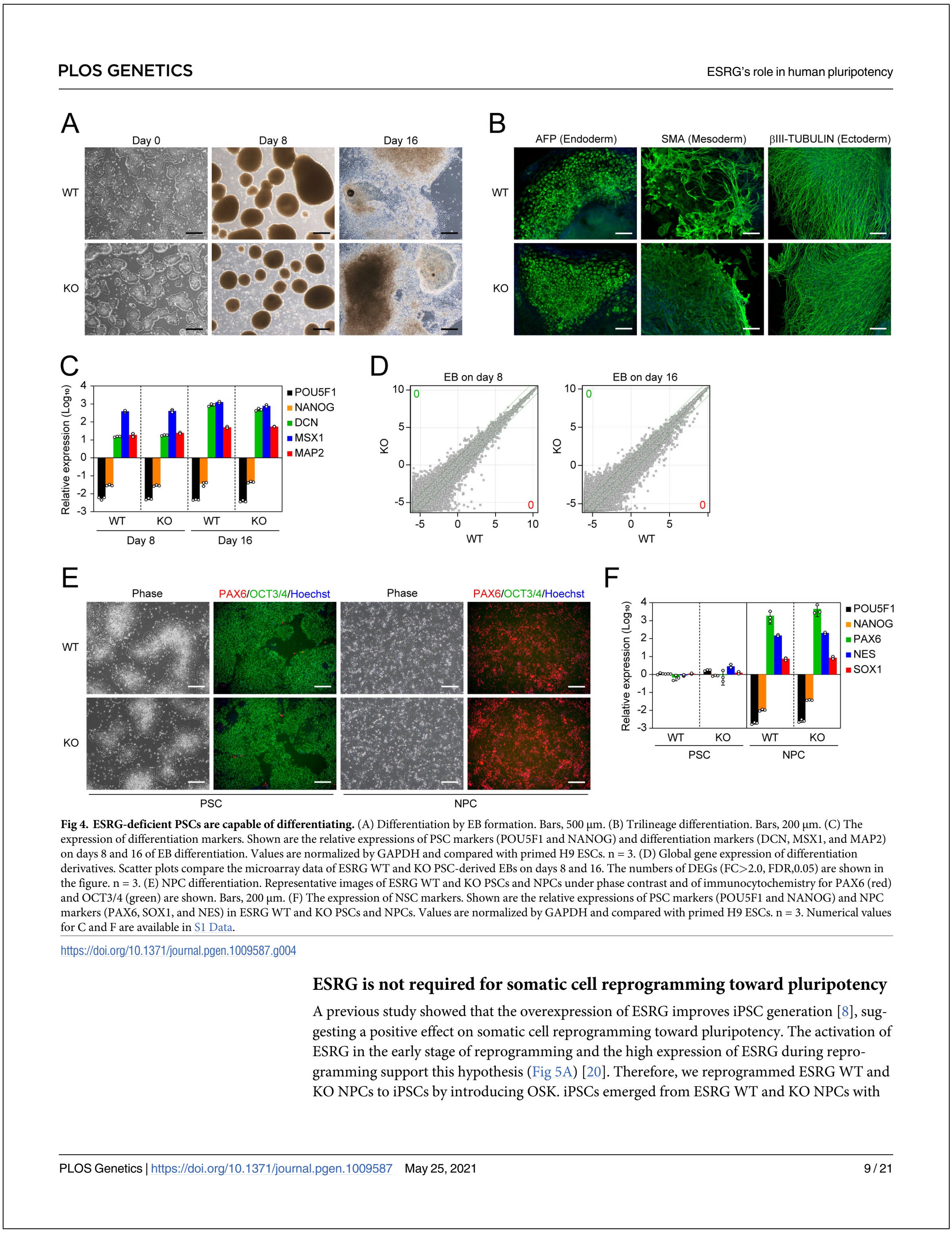


Fig 4. ESRG-deficient PSCs are capable of differentiating. (A) Differentiation by EB formation. Bars, 500 μm. (B) Trilineage differentiation. Bars, 200 μm. (C) The expression of differentiation markers. Shown are the relative expressions of PSC markers (POU5F1 and NANOG) and differentiation markers (DCN, MSX1, and MAP2) on days 8 and 16 of EB differentiation. Values are normalized by GAPDH and compared with primed H9 ESCs. n = 3. (D) Global gene expression of differentiation derivatives. Scatter plots compare the microarray data of ESRG WT and KO PSC-derived EBs on days 8 and 16. The numbers of DEGs (FC>2.0, FDR,0.05) are shown in the figure. n = 3. (E) NPC differentiation. Representative images of ESRG WT and KO PSCs and NPCs under phase contrast and of immunocytochemistry for PAX6 (red) and OCT3/4 (green) are shown. Bars, 200 μm. (F) The expression of NSC markers. Shown are the relative expressions of PSC markers (POU5F1 and NANOG) and NPC markers (PAX6, SOX1, and NES) in ESRG WT and KO PSCs and NPCs. Values are normalized by GAPDH and compared with primed H9 ESCs. n = 3. Numerical values for C and F are available in S1 Data.

https://doi.org/10.1371/journal.pgen.1009587.g004

ESRG is not required for somatic cell reprogramming toward pluripotency

A previous study showed that the overexpression of ESRG improves iPSC generation [8], suggesting a positive effect on somatic cell reprogramming toward pluripotency. The activation of ESRG in the early stage of reprogramming and the high expression of ESRG during reprogramming support this hypothesis (Fig 5A) [20]. Therefore, we reprogrammed ESRG WT and KO NPCs to iPSCs by introducing OSK. iPSCs emerged from ESRG WT and KO NPCs with

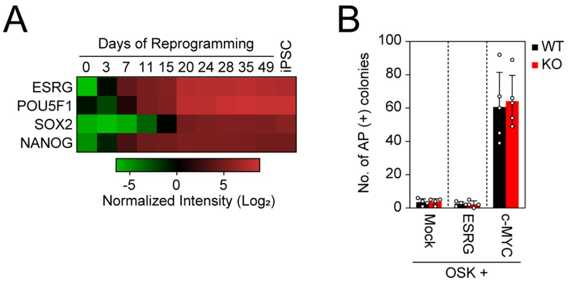


Fig 5. ESRG is dispensable for iPSC reprogramming. (A) The expression of ESRG during reprogramming. The heatmap generated by using the dataset (GSE54848) shows the normalized intensities of ESRG, POU5F1 (endogenous), SOX2 (endogenous), and NANOG expression from microarray data in the time course of iPSC reprogramming (days 0–49) and established iPSCs (far right). n = 3. (B) The effect of ESRG on iPSC generation. Shown are the numbers of AP (+) iPSC colonies 24 days after the transduction of OSK along with Mock (n = 4), ESRG (n = 4), and c-MYC (n = 5). Numerical values for A and B are available in S1 Data.

https://doi.org/10.1371/journal.pgen.1009587.g005

comparable efficiency (Fig 5B). This observation suggests that ESRG is dispensable for iPSC generation. In addition, along with OSK, we transduced c-MYC, a potent enhancer of iPSC generation [37,38], or exogenous ESRG. c-MYC but not exogenous ESRG increased the efficiency of the iPSC generation from ESRG WT and KO NPCs equally (Fig 5B). Taken together, these data suggest that ESRG has no impact on somatic cell reprogramming toward iPSCs.

Discussion

In this study, we completely excised the entire ESRG gene to understand its role in human PSCs while avoiding residual expression and off-target effects. As a result, ESRG KO PSCs showed no apparent phenotypes in self-renewal and differentiation potential. A previous study showed the importance of ESRG in human PSC identity by using an shRNA-mediated KD approach [8]. Although we used the same H9 ESC line as that study, the different strategies for the loss of function and subsequent experiments, such as KD and KO, may explain the different results. Therefore, this study revisited the ESRG KD by using three shRNAs including published sequences [8]. Indeed, two published shRNAs (shESRG-4 and 5) decreased POU5F1 (84.28 and 55.28% of the parental line) and NANOG (52.66 and 67.14% of the parental line), respectively, whereas shESRG-2 that is newly designed in this study did not change their expression (103.54 (POU5F1) and 106.64% (NANOG) of the parental line) (S5A Fig). The reduction of PSC marker expression that varied among shRNAs was not enough to induce the differentiation of human PSCs (S5C Fig). In addition to the ESRG KD, we also showed the effects of pan HERV-H KD in human PSCs in primed condition (S6 Fig). We previously showed that the suppression of HERV-H expression using shRNA did not disrupt the self-renewal of human PSCs [10,34]. A recent paper by Zhang et al. showed that pan-HERV-H KD in human PSCs by using CRISPR interference did not induce spontaneous differentiation like we observed [39]. However, since other groups concluded that HERV-H KD induced differentiation [8,9], further studies are required to understand what HERV-H is doing. One possibility that may explain the discrepancy of the results between previous and current studies [8] is the off-target effect of RNAi. Similar observations have been found for the role of lncRNA Cyrano that is highly conserved in mice and humans. Knockdown by using shRNA suggested Cyrano lncRNA maintains mouse PSC identity [40], but targeted deletion of the Cyrano gene and gene silencing by CRISPR interference demonstrated no impact on

the mouse or human PSC identity [41–43]. Further, it has been argued that the shRNA-mediated KD of nuclear lncRNAs might be difficult or inefficient compared to cytoplasmic RNAs such as mRNAs [44,45]. In addition, while small nucleotide insertions or deletions causing frameshift of the reading frames work well for the loss of function of protein-coding genes, the same is not true for non-coding RNAs. In this context, our study succeeded in generating the complete deletion of ESRG gene alleles, providing highly reliable results.

This study clearly demonstrated that ESRG is dispensable for human PSC identity. Neither primed nor naïve PSCs require ESRG for their identities, such as colony morphology or gene expression signatures, meaning ESRG is dispensable for human pluripotency, at least in an in vitro culture environment. However, since ESRG is expressed in epiblast-stage human embryos [8,46], it might be involved in early human embryogenesis.

ESRG is stochastically activated by OSK in rare reprogrammed intermediates that have the potential to become bona fide iPSCs and is highly expressed throughout the process of reprogramming toward iPSCs [20]. In the present study, we showed that ESRG KO NPCs can be reprogrammed with the same efficiency as ESRG WT NPCs. These data suggest that ESRG is a good marker of the intermediate cells in the early stage of reprogramming rather than a functional molecule that is needed for iPSC generation.

In summary, this study provides clear evidence of the dispensability of ESRG for human PSC identities, such as global gene expressions and differentiation potentials, in two distinct types of pluripotent states. We also demonstrated that the function of ESRG is not required for recapturing pluripotency via somatic cell reprogramming. Finally, the tightly regulated and high expression of ESRG promises to make an excellent marker of undifferentiated human PSCs both in basic research and clinical application [20,47].

Methods

Expression conservation

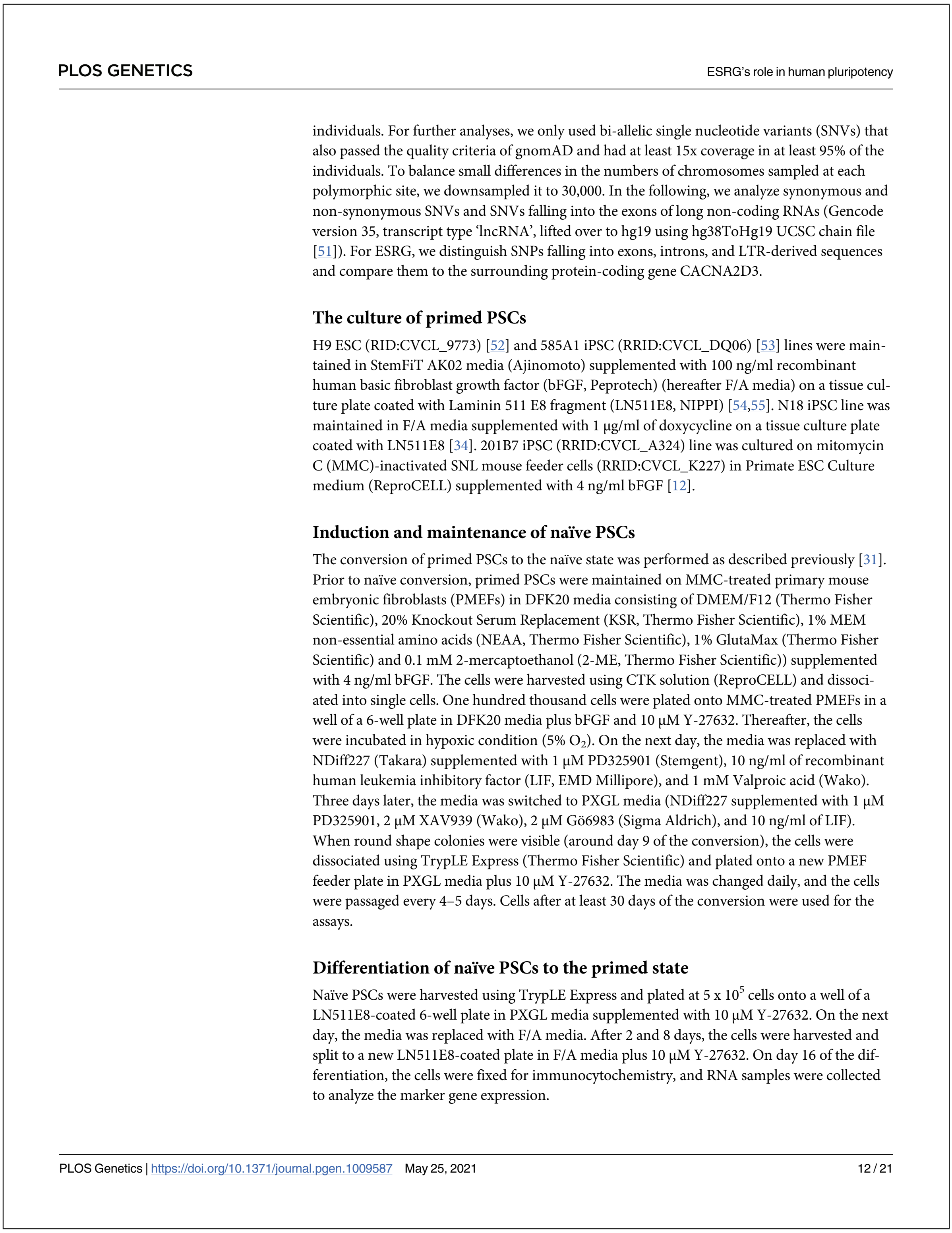
To investigate ESRG expression, we used an RNA-seq data set that investigated cardiomyocyte differentiation from human and chimpanzee iPSCs [24]. Read count matrices were downloaded from Gene Expression Omnibus (GSE110471). We selected iPSC and iPSC-derived cardiomyocyte samples and filtered the data for genes that were detected in at least 40% of the samples and had an average expression of at least 5 counts, yielding a final matrix with 17,213 genes. Differential expression analyses and variance-stabilizing transformation were performed using DESeq2 v.1.30.0 [48], using a model including the factors ~cell type: species + species. iPSC-specific differential expression between human and chimpanzee was inferred via the interaction term identifying iPSC-specific differences between human and chimpanzee.

Multiple sequence alignment

We used the human ESRG sequence (+20 kb in each direction) (NCBI 105.20190906 Reference Sequence NR_027122.1; hg19) to search orthologous sequence in the great apes genomes: chimpanzee (Pan troglodytes, GCF_002880755.1), bonobo (Pan paniscus, GCF_013052645.1), gorilla (Gorilla gorilla, GCA_900006655.3) and orang (Pongo abelii, GCF_002880775.1) using dc-megablast with default options [49]. Finally, the identified regions were aligned into a multiple sequence alignment using mafft [50] and manual inspection.

Human polymorphism data

We identified the polymorphic sites based on gnomAD v2.1.1 database [28]. We downloaded the vcf-file and tsv coverage files derived from whole-genome sequencing of 15,708 unrelated



individuals. For further analyses, we only used bi-allelic single nucleotide variants (SNVs) that also passed the quality criteria of gnomAD and had at least 15x coverage in at least 95% of the individuals. To balance small differences in the numbers of chromosomes sampled at each polymorphic site, we downsampled it to 30,000. In the following, we analyze synonymous and non-synonymous SNVs and SNVs falling into the exons of long non-coding RNAs (Gencode version 35, transcript type 'lncRNA', lifted over to hg19 using hg38ToHg19 UCSC chain file [51]). For ESRG, we distinguish SNPs falling into exons, introns, and LTR-derived sequences and compare them to the surrounding protein-coding gene CACNA2D3.

The culture of primed PSCs

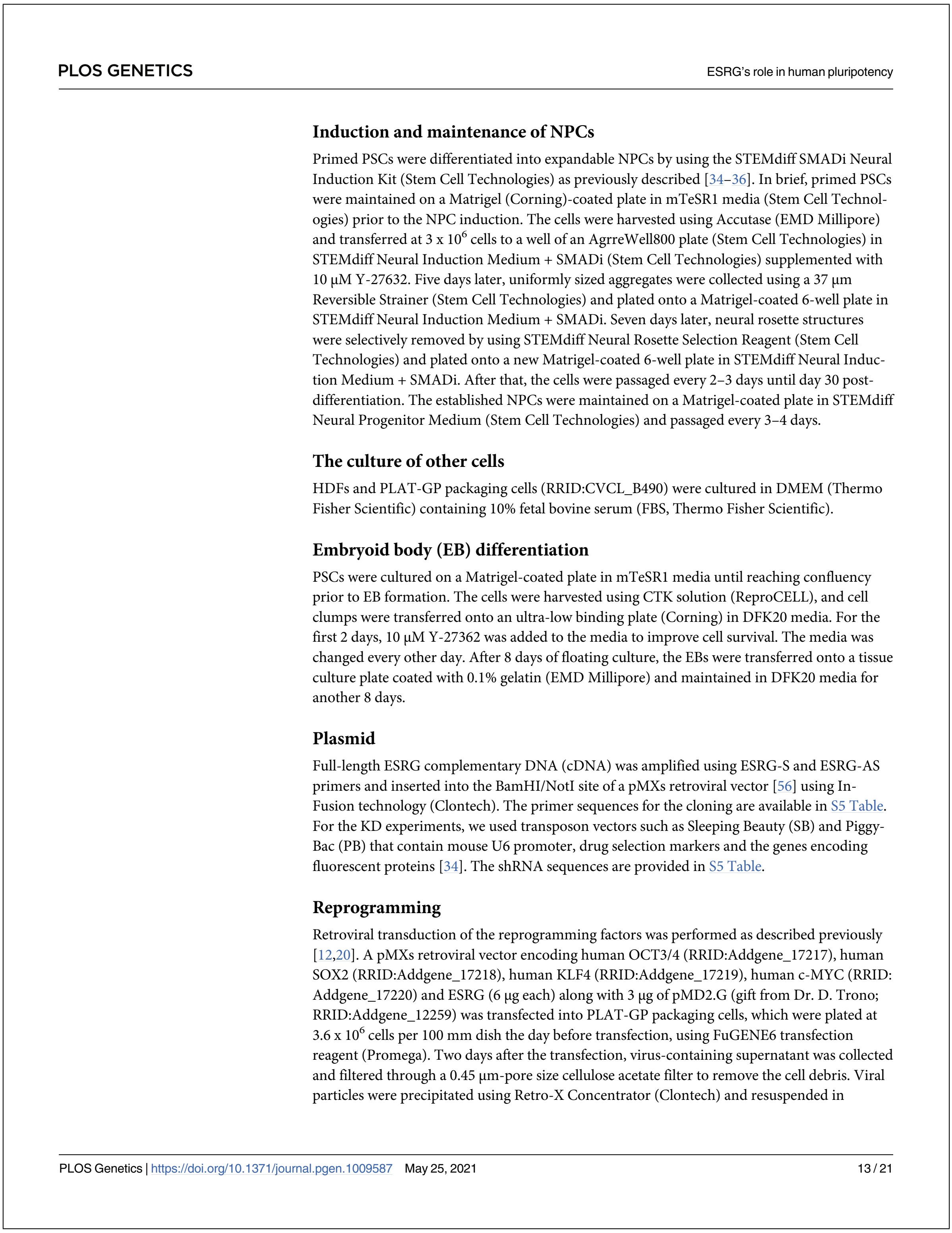
H9 ESC (RID:CVCL_9773) [52] and 585A1 iPSC (RRID:CVCL_DQ06) [53] lines were maintained in StemFiT AK02 media (Ajinomoto) supplemented with 100 ng/ml recombinant human basic fibroblast growth factor (bFGF, Peprotech) (hereafter F/A media) on a tissue culture plate coated with Laminin 511 E8 fragment (LN511E8, NIPPI) [54,55]. N18 iPSC line was maintained in F/A media supplemented with 1 μg/ml of doxycycline on a tissue culture plate coated with LN511E8 [34]. 201B7 iPSC (RRID:CVCL_A324) line was cultured on mitomycin C (MMC)-inactivated SNL mouse feeder cells (RRID:CVCL_K227) in Primate ESC Culture medium (ReproCELL) supplemented with 4 ng/ml bFGF [12].

Induction and maintenance of naïve PSCs

The conversion of primed PSCs to the naïve state was performed as described previously [31]. Prior to naïve conversion, primed PSCs were maintained on MMC-treated primary mouse embryonic fibroblasts (PMEFs) in DFK20 media consisting of DMEM/F12 (Thermo Fisher Scientific), 20% Knockout Serum Replacement (KSR, Thermo Fisher Scientific), 1% MEM non-essential amino acids (NEAA, Thermo Fisher Scientific), 1% GlutaMax (Thermo Fisher Scientific) and 0.1 mM 2-mercaptoethanol (2-ME, Thermo Fisher Scientific)) supplemented with 4 ng/ml bFGF. The cells were harvested using CTK solution (ReproCELL) and dissociated into single cells. One hundred thousand cells were plated onto MMC-treated PMEFs in a well of a 6-well plate in DFK20 media plus bFGF and 10 μM Y-27632. Thereafter, the cells were incubated in hypoxic condition (5% O_2). On the next day, the media was replaced with NDiff227 (Takara) supplemented with 1 μM PD325901 (Stemgent), 10 ng/ml of recombinant human leukemia inhibitory factor (LIF, EMD Millipore), and 1 mM Valproic acid (Wako). Three days later, the media was switched to PXGL media (NDiff227 supplemented with 1 μM PD325901, 2 μM XAV939 (Wako), 2 μM Gö6983 (Sigma Aldrich), and 10 ng/ml of LIF). When round shape colonies were visible (around day 9 of the conversion), the cells were dissociated using TrypLE Express (Thermo Fisher Scientific) and plated onto a new PMEF feeder plate in PXGL media plus 10 μM Y-27632. The media was changed daily, and the cells were passaged every 4–5 days. Cells after at least 30 days of the conversion were used for the assays.

Differentiation of naïve PSCs to the primed state

Naïve PSCs were harvested using TrypLE Express and plated at 5 x 10^5 cells onto a well of a LN511E8-coated 6-well plate in PXGL media supplemented with 10 μM Y-27632. On the next day, the media was replaced with F/A media. After 2 and 8 days, the cells were harvested and split to a new LN511E8-coated plate in F/A media plus 10 μM Y-27632. On day 16 of the differentiation, the cells were fixed for immunocytochemistry, and RNA samples were collected to analyze the marker gene expression.



Induction and maintenance of NPCs

Primed PSCs were differentiated into expandable NPCs by using the STEMdiff SMADi Neural Induction Kit (Stem Cell Technologies) as previously described [34–36]. In brief, primed PSCs were maintained on a Matrigel (Corning)-coated plate in mTeSR1 media (Stem Cell Technologies) prior to the NPC induction. The cells were harvested using Accutase (EMD Millipore) and transferred at 3 x 10^6 cells to a well of an AgrreWell800 plate (Stem Cell Technologies) in STEMdiff Neural Induction Medium + SMADi (Stem Cell Technologies) supplemented with 10 μM Y-27632. Five days later, uniformly sized aggregates were collected using a 37 μm Reversible Strainer (Stem Cell Technologies) and plated onto a Matrigel-coated 6-well plate in STEMdiff Neural Induction Medium + SMADi. Seven days later, neural rosette structures were selectively removed by using STEMdiff Neural Rosette Selection Reagent (Stem Cell Technologies) and plated onto a new Matrigel-coated 6-well plate in STEMdiff Neural Induction Medium + SMADi. After that, the cells were passaged every 2–3 days until day 30 post-differentiation. The established NPCs were maintained on a Matrigel-coated plate in STEMdiff Neural Progenitor Medium (Stem Cell Technologies) and passaged every 3–4 days.

The culture of other cells

HDFs and PLAT-GP packaging cells (RRID:CVCL_B490) were cultured in DMEM (Thermo Fisher Scientific) containing 10% fetal bovine serum (FBS, Thermo Fisher Scientific).

Embryoid body (EB) differentiation

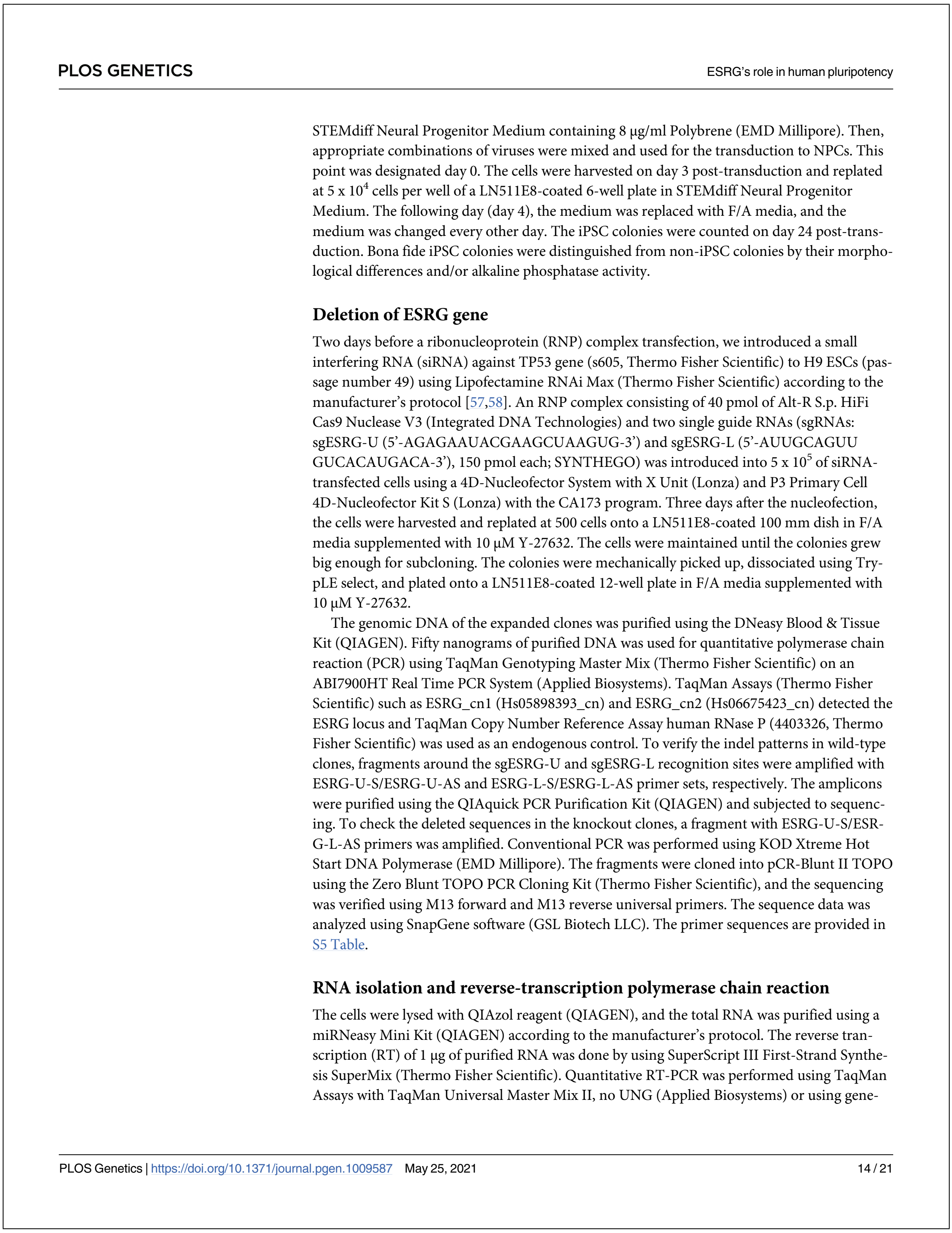
PSCs were cultured on a Matrigel-coated plate in mTeSR1 media until reaching confluency prior to EB formation. The cells were harvested using CTK solution (ReproCELL), and cell clumps were transferred onto an ultra-low binding plate (Corning) in DFK20 media. For the first 2 days, 10 μM Y-27362 was added to the media to improve cell survival. The media was changed every other day. After 8 days of floating culture, the EBs were transferred onto a tissue culture plate coated with 0.1% gelatin (EMD Millipore) and maintained in DFK20 media for another 8 days.

Plasmid

Full-length ESRG complementary DNA (cDNA) was amplified using ESRG-S and ESRG-AS primers and inserted into the BamHI/NotI site of a pMXs retroviral vector [56] using In-Fusion technology (Clontech). The primer sequences for the cloning are available in S5 Table. For the KD experiments, we used transposon vectors such as Sleeping Beauty (SB) and PiggyBac (PB) that contain mouse U6 promoter, drug selection markers and the genes encoding fluorescent proteins [34]. The shRNA sequences are provided in S5 Table.

Reprogramming

Retroviral transduction of the reprogramming factors was performed as described previously [12,20]. A pMXs retroviral vector encoding human OCT3/4 (RRID:Addgene_17217), human SOX2 (RRID:Addgene_17218), human KLF4 (RRID:Addgene_17219), human c-MYC (RRID: Addgene_17220) and ESRG (6 μg each) along with 3 μg of pMD2.G (gift from Dr. D. Trono; RRID:Addgene_12259) was transfected into PLAT-GP packaging cells, which were plated at 3.6 x 10^6 cells per 100 mm dish the day before transfection, using FuGENE6 transfection reagent (Promega). Two days after the transfection, virus-containing supernatant was collected and filtered through a 0.45 μm-pore size cellulose acetate filter to remove the cell debris. Viral particles were precipitated using Retro-X Concentrator (Clontech) and resuspended in



STEMdiff Neural Progenitor Medium containing 8 μg/ml Polybrene (EMD Millipore). Then, appropriate combinations of viruses were mixed and used for the transduction to NPCs. This point was designated day 0. The cells were harvested on day 3 post-transduction and replated at 5 x 10^4 cells per well of a LN511E8-coated 6-well plate in STEMdiff Neural Progenitor Medium. The following day (day 4), the medium was replaced with F/A media, and the medium was changed every other day. The iPSC colonies were counted on day 24 post-transduction. Bona fide iPSC colonies were distinguished from non-iPSC colonies by their morphological differences and/or alkaline phosphatase activity.

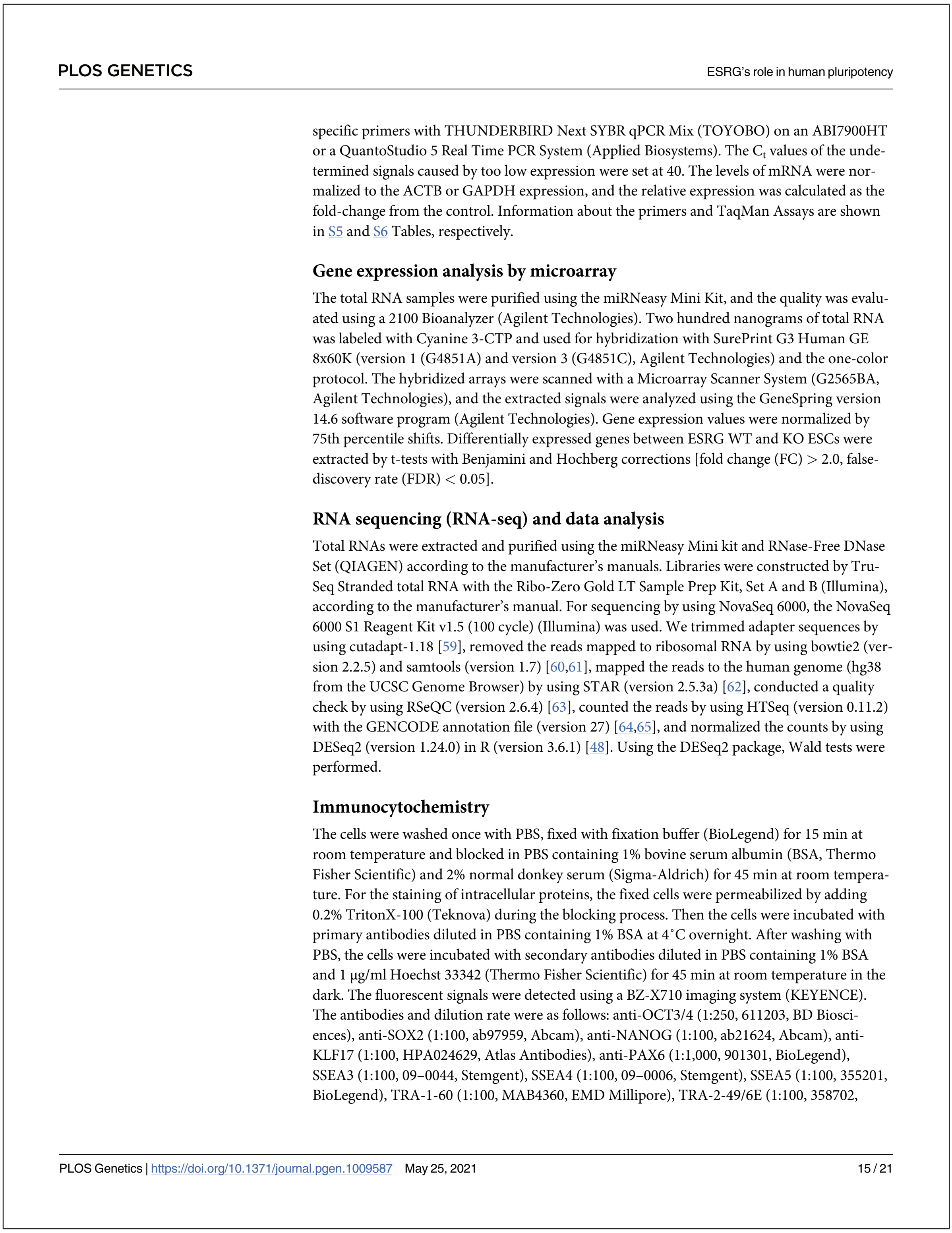
Deletion of ESRG gene

Two days before a ribonucleoprotein (RNP) complex transfection, we introduced a small interfering RNA (siRNA) against TP53 gene (s605, Thermo Fisher Scientific) to H9 ESCs (passage number 49) using Lipofectamine RNAi Max (Thermo Fisher Scientific) according to the manufacturer's protocol [57,58]. An RNP complex consisting of 40 pmol of Alt-R S.p. HiFi Cas9 Nuclease V3 (Integrated DNA Technologies) and two single guide RNAs (sgRNAs: sgESRG-U (5'-AGAGAAUACGAAGCUAAGUG-3') and sgESRG-L (5'-AUUGCAGUU GUCACAUGACA-3'), 150 pmol each; SYNTHEGO) was introduced into 5 x 10^5 of siRNA-transfected cells using a 4D-Nucleofector System with X Unit (Lonza) and P3 Primary Cell 4D-Nucleofector Kit S (Lonza) with the CA173 program. Three days after the nucleofection, the cells were harvested and replated at 500 cells onto a LN511E8-coated 100 mm dish in F/A media supplemented with 10 μM Y-27632. The cells were maintained until the colonies grew big enough for subcloning. The colonies were mechanically picked up, dissociated using TrypLE select, and plated onto a LN511E8-coated 12-well plate in F/A media supplemented with 10 μM Y-27632.

The genomic DNA of the expanded clones was purified using the DNeasy Blood & Tissue Kit (QIAGEN). Fifty nanograms of purified DNA was used for quantitative polymerase chain reaction (PCR) using TaqMan Genotyping Master Mix (Thermo Fisher Scientific) on an ABI7900HT Real Time PCR System (Applied Biosystems). TaqMan Assays (Thermo Fisher Scientific) such as ESRG_cn1 (Hs05898393_cn) and ESRG_cn2 (Hs06675423_cn) detected the ESRG locus and TaqMan Copy Number Reference Assay human RNase P (4403326, Thermo Fisher Scientific) was used as an endogenous control. To verify the indel patterns in wild-type clones, fragments around the sgESRG-U and sgESRG-L recognition sites were amplified with ESRG-U-S/ESRG-U-AS and ESRG-L-S/ESRG-L-AS primer sets, respectively. The amplicons were purified using the QIAquick PCR Purification Kit (QIAGEN) and subjected to sequencing. To check the deleted sequences in the knockout clones, a fragment with ESRG-U-S/ESRG-L-AS primers was amplified. Conventional PCR was performed using KOD Xtreme Hot Start DNA Polymerase (EMD Millipore). The fragments were cloned into pCR-Blunt II TOPO using the Zero Blunt TOPO PCR Cloning Kit (Thermo Fisher Scientific), and the sequencing was verified using M13 forward and M13 reverse universal primers. The sequence data was analyzed using SnapGene software (GSL Biotech LLC). The primer sequences are provided in S5 Table.

RNA isolation and reverse-transcription polymerase chain reaction

The cells were lysed with QIAzol reagent (QIAGEN), and the total RNA was purified using a miRNeasy Mini Kit (QIAGEN) according to the manufacturer's protocol. The reverse transcription (RT) of 1 μg of purified RNA was done by using SuperScript III First-Strand Synthesis SuperMix (Thermo Fisher Scientific). Quantitative RT-PCR was performed using TaqMan Assays with TaqMan Universal Master Mix II, no UNG (Applied Biosystems) or using gene-



specific primers with THUNDERBIRD Next SYBR qPCR Mix (TOYOBO) on an ABI7900HT or a QuantoStudio 5 Real Time PCR System (Applied Biosystems). The C_t values of the undetermined signals caused by too low expression were set at 40. The levels of mRNA were normalized to the ACTB or GAPDH expression, and the relative expression was calculated as the fold-change from the control. Information about the primers and TaqMan Assays are shown in S5 and S6 Tables, respectively.

Gene expression analysis by microarray

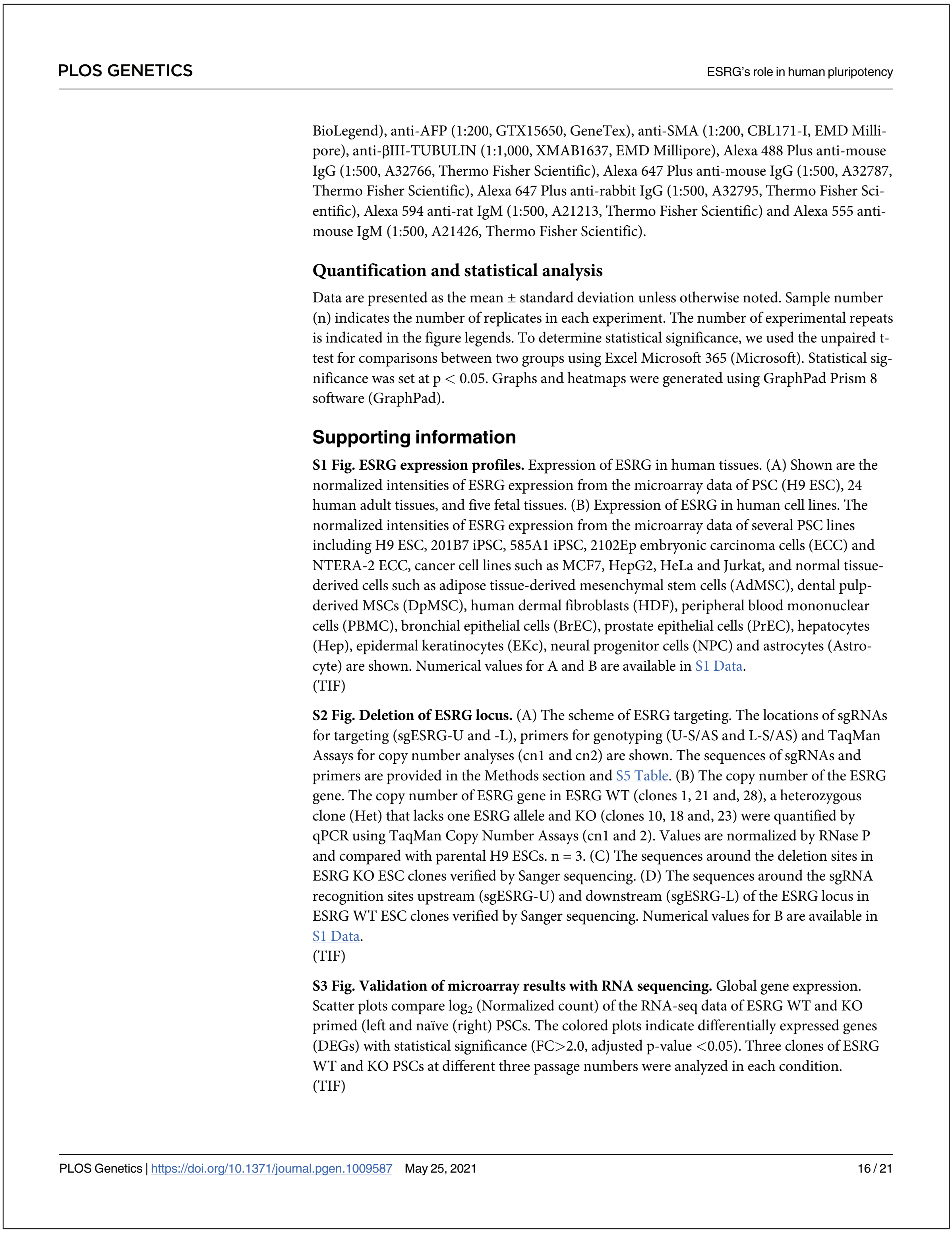
The total RNA samples were purified using the miRNeasy Mini Kit, and the quality was evaluated using a 2100 Bioanalyzer (Agilent Technologies). Two hundred nanograms of total RNA was labeled with Cyanine 3-CTP and used for hybridization with SurePrint G3 Human GE 8x60K (version 1 (G4851A) and version 3 (G4851C), Agilent Technologies) and the one-color protocol. The hybridized arrays were scanned with a Microarray Scanner System (G2565BA, Agilent Technologies), and the extracted signals were analyzed using the GeneSpring version 14.6 software program (Agilent Technologies). Gene expression values were normalized by 75th percentile shifts. Differentially expressed genes between ESRG WT and KO ESCs were extracted by t-tests with Benjamini and Hochberg corrections [fold change (FC) > 2.0, false-discovery rate (FDR) < 0.05].

RNA sequencing (RNA-seq) and data analysis

Total RNAs were extracted and purified using the miRNeasy Mini kit and RNase-Free DNase Set (QIAGEN) according to the manufacturer's manuals. Libraries were constructed by Tru-Seq Stranded total RNA with the Ribo-Zero Gold LT Sample Prep Kit, Set A and B (Illumina), according to the manufacturer's manual. For sequencing by using NovaSeq 6000, the NovaSeq 6000 S1 Reagent Kit v1.5 (100 cycle) (Illumina) was used. We trimmed adapter sequences by using cutadapt-1.18 [59], removed the reads mapped to ribosomal RNA by using bowtie2 (version 2.2.5) and samtools (version 1.7) [60,61], mapped the reads to the human genome (hg38 from the UCSC Genome Browser) by using STAR (version 2.5.3a) [62], conducted a quality check by using RSeQC (version 2.6.4) [63], counted the reads by using HTSeq (version 0.11.2) with the GENCODE annotation file (version 27) [64,65], and normalized the counts by using DESeq2 (version 1.24.0) in R (version 3.6.1) [48]. Using the DESeq2 package, Wald tests were performed.

Immunocytochemistry

The cells were washed once with PBS, fixed with fixation buffer (BioLegend) for 15 min at room temperature and blocked in PBS containing 1% bovine serum albumin (BSA, Thermo Fisher Scientific) and 2% normal donkey serum (Sigma-Aldrich) for 45 min at room temperature. For the staining of intracellular proteins, the fixed cells were permeabilized by adding 0.2% TritonX-100 (Teknova) during the blocking process. Then the cells were incubated with primary antibodies diluted in PBS containing 1% BSA at 4°C overnight. After washing with PBS, the cells were incubated with secondary antibodies diluted in PBS containing 1% BSA and 1 μg/ml Hoechst 33342 (Thermo Fisher Scientific) for 45 min at room temperature in the dark. The fluorescent signals were detected using a BZ-X710 imaging system (KEYENCE). The antibodies and dilution rate were as follows: anti-OCT3/4 (1:250, 611203, BD Biosciences), anti-SOX2 (1:100, ab97959, Abcam), anti-NANOG (1:100, ab21624, Abcam), anti-KLF17 (1:100, HPA024629, Atlas Antibodies), anti-PAX6 (1:1,000, 901301, BioLegend), SSEA3 (1:100, 09–0044, Stemgent), SSEA4 (1:100, 09–0006, Stemgent), SSEA5 (1:100, 355201, BioLegend), TRA-1-60 (1:100, MAB4360, EMD Millipore), TRA-2-49/6E (1:100, 358702,



BioLegend), anti-AFP (1:200, GTX15650, GeneTex), anti-SMA (1:200, CBL171-I, EMD Millipore), anti-βIII-TUBULIN (1:1,000, XMAB1637, EMD Millipore), Alexa 488 Plus anti-mouse IgG (1:500, A32766, Thermo Fisher Scientific), Alexa 647 Plus anti-mouse IgG (1:500, A32787, Thermo Fisher Scientific), Alexa 647 Plus anti-rabbit IgG (1:500, A32795, Thermo Fisher Scientific), Alexa 594 anti-rat IgM (1:500, A21213, Thermo Fisher Scientific) and Alexa 555 anti-mouse IgM (1:500, A21426, Thermo Fisher Scientific).

Quantification and statistical analysis

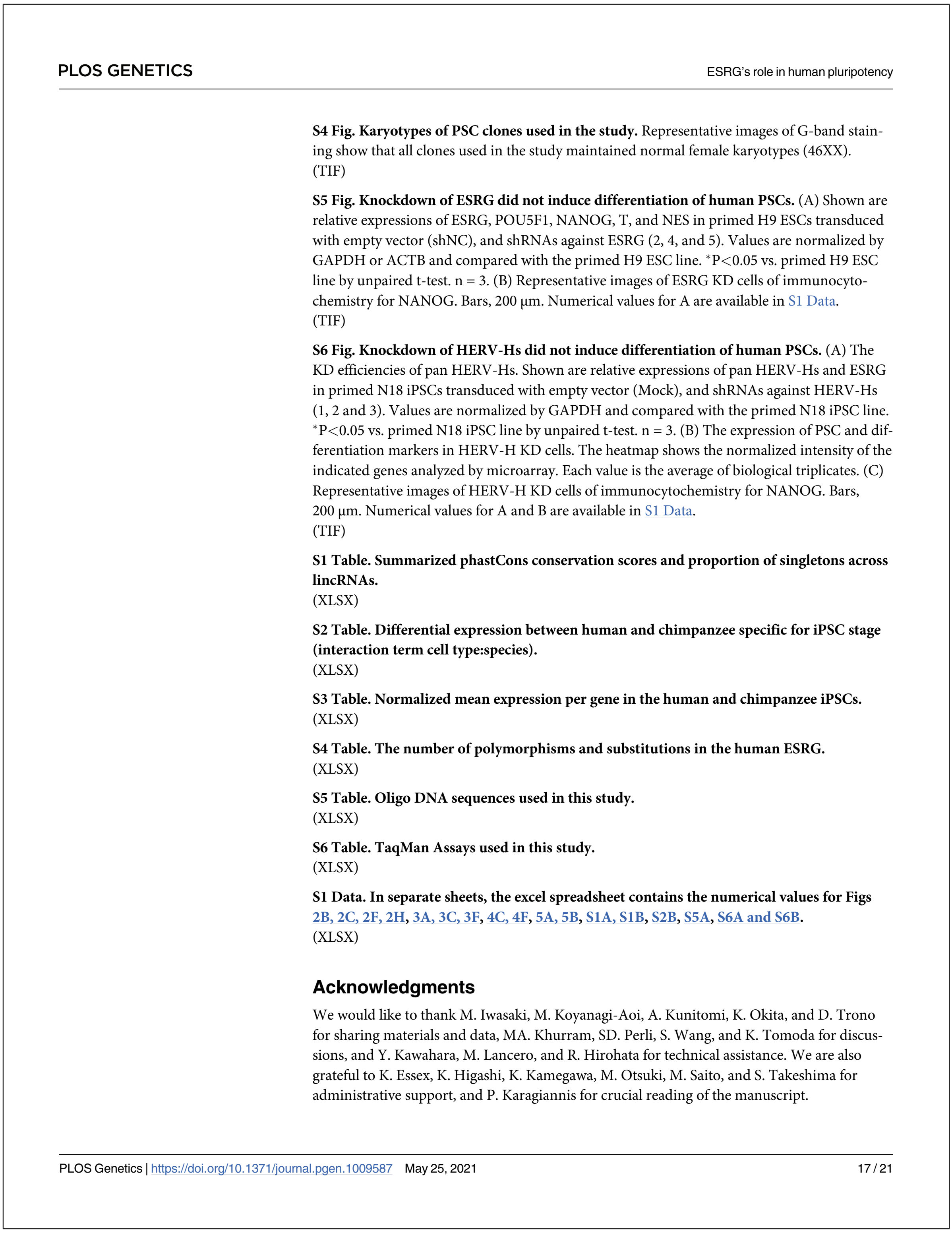
Data are presented as the mean ± standard deviation unless otherwise noted. Sample number (n) indicates the number of replicates in each experiment. The number of experimental repeats is indicated in the figure legends. To determine statistical significance, we used the unpaired t-test for comparisons between two groups using Excel Microsoft 365 (Microsoft). Statistical significance was set at $p < 0.05$. Graphs and heatmaps were generated using GraphPad Prism 8 software (GraphPad).

Supporting information

S1 Fig. ESRG expression profiles. Expression of ESRG in human tissues. (A) Shown are the normalized intensities of ESRG expression from the microarray data of PSC (H9 ESC), 24 human adult tissues, and five fetal tissues. (B) Expression of ESRG in human cell lines. The normalized intensities of ESRG expression from the microarray data of several PSC lines including H9 ESC, 201B7 iPSC, 585A1 iPSC, 2102Ep embryonic carcinoma cells (ECC) and NTERA-2 ECC, cancer cell lines such as MCF7, HepG2, HeLa and Jurkat, and normal tissue-derived cells such as adipose tissue-derived mesenchymal stem cells (AdMSC), dental pulp-derived MSCs (DpMSC), human dermal fibroblasts (HDF), peripheral blood mononuclear cells (PBMC), bronchial epithelial cells (BrEC), prostate epithelial cells (PrEC), hepatocytes (Hep), epidermal keratinocytes (EKc), neural progenitor cells (NPC) and astrocytes (Astrocyte) are shown. Numerical values for A and B are available in S1 Data.
(TIF)

S2 Fig. Deletion of ESRG locus. (A) The scheme of ESRG targeting. The locations of sgRNAs for targeting (sgESRG-U and -L), primers for genotyping (U-S/AS and L-S/AS) and TaqMan Assays for copy number analyses (cn1 and cn2) are shown. The sequences of sgRNAs and primers are provided in the Methods section and S5 Table. (B) The copy number of the ESRG gene. The copy number of ESRG gene in ESRG WT (clones 1, 21 and, 28), a heterozygous clone (Het) that lacks one ESRG allele and KO (clones 10, 18 and, 23) were quantified by qPCR using TaqMan Copy Number Assays (cn1 and 2). Values are normalized by RNase P and compared with parental H9 ESCs. n = 3. (C) The sequences around the deletion sites in ESRG KO ESC clones verified by Sanger sequencing. (D) The sequences around the sgRNA recognition sites upstream (sgESRG-U) and downstream (sgESRG-L) of the ESRG locus in ESRG WT ESC clones verified by Sanger sequencing. Numerical values for B are available in S1 Data.
(TIF)

S3 Fig. Validation of microarray results with RNA sequencing. Global gene expression. Scatter plots compare $\log_2$ (Normalized count) of the RNA-seq data of ESRG WT and KO primed (left and naïve (right) PSCs. The colored plots indicate differentially expressed genes (DEGs) with statistical significance (FC>2.0, adjusted p-value <0.05). Three clones of ESRG WT and KO PSCs at different three passage numbers were analyzed in each condition.
(TIF)



S4 Fig. Karyotypes of PSC clones used in the study. Representative images of G-band staining show that all clones used in the study maintained normal female karyotypes (46XX).
(TIF)

S5 Fig. Knockdown of ESRG did not induce differentiation of human PSCs. (A) Shown are relative expressions of ESRG, POU5F1, NANOG, T, and NES in primed H9 ESCs transduced with empty vector (shNC), and shRNAs against ESRG (2, 4, and 5). Values are normalized by GAPDH or ACTB and compared with the primed H9 ESC line. $^*P<0.05$ vs. primed H9 ESC line by unpaired t-test. n = 3. (B) Representative images of ESRG KD cells of immunocytochemistry for NANOG. Bars, 200 μm. Numerical values for A are available in S1 Data.
(TIF)

S6 Fig. Knockdown of HERV-Hs did not induce differentiation of human PSCs. (A) The KD efficiencies of pan HERV-Hs. Shown are relative expressions of pan HERV-Hs and ESRG in primed N18 iPSCs transduced with empty vector (Mock), and shRNAs against HERV-Hs (1, 2 and 3). Values are normalized by GAPDH and compared with the primed N18 iPSC line. $^*P<0.05$ vs. primed N18 iPSC line by unpaired t-test. n = 3. (B) The expression of PSC and differentiation markers in HERV-H KD cells. The heatmap shows the normalized intensity of the indicated genes analyzed by microarray. Each value is the average of biological triplicates. (C) Representative images of HERV-H KD cells of immunocytochemistry for NANOG. Bars, 200 μm. Numerical values for A and B are available in S1 Data.
(TIF)

S1 Table. Summarized phastCons conservation scores and proportion of singletons across lincRNAs.
(XLSX)

S2 Table. Differential expression between human and chimpanzee specific for iPSC stage (interaction term cell type:species).
(XLSX)

S3 Table. Normalized mean expression per gene in the human and chimpanzee iPSCs.
(XLSX)

S4 Table. The number of polymorphisms and substitutions in the human ESRG.
(XLSX)

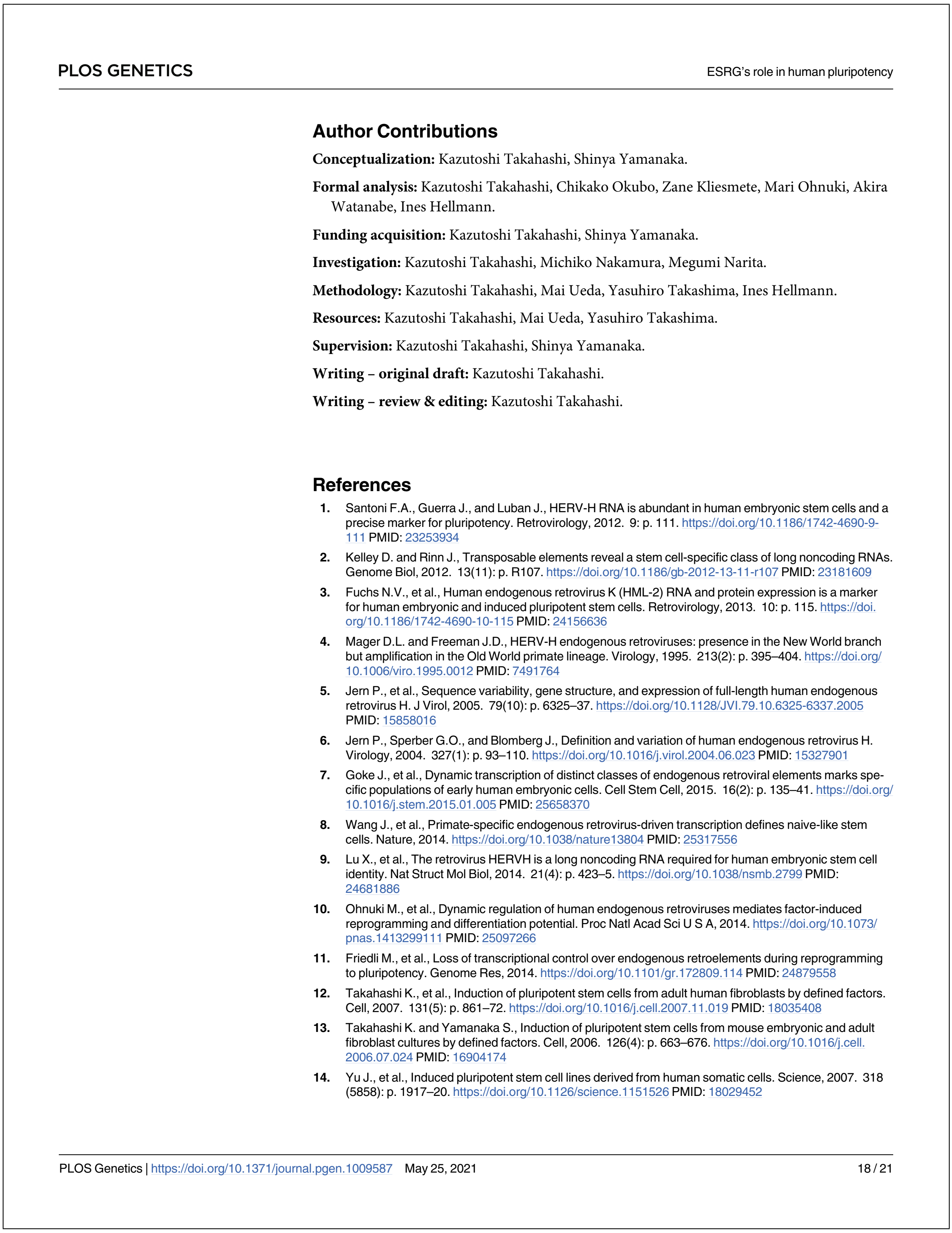
S5 Table. Oligo DNA sequences used in this study.
(XLSX)

S6 Table. TaqMan Assays used in this study.
(XLSX)

S1 Data. In separate sheets, the excel spreadsheet contains the numerical values for Figs 2B, 2C, 2F, 2H, 3A, 3C, 3F, 4C, 4F, 5A, 5B, S1A, S1B, S2B, S5A, S6A and S6B.
(XLSX)

Acknowledgments

We would like to thank M. Iwasaki, M. Koyanagi-Aoi, A. Kunitomi, K. Okita, and D. Trono for sharing materials and data, MA. Khurram, SD. Perli, S. Wang, and K. Tomoda for discussions, and Y. Kawahara, M. Lancero, and R. Hirohata for technical assistance. We are also grateful to K. Essex, K. Higashi, K. Kamegawa, M. Otsuki, M. Saito, and S. Takeshima for administrative support, and P. Karagiannis for crucial reading of the manuscript.



Author Contributions

Conceptualization: Kazutoshi Takahashi, Shinya Yamanaka.

Formal analysis: Kazutoshi Takahashi, Chikako Okubo, Zane Kliesmete, Mari Ohnuki, Akira Watanabe, Ines Hellmann.

Funding acquisition: Kazutoshi Takahashi, Shinya Yamanaka.

Investigation: Kazutoshi Takahashi, Michiko Nakamura, Megumi Narita.

Methodology: Kazutoshi Takahashi, Mai Ueda, Yasuhiro Takashima, Ines Hellmann.

Resources: Kazutoshi Takahashi, Mai Ueda, Yasuhiro Takashima.

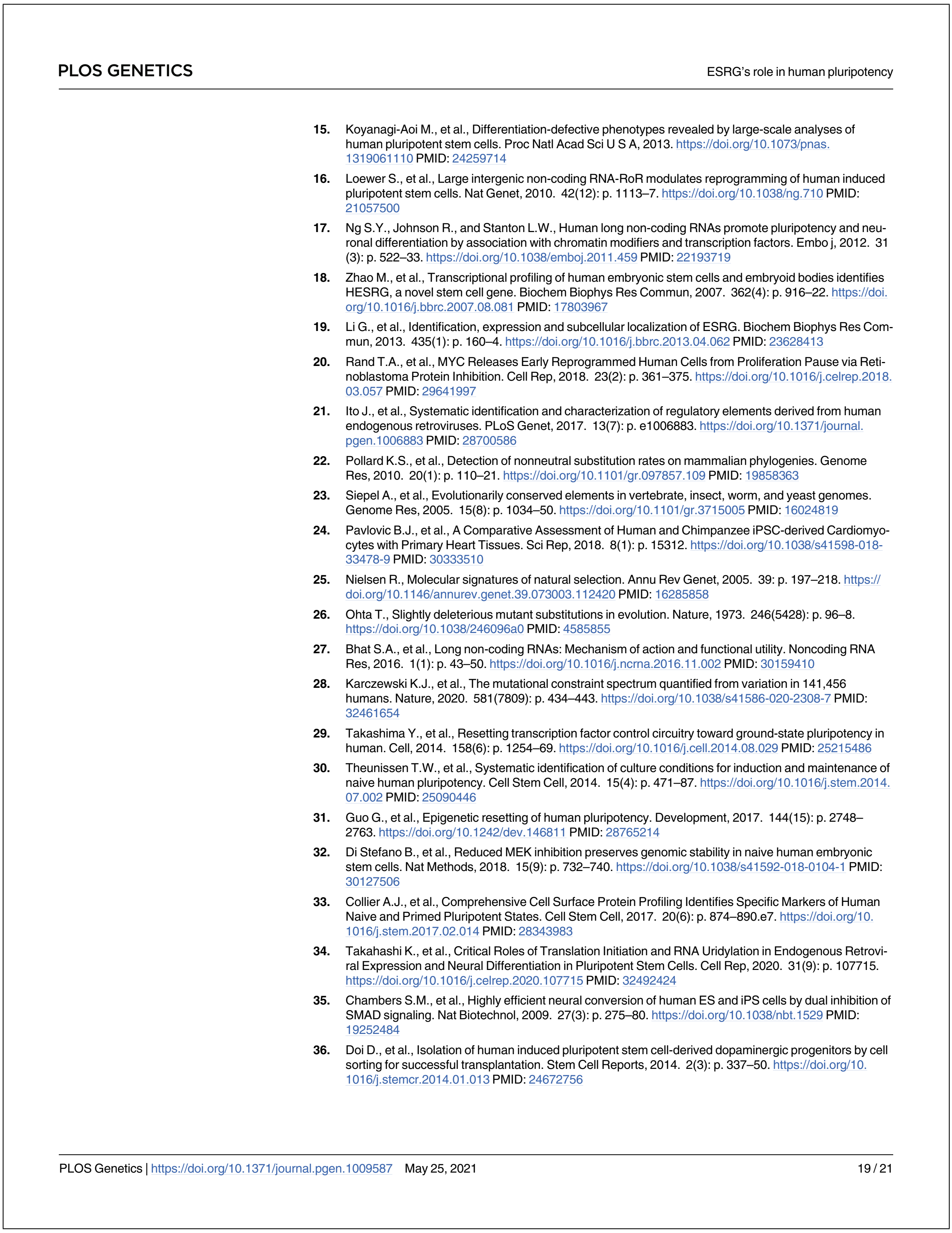
Supervision: Kazutoshi Takahashi, Shinya Yamanaka.

Writing – original draft: Kazutoshi Takahashi.

Writing – review & editing: Kazutoshi Takahashi.

References

1. Santoni F.A., Guerra J., and Luban J., HERV-H RNA is abundant in human embryonic stem cells and a precise marker for pluripotency. Retrovirology, 2012. 9: p. 111. https://doi.org/10.1186/1742-4690-9-111 PMID: 23253934
2. Kelley D. and Rinn J., Transposable elements reveal a stem cell-specific class of long noncoding RNAs. Genome Biol, 2012. 13(11): p. R107. https://doi.org/10.1186/gb-2012-13-11-r107 PMID: 23181609
3. Fuchs N.V., et al., Human endogenous retrovirus K (HML-2) RNA and protein expression is a marker for human embryonic and induced pluripotent stem cells. Retrovirology, 2013. 10: p. 115. https://doi.org/10.1186/1742-4690-10-115 PMID: 24156636
4. Mager D.L. and Freeman J.D., HERV-H endogenous retroviruses: presence in the New World branch but amplification in the Old World primate lineage. Virology, 1995. 213(2): p. 395–404. https://doi.org/10.1006/viro.1995.0012 PMID: 7491764
5. Jern P., et al., Sequence variability, gene structure, and expression of full-length human endogenous retrovirus H. J Virol, 2005. 79(10): p. 6325–37. https://doi.org/10.1128/JVI.79.10.6325-6337.2005 PMID: 15858016
6. Jern P., Sperber G.O., and Blomberg J., Definition and variation of human endogenous retrovirus H. Virology, 2004. 327(1): p. 93–110. https://doi.org/10.1016/j.virol.2004.06.023 PMID: 15327901
7. Goke J., et al., Dynamic transcription of distinct classes of endogenous retroviral elements marks specific populations of early human embryonic cells. Cell Stem Cell, 2015. 16(2): p. 135–41. https://doi.org/10.1016/j.stem.2015.01.005 PMID: 25658370
8. Wang J., et al., Primate-specific endogenous retrovirus-driven transcription defines naive-like stem cells. Nature, 2014. https://doi.org/10.1038/nature13804 PMID: 25317556
9. Lu X., et al., The retrovirus HERVH is a long noncoding RNA required for human embryonic stem cell identity. Nat Struct Mol Biol, 2014. 21(4): p. 423–5. https://doi.org/10.1038/nsmb.2799 PMID: 24681886
10. Ohnuki M., et al., Dynamic regulation of human endogenous retroviruses mediates factor-induced reprogramming and differentiation potential. Proc Natl Acad Sci U S A, 2014. https://doi.org/10.1073/pnas.1413299111 PMID: 25097266
11. Friedli M., et al., Loss of transcriptional control over endogenous retroelements during reprogramming to pluripotency. Genome Res, 2014. https://doi.org/10.1101/gr.172809.114 PMID: 24879558
12. Takahashi K., et al., Induction of pluripotent stem cells from adult human fibroblasts by defined factors. Cell, 2007. 131(5): p. 861–72. https://doi.org/10.1016/j.cell.2007.11.019 PMID: 18035408
13. Takahashi K. and Yamanaka S., Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. Cell, 2006. 126(4): p. 663–676. https://doi.org/10.1016/j.cell.2006.07.024 PMID: 16904174
14. Yu J., et al., Induced pluripotent stem cell lines derived from human somatic cells. Science, 2007. 318 (5858): p. 1917–20. https://doi.org/10.1126/science.1151526 PMID: 18029452



15. Koyanagi-Aoi M., et al., Differentiation-defective phenotypes revealed by large-scale analyses of human pluripotent stem cells. Proc Natl Acad Sci U S A, 2013. https://doi.org/10.1073/pnas.1319061110 PMID: 24259714

16. Loewer S., et al., Large intergenic non-coding RNA-RoR modulates reprogramming of human induced pluripotent stem cells. Nat Genet, 2010. 42(12): p. 1113–7. https://doi.org/10.1038/ng.710 PMID: 21057500

17. Ng S.Y., Johnson R., and Stanton L.W., Human long non-coding RNAs promote pluripotency and neuronal differentiation by association with chromatin modifiers and transcription factors. Embo j, 2012. 31 (3): p. 522–33. https://doi.org/10.1038/emboj.2011.459 PMID: 22193719

18. Zhao M., et al., Transcriptional profiling of human embryonic stem cells and embryoid bodies identifies HESRG, a novel stem cell gene. Biochem Biophys Res Commun, 2007. 362(4): p. 916–22. https://doi.org/10.1016/j.bbrc.2007.08.081 PMID: 17803967

19. Li G., et al., Identification, expression and subcellular localization of ESRG. Biochem Biophys Res Commun, 2013. 435(1): p. 160–4. https://doi.org/10.1016/j.bbrc.2013.04.062 PMID: 23628413

20. Rand T.A., et al., MYC Releases Early Reprogrammed Human Cells from Proliferation Pause via Retinoblastoma Protein Inhibition. Cell Rep, 2018. 23(2): p. 361–375. https://doi.org/10.1016/j.celrep.2018.03.057 PMID: 29641997

21. Ito J., et al., Systematic identification and characterization of regulatory elements derived from human endogenous retroviruses. PLoS Genet, 2017. 13(7): p. e1006883. https://doi.org/10.1371/journal.pgen.1006883 PMID: 28700586

22. Pollard K.S., et al., Detection of nonneutral substitution rates on mammalian phylogenies. Genome Res, 2010. 20(1): p. 110–21. https://doi.org/10.1101/gr.097857.109 PMID: 19858363

23. Siepel A., et al., Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res, 2005. 15(8): p. 1034–50. https://doi.org/10.1101/gr.3715005 PMID: 16024819

24. Pavlovic B.J., et al., A Comparative Assessment of Human and Chimpanzee iPSC-derived Cardiomyocytes with Primary Heart Tissues. Sci Rep, 2018. 8(1): p. 15312. https://doi.org/10.1038/s41598-018-33478-9 PMID: 30333510

25. Nielsen R., Molecular signatures of natural selection. Annu Rev Genet, 2005. 39: p. 197–218. https://doi.org/10.1146/annurev.genet.39.073003.112420 PMID: 16285858

26. Ohta T., Slightly deleterious mutant substitutions in evolution. Nature, 1973. 246(5428): p. 96–8. https://doi.org/10.1038/246096a0 PMID: 4585855

27. Bhat S.A., et al., Long non-coding RNAs: Mechanism of action and functional utility. Noncoding RNA Res, 2016. 1(1): p. 43–50. https://doi.org/10.1016/j.ncrna.2016.11.002 PMID: 30159410

28. Karczewski K.J., et al., The mutational constraint spectrum quantified from variation in 141,456 humans. Nature, 2020. 581(7809): p. 434–443. https://doi.org/10.1038/s41586-020-2308-7 PMID: 32461654

29. Takashima Y., et al., Resetting transcription factor control circuitry toward ground-state pluripotency in human. Cell, 2014. 158(6): p. 1254–69. https://doi.org/10.1016/j.cell.2014.08.029 PMID: 25215486

30. Theunissen T.W., et al., Systematic identification of culture conditions for induction and maintenance of naive human pluripotency. Cell Stem Cell, 2014. 15(4): p. 471–87. https://doi.org/10.1016/j.stem.2014.07.002 PMID: 25090446

31. Guo G., et al., Epigenetic resetting of human pluripotency. Development, 2017. 144(15): p. 2748–2763. https://doi.org/10.1242/dev.146811 PMID: 28765214

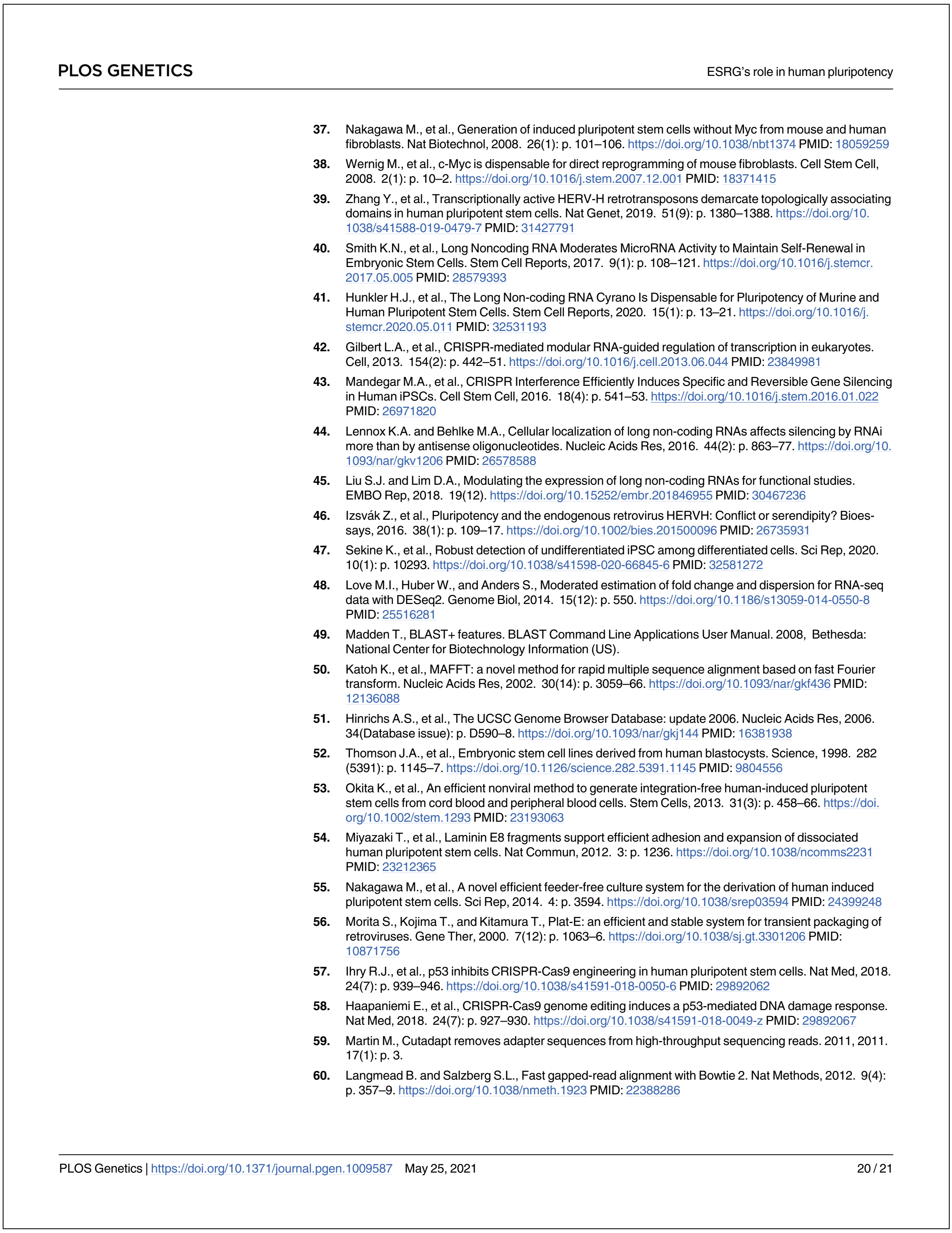
32. Di Stefano B., et al., Reduced MEK inhibition preserves genomic stability in naive human embryonic stem cells. Nat Methods, 2018. 15(9): p. 732–740. https://doi.org/10.1038/s41592-018-0104-1 PMID: 30127506

33. Collier A.J., et al., Comprehensive Cell Surface Protein Profiling Identifies Specific Markers of Human Naive and Primed Pluripotent States. Cell Stem Cell, 2017. 20(6): p. 874–890.e7. https://doi.org/10.1016/j.stem.2017.02.014 PMID: 28343983

34. Takahashi K., et al., Critical Roles of Translation Initiation and RNA Uridylation in Endogenous Retroviral Expression and Neural Differentiation in Pluripotent Stem Cells. Cell Rep, 2020. 31(9): p. 107715. https://doi.org/10.1016/j.celrep.2020.107715 PMID: 32492424

35. Chambers S.M., et al., Highly efficient neural conversion of human ES and iPS cells by dual inhibition of SMAD signaling. Nat Biotechnol, 2009. 27(3): p. 275–80. https://doi.org/10.1038/nbt.1529 PMID: 19252484

36. Doi D., et al., Isolation of human induced pluripotent stem cell-derived dopaminergic progenitors by cell sorting for successful transplantation. Stem Cell Reports, 2014. 2(3): p. 337–50. https://doi.org/10.1016/j.stemcr.2014.01.013 PMID: 24672756



37. Nakagawa M., et al., Generation of induced pluripotent stem cells without Myc from mouse and human fibroblasts. Nat Biotechnol, 2008. 26(1): p. 101–106. https://doi.org/10.1038/nbt1374 PMID: 18059259

38. Wernig M., et al., c-Myc is dispensable for direct reprogramming of mouse fibroblasts. Cell Stem Cell, 2008. 2(1): p. 10–2. https://doi.org/10.1016/j.stem.2007.12.001 PMID: 18371415

39. Zhang Y., et al., Transcriptionally active HERV-H retrotransposons demarcate topologically associating domains in human pluripotent stem cells. Nat Genet, 2019. 51(9): p. 1380–1388. https://doi.org/10.1038/s41588-019-0479-7 PMID: 31427791

40. Smith K.N., et al., Long Noncoding RNA Moderates MicroRNA Activity to Maintain Self-Renewal in Embryonic Stem Cells. Stem Cell Reports, 2017. 9(1): p. 108–121. https://doi.org/10.1016/j.stemcr.2017.05.005 PMID: 28579393

41. Hunkler H.J., et al., The Long Non-coding RNA Cyrano Is Dispensable for Pluripotency of Murine and Human Pluripotent Stem Cells. Stem Cell Reports, 2020. 15(1): p. 13–21. https://doi.org/10.1016/j.stemcr.2020.05.011 PMID: 32531193

42. Gilbert L.A., et al., CRISPR-mediated modular RNA-guided regulation of transcription in eukaryotes. Cell, 2013. 154(2): p. 442–51. https://doi.org/10.1016/j.cell.2013.06.044 PMID: 23849981

43. Mandegar M.A., et al., CRISPR Interference Efficiently Induces Specific and Reversible Gene Silencing in Human iPSCs. Cell Stem Cell, 2016. 18(4): p. 541–53. https://doi.org/10.1016/j.stem.2016.01.022 PMID: 26971820

44. Lennox K.A. and Behlke M.A., Cellular localization of long non-coding RNAs affects silencing by RNAi more than by antisense oligonucleotides. Nucleic Acids Res, 2016. 44(2): p. 863–77. https://doi.org/10.1093/nar/gkv1206 PMID: 26578588

45. Liu S.J. and Lim D.A., Modulating the expression of long non-coding RNAs for functional studies. EMBO Rep, 2018. 19(12). https://doi.org/10.15252/embr.201846955 PMID: 30467236

46. Izsvák Z., et al., Pluripotency and the endogenous retrovirus HERVH: Conflict or serendipity? Bioessays, 2016. 38(1): p. 109–17. https://doi.org/10.1002/bies.201500096 PMID: 26735931

47. Sekine K., et al., Robust detection of undifferentiated iPSC among differentiated cells. Sci Rep, 2020. 10(1): p. 10293. https://doi.org/10.1038/s41598-020-66845-6 PMID: 32581272

48. Love M.I., Huber W., and Anders S., Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol, 2014. 15(12): p. 550. https://doi.org/10.1186/s13059-014-0550-8 PMID: 25516281

49. Madden T., BLAST+ features. BLAST Command Line Applications User Manual. 2008, Bethesda: National Center for Biotechnology Information (US).

50. Katoh K., et al., MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res, 2002. 30(14): p. 3059–66. https://doi.org/10.1093/nar/gkf436 PMID: 12136088

51. Hinrichs A.S., et al., The UCSC Genome Browser Database: update 2006. Nucleic Acids Res, 2006. 34(Database issue): p. D590–8. https://doi.org/10.1093/nar/gkj144 PMID: 16381938

52. Thomson J.A., et al., Embryonic stem cell lines derived from human blastocysts. Science, 1998. 282 (5391): p. 1145–7. https://doi.org/10.1126/science.282.5391.1145 PMID: 9804556

53. Okita K., et al., An efficient nonviral method to generate integration-free human-induced pluripotent stem cells from cord blood and peripheral blood cells. Stem Cells, 2013. 31(3): p. 458–66. https://doi.org/10.1002/stem.1293 PMID: 23193063

54. Miyazaki T., et al., Laminin E8 fragments support efficient adhesion and expansion of dissociated human pluripotent stem cells. Nat Commun, 2012. 3: p. 1236. https://doi.org/10.1038/ncomms2231 PMID: 23212365

55. Nakagawa M., et al., A novel efficient feeder-free culture system for the derivation of human induced pluripotent stem cells. Sci Rep, 2014. 4: p. 3594. https://doi.org/10.1038/srep03594 PMID: 24399248

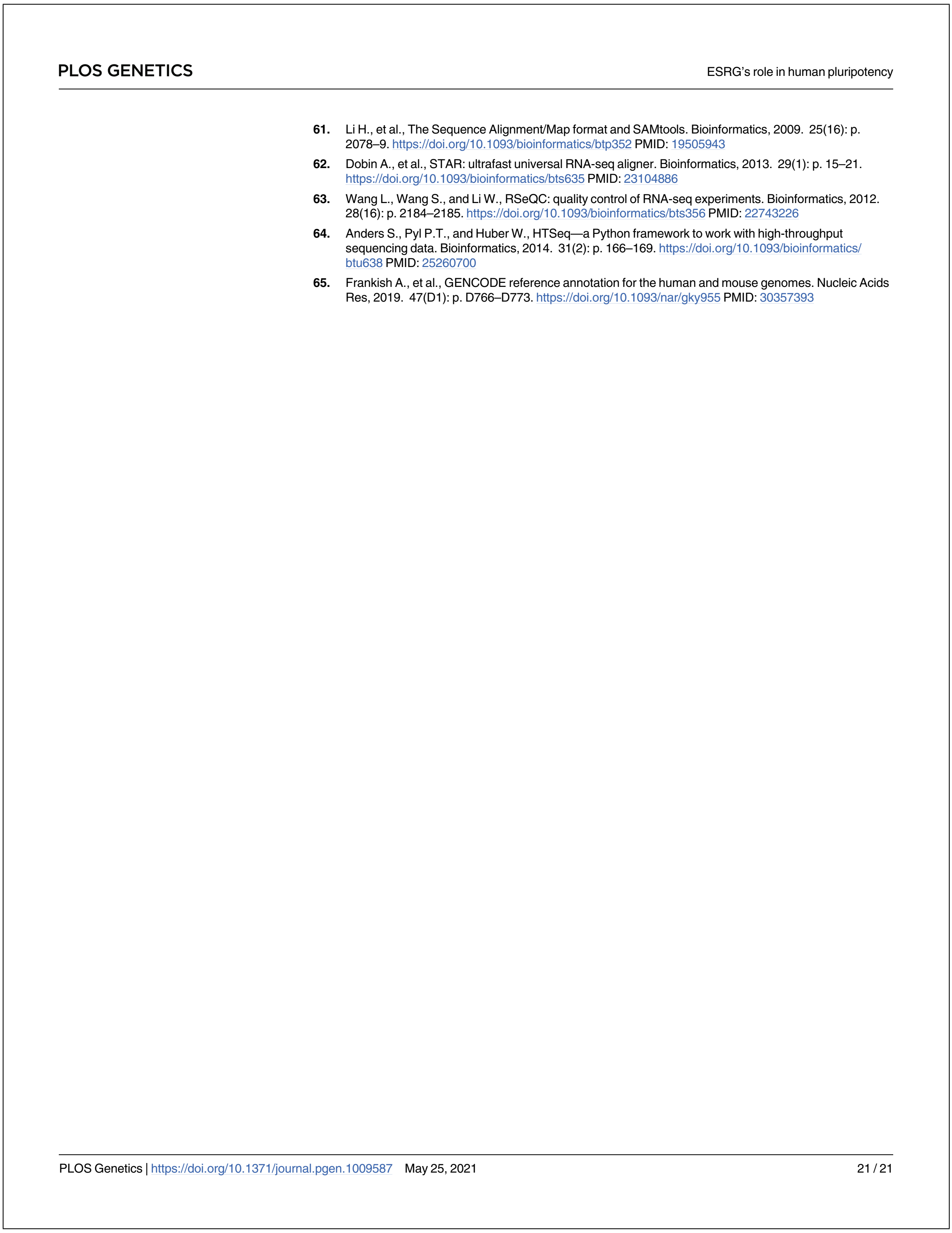
56. Morita S., Kojima T., and Kitamura T., Plat-E: an efficient and stable system for transient packaging of retroviruses. Gene Ther, 2000. 7(12): p. 1063–6. https://doi.org/10.1038/sj.gt.3301206 PMID: 10871756

57. Ihry R.J., et al., p53 inhibits CRISPR-Cas9 engineering in human pluripotent stem cells. Nat Med, 2018. 24(7): p. 939–946. https://doi.org/10.1038/s41591-018-0050-6 PMID: 29892062

58. Haapaniemi E., et al., CRISPR-Cas9 genome editing induces a p53-mediated DNA damage response. Nat Med, 2018. 24(7): p. 927–930. https://doi.org/10.1038/s41591-018-0049-z PMID: 29892067

59. Martin M., Cutadapt removes adapter sequences from high-throughput sequencing reads. 2011, 2011. 17(1): p. 3.

60. Langmead B. and Salzberg S.L., Fast gapped-read alignment with Bowtie 2. Nat Methods, 2012. 9(4): p. 357–9. https://doi.org/10.1038/nmeth.1923 PMID: 22388286



61. Li H., et al., The Sequence Alignment/Map format and SAMtools. Bioinformatics, 2009. 25(16): p. 2078–9. https://doi.org/10.1093/bioinformatics/btp352 PMID: 19505943

62. Dobin A., et al., STAR: ultrafast universal RNA-seq aligner. Bioinformatics, 2013. 29(1): p. 15–21. https://doi.org/10.1093/bioinformatics/bts635 PMID: 23104886

63. Wang L., Wang S., and Li W., RSeQC: quality control of RNA-seq experiments. Bioinformatics, 2012. 28(16): p. 2184–2185. https://doi.org/10.1093/bioinformatics/bts356 PMID: 22743226

64. Anders S., Pyl P.T., and Huber W., HTSeq—a Python framework to work with high-throughput sequencing data. Bioinformatics, 2014. 31(2): p. 166–169. https://doi.org/10.1093/bioinformatics/btu638 PMID: 25260700

65. Frankish A., et al., GENCODE reference annotation for the human and mouse genomes. Nucleic Acids Res, 2019. 47(D1): p. D766–D773. https://doi.org/10.1093/nar/gky955 PMID: 30357393

2.4 Regulatory and coding sequences of TRNP1 co-evolve with brain size and cortical folding in mammals

Kliesmete Z, Wange LE, Vieth B, Esgleas M, Radmer J, Hülsmann M, Geuder J, Richter D, Ohnuki M, Götz M, Hellmann I, Enard W:

"Regulatory and coding sequences of TRNP1 co-evolve with brain size and cortical folding in mammals." (2023)

Elife 12:e83593.

doi: 10.7554/eLife.83593

Supplementary Information is freely available at the publisher's website:

`https://elifesciences.org/articles/83593`

eLife

RESEARCH ARTICLE

Regulatory and coding sequences of TRNP1 co-evolve with brain size and cortical folding in mammals

Zane Kliesmete[1†], Lucas Esteban Wange[1†], Beate Vieth[1], Miriam Esgleas[2,3], Jessica Radmer[1], Matthias Hülsmann[1,4,5], Johanna Geuder[1], Daniel Richter[1], Mari Ohnuki[1], Magdelena Götz[2,3,6], Ines Hellmann[1*‡], Wolfgang Enard[1*‡]

[1]Anthropology and Human Genomics, Faculty of Biology, Ludwig-Maximilians-Universität, Munich, Germany; [2]Physiological Genomics, BioMedical Center - BMC, Ludwig-Maximilians-Universität, Munich, Germany; [3]Institute for Stem Cell Research, Helmholtz Zentrum München, Germany Research Center for Environmental Health, Munich, Germany; [4]Department of Environmental Microbiology, Eawag, Dübendorf, Switzerland; [5]Department of Environmental Systems Science, ETH Zurich, Zurich, Switzerland; [6]SYNERGY, Excellence Cluster of Systems Neurology, BioMedical Center (BMC), Ludwig-Maximilians-Universität München, Munich, Germany

Abstract Brain size and cortical folding have increased and decreased recurrently during mammalian evolution. Identifying genetic elements whose sequence or functional properties co-evolve with these traits can provide unique information on evolutionary and developmental mechanisms. A good candidate for such a comparative approach is *TRNP1*, as it controls proliferation of neural progenitors in mice and ferrets. Here, we investigate the contribution of both regulatory and coding sequences of *TRNP1* to brain size and cortical folding in over 30 mammals. We find that the rate of TRNP1 protein evolution (ω) significantly correlates with brain size, slightly less with cortical folding and much less with body size. This brain correlation is stronger than for >95% of random control proteins. This co-evolution is likely affecting TRNP1 activity, as we find that TRNP1 from species with larger brains and more cortical folding induce higher proliferation rates in neural stem cells. Furthermore, we compare the activity of putative cis-regulatory elements (CREs) of *TRNP1* in a massively parallel reporter assay and identify one CRE that likely co-evolves with cortical folding in Old World monkeys and apes. Our analyses indicate that coding and regulatory changes that increased *TRNP1* activity were positively selected either as a cause or a consequence of increases in brain size and cortical folding. They also provide an example how phylogenetic approaches can inform biological mechanisms, especially when combined with molecular phenotypes across several species.

Editor's evaluation

This is an important paper that combines comparative analysis and experimental assays to investigate the role of protein-coding and regulatory changes at TRNP1 in mammalian brain evolution. The evidence supporting a contribution of TRNP1 is convincing, although the strength of the link between protein-coding changes and trait evolution is stronger and more readily interpretable than the data on gene regulation. The work will be of interest to researchers interested in mammalian evolution, brain evolution, and evolutionary genetics.

***For correspondence:** hellmann@bio.lmu.de (IH); enard@bio.lmu.de (WE)

†These authors contributed equally to this work

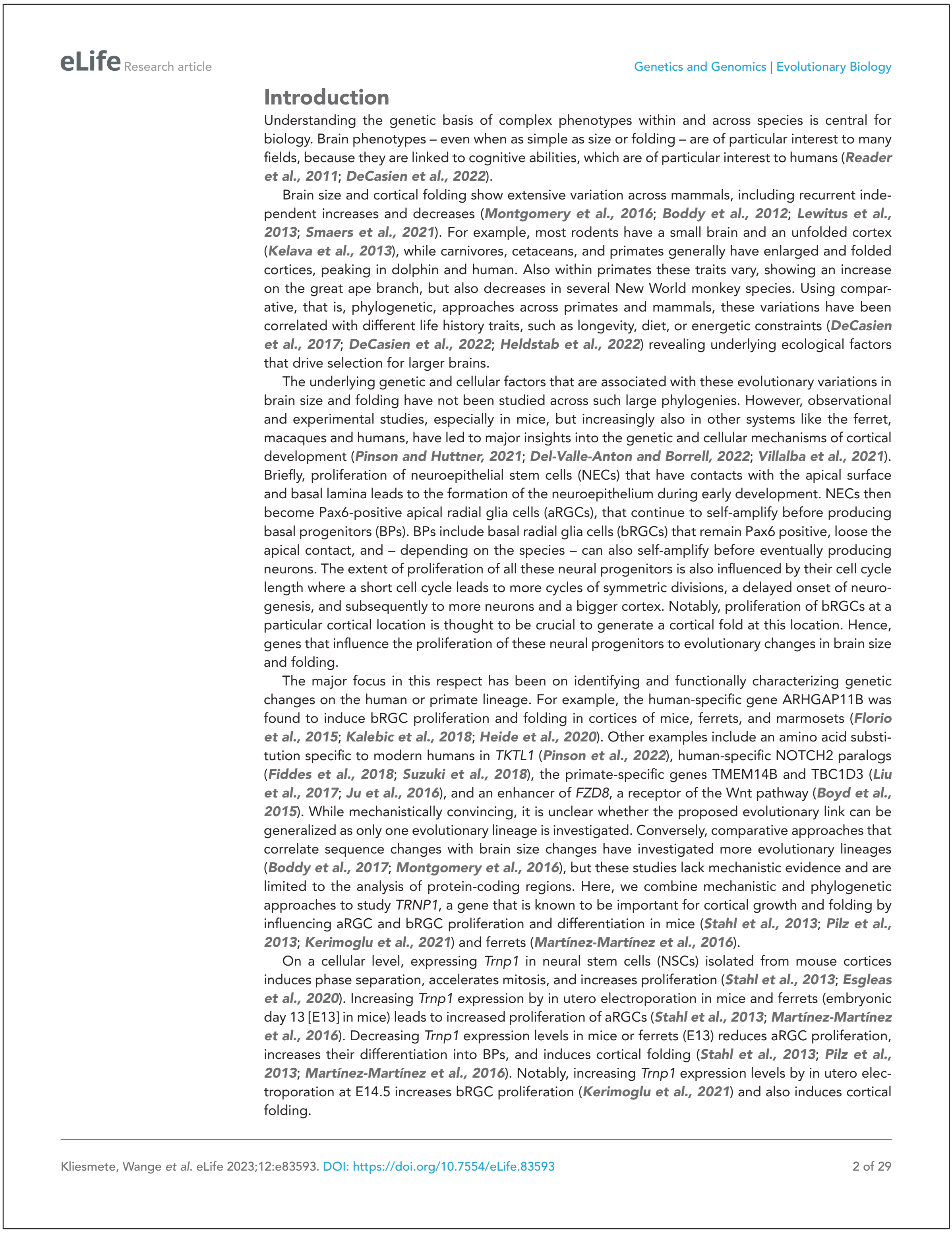
‡These authors also contributed equally to this work

Competing interest: The authors declare that no competing interests exist.

Funding: See page 19

Preprinted: 05 February 2021
Received: 20 September 2022
Accepted: 01 March 2023
Published: 22 March 2023

Reviewing Editor: Jenny Tung, Duke University, United States



Introduction

Understanding the genetic basis of complex phenotypes within and across species is central for biology. Brain phenotypes – even when as simple as size or folding – are of particular interest to many fields, because they are linked to cognitive abilities, which are of particular interest to humans (***Reader et al., 2011***; ***DeCasien et al., 2022***).

Brain size and cortical folding show extensive variation across mammals, including recurrent independent increases and decreases (***Montgomery et al., 2016***; ***Boddy et al., 2012***; ***Lewitus et al., 2013***; ***Smaers et al., 2021***). For example, most rodents have a small brain and an unfolded cortex (***Kelava et al., 2013***), while carnivores, cetaceans, and primates generally have enlarged and folded cortices, peaking in dolphin and human. Also within primates these traits vary, showing an increase on the great ape branch, but also decreases in several New World monkey species. Using comparative, that is, phylogenetic, approaches across primates and mammals, these variations have been correlated with different life history traits, such as longevity, diet, or energetic constraints (***DeCasien et al., 2017***; ***DeCasien et al., 2022***; ***Heldstab et al., 2022***) revealing underlying ecological factors that drive selection for larger brains.

The underlying genetic and cellular factors that are associated with these evolutionary variations in brain size and folding have not been studied across such large phylogenies. However, observational and experimental studies, especially in mice, but increasingly also in other systems like the ferret, macaques and humans, have led to major insights into the genetic and cellular mechanisms of cortical development (***Pinson and Huttner, 2021***; ***Del-Valle-Anton and Borrell, 2022***; ***Villalba et al., 2021***). Briefly, proliferation of neuroepithelial stem cells (NECs) that have contacts with the apical surface and basal lamina leads to the formation of the neuroepithelium during early development. NECs then become Pax6-positive apical radial glia cells (aRGCs), that continue to self-amplify before producing basal progenitors (BPs). BPs include basal radial glia cells (bRGCs) that remain Pax6 positive, loose the apical contact, and – depending on the species – can also self-amplify before eventually producing neurons. The extent of proliferation of all these neural progenitors is also influenced by their cell cycle length where a short cell cycle leads to more cycles of symmetric divisions, a delayed onset of neurogenesis, and subsequently to more neurons and a bigger cortex. Notably, proliferation of bRGCs at a particular cortical location is thought to be crucial to generate a cortical fold at this location. Hence, genes that influence the proliferation of these neural progenitors to evolutionary changes in brain size and folding.

The major focus in this respect has been on identifying and functionally characterizing genetic changes on the human or primate lineage. For example, the human-specific gene ARHGAP11B was found to induce bRGC proliferation and folding in cortices of mice, ferrets, and marmosets (***Florio et al., 2015***; ***Kalebic et al., 2018***; ***Heide et al., 2020***). Other examples include an amino acid substitution specific to modern humans in *TKTL1* (***Pinson et al., 2022***), human-specific NOTCH2 paralogs (***Fiddes et al., 2018***; ***Suzuki et al., 2018***), the primate-specific genes TMEM14B and TBC1D3 (***Liu et al., 2017***; ***Ju et al., 2016***), and an enhancer of *FZD8*, a receptor of the Wnt pathway (***Boyd et al., 2015***). While mechanistically convincing, it is unclear whether the proposed evolutionary link can be generalized as only one evolutionary lineage is investigated. Conversely, comparative approaches that correlate sequence changes with brain size changes have investigated more evolutionary lineages (***Boddy et al., 2017***; ***Montgomery et al., 2016***), but these studies lack mechanistic evidence and are limited to the analysis of protein-coding regions. Here, we combine mechanistic and phylogenetic approaches to study *TRNP1*, a gene that is known to be important for cortical growth and folding by influencing aRGC and bRGC proliferation and differentiation in mice (***Stahl et al., 2013***; ***Pilz et al., 2013***; ***Kerimoglu et al., 2021***) and ferrets (***Martínez-Martínez et al., 2016***).

On a cellular level, expressing *Trnp1* in neural stem cells (NSCs) isolated from mouse cortices induces phase separation, accelerates mitosis, and increases proliferation (***Stahl et al., 2013***; ***Esgleas et al., 2020***). Increasing *Trnp1* expression by in utero electroporation in mice and ferrets (embryonic day 13 [E13] in mice) leads to increased proliferation of aRGCs (***Stahl et al., 2013***; ***Martínez-Martínez et al., 2016***). Decreasing *Trnp1* expression levels in mice or ferrets (E13) reduces aRGC proliferation, increases their differentiation into BPs, and induces cortical folding (***Stahl et al., 2013***; ***Pilz et al., 2013***; ***Martínez-Martínez et al., 2016***). Notably, increasing *Trnp1* expression levels by in utero electroporation at E14.5 increases bRGC proliferation (***Kerimoglu et al., 2021***) and also induces cortical folding.

Hence, *Trnp1* levels can alter proliferation and differentiation of neural progenitors and in turn alter brain size and folding in mice and ferrets. However, whether genetic changes in *TRNP1* did alter cortical size and folding during mammalian evolution is unclear. Here, we analyse the evolution of *TRNP1* regulatory and coding sequences across mammals and investigate their link to the evolution of brain size and cortical folding.

Results

TRNP1 amino acid substitution rates co-evolve with rates of change in brain size and cortical folding in mammals

We experimentally and computationally collected (***Camacho et al., 2009***) and aligned (***Löytynoja, 2021***) 45 mammalian TRNP1 coding sequences, including dolphin and 18 primates (99.0% completeness, ***Figure 1—figure supplement 1A***). For 30 of those species, we could also compile estimates for brain size and cortical folding, as well as body mass as a potentially confounding parameter (***Figure 1A***; ***Supplementary file 1c***). We quantify brain size as its weight and cortical folding as the ratio of the cortical surface over the perimeter of the brain surface, the gyrification index (GI), where a GI $= 1$ indicates a completely smooth brain and a GI gt_1 indicates higher levels of cortical folding (***Zilles et al., 1989***). This phenotypic data together with the coding sequences are the basis for our investigation in the evolutionary relation between the rate of TRNP1 protein evolution and the evolution of brain size and gyrification.

The ratio of the non-synonymous (non-neutral) and the synonymous substitution rates, ω, is easily accessible and hence one of the most widespread measures of selection on protein-coding sequences, despite its limitations (***Yang, 2006***; ***Nei et al., 2000***). In the absence of additional evidence, only an $\omega > 1$ can be interpreted as proof of positive selection. However, an ωgt_1 requires many recurrent selective events and hence is underpowered to detect moderate amounts of positive selection. Therefore, it has become common practice to identify increases of ω on certain branches or subtrees relative to the remainder of the tree. For our question, we are analyzing the variation of ω across branches. To this end, we use the software Coevol that allows estimating the co-variance between rates of phenotypic and evolutionary sequence changes (ω), while both types of information go into the optimization of branch length estimates of the underlying phylogenetic tree (***Lartillot and Poujol, 2011***). This allows to detect a correlation between the strength of selection (ω) and a phenotypic trait. The question remains whether this correlation is directly caused by selection on that trait, or what we observe are indirect effects. This is not uncommon, because the strength of selection depends on the effective population size (N_e) of a species, which is often linked to life history traits and body size (***Ohta, 1987***; ***Lynch and Walsh, 2007***). For example, species with a large body size tend to have a small N_e and thus a low efficacy of selection (***Figuet et al., 2016***; ***Lartillot and Poujol, 2011***). With purifying selection being the dominant force in protein sequence evolution, we would thus expect a positive correlation between ω and body size due to indirect effects of N_e. However, in contrast to directed selection on one trait which is targeted to specific genes, a lower efficacy in purifying selection due to N_e will have an impact on all genes.

Therefore, we compiled a set of control genes in the same 30 species for which we have *TRNP1* sequences and phenotypic data. We started with all human autosomal genes that – as TRNP1 – have only one coding exon (*n*=1997; Human CCDS; ***Pujar et al., 2018***) and a similar length (*n*=1088; 291–999 bp vs. 682 bp of TRNP1). For 133 (12.3%) of these we could find full-length high-quality one-to-one orthologous sequences for all 30 species (***Figure 1—figure supplement 3A***; ***Supplementary file 1f***; Materials and methods). To ensure the quality of the resulting multiple sequence alignments, all of them were manually inspected. Based on the overall tree length we removed one outlier ($\sigma_{log(dS)} > 3$) leaving us with 132 control proteins that are well comparable to TRNP1 with respect to tree length, alignment quality, and ω (***Figure 1—figure supplement 3B***). Eight rather conserved genes (six with ω<0.04 and two with ω<0.19) did not show an acceptable parameter convergence between runs of Coevol, leaving 124 control genes well comparable to TRNP1 (***Supplementary file 1f***). If a species such as human or dolphin evolved a large, gyrified brain due to positive selection on TRNP1, we expect those lineages to show an increased rate of phenotype (brain size and GI) change and an increased ω. If this pattern is consistent across the majority of branches, Coevol would infer a

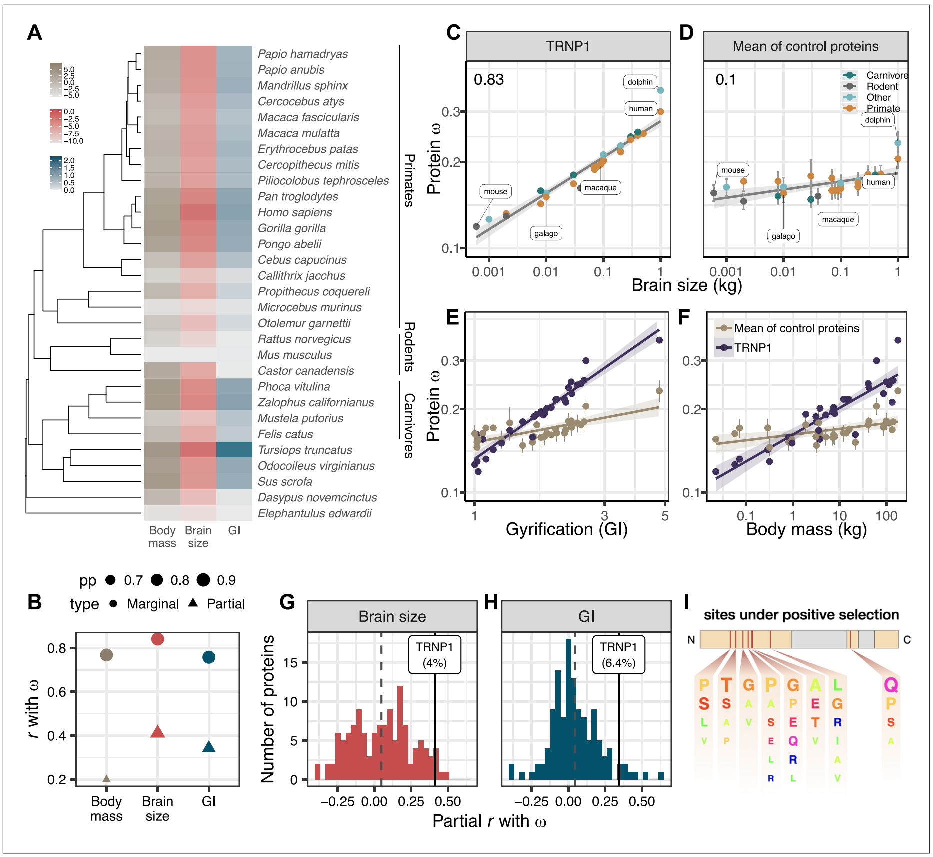


Figure 1. TRNP1 amino acid substitution rates co-evolve with brain size and cortical folding in mammals. (**A**) Mammalian species for which body mass, brain size, gyrification index (GI) measurements, and TRNP1 coding sequences were available (n=30)(***Figure 1—figure supplement 1***). Log2-transformed units: body mass and brain size in kg; GI is a ratio (cortical surface/perimeter of the brain surface). (**B**) Estimated marginal and partial correlation between ω of TRNP1 and the three traits using Coevol (***Lartillot and Poujol, 2011***). Size indicates posterior probability (pp). (**C**) TRNP1 protein substitution rates (ω) significantly correlate with brain size ($r = 0.83,\ pp$ = 0.97).(**D**) The average correlation across 124 control proteins with brain size ($\bar{r}$=0.10). (**E**) TRNP1 ω correlation with GI compared to the average across control proteins. (**F**) TRNP1 ω correlation with body mass compared to the average across control proteins. (**C, D, E, F**) Error bars indicate standard errors. (**G**) Distribution of partial correlations between ω and brain size of the control proteins and TRNP1. (**H**) Distribution of partial correlations between ω and GI of the control proteins and TRNP1. (**I**) Scheme of the mouse TRNP1 protein (223 amino acids [AAs]) with intrinsically disordered regions (orange) and sites (red lines) subject to positive selection in mammals (ω > 1, $pp > 0.95$***Figure 1—figure supplement 1***). Letter size of the depicted AAs represents the abundance of AAs at the positively selected sites.

The online version of this article includes the following figure supplement(s) for figure 1:

Figure supplement 1. TRNP1 protein-coding sequence analysis.

Figure supplement 2. Estimated marginal (**A**) and partial (**B**) correlation matrices of the combined Coevol model including the three traits and substitution rates of TRNP1.

Figure supplement 3. Control protein evolution rate correlation with brain size, gyrification, and body mass.

positive correlation between ω and the trait. Moreover, if this correlation is stronger than that for the average control protein, we can exclude that this is solely due to variation in the efficacy of selection.

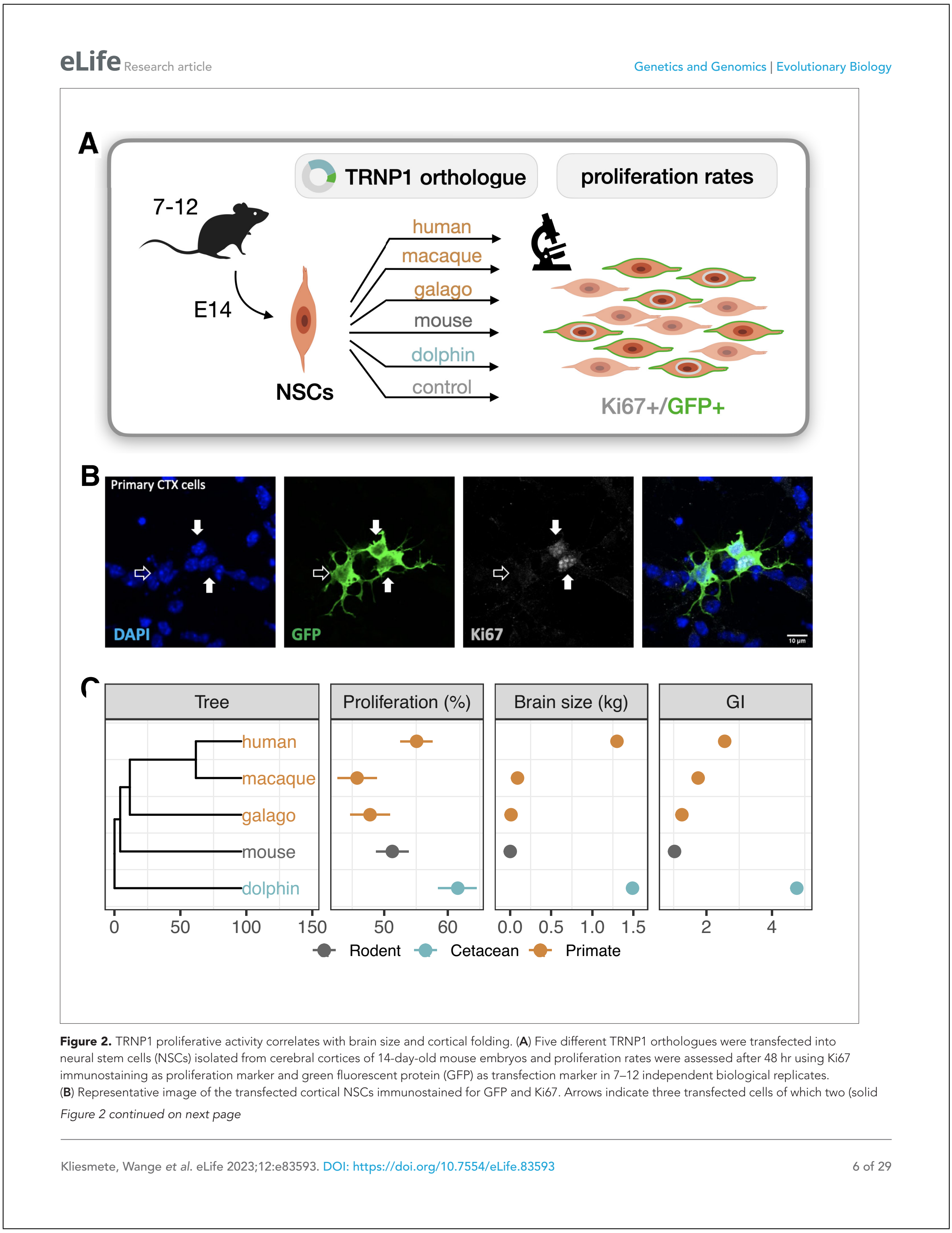
Indeed, we find that ω of TRNP1 positively correlates with brain size (*r*=0.83; p=0.97), GI (*r*=0.75; p=0.98), and also body mass (*r*=0.76; p=0.97) and that these correlations are stronger than those of the average control protein (***Figure 1C–F***, ***Figure 1—figure supplement 3C***), showing that the interaction between TRNP1 and the phenotypes goes beyond pure efficacy of selection effects. All three traits are highly correlated with one another. It is well known that brain and body size are not independent, and the same is true for GI and brain size (***Montgomery et al., 2016***; ***Smaers et al., 2021***). To disentangle which trait is most likely to be causal for the observed correlation with ω, we compare the partial correlations and find that brain size has the highest partial correlation (*r*=0.4), followed by GI (*r*=0.34), while the partial correlation with body mass (*r*=0.19) has a much larger drop compared to the marginals (***Figure 1B***, ***Figure 1—figure supplement 3C***), making selection on brain size and/or GI the more likely causes for the variation in ω. This said, TRNP1 is unlikely to be the sole evolutionary modifier of such an important and complex phenotype as brain size and gyrification. Because our control proteins represent a random selection of genes that based on sequence properties should give us comparable power to detect a link to these phenotypes, we can use the distribution of partial correlations of ω of the controls with brain size and GI to gauge the relative importance of TRNP1 for brain evolution (***Figure 1G and H***; ***Supplementary file 1g***). We find that TRNP1 protein evolution is among 4.0% and 6.4% of the most correlated proteins for brain size and GI, respectively.

Having established that the rate of protein evolution of TRNP1 is linked to brain size evolution, we now want to pinpoint the relevant sites or domains in the protein to facilitate further functional studies. Using the site model of PAML (***Yang, 1997***), we find 9.8% of the codons to show signs of recurrent positive selection (i.e., $\omega > 1$, site models M8 vs. M7, χ^2-value <0.001, *df* = 2). Eight codons with a selection signature could be pinpointed with high confidence (***Supplementary file 1d***). Seven out of those eight reside within the first intrinsically disordered region (IDR) and one in the second IDR of the protein (***Figure 1I***; ***Figure 1—figure supplement 1B***). The IDRs of TRNP1 are thought to mediate homotypic and heterotypic protein-protein interactions and are relevant for TRNP1-dependent phase separation, nuclear compartment size regulation, and M-phase length regulation (***Esgleas et al., 2020***). Hence, the positively selected sites indicate that these IDR-mediated TRNP1 functions were repeatedly adapted during mammalian evolution and the identified sites are candidates for further functional studies.

TRNP1 proliferative activity co-evolves with brain size and cortical folding in mammals

Next, we investigated whether the correlation between TRNP1 protein evolution and cortical phenotypes can be linked to functional properties of TRNP1 at a cellular level. A central property of TRNP1 is to promote proliferation of aRGC (***Stahl et al., 2013***; ***Esgleas et al., 2020***) and also of BPs (***Kerimoglu et al., 2021***). This proliferative activity can be assessed in an in vitro assay in which *TRNP1* is transfected into NSCs isolated from E14 mouse cortices (***Stahl et al., 2013***; ***Esgleas et al., 2020***).

To compare TRNP1 orthologues in this assay, we synthesized and cloned the TRNP1 coding sequence of human, rhesus macaque, galago, mouse, and dolphin that cover the observed range of ω (***Figure 1C***). After co-transfection with green fluorescent protein (GFP), we quantified the number of proliferating (Ki67+, GFP+) over all transfected (GFP+) NSCs for each *TRNP1* orthologue in ≥7 replicates (***Figure 2A and B***). We confirmed that *TRNP1* transfection does increase proliferation compared to a GFP-only control (p-value $< 2 \times 10^{-16}$; ***Figure 2—figure supplement 1A***) as shown in previous studies (***Stahl et al., 2013***; ***Esgleas et al., 2020***). Remarkably, the proportion of proliferating cells was highest in cells transfected with dolphin TRNP1 followed by human, which was significantly higher than the two other primates, galago and macaque (***Figure 2C***; ***Figure 2—figure supplement 1B***; ***Supplementary file 2a-c***). Indeed, the proliferative activity of TRNP1 is a significant predictor for brain size (BH-adjusted p-value = 0.0018, $R^2 = 0.89$) and GI (BH-adjusted p-value = 0.016, $R^2 = 0.69$) of its species of origin (phylogenetic generalized least squares [PGLS], likelihood ratio test [LRT]; ***Figure 2C***). Note that the three primates and the dolphin are phylogenetically equally distant to the mouse (***Figure 2C***) and hence a bias due to the murine assay system cannot explain the observed correlations with brain size and GI. Hence, these results further support that the TRNP1 protein co-evolves with brain size and cortical folding.



eLife Research article Genetics and Genomics | Evolutionary Biology

Figure 2. TRNP1 proliferative activity correlates with brain size and cortical folding. (**A**) Five different TRNP1 orthologues were transfected into neural stem cells (NSCs) isolated from cerebral cortices of 14-day-old mouse embryos and proliferation rates were assessed after 48 hr using Ki67 immunostaining as proliferation marker and green fluorescent protein (GFP) as transfection marker in 7–12 independent biological replicates. (**B**) Representative image of the transfected cortical NSCs immunostained for GFP and Ki67. Arrows indicate three transfected cells of which two (solid

Figure 2 continued on next page

Figure 2 continued

arrows) are Ki67-positive (***Figure 2—figure supplement 1***). (**C**) Induced proliferation in NSCs transfected with TRNP1 orthologues from five different species (***Supplementary file 2***). Proliferation rates are a significant predictor for brain size (χ^2=10.04, df = 1, BH-adjusted p-value = 0.0018 = 11.75 ± 2.412, R^2 = 0.89) and GI (χ^2=5.85, df = 1, BH-adjusted p-value = 0.016 = 16.97 ± 6.568, R^2 = 0.69) in the respective species (phylogenetic generalized least squares [PGLS], likelihood ratio test [LRT]). Error bars indicate standard errors. Included species: human (*Homo sapiens*), rhesus macaque (*Macaca mulatta*), northern greater galago (*Otolemur garnettii*), house mouse (*Mus musculus*), common bottlenose dolphin (*Tursiops truncatus*).

The online version of this article includes the following figure supplement(s) for figure 2:

Figure supplement 1. Proliferation induced by TRNP1.

Activity of a cis-regulatory element of *TRNP1* likely co-evolves with cortical folding in catarrhines

Experimental manipulation of *Trnp1* expression levels alters proliferation and differentiation of aRGC and bRGC in mice and ferrets (***Stahl et al., 2013***; ***Martínez-Martínez et al., 2016***; ***Kerimoglu et al., 2021***). Therefore, we next investigated whether changes in *TRNP1* regulation may also be associated with the evolution of cortical folding and brain size by analyzing co-variation in the activity of *TRNP1* associated cis-regulatory elements (CREs), using massively parallel reporter assays (MPRAs). To this end, a library of putative regulatory sequences is cloned into a reporter vector and their activity is quantified simultaneously by the expression levels of element-specific barcodes (***Inoue and Ahituv, 2015***). To identify putative CREs of *TRNP1*, we used DNase hypersensitive sites (DHS) from human foetal brain (***Bernstein et al., 2010***) and found three upstream CREs, the promoter-including exon 1, an intron CRE, one CRE overlapping the second exon, and one downstream CRE (***Figure 3A***). We obtained the orthologous sequences of the human CREs using a reciprocal best blat (RBB) strategy across additional mammalian species either from genome databases or by sequencing, yielding a total of 351 putative CREs in a panel of 75 mammalian species (***Figure 3—figure supplement 1***).

Due to limitations in the length of oligonucleotide synthesis, we split each orthologous putative CRE into highly overlapping, 94 bp fragments. The resulting 4950 sequence tiles were synthesized together with a barcode unique for each tile. From those, we constructed a complex and unbiased lentiviral plasmid library containing at least 4251 (86%) CRE sequence tiles (***Figure 3B and C***). Next, we stably transduced this library into neural progenitor cells (NPCs) derived from two humans and one cynomolgus macaque (***Geuder et al., 2021***). We calculated the activity per CRE sequence tile as the read-normalized reporter gene expression over the read-normalized input plasmid DNA (***Figure 3A***, Materials and methods). Finally, we use the per-tile activities (***Figure 3—figure supplement 2A***) to reconstruct the activities of the putative CREs. To this end, we summed all tile sequence activities for a given CRE while correcting for the built-in sequence overlap (***Figure 3D***; Materials and methods). CRE activities correlate well within the two human NPC lines and between the human and cynomolgus macaque NPC lines, indicating that the assay is robust across replicates and species (Pearson's r 0.85–0.88; ***Figure 3—figure supplement 2B***). The CREs covering exon 1, the intron, and the CRE downstream of *TRNP1* show the highest total activity across species while the CREs upstream of *TRNP1* show the lowest activity (***Figure 3E***).

Next, we tested whether CRE activity is associated with either brain size or GI across the 45 of the 75 mammalian species for which these phenotypes were available (***Figure 3D***). None of the CREs showed a significant association with brain size or GI (PGLS, LRT uncorrected p-value > 0.05) and only the intron CRE had a tendency to be positively associated with gyrification (PGLS, uncorrected LRT p-value=0.097, ***Figure 3F***, left; ***Supplementary file 3b***). Our power to detect such associations might be considerably lower than for coding sequences also because regulatory elements have a high turn-over rate (***Danko et al., 2018***; ***Berthelot et al., 2018***; ***Huber et al., 2020***). Hence, we expect that some orthologous DNA sequences that are CREs in one species do not function as CREs in others and can even be lost. The latter effect might explain why the sequences orthologous to human CREs are shorter in non-primate species more distantly related to humans (***Figure 3—figure supplement 1***). So phylogenetic comparisons of regulatory elements might be more powerful when restricted to species closely related to the species from which the CRE annotation is derived (humans in our case). Indeed, when we restrict our analysis to the catarrhine clade that encompasses Old World monkeys, great apes, and humans, the association between intron CRE activity and GI becomes considerably stronger (PGLS, uncorrected LRT p-value=0.003, Bonferroni-corrected for seven regions

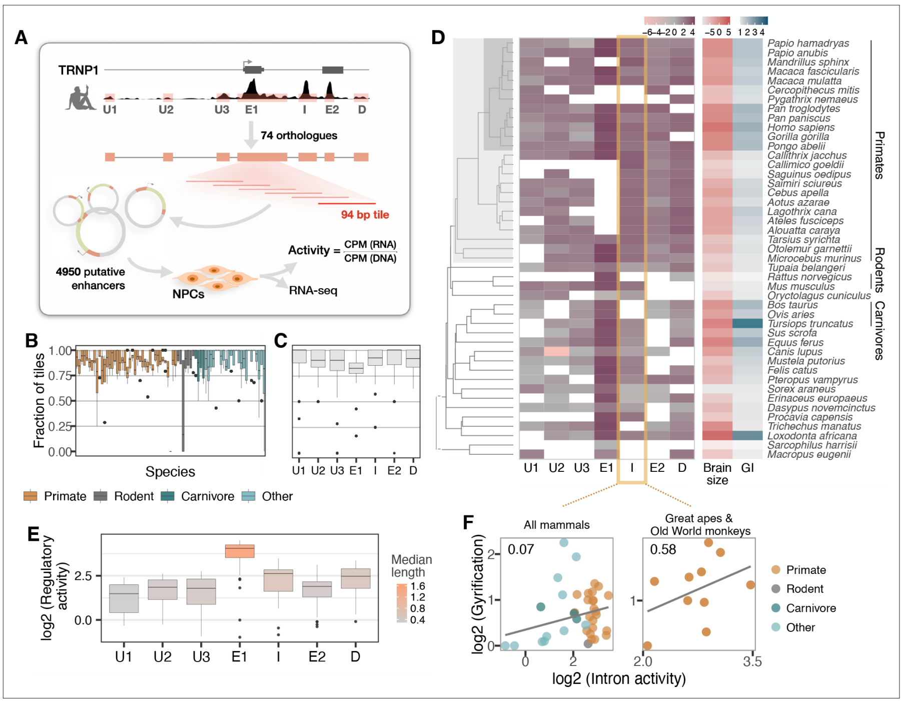


Figure 3. Activity of a cis-regulatory element (CRE) of *TRNP1* correlates with cortical folding in catarrhines. (**A**) Experimental setup of the massively parallel reporter assay (MPRA). Regulatory activity of seven putative TRNP1 CREs from 75 species were assayed in neural progenitor cells (NPCs) derived from human and cynomolgus macaque induced pluripotent stem cells. (***Figure 3—figure supplement 1***). (**B**) Fraction of the detected CRE tiles in the plasmid library per species across regions. The detection rates are unbiased and uniformly distributed across species and clades with only one extreme outlier *Dipodomys ordii*. (**C**) Fraction of the detected CRE tiles in the plasmid library per region across species. (**D**) Log-transformed total regulatory activity per CRE in human NPCs across species with available brain size and gyrification index (GI) measurements (*n*=45). (**E**) Total activity per CRE across species. Exon 1 (E1), intron (I), and the downstream (**D**) regions are more active and longer than other regions. (**B, C, E**) Each box represents the median and first and third quartiles with the whiskers indicating the furthest value no further than 1.5 * IQR from the box. Individual points indicate outliers. ***Figure 3—figure supplement 2*** (**F**) Regulatory activity of the intron CRE is weakly associated with gyrification across mammals (phylogenetic generalized least squares [PGLS], likelihood ratio test [LRT] p-value=0.097, R^2=0.07, *n*=37) and strongest across great apes and Old World monkeys, that is, catarrhines (PGLS, LRT p-value=0.003, R^2=0.58, *n*=10).

The online version of this article includes the following figure supplement(s) for figure 3:

Figure supplement 1. Length of the covered cis-regulatory element (CRE) sequences in the massively parallel reporter assay (MPRA) library across the tree.

Figure supplement 2. Analysis of massively parallel reporter assay (MPRA) data.

p-value=0.02, ***Figure 3F***, right; ***Supplementary file 3***). To validate that our model results are rather specific, we generated a null distribution for the observed correlation across catharrines, permuting the activities of all other CREs of this study. In agreement with our model results, we find 8/1000 (0.8%) of the random CRE combinations to have such a significant association of $p \leq 0.003$. Moreover, the intron CRE activity-GI association was consistently detected across all three cell lines including the cynomolgus macaque NPCs (***Supplementary file 3***). Furthermore, Reilly et al. compared enhancer activity by histone modifications in the developing cortex of humans, rhesus macaques, and mice and found a gain in activity on the human lineage in a region overlapping the intron CRE (***Reilly et al., 2015***). Thus, while the statistical evidence from our MPRA data alone is limited, we consider the

GI association in catarrhines together with the additional evidence from ***Reilly et al., 2015***, strong enough to warrant a more detailed analysis of the intron CRE.

Transcription factors with binding site enrichment on intron CREs regulate cell proliferation and are candidates to explain the observed activity across catarrhines

Reasoning that differences in CRE activities will likely be mediated by differences in their interactions with transcription factors (TF), we analysed the sequence evolution of putative TF binding sites (***Figure 4A***). First, we performed RNA-seq on the same samples that were used for the MPRA. Notably, also *TRNP1* was expressed (***Figure 4B***), supporting the relevance of our cellular system. Moreover, *TRNP1* expression was significantly higher in human NPC lines than that of cynomolgus macaque's (BH-adjusted p-value <0.05, ***Figure 4—figure supplement 1A–C***), consistent with higher intron CRE activity. Among the 392 expressed TFs with known binding motifs, we identified 22 with an excess of binding sites (***Frith et al., 2003***) within the catarrhine intron CRE sequences (***Figure 4B and D***). In agreement with TRNP1 itself being involved in the regulation of cell proliferation (***Volpe et al., 2006***; ***Stahl et al., 2013***; ***Esgleas et al., 2020***), these 22 TFs are enriched in biological processes regulating cell proliferation, neuron apoptotic process, and hormone levels (Gene Ontology, Fisher's exact p-value <0.05, background: 392 expressed TFs; ***Figure 4C***; ***Supplementary file 3***).

To further prioritize these 22 TFs, we used the motif binding scores in the 10 catarrhine intron CREs to predict the observed intron CRE activity in the MPRA and to predict the GI of the respective species. We found three TFs (CTCF, ZBTB26, SOX8) to be the best candidates to explain the variation in the intron CRE activity and one TF (CTCF) to co-vary with GI (PGLS, uncorrected LRT p-value <0.05, ***Figure 4D–F***). While the statistical support for this association is not strong, which is expected given that we were screening 22 candidate TFs in only 10 species, CTCF ChIP-seq data from the relevant cell types suggests that this particular CTCF binding site is indeed bound by CTCF in human NPCs (ChIP-seq, ***Encode Project Consortium, 2012***, ***Figure 4—figure supplement 2***). Moreover, HiC data show a topologically associated domain (TAD) boundary just upstream of *TRNP1* in the germinal zone of the developing human brain (postconception week 8, ***Won et al., 2016***). Hence, variations in the binding strength of CTCF across species might likely have consequences for the stability of the TAD boundary and *TRNP1* expression, affecting the associated phenotypes given its crucial role for brain development (***Stahl et al., 2013***).

In summary, we find a suggestive correlation between the activity of the intron CRE and gyrification in catarrhines, indicating that also regulatory changes of *TRNP1* might have contributed to the evolution of gyrification.

Discussion

Previous studies in mice and ferrets have elucidated mechanisms how Trnp1 is necessary for proliferation and differentiation of neural progenitors and how it could contribute to the evolution of brain size and cortical folding. We applied phylogenetic methods to explore associations between sequence and trait evolution and found that the rate of protein evolution and the proliferative activity of TRNP1 positively correlate with brain size and gyrification in mammals. Moreover, we find tentative evidence that the activity of a regulatory element in the intron of *TRNP1* might be associated with gyrification in catarrhines. At the sequence level, such a correlation could also be caused by confounding factors that affect the efficacy of natural selection such as the effective population size (***Ohta, 1987***; ***Lynch and Walsh, 2007***). However, body size – a reasonable proxy for effective population size (***Figuet et al., 2016***; ***Lartillot and Poujol, 2011***) – correlates much less with TRNP1 protein evolution than brain size or gyrification. Even more convincingly, the correlation of TRNP1 with brain size and gyrification is much stronger than the average correlation of these traits with the evolution of other proteins, that would have had to experience the same population size changes. Furthermore, it is unclear how an increased proliferative activity of TRNP1 or an increased CRE activity could be caused by a reduced efficacy of selection or other confounding factors. Together with the known role of TRNP1 in brain development, we think that the observed correlations are best interpreted as co-evolution of TRNP1 activity with brain size and gyrification, that is, that more active TRNP1 alleles were selected because they were advantageous to increase brain size and/or gyrification.

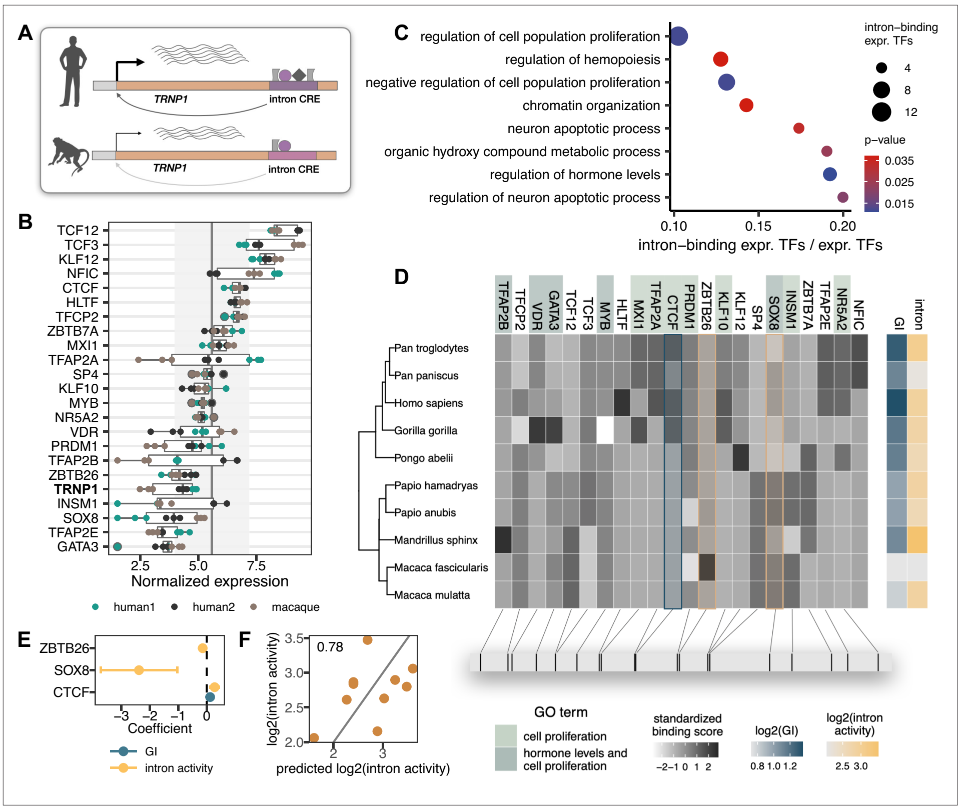


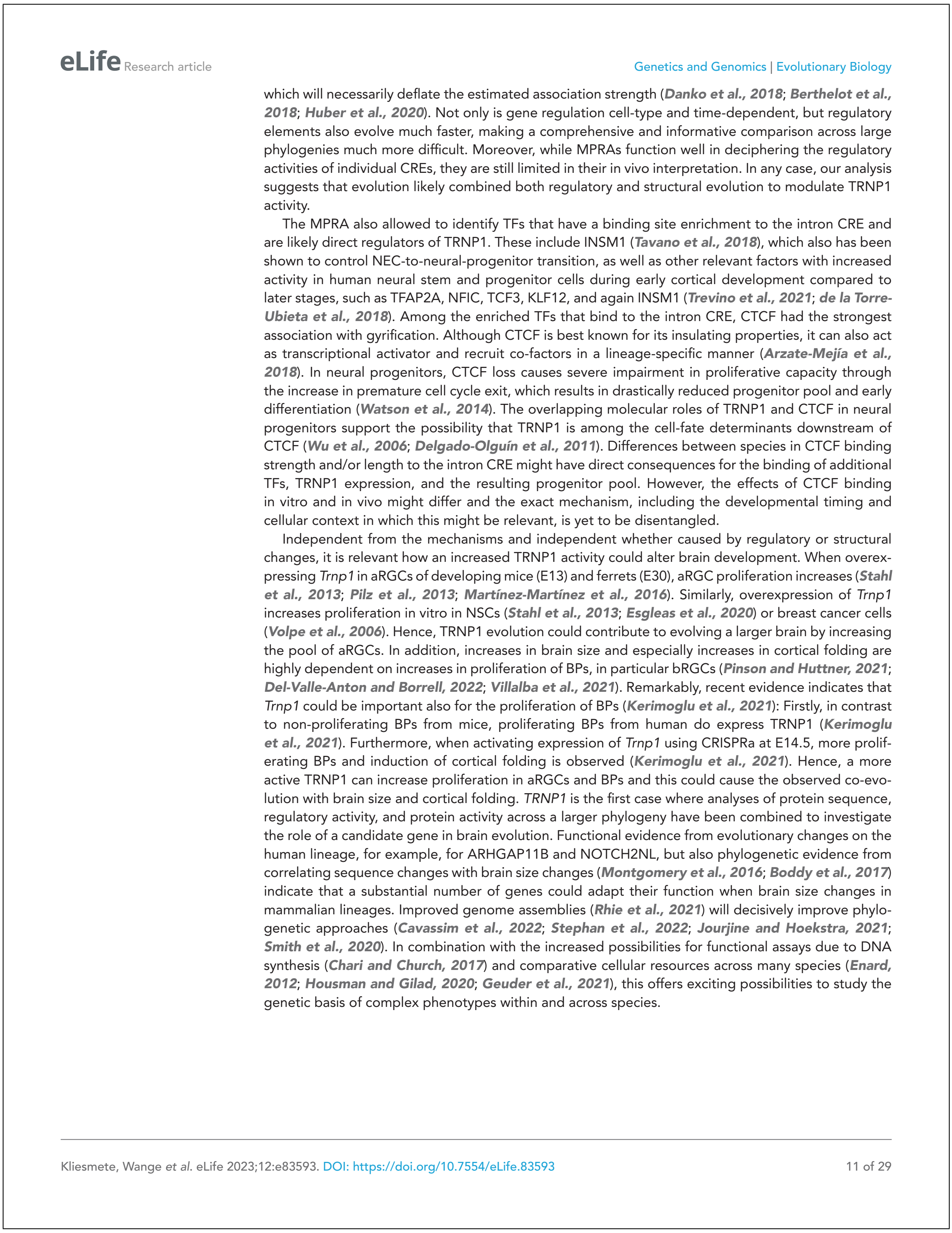
Figure 4. Transcription factors (TFs) with binding site enrichment on intron cis-regulatory elements (CREs) regulate cell proliferation and are candidates to explain the observed activity across catarrhines. (**A**) Orthologous intron CRE sequences show different regulatory activities under the same cellular conditions, suggesting variation in cis regulation across species. (**B**) Variance-stabilized expression in neural progenitor cells (NPCs) of *TRNP1* and the 22 TFs with enriched binding sites (motif weight ≥ 1) on the intron CREs. Each box represents the median, first and third quartiles with the whiskers indicating the furthest value no further than 1.5 * IQR from the box. Points indicate individual expression values. Vertical line indicates average expression across all 392 TFs (5.58), grey area: standard deviation (1.61). (**C**) Eight top enriched biological processes (Gene Ontology, Fisher's exact test p-value <0.05) of the 22 TFs. Background: all expressed TFs (392). (**D**) Variation in binding scores of the enriched TFs across catarrhines. Heatmaps indicate standardized binding scores (grey), gyrification index (GI) values (blue) and intron CRE activities (yellow) from the respective species. TF background colour indicates gene ontology assignment of the TFs to the two most significant biological processes. The bottom panel indicates the spatial position of the top binding site (motif score >3) for each TF on the human sequence. (**E**) Binding scores of three TFs (CTCF, ZBTB26, SOX8) are the best candidates to explain intron CRE activity, whereas only CTCF binding shows an association with the GI (phylogenetic generalized least squares [PGLS], likelihood ratio test [LRT] p-value <0.05). (**F**) Predicted intron CRE activity by the binding scores of the three TFs vs. the measured intron CRE activity across catarrhines.

The online version of this article includes the following figure supplement(s) for figure 4:

Figure supplement 1. TRNP1 expression in human and cynomolgus macaque (*Macaca fascicularis*) cell lines.

Figure supplement 2. Human genome tracks for the TRNP1 locus (hg19).

Of note, the effect of structural changes appears stronger than the effect of regulatory changes. This is contrary to the notion that regulatory changes should be the more likely targets of selection as they are more cell-type specific (***Carroll, 2008***) (but see also ***Hoekstra and Coyne, 2007***). However, current measures of regulatory activity are inherently less precise than counting amino acid changes,



which will necessarily deflate the estimated association strength (***Danko et al., 2018***; ***Berthelot et al., 2018***; ***Huber et al., 2020***). Not only is gene regulation cell-type and time-dependent, but regulatory elements also evolve much faster, making a comprehensive and informative comparison across large phylogenies much more difficult. Moreover, while MPRAs function well in deciphering the regulatory activities of individual CREs, they are still limited in their in vivo interpretation. In any case, our analysis suggests that evolution likely combined both regulatory and structural evolution to modulate TRNP1 activity.

The MPRA also allowed to identify TFs that have a binding site enrichment to the intron CRE and are likely direct regulators of TRNP1. These include INSM1 (***Tavano et al., 2018***), which also has been shown to control NEC-to-neural-progenitor transition, as well as other relevant factors with increased activity in human neural stem and progenitor cells during early cortical development compared to later stages, such as TFAP2A, NFIC, TCF3, KLF12, and again INSM1 (***Trevino et al., 2021***; ***de la Torre-Ubieta et al., 2018***). Among the enriched TFs that bind to the intron CRE, CTCF had the strongest association with gyrification. Although CTCF is best known for its insulating properties, it can also act as transcriptional activator and recruit co-factors in a lineage-specific manner (***Arzate-Mejía et al., 2018***). In neural progenitors, CTCF loss causes severe impairment in proliferative capacity through the increase in premature cell cycle exit, which results in drastically reduced progenitor pool and early differentiation (***Watson et al., 2014***). The overlapping molecular roles of TRNP1 and CTCF in neural progenitors support the possibility that TRNP1 is among the cell-fate determinants downstream of CTCF (***Wu et al., 2006***; ***Delgado-Olguín et al., 2011***). Differences between species in CTCF binding strength and/or length to the intron CRE might have direct consequences for the binding of additional TFs, TRNP1 expression, and the resulting progenitor pool. However, the effects of CTCF binding in vitro and in vivo might differ and the exact mechanism, including the developmental timing and cellular context in which this might be relevant, is yet to be disentangled.

Independent from the mechanisms and independent whether caused by regulatory or structural changes, it is relevant how an increased TRNP1 activity could alter brain development. When overexpressing *Trnp1* in aRGCs of developing mice (E13) and ferrets (E30), aRGC proliferation increases (***Stahl et al., 2013***; ***Pilz et al., 2013***; ***Martínez-Martínez et al., 2016***). Similarly, overexpression of *Trnp1* increases proliferation in vitro in NSCs (***Stahl et al., 2013***; ***Esgleas et al., 2020***) or breast cancer cells (***Volpe et al., 2006***). Hence, TRNP1 evolution could contribute to evolving a larger brain by increasing the pool of aRGCs. In addition, increases in brain size and especially increases in cortical folding are highly dependent on increases in proliferation of BPs, in particular bRGCs (***Pinson and Huttner, 2021***; ***Del-Valle-Anton and Borrell, 2022***; ***Villalba et al., 2021***). Remarkably, recent evidence indicates that *Trnp1* could be important also for the proliferation of BPs (***Kerimoglu et al., 2021***): Firstly, in contrast to non-proliferating BPs from mice, proliferating BPs from human do express TRNP1 (***Kerimoglu et al., 2021***). Furthermore, when activating expression of *Trnp1* using CRISPRa at E14.5, more proliferating BPs and induction of cortical folding is observed (***Kerimoglu et al., 2021***). Hence, a more active TRNP1 can increase proliferation in aRGCs and BPs and this could cause the observed co-evolution with brain size and cortical folding. *TRNP1* is the first case where analyses of protein sequence, regulatory activity, and protein activity across a larger phylogeny have been combined to investigate the role of a candidate gene in brain evolution. Functional evidence from evolutionary changes on the human lineage, for example, for ARHGAP11B and NOTCH2NL, but also phylogenetic evidence from correlating sequence changes with brain size changes (***Montgomery et al., 2016***; ***Boddy et al., 2017***) indicate that a substantial number of genes could adapt their function when brain size changes in mammalian lineages. Improved genome assemblies (***Rhie et al., 2021***) will decisively improve phylogenetic approaches (***Cavassim et al., 2022***; ***Stephan et al., 2022***; ***Jourjine and Hoekstra, 2021***; ***Smith et al., 2020***). In combination with the increased possibilities for functional assays due to DNA synthesis (***Chari and Church, 2017***) and comparative cellular resources across many species (***Enard, 2012***; ***Housman and Gilad, 2020***; ***Geuder et al., 2021***), this offers exciting possibilities to study the genetic basis of complex phenotypes within and across species.

Materials and methods

Sample collection and cell culture

Mouse strain and handling

Mouse handling and experimental procedures were performed in accordance with German and European Union guidelines and were approved by the State of upper Bavaria. All efforts were made to minimize suffering and number of animals. Two- to three-month female C57BL/6J wild-type mice were maintained in specific pathogen-free conditions in the animal facility, in 12:12 hr light/dark cycles and bred under standard housing conditions in the animal facility of the Helmholtz Center Munich and the Biomedical Center Munich. The day of the vaginal plug was considered E0.

Primary cerebral cortex harvesting and culture

E14 mouse (*M. musculus*) cerebral cortices were dissected, removing the ganglionic eminence, the olfactory bulb, the hippocampal anlage, and the meninges. Cells were mechanically dissociated with a fire polish Pasteur pipette. Cells were then seeded onto poly-D-lysine (PDL)-coated glass coverslips in DMEM-GlutaMAX (Dulbeccos's modified Eagles's medium) supplemented with 10% foetal calf serum (FCS) and 100 µg/mL Pen. Strep. and cultured at 37°C in a 5% CO_2 incubator.

Culture of HEK293T cells

HEK 293T cells (*H. sapiens*) were grown in DMEM supplemented with 10% FCS and 1% Pen. Strep. Cells were cultured in 10 cm flat-bottom dishes at 37°C in a 5% CO_2 environment and split every 2–3 days in a 1:10 ratio using 5 mL PBS to wash and 0.5 mL 0.25% Trypsin to detach the cells.

Culture of Neuro-2A cells

Neuro-2A cells (N2A) (ATCC; CCL-131, *M. musculus*) were cultured in Eagle's minimum essential medium (Thermo Fisher Scientific) with 10% FCS (Thermo Fisher Scientific) at 37°C in a 5% CO_2 incubator and split every 2–3 days in a 1:5 ratio using 5 mL PBS (Thermo Fisher Scientific) to wash and 0.5 mL 0.25% Trypsin (Thermo Fisher Scientific) to detach the cells.

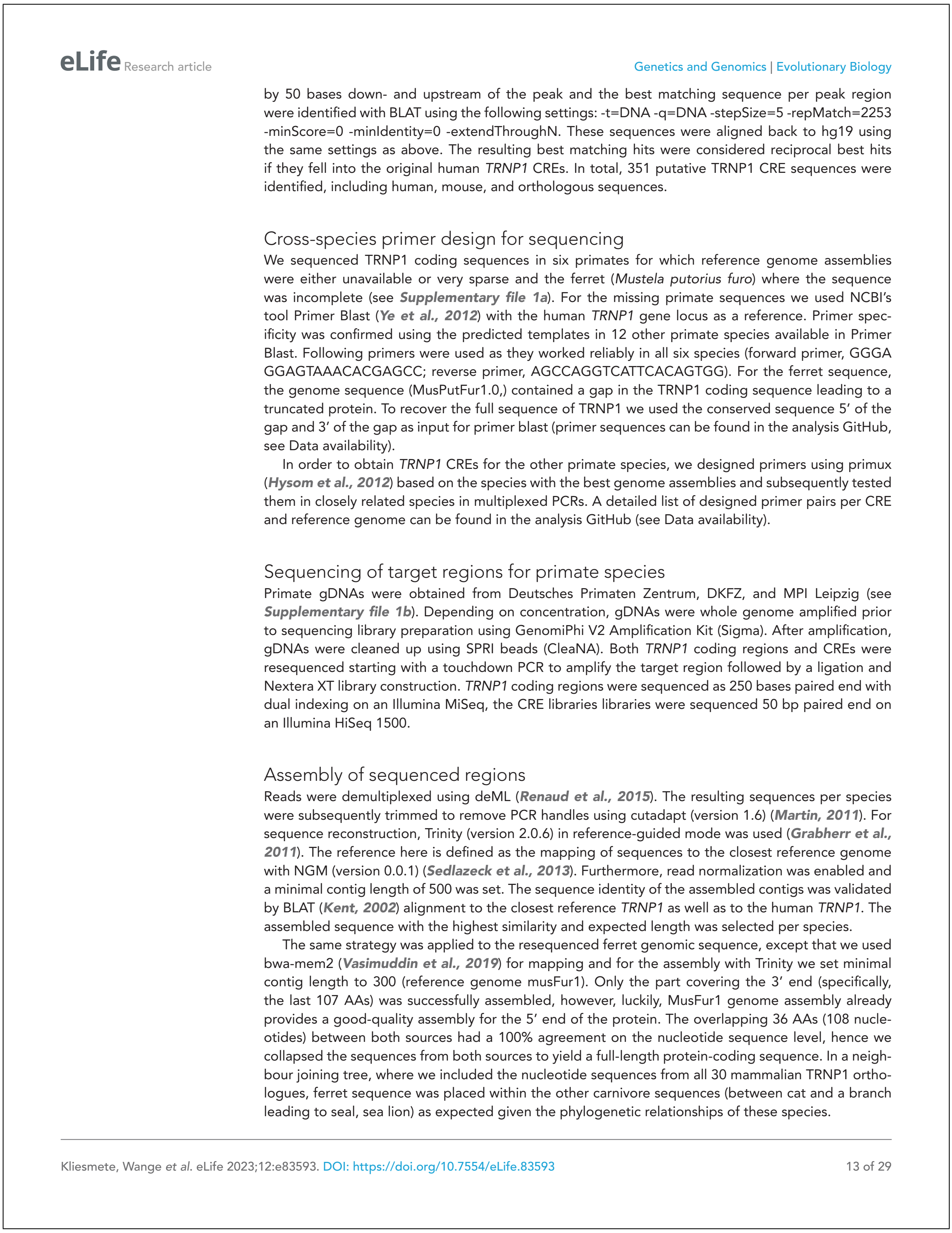
Culture of neural progenitor cells

Neural progenitor cells of two human (*H. sapiens*) and one cynomolgus monkey (*M. fascicularis*) cell line (***Geuder et al., 2021***) were cultured at 37°C in a 5% CO_2 incubator on Geltrex (Thermo Fisher Scientific) in DMEM F12 (Fisher Scientific) supplemented with 2 mM GlutaMAX-I (Fisher Scientific), 20 ng/µL bFGF (Peprotech), 20 ng/µL hEGF (Miltenyi Biotec), 2% B-27 supplement (50×) minus vitamin A (Gibco), 1% N2 supplement 100× (Gibco), 200 µM L-ascorbic acid 2-phosphate (Sigma), and 100 µg/mL penicillin-streptomycin (Pen. Strep.) with medium change every second day. For passaging, NPCs were washed with PBS and then incubated with TrypLE Select (Thermo Fisher Scientific) for 5 min at 37°C. Culture medium was added and cells were centrifuged at 200 × *g* for 5 min. Supernatant was replaced by fresh culture medium and cells were transferred to a new Geltrex-coated dish. The cells were split every 2–3 days in a ratio of 1:3. All cell lines have been authenticated using RNA sequencing (RNA-seq), see ***Geuder et al., 2021***, and the current study. Mycoplasma is regularly tested for using PCR-based test.

Sequencing of *TRNP1* for primate species

Identification of CREs of *TRNP1*

DHS in the proximity to *TRNP1* (25 kb upstream, 3 kb downstream) were identified in human foetal brain and mouse embryonic brain DNase-seq datasets (***Vierstra et al., 2014***; ***Bernstein et al., 2010***) downloaded from NCBI's Sequence Read Archive (see Appendix 1—key resources table). Reads were mapped to human genome version hg19 and mouse genome version mm10 using NextGenMap with default parameters (NGM; version 0.0.1) (***Sedlazeck et al., 2013***). Peaks were identified with Hotspot version 4.0.0 using default parameters (***John et al., 2011***). Overlapping peaks were merged, and the union per species was taken as putative CREs of *TRNP1* (***Supplementary file 3a***). The orthologous regions of human *TRNP1* DNase peaks in 49 mammalian species were identified with reciprocal best hit using BLAT (v. 35x1) (***Kent, 2002***). Firstly, sequences of human *TRNP1* DNase peaks were extended



by 50 bases down- and upstream of the peak and the best matching sequence per peak region were identified with BLAT using the following settings: -t=DNA -q=DNA -stepSize=5 -repMatch=2253 -minScore=0 -minIdentity=0 -extendThroughN. These sequences were aligned back to hg19 using the same settings as above. The resulting best matching hits were considered reciprocal best hits if they fell into the original human *TRNP1* CREs. In total, 351 putative TRNP1 CRE sequences were identified, including human, mouse, and orthologous sequences.

Cross-species primer design for sequencing

We sequenced TRNP1 coding sequences in six primates for which reference genome assemblies were either unavailable or very sparse and the ferret (*Mustela putorius furo*) where the sequence was incomplete (see ***Supplementary file 1a***). For the missing primate sequences we used NCBI's tool Primer Blast (***Ye et al., 2012***) with the human *TRNP1* gene locus as a reference. Primer specificity was confirmed using the predicted templates in 12 other primate species available in Primer Blast. Following primers were used as they worked reliably in all six species (forward primer, GGGA GGAGTAAACACGAGCC; reverse primer, AGCCAGGTCATTCACAGTGG). For the ferret sequence, the genome sequence (MusPutFur1.0,) contained a gap in the TRNP1 coding sequence leading to a truncated protein. To recover the full sequence of TRNP1 we used the conserved sequence 5' of the gap and 3' of the gap as input for primer blast (primer sequences can be found in the analysis GitHub, see Data availability).

In order to obtain *TRNP1* CREs for the other primate species, we designed primers using primux (***Hysom et al., 2012***) based on the species with the best genome assemblies and subsequently tested them in closely related species in multiplexed PCRs. A detailed list of designed primer pairs per CRE and reference genome can be found in the analysis GitHub (see Data availability).

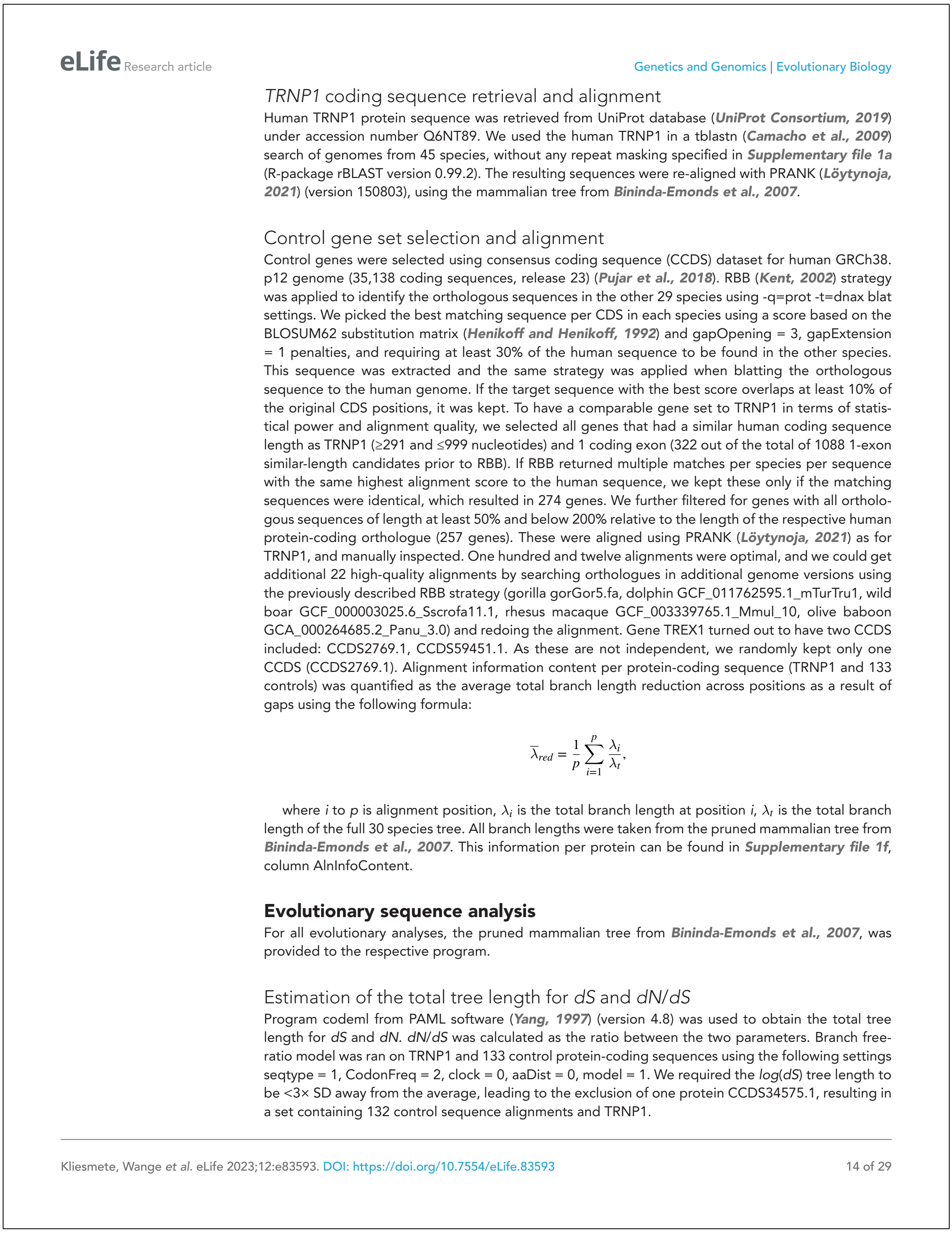
Sequencing of target regions for primate species

Primate gDNAs were obtained from Deutsches Primaten Zentrum, DKFZ, and MPI Leipzig (see ***Supplementary file 1b***). Depending on concentration, gDNAs were whole genome amplified prior to sequencing library preparation using GenomiPhi V2 Amplification Kit (Sigma). After amplification, gDNAs were cleaned up using SPRI beads (CleaNA). Both *TRNP1* coding regions and CREs were resequenced starting with a touchdown PCR to amplify the target region followed by a ligation and Nextera XT library construction. *TRNP1* coding regions were sequenced as 250 bases paired end with dual indexing on an Illumina MiSeq, the CRE libraries libraries were sequenced 50 bp paired end on an Illumina HiSeq 1500.

Assembly of sequenced regions

Reads were demultiplexed using deML (***Renaud et al., 2015***). The resulting sequences per species were subsequently trimmed to remove PCR handles using cutadapt (version 1.6) (***Martin, 2011***). For sequence reconstruction, Trinity (version 2.0.6) in reference-guided mode was used (***Grabherr et al., 2011***). The reference here is defined as the mapping of sequences to the closest reference genome with NGM (version 0.0.1) (***Sedlazeck et al., 2013***). Furthermore, read normalization was enabled and a minimal contig length of 500 was set. The sequence identity of the assembled contigs was validated by BLAT (***Kent, 2002***) alignment to the closest reference *TRNP1* as well as to the human *TRNP1*. The assembled sequence with the highest similarity and expected length was selected per species.

The same strategy was applied to the resequenced ferret genomic sequence, except that we used bwa-mem2 (***Vasimuddin et al., 2019***) for mapping and for the assembly with Trinity we set minimal contig length to 300 (reference genome musFur1). Only the part covering the 3' end (specifically, the last 107 AAs) was successfully assembled, however, luckily, MusFur1 genome assembly already provides a good-quality assembly for the 5' end of the protein. The overlapping 36 AAs (108 nucleotides) between both sources had a 100% agreement on the nucleotide sequence level, hence we collapsed the sequences from both sources to yield a full-length protein-coding sequence. In a neighbour joining tree, where we included the nucleotide sequences from all 30 mammalian TRNP1 orthologues, ferret sequence was placed within the other carnivore sequences (between cat and a branch leading to seal, sea lion) as expected given the phylogenetic relationships of these species.



TRNP1 coding sequence retrieval and alignment

Human TRNP1 protein sequence was retrieved from UniProt database (***UniProt Consortium, 2019***) under accession number Q6NT89. We used the human TRNP1 in a tblastn (***Camacho et al., 2009***) search of genomes from 45 species, without any repeat masking specified in ***Supplementary file 1a*** (R-package rBLAST version 0.99.2). The resulting sequences were re-aligned with PRANK (***Löytynoja, 2021***) (version 150803), using the mammalian tree from ***Bininda-Emonds et al., 2007***.

Control gene set selection and alignment

Control genes were selected using consensus coding sequence (CCDS) dataset for human GRCh38.p12 genome (35,138 coding sequences, release 23) (***Pujar et al., 2018***). RBB (***Kent, 2002***) strategy was applied to identify the orthologous sequences in the other 29 species using -q=prot -t=dnax blat settings. We picked the best matching sequence per CDS in each species using a score based on the BLOSUM62 substitution matrix (***Henikoff and Henikoff, 1992***) and gapOpening = 3, gapExtension = 1 penalties, and requiring at least 30% of the human sequence to be found in the other species. This sequence was extracted and the same strategy was applied when blatting the orthologous sequence to the human genome. If the target sequence with the best score overlaps at least 10% of the original CDS positions, it was kept. To have a comparable gene set to TRNP1 in terms of statistical power and alignment quality, we selected all genes that had a similar human coding sequence length as TRNP1 (≥291 and ≤999 nucleotides) and 1 coding exon (322 out of the total of 1088 1-exon similar-length candidates prior to RBB). If RBB returned multiple matches per species per sequence with the same highest alignment score to the human sequence, we kept these only if the matching sequences were identical, which resulted in 274 genes. We further filtered for genes with all orthologous sequences of length at least 50% and below 200% relative to the length of the respective human protein-coding orthologue (257 genes). These were aligned using PRANK (***Löytynoja, 2021***) as for TRNP1, and manually inspected. One hundred and twelve alignments were optimal, and we could get additional 22 high-quality alignments by searching orthologues in additional genome versions using the previously described RBB strategy (gorilla gorGor5.fa, dolphin GCF_011762595.1_mTurTru1, wild boar GCF_000003025.6_Sscrofa11.1, rhesus macaque GCF_003339765.1_Mmul_10, olive baboon GCA_000264685.2_Panu_3.0) and redoing the alignment. Gene TREX1 turned out to have two CCDS included: CCDS2769.1, CCDS59451.1. As these are not independent, we randomly kept only one CCDS (CCDS2769.1). Alignment information content per protein-coding sequence (TRNP1 and 133 controls) was quantified as the average total branch length reduction across positions as a result of gaps using the following formula:

$$\overline{\lambda}_{red} = \frac{1}{p}\sum_{i=1}^{p}\frac{\lambda_i}{\lambda_t},$$

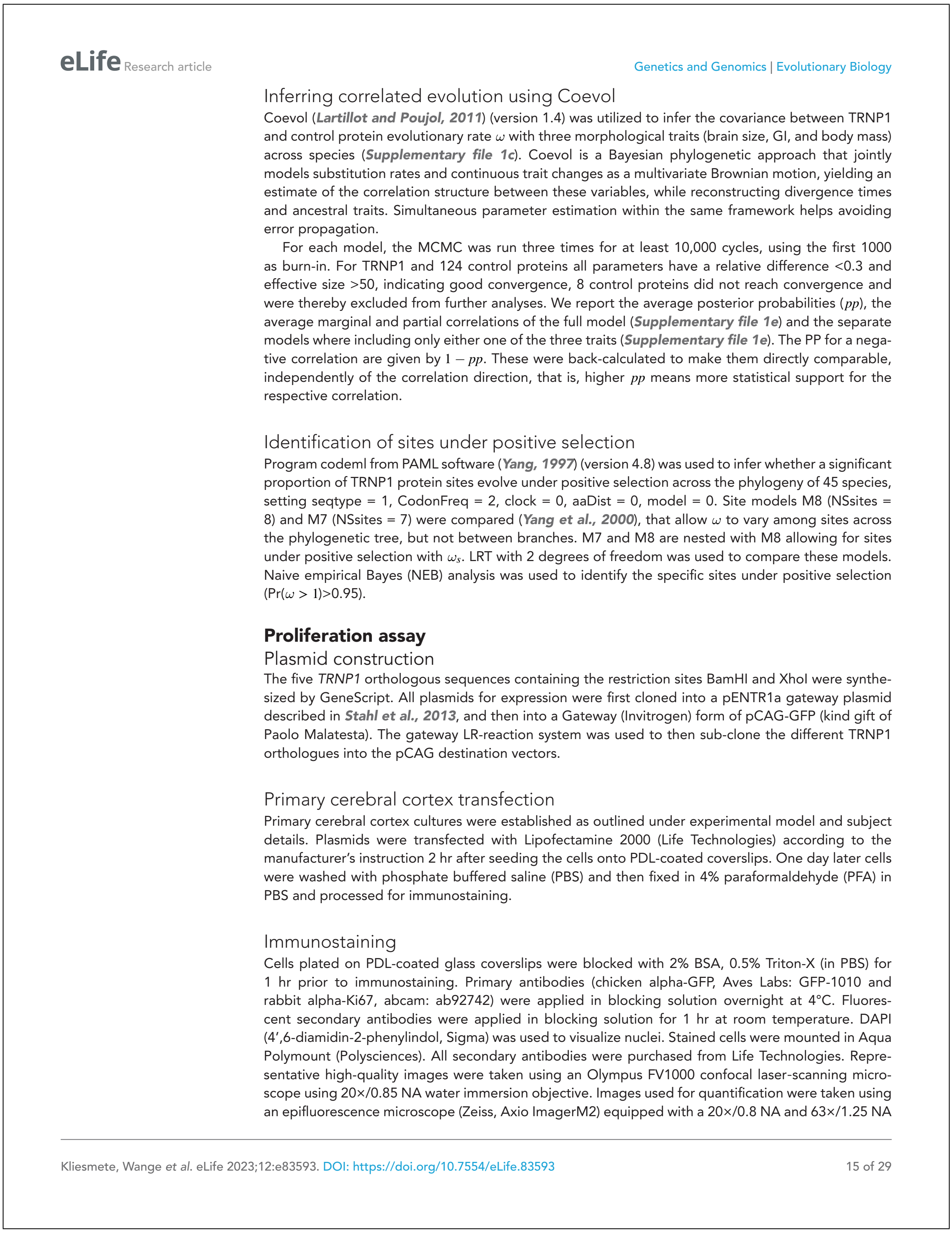
where i to p is alignment position, λ_i is the total branch length at position i, λ_t is the total branch length of the full 30 species tree. All branch lengths were taken from the pruned mammalian tree from ***Bininda-Emonds et al., 2007***. This information per protein can be found in ***Supplementary file 1f***, column AlnInfoContent.

Evolutionary sequence analysis

For all evolutionary analyses, the pruned mammalian tree from ***Bininda-Emonds et al., 2007***, was provided to the respective program.

Estimation of the total tree length for *dS* and *dN*/*dS*

Program codeml from PAML software (***Yang, 1997***) (version 4.8) was used to obtain the total tree length for *dS* and *dN*. *dN*/*dS* was calculated as the ratio between the two parameters. Branch free-ratio model was ran on TRNP1 and 133 control protein-coding sequences using the following settings seqtype = 1, CodonFreq = 2, clock = 0, aaDist = 0, model = 1. We required the *log*(*dS*) tree length to be <3× SD away from the average, leading to the exclusion of one protein CCDS34575.1, resulting in a set containing 132 control sequence alignments and TRNP1.



Inferring correlated evolution using Coevol

Coevol (***Lartillot and Poujol, 2011***) (version 1.4) was utilized to infer the covariance between TRNP1 and control protein evolutionary rate ω with three morphological traits (brain size, GI, and body mass) across species (***Supplementary file 1c***). Coevol is a Bayesian phylogenetic approach that jointly models substitution rates and continuous trait changes as a multivariate Brownian motion, yielding an estimate of the correlation structure between these variables, while reconstructing divergence times and ancestral traits. Simultaneous parameter estimation within the same framework helps avoiding error propagation.

For each model, the MCMC was run three times for at least 10,000 cycles, using the first 1000 as burn-in. For TRNP1 and 124 control proteins all parameters have a relative difference <0.3 and effective size >50, indicating good convergence, 8 control proteins did not reach convergence and were thereby excluded from further analyses. We report the average posterior probabilities (pp), the average marginal and partial correlations of the full model (***Supplementary file 1e***) and the separate models where including only either one of the three traits (***Supplementary file 1e***). The PP for a negative correlation are given by $1 - pp$. These were back-calculated to make them directly comparable, independently of the correlation direction, that is, higher pp means more statistical support for the respective correlation.

Identification of sites under positive selection

Program codeml from PAML software (***Yang, 1997***) (version 4.8) was used to infer whether a significant proportion of TRNP1 protein sites evolve under positive selection across the phylogeny of 45 species, setting seqtype = 1, CodonFreq = 2, clock = 0, aaDist = 0, model = 0. Site models M8 (NSsites = 8) and M7 (NSsites = 7) were compared (***Yang et al., 2000***), that allow ω to vary among sites across the phylogenetic tree, but not between branches. M7 and M8 are nested with M8 allowing for sites under positive selection with ω_S. LRT with 2 degrees of freedom was used to compare these models. Naive empirical Bayes (NEB) analysis was used to identify the specific sites under positive selection (Pr($\omega > 1$)>0.95).

Proliferation assay

Plasmid construction

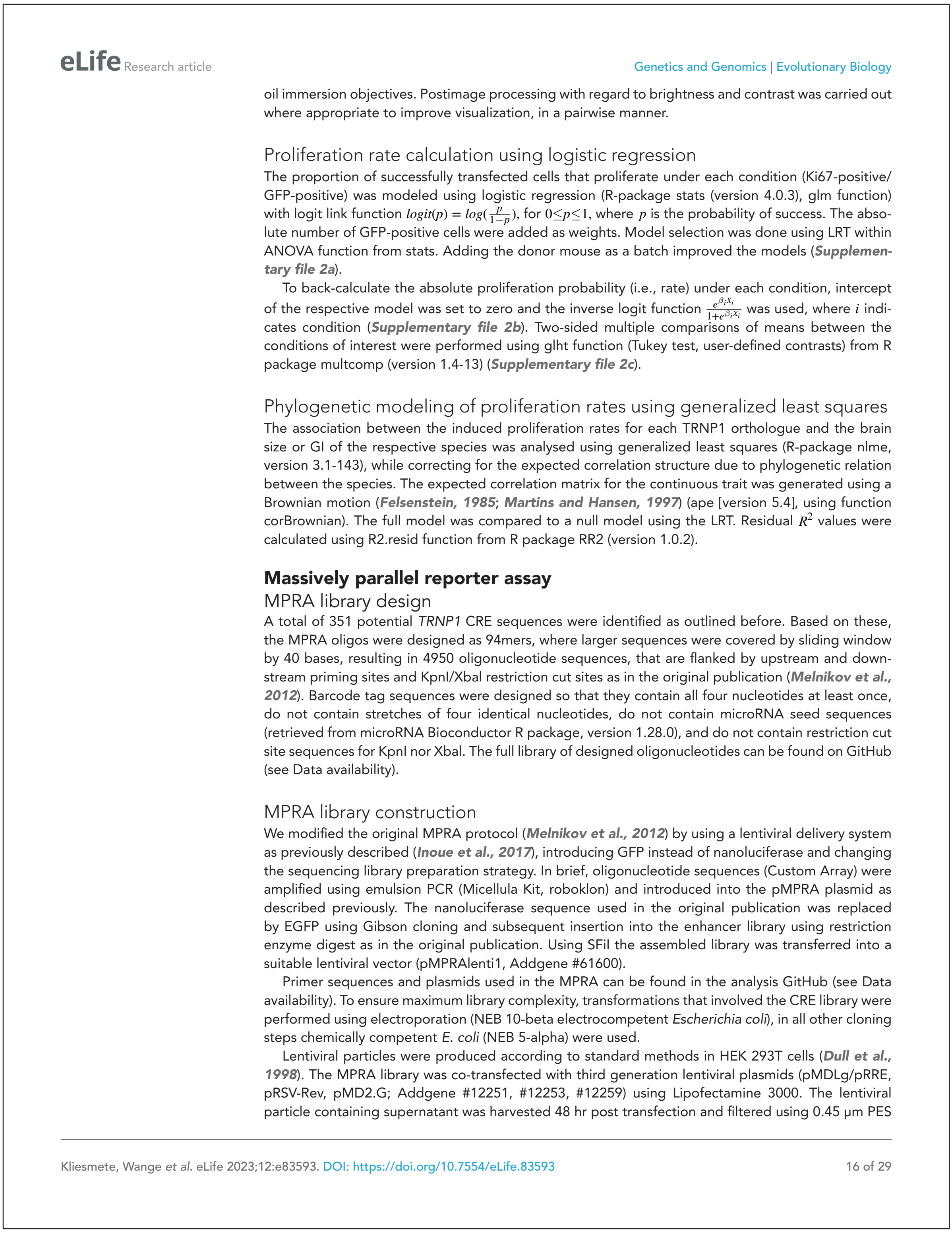
The five *TRNP1* orthologous sequences containing the restriction sites BamHI and XhoI were synthesized by GeneScript. All plasmids for expression were first cloned into a pENTR1a gateway plasmid described in ***Stahl et al., 2013***, and then into a Gateway (Invitrogen) form of pCAG-GFP (kind gift of Paolo Malatesta). The gateway LR-reaction system was used to then sub-clone the different TRNP1 orthologues into the pCAG destination vectors.

Primary cerebral cortex transfection

Primary cerebral cortex cultures were established as outlined under experimental model and subject details. Plasmids were transfected with Lipofectamine 2000 (Life Technologies) according to the manufacturer's instruction 2 hr after seeding the cells onto PDL-coated coverslips. One day later cells were washed with phosphate buffered saline (PBS) and then fixed in 4% paraformaldehyde (PFA) in PBS and processed for immunostaining.

Immunostaining

Cells plated on PDL-coated glass coverslips were blocked with 2% BSA, 0.5% Triton-X (in PBS) for 1 hr prior to immunostaining. Primary antibodies (chicken alpha-GFP, Aves Labs: GFP-1010 and rabbit alpha-Ki67, abcam: ab92742) were applied in blocking solution overnight at 4°C. Fluorescent secondary antibodies were applied in blocking solution for 1 hr at room temperature. DAPI (4',6-diamidin-2-phenylindol, Sigma) was used to visualize nuclei. Stained cells were mounted in Aqua Polymount (Polysciences). All secondary antibodies were purchased from Life Technologies. Representative high-quality images were taken using an Olympus FV1000 confocal laser-scanning microscope using 20×/0.85 NA water immersion objective. Images used for quantification were taken using an epifluorescence microscope (Zeiss, Axio ImagerM2) equipped with a 20×/0.8 NA and 63×/1.25 NA



oil immersion objectives. Postimage processing with regard to brightness and contrast was carried out where appropriate to improve visualization, in a pairwise manner.

Proliferation rate calculation using logistic regression

The proportion of successfully transfected cells that proliferate under each condition (Ki67-positive/GFP-positive) was modeled using logistic regression (R-package stats (version 4.0.3), glm function) with logit link function $logit(p) = log(\frac{p}{1-p})$, for $0 \leq p \leq 1$, where p is the probability of success. The absolute number of GFP-positive cells were added as weights. Model selection was done using LRT within ANOVA function from stats. Adding the donor mouse as a batch improved the models (***Supplementary file 2a***).

To back-calculate the absolute proliferation probability (i.e., rate) under each condition, intercept of the respective model was set to zero and the inverse logit function $\frac{e^{\beta_i X_i}}{1+e^{\beta_i X_i}}$ was used, where i indicates condition (***Supplementary file 2b***). Two-sided multiple comparisons of means between the conditions of interest were performed using glht function (Tukey test, user-defined contrasts) from R package multcomp (version 1.4-13) (***Supplementary file 2c***).

Phylogenetic modeling of proliferation rates using generalized least squares

The association between the induced proliferation rates for each TRNP1 orthologue and the brain size or GI of the respective species was analysed using generalized least squares (R-package nlme, version 3.1-143), while correcting for the expected correlation structure due to phylogenetic relation between the species. The expected correlation matrix for the continuous trait was generated using a Brownian motion (***Felsenstein, 1985***; ***Martins and Hansen, 1997***) (ape [version 5.4], using function corBrownian). The full model was compared to a null model using the LRT. Residual R^2 values were calculated using R2.resid function from R package RR2 (version 1.0.2).

Massively parallel reporter assay

MPRA library design

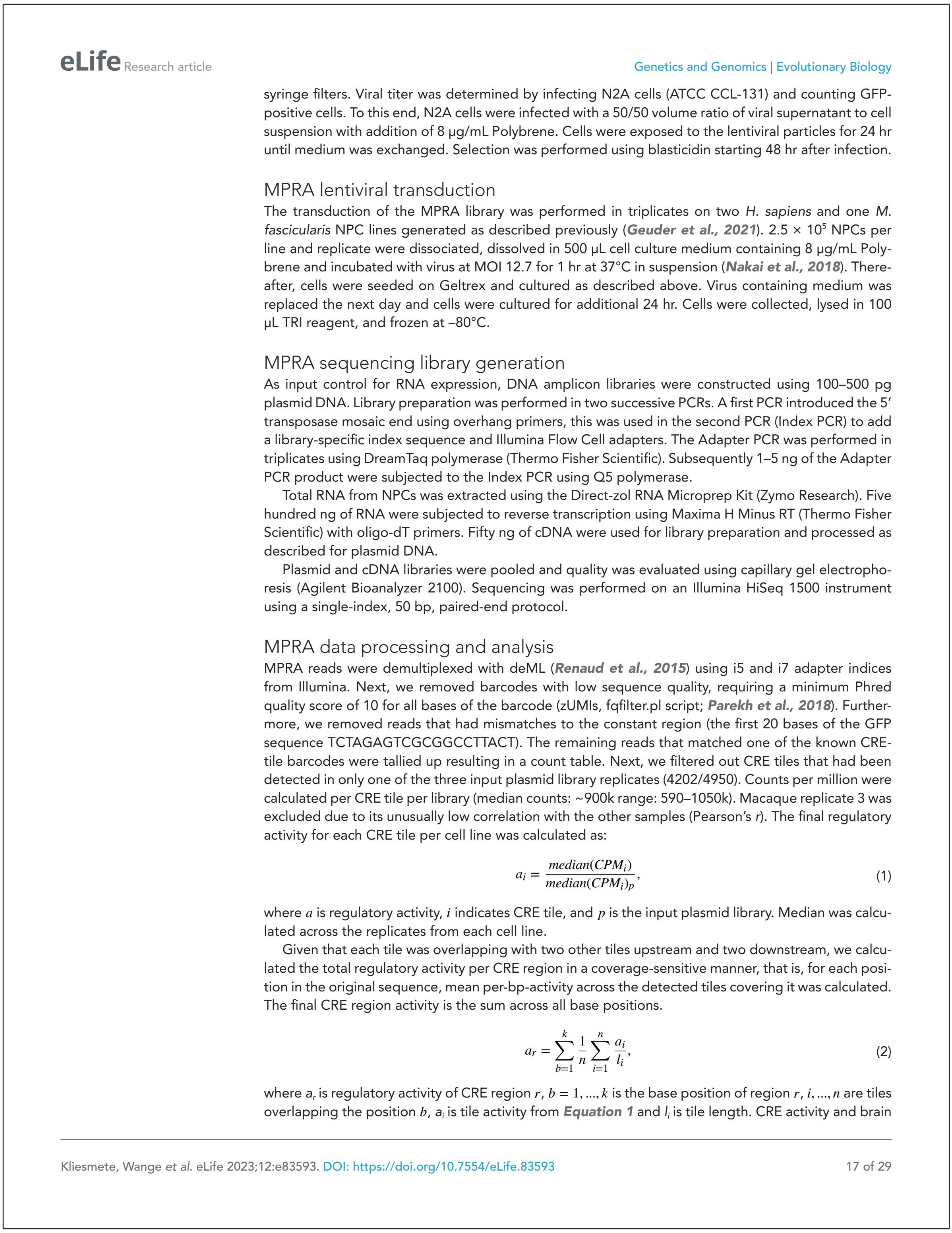
A total of 351 potential *TRNP1* CRE sequences were identified as outlined before. Based on these, the MPRA oligos were designed as 94mers, where larger sequences were covered by sliding window by 40 bases, resulting in 4950 oligonucleotide sequences, that are flanked by upstream and downstream priming sites and KpnI/XbaI restriction cut sites as in the original publication (***Melnikov et al., 2012***). Barcode tag sequences were designed so that they contain all four nucleotides at least once, do not contain stretches of four identical nucleotides, do not contain microRNA seed sequences (retrieved from microRNA Bioconductor R package, version 1.28.0), and do not contain restriction cut site sequences for KpnI nor XbaI. The full library of designed oligonucleotides can be found on GitHub (see Data availability).

MPRA library construction

We modified the original MPRA protocol (***Melnikov et al., 2012***) by using a lentiviral delivery system as previously described (***Inoue et al., 2017***), introducing GFP instead of nanoluciferase and changing the sequencing library preparation strategy. In brief, oligonucleotide sequences (Custom Array) were amplified using emulsion PCR (Micellula Kit, roboklon) and introduced into the pMPRA plasmid as described previously. The nanoluciferase sequence used in the original publication was replaced by EGFP using Gibson cloning and subsequent insertion into the enhancer library using restriction enzyme digest as in the original publication. Using SFiI the assembled library was transferred into a suitable lentiviral vector (pMPRAlenti1, Addgene #61600).

Primer sequences and plasmids used in the MPRA can be found in the analysis GitHub (see Data availability). To ensure maximum library complexity, transformations that involved the CRE library were performed using electroporation (NEB 10-beta electrocompetent *Escherichia coli*), in all other cloning steps chemically competent *E. coli* (NEB 5-alpha) were used.

Lentiviral particles were produced according to standard methods in HEK 293T cells (***Dull et al., 1998***). The MPRA library was co-transfected with third generation lentiviral plasmids (pMDLg/pRRE, pRSV-Rev, pMD2.G; Addgene #12251, #12253, #12259) using Lipofectamine 3000. The lentiviral particle containing supernatant was harvested 48 hr post transfection and filtered using 0.45 µm PES



syringe filters. Viral titer was determined by infecting N2A cells (ATCC CCL-131) and counting GFP-positive cells. To this end, N2A cells were infected with a 50/50 volume ratio of viral supernatant to cell suspension with addition of 8 µg/mL Polybrene. Cells were exposed to the lentiviral particles for 24 hr until medium was exchanged. Selection was performed using blasticidin starting 48 hr after infection.

MPRA lentiviral transduction

The transduction of the MPRA library was performed in triplicates on two *H. sapiens* and one *M. fascicularis* NPC lines generated as described previously (***Geuder et al., 2021***). 2.5 × 10^5 NPCs per line and replicate were dissociated, dissolved in 500 µL cell culture medium containing 8 µg/mL Polybrene and incubated with virus at MOI 12.7 for 1 hr at 37°C in suspension (***Nakai et al., 2018***). Thereafter, cells were seeded on Geltrex and cultured as described above. Virus containing medium was replaced the next day and cells were cultured for additional 24 hr. Cells were collected, lysed in 100 µL TRI reagent, and frozen at –80°C.

MPRA sequencing library generation

As input control for RNA expression, DNA amplicon libraries were constructed using 100–500 pg plasmid DNA. Library preparation was performed in two successive PCRs. A first PCR introduced the 5′ transposase mosaic end using overhang primers, this was used in the second PCR (Index PCR) to add a library-specific index sequence and Illumina Flow Cell adapters. The Adapter PCR was performed in triplicates using DreamTaq polymerase (Thermo Fisher Scientific). Subsequently 1–5 ng of the Adapter PCR product were subjected to the Index PCR using Q5 polymerase.

Total RNA from NPCs was extracted using the Direct-zol RNA Microprep Kit (Zymo Research). Five hundred ng of RNA were subjected to reverse transcription using Maxima H Minus RT (Thermo Fisher Scientific) with oligo-dT primers. Fifty ng of cDNA were used for library preparation and processed as described for plasmid DNA.

Plasmid and cDNA libraries were pooled and quality was evaluated using capillary gel electrophoresis (Agilent Bioanalyzer 2100). Sequencing was performed on an Illumina HiSeq 1500 instrument using a single-index, 50 bp, paired-end protocol.

MPRA data processing and analysis

MPRA reads were demultiplexed with deML (***Renaud et al., 2015***) using i5 and i7 adapter indices from Illumina. Next, we removed barcodes with low sequence quality, requiring a minimum Phred quality score of 10 for all bases of the barcode (zUMIs, fqfilter.pl script; ***Parekh et al., 2018***). Furthermore, we removed reads that had mismatches to the constant region (the first 20 bases of the GFP sequence TCTAGAGTCGCGGCCTTACT). The remaining reads that matched one of the known CRE-tile barcodes were tallied up resulting in a count table. Next, we filtered out CRE tiles that had been detected in only one of the three input plasmid library replicates (4202/4950). Counts per million were calculated per CRE tile per library (median counts: ~900k range: 590–1050k). Macaque replicate 3 was excluded due to its unusually low correlation with the other samples (Pearson's *r*). The final regulatory activity for each CRE tile per cell line was calculated as:

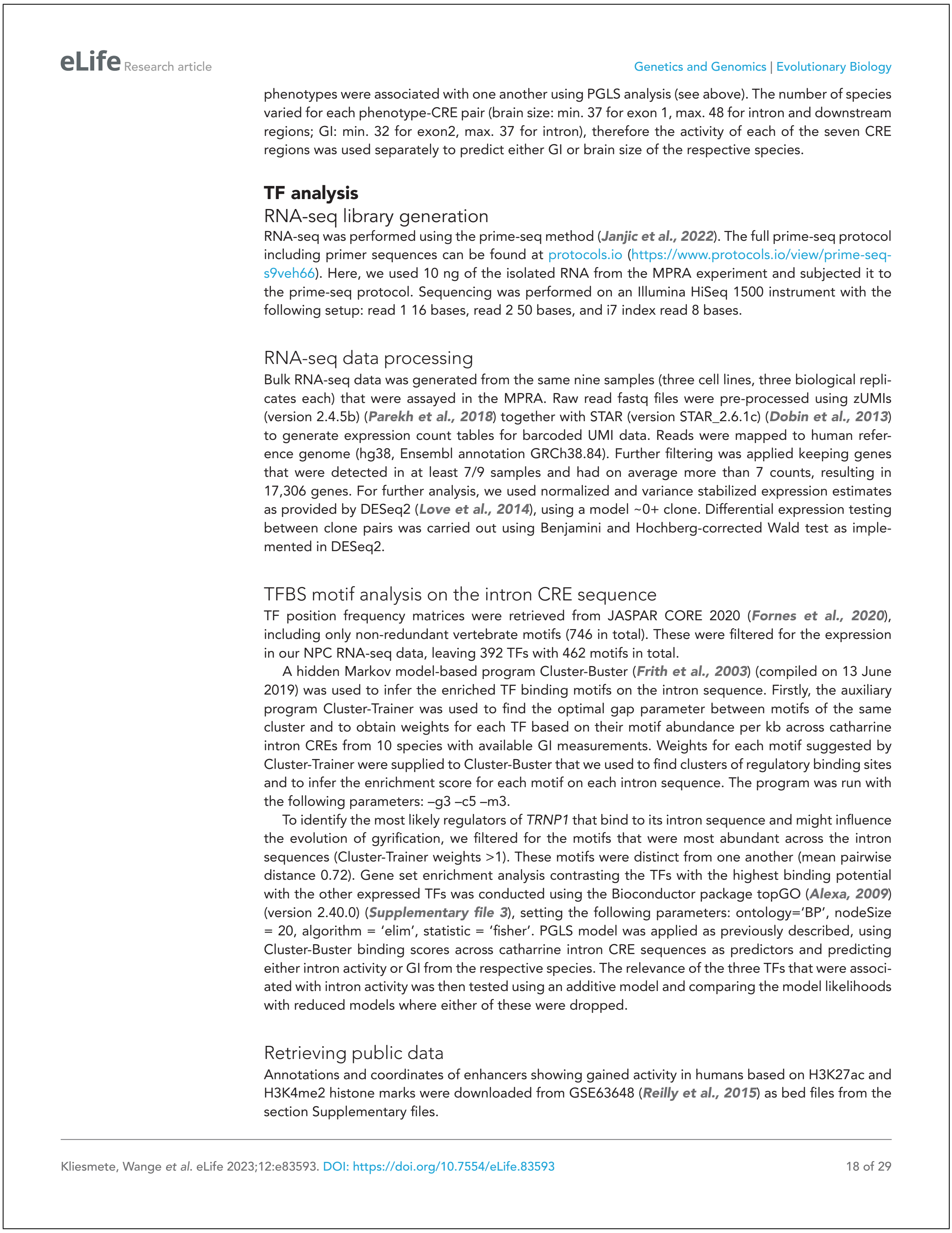
$$a_i = \frac{median(CPM_i)}{median(CPM_i)_p}, \tag{1}$$

where a is regulatory activity, i indicates CRE tile, and p is the input plasmid library. Median was calculated across the replicates from each cell line.

Given that each tile was overlapping with two other tiles upstream and two downstream, we calculated the total regulatory activity per CRE region in a coverage-sensitive manner, that is, for each position in the original sequence, mean per-bp-activity across the detected tiles covering it was calculated. The final CRE region activity is the sum across all base positions.

$$a_r = \sum_{b=1}^{k} \frac{1}{n} \sum_{i=1}^{n} \frac{a_i}{l_i}, \tag{2}$$

where a_r is regulatory activity of CRE region r, $b = 1, ..., k$ is the base position of region r, $i, ..., n$ are tiles overlapping the position b, a_i is tile activity from ***Equation 1*** and l_i is tile length. CRE activity and brain



phenotypes were associated with one another using PGLS analysis (see above). The number of species varied for each phenotype-CRE pair (brain size: min. 37 for exon 1, max. 48 for intron and downstream regions; GI: min. 32 for exon2, max. 37 for intron), therefore the activity of each of the seven CRE regions was used separately to predict either GI or brain size of the respective species.

TF analysis

RNA-seq library generation

RNA-seq was performed using the prime-seq method (***Janjic et al., 2022***). The full prime-seq protocol including primer sequences can be found at protocols.io (https://www.protocols.io/view/prime-seq-s9veh66). Here, we used 10 ng of the isolated RNA from the MPRA experiment and subjected it to the prime-seq protocol. Sequencing was performed on an Illumina HiSeq 1500 instrument with the following setup: read 1 16 bases, read 2 50 bases, and i7 index read 8 bases.

RNA-seq data processing

Bulk RNA-seq data was generated from the same nine samples (three cell lines, three biological replicates each) that were assayed in the MPRA. Raw read fastq files were pre-processed using zUMIs (version 2.4.5b) (***Parekh et al., 2018***) together with STAR (version STAR_2.6.1c) (***Dobin et al., 2013***) to generate expression count tables for barcoded UMI data. Reads were mapped to human reference genome (hg38, Ensembl annotation GRCh38.84). Further filtering was applied keeping genes that were detected in at least 7/9 samples and had on average more than 7 counts, resulting in 17,306 genes. For further analysis, we used normalized and variance stabilized expression estimates as provided by DESeq2 (***Love et al., 2014***), using a model ~0+ clone. Differential expression testing between clone pairs was carried out using Benjamini and Hochberg-corrected Wald test as implemented in DESeq2.

TFBS motif analysis on the intron CRE sequence

TF position frequency matrices were retrieved from JASPAR CORE 2020 (***Fornes et al., 2020***), including only non-redundant vertebrate motifs (746 in total). These were filtered for the expression in our NPC RNA-seq data, leaving 392 TFs with 462 motifs in total.

A hidden Markov model-based program Cluster-Buster (***Frith et al., 2003***) (compiled on 13 June 2019) was used to infer the enriched TF binding motifs on the intron sequence. Firstly, the auxiliary program Cluster-Trainer was used to find the optimal gap parameter between motifs of the same cluster and to obtain weights for each TF based on their motif abundance per kb across catharrine intron CREs from 10 species with available GI measurements. Weights for each motif suggested by Cluster-Trainer were supplied to Cluster-Buster that we used to find clusters of regulatory binding sites and to infer the enrichment score for each motif on each intron sequence. The program was run with the following parameters: –g3 –c5 –m3.

To identify the most likely regulators of *TRNP1* that bind to its intron sequence and might influence the evolution of gyrification, we filtered for the motifs that were most abundant across the intron sequences (Cluster-Trainer weights >1). These motifs were distinct from one another (mean pairwise distance 0.72). Gene set enrichment analysis contrasting the TFs with the highest binding potential with the other expressed TFs was conducted using the Bioconductor package topGO (***Alexa, 2009***) (version 2.40.0) (***Supplementary file 3***), setting the following parameters: ontology='BP', nodeSize = 20, algorithm = 'elim', statistic = 'fisher'. PGLS model was applied as previously described, using Cluster-Buster binding scores across catharrine intron CRE sequences as predictors and predicting either intron activity or GI from the respective species. The relevance of the three TFs that were associated with intron activity was then tested using an additive model and comparing the model likelihoods with reduced models where either of these were dropped.

Retrieving public data

Annotations and coordinates of enhancers showing gained activity in humans based on H3K27ac and H3K4me2 histone marks were downloaded from GSE63648 (***Reilly et al., 2015***) as bed files from the section Supplementary files.

CTCF ChiP-seq data from human neural progenitor cells (line H9) was retrieved from ENCODE (***Encode Project Consortium, 2012***) (doi:10.17989/ENCSR125NBL). All samples were consistent regarding TRNP1 CTCF ChIP-seq landscape. We depict read distribution using BigWig file of sample ENCFF896TQG.

Human Hi-C data (***Won et al., 2016***) on TAD positions in germinal zone at week 8 was retrieved as a coordinate file in bed format using GEO accession GSE77565.

Quantification and statistical analysis

Data visualizations and statistical analysis was performed using R (version 4.0) (***R Development Core Team, 2019***). Details of the statistical tests performed in this study can be found in the main text as well as the Materials and methods section and ***Supplementary files 1–3***. For display items all relevant parameters like sample size (*n*), type of statistical test, significance thresholds, degrees of freedom, as well as standard deviations can be found in the figure legends.

Resource availability

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Wolfgang Enard (enard@bio.lmu.de).

Materials availability

Plasmids and cell lines used in this work will be available upon request.

Acknowledgements

This work was supported by the Deutsche Forschungsgemeinschaft (DFG) through LMUexcellent, SFB1243 (Subproject A14/A15 to WE and IH, respectively), DFG grant HE 7669/1-1 (to IH) and the advanced ERC grants ChroNeuroRepair and NeuroCentro (to MG) and the Cyliax foundation (to WE). We want to thank Christian Roos from the German Primate Center for providing genomic DNA from primates, Deeksha for providing Ferret fibroblast cells, project students Gunnar Kuut and Fatih Sarigoel for helping to generate TRNP1 orthologous sequences, Nikola Vuković for helping to establish the MPRA assay, Nika Foglar and Reza Rifat for helping with the proliferation assays and Christoph Neumayr, Tamina Dietl for helping in data analysis.

Additional information

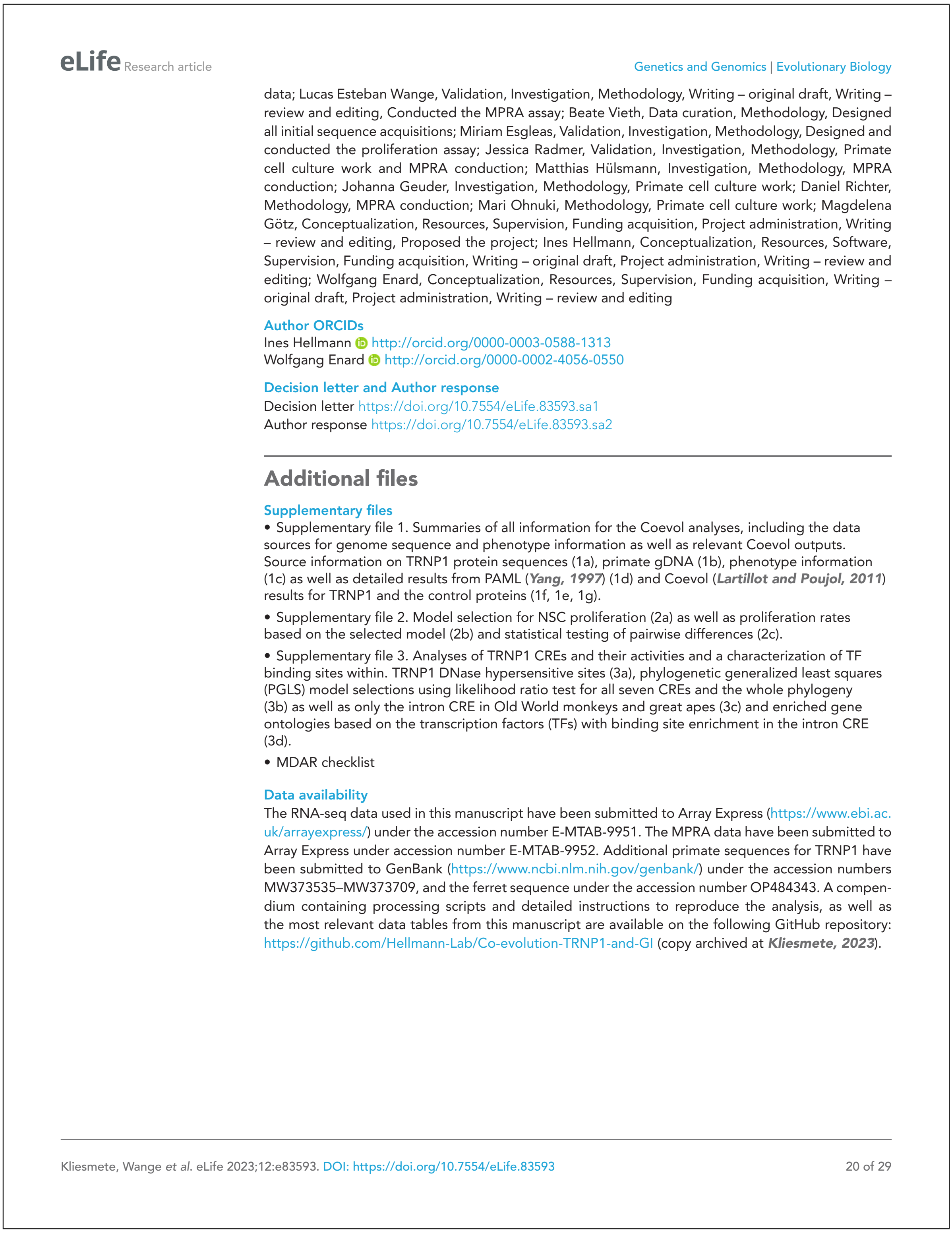
Funding

Funder	Grant reference number	Author
Cyliax Foundation		Wolfgang Enard
Deutsche Forschungsgemeinschaft	458247426	Wolfgang Enard Ines Hellmann
Deutsche Forschungsgemeinschaft	407541155	Ines Hellmann
European Research Council	ChroNeuroRepair	Magdelena Götz
European Research Council	NeuroCentro	Magdelena Götz

The funders had no role in study design, data collection and interpretation, or the decision to submit the work for publication.

Author contributions

Zane Kliesmete, Data curation, Software, Formal analysis, Validation, Investigation, Visualization, Methodology, Writing – original draft, Writing – review and editing, Collected, integrated and analysed all



data; Lucas Esteban Wange, Validation, Investigation, Methodology, Writing – original draft, Writing – review and editing, Conducted the MPRA assay; Beate Vieth, Data curation, Methodology, Designed all initial sequence acquisitions; Miriam Esgleas, Validation, Investigation, Methodology, Designed and conducted the proliferation assay; Jessica Radmer, Validation, Investigation, Methodology, Primate cell culture work and MPRA conduction; Matthias Hülsmann, Investigation, Methodology, MPRA conduction; Johanna Geuder, Investigation, Methodology, Primate cell culture work; Daniel Richter, Methodology, MPRA conduction; Mari Ohnuki, Methodology, Primate cell culture work; Magdelena Götz, Conceptualization, Resources, Supervision, Funding acquisition, Project administration, Writing – review and editing, Proposed the project; Ines Hellmann, Conceptualization, Resources, Software, Supervision, Funding acquisition, Writing – original draft, Project administration, Writing – review and editing; Wolfgang Enard, Conceptualization, Resources, Supervision, Funding acquisition, Writing – original draft, Project administration, Writing – review and editing

Author ORCIDs

Ines Hellmann http://orcid.org/0000-0003-0588-1313
Wolfgang Enard http://orcid.org/0000-0002-4056-0550

Decision letter and Author response

Decision letter https://doi.org/10.7554/eLife.83593.sa1
Author response https://doi.org/10.7554/eLife.83593.sa2

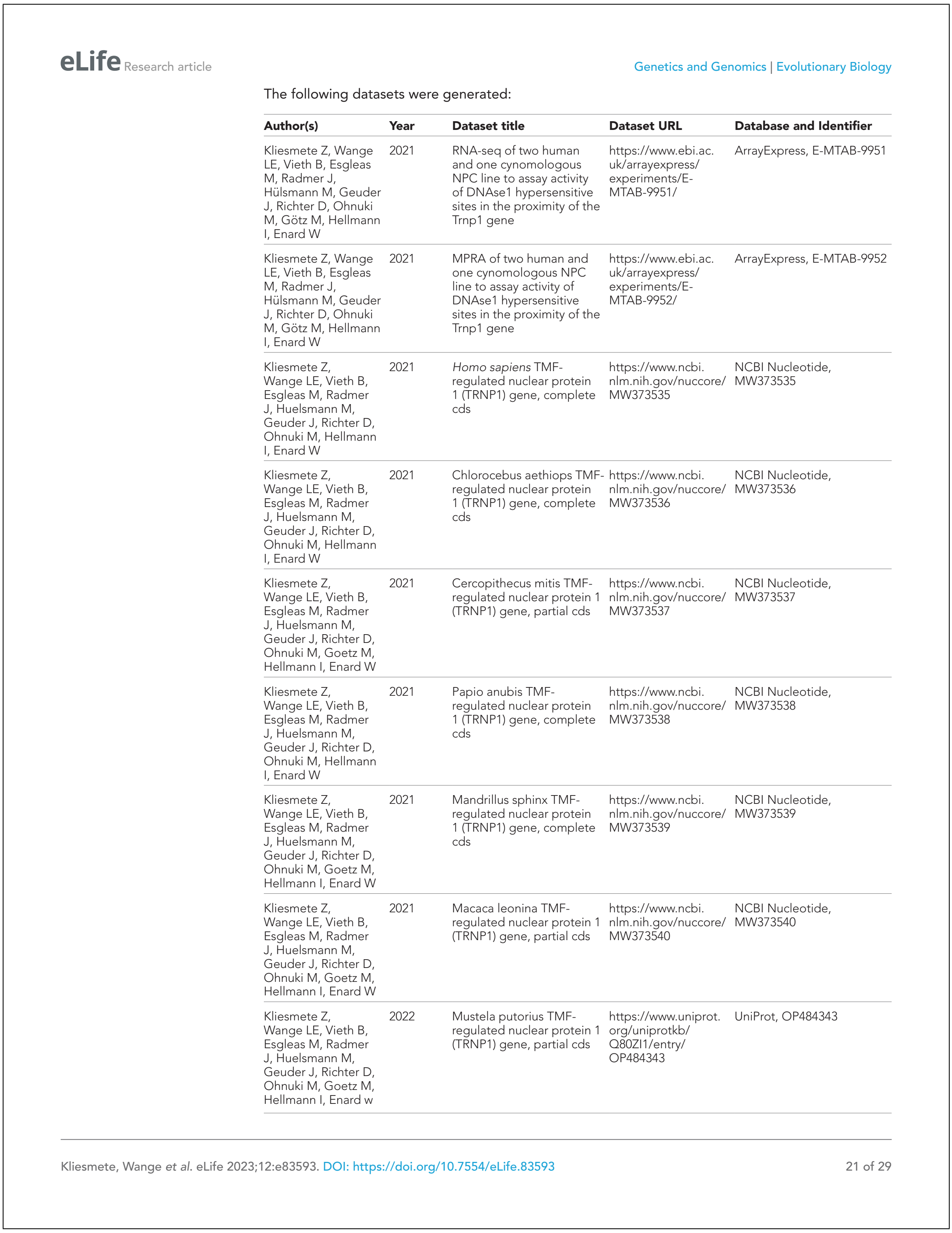
Additional files

Supplementary files

- Supplementary file 1. Summaries of all information for the Coevol analyses, including the data sources for genome sequence and phenotype information as well as relevant Coevol outputs. Source information on TRNP1 protein sequences (1a), primate gDNA (1b), phenotype information (1c) as well as detailed results from PAML (***Yang, 1997***) (1d) and Coevol (***Lartillot and Poujol, 2011***) results for TRNP1 and the control proteins (1f, 1e, 1g).
- Supplementary file 2. Model selection for NSC proliferation (2a) as well as proliferation rates based on the selected model (2b) and statistical testing of pairwise differences (2c).
- Supplementary file 3. Analyses of TRNP1 CREs and their activities and a characterization of TF binding sites within. TRNP1 DNase hypersensitive sites (3a), phylogenetic generalized least squares (PGLS) model selections using likelihood ratio test for all seven CREs and the whole phylogeny (3b) as well as only the intron CRE in Old World monkeys and great apes (3c) and enriched gene ontologies based on the transcription factors (TFs) with binding site enrichment in the intron CRE (3d).
- MDAR checklist

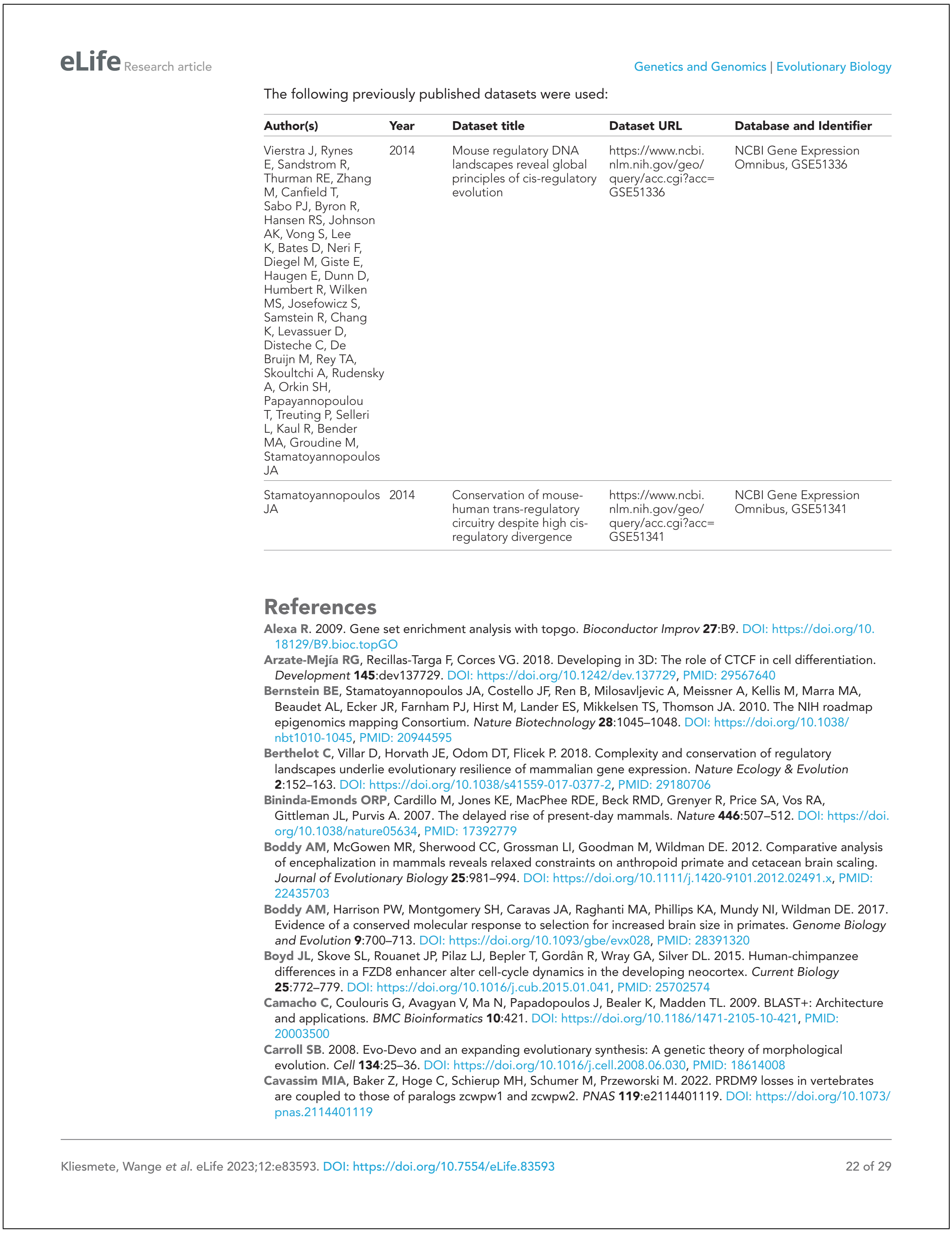
Data availability

The RNA-seq data used in this manuscript have been submitted to Array Express (https://www.ebi.ac.uk/arrayexpress/) under the accession number E-MTAB-9951. The MPRA data have been submitted to Array Express under accession number E-MTAB-9952. Additional primate sequences for TRNP1 have been submitted to GenBank (https://www.ncbi.nlm.nih.gov/genbank/) under the accession numbers MW373535–MW373709, and the ferret sequence under the accession number OP484343. A compendium containing processing scripts and detailed instructions to reproduce the analysis, as well as the most relevant data tables from this manuscript are available on the following GitHub repository: https://github.com/Hellmann-Lab/Co-evolution-TRNP1-and-GI (copy archived at ***Kliesmete, 2023***).



The following datasets were generated:

Author(s)	Year	Dataset title	Dataset URL	Database and Identifier
Kliesmete Z, Wange LE, Vieth B, Esgleas M, Radmer J, Hülsmann M, Geuder J, Richter D, Ohnuki M, Götz M, Hellmann I, Enard W	2021	RNA-seq of two human and one cynomologous NPC line to assay activity of DNAse1 hypersensitive sites in the proximity of the Trnp1 gene	https://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-9951/	ArrayExpress, E-MTAB-9951
Kliesmete Z, Wange LE, Vieth B, Esgleas M, Radmer J, Hülsmann M, Geuder J, Richter D, Ohnuki M, Götz M, Hellmann I, Enard W	2021	MPRA of two human and one cynomologous NPC line to assay activity of DNAse1 hypersensitive sites in the proximity of the Trnp1 gene	https://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-9952/	ArrayExpress, E-MTAB-9952
Kliesmete Z, Wange LE, Vieth B, Esgleas M, Radmer J, Huelsmann M, Geuder J, Richter D, Ohnuki M, Hellmann I, Enard W	2021	*Homo sapiens* TMF-regulated nuclear protein 1 (TRNP1) gene, complete cds	https://www.ncbi.nlm.nih.gov/nuccore/MW373535	NCBI Nucleotide, MW373535
Kliesmete Z, Wange LE, Vieth B, Esgleas M, Radmer J, Huelsmann M, Geuder J, Richter D, Ohnuki M, Hellmann I, Enard W	2021	Chlorocebus aethiops TMF-regulated nuclear protein 1 (TRNP1) gene, complete cds	https://www.ncbi.nlm.nih.gov/nuccore/MW373536	NCBI Nucleotide, MW373536
Kliesmete Z, Wange LE, Vieth B, Esgleas M, Radmer J, Huelsmann M, Geuder J, Richter D, Ohnuki M, Goetz M, Hellmann I, Enard W	2021	Cercopithecus mitis TMF-regulated nuclear protein 1 (TRNP1) gene, partial cds	https://www.ncbi.nlm.nih.gov/nuccore/MW373537	NCBI Nucleotide, MW373537
Kliesmete Z, Wange LE, Vieth B, Esgleas M, Radmer J, Huelsmann M, Geuder J, Richter D, Ohnuki M, Hellmann I, Enard W	2021	Papio anubis TMF-regulated nuclear protein 1 (TRNP1) gene, complete cds	https://www.ncbi.nlm.nih.gov/nuccore/MW373538	NCBI Nucleotide, MW373538
Kliesmete Z, Wange LE, Vieth B, Esgleas M, Radmer J, Huelsmann M, Geuder J, Richter D, Ohnuki M, Goetz M, Hellmann I, Enard W	2021	Mandrillus sphinx TMF-regulated nuclear protein 1 (TRNP1) gene, complete cds	https://www.ncbi.nlm.nih.gov/nuccore/MW373539	NCBI Nucleotide, MW373539
Kliesmete Z, Wange LE, Vieth B, Esgleas M, Radmer J, Huelsmann M, Geuder J, Richter D, Ohnuki M, Goetz M, Hellmann I, Enard W	2021	Macaca leonina TMF-regulated nuclear protein 1 (TRNP1) gene, partial cds	https://www.ncbi.nlm.nih.gov/nuccore/MW373540	NCBI Nucleotide, MW373540
Kliesmete Z, Wange LE, Vieth B, Esgleas M, Radmer J, Huelsmann M, Geuder J, Richter D, Ohnuki M, Goetz M, Hellmann I, Enard w	2022	Mustela putorius TMF-regulated nuclear protein 1 (TRNP1) gene, partial cds	https://www.uniprot.org/uniprotkb/Q80ZI1/entry/OP484343	UniProt, OP484343



The following previously published datasets were used:

Author(s)	Year	Dataset title	Dataset URL	Database and Identifier
Vierstra J, Rynes E, Sandstrom R, Thurman RE, Zhang M, Canfield T, Sabo PJ, Byron R, Hansen RS, Johnson AK, Vong S, Lee K, Bates D, Neri F, Diegel M, Giste E, Haugen E, Dunn D, Humbert R, Wilken MS, Josefowicz S, Samstein R, Chang K, Levassuer D, Disteche C, De Bruijn M, Rey TA, Skoultchi A, Rudensky A, Orkin SH, Papayannopoulou T, Treuting P, Selleri L, Kaul R, Bender MA, Groudine M, Stamatoyannopoulos JA	2014	Mouse regulatory DNA landscapes reveal global principles of cis-regulatory evolution	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE51336	NCBI Gene Expression Omnibus, GSE51336
Stamatoyannopoulos JA	2014	Conservation of mouse-human trans-regulatory circuitry despite high cis-regulatory divergence	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE51341	NCBI Gene Expression Omnibus, GSE51341

References

Alexa R. 2009. Gene set enrichment analysis with topgo. *Bioconductor Improv* **27**:B9. DOI: https://doi.org/10.18129/B9.bioc.topGO

Arzate-Mejía RG, Recillas-Targa F, Corces VG. 2018. Developing in 3D: The role of CTCF in cell differentiation. *Development* **145**:dev137729. DOI: https://doi.org/10.1242/dev.137729, PMID: 29567640

Bernstein BE, Stamatoyannopoulos JA, Costello JF, Ren B, Milosavljevic A, Meissner A, Kellis M, Marra MA, Beaudet AL, Ecker JR, Farnham PJ, Hirst M, Lander ES, Mikkelsen TS, Thomson JA. 2010. The NIH roadmap epigenomics mapping Consortium. *Nature Biotechnology* **28**:1045–1048. DOI: https://doi.org/10.1038/nbt1010-1045, PMID: 20944595

Berthelot C, Villar D, Horvath JE, Odom DT, Flicek P. 2018. Complexity and conservation of regulatory landscapes underlie evolutionary resilience of mammalian gene expression. *Nature Ecology & Evolution* **2**:152–163. DOI: https://doi.org/10.1038/s41559-017-0377-2, PMID: 29180706

Bininda-Emonds ORP, Cardillo M, Jones KE, MacPhee RDE, Beck RMD, Grenyer R, Price SA, Vos RA, Gittleman JL, Purvis A. 2007. The delayed rise of present-day mammals. *Nature* **446**:507–512. DOI: https://doi.org/10.1038/nature05634, PMID: 17392779

Boddy AM, McGowen MR, Sherwood CC, Grossman LI, Goodman M, Wildman DE. 2012. Comparative analysis of encephalization in mammals reveals relaxed constraints on anthropoid primate and cetacean brain scaling. *Journal of Evolutionary Biology* **25**:981–994. DOI: https://doi.org/10.1111/j.1420-9101.2012.02491.x, PMID: 22435703

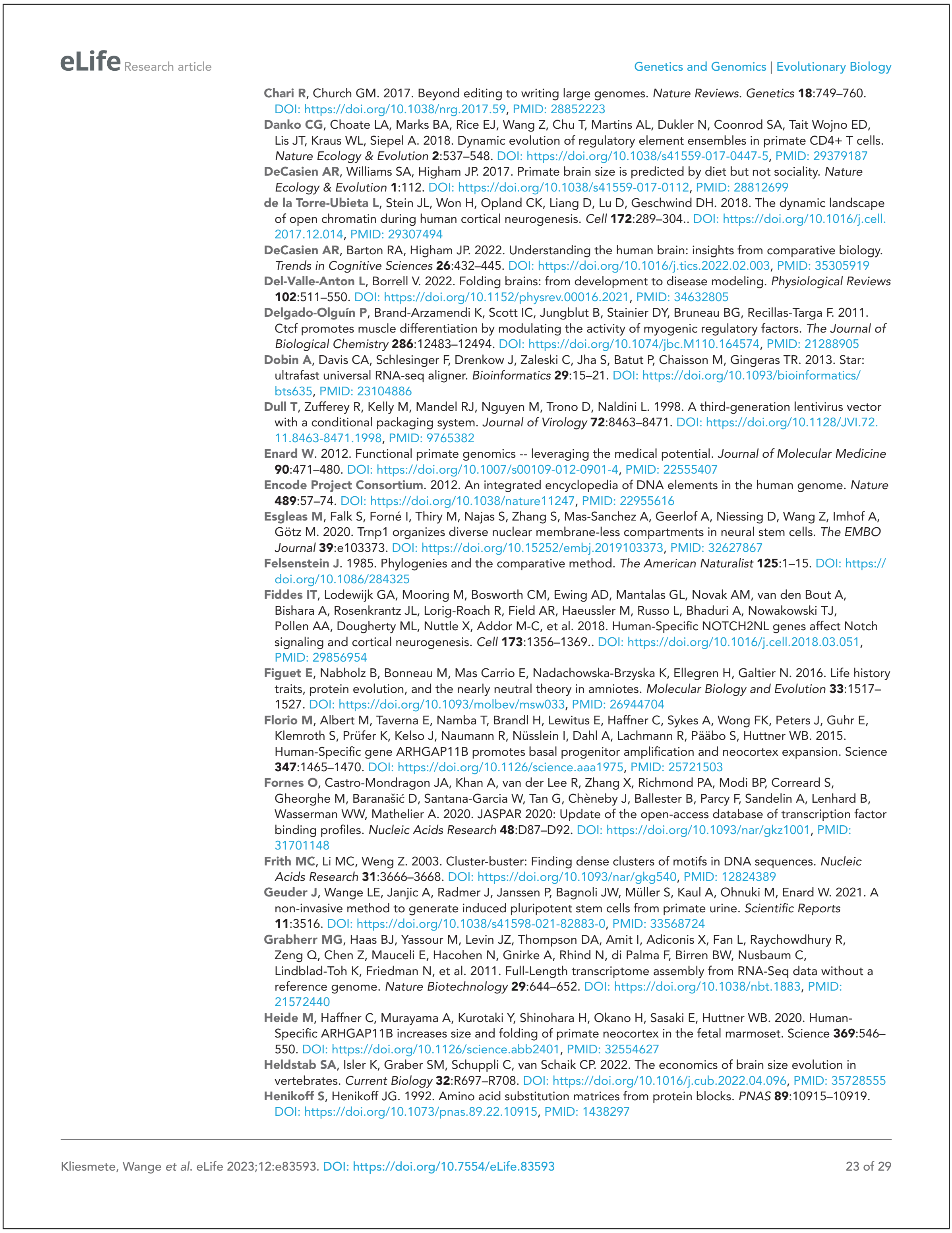
Boddy AM, Harrison PW, Montgomery SH, Caravas JA, Raghanti MA, Phillips KA, Mundy NI, Wildman DE. 2017. Evidence of a conserved molecular response to selection for increased brain size in primates. *Genome Biology and Evolution* **9**:700–713. DOI: https://doi.org/10.1093/gbe/evx028, PMID: 28391320

Boyd JL, Skove SL, Rouanet JP, Pilaz LJ, Bepler T, Gordân R, Wray GA, Silver DL. 2015. Human-chimpanzee differences in a FZD8 enhancer alter cell-cycle dynamics in the developing neocortex. *Current Biology* **25**:772–779. DOI: https://doi.org/10.1016/j.cub.2015.01.041, PMID: 25702574

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. BLAST+: Architecture and applications. *BMC Bioinformatics* **10**:421. DOI: https://doi.org/10.1186/1471-2105-10-421, PMID: 20003500

Carroll SB. 2008. Evo-Devo and an expanding evolutionary synthesis: A genetic theory of morphological evolution. *Cell* **134**:25–36. DOI: https://doi.org/10.1016/j.cell.2008.06.030, PMID: 18614008

Cavassim MIA, Baker Z, Hoge C, Schierup MH, Schumer M, Przeworski M. 2022. PRDM9 losses in vertebrates are coupled to those of paralogs zcwpw1 and zcwpw2. *PNAS* **119**:e2114401119. DOI: https://doi.org/10.1073/pnas.2114401119



Chari R, Church GM. 2017. Beyond editing to writing large genomes. *Nature Reviews. Genetics* **18**:749–760. DOI: https://doi.org/10.1038/nrg.2017.59, PMID: 28852223

Danko CG, Choate LA, Marks BA, Rice EJ, Wang Z, Chu T, Martins AL, Dukler N, Coonrod SA, Tait Wojno ED, Lis JT, Kraus WL, Siepel A. 2018. Dynamic evolution of regulatory element ensembles in primate CD4+ T cells. *Nature Ecology & Evolution* **2**:537–548. DOI: https://doi.org/10.1038/s41559-017-0447-5, PMID: 29379187

DeCasien AR, Williams SA, Higham JP. 2017. Primate brain size is predicted by diet but not sociality. *Nature Ecology & Evolution* **1**:112. DOI: https://doi.org/10.1038/s41559-017-0112, PMID: 28812699

de la Torre-Ubieta L, Stein JL, Won H, Opland CK, Liang D, Lu D, Geschwind DH. 2018. The dynamic landscape of open chromatin during human cortical neurogenesis. *Cell* **172**:289–304.. DOI: https://doi.org/10.1016/j.cell.2017.12.014, PMID: 29307494

DeCasien AR, Barton RA, Higham JP. 2022. Understanding the human brain: insights from comparative biology. *Trends in Cognitive Sciences* **26**:432–445. DOI: https://doi.org/10.1016/j.tics.2022.02.003, PMID: 35305919

Del-Valle-Anton L, Borrell V. 2022. Folding brains: from development to disease modeling. *Physiological Reviews* **102**:511–550. DOI: https://doi.org/10.1152/physrev.00016.2021, PMID: 34632805

Delgado-Olguín P, Brand-Arzamendi K, Scott IC, Jungblut B, Stainier DY, Bruneau BG, Recillas-Targa F. 2011. Ctcf promotes muscle differentiation by modulating the activity of myogenic regulatory factors. *The Journal of Biological Chemistry* **286**:12483–12494. DOI: https://doi.org/10.1074/jbc.M110.164574, PMID: 21288905

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. 2013. Star: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**:15–21. DOI: https://doi.org/10.1093/bioinformatics/bts635, PMID: 23104886

Dull T, Zufferey R, Kelly M, Mandel RJ, Nguyen M, Trono D, Naldini L. 1998. A third-generation lentivirus vector with a conditional packaging system. *Journal of Virology* **72**:8463–8471. DOI: https://doi.org/10.1128/JVI.72.11.8463-8471.1998, PMID: 9765382

Enard W. 2012. Functional primate genomics -- leveraging the medical potential. *Journal of Molecular Medicine* **90**:471–480. DOI: https://doi.org/10.1007/s00109-012-0901-4, PMID: 22555407

Encode Project Consortium. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**:57–74. DOI: https://doi.org/10.1038/nature11247, PMID: 22955616

Esgleas M, Falk S, Forné I, Thiry M, Najas S, Zhang S, Mas-Sanchez A, Geerlof A, Niessing D, Wang Z, Imhof A, Götz M. 2020. Trnp1 organizes diverse nuclear membrane-less compartments in neural stem cells. *The EMBO Journal* **39**:e103373. DOI: https://doi.org/10.15252/embj.2019103373, PMID: 32627867

Felsenstein J. 1985. Phylogenies and the comparative method. *The American Naturalist* **125**:1–15. DOI: https://doi.org/10.1086/284325

Fiddes IT, Lodewijk GA, Mooring M, Bosworth CM, Ewing AD, Mantalas GL, Novak AM, van den Bout A, Bishara A, Rosenkrantz JL, Lorig-Roach R, Field AR, Haeussler M, Russo L, Bhaduri A, Nowakowski TJ, Pollen AA, Dougherty ML, Nuttle X, Addor M-C, et al. 2018. Human-Specific NOTCH2NL genes affect Notch signaling and cortical neurogenesis. *Cell* **173**:1356–1369.. DOI: https://doi.org/10.1016/j.cell.2018.03.051, PMID: 29856954

Figuet E, Nabholz B, Bonneau M, Mas Carrio E, Nadachowska-Brzyska K, Ellegren H, Galtier N. 2016. Life history traits, protein evolution, and the nearly neutral theory in amniotes. *Molecular Biology and Evolution* **33**:1517–1527. DOI: https://doi.org/10.1093/molbev/msw033, PMID: 26944704

Florio M, Albert M, Taverna E, Namba T, Brandl H, Lewitus E, Haffner C, Sykes A, Wong FK, Peters J, Guhr E, Klemroth S, Prüfer K, Kelso J, Naumann R, Nüsslein I, Dahl A, Lachmann R, Pääbo S, Huttner WB. 2015. Human-Specific gene ARHGAP11B promotes basal progenitor amplification and neocortex expansion. Science **347**:1465–1470. DOI: https://doi.org/10.1126/science.aaa1975, PMID: 25721503

Fornes O, Castro-Mondragon JA, Khan A, van der Lee R, Zhang X, Richmond PA, Modi BP, Correard S, Gheorghe M, Baranašić D, Santana-Garcia W, Tan G, Chèneby J, Ballester B, Parcy F, Sandelin A, Lenhard B, Wasserman WW, Mathelier A. 2020. JASPAR 2020: Update of the open-access database of transcription factor binding profiles. *Nucleic Acids Research* **48**:D87–D92. DOI: https://doi.org/10.1093/nar/gkz1001, PMID: 31701148

Frith MC, Li MC, Weng Z. 2003. Cluster-buster: Finding dense clusters of motifs in DNA sequences. *Nucleic Acids Research* **31**:3666–3668. DOI: https://doi.org/10.1093/nar/gkg540, PMID: 12824389

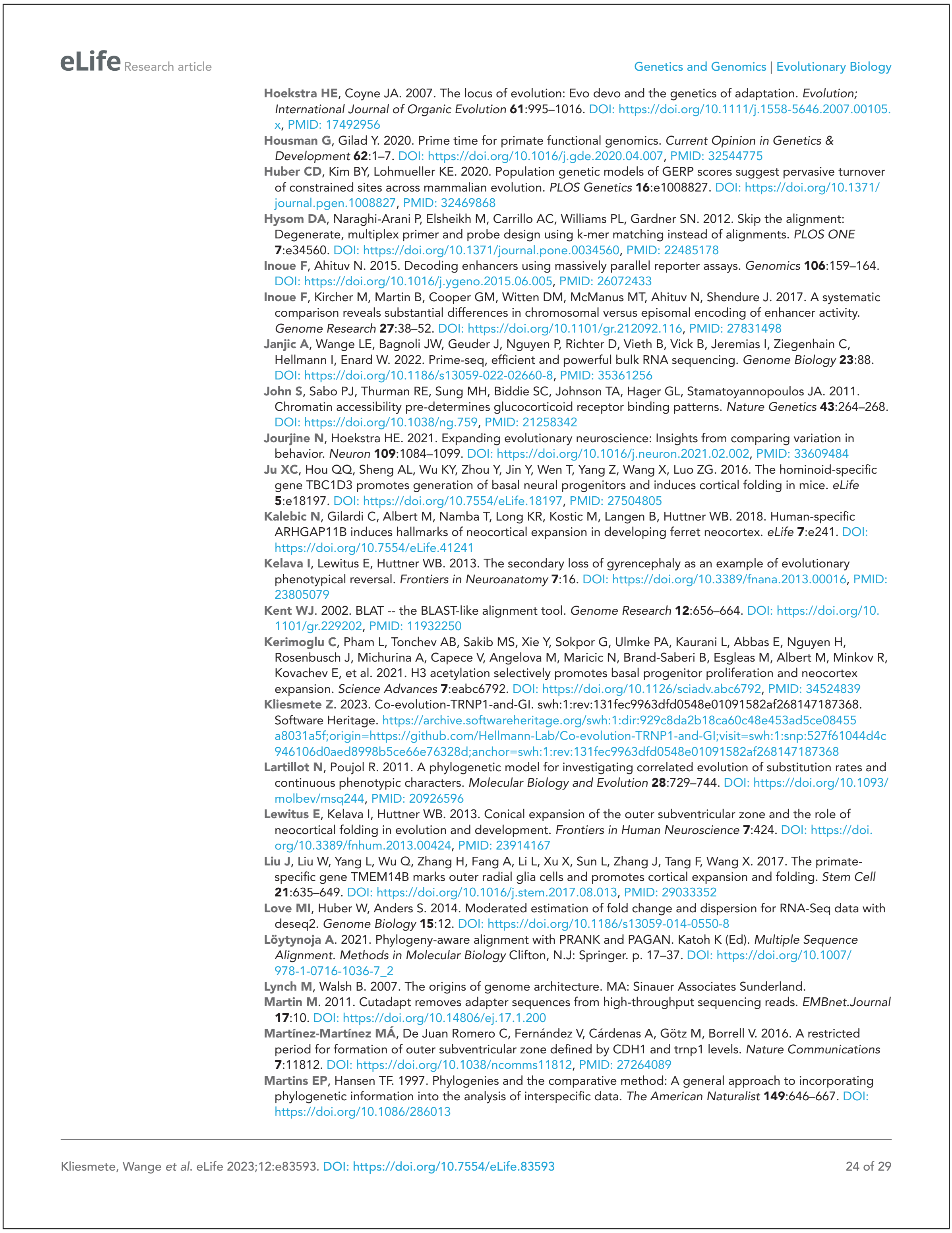
Geuder J, Wange LE, Janjic A, Radmer J, Janssen P, Bagnoli JW, Müller S, Kaul A, Ohnuki M, Enard W. 2021. A non-invasive method to generate induced pluripotent stem cells from primate urine. *Scientific Reports* **11**:3516. DOI: https://doi.org/10.1038/s41598-021-82883-0, PMID: 33568724

Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, Chen Z, Mauceli E, Hacohen N, Gnirke A, Rhind N, di Palma F, Birren BW, Nusbaum C, Lindblad-Toh K, Friedman N, et al. 2011. Full-Length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology* **29**:644–652. DOI: https://doi.org/10.1038/nbt.1883, PMID: 21572440

Heide M, Haffner C, Murayama A, Kurotaki Y, Shinohara H, Okano H, Sasaki E, Huttner WB. 2020. Human-Specific ARHGAP11B increases size and folding of primate neocortex in the fetal marmoset. Science **369**:546–550. DOI: https://doi.org/10.1126/science.abb2401, PMID: 32554627

Heldstab SA, Isler K, Graber SM, Schuppli C, van Schaik CP. 2022. The economics of brain size evolution in vertebrates. *Current Biology* **32**:R697–R708. DOI: https://doi.org/10.1016/j.cub.2022.04.096, PMID: 35728555

Henikoff S, Henikoff JG. 1992. Amino acid substitution matrices from protein blocks. *PNAS* **89**:10915–10919. DOI: https://doi.org/10.1073/pnas.89.22.10915, PMID: 1438297



Hoekstra HE, Coyne JA. 2007. The locus of evolution: Evo devo and the genetics of adaptation. *Evolution; International Journal of Organic Evolution* **61**:995–1016. DOI: https://doi.org/10.1111/j.1558-5646.2007.00105.x, PMID: 17492956

Housman G, Gilad Y. 2020. Prime time for primate functional genomics. *Current Opinion in Genetics & Development* **62**:1–7. DOI: https://doi.org/10.1016/j.gde.2020.04.007, PMID: 32544775

Huber CD, Kim BY, Lohmueller KE. 2020. Population genetic models of GERP scores suggest pervasive turnover of constrained sites across mammalian evolution. *PLOS Genetics* **16**:e1008827. DOI: https://doi.org/10.1371/journal.pgen.1008827, PMID: 32469868

Hysom DA, Naraghi-Arani P, Elsheikh M, Carrillo AC, Williams PL, Gardner SN. 2012. Skip the alignment: Degenerate, multiplex primer and probe design using k-mer matching instead of alignments. *PLOS ONE* **7**:e34560. DOI: https://doi.org/10.1371/journal.pone.0034560, PMID: 22485178

Inoue F, Ahituv N. 2015. Decoding enhancers using massively parallel reporter assays. *Genomics* **106**:159–164. DOI: https://doi.org/10.1016/j.ygeno.2015.06.005, PMID: 26072433

Inoue F, Kircher M, Martin B, Cooper GM, Witten DM, McManus MT, Ahituv N, Shendure J. 2017. A systematic comparison reveals substantial differences in chromosomal versus episomal encoding of enhancer activity. *Genome Research* **27**:38–52. DOI: https://doi.org/10.1101/gr.212092.116, PMID: 27831498

Janjic A, Wange LE, Bagnoli JW, Geuder J, Nguyen P, Richter D, Vieth B, Vick B, Jeremias I, Ziegenhain C, Hellmann I, Enard W. 2022. Prime-seq, efficient and powerful bulk RNA sequencing. *Genome Biology* **23**:88. DOI: https://doi.org/10.1186/s13059-022-02660-8, PMID: 35361256

John S, Sabo PJ, Thurman RE, Sung MH, Biddie SC, Johnson TA, Hager GL, Stamatoyannopoulos JA. 2011. Chromatin accessibility pre-determines glucocorticoid receptor binding patterns. *Nature Genetics* **43**:264–268. DOI: https://doi.org/10.1038/ng.759, PMID: 21258342

Jourjine N, Hoekstra HE. 2021. Expanding evolutionary neuroscience: Insights from comparing variation in behavior. *Neuron* **109**:1084–1099. DOI: https://doi.org/10.1016/j.neuron.2021.02.002, PMID: 33609484

Ju XC, Hou QQ, Sheng AL, Wu KY, Zhou Y, Jin Y, Wen T, Yang Z, Wang X, Luo ZG. 2016. The hominoid-specific gene TBC1D3 promotes generation of basal neural progenitors and induces cortical folding in mice. *eLife* **5**:e18197. DOI: https://doi.org/10.7554/eLife.18197, PMID: 27504805

Kalebic N, Gilardi C, Albert M, Namba T, Long KR, Kostic M, Langen B, Huttner WB. 2018. Human-specific ARHGAP11B induces hallmarks of neocortical expansion in developing ferret neocortex. *eLife* **7**:e241. DOI: https://doi.org/10.7554/eLife.41241

Kelava I, Lewitus E, Huttner WB. 2013. The secondary loss of gyrencephaly as an example of evolutionary phenotypical reversal. *Frontiers in Neuroanatomy* **7**:16. DOI: https://doi.org/10.3389/fnana.2013.00016, PMID: 23805079

Kent WJ. 2002. BLAT -- the BLAST-like alignment tool. *Genome Research* **12**:656–664. DOI: https://doi.org/10.1101/gr.229202, PMID: 11932250

Kerimoglu C, Pham L, Tonchev AB, Sakib MS, Xie Y, Sokpor G, Ulmke PA, Kaurani L, Abbas E, Nguyen H, Rosenbusch J, Michurina A, Capece V, Angelova M, Maricic N, Brand-Saberi B, Esgleas M, Albert M, Minkov R, Kovachev E, et al. 2021. H3 acetylation selectively promotes basal progenitor proliferation and neocortex expansion. *Science Advances* **7**:eabc6792. DOI: https://doi.org/10.1126/sciadv.abc6792, PMID: 34524839

Kliesmete Z. 2023. Co-evolution-TRNP1-and-GI. swh:1:rev:131fec9963dfd0548e01091582af268147187368. Software Heritage. https://archive.softwareheritage.org/swh:1:dir:929c8da2b18ca60c48e453ad5ce08455a8031a5f;origin=https://github.com/Hellmann-Lab/Co-evolution-TRNP1-and-GI;visit=swh:1:snp:527f61044d4c946106d0aed8998b5ce66e76328d;anchor=swh:1:rev:131fec9963dfd0548e01091582af268147187368

Lartillot N, Poujol R. 2011. A phylogenetic model for investigating correlated evolution of substitution rates and continuous phenotypic characters. *Molecular Biology and Evolution* **28**:729–744. DOI: https://doi.org/10.1093/molbev/msq244, PMID: 20926596

Lewitus E, Kelava I, Huttner WB. 2013. Conical expansion of the outer subventricular zone and the role of neocortical folding in evolution and development. *Frontiers in Human Neuroscience* **7**:424. DOI: https://doi.org/10.3389/fnhum.2013.00424, PMID: 23914167

Liu J, Liu W, Yang L, Wu Q, Zhang H, Fang A, Li L, Xu X, Sun L, Zhang J, Tang F, Wang X. 2017. The primate-specific gene TMEM14B marks outer radial glia cells and promotes cortical expansion and folding. *Stem Cell* **21**:635–649. DOI: https://doi.org/10.1016/j.stem.2017.08.013, PMID: 29033352

Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for RNA-Seq data with deseq2. *Genome Biology* **15**:12. DOI: https://doi.org/10.1186/s13059-014-0550-8

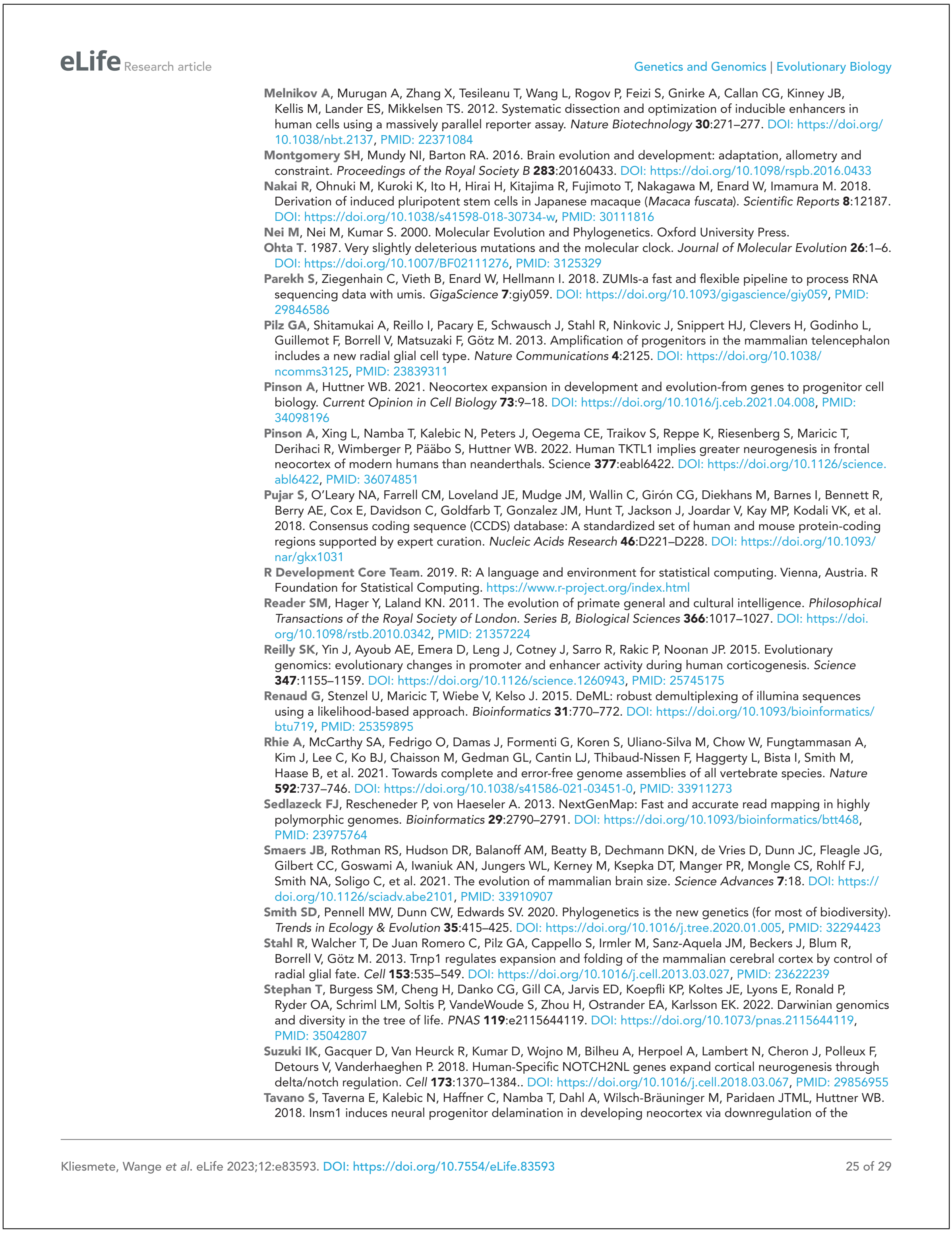
Löytynoja A. 2021. Phylogeny-aware alignment with PRANK and PAGAN. Katoh K (Ed). *Multiple Sequence Alignment. Methods in Molecular Biology* Clifton, N.J: Springer. p. 17–37. DOI: https://doi.org/10.1007/978-1-0716-1036-7_2

Lynch M, Walsh B. 2007. The origins of genome architecture. MA: Sinauer Associates Sunderland.

Martin M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.Journal* **17**:10. DOI: https://doi.org/10.14806/ej.17.1.200

Martínez-Martínez MÁ, De Juan Romero C, Fernández V, Cárdenas A, Götz M, Borrell V. 2016. A restricted period for formation of outer subventricular zone defined by CDH1 and trnp1 levels. *Nature Communications* **7**:11812. DOI: https://doi.org/10.1038/ncomms11812, PMID: 27264089

Martins EP, Hansen TF. 1997. Phylogenies and the comparative method: A general approach to incorporating phylogenetic information into the analysis of interspecific data. *The American Naturalist* **149**:646–667. DOI: https://doi.org/10.1086/286013



Melnikov A, Murugan A, Zhang X, Tesileanu T, Wang L, Rogov P, Feizi S, Gnirke A, Callan CG, Kinney JB, Kellis M, Lander ES, Mikkelsen TS. 2012. Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. *Nature Biotechnology* **30**:271–277. DOI: https://doi.org/10.1038/nbt.2137, PMID: 22371084

Montgomery SH, Mundy NI, Barton RA. 2016. Brain evolution and development: adaptation, allometry and constraint. *Proceedings of the Royal Society B* **283**:20160433. DOI: https://doi.org/10.1098/rspb.2016.0433

Nakai R, Ohnuki M, Kuroki K, Ito H, Hirai H, Kitajima R, Fujimoto T, Nakagawa M, Enard W, Imamura M. 2018. Derivation of induced pluripotent stem cells in Japanese macaque (*Macaca fuscata*). *Scientific Reports* **8**:12187. DOI: https://doi.org/10.1038/s41598-018-30734-w, PMID: 30111816

Nei M, Nei M, Kumar S. 2000. Molecular Evolution and Phylogenetics. Oxford University Press.

Ohta T. 1987. Very slightly deleterious mutations and the molecular clock. *Journal of Molecular Evolution* **26**:1–6. DOI: https://doi.org/10.1007/BF02111276, PMID: 3125329

Parekh S, Ziegenhain C, Vieth B, Enard W, Hellmann I. 2018. ZUMIs-a fast and flexible pipeline to process RNA sequencing data with umis. *GigaScience* **7**:giy059. DOI: https://doi.org/10.1093/gigascience/giy059, PMID: 29846586

Pilz GA, Shitamukai A, Reillo I, Pacary E, Schwausch J, Stahl R, Ninkovic J, Snippert HJ, Clevers H, Godinho L, Guillemot F, Borrell V, Matsuzaki F, Götz M. 2013. Amplification of progenitors in the mammalian telencephalon includes a new radial glial cell type. *Nature Communications* **4**:2125. DOI: https://doi.org/10.1038/ncomms3125, PMID: 23839311

Pinson A, Huttner WB. 2021. Neocortex expansion in development and evolution-from genes to progenitor cell biology. *Current Opinion in Cell Biology* **73**:9–18. DOI: https://doi.org/10.1016/j.ceb.2021.04.008, PMID: 34098196

Pinson A, Xing L, Namba T, Kalebic N, Peters J, Oegema CE, Traikov S, Reppe K, Riesenberg S, Maricic T, Derihaci R, Wimberger P, Pääbo S, Huttner WB. 2022. Human TKTL1 implies greater neurogenesis in frontal neocortex of modern humans than neanderthals. Science **377**:eabl6422. DOI: https://doi.org/10.1126/science.abl6422, PMID: 36074851

Pujar S, O'Leary NA, Farrell CM, Loveland JE, Mudge JM, Wallin C, Girón CG, Diekhans M, Barnes I, Bennett R, Berry AE, Cox E, Davidson C, Goldfarb T, Gonzalez JM, Hunt T, Jackson J, Joardar V, Kay MP, Kodali VK, et al. 2018. Consensus coding sequence (CCDS) database: A standardized set of human and mouse protein-coding regions supported by expert curation. *Nucleic Acids Research* **46**:D221–D228. DOI: https://doi.org/10.1093/nar/gkx1031

R Development Core Team. 2019. R: A language and environment for statistical computing. Vienna, Austria. R Foundation for Statistical Computing. https://www.r-project.org/index.html

Reader SM, Hager Y, Laland KN. 2011. The evolution of primate general and cultural intelligence. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* **366**:1017–1027. DOI: https://doi.org/10.1098/rstb.2010.0342, PMID: 21357224

Reilly SK, Yin J, Ayoub AE, Emera D, Leng J, Cotney J, Sarro R, Rakic P, Noonan JP. 2015. Evolutionary genomics: evolutionary changes in promoter and enhancer activity during human corticogenesis. *Science* **347**:1155–1159. DOI: https://doi.org/10.1126/science.1260943, PMID: 25745175

Renaud G, Stenzel U, Maricic T, Wiebe V, Kelso J. 2015. DeML: robust demultiplexing of illumina sequences using a likelihood-based approach. *Bioinformatics* **31**:770–772. DOI: https://doi.org/10.1093/bioinformatics/btu719, PMID: 25359895

Rhie A, McCarthy SA, Fedrigo O, Damas J, Formenti G, Koren S, Uliano-Silva M, Chow W, Fungtammasan A, Kim J, Lee C, Ko BJ, Chaisson M, Gedman GL, Cantin LJ, Thibaud-Nissen F, Haggerty L, Bista I, Smith M, Haase B, et al. 2021. Towards complete and error-free genome assemblies of all vertebrate species. *Nature* **592**:737–746. DOI: https://doi.org/10.1038/s41586-021-03451-0, PMID: 33911273

Sedlazeck FJ, Rescheneder P, von Haeseler A. 2013. NextGenMap: Fast and accurate read mapping in highly polymorphic genomes. *Bioinformatics* **29**:2790–2791. DOI: https://doi.org/10.1093/bioinformatics/btt468, PMID: 23975764

Smaers JB, Rothman RS, Hudson DR, Balanoff AM, Beatty B, Dechmann DKN, de Vries D, Dunn JC, Fleagle JG, Gilbert CC, Goswami A, Iwaniuk AN, Jungers WL, Kerney M, Ksepka DT, Manger PR, Mongle CS, Rohlf FJ, Smith NA, Soligo C, et al. 2021. The evolution of mammalian brain size. *Science Advances* **7**:18. DOI: https://doi.org/10.1126/sciadv.abe2101, PMID: 33910907

Smith SD, Pennell MW, Dunn CW, Edwards SV. 2020. Phylogenetics is the new genetics (for most of biodiversity). *Trends in Ecology & Evolution* **35**:415–425. DOI: https://doi.org/10.1016/j.tree.2020.01.005, PMID: 32294423

Stahl R, Walcher T, De Juan Romero C, Pilz GA, Cappello S, Irmler M, Sanz-Aquela JM, Beckers J, Blum R, Borrell V, Götz M. 2013. Trnp1 regulates expansion and folding of the mammalian cerebral cortex by control of radial glial fate. *Cell* **153**:535–549. DOI: https://doi.org/10.1016/j.cell.2013.03.027, PMID: 23622239

Stephan T, Burgess SM, Cheng H, Danko CG, Gill CA, Jarvis ED, Koepfli KP, Koltes JE, Lyons E, Ronald P, Ryder OA, Schriml LM, Soltis P, VandeWoude S, Zhou H, Ostrander EA, Karlsson EK. 2022. Darwinian genomics and diversity in the tree of life. *PNAS* **119**:e2115644119. DOI: https://doi.org/10.1073/pnas.2115644119, PMID: 35042807

Suzuki IK, Gacquer D, Van Heurck R, Kumar D, Wojno M, Bilheu A, Herpoel A, Lambert N, Cheron J, Polleux F, Detours V, Vanderhaeghen P. 2018. Human-Specific NOTCH2NL genes expand cortical neurogenesis through delta/notch regulation. *Cell* **173**:1370–1384.. DOI: https://doi.org/10.1016/j.cell.2018.03.067, PMID: 29856955

Tavano S, Taverna E, Kalebic N, Haffner C, Namba T, Dahl A, Wilsch-Bräuninger M, Paridaen JTML, Huttner WB. 2018. Insm1 induces neural progenitor delamination in developing neocortex via downregulation of the

adherens junction belt-specific protein PLEKHA7. *Neuron* **97**:1299–1314.. DOI: https://doi.org/10.1016/j.neuron.2018.01.052, PMID: 29503187

Trevino AE, Müller F, Andersen J, Sundaram L, Kathiria A, Shcherbina A, Farh K, Chang HY, Paşca AM, Kundaje A, Paşca SP, Greenleaf WJ. 2021. Chromatin and gene-regulatory dynamics of the developing human cerebral cortex at single-cell resolution. *Cell* **184**:5053–5069.. DOI: https://doi.org/10.1016/j.cell.2021.07.039, PMID: 34390642

UniProt Consortium. 2019. UniProt: A worldwide hub of protein knowledge. *Nucleic Acids Research* **47**:D506–D515. DOI: https://doi.org/10.1093/nar/gky1049, PMID: 30395287

Vasimuddin M, Misra S, Li H, Aluru S. 2019. Efficient Architecture-Aware Acceleration of BWA-MEM for Multicore Systems. 2019 IEEE International Parallel and Distributed Processing Symposium (IPDPS. Rio de Janeiro, Brazil. DOI: https://doi.org/10.1109/IPDPS.2019.00041

Vierstra J, Rynes E, Sandstrom R, Zhang M, Canfield T, Hansen RS, Stehling-Sun S, Sabo PJ, Byron R, Humbert R, Thurman RE, Johnson AK, Vong S, Lee K, Bates D, Neri F, Diegel M, Giste E, Haugen E, Dunn D, et al. 2014. Mouse regulatory DNA landscapes reveal global principles of cis-regulatory evolution. Science **346**:1007–1012. DOI: https://doi.org/10.1126/science.1246426, PMID: 25411453

Villalba A, Götz M, Borrell V. 2021. The regulation of cortical neurogenesis. *Current Topics in Developmental Biology* **142**:1–66. DOI: https://doi.org/10.1016/bs.ctdb.2020.10.003, PMID: 33706916

Volpe M, Shpungin S, Barbi C, Abrham G, Malovani H, Wides R, Nir U. 2006. Trnp: a conserved mammalian gene encoding a nuclear protein that accelerates cell-cycle progression. *DNA and Cell Biology* **25**:331–339. DOI: https://doi.org/10.1089/dna.2006.25.331, PMID: 16792503

Watson LA, Wang X, Elbert A, Kernohan KD, Galjart N, Bérubé NG. 2014. Dual effect of CTCF loss on neuroprogenitor differentiation and survival. *The Journal of Neuroscience* **34**:2860–2870. DOI: https://doi.org/10.1523/JNEUROSCI.3769-13.2014, PMID: 24553927

Won H, de la Torre-Ubieta L, Stein JL, Parikshak NN, Huang J, Opland CK, Gandal MJ, Sutton GJ, Hormozdiari F, Lu D, Lee C, Eskin E, Voineagu I, Ernst J, Geschwind DH. 2016. Chromosome conformation elucidates regulatory relationships in developing human brain. *Nature* **538**:523–527. DOI: https://doi.org/10.1038/nature19847, PMID: 27760116

Wu D, Li T, Lu Z, Dai W, Xu M, Lu L. 2006. Effect of CTCF-binding motif on regulation of Pax6 transcription. *Investigative Ophthalmology & Visual Science* **47**:2422–2429. DOI: https://doi.org/10.1167/iovs.05-0536, PMID: 16723452

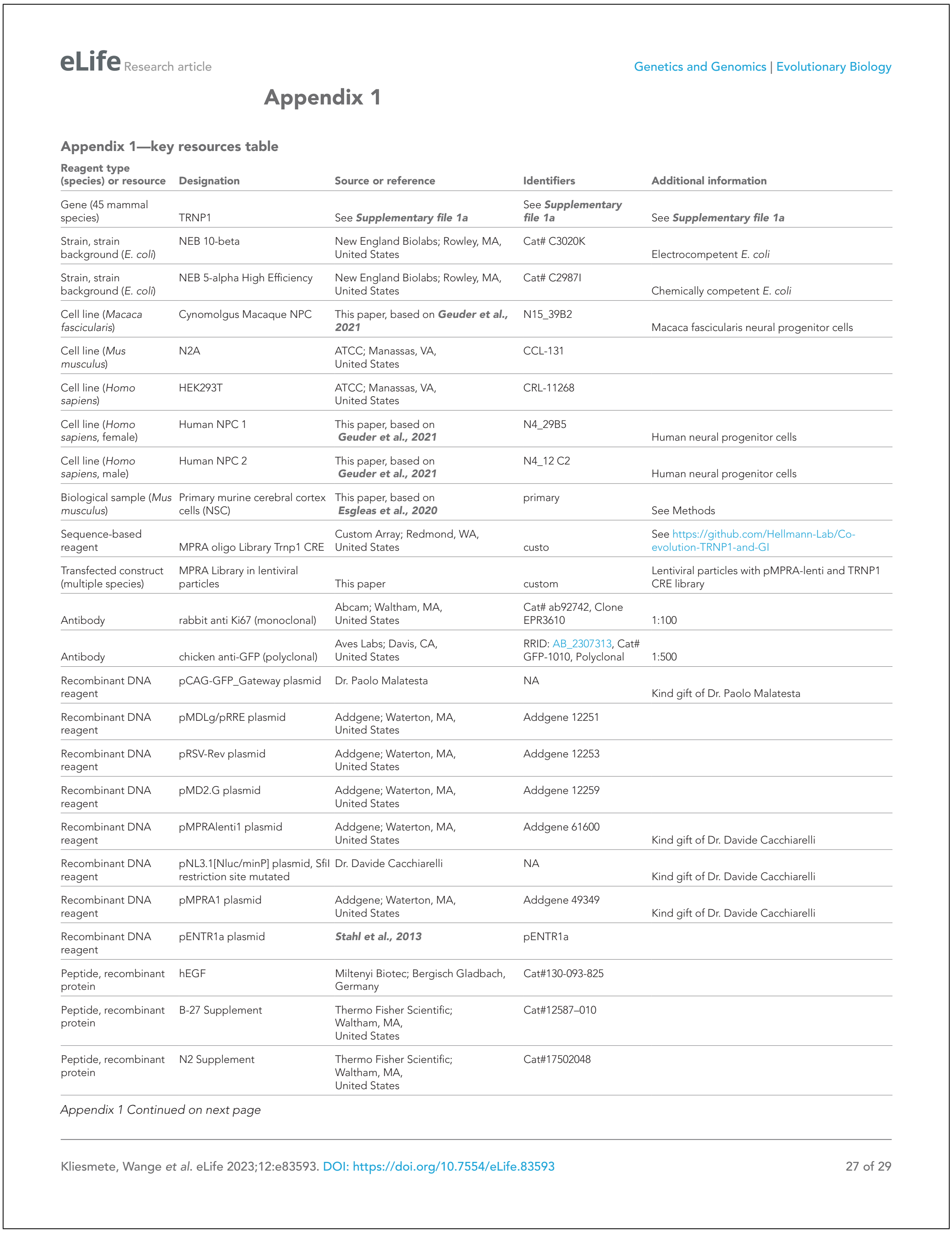
Yang Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. *Computer Applications in the Biosciences* **13**:555–556. DOI: https://doi.org/10.1093/bioinformatics/13.5.555, PMID: 9367129

Yang Z, Nielsen R, Goldman N, Pedersen AM. 2000. Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics* **155**:431–449. DOI: https://doi.org/10.1093/genetics/155.1.431, PMID: 10790415

Yang Z. 2006. Computational Molecular Evolution. Oxford: OUP. DOI: https://doi.org/10.1093/acprof:oso/9780198567028.001.0001

Ye J, Coulouris G, Zaretskaya I, Cutcutache I, Rozen S, Madden TL. 2012. Primer-BLAST: A tool to design target-specific primers for polymerase chain reaction. *BMC Bioinformatics* **13**:134. DOI: https://doi.org/10.1186/1471-2105-13-134, PMID: 22708584

Zilles K, Armstrong E, Moser KH, Schleicher A, Stephan H. 1989. Gyrification in the cerebral cortex of primates. *Brain, Behavior and Evolution* **34**:143–150. DOI: https://doi.org/10.1159/000116500, PMID: 2512000

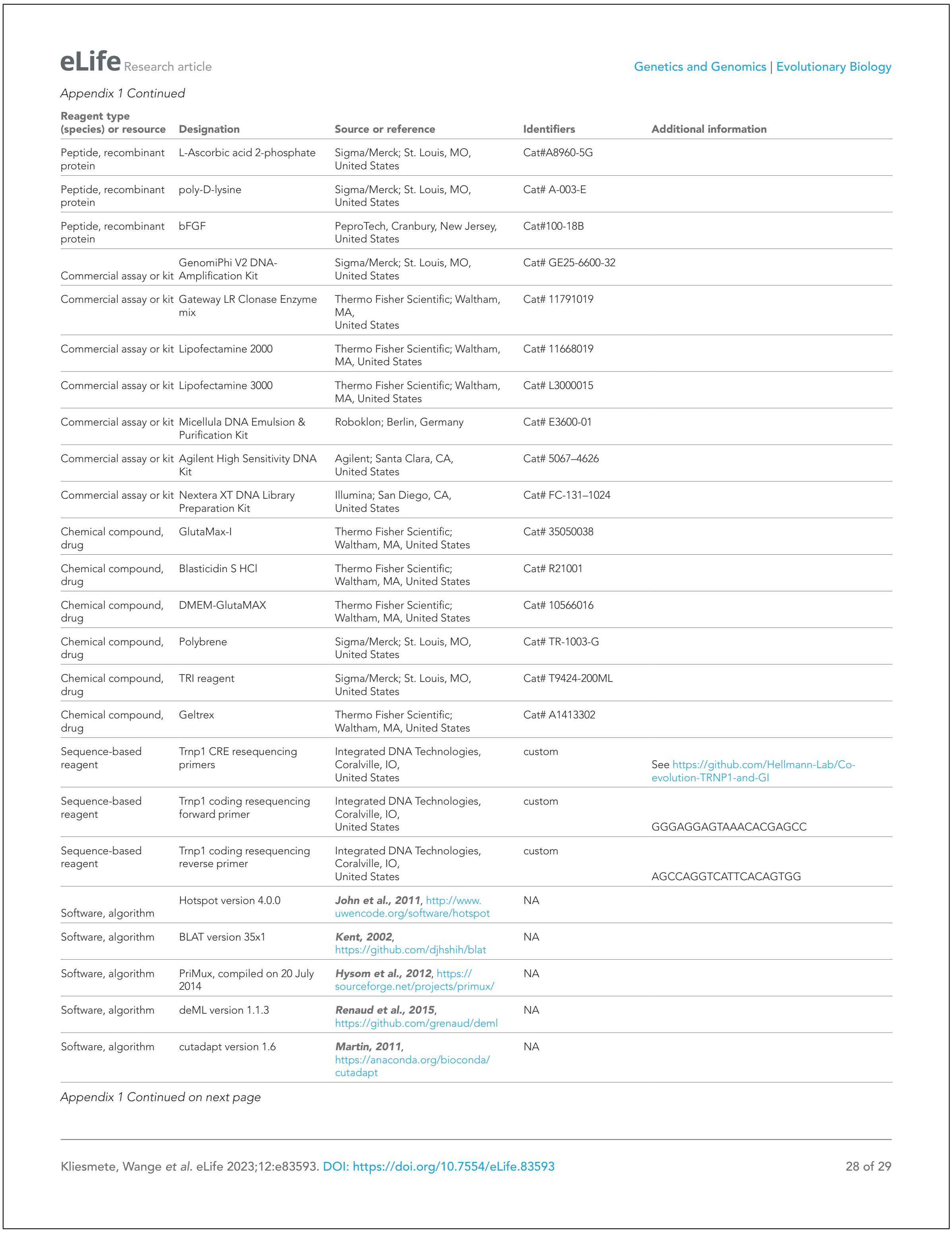


Appendix 1

Appendix 1—key resources table

Reagent type (species) or resource	Designation	Source or reference	Identifiers	Additional information
Gene (45 mammal species)	TRNP1	See ***Supplementary file 1a***	See ***Supplementary file 1a***	See ***Supplementary file 1a***
Strain, strain background (*E. coli*)	NEB 10-beta	New England Biolabs; Rowley, MA, United States	Cat# C3020K	Electrocompetent *E. coli*
Strain, strain background (*E. coli*)	NEB 5-alpha High Efficiency	New England Biolabs; Rowley, MA, United States	Cat# C2987I	Chemically competent *E. coli*
Cell line (*Macaca fascicularis*)	Cynomolgus Macaque NPC	This paper, based on ***Geuder et al., 2021***	N15_39B2	Macaca fascicularis neural progenitor cells
Cell line (*Mus musculus*)	N2A	ATCC; Manassas, VA, United States	CCL-131	
Cell line (*Homo sapiens*)	HEK293T	ATCC; Manassas, VA, United States	CRL-11268	
Cell line (*Homo sapiens*, female)	Human NPC 1	This paper, based on ***Geuder et al., 2021***	N4_29B5	Human neural progenitor cells
Cell line (*Homo sapiens*, male)	Human NPC 2	This paper, based on ***Geuder et al., 2021***	N4_12 C2	Human neural progenitor cells
Biological sample (*Mus musculus*)	Primary murine cerebral cortex cells (NSC)	This paper, based on ***Esgleas et al., 2020***	primary	See Methods
Sequence-based reagent	MPRA oligo Library Trnp1 CRE	Custom Array; Redmond, WA, United States	custo	See https://github.com/Hellmann-Lab/Co-evolution-TRNP1-and-GI
Transfected construct (multiple species)	MPRA Library in lentiviral particles	This paper	custom	Lentiviral particles with pMPRA-lenti and TRNP1 CRE library
Antibody	rabbit anti Ki67 (monoclonal)	Abcam; Waltham, MA, United States	Cat# ab92742, Clone EPR3610	1:100
Antibody	chicken anti-GFP (polyclonal)	Aves Labs; Davis, CA, United States	RRID: AB_2307313, Cat# GFP-1010, Polyclonal	1:500
Recombinant DNA reagent	pCAG-GFP_Gateway plasmid	Dr. Paolo Malatesta	NA	Kind gift of Dr. Paolo Malatesta
Recombinant DNA reagent	pMDLg/pRRE plasmid	Addgene; Waterton, MA, United States	Addgene 12251	
Recombinant DNA reagent	pRSV-Rev plasmid	Addgene; Waterton, MA, United States	Addgene 12253	
Recombinant DNA reagent	pMD2.G plasmid	Addgene; Waterton, MA, United States	Addgene 12259	
Recombinant DNA reagent	pMPRAlenti1 plasmid	Addgene; Waterton, MA, United States	Addgene 61600	Kind gift of Dr. Davide Cacchiarelli
Recombinant DNA reagent	pNL3.1[Nluc/minP] plasmid, SfiI restriction site mutated	Dr. Davide Cacchiarelli	NA	Kind gift of Dr. Davide Cacchiarelli
Recombinant DNA reagent	pMPRA1 plasmid	Addgene; Waterton, MA, United States	Addgene 49349	Kind gift of Dr. Davide Cacchiarelli
Recombinant DNA reagent	pENTR1a plasmid	***Stahl et al., 2013***	pENTR1a	
Peptide, recombinant protein	hEGF	Miltenyi Biotec; Bergisch Gladbach, Germany	Cat#130-093-825	
Peptide, recombinant protein	B-27 Supplement	Thermo Fisher Scientific; Waltham, MA, United States	Cat#12587–010	
Peptide, recombinant protein	N2 Supplement	Thermo Fisher Scientific; Waltham, MA, United States	Cat#17502048	

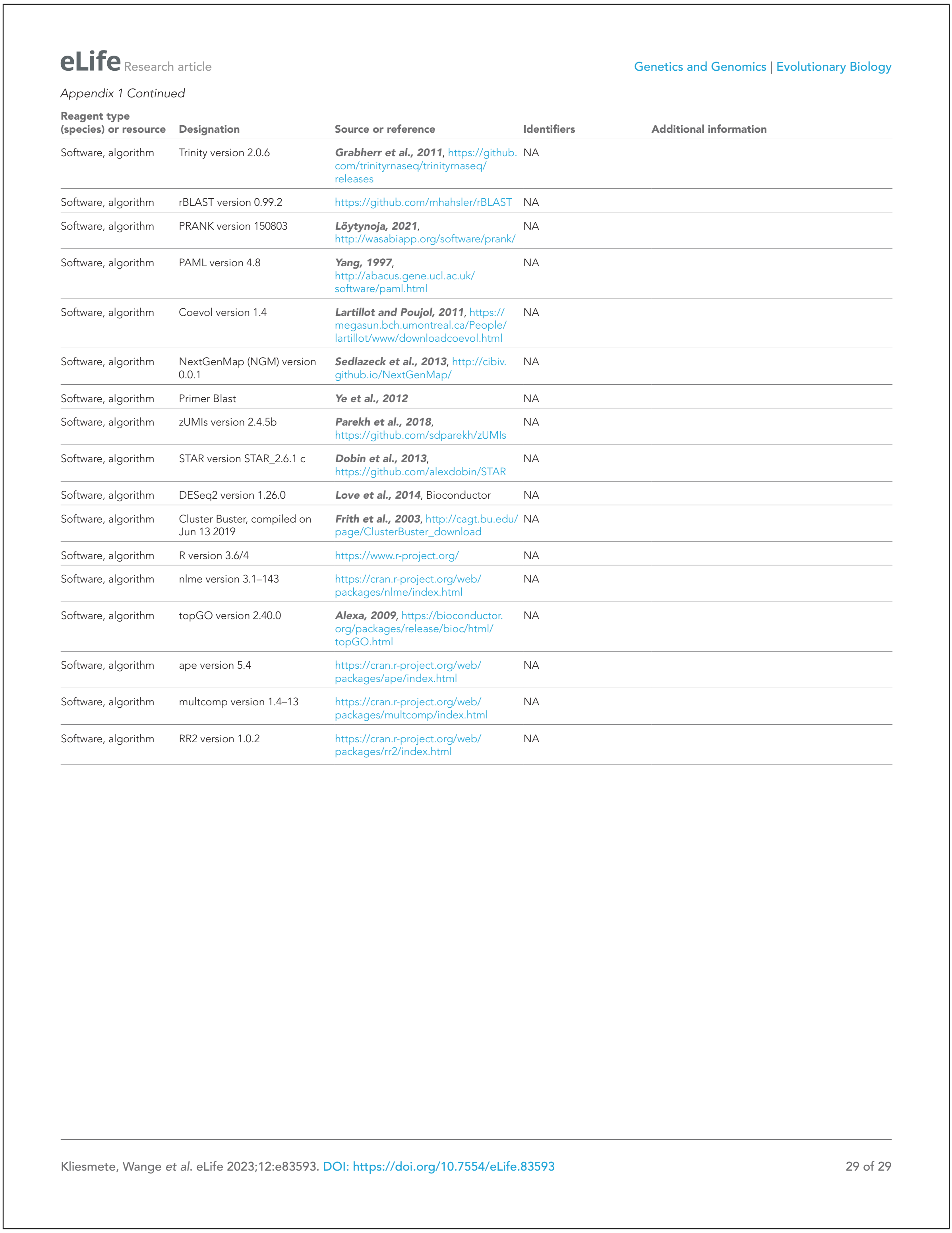
Appendix 1 Continued on next page



Appendix 1 Continued

Reagent type (species) or resource	Designation	Source or reference	Identifiers	Additional information
Peptide, recombinant protein	L-Ascorbic acid 2-phosphate	Sigma/Merck; St. Louis, MO, United States	Cat#A8960-5G	
Peptide, recombinant protein	poly-D-lysine	Sigma/Merck; St. Louis, MO, United States	Cat# A-003-E	
Peptide, recombinant protein	bFGF	PeproTech, Cranbury, New Jersey, United States	Cat#100-18B	
Commercial assay or kit	GenomiPhi V2 DNA-Amplification Kit	Sigma/Merck; St. Louis, MO, United States	Cat# GE25-6600-32	
Commercial assay or kit	Gateway LR Clonase Enzyme mix	Thermo Fisher Scientific; Waltham, MA, United States	Cat# 11791019	
Commercial assay or kit	Lipofectamine 2000	Thermo Fisher Scientific; Waltham, MA, United States	Cat# 11668019	
Commercial assay or kit	Lipofectamine 3000	Thermo Fisher Scientific; Waltham, MA, United States	Cat# L3000015	
Commercial assay or kit	Micellula DNA Emulsion & Purification Kit	Roboklon; Berlin, Germany	Cat# E3600-01	
Commercial assay or kit	Agilent High Sensitivity DNA Kit	Agilent; Santa Clara, CA, United States	Cat# 5067–4626	
Commercial assay or kit	Nextera XT DNA Library Preparation Kit	Illumina; San Diego, CA, United States	Cat# FC-131–1024	
Chemical compound, drug	GlutaMax-I	Thermo Fisher Scientific; Waltham, MA, United States	Cat# 35050038	
Chemical compound, drug	Blasticidin S HCl	Thermo Fisher Scientific; Waltham, MA, United States	Cat# R21001	
Chemical compound, drug	DMEM-GlutaMAX	Thermo Fisher Scientific; Waltham, MA, United States	Cat# 10566016	
Chemical compound, drug	Polybrene	Sigma/Merck; St. Louis, MO, United States	Cat# TR-1003-G	
Chemical compound, drug	TRI reagent	Sigma/Merck; St. Louis, MO, United States	Cat# T9424-200ML	
Chemical compound, drug	Geltrex	Thermo Fisher Scientific; Waltham, MA, United States	Cat# A1413302	
Sequence-based reagent	Trnp1 CRE resequencing primers	Integrated DNA Technologies, Coralville, IO, United States	custom	See https://github.com/Hellmann-Lab/Co-evolution-TRNP1-and-GI
Sequence-based reagent	Trnp1 coding resequencing forward primer	Integrated DNA Technologies, Coralville, IO, United States	custom	GGGAGGAGTAAACACGAGCC
Sequence-based reagent	Trnp1 coding resequencing reverse primer	Integrated DNA Technologies, Coralville, IO, United States	custom	AGCCAGGTCATTCACAGTGG
Software, algorithm	Hotspot version 4.0.0	***John et al., 2011***, http://www.uwencode.org/software/hotspot	NA	
Software, algorithm	BLAT version 35x1	***Kent, 2002***, https://github.com/djhshih/blat	NA	
Software, algorithm	PriMux, compiled on 20 July 2014	***Hysom et al., 2012***, https://sourceforge.net/projects/primux/	NA	
Software, algorithm	deML version 1.1.3	***Renaud et al., 2015***, https://github.com/grenaud/deml	NA	
Software, algorithm	cutadapt version 1.6	***Martin, 2011***, https://anaconda.org/bioconda/cutadapt	NA	

Appendix 1 Continued on next page



Appendix 1 Continued

Reagent type (species) or resource	Designation	Source or reference	Identifiers	Additional information
Software, algorithm	Trinity version 2.0.6	***Grabherr et al., 2011***, https://github.com/trinityrnaseq/trinityrnaseq/releases	NA	
Software, algorithm	rBLAST version 0.99.2	https://github.com/mhahsler/rBLAST	NA	
Software, algorithm	PRANK version 150803	***Löytynoja, 2021***, http://wasabiapp.org/software/prank/	NA	
Software, algorithm	PAML version 4.8	***Yang, 1997***, http://abacus.gene.ucl.ac.uk/software/paml.html	NA	
Software, algorithm	Coevol version 1.4	***Lartillot and Poujol, 2011***, https://megasun.bch.umontreal.ca/People/lartillot/www/downloadcoevol.html	NA	
Software, algorithm	NextGenMap (NGM) version 0.0.1	***Sedlazeck et al., 2013***, http://cibiv.github.io/NextGenMap/	NA	
Software, algorithm	Primer Blast	***Ye et al., 2012***	NA	
Software, algorithm	zUMIs version 2.4.5b	***Parekh et al., 2018***, https://github.com/sdparekh/zUMIs	NA	
Software, algorithm	STAR version STAR_2.6.1 c	***Dobin et al., 2013***, https://github.com/alexdobin/STAR	NA	
Software, algorithm	DESeq2 version 1.26.0	***Love et al., 2014***, Bioconductor	NA	
Software, algorithm	Cluster Buster, compiled on Jun 13 2019	***Frith et al., 2003***, http://cagt.bu.edu/page/ClusterBuster_download	NA	
Software, algorithm	R version 3.6/4	https://www.r-project.org/	NA	
Software, algorithm	nlme version 3.1–143	https://cran.r-project.org/web/packages/nlme/index.html	NA	
Software, algorithm	topGO version 2.40.0	***Alexa, 2009***, https://bioconductor.org/packages/release/bioc/html/topGO.html	NA	
Software, algorithm	ape version 5.4	https://cran.r-project.org/web/packages/ape/index.html	NA	
Software, algorithm	multcomp version 1.4–13	https://cran.r-project.org/web/packages/multcomp/index.html	NA	
Software, algorithm	RR2 version 1.0.2	https://cran.r-project.org/web/packages/rr2/index.html	NA	

3 | Discussion

There has been a long standing interest for biologists from different fields to answer questions related to the role of species-specific elements in generating species-specific functional novelties, deriving the function and importance of tissue-specific and pleiotropic elements for complex multicellular organisms and connecting genetic changes to molecular or organismal phenotypes. With the technological advances made in the last decades and the large amounts of recently available data, we can now revisit these questions on a genome-wide scale in a less biased manner than ever before. Genome-wide assays on different functional levels can be combined to get a more complete picture of the patterns of genome evolution, not limited to a few selected genetic regions or model organisms.

However, analysing such data also proposes challenges of its own. Firstly, the validity and the amount of error made using recent technologies need to be assessed. Secondly, workflows for unbiased cross-species comparisons using the newly emerging data types need to be established. Thirdly, the information from different modalities needs to be incorporated in a meaningful, informative way. Finally, evolutionary frameworks have to be selected to accommodate the types of available omics data. In this thesis, I tackled these challenges and addressed case-specific questions on the evolution and importance of regulatory pleiotropy, newly emerging elements and the association between genetic and phenotypic change.

3.1 Error rate estimation in RNA-seq assays using cross-species genetic variation

When interpreting data from high-throughput assays like (single-cell) RNA-seq that simultaneously yield information on expression from multiple cells/samples and conditions, we generally assume that the counted RNA molecules come from the cell or sample they are assigned to. However, as any method, RNA-seq is not perfect. Errors affecting the precision in the measurements can come from the process of amplification of the cDNA molecules during library generation and in droplet-based methods, such as 10x Chromium, also from freely swimming RNA molecules present in the sample that arise through damaged cell bursts [209,210]. If these errors are random across the different cells/samples and RNA molecules, this should lead to shifts in the detected transcripts towards the mean. Random noise can decrease the power to detect differentially expressed genes between conditions or marker genes of certain cell types [211,212]. In addition to random errors, some RNA molecules might be more likely to swap during the amplification process, potentially generating non-uniformly distributed presence of chimeric molecules [213,214]. Such non-random errors across genes or cells can lead to biases in the expression profiles.

To account for unequal amplification of sequences, adding a random RNA molecule-specific barcode, called unique molecular identifier (UMI), during cDNA generation serves as a molecular stamp [215,216,217]. This allows to trace back the original RNA molecule and thereby avoid counting the same molecule multiple times. However, it does not correct for the RNA molecules that are swapped during the amplification or that get assigned a cellular barcode although they come from extracellular sources. Available approaches to quantify background noise rely on marker gene expression [218], BC-UMI-gene complexity [214] or RNA quantification in empty droplets [209,210] for droplet-based methods. We used a cross-species setup [219,210] that can offer further insight by combining samples from different closely related species: Based on the sequence of the transcript and (known) substitutions, the (sub-)species and thereby the sample origin of the RNA molecule can be identified. If it mismatches the sample origin of the majority of the reads carrying the same barcode, this molecule likely did not originally come from the same sample/cell.

Using these principles, we quantified the error made across samples generated using droplet-based RNA-seq and benchmarked methods that aim to remove background noise. Hence, genetic divergence is not only informative for understanding selective forces acting on genetic elements and thereby inferring functional importance, but it can also help in answering more technical questions regarding the reliability of the recent techniques we use for measuring different modalities. Using a similar approach, the error present in other genome-wide assays such as (single-cell) ATAC-seq could be further estimated in the future, including the benchmarking of recently emerging background removal methods[220].

3.2 Regulatory code as revealed through stratification of tissue-specificity

As a part of this thesis, I systematically studied the effects of pleiotropy on CRE conservation in primates across multiple functional levels. Contrary to genes, pleiotropic CREs show lower sequence conservation than tissue-specific CREs. Pleiotropic degree (PD) also goes along with increasing CRE width and CpG island content. Noteworthy, by distinguishing between types of di-nucleotide substitutions, we found that pleiotropic CREs show a decrease in CpG-depleting and an increase in non-CpG-related and CpG-creating substitutions, suggesting an underlying mechanism that facilitates constant di-nucleotide content of the sequences. A comparison of CpG observed / expected ratio between orthologous CREs from human and macaque validates that this property is indeed better conserved in the pleiotropic CREs. I further investigated transcription factor binding site conservation between orthologous CREs and found that the exact binding positions are also less conserved in pleiotropic CREs than in any other PD group, including tissue-specific CREs. According to the Billboard model [184,185,188,221], whether the required TFs bind a particular CRE a few tens of bases up- or downstream within the CRE sequence might not make much of a difference in many cases, as long as the TF repertoire is contained. Indeed, TFBS repertoire, measured as the cumulative binding potential per motif across the different TFs, is highly conserved at the pleiotropic CREs. This higher-level TFBS property also appears to induce highly conserved downstream gene expression, most of which also show pleiotropic expression patterns across tissues.

Pleiotropic CREs are enriched in GC-rich TF binding motifs belonging to the 'Stripe factor' class[32], that were experimentally shown to stabilize and prolong the binding of other TFs and thereby CRE accessibility. The binding of these and similar TFs could be the key to connect the different patterns we observe: CpG conservation facilitates conserved binding of the accessibility-stabilizing TFs leading to conserved expression. The larger width of the pleiotropic CREs assures the presence of sufficiently strong binding sites for the expressed TFs across different cellular environments. The exact binding position does not matter, hence the sequence conservation is relatively low[222].

At this point, we can speculate about the likely evolutionary mechanisms underlying the observed patterns. Given the known features of CRE landscapes, the expected selection coefficients associated with each individual binding site are, in average, not high. This creates a fertile ground for compensatory evolution[223], where a weakly deleterious loss of a binding site due to drift might be compensated by fixation of one out of multiple possible compensatory binding sites, each of which could lead to an equally good fitness [224]. This facilitates a potential existance of multiple equally fit haplotypes. During gradual species divergence, it is not unlikely that for many CREs, a different similarly fit haplotype accidentally becomes the most frequent one. Experimental evidence supports such a scenario, where orthologous CRE sequences from related species with diverged binding positions but conserved binding repertoires lead to highly conserved downstream expression[225,226]. When assessing a hybrid sequence containing half of each orthologous CRE, and thereby both binding sites, it leads to over-expression. This is in agreement with the idea that the total binding potential matters more than the exact position of individual sites. Such a general mechanism would also explain how the seemingly lowly conserved CRE sequences can achieve highly conserved expression patterns in the case of pleiotropic house-keeping genes that are governed by particularly diverged CRE sequences. The existing theoretical models simulating CRE evolution and turnover of TFBS support the notion that the drafted mechanism is common[120,36]. Further simulations that incorporate the PD and our multi-level characterization of primate CREs, including the downstream gene expression, would be an informative further step to understand the within-CRE functional compensation.

The high evolutionary turnover of tissue-specific elements is another highly interesting aspect that could be further studied. Our and many other studies[198,226,37,202,48] have found hints towards potential between-CRE compensation across tissue-specific elements. In order to understand to what extent a species- and tissue-specific CRE is compensated by another CRE in another species or cellular context, we need to map and characterize the CRE landscapes associated with a specific gene across multiple species and employ more sophisticated models that build up on previous work[120,227]. By simultaneously considering various aspects of higher and lower level functional similarity between CREs could help identify functionally-orthologous CREs that might not necessarily be the sequence orthologues. Also the quantification of CRE activity of each individual element of the regulatory landscape is a helpful aspect to include in future statistical models[228].

3.3 The role of TE-derived elements in species-specific rewiring of gene regulation

In the previous section, I discussed how compensatory evolution acting on CRE sequences shapes the landscape of gene regulation. Compensatory evolution can also be relevant for coping with the disruptions imposed by transposable elements (TEs)[229]. There is an ongoing and controversial discussion in the scientific community about the role of TE elements in general, and LTR elements specifically, in rewiring gene expression networks in primates. There are different views, some proposing that successful adaptation of the genome is highly dependent on TE-derived sequences and the molecular novelties they induce[76,103,101,230]. Arguments supporting this possibility include the fact that the *de novo* inserted sequences initially do not have a concrete, important role in gene regulatory networks. Therefore, these might be more amenable to drift and positive selection to facilitate adaptations to changing environments or to compensate for other slightly deleterious changes in the genome. Another, more sceptical view mainly considers TE insertions as neutral or destabilizing events for the genome and its evolution. While TE activity might lead to new transcripts in some cells, it is not necessarily indicative of function. Even if the TE insertion in a few cases indeed leads to changes in expression networks, it is merely to compensate for the inconveniences that were introduced through the perturbations. Accidentally, the newly generated CRE or transcript might even be chosen over the original pathway, however this rewiring is more of an obstacle rather than an innovation[229,231].

Given the fitness landscapes observed in lower complexity organisms that are easier to manipulate in the lab, some truth probably lies in both of these views. The expected proportion of the alleles that are immediately fixed due to positive selection is low[79]. The considerable expansion of TE elements in primates is likely a consequence of the small effective population sizes that enable higher fixation of slightly-deleterious variants because of the large impact of genetic drift. Still, fixation of slightly deleterious alleles can also lead to secondary fixations that might in long term prove advantageous for the species. In addition, the estimates that as much as 75% of our genome might actually have emerged as a result of TE activity allows for the possibility that we still underestimate the contribution of TE sequences for regulatory and gene coding sequences in a longer evolutionary run. Moreover, as discussed in the previous chapter, the constraint on the exact genetic location or exact sequence in the case of CREs appears to be rather low. This sets a perfect stage for the emergence of potentially functional TE-derived CREs, particularly from LTR elements that carry TFBS.

I contributed to deciphering the evolutionary and functional importance of a human endogenous retrovirus type-H (HERVH)-derived long non-coding RNA called Embryonic Stem Cell Related Gene (*ESRG*). Its promoter, the whole exon 1 and a part of exon 2 are LTR7-derived and contain binding sites for pluripotency factors such as OCT3/4. *ESRG* was identified through its high and specific expression in human ESCs and iPSCs[232,233,234]. Previous studies, based on knock-downs of *ESRG*, had concluded that its expression is required for the maintenance of pluripotency and self-renewal[232,233]. In this study, *ESRG* contribution to pluripotency was investigated using knock-outs combined with differentiation assays and evolutionary approaches. None of the previously suggested molecular phenotypes could be experimentally captured using independent replicates of the complete *ESRG* locus deletion. Still, it can be argumented that differences in experimental setups or the knock-out strategy could affect the cell state to begin with[235]. Here, evolutionary and population genetic approaches can be helpful to further investigate the functional relevance of an element beyond relying on a specific cellular context. Like many TE-derived elements proposed to possess clade- or species-specific function, *ESRG* is only present in few species - humans, bonobos and chimpanzees, but not in the other primates. This limits the analysis and thereby the statistical power to frequency-based metrics and contrasting rates of substitutions to rates of polymorphisms. Using the available large human-polymorphism database gnomAD[236], we compared polymorphism frequencies and divergence between *ESRG* exons vs. introns and to coding gene and other lncRNA sequences to detect signatures of selection. None of the comparisons showed compelling evidence for selection, aligning with the lack of experimental evidence for *ESRG* role in pluripotency networks. Moreover, although *ESRG* shows among the 5% highest expression levels across all genes in human iPCSs, its high expression is not present in the chimpanzee iPSCs, suggesting non-conserved expression patterns.

Although the functional relevance of *ESRG* is still being debated[235,237], this study exemplifies the general importance of multiple functional validations, ideally by independent research groups using independent experimental setups and the added value of evolutionary and population genetic approaches. It also shows that the presence of TF binding sites and high expression levels is not necessarily equivalent to function[231]. Overall, the extent to which TE-derived species- or clade-specific elements are responsible for the emergence of species-specific molecular networks and phenotypes is still unclear. Improved long read sequencing that allows for improved mapping of these elements across different species, as well as cellular assays and silencing technologies like CRISPR/Cas9 are aiding in rapid progress to further clarify the roles of many of these elements.

3.4 The central task of studying genotype contribution to phenotypes

The mammalian brain is arguably among the most interesting tissues to investigate using comparative approaches, because it is linked to cognitive abilities and behavioural complexity [238] and because of its remarkable diversity. Particularly the outer layer called cerebral cortex shows an extraordinary phenotypic diversity in size and shape across vertebrates[239], reaching its highest complexity on the mammalian branch where cortical folding has emerged [240,241]. Also within mammals, brain size and folding show extensive variation, including recurrent independent increases and decreases[242,243,244]. This natural variability can be used to investigate the genetic sources of this intricate phenotype[245].

TMF-regulated nuclear protein 1 (TRNP1) is known to be essential for cortical development in model organisms like ferrets[246] and mice[247,248,249] by controlling neural stem cell proliferation[250,251]. Its knock-down as well as over-expression have clear phenotypic effects on the resulting cortical size and folding. The prior knowledge on the decisive role of TRNP1 in brain developmental processes puts TRNP1 among the prime candidates to study across the mammalian phylogeny. Cross-species genotype-phenotype association studies allow to investigate to what extent conclusions from experimental findings in few species can be extended to a larger phylogeny by performing sequence, regulatory and cellular activity analyses.

These types of analyses impose multiple challenges. To study protein-coding sequence co-evolution with a trait[157], high-quality coding sequences and evolutionarily most plausible alignments need to be generated[252]. It is also necessary to control for the potential confounding effects emerging through differing effective population sizes between species[157,138], which is essential for investigating traits like brain size that correlate with body size[253]. To get the full sequence of *TRNP1*, we resequenced *TRNP1* of many primates. I chose control proteins with similar turn-over rates and length, coming from genomic regions of good sequence quality in all included species of the phylogeny. In all comparisons, I also included body size as a control trait. Moreover, to establish functional evolutionary relevance, I quantified TRNP1 cellular activity of six different orthologues *in vivo* at the relevant developmental stage.

To investigate the evolution of regulatory activity, we need to identify orthologous CREs and develop unbiased assays. We assayed the activity of orthologous CREs in cellular trans-environments of humans and cynomolgus macaques in a cell type that is close to the relevant *in vivo* cell type - neural progenitor cells (NPCs). The most recent assay for regulatory activity at the time was MPRA[204], which is limited in length, therefore requiring tiling of the sequences. The back-calculated total CRE activity is only an approximation of the exact activity of the whole element. Another limitation of our approach is imposed by the lack of full regulatory landscape from each species, thereby likely leading to missing other *TRNP1* CREs active in species for which we did not have accessibility or histone modification data. In this context, a helpful additional confirmation of the observed activity patterns in our MPRA data is the fact that cellular *TRNP1* expression levels differ between NPCs of humans and macaques and across the brain organoids of humans and two other primates in the expected direction[254].

Combining these different approaches and lines of evidence, I found that the evolution of TRNP1 coding sequences correlate with brain size and folding across mammalian phylogeny strongly above the average protein association (top 5%), also reflected in a correlated change of its cellular activity. Because *TRNP1* is expressed in various proliferating cell types, it is important that the association with body size is considerably lower than with brain, indicating certain specificity in brain-related evolutionary change. In addition, I identified one CRE located in the intron of *TRNP1* with correlated regulatory activity with brain folding across primates. I also pinpoint candidate TFs, the binding of which might generate the observed regulatory divergence. These findings strengthen the proposed evolutionary role of TRNP1 for cortical evolution across mammals (Figure 3.1).

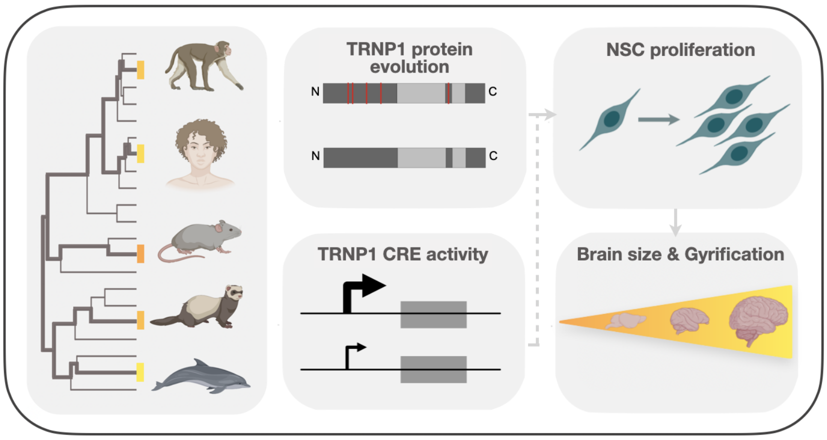


Figure 3.1. Summary of the findings regarding *TRNP1* protein and regulatory sequence co-evolution with brain size and folding across mammals.

There has been a long existing discussion on what type of genetic change is more central for phenotypic evolution: Gene or regulatory evolution, where regulation is proposed to have a larger total contribution due to less-constrained adaptation to different cellular and temporal contexts[162]. Among the reasons for lower flexibility in protein-coding evolution is the constrained functional 3D protein structure that can affect their folding, activity and interactions with other proteins or DNA[255,256,257,258]. While the total absolute contributions of protein and regulatory evolution are still difficult to quantify and therefore up to debate, it is clear that these are not necessarily mutually exclusive scenarios. Instead, if a certain gene facilitates a certain phenotype, it is possible that the resulting protein's activity as well as the amount of the protein product are tuned, thereby presenting two sides of the same coin. The study on *TRNP1* represents such a case, where both the protein and the regulatory activity facilitate the same phenotype. According to our findings, change in the protein is correlated with brain phenotypes across a deeper phylogeny, i.e., larger time scales, whereas the consistent co-variation for regulation was detected to be considerably stronger within apes and Old world monkeys, i.e., shorter time scales. This is in agreement with the expectations that regulatory turnover is much higher due to less unconstrained genomic positions. The observed evolutionary flexibility of TRNP1 could be partially explained by the high amount of intrinsically disordered regions (IDRs) in its sequence[118]. IDR-rich proteins have been shown to lack a fixed or ordered 3D structure, thereby allowing for more relaxed evolutionary modes and enabling positive selection to, e.g., facilitate interaction with other proteins of the particular cellular environment[259,260,261,262]. Further improvement in protein-coding sequence alignments, boosted by better genome and annotation qualities, will allow comparisons of the less conserved parts of proteins. Combined with accumulation of quantitative phenotype annotations, this could accelerate the identification of other genotypic and phenotypic cross-species associations.

4 | Outlook

The progress that we have made in the last decades in clarifying and mapping the functionality of our genome is already immense. Still, there are several limitations that I encountered in my attempts to quantify regulatory and gene evolution rates across primates and other mammals and in connecting these to phenotypes, which require further improvements in the future.

To have more high-throughput screens of protein sequence co-evolution with traits, more sophisticated codon-aware alignment algorithms will be necessary. Algorithms like PRANK [252] are readily aware of the phylogeny when finding the optimal alignment of the protein-coding sequences. However, this is after manually removing the intronic sequences by the user. Further extension of the phylogenetic alignment frameworks boosted by expectation maximization algorithms, Bayesian frameworks or machine learning could add the recognition of potential exon-intron boundaries, where the known protein sequence and exon-intron structure from some species and the known typical bases at these boundaries could be used to automatically extract the orthologous protein-coding sequences from the orthologous DNA. To my knowledge, the current implementation of evolutionary models that accommodate different modes of evolution, such as Brownian Motion or Ornstein Uhlenbeck process [263,264], do not allow for multiple replicates per species. These frameworks could be extended to replicated experiments and measurements in the future. As an example, for a more quantitative detection of evolutionary modes of expression, the current developments in high-throughput single-cell RNA sequencing are offering exciting possibilities to compare orthologous cell type expression profiles across orthologous developmental stages between species. Evolutionary longitudinal models that account for the non-independence that arises through the species and subject source of the cell will be essential to unbiasedly identify diverged and conserved gene expression by appropriately utilizing the statistical power offered by repeated measurements, while accounting for the phylogenetic relationships.

Finally, there is a lot of exciting work in the future to better understand CRE evolution. Since regulatory sequence turnover does not appear to be a good predictor for the activity, focusing on assays that measure TF binding or CRE activity turn-over might be more informative. ATAC-STARR[41,265,266,267] proposes a more inclusive approach than a classic MPRA to assay the regulatory activity of individual CREs from certain species and cellular contexts. ATAC-STARR combines and alleviates two steps that have been part of a common workflow during my thesis: 1) Analyse ATAC-seq data from a species and cell-type of interest, 2) Select certain CRE sequences that will be assayed for their activity. ATAC-STARR uses enrichment technique by transposase as in ATAC-seq, followed by cloning of the sequences into plasmids containing a reporter gene to assay their activity. Hence, automatically, more accessible peaks will be more present in the assay, requiring no prior selection of individual CREs. If a sufficient number of cells is available from each species (and a sufficiently large budget for sequencing), this might be the most unbiased CRE activity approach to date, as long as orthologous cell types are available. Such data allows to characterize the genomic location, sequence, TFBS repertoire and activity of whole regulatory landscapes. This combined information could be further used to identify functionally-orthologous CREs, that are not necessary sequence orthologues, using a combination of advanced clustering and, potentially, deep learning approaches.

Bibliography

1. Luke Rendell, Robert Boyd, Daniel Cownden, Marquist Enquist, Kimmo Eriksson, Marc W Feldman, Laurel Fogarty, Stefano Ghirlanda, Timothy Lillicrap, and Kevin N Laland. Why copy others? insights from the social learning strategies tournament. *Science*, 328(5975):208–213, 2010.

2. Lewis G Dean, Rachel L Kendal, Steven J Schapiro, Bernard Thierry, and Kevin N Laland. Identification of the social and cognitive processes underlying human cumulative culture. *Science*, 335(6072):1114–1118, 2012.

3. Firas Gerges, Germain Zouein, and Danielle Azar. Genetic algorithms with local optima handling to solve sudoku puzzles. In *Proceedings of the 2018 international conference on computing and artificial intelligence*, pages 19–22, 2018.

4. Carlos Echegoyen, Alexander Mendiburu, Roberto Santana, and Jose A Lozano. On the taxonomy of optimization problems under estimation of distribution algorithms. *Evolutionary computation*, 21(3):471–495, 2013.

5. Nancy Forbes. *Imitation of life: how biology is inspiring computing*. Mit Press, 2004.

6. Wolfgang Enard. The molecular basis of human brain evolution. *Curr. Biol.*, 26(20):R1109–R1117, October 2016.

7. Maxim VC Greenberg and Deborah Bourc'his. The diverse roles of dna methylation in mammalian development and disease. *Nature reviews Molecular cell biology*, 20(10):590–607, 2019.

8. Andrew J Bannister and Tony Kouzarides. Regulation of chromatin by histone modifications. *Cell research*, 21(3):381–395, 2011.

9. Luciano Di Croce and Kristian Helin. Transcriptional regulation by polycomb group proteins. *Nature structural & molecular biology*, 20(10):1147–1155, 2013.

10. Tong Ihn Lee and Richard A Young. Transcription of eukaryotic protein-coding genes. *Annual review of genetics*, 34(1):77–137, 2000.

11. Robert G Roeder. 50+ years of eukaryotic transcription: an expanding universe of factors and mechanisms. *Nat. Struct. Mol. Biol.*, 26(9):783–791, September 2019.

12. Sumantra Chatterjee and Nadav Ahituv. Gene regulatory elements, major drivers of human disease. *Annu. Rev. Genomics Hum. Genet.*, 18:45–63, August 2017.

13. Samuel A Lambert, Arttu Jolma, Laura F Campitelli, Pratyush K Das, Yimeng Yin, Mihai Albu, Xiaoting Chen, Jussi Taipale, Timothy R Hughes, and Matthew T Weirauch. The human transcription factors. *Cell*, 172(4):650–665, 2018.

14. Fatih Ozsolak and Patrice M Milos. Rna sequencing: advances, challenges and opportunities. *Nature reviews genetics*, 12(2):87–98, 2011.

15. Rory Stark, Marta Grzelak, and James Hadfield. Rna sequencing: the teenage years. *Nature Reviews Genetics*, 20(11):631–656, 2019.

16. Juan M Vaquerizas, Sarah K Kummerfeld, Sarah A Teichmann, and Nicholas M Luscombe. A census of human transcription factors: function, expression and evolution. *Nature Reviews Genetics*, 10(4):252–263, 2009.

17. Martin Fischer, Lydia Steiner, and Kurt Engeland. The transcription factor p53: not a repressor, solely an activator. *Cell cycle*, 13(19):3037–3058, 2014.

18. Seth Frietze and Peggy J Farnham. Transcription factor effector domains. *A handbook of transcription factors*, pages 261–277, 2011.

19. Michael G Rosenfeld, Victoria V Lunyak, and Christopher K Glass. Sensors and signals: a coactivator/-corepressor/epigenetic code for integrating signal-dependent programs of transcriptional response. *Genes & development*, 20(11):1405–1428, 2006.

20. Jun Ma. Crossing the line between activation and repression. *Trends in genetics*, 21(1):54–59, 2005.

21. Arttu Jolma, Jian Yan, Thomas Whitington, Jarkko Toivonen, Kazuhiro R Nitta, Pasi Rastas, Ekaterina Morgunova, Martin Enge, Mikko Taipale, Gonghong Wei, et al. Dna-binding specificities of human transcription factors. *Cell*, 152(1):327–339, 2013.

22. Arttu Jolma, Teemu Kivioja, Jarkko Toivonen, Lu Cheng, Gonghong Wei, Martin Enge, Mikko Taipale, Juan M Vaquerizas, Jian Yan, Mikko J Sillanpää, et al. Multiplexed massively parallel selex for characterization of human transcription factor binding specificities. *Genome research*, 20(6):861–873, 2010.

23. Jaime A Castro-Mondragon, Rafael Riudavets-Puig, Ieva Rauluseviciute, Roza Berhanu Lemma, Laura Turchi, Romain Blanc-Mathieu, Jeremy Lucas, Paul Boddie, Aziz Khan, Nicolás Manosalva Pérez, et al. Jaspar 2022: the 9th release of the open-access database of transcription factor binding profiles. *Nucleic acids research*, 50(D1):D165–D173, 2022.

24. Justin Crocker, Ella Preger-Ben Noon, and David L Stern. The soft touch: Low-affinity transcription factor binding sites in development and evolution. *Curr. Top. Dev. Biol.*, 117:455–469, January 2016.

25. Connor A Horton, Amr M Alexandari, Michael G B Hayes, Emil Marklund, Julia M Schaepe, Arjun K Aditham, Nilay Shah, Peter H Suzuki, Avanti Shrikumar, Ariel Afek, William J Greenleaf, Raluca Gordân, Julia Zeitlinger, Anshul Kundaje, and Polly M Fordyce. Short tandem repeats bind transcription factors to tune eukaryotic gene expression. *Science*, 381(6664):eadd1250, September 2023.

26. N M Luscombe, S E Austin, H M Berman, and J M Thornton. An overview of the structures of protein-DNA complexes. *Genome Biol.*, 1(1):REVIEWS001, June 2000.

27. Julie Carnesecchi, Pedro B Pinto, and Ingrid Lohmann. Hox transcription factors: an overview of multi-step regulators of gene expression. *Int. J. Dev. Biol.*, 62(11-12):723–732, 2018.

28. Jonathan N Wells, Ni-Chen Chang, John McCormick, Caitlyn Coleman, Nathalie Ramos, Bozhou Jin, and Cédric Feschotte. Transposable elements drive the evolution of metazoan zinc finger genes. *Genome Res.*, 33 (8):1325–1339, August 2023.

29. Ivana L de la Serna, Yasuyuki Ohkawa, Charlotte A Berkes, Donald A Bergstrom, Caroline S Dacwag, Stephen J Tapscott, and Anthony N Imbalzano. MyoD targets chromatin remodeling complexes to the myogenin locus prior to forming a stable DNA-bound complex. *Mol. Cell. Biol.*, 25(10):3997–4009, May 2005.

30. Shane McManus, Anja Ebert, Giorgia Salvagiotto, Jasna Medvedovic, Qiong Sun, Ido Tamir, Markus Jaritz, Hiromi Tagoh, and Meinrad Busslinger. The transcription factor pax5 regulates its target genes by recruiting chromatin-modifying proteins in committed B cells. *EMBO J.*, 30(12):2388–2404, May 2011.

31. Mathieu Lupien, Jérôme Eeckhoute, Clifford A Meyer, Qianben Wang, Yong Zhang, Wei Li, Jason S Carroll, X Shirley Liu, and Myles Brown. FoxA1 translates epigenetic signatures into enhancer-driven lineage-specific transcription. *Cell*, 132(6):958–970, March 2008.

32. Yongbing Zhao, Supriya V Vartak, Andrea Conte, Xiang Wang, David A Garcia, Evan Stevens, Seol Kyoung Jung, Kyong-Rim Kieffer-Kwon, Laura Vian, Timothy Stodola, et al. "stripe" transcription factors provide accessibility to co-binding partners in mammalian genomes. *Molecular cell*, 82(18):3398–3411, 2022.

33. Kazutoshi Takahashi, Michiko Nakamura, Chikako Okubo, Zane Kliesmete, Mari Ohnuki, Megumi Narita, Akira Watanabe, Mai Ueda, Yasuhiro Takashima, Ines Hellmann, et al. The pluripotent stem cell-specific transcript esrg is dispensable for human pluripotency. *PLoS genetics*, 17(5):e1009587, 2021.

34. Adam Woolfe, Martin Goodson, Debbie K Goode, Phil Snell, Gayle K McEwen, Tanya Vavouri, Sarah F Smith, Phil North, Heather Callaway, Krys Kelly, Klaudia Walter, Irina Abnizova, Walter Gilks, Yvonne J K Edwards, Julie E Cooke, and Greg Elgar. Highly conserved non-coding sequences are associated with vertebrate development. *PLoS Biol.*, 3(1):e7, January 2005.

35. Robert E Thurman, Eric Rynes, Richard Humbert, Jeff Vierstra, Matthew T Maurano, Eric Haugen, Nathan C Sheffield, Andrew B Stergachis, Hao Wang, Benjamin Vernot, Kavita Garg, Sam John, Richard Sandstrom, Daniel Bates, Lisa Boatman, Theresa K Canfield, Morgan Diegel, Douglas Dunn, Abigail K Ebersol, Tristan Frum, Erika Giste, Audra K Johnson, Ericka M Johnson, Tanya Kutyavin, Bryan Lajoie, Bum-Kyu Lee, Kristen Lee, Darin London, Dimitra Lotakis, Shane Neph, Fidencio Neri, Eric D Nguyen, Hongzhu Qu, Alex P Reynolds, Vaughn Roach, Alexias Safi, Minerva E Sanchez, Amartya Sanyal, Anthony Shafer, Jeremy M Simon, Lingyun Song, Shinny Vong, Molly Weaver, Yongqi Yan, Zhancheng Zhang, Zhuzhu Zhang, Boris Lenhard, Muneesh Tewari, Michael O Dorschner, R Scott Hansen, Patrick A Navas, George Stamatoyannopoulos, Vishwanath R Iyer, Jason D Lieb, Shamil R Sunyaev, Joshua M Akey, Peter J Sabo, Rajinder Kaul, Terrence S Furey, Job Dekker, Gregory E Crawford, and John A Stamatoyannopoulos. The accessible chromatin landscape of the human genome. *Nature*, 489(7414):75–82, September 2012.

36. Scott W Doniger and Justin C Fay. Frequent gain and loss of functional transcription factor binding sites. *PLoS computational biology*, 3(5):e99, 2007.

37. Diego Villar, Camille Berthelot, Sarah Aldridge, Tim F Rayner, Margus Lukk, Miguel Pignatelli, Thomas J Park, Robert Deaville, Jonathan T Erichsen, Anna J Jasinska, James M A Turner, Mads F Bertelsen, Elizabeth P Murchison, Paul Flicek, and Duncan T Odom. Enhancer evolution across 20 mammalian species. *Cell*, 160(3):554–566, January 2015.

38. Lingyun Song and Gregory E Crawford. Dnase-seq: a high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells. *Cold Spring Harbor Protocols*, 2010(2): pdb–prot5384, 2010.

39. Jason D Buenrostro, Beijing Wu, Howard Y Chang, and William J Greenleaf. Atac-seq: a method for assaying chromatin accessibility genome-wide. *Current protocols in molecular biology*, 109(1):21–29, 2015.

40. Fiorella C Grandi, Hailey Modi, Lucas Kampman, and M Ryan Corces. Chromatin accessibility profiling by ATAC-seq. *Nat. Protoc.*, 17(6):1518–1552, June 2022.

41. Jason Ernst, Alexandre Melnikov, Xiaolan Zhang, Li Wang, Peter Rogov, Tarjei S Mikkelsen, and Manolis Kellis. Genome-scale high-resolution mapping of activating and repressive nucleotides in regulatory regions. *Nature biotechnology*, 34(11):1180–1190, 2016.

42. David S Johnson, Ali Mortazavi, Richard M Myers, and Barbara Wold. Genome-wide mapping of in vivo protein-dna interactions. *Science*, 316(5830):1497–1502, 2007.

43. Artem Barski, Suresh Cuddapah, Kairong Cui, Tae-Young Roh, Dustin E Schones, Zhibin Wang, Gang Wei, Iouri Chepelev, and Keji Zhao. High-resolution profiling of histone methylations in the human genome. *Cell*, 129(4):823–837, 2007.

44. Alan P Boyle, Sean Davis, Hennady P Shulha, Paul Meltzer, Elliott H Margulies, Zhiping Weng, Terrence S Furey, and Gregory E Crawford. High-resolution mapping and characterization of open chromatin across the genome. *Cell*, 132(2):311–322, January 2008.

45. Piero Carninci, Albin Sandelin, Boris Lenhard, Shintaro Katayama, Kazuro Shimokawa, Jasmina Ponjavic, Colin A M Semple, Martin S Taylor, Pär G Engström, Martin C Frith, Alistair R R Forrest, Wynand B Alkema, Sin Lam Tan, Charles Plessy, Rimantas Kodzius, Timothy Ravasi, Takeya Kasukawa, Shiro Fukuda, Mutsumi Kanamori-Katayama, Yayoi Kitazume, Hideya Kawaji, Chikatoshi Kai, Mari Nakamura, Hideaki Konno, Kenji Nakano, Salim Mottagui-Tabar, Peter Arner, Alessandra Chesi, Stefano Gustincich, Francesca Persichetti, Harukazu Suzuki, Sean M Grimmond, Christine A Wells, Valerio Orlando, Claes Wahlestedt, Edison T Liu, Matthias Harbers, Jun Kawai, Vladimir B Bajic, David A Hume, and Yoshihide Hayashizaki. Genome-wide analysis of mammalian promoter architecture and evolution. *Nat. Genet.*, 38(6):626–635, June 2006.

46. Michael Bulger and Mark Groudine. Enhancers: the abundance and function of regulatory sequences beyond promoters. *Dev. Biol.*, 339(2):250–257, March 2010.

47. Elinore M Mercer, Yin C Lin, Christopher Benner, Suchit Jhunjhunwala, Janusz Dutkowski, Martha Flores, Mikael Sigvardsson, Trey Ideker, Christopher K Glass, and Cornelis Murre. Multilineage priming of enhancer repertoires precedes commitment to the B and myeloid cell lineages in hematopoietic progenitors. *Immunity*, 35(3):413–425, September 2011.

48. Molly Gasperini, Jacob M Tome, and Jay Shendure. Towards a comprehensive catalogue of validated and target-linked human enhancers. *Nat. Rev. Genet.*, 21(5):292–310, May 2020.

49. Di Huang, Hanna M Petrykowska, Brendan F Miller, Laura Elnitski, and Ivan Ovcharenko. Identification of human silencers by correlating cross-tissue epigenetic profiles and gene expression. *Genome Res.*, 29(4): 657–667, April 2019.

50. Ying Zhang, Yi Xiang See, Vinay Tergaonkar, and Melissa Jane Fullwood. Long-Distance repression by human silencers: Chromatin interactions and phase separation in silencers. *Cells*, 11(9), May 2022.

51. Robin Andersson and Albin Sandelin. Determinants of enhancer and promoter activities of regulatory elements. *Nat. Rev. Genet.*, 21(2):71–87, February 2020.

52. Maša Roller, Ericca Stamper, Diego Villar, Osagie Izuogu, Fergal Martin, Aisling M Redmond, Raghavendra Ramachanderan, Louise Harewood, Duncan T Odom, and Paul Flicek. LINE retrotransposons characterize mammalian tissue-specific and evolutionarily dynamic regulatory regions. *Genome Biol.*, 22(1):62, February 2021.

53. Marco Osterwalder, Iros Barozzi, Virginie Tissières, Yoko Fukuda-Yuzawa, Brandon J Mannion, Sarah Y Afzal, Elizabeth A Lee, Yiwen Zhu, Ingrid Plajzer-Frick, Catherine S Pickle, Momoe Kato, Tyler H Garvin, Quan T Pham, Anne N Harrington, Jennifer A Akiyama, Veena Afzal, Javier Lopez-Rios, Diane E Dickel, Axel Visel, and Len A Pennacchio. Enhancer redundancy provides phenotypic robustness in mammalian development. *Nature*, 554(7691):239–243, February 2018.

54. Joung-Woo Hong, David A Hendrix, and Michael S Levine. Shadow enhancers as a source of evolutionary novelty. *Science*, 321(5894):1314, September 2008.

55. Scott Barolo. Shadow enhancers: frequently asked questions about distributed cis-regulatory information and enhancer redundancy. *Bioessays*, 34(2):135–141, February 2012.

56. Alexandre Laverré, Eric Tannier, and Anamaria Necsulea. Long-range promoter-enhancer contacts are conserved during evolution and contribute to gene expression robustness. *Genome Res.*, 32(2):280–296, February 2022.

57. Erez Lieberman-Aiden, Nynke L van Berkum, Louise Williams, Maxim Imakaev, Tobias Ragoczy, Agnes Telling, Ido Amit, Bryan R Lajoie, Peter J Sabo, Michael O Dorschner, Richard Sandstrom, Bradley Bernstein, M A Bender, Mark Groudine, Andreas Gnirke, John Stamatoyannopoulos, Leonid A Mirny, Eric S Lander, and Job Dekker. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, 326(5950):289–293, October 2009.

58. Jesse R Dixon, Siddarth Selvaraj, Feng Yue, Audrey Kim, Yan Li, Yin Shen, Ming Hu, Jun S Liu, and Bing Ren. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, 485(7398):376–380, April 2012.

59. Suhas S P Rao, Miriam H Huntley, Neva C Durand, Elena K Stamenova, Ivan D Bochkov, James T Robinson, Adrian L Sanborn, Ido Machol, Arina D Omer, Eric S Lander, and Erez Lieberman Aiden. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*, 159(7):1665–1680, December 2014.

60. Somi Kim, Nam-Kyung Yu, and Bong-Kiun Kaang. CTCF as a multifunctional protein in genome regulation and gene expression. *Exp. Mol. Med.*, 47(6):e166, June 2015.

61. C Y Mclean and M Bristor. GREAT improves functional interpretation cis-regulatory regions". *Nat. Biotechnol*, 926(5):495–501, 2010.

62. Simon Fishilevich, Ron Nudel, Noa Rappaport, Rotem Hadar, Inbar Plaschkes, Tsippi Iny Stein, Naomi Rosen, Asher Kohn, Michal Twik, Marilyn Safran, Doron Lancet, and Dana Cohen. GeneHancer: genome-wide integration of enhancers and target genes in GeneCards. *Database (Oxford)*, 2017, January 2017.

63. Hannah A Pliner, Jonathan S Packer, José L McFaline-Figueroa, Darren A Cusanovich, Riza M Daza, Delasa Aghamirzaie, Sanjay Srivatsan, Xiaojie Qiu, Dana Jackson, Anna Minkina, Andrew C Adey, Frank J Steemers, Jay Shendure, and Cole Trapnell. Cicero predicts cis-regulatory DNA interactions from single-cell chromatin accessibility data. *Mol. Cell*, 71(5):858–871.e8, September 2018.

64. Sai Ma, Bing Zhang, Lindsay M LaFave, Andrew S Earl, Zachary Chiang, Yan Hu, Jiarui Ding, Alison Brack, Vinay K Kartha, Tristan Tay, Travis Law, Caleb Lareau, Ya-Chieh Hsu, Aviv Regev, and Jason D Buenrostro. Chromatin potential identified by shared single-cell profiling of RNA and chromatin. *Cell*, 183 (4):1103–1116.e20, November 2020.

65. Tim Stuart, Avi Srivastava, Shaista Madad, Caleb A Lareau, and Rahul Satija. Single-cell chromatin state analysis with signac. *Nat. Methods*, 18(11):1333–1341, November 2021.

66. Sean Whalen, Rebecca M Truty, and Katherine S Pollard. Enhancer-promoter interactions are encoded by complex genomic signatures on looping chromatin. *Nat. Genet.*, 48(5):488–496, May 2016.

67. Dina Hafez, Aslihan Karabacak, Sabrina Krueger, Yih-Chii Hwang, Li-San Wang, Robert P Zinzen, and Uwe Ohler. McEnhancer: predicting gene expression via semi-supervised assignment of enhancers to target genes. *Genome Biol.*, 18(1):199, October 2017.

68. Kenneth EM Hastings. Strong evolutionary conservation of broadly expressed protein isoforms in the troponin i gene family and other vertebrate gene families. *Journal of Molecular Evolution*, 42:631–640, 1996.

69. Laurent Duret and Dominique Mouchiroud. Determinants of substitution rates in mammalian genes: expression pattern affects selection intensity but not mutation rate. *Molecular biology and evolution*, 17(1): 68–070, 2000.

70. Margarida Cardoso-Moreira, Jean Halbert, Delphine Valloton, Britta Velten, Chunyan Chen, Yi Shao, Angélica Liechti, Kelly Ascenção, Coralie Rummel, Svetlana Ovchinnikova, et al. Gene expression across mammalian organ development. *Nature*, 571(7766):505–509, 2019.

71. Philipp Khaitovich, Ines Hellmann, Wolfgang Enard, Katja Nowick, Marcus Leinweber, Henriette Franz, Gunter Weiss, Michael Lachmann, and Svante Pääbo. Parallel patterns of evolution in the genomes and transcriptomes of humans and chimpanzees. *Science*, 309(5742):1850–1854, September 2005.

72. David Brawand, Magali Soumillon, Anamaria Necsulea, Philippe Julien, Gábor Csárdi, Patrick Harrigan, Manuela Weier, Angélica Liechti, Ayinuer Aximu-Petri, Martin Kircher, Frank W Albert, Ulrich Zeller, Philipp Khaitovich, Frank Grützner, Sven Bergmann, Rasmus Nielsen, Svante Pääbo, and Henrik Kaessmann. The evolution of gene expression levels in mammalian organs. *Nature*, 478(7369):343–348, October 2011.

73. Zhong-Yi Wang, Evgeny Leushkin, Angélica Liechti, Svetlana Ovchinnikova, Katharina Mößinger, Thoomke Brüning, Coralie Rummel, Frank Grützner, Margarida Cardoso-Moreira, Peggy Janich, David Gatfield, Boubou Diagouraga, Bernard de Massy, Mark E Gill, Antoine H F M Peters, Simon Anders, and Henrik Kaessmann. Transcriptome and translatome co-evolution in mammals. *Nature*, 588(7839):642–647, December 2020.

74. Kyoko Watanabe, Sven Stringer, Oleksandr Frei, Maša Umićević Mirkov, Christiaan de Leeuw, Tinca J C Polderman, Sophie van der Sluis, Ole A Andreassen, Benjamin M Neale, and Danielle Posthuma. A global overview of pleiotropy and genetic architecture in complex traits. *Nat. Genet.*, 51(9):1339–1348, September 2019.

75. Haig H Kazazian, Jr. Mobile elements: drivers of genome evolution. *Science*, 303(5664):1626–1632, March 2004.

76. Anna D Senft and Todd S Macfarlan. Transposable elements shape the evolution of mammalian development. *Nature Reviews Genetics*, 22(11):691–711, 2021.

77. E S Lander, L M Linton, B Birren, C Nusbaum, M C Zody, J Baldwin, K Devon, K Dewar, M Doyle, W FitzHugh, R Funke, D Gage, K Harris, A Heaford, J Howland, L Kann, J Lehoczky, R LeVine, P McEwan, K McKernan, J Meldrim, J P Mesirov, C Miranda, W Morris, J Naylor, C Raymond, M Rosetti, R Santos, A Sheridan, C Sougnez, Y Stange-Thomann, N Stojanovic, A Subramanian, D Wyman, J Rogers, J Sulston, R Ainscough, S Beck, D Bentley, J Burton, C Clee, N Carter, A Coulson, R Deadman, P Deloukas, A Dunham, I Dunham, R Durbin, L French, D Grafham, S Gregory, T Hubbard, S Humphray, A Hunt, M Jones, C Lloyd, A McMurray, L Matthews, S Mercer, S Milne, J C Mullikin, A Mungall, R Plumb, M Ross, R Shownkeen, S Sims, R H Waterston, R K Wilson, L W Hillier, J D McPherson, M A Marra, E R Mardis, L A Fulton, A T Chinwalla, K H Pepin, W R Gish, S L Chissoe, M C Wendl, K D Delehaunty, T L Miner, A Delehaunty, J B Kramer, L L Cook, R S Fulton, D L Johnson, P J Minx, S W Clifton, T Hawkins, E Branscomb, P Predki, P Richardson, S Wenning, T Slezak, N Doggett, J F Cheng, A Olsen, S Lucas, C Elkin, E Uberbacher, M Frazier, R A Gibbs, D M Muzny, S E Scherer, J B Bouck, E J Sodergren, K C Worley, C M Rives, J H Gorrell, M L Metzker, S L Naylor, R S Kucherlapati, D L Nelson, G M Weinstock, Y Sakaki, A Fujiyama, M Hattori, T Yada, A Toyoda, T Itoh, C Kawagoe, H Watanabe, Y Totoki, T Taylor, J Weissenbach, R Heilig, W Saurin, F Artiguenave, P Brottier, T Bruls, E Pelletier, C Robert, P Wincker, D R Smith, L Doucette-Stamm, M Rubenfield, K Weinstock, H M Lee, J Dubois, A Rosenthal, M Platzer, G Nyakatura, S Taudien, A Rump, H Yang, J Yu, J Wang, G Huang, J Gu, L Hood, L Rowen, A Madan, S Qin, R W Davis, N A Federspiel, A P Abola, M J Proctor, R M Myers, J Schmutz, M Dickson, J Grimwood, D R Cox, M V Olson, R Kaul, C Raymond, N Shimizu, K Kawasaki, S Minoshima, G A Evans, M Athanasiou, R Schultz, B A Roe, F Chen, H Pan, J Ramser, H Lehrach, R Reinhardt, W R McCombie, M de la Bastide, N Dedhia, H Blöcker, K Hornischer, G Nordsiek, R Agarwala, L Aravind, J A Bailey, A Bateman, S Batzoglou, E Birney, P Bork, D G Brown, C B Burge, L Cerutti, H C Chen, D Church, M Clamp, R R Copley, T Doerks, S R Eddy, E E Eichler, T S Furey, J Galagan, J G Gilbert, C Harmon, Y Hayashizaki, D Haussler, H Hermjakob, K Hokamp, W Jang, L S Johnson, T A Jones, S Kasif, A Kaspryzk, S Kennedy, W J Kent, P Kitts, E V Koonin, I Korf, D Kulp, D Lancet, T M Lowe, A McLysaght, T Mikkelsen, J V Moran, N Mulder, V J Pollara, C P Ponting, G Schuler, J Schultz, G Slater, A F Smit, E Stupka, J Szustakowki, D Thierry-Mieg, J Thierry-Mieg, L Wagner, J Wallis, R Wheeler, A Williams, Y I Wolf, K H Wolfe, S P Yang, R F Yeh, F Collins, M S Guyer, J Peterson, A Felsenfeld, K A Wetterstrand, A Patrinos, M J Morgan, P de Jong, J J Catanese, K Osoegawa, H Shizuya, S Choi, Y J Chen, J Szustakowki, and International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, February 2001.

78. Sophie Lanciano and Gael Cristofari. Measuring and interpreting transposable element expression. *Nat. Rev. Genet.*, 21(12):721–736, December 2020.

79. Michael Lynch and Bruce Walsh. *The origins of genome architecture*, volume 98. Sinauer associates Sunderland, MA, 2007.

80. Olympia Gianfrancesco, Bethany Geary, Abigail L Savage, Kimberley J Billingsley, Vivien J Bubb, and John P Quinn. The role of SINE-VNTR-alu (SVA) retrotransposons in shaping the human genome. *Int. J. Mol. Sci.*, 20(23):5977, November 2019.

81. RepeatMasker home page. `http://www.repeatmasker.org/`. Accessed: 2024-6-4.

82. Guillaume Bourque, Kathleen H Burns, Mary Gehring, Vera Gorbunova, Andrei Seluanov, Molly Hammell, Michaël Imbeault, Zsuzsanna Izsvák, Henry L Levin, Todd S Macfarlan, et al. Ten things you should know about transposable elements. *Genome biology*, 19:1–12, 2018.

83. Rita Rebollo, Mark T Romanish, and Dixie L Mager. Transposable elements: an abundant and natural source of regulatory sequences for host genes. *Annu. Rev. Genet.*, 46:21–42, August 2012.

84. Gregory Andrews, Kaili Fan, Henry E Pratt, Nishigandha Phalke, Zoonomia Consortium§, Elinor K Karlsson, Kerstin Lindblad-Toh, Steven Gazal, Jill E Moore, and Zhiping Weng. Mammalian evolution of human cis-regulatory elements and transcription factor binding sites. *Science*, 380(6643):eabn7930, April 2023.

85. M Naville, I A Warren, Z Haftek-Terreau, D Chalopin, F Brunet, P Levin, D Galiana, and J-N Volff. Not so bad after all: retroviruses and long terminal repeat retrotransposons as a source of new genes in vertebrates. *Clin. Microbiol. Infect.*, 22(4):312–323, April 2016.

86. Raquel Fueyo, Julius Judd, Cedric Feschotte, and Joanna Wysocka. Roles of transposable elements in the regulation of mammalian transcription. *Nat. Rev. Mol. Cell Biol.*, 23(7):481–497, July 2022.

87. Y Amy Tang, Derek Huntley, Giovanni Montana, Andrea Cerase, Tatyana B Nesterova, and Neil Brockdorff. Efficiency of xist-mediated silencing on autosomes is linked to chromosomal domain organisation. *Epigenetics & chromatin*, 3:1–12, 2010.

88. Carolyn J Brown, Brian D Hendrich, Jim L Rupert, Ronald G Lafreniere, Yigong Xing, Jeanne Lawrence, and Huntington F Willard. The human xist gene: analysis of a 17 kb inactive x-specific rna that contains conserved repeats and is highly localized within the nucleus. *Cell*, 71(3):527–542, 1992.

89. Neil Brockdorff, Alan Ashworth, Graham F Kay, Veronica M McCabe, Dominic P Norris, Penny J Cooper, Sally Swift, and Sohaila Rastan. The product of the mouse xist gene is a 15 kb inactive x-specific transcript containing no conserved orf and located in the nucleus. *Cell*, 71(3):515–526, 1992.

90. Graeme D Penny, Graham F Kay, Steven A Sheardown, Sohaila Rastan, and Neil Brockdorff. Requirement for xist in x chromosome inactivation. *Nature*, 379(6561):131–137, 1996.

91. Jean-Louis Frendo, Delphine Olivier, Valérie Cheynet, Jean-Luc Blond, Olivier Bouton, Michel Vidaud, Michèle Rabreau, Danièle Evain-Brion, and François Mallet. Direct involvement of herv-w env glycoprotein in human trophoblast cell fusion and differentiation. *Molecular and cellular biology*, 23(10):3566–3574, 2003.

92. François Mallet, Olivier Bouton, Sarah Prudhomme, Valérie Cheynet, Guy Oriol, Bertrand Bonnaud, Gérard Lucotte, Laurent Duret, and Bernard Mandrand. The endogenous retroviral locus ervwe1 is a bona fide gene involved in hominoid placental physiology. *Proceedings of the National Academy of Sciences*, 101(6): 1731–1736, 2004.

93. Abdulrahman Mohammed Alhowikan. Activity-regulated cytoskeleton-associated protein dysfunction may contribute to memory disorder and earlier detection of autism spectrum disorders. *Medical Principles and Practice*, 25(4):350–354, 2016.

94. Elissa D Pastuzyn, Cameron E Day, Rachel B Kearns, Madeleine Kyrke-Smith, Andrew V Taibi, John McCormick, Nathan Yoder, David M Belnap, Simon Erlendsson, Dustin R Morado, et al. The neuronal gene arc encodes a repurposed retrotransposon gag protein that mediates intercellular rna transfer. *Cell*, 172(1): 275–288, 2018.

95. Norbert Bannert and Reinhard Kurth. The evolutionary dynamics of human endogenous retroviral families. *Annu. Rev. Genomics Hum. Genet.*, 7(1):149–173, 2006.

96. Jumpei Ito, Ryota Sugimoto, Hirofumi Nakaoka, Shiro Yamada, Tetsuaki Kimura, Takahide Hayano, and Ituro Inoue. Systematic identification and characterization of regulatory elements derived from human endogenous retroviruses. *PLoS Genet.*, 13(7):e1006883, July 2017.

97. Thomas A Carter, Manvendra Singh, Gabrijela Dumbović, Jason D Chobirko, John L Rinn, and Cédric Feschotte. Mosaic cis-regulatory evolution drives transcriptional partitioning of HERVH endogenous retrovirus in the human embryo. *Elife*, 11, February 2022.

98. Mari Ohnuki, Koji Tanabe, Kenta Sutou, Ito Teramoto, Yuka Sawamura, Megumi Narita, Michiko Nakamura, Yumie Tokunaga, Masahiro Nakamura, Akira Watanabe, Shinya Yamanaka, and Kazutoshi Takahashi. Dynamic regulation of human endogenous retroviruses mediates factor-induced reprogramming and differentiation potential. *Proc. Natl. Acad. Sci. U. S. A.*, 111(34):12426–12431, August 2014.

99. Pierre-Étienne Jacques, Justin Jeyakani, and Guillaume Bourque. The majority of primate-specific regulatory sequences are derived from transposable elements. *PLoS Genet.*, 9(5):e1003504, May 2013.

100. Julien Pontis, Evarist Planet, Sandra Offner, Priscilla Turelli, Julien Duc, Alexandre Coudray, Thorold W Theunissen, Rudolf Jaenisch, and Didier Trono. Hominoid-specific transposable elements and KZFPs facilitate human embryonic genome activation and control transcription in naive human ESCs. *Cell Stem Cell*, 24(5): 724–735.e5, May 2019.

101. Julien Pontis, Cyril Pulver, Christopher J Playfoot, Evarist Planet, Delphine Grun, Sandra Offner, Julien Duc, Andrea Manfrin, Matthias P Lutolf, and Didier Trono. Primate-specific transposable elements shape transcriptional networks during human development. *Nature Communications*, 13(1):7178, 2022.

102. Galih Kunarso, Na-Yu Chia, Justin Jeyakani, Catalina Hwang, Xinyi Lu, Yun-Shen Chan, Huck-Hui Ng, and Guillaume Bourque. Transposable elements have rewired the core regulatory network of human embryonic stem cells. *Nat. Genet.*, 42(7):631–634, July 2010.

103. Marco Trizzino, YoSon Park, Marcia Holsbach-Beltrame, Katherine Aracena, Katelyn Mika, Minal Caliskan, George H Perry, Vincent J Lynch, and Christopher D Brown. Transposable elements are the primary source of novelty in primate gene regulation. *Genome research*, 27(10):1623–1633, 2017.

104. Michelle C Ward, Siming Zhao, Kaixuan Luo, Bryan J Pavlovic, Mohammad M Karimi, Matthew Stephens, and Yoav Gilad. Silencing of transposable elements may not be a major driver of regulatory evolution in primate iPSCs. *Elife*, 7, April 2018.

105. Henry L Levin and John V Moran. Dynamic interactions between transposable elements and their hosts. *Nat. Rev. Genet.*, 12(9):615–627, August 2011.

106. A Caballero. Developments in the prediction of effective population size. *Heredity*, 73 (Pt 6):657–679, December 1994.

107. M C Whitlock and N H Barton. The effective size of a subdivided population. *Genetics*, 146(1):427–441, May 1997.

108. F Rousset. Effective size in simple metapopulation models. *Heredity*, 91(2):107–111, August 2003.

109. Philip Hedrick. Large variance in reproductive success and the Ne/N ratio. *Evolution*, 59(7):1596–1599, July 2005.

110. Richard Frankham. Effective population size/adult population size ratios in wildlife: a review. *Genet. Res.*, 89(5-6):491–503, December 2007.

111. W H Li. Molecular evolution. 1997.

112. Lucie A Bergeron, Søren Besenbacher, Jiao Zheng, Panyi Li, Mads Frost Bertelsen, Benoit Quintard, Joseph I Hoffman, Zhipeng Li, Judy St Leger, Changwei Shao, Josefin Stiller, M Thomas P Gilbert, Mikkel H Schierup, and Guojie Zhang. Evolution of the germline mutation rate across vertebrates. *Nature*, 615(7951):285–291, March 2023.

113. Ines Hellmann, Ingo Ebersberger, Susan E Ptak, Svante Pääbo, and Molly Przeworski. A neutral explanation for the correlation of diversity with recombination rates in humans. *Am. J. Hum. Genet.*, 72(6):1527–1535, June 2003.

114. Svitlana Tyekucheva, Kateryna D Makova, John E Karro, Ross C Hardison, Webb Miller, and Francesca Chiaromonte. Human-macaque comparisons illuminate variation in neutral substitution rates. *Genome Biol.*, 9(4):R76, April 2008.

115. Kateryna D Makova and Ross C Hardison. The effects of chromatin organization on variation in mutation rates in the genome. *Nat. Rev. Genet.*, 16(4):213–223, April 2015.

116. Tomoko Ohta. The nearly neutral theory of molecular evolution. *Annu. Rev. Ecol. Syst.*, 23(1):263–286, November 1992.

117. Justin C Fay and Chung-I Wu. Sequence divergence, functional constraint, and selection in protein evolution. *Annu. Rev. Genomics Hum. Genet.*, 4(1):213–235, 2003.

118. Eugene V Koonin and Yuri I Wolf. Constraints and plasticity in genome and molecular-phenome evolution. *Nature Reviews Genetics*, 11(7):487–498, 2010.

119. Valentina Snetkova, Len A Pennacchio, Axel Visel, and Diane E Dickel. Perfect and imperfect views of ultraconserved sequences. *Nat. Rev. Genet.*, 23(3):182–194, March 2022.

120. Murat Tuğrul, Tiago Paixao, Nicholas H Barton, and Gašper Tkačik. Dynamics of transcription factor binding site evolution. *PLoS genetics*, 11(11):e1005639, 2015.

121. Deena Emera, Jun Yin, Steven K Reilly, Jake Gockley, and James P Noonan. Origin and evolution of developmental enhancers in the mammalian neocortex. *Proc. Natl. Acad. Sci. U. S. A.*, 113(19):E2617–26, May 2016.

122. Jianzhi Zhang and Jian-Rong Yang. Determinants of the rate of protein sequence evolution. *Nat. Rev. Genet.*, 16(7):409–420, July 2015.

123. Naoki Osada and Hiroshi Akashi. Mitochondrial-nuclear interactions and accelerated compensatory evolution: evidence from the primate cytochrome C oxidase complex. *Mol. Biol. Evol.*, 29(1):337–346, January 2012.

124. Dan I Andersson and Diarmaid Hughes. Antibiotic resistance and its cost: is it possible to reverse resistance? *Nat. Rev. Microbiol.*, 8(4):260–271, April 2010.

125. J C Fay and C I Wu. Hitchhiking under positive darwinian selection. *Genetics*, 155(3):1405–1413, July 2000.

126. M Kimura. On the probability of fixation of mutant genes in a population. *Genetics*, 47(6):713–719, June 1962.

127. Reinhard Bürger and Warren J Ewens. Fixation probabilities of additive alleles in diploid populations. *J. Math. Biol.*, 33(5):557–575, April 1995.

128. Hiroshi Akashi, Naoki Osada, and Tomoko Ohta. Weak selection and protein evolution. *Genetics*, 192(1):15–31, September 2012.

129. T Ohta. Slightly deleterious mutant substitutions in evolution. *Nature*, 246(5428):96–98, November 1973.

130. Motoo Kimura. Genetic variability maintained in a finite population due to mutational production of neutral and nearly neutral isoalleles. *Genetics research*, 11(3):247–270, 1968.

131. Motoo Kimura. *The neutral theory of molecular evolution.* Cambridge University Press, Cambridge, England, January 1985.

132. E E Harris and J Hey. Human populations show reduced DNA sequence variation at the factor IX locus. *Curr. Biol.*, 11(10):774–778, May 2001.

133. M Przeworski, R R Hudson, and A Di Rienzo. Adjusting the focus on human variation. *Trends Genet.*, 16 (7):296–302, July 2000.

134. N Yu, Z Zhao, Y X Fu, N Sambuughin, M Ramsay, T Jenkins, E Leskinen, L Patthy, L B Jorde, T Kuromori, and W H Li. Global patterns of human DNA sequence variation in a 10-kb region on chromosome 1. *Mol. Biol. Evol.*, 18(2):214–222, February 2001.

135. Jeffrey D Wall and Jonathan K Pritchard. Haplotype blocks and linkage disequilibrium in the human genome. *Nat. Rev. Genet.*, 4(8):587–597, August 2003.

136. Feng-Chi Chen and Wen-Hsiung Li. Genomic divergences between humans and other hominoids and the effective population size of the common ancestor of humans and chimpanzees. *Am. J. Hum. Genet.*, 68(2): 444–456, February 2001.

137. Ralph Burgess and Ziheng Yang. Estimation of hominoid ancestral population sizes under bayesian coalescent models incorporating mutation rate variation and sequencing errors. *Mol. Biol. Evol.*, 25(9):1979–1994, September 2008.

138. Mathieu Brevet and Nicolas Lartillot. Reconstructing the history of variation in effective population size along phylogenies. *Genome Biology and Evolution*, 13(8):evab150, 2021.

139. N Takahata. Allelic genealogy and human evolution. *Mol. Biol. Evol.*, 10(1):2–22, January 1993.

140. Albert Tenesa, Pau Navarro, Ben J Hayes, David L Duffy, Geraldine M Clarke, Mike E Goddard, and Peter M Visscher. Recent human effective population size estimated from linkage disequilibrium. *Genome Res.*, 17 (4):520–526, April 2007.

141. Ryan D Hernandez, Melissa J Hubisz, David A Wheeler, David G Smith, Betsy Ferguson, Jeffrey Rogers, Lynne Nazareth, Amit Indap, Traci Bourquin, John McPherson, Donna Muzny, Richard Gibbs, Rasmus Nielsen, and Carlos D Bustamante. Demographic histories and patterns of linkage disequilibrium in chinese and indian rhesus macaques. *Science*, 316(5822):240–243, April 2007.

142. Adam Eyre-Walker. Changing effective population size and the McDonald-Kreitman test. *Genetics*, 162(4): 2017–2024, December 2002.

143. Eugene E Harris. Nonadaptive processes in primate and human evolution. *Am. J. Phys. Anthropol.*, 143 Suppl 51:13–45, 2010.

144. Francis Crick, Leslie Barnett, Sydney Brenner, Richard J Watts-Tobin, et al. General nature of the genetic code for proteins. 1961.

145. Marvin B Shapiro and Periannan Senapathy. RNA splice junctions of different classes of eukaryotes: sequence statistics and functional implications in gene expression. *Nucleic Acids Res.*, 15(17):7155–7174, 1987.

146. Liqing Zhang and Wen-Hsiung Li. Mammalian housekeeping genes evolve more slowly than tissue-specific genes. *Mol. Biol. Evol.*, 21(2):236–239, February 2004.

147. Nick Goldman and Ziheng Yang. A codon-based model of nucleotide substitution for protein-coding dna sequences. *Molecular biology and evolution*, 11(5):725–736, 1994.

148. R Nielsen and Z Yang. Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. *Genetics*, 148(3):929–936, March 1998.

149. Ziheng Yang. Paml 4: phylogenetic analysis by maximum likelihood. *Molecular biology and evolution*, 24(8): 1586–1591, 2007.

150. Carolin Kosiol, Tomáš Vinař, Rute R da Fonseca, Melissa J Hubisz, Carlos D Bustamante, Rasmus Nielsen, and Adam Siepel. Patterns of positive selection in six mammalian genomes. *PLoS genetics*, 4(8):e1000144, 2008.

151. Ziheng Yang. PAML: a program package for phylogenetic analysis by maximum likelihood. *Bioinformatics*, 13(5):555–556, 1997.

152. Austin L Hughes and Masatoshi Nei. Pattern of nucleotide substitution at major histocompatibility complex class i loci reveals overdominant selection. *Nature*, 335(6186):167–170, 1988.

153. Z Yang and J P Bielawski. Statistical methods for detecting molecular adaptation. *Trends Ecol. Evol.*, 15 (12):496–503, December 2000.

154. John H McDonald and Martin Kreitman. Adaptive protein evolution at the adh locus in drosophila. *Nature*, 351(6328):652–654, 1991.

155. Jane Charlesworth and Adam Eyre-Walker. The mcdonald–kreitman test and slightly deleterious mutations. *Molecular biology and evolution*, 25(6):1007–1015, 2008.

156. Philipp W Messer and Dmitri A Petrov. Frequent adaptation and the mcdonald–kreitman test. *Proceedings of the National Academy of Sciences*, 110(21):8615–8620, 2013.

157. Nicolas Lartillot and Raphaël Poujol. A phylogenetic model for investigating correlated evolution of substitution rates and continuous phenotypic characters. *Molecular biology and evolution*, 28(1):729–744, 2011.

158. Emilia P Martins and Thomas F Hansen. Phylogenies and the comparative method: A general approach to incorporating phylogenetic information into the analysis of interspecific data. *Am. Nat.*, 149(4):646–667, April 1997.

159. Xavier Farré, Ruben Molina, Fabio Barteri, Paul R H J Timmers, Peter K Joshi, Baldomero Oliva, Sandra Acosta, Borja Esteve-Altava, Arcadi Navarro, and Gerard Muntané. Comparative analysis of mammal genomes unveils key genomic variability for human life span. *Mol. Biol. Evol.*, 38(11):4948–4961, October 2021.

160. Sacha Vignieri. Zoonomia. *Science*, 380(6643):356–357, April 2023.

161. Genome 10K Community of Scientists. Genome 10K: a proposal to obtain whole-genome sequence for 10,000 vertebrate species. *J. Hered.*, 100(6):659–674, November 2009.

162. Mary-Claire King and Allan C Wilson. Evolution at two levels in humans and chimpanzees: Their macromolecules are so alike that regulatory mutations may account for their biological differences. *Science*, 188 (4184):107–116, 1975.

163. Wolfgang Enard, Philipp Khaitovich, Joachim Klose, Sebastian Zöllner, Florian Heissig, Patrick Giavalisco, Kay Nieselt-Struwe, Elaine Muchmore, Ajit Varki, Rivka Ravid, Gaby M Doxiadis, Ronald E Bontrop, and Svante Pääbo. Intra- and interspecific variation in primate gene expression patterns. *Science*, 296(5566): 340–343, April 2002.

164. Peter D Price, Daniela H Palmer Droguett, Jessica A Taylor, Dong Won Kim, Elsie S Place, Thea F Rogers, Judith E Mank, Christopher R Cooney, and Alison E Wright. Detecting signatures of selection on gene expression. *Nat Ecol Evol*, 6(7):1035–1045, July 2022.

165. Zhong Wang, Mark Gerstein, and Michael Snyder. RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.*, 10(1):57–63, January 2009.

166. Aleksandar Janjic, Lucas E Wange, Johannes W Bagnoli, Johanna Geuder, Phong Nguyen, Daniel Richter, Beate Vieth, Binje Vick, Irmela Jeremias, Christoph Ziegenhain, Ines Hellmann, and Wolfgang Enard. Prime-seq, efficient and powerful bulk RNA sequencing. *Genome Biol.*, 23(1):88, March 2022.

167. M Love, S Anders, and W Huber. Differential analysis of count data–the DESeq2 package. *Genome Biol.*, 2014.

168. Matthew E Ritchie, Belinda Phipson, Di Wu, Yifang Hu, Charity W Law, Wei Shi, and Gordon K Smyth. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.*, 43(7):e47, April 2015.

169. Matthew Chung, Vincent M Bruno, David A Rasko, Christina A Cuomo, José F Muñoz, Jonathan Livny, Amol C Shetty, Anup Mahurkar, and Julie C Dunning Hotopp. Best practices on the differential expression analysis of multi-species RNA-seq. *Genome Biol.*, 22(1):121, April 2021.

170. Paul Bastide, Charlotte Soneson, David B Stern, Olivier Lespinet, and Mélina Gallopin. A phylogenetic framework to simulate synthetic interspecies RNA-Seq data. *Mol. Biol. Evol.*, 40(1), January 2023.

171. Yoav Gilad and Orna Mizrahi-Man. A reanalysis of mouse ENCODE comparative gene expression data. *F1000Res.*, 4:121, May 2015.

172. Gabriel E Hoffman and Panos Roussos. Dream: powerful differential expression analysis for repeated measures designs. *Bioinformatics*, 37(2):192–201, April 2021.

173. Rori V Rohlfs and Rasmus Nielsen. Phylogenetic ANOVA: The expression variance and evolution model for quantitative trait evolution. *Syst. Biol.*, 64(5):695–708, September 2015.

174. Lam si Tung Ho and Cécile Ané. A linear-time algorithm for gaussian and non-gaussian trait evolution models. *Syst. Biol.*, 63(3):397–408, May 2014.

175. Matthew R E Symonds and Simon P Blomberg. A primer on phylogenetic generalised least squares. In László Zsolt Garamszegi, editor, *Modern Phylogenetic Comparative Methods and Their Application in Evolutionary Biology: Concepts and Practice*, pages 105–130. Springer Berlin Heidelberg, Berlin, Heidelberg, 2014.

176. Jesualdo A Fuentes-G, Paul David Polly, and Emília P Martins. A bayesian extension of phylogenetic generalized least squares: Incorporating uncertainty in the comparative study of trait relationships and evolutionary rates. *Evolution*, 74(2):311–325, February 2020.

177. Marguerite A Butler and Aaron A King. Phylogenetic comparative analysis: A modeling approach for adaptive evolution. *Am. Nat.*, 164(6):683–695, December 2004.

178. Trevor Bedford and Daniel L Hartl. Optimization of gene expression by natural selection. *Proc. Natl. Acad. Sci. U. S. A.*, 106(4):1133–1138, January 2009.

179. Jenny Chen, Ross Swofford, Jeremy Johnson, Beryl B Cummings, Noga Rogel, Kerstin Lindblad-Toh, Wilfried Haerty, Federica di Palma, and Aviv Regev. A quantitative framework for characterizing the evolutionary history of mammalian gene expression. *Genome Res.*, 29(1):53–63, January 2019.

180. Roger Mundry. Statistical issues and assumptions of phylogenetic generalized least squares. In László Zsolt Garamszegi, editor, *Modern Phylogenetic Comparative Methods and Their Application in Evolutionary Biology: Concepts and Practice*, pages 131–153. Springer Berlin Heidelberg, Berlin, Heidelberg, 2014.

181. Casey W Dunn, Felipe Zapata, Catriona Munro, Stefan Siebert, and Andreas Hejnol. Pairwise comparisons across species are problematic when analyzing functional genomic data. *Proc. Natl. Acad. Sci. U. S. A.*, 115 (3):E409–E417, January 2018.

182. Gwenael Badis, Michael F Berger, Anthony A Philippakis, Shaheynoor Talukder, Andrew R Gehrke, Savina A Jaeger, Esther T Chan, Genita Metzler, Anastasia Vedenko, Xiaoyu Chen, Hanna Kuznetsov, Chi-Fong Wang, David Coburn, Daniel E Newburger, Quaid Morris, Timothy R Hughes, and Martha L Bulyk. Diversity and complexity in DNA recognition by transcription factors. *Science*, 324(5935):1720–1723, June 2009.

183. Arttu Jolma, Jian Yan, Thomas Whitington, Jarkko Toivonen, Kazuhiro R Nitta, Pasi Rastas, Ekaterina Morgunova, Martin Enge, Mikko Taipale, Gonghong Wei, Kimmo Palin, Juan M Vaquerizas, Renaud Vincentelli, Nicholas M Luscombe, Timothy R Hughes, Patrick Lemaire, Esko Ukkonen, Teemu Kivioja, and Jussi Taipale. DNA-binding specificities of human transcription factors. *Cell*, 152(1-2):327–339, January 2013.

184. Meghana M Kulkarni and David N Arnosti. Information display by transcriptional enhancers. 2003.

185. David N Arnosti and Meghana M Kulkarni. Transcriptional enhancers: Intelligent enhanceosomes or flexible billboards? *Journal of cellular biochemistry*, 94(5):890–898, 2005.

186. Klara Stefflova, David Thybert, Michael D Wilson, Ian Streeter, Jelena Aleksic, Panagiota Karagianni, Alvis Brazma, David J Adams, Iannis Talianidis, John C Marioni, Paul Flicek, and Duncan T Odom. Cooperativity and rapid evolution of cobound transcription factors in closely related mammals. *Cell*, 154(3):530–540, August 2013.

187. Ryan Rickels and Ali Shilatifard. Enhancer logic and mechanics in development and disease. *Trends Cell Biol.*, 28(8):608–630, August 2018.

188. Granton A Jindal and Emma K Farley. Enhancer grammar in development, evolution, and disease: dependencies and interplay. *Developmental cell*, 56(5):575–587, 2021.

189. Swann Floc'hlay, Emily S Wong, Bingqing Zhao, Rebecca R Viales, Morgane Thomas-Chollier, Denis Thieffry, David A Garfield, and Eileen E M Furlong. Cis-acting variation is common across regulatory layers but is often buffered during embryonic development. *Genome Res.*, 31(2):211–224, February 2021.

190. Adam Siepel, Gill Bejerano, Jakob S Pedersen, Angie S Hinrichs, Minmei Hou, Kate Rosenbloom, Hiram Clawson, John Spieth, Ladeana W Hillier, Stephen Richards, George M Weinstock, Richard K Wilson, Richard A Gibbs, W James Kent, Webb Miller, and David Haussler. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.*, 15(8):1034–1050, August 2005.

191. Katherine S Pollard, Melissa J Hubisz, Kate R Rosenbloom, and Adam Siepel. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res.*, 20(1):110–121, January 2010.

192. Ilan Gronau, Leonardo Arbiza, Jaaved Mohammed, and Adam Siepel. Inference of natural selection from interspersed genomic elements based on polymorphism and divergence. *Mol. Biol. Evol.*, 30(5):1159–1171, May 2013.

193. Leonardo Arbiza, Ilan Gronau, Bulent A Aksoy, Melissa J Hubisz, Brad Gulko, Alon Keinan, and Adam Siepel. Genome-wide inference of natural selection on human transcription factor binding sites. *Nat. Genet.*, 45(7):723–729, July 2013.

194. Dara G Torgerson, Adam R Boyko, Ryan D Hernandez, Amit Indap, Xiaolan Hu, Thomas J White, John J Sninsky, Michele Cargill, Mark D Adams, Carlos D Bustamante, and Andrew G Clark. Evolutionary processes acting on candidate cis-regulatory regions in humans inferred from patterns of polymorphism and divergence. *PLoS Genet.*, 5(8):e1000592, August 2009.

195. Martin C Frith, Michael C Li, and Zhiping Weng. Cluster-Buster: Finding dense clusters of motifs in DNA sequences. *Nucleic Acids Res.*, 31(13):3666–3668, July 2003.

196. V Matys, E Fricke, R Geffers, E Gößling, M Haubrock, R Hehl, K Hornischer, D Karas, A E Kel, O V Kel-Margoulis, D-U Kloos, S Land, B Lewicki-Potapov, H Michael, R Münch, I Reuter, S Rotert, H Saxel, M Scheer, S Thiele, and E Wingender. TRANSFAC ® : transcriptional regulation, from patterns to profiles. *Nucleic Acids Res.*, 31(1):374–378, January 2003.

197. Charles E Grant, Timothy L Bailey, and William Stafford Noble. FIMO: scanning for occurrences of a given motif. *Bioinformatics*, 27(7):1017–1018, April 2011.

198. Dominic Schmidt, Michael D Wilson, Benoit Ballester, Petra C Schwalie, Gordon D Brown, Aileen Marshall, Claudia Kutter, Stephen Watt, Celia P Martinez-Jimenez, Sarah Mackay, Iannis Talianidis, Paul Flicek, and Duncan T Odom. Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding. *Science*, 328(5981):1036–1040, May 2010.

199. Benoit Ballester, Alejandra Medina-Rivera, Dominic Schmidt, Mar Gonzàlez-Porta, Matthew Carlucci, Xiaoting Chen, Kyle Chessman, Andre J Faure, Alister P W Funnell, Angela Goncalves, Claudia Kutter, Margus Lukk, Suraj Menon, William M McLaren, Klara Stefflova, Stephen Watt, Matthew T Weirauch, Merlin Crossley, John C Marioni, Duncan T Odom, Paul Flicek, and Michael D Wilson. Multi-species, multi-transcription factor binding highlights conserved control of tissue-specific biological pathways. *Elife*, 3: e02626, October 2014.

200. Duncan T Odom, Robin D Dowell, Elizabeth S Jacobsen, William Gordon, Timothy W Danford, Kenzie D MacIsaac, P Alexander Rolfe, Caitlin M Conboy, David K Gifford, and Ernest Fraenkel. Tissue-specific transcriptional regulation has diverged significantly between human and mouse. *Nat. Genet.*, 39(6):730–732, June 2007.

201. Charles G Danko, Lauren A Choate, Brooke A Marks, Edward J Rice, Zhong Wang, Tinyi Chu, Andre L Martins, Noah Dukler, Scott A Coonrod, Elia D Tait Wojno, John T Lis, W Lee Kraus, and Adam Siepel.

Dynamic evolution of regulatory element ensembles in primate CD4+ T cells. *Nat Ecol Evol*, 2(3):537–548, March 2018.

202. Camille Berthelot, Diego Villar, Julie E Horvath, Duncan T Odom, and Paul Flicek. Complexity and conservation of regulatory landscapes underlie evolutionary resilience of mammalian gene expression. *Nat. Ecol. Evol.*, 2(1):152–163, January 2018.

203. Julian Banerji, Sandro Rusconi, and Walter Schaffner. Expression of a β-globin gene is enhanced by remote sv40 dna sequences. *Cell*, 27(2):299–308, 1981.

204. Fumitaka Inoue and Nadav Ahituv. Decoding enhancers using massively parallel reporter assays. *Genomics*, 106(3):159–164, 2015.

205. Robin Andersson, Claudia Gebhard, Irene Miguel-Escalada, Ilka Hoof, Jette Bornholdt, Mette Boyd, Yun Chen, Xiaobei Zhao, Christian Schmidl, Takahiro Suzuki, et al. An atlas of active enhancers across human cell types and tissues. *Nature*, 507(7493):455–461, 2014.

206. Caitlin Mills, Anushya Muruganujan, Dustin Ebert, Crystal N Marconett, Juan Pablo Lewinger, Paul D Thomas, and Huaiyu Mi. Peregrine: a genome-wide prediction of enhancer to gene relationships supported by experimental evidence. *PloS one*, 15(12):e0243791, 2020.

207. Zane Kliesmete, Peter Orchard, Victor Yan Kin Lee, Johanna Geuder, Simon M Krauss, Mari Ohnuki, Jessica Jocher, Beate Vieth, Wolfgang Enard, and Ines Hellmann. Evidence for compensatory evolution within pleiotropic regulatory elements. *bioRxiv*, pages 2024–01, 2024.

208. Carl G de Boer, Eeshit Dhaval Vaishnav, Ronen Sadeh, Esteban Luis Abeyta, Nir Friedman, and Aviv Regev. Deciphering eukaryotic gene-regulatory logic with 100 million random promoters. *Nat. Biotechnol.*, 38(1): 56–65, January 2020.

209. Aaron T L Lun, Samantha Riesenfeld, Tallulah Andrews, The Phuong Dao, Tomas Gomes, participants in the 1st Human Cell Atlas Jamboree, and John C Marioni. EmptyDrops: distinguishing cells from empty droplets in droplet-based single-cell RNA sequencing data. *Genome Biol.*, 20(1):63, March 2019.

210. Stephen J Fleming, Mark D Chaffin, Alessandro Arduini, Amer-Denis Akkad, Eric Banks, John C Marioni, Anthony A Philippakis, Patrick T Ellinor, and Mehrtash Babadi. Unsupervised removal of systematic background noise from droplet-based single-cell experiments using CellBender. *Nat. Methods*, 20(9):1323–1335, September 2023.

211. Shiyi Yang, Sean E Corbett, Yusuke Koga, Zhe Wang, W Evan Johnson, Masanao Yajima, and Joshua D Campbell. Decontamination of ambient RNA in single-cell RNA-seq with DecontX. *Genome Biol.*, 21(1), December 2020.

212. Emre Caglayan, Yuxiang Liu, and Genevieve Konopka. Neuronal ambient RNA contamination causes misinterpreted and masked cell types in brain single-nuclei datasets. *Neuron*, 110(24):4043–4056.e5, December 2022.

213. Jonathan A Griffiths, Arianne C Richard, Karsten Bach, Aaron T L Lun, and John C Marioni. Detection and removal of barcode swapping in single-cell RNA-seq data. *Nat. Commun.*, 9(1):1–6, July 2018.

214. Atray Dixit. Correcting chimeric crosstalk in single cell RNA-seq experiments. *bioRxiv*, December 2016.

215. Swati Parekh, Christoph Ziegenhain, Beate Vieth, Wolfgang Enard, and Ines Hellmann. The impact of amplification on differential expression analyses by RNA-seq. *Sci. Rep.*, 6(1):25533, May 2016.

216. Christoph Ziegenhain, Beate Vieth, Swati Parekh, Björn Reinius, Amy Guillaumet-Adkins, Martha Smets, Heinrich Leonhardt, Holger Heyn, Ines Hellmann, and Wolfgang Enard. Comparative analysis of single-cell RNA sequencing methods. *Mol. Cell*, 65(4):631–643.e4, February 2017.

217. Grace X Y Zheng, Jessica M Terry, Phillip Belgrader, Paul Ryvkin, Zachary W Bent, Ryan Wilson, Solongo B Ziraldo, Tobias D Wheeler, Geoff P McDermott, Junjie Zhu, Mark T Gregory, Joe Shuga, Luz Montesclaros, Jason G Underwood, Donald A Masquelier, Stefanie Y Nishimura, Michael Schnall-Levin, Paul W Wyatt, Christopher M Hindson, Rajiv Bharadwaj, Alexander Wong, Kevin D Ness, Lan W Beppu, H Joachim Deeg, Christopher McFarland, Keith R Loeb, William J Valente, Nolan G Ericson, Emily A Stevens, Jerald P Radich, Tarjei S Mikkelsen, Benjamin J Hindson, and Jason H Bielas. Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.*, 8(1):14049, January 2017.

218. Matthew D Young and Sam Behjati. SoupX removes ambient RNA contamination from droplet-based single-cell RNA sequencing data. *Gigascience*, 9(12):giaa151, December 2020.

219. J Ding, X Adiconis, S K Simmons, M S Kowalczyk, C C Hession, and N D Marjanovic. Systematic comparison single- cell single-nucleus RNA-sequencing methods. *Nat Biotechnol*, 38(6):737–746, 2020.

220. Stathis Megas, Valentina Lorenzi, and John C Marioni. EmptyDropsMultiome discriminates real cells from background in single-cell multiomics assays. *Genome Biol.*, 25(1):121, May 2024.

221. Guillaume Junion, Mikhail Spivakov, Charles Girardot, Martina Braun, E Hilary Gustafson, Ewan Birney, and Eileen EM Furlong. A transcription factor collective defines cardiac cell fate and reflects lineage history. *Cell*, 148(3):473–486, 2012.

222. Sabina Domené, Viviana F Bumaschny, Flávio SJ de Souza, Lucía F Franchini, Sofía Nasif, Malcolm J Low, and Marcelo Rubinstein. Enhancer turnover and conserved regulatory function in vertebrate evolution. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 368(1632):20130027, 2013.

223. Zeba Wunderlich, Meghan DJ Bragdon, Ben J Vincent, Jonathan A White, Javier Estrada, and Angela H DePace. Krüppel expression levels are maintained through compensatory evolution of shadow enhancers. *Cell reports*, 12(11):1740–1747, 2015.

224. José Aguilar-Rodríguez, Joshua L Payne, and Andreas Wagner. A thousand empirical adaptive landscapes and their navigability. *Nature ecology & evolution*, 1(2):0045, 2017.

225. Michael Z Ludwig, Casey Bergman, Nipam H Patel, and Martin Kreitman. Evidence for stabilizing selection in a eukaryotic enhancer element. *Nature*, 403(6769):564–567, 2000.

226. Cosmas D Arnold, Daniel Gerlach, Daniel Spies, Jessica A Matts, Yuliya A Sytnikova, Michaela Pagani, Nelson C Lau, and Alexander Stark. Quantitative genome-wide enhancer activity maps for five drosophila species show functional enhancer conservation and turnover during cis-regulatory evolution. *Nature genetics*, 46(7):685–692, 2014.

227. Noah Dukler, Yi-Fei Huang, and Adam Siepel. Phylogenetic modeling of regulatory element turnover based on epigenomic data. *Mol. Biol. Evol.*, 37(7):2137–2152, July 2020.

228. Irene Gallego Romero and Amanda J Lea. Leveraging massively parallel reporter assays for evolutionary questions. *Genome Biology*, 24(1):26, 2023.

229. Roy N Platt, Michael W Vandewege, and David A Ray. Mammalian transposable elements and their impacts on genome evolution. *Chromosome Research*, 26:25–43, 2018.

230. Barbara McClintock. Controlling elements and the gene. In *Cold Spring Harbor symposia on quantitative biology*, volume 21, pages 197–216. Cold Spring Harbor Laboratory Press, 1956.

231. Chris P Ponting and Wilfried Haerty. Genome-wide analysis of human long noncoding rnas: a provocative review. *Annual review of genomics and human genetics*, 23:153–172, 2022.

232. Ming Zhao, Caiping Ren, Hong Yang, Xiangling Feng, Xingjun Jiang, Bin Zhu, Wen Zhou, Lei Wang, Ying Zeng, and Kaitai Yao. Transcriptional profiling of human embryonic stem cells and embryoid bodies identifies hesrg, a novel stem cell gene. *Biochemical and biophysical research communications*, 362(4):916–922, 2007.

233. Guifei Li, Caiping Ren, Jia Shi, Wei Huang, Hui Liu, Xiangling Feng, Weidong Liu, Bin Zhu, Chang Zhang, Lei Wang, et al. Identification, expression and subcellular localization of esrg. *Biochemical and biophysical research communications*, 435(1):160–164, 2013.

234. Jichang Wang, Gangcai Xie, Manvendra Singh, Avazeh T Ghanbarian, Tamás Raskó, Attila Szvetnik, Huiqiang Cai, Daniel Besser, Alessandro Prigione, Nina V Fuchs, Gerald G Schumann, Wei Chen, Matthew C Lorincz, Zoltán Ivics, Laurence D Hurst, and Zsuzsanna Izsvák. Primate-specific endogenous retrovirus-driven transcription defines naive-like stem cells. *Nature*, 516(7531):405–409, December 2014.

235. Shasha Li, Hui Liu, Weidong Liu, Ning Shi, Ming Zhao, Siyi Wanggou, Weiren Luo, Lei Wang, Bin Zhu, Xiang Zuo, Wen Xie, Cong Zhao, Yao Zhou, Longlong Luo, Xiang Gao, Xingjun Jiang, and Caiping Ren. ESRG is critical to maintain the cell survival and self-renewal/pluripotency of hPSCs by collaborating with MCM2 to suppress p53 pathway. *Int. J. Biol. Sci.*, 19(3):916–935, January 2023.

236. Konrad J Karczewski, Laurent C Francioli, Grace Tiao, Beryl B Cummings, Jessica Alföldi, Qingbo Wang, Ryan L Collins, Kristen M Laricchia, Andrea Ganna, Daniel P Birnbaum, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*, 581(7809):434–443, 2020.

237. Wen Xie, Weidong Liu, Lei Wang, Shasha Li, Zilin Liao, Hongjuan Xu, Yihan Li, Xingjun Jiang, and Caiping Ren. Embryonic stem cell related gene regulates alternative splicing of transcription factor 3 to maintain human embryonic stem cells' self-renewal and pluripotency. *Stem Cells*, 42(6):540–553, June 2024.

238. Simon M Reader, Yfke Hager, and Kevin N Laland. The evolution of primate general and cultural intelligence. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 366(1567):1017–1027, 2011.

239. URL http://www.brainmuseum.org/.

240. Wally Welker. Why does cerebral cortex fissure and fold? In *Cerebral cortex*, pages 3–136. Springer, 1990.

241. Eric Lewitus, Iva Kelava, Alex T Kalinka, Pavel Tomancak, and Wieland B Huttner. An adaptive threshold in mammalian neocortical evolution. *PLoS biology*, 12(11):e1002000, 2014.

242. Stephen H Montgomery, Isabella Capellini, Chris Venditti, Robert A Barton, and Nicholas I Mundy. Adaptive evolution of four microcephaly genes and the evolution of brain size in anthropoid primates. *Molecular biology and evolution*, 28(1):625–638, 2010.

243. AM Boddy, MR McGowen, CC Sherwood, LI Grossman, M Goodman, and DE Wildman. Comparative analysis of encephalization in mammals reveals relaxed constraints on anthropoid primate and cetacean brain scaling. *Journal of evolutionary biology*, 25(5):981–994, 2012.

244. Eric Lewitus, Iva Kelava, and Wieland B Huttner. Conical expansion of the outer subventricular zone and the role of neocortical folding in evolution and development. *Frontiers in human neuroscience*, 7:424, 2013.

245. Wolfgang Enard. Comparative genomics of brain size evolution. *Frontiers in human neuroscience*, 8:345, 2014.

246. Maria Ángeles Martínez-Martínez, Camino De Juan Romero, Virginia Fernández, Adrián Cárdenas, Magdalena Götz, and Víctor Borrell. A restricted period for formation of outer subventricular zone defined by cdh1 and trnp1 levels. *Nature communications*, 7:11812, 2016.

247. Ronny Stahl, Tessa Walcher, Camino De Juan Romero, Gregor Alexander Pilz, Silvia Cappello, Martin Irmler, José Miguel Sanz-Aquela, Johannes Beckers, Robert Blum, Víctor Borrell, and Magdalena Götz. Trnp1 regulates expansion and folding of the mammalian cerebral cortex by control of radial glial fate. *Cell*, 153(3):535–549, April 2013. ISSN 0092-8674. doi: 10.1016/j.cell.2013.03.027.

248. G A Pilz, A Shitamukai, I Reillo, E Pacary, J Schwausch, R Stahl, J Ninkovic, H J Snippert, H Clevers, L Godinho, F Guillemot, V Borrell, F Matsuzaki, and M Gotz. Amplification of progenitors in the mammalian telencephalon includes a new radial glial cell type. *Nat. Commun.*, 4:2125, 2013.

249. Cemil Kerimoglu, Linh Pham, Anton B Tonchev, M Sadman Sakib, Yuanbin Xie, Godwin Sokpor, Pauline Antonie Ulmke, Lalit Kaurani, Eman Abbas, Huong Nguyen, et al. H3 acetylation selectively promotes basal progenitor proliferation and neocortex expansion. *Science advances*, 7(38):eabc6792, 2021.

250. Víctor Borrell and Isabel Reillo. Emerging roles of neural stem cells in cerebral cortex development and evolution. *Developmental neurobiology*, 72(7):955–971, 2012.

251. Victor Borrell and Federico Calegari. Mechanisms of brain evolution: regulation of neural progenitor cell diversity and cell cycle length. *Neuroscience research*, 86:14–24, 2014.

252. Ari Löytynoja. Phylogeny-aware alignment with prank. *Multiple sequence alignment methods*, pages 155–170, 2014.

253. Susan D Healy and Candy Rowe. A critique of comparative studies of brain size. *Proceedings of the Royal Society B: Biological Sciences*, 274(1609):453–464, 2007.

254. Sabina Kanton, Michael James Boyle, Zhisong He, Malgorzata Santel, Anne Weigert, Fátima Sanchís-Calleja, Patricia Guijarro, Leila Sidow, Jonas Simon Fleck, Dingding Han, et al. Organoid single-cell genomic atlas uncovers human-specific features of brain development. *Nature*, 574(7778):418–422, 2019.

255. D Allan Drummond, Jesse D Bloom, Christoph Adami, Claus O Wilke, and Frances H Arnold. Why highly expressed proteins evolve slowly. *Proceedings of the National Academy of Sciences*, 102(40):14338–14343, 2005.

256. Adrian WR Serohijos, Zilvinas Rimas, and Eugene I Shakhnovich. Protein biophysics explains why highly abundant proteins evolve slowly. *Cell reports*, 2(2):249–256, 2012.

257. Jian-Rong Yang, Ben-Yang Liao, Shi-Mei Zhuang, and Jianzhi Zhang. Protein misinteraction avoidance causes highly expressed proteins to evolve slowly. *Proceedings of the National Academy of Sciences*, 109 (14):E831–E840, 2012.

258. Julian Echave and Claus O Wilke. Biophysical models of protein evolution: understanding the patterns of evolutionary sequence divergence. *Annual review of biophysics*, 46:85–103, 2017.

259. A Keith Dunker, Celeste J Brown, J David Lawson, Lilia M Iakoucheva, and Zoran Obradović. Intrinsic disorder and protein function. *Biochemistry*, 41(21):6573–6582, 2002.

260. Peter Tompa. Intrinsically unstructured proteins. *Trends in biochemical sciences*, 27(10):527–533, 2002.

261. Tobias Sikosek and Hue Sun Chan. Biophysics of protein evolution and evolutionary protein biophysics. *Journal of The Royal Society Interface*, 11(100):20140419, 2014.

262. Zhirong Liu and Yongqi Huang. Advantages of proteins being disordered. *Protein Science*, 23(5):539–550, 2014.

263. Emmanuel Paradis and Klaus Schliep. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in r. *Bioinformatics*, 35(3):526–528, 2019.

264. Robert P Freckleton, Paul H Harvey, and Mark Pagel. Phylogenetic analysis and comparative data: a test and review of evidence. *The American Naturalist*, 160(6):712–726, 2002.

265. Cosmas D Arnold, Daniel Gerlach, Christoph Stelzer, Łukasz M Boryń, Martina Rath, and Alexander Stark. Genome-wide quantitative enhancer activity maps identified by starr-seq. *Science*, 339(6123):1074–1077, 2013.

266. Jason D Buenrostro, Paul G Giresi, Lisa C Zaba, Howard Y Chang, and William J Greenleaf. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, dna-binding proteins and nucleosome position. *Nature methods*, 10(12):1213–1218, 2013.

267. Xinchen Wang, Liang He, Sarah M Goggin, Alham Saadat, Li Wang, Nasa Sinnott-Armstrong, Melina Claussnitzer, and Manolis Kellis. High-resolution genome-wide functional dissection of transcriptional regulatory regions and nucleotides in human. *Nature communications*, 9(1):5380, 2018.

List of Figures

Acknowledgements

With this I would like to thank so many people that were there for me, that believed in me and my ideas and, most importantly, that made these past years so fun. First and foremost, I want to thank Ines Hellmann for the fruitful and insightful guidance and all the discussions, I don't think I will ever stop learning from you. I also want to thank Wolfgang Enard for the amount of strength and inspiration you are giving me without noticing it. The two of you and your way of conducting science has had a very strong impact on my scientific and personal development throughout the last eight years.

I would also like to thank Johanna, Philipp and Lucas for being the best PhD buddies ever. You have always been there for me and strongly enriched my scientific and personal views. I don't think there are many people that know me the way you do. I would furthermore like to thank Anita, Paulina, Eva, Fiona, Daniel, Beate and Victor for being the awesome scientists and friends that you are: Impressive, strong and supportive. To all of the mentioned people, I would like to give a round of applause for tolerating my spontaneous bursts of thought spams in verbal and written form. Moreover, to the whole Enard/Hellmann group, including - but not limited to - Aleks, Mari, Jessy, Ines B., Karin, Sara, Felix, Dana, Antonia and many, many great previous students and colleagues: I could have not wished for a more warm and vibrating environment to pursue science and party.

I also want to thank my fellow PhD candidates from outside the group that were the perfect company to have a Friday beer with and laugh our assess off: Alessa, Liza, Lorenz, Miri, Santi, Constance and others. Furthermore, Julita, Maxi, Stefan and Hannah, although we each went into a different direction, your ideas and warm hearts still keep amazing me. I would also like to thank my dearest friends at home, Ilva, Ildze and Laura, for making home feel like home. Most of all, I want to thank my parents, sister and brother for being my strong rocks and helping me make my dreams come true. Last but not least, Gabi, thank you for being my best friend and providing strong support in whatever adventures I decide are appropriate.