On Evolution of Gene Regulation in *Mus*

Dissertation

der Mathematisch-Naturwissenschaftlichen Fakultät

der Eberhard Karls Universität Tübingen

zur Erlangung des Grades eines

Doktors der Naturwissenschaften

(Dr. rer. nat.)

vorgelegt von

Volker Soltys

aus Bietigheim-Bissingen

Tübingen

2024

Gedruckt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät der Eberhard Karls Universität Tübingen.

Tag der mündlichen Qualifikation:	28.06.2024
Dekan:	Prof. Dr. Thilo Stehle
1. Berichterstatter/-in:	Prof. Dr. Yingguang Frank Chan
2. Berichterstatter/-in:	Prof. Dr. Alfred Nordheim

Table of Contents

Summary

What constitutes the genetic basis of adaptation is a fundamental question in evolutionary biology. Evolution of gene regulation is a major contributor to phenotypic variation and as such plays a critical role in adaptation. In general, regulatory changes can be facilitated through genetic changes in either *cis*-regulatory elements (CREs), which modify gene expression of local genes on their own allele, or in *trans*-regulatory factors such as transcription factors, which can affect any gene in the genome. Pinpointing genetic variation facilitating expression changes is challenging, but current evidence points to *cis*-regulatory changes being the main contributor to regulatory evolution, though this is debated. However, much remains unknown about how gene regulation at the chromatin, transcript and cellular level evolves in mammals, especially at the crucial transition from individual to species-level differences. In this thesis, I investigated cell-type specific regulatory evolution between several *Mus* species with particular focus on the role of CREs. In the first chapter, I developed easySHARE-seq, a single-cell technique simultaneously measuring gene expression and chromatin accessibility. I show that easySHARE-seq generates high-quality datasets and removes cost-prohibitive barriers, allowing diverse and flexible study design. I further demonstrate how the simultaneous measurements can be exploited to survey the *cis*-regulatory landscape of cell types and link CREs to their target gene. In the second chapter, I apply easySHARE-seq to four different species and their F1 hybrids from *Mus* to investigate how gene regulation evolved across them. I find that in all cell types *cis*-regulatory changes become pervasive with increasing evolutionary divergence. However, between closer related species the majority of regulatory changes occur in *trans*. Furthermore, I argue that some cell types might follow common evolutionary trajectories independent of species, possibly due to similar selective pressures. Lastly, I link CREs to their target gene in each species and cell type and show that these linked CREs are generally under purifying selection yet those linked to *cis*-regulated genes show signatures of adaptive evolution. These results contribute to uncovering the genetic basis of adaptation by demonstrating that CREs are the dominant driver of regulatory evolution across *Mus*. They also provide a novel approach in identifying genetic variants underlying regulatory changes in CREs.

Zusammenfassung

Was genau die genetische Grundlage von Adaptation bildet ist eine grundlegende Frage der Evolutionsbiologie. Die Evolution von Genregulation trägt maßgeblich zu phänotypischer Variation bei und spielt deshalb eine wichtige Rolle in der Adaptation. Generell können regulatorische Veränderungen durch genetische Veränderungen in entweder *cis*-regulatorischen Elementen (CREs), welche Genexpression von lokalen Genen auf ihrem Allele modifizieren, oder in *trans*-regulatorischen Faktoren wie Transkriptionsfaktoren ermöglicht werden. Die exakten genetischen Varianten welche regulatorische Veränderungen auslösen zu lokalisieren ist herausfordernd, aber aktuelle Studien deuten darauf hin, dass *cis*-regulatorische Veränderungen den Hauptbeitrag zu regulatorischer Evolution leisten, jedoch existiert darüber eine lebhafte Debatte. Wenig ist darüber bekannt wie Genregulation auf Chromatin, Transkript oder zellulärer Ebene in Säugetieren evolviert, vor allem am entscheidenden Übergang von individuellen zu Spezies-spezifischen Unterschieden. In dieser Thesis habe ich die Zelltyp-spezifische Evolution von Genregulation zwischen verschiedenen Mausspezies untersucht mit speziellem Fokus auf die Rolle von CREs. Im ersten Kapitel habe ich easySHARE-seq entwickelt, eine Einzelzellmethode welche Genexpression und Chromatinzugänglichkeit gleichzeitig misst. Ich zeige, dass easySHARE-seq mit geringem Kostenaufwand hochwertige Datensätze generiert und somit vielfältiges und flexibles Studiendesign ermöglicht. Des Weiteren zeige ich, wie man die simultanen Messungen ausnutzen kann um die *cis*-regulatorische Landschaft zu untersuchen und um CREs mit ihrem Zielgenen zu verknüpfen. Im zweiten Kapitel wende ich easySHARE-seq an vier verschiedenen Mausarten und ihren F1 Hybriden an um zu untersuchen, wie Genregulation zwischen diesen evolviert ist. Ich zeige, dass in allen Zelltypen *cis*-regulatorische Veränderungen mit zunnehmender evolutionärer Divergenz dominant werden jedoch zwischen näher verwandten Spezies die Mehrheit an regulatorischen Veränderungen in *trans* geschieht. Außerdem deuten meine Ergebnisse darauf hin, dass manche Zelltypen ähnlichen evolutionären Dynamiken unabhängig der Spezies folgen, möglicherweise aufgrund von ähnlichem Selektionsdruck. Als letztes verknüpfe ich CREs mit ihren Zielgenen in jeder Spezies und Zelltyp und zeige, dass die verknüpften CREs generell unter negativer Selektion stehen jedoch diese welche mit *cis*-regulierten Genen verknüpft sind Merkmale von Adaptation aufweisen. Diese

Ergebnisse tragen dazu bei, die genetische Grundlage von Adaptation aufzudecken indem sie zeigen, dass CREs die dominanten Antreiber von regulatorischer Evolution in *Mus* sind. Sie zeigen auch eine neue Herangehensweise in der Identifizierung genetischer Varianten auf, welche regulatorische Veränderungen durch CREs verursachen.

Publications

Unpublished:

The evolutionary dynamics of cell-type specific regulatory evolution in Mus;
Volker Soltys, Moritz Peters, Dingwen Su, Yingguang Frank Chan
Manuscript in preparation, ready to submit

Published:

Flexible and high-throughput simultaneous profiling of gene expression and chromatin accessibility in single cells;
Volker Soltys, Moritz Peters, Dingwen Su, Marek Kučka, Yingguang Frank Chan;
Pre-print in bioRxiv [doi: https://doi.org/10.1101/2024.02.26.581705]

Genetic studies of human–chimpanzee divergence using stem cell fusions;
Janet H. T. Song, Rachel L. Grant, Veronica C. Behrens, Marek Kučka, Garrett A. Roberts Kingman, **Volker Soltys**, Yingguang Frank Chan and David M. Kingsley;
Proceedings of the National Academy of Sciences 118 No. 51 e2117557118 [https://doi.org/10.1073/pnas.2117557118]

Not included in this manuscript:

Nuclear dualism without extensive DNA elimination in the ciliate *Loxodes magnus*;
Brandon K. B. Seah, Aditi Singh, David E. Vetter, Christiane Emmerich, Moritz Peters, **Volker Soltys**, Bruno Huettel, Estienne Swart
bioRxiv [doi: https://doi.org/10.1101/2023.11.09.566212]

Copy number normalization distinguishes differential signals driven by copy number differences in ATAC-seq and ChIP-seq;
Dingwen Su, Moritz A. Peters, **Volker Soltys**, Yingguang Frank Chan
bioRxiv [doi: https://doi.org/10.1101/2024.04.11.588815]

Introduction

What constitutes the genetic basis of phenotypic variation? This is a longstanding question that has motivated genetics research for more than a century[1–3]. Many researchers tackle this problem by investigating how heritable differences in gene expression lead to phenotypic variation and thus provide a substrate for natural selection. One of the most of debated aspects concerns the relative position of genetic variation that ultimately confer expression changes: Do the proteins themselves change or is it the non-coding DNA, thereby altering their regulation? These two mechanisms differ strongly in their implications: If a proteins' structure is changed, besides resulting in the differential expression of the focal gene, this can lead to pleiotropic regulatory changes of genes in every cell type the protein is expressed. In contrast, since non-coding DNA elements are usually context specific, this type of change should be more specific and come with potentially less deleterious pleiotropic effects. Even though many studies investigated these questions, our knowledge about regulatory evolution in mammalian systems is sparse. A further challenge limiting our understanding of regulatory evolution is to determine if a gene expression change is adaptive.

In this thesis, I use several species within the genus *Mus* to investigate the genetic basis for the evolution of gene expression across these species, with particular focus on the role of non-coding regulatory sequences called *cis*-regulatory elements. To contextualize my work, I will briefly talk about our current understanding of how gene regulation works, the theoretical framework upon which my work is built, introduce core concepts needed to understand it and describe the enormous potential that advancements in new methodologies hold. In Chapter One, I report a newly developed single-cell method called easySHARE-seq, measuring both gene expression and chromatin accessibility within single-cells. By measuring both simultaneously, we can assess their intermolecular dynamics and establish links between these two modalities. Additionally, this method significantly improves upon current single-cell techniques as it provides a highly flexible framework, improves upon data quality and reduces costs. In Chapter Two, I make use of easySHARE-seq to dissect the evolutionary dynamics of regulatory evolution in *Mus*. I show what role *cis*-regulatory elements (CREs) played in their evolution of gene expression, argue that a cell type

can follow a common evolutionary trajectory even across several species, and show how new genomic techniques can be used to connect *cis*-regulatory elements to their target genes.

Overview of Gene Regulation

Over 60 years ago, Francis Crick formulated the 'Central Dogma'[4], stating that once information has passed into a protein, it cannot get out again. James Watson then further modified it into how we understand it today: That information from DNA gets transcribed into mRNA, which is then translated into proteins[5]. The main purpose of DNA is to store information, the main purpose of proteins is to fulfill their specific function. What is a key aspect then is which and how much information is ultimately turned into function, which is to a large degree (but not exclusively) controlled by mRNA levels and thus, gene expression. The layers in which gene expression can be regulated are manifold, but in broad terms there are two types of mechanisms, *cis* and *trans. Cis*-regulation directly drives transcription itself and is linked to its own molecule via *cis*-regulatory elements such as enhancers, promotors or silencers (for a review, see Gasperini et al.[6]). These are stretches of non-coding DNA, typically a few hundred basepair (bp) long and in proximity to their target gene[7] but there have been reports of cis-regulatory elements (CREs) regulating genes over nearly 1 Mb distance ($1x10^9$ bp)[8]. *Trans*-regulation covers all other mechanisms of gene regulation, including any indirect mechanism, from amino acid changes in proteins, e.g. transcription factors (TF), to long-non-coding RNA to DNA conformation to post-translational modification of the protein itself and any other transcriptional or translational mechanism[9–12]. Therefore, *cis*- and *trans*-regulation together cover all types of gene regulatory mechanisms and are intertwined into several regulatory layers.

First off, CREs can only be functional when they are free of histones and thereby accessible for all kinds of proteins such as TFs[13]. The current view is that so-called 'pioneer factors' bind to heterochromatin and open up the closed chromatin[14]. How CREs exactly confer their function is subject of many studies and not entirely clear. Often, CREs contain multiple TF binding sites consisting of lengths between 6-12 bp to which then multiple TFs might bind simultaneously in a cooperative manner[15,16]. This would suggest that the directionality, affinity, spacing and arrangement of these TF binding sites can have a defined influence on CRE functionality, a concept which

sometimes is summarized under the term of (enhancer) grammar[17]. That might be of particular importance within their native, cellular context since the use of enhancers in transgenic assays suggest that they can function independent of their directionality[18]. One enticing discovery was that TF binding sites seem to be rarely optimized for binding affinity. By examining millions of synthetic variants of a specific developmental enhancer in the sea squirt *Ciona intestinalis*, Farley et al. could show that increased binding affinity of TFs frequently leads to ectopic expression of the target gene[19]. Thus, the authors proposed that the general low-affinity binding in enhancers confers tissue specificity by ensuring combinatorial control of gene expression[20].

Next, the bound TFs recruit general co-factors, which can function as either activators or repressors[16]. These co-factors can have diverse functions such as nucleosome remodelers[21], histone modifiers[22], mediator complexes[23] or function as a scaffold[24]. As diverse as their function as numerous their possibilities to influence and regulate gene expression, far more than can be summarized here (for further reviews, see [16,23,25]). One intriguing study demonstrated how the rate of transcription directly depends and changes with the presence or absence of TFs and thus co-factors at CREs, showcasing the combinatorial nature of enhancers and how mutations in TF binding sites might directly lead to gene expression changes[26]. Through these recruited co-factors, CREs then come into physical contact with their target gene where in the case of enhancers, core transcriptional machinery is recruited and transcription initiated. Lastly, CREs are imbedded in their 3D regulatory landscape commonly known as Topologically Associated Domain (TAD)[27,28], which are genomic regions that are interacting with themselves and are mostly insulated from neighboring genomic regions. Therefore, they are presumed to shape which gene a CRE can interact with. In general tough, much remains unknown about how gene regulation functions. Therefore, conflicting ideas about different aspects of gene regulation exist, which can only be resolved by further investigations.

The short overview of gene regulation given here is by no means exhaustive and there are several aspects of gene regulation not discussed here. For example, the entire range of post-translational modifications to proteins, which also serve regulatory functions, especially in the context of TFs[29]. Additionally, there are many other processes that either directly or indirectly can regulate transcriptional output of a gene, such as chromatin state[30], how physical proximity between a CRE and its target gene is established, modifications of the rate of transcription[31], alternative splicing[32] as well

as a whole array of post-transcriptional modification such as modifying the rate of mRNA decay[33] or miRNAs[34]. Furthermore, several CREs can influence the same gene simultaneously, adding an additional layer of complexity[35]. In summary, gene regulation is an interplay between several complex processes and by no means fully understood. In consequence, investigation of gene regulation comprises one of the most active fields in molecular biology, also because its dysregulation is implicated in a plethora of diseases. In the year 2023 alone, 350 articles were published in the journal *Nature* that contained the term 'gene regulation' in their title (www.nature.com).

However, measuring gene expression and possibly other layers of gene regulation such as chromatin status allows to simultaneously capture *cis*- and *trans*-regulatory modes. Since transcription is key in converting genetic variation into functional changes, it strongly influences traits and ultimately plays a major role during selection and adaptation. The systematic assessment of *cis*- and *trans*-regulatory changes would therefore provide vital insight into how gene regulation may evolve – either through genetic changes directly linked to transcriptional activity (*cis*) or through other indirect means (trans).

Methodologies to investigate Gene Regulation

Scientific breakthroughs and discoveries are, among other things, often tied to opportunity. To this end, the methodologies how gene expression and CREs and thus gene regulation can be studied have improved and diversified tremendously in terms of quality, power and modality they measure since genome-wide assays became available. However, as I will describe down below, to gain a better understanding of gene regulation further improvements are needed.

In general, we can distinguish between assays measuring gene expression (transcriptomics) and those investigating epigenetic features such as chromatin accessibility or histone marks (epigenomics). For transcriptomics, the most commonly used method is RNA-seq. This measures the abundance of mRNA transcripts by either capturing them at their polyA-tail, which results in a dataset that tends to be enriched toward the 3' end of the transcript, or by using random primers, which is often used to capture full-length transcripts and thus investigate alternative splicing. In addition, there are transcriptomic assays measuring different regulatory layers of gene expression. For example, PRO-seq (and its successor ChRO-seq) measures nascent

transcripts that are actively transcribed and therefore enables investigation of e.g. polymerase pausing[36]. A further technique called Ribo-seq measures transcripts that are actively translated, which allows assessment of the extent of post-transcriptional regulation or buffering[37].

In contrast to transcriptomics, epigenomics and methods profiling CREs are far more diverse, reflecting the multiple regulatory layers. The possibly most widely used class of assays is measuring which parts of the genome are not occupied by histones. As described above, CREs are only functional when chromatin is accessible so these techniques provide a genome-wide readout of potentially active CREs, including promotors of actively transcribed genes. However, it is important to keep in mind that not all accessible regions are automatically CREs. Early types of these assays used DNase[38] or MNase[39] to digest open chromatin regions, but nowadays the most popular method is ATAC-seq[40], due to its great sensitivity, resolution and ease of use. ATAC-seq has been used to profile the *cis*-regulatory landscape of a vast number of tissues and developmental processes across diverse taxa[41–44], thereby underscoring the context dependency and evolutionary importance of CREs. A second class of assays investigates protein binding to regulatory DNA, the most common of which are ChIP-seq or CUT&Tag. Both work similarly in that DNA to which the protein in question binds is sequenced by enriching for it using an antibody. These methods revealed for example that different classes of CREs are usually marked by different combinations of histone modifications[45]. Other applications include measuring which TFs binds to a given CRE[46]. More classes of assays investigating different regulatory layers exist. For example, there is a whole array of methodologies (generally denoted by “C” as in conformation) measuring the 3D landscape of the genome using so-called ‘contact-maps’ [47–49]. This is useful for defining TADs or showcasing how differential TADs might facilitate and accelerate evolutionary processes[50,51]. Another exciting technology relies on the CRISPR system to directly edit CREs in their native context and measure resulting expression changes[52], potentially allowing the direct linking of a CRE to its target gene.

Altogether, most described methods rose in popularity in the wake of high-throughput sequencing in the early 2000s as this made them economically feasible. Over the following years however, it became clear that their lack of resolution placed clear hurdles for further scientific advancements. Namely, all these methods average their signal over multiple cell types or even entire tissues, only occasionally singular cell

populations can be assayed. This has clear limitations. For one, the averaged signal might not be representative of any cell type within a tissue[53]. Second, rare or infrequent cell types are difficult to assay, especially if it is not possible to enrich for them. However, these can have defining regulatory or developmental functions[54,55]. Third, interactions between cell types are very challenging to measure. And lastly, developmental processes and cell fate decisions cannot be resolved properly since cell types in these processes often exist in a continuum and cells from the same cell type can display considerable heterogeneity[56].

Single-cell methodologies can overcome those limitations by providing readouts for individual cells. These methodologies emerged around 15 years ago and matured rapidly alongside a suite of computational tools[57]. Today there are a large number of protocols and platforms available, often several for each of the above-described assays (e.g. RNA-seq, ATAC-seq etc.), both commercially and custom. Commercially, the most widely-used platform is droplet-based, where individual cells are incapsulated in droplets and barcoded. However, this comes with the drawbacks of high costs for both instrumentation and kits as well as limited throughput, typically around 10.000 per reaction. Therefore, custom protocols have been developed, many of which rely upon a concept called combinatorial indexing or 'split-and-pool'. Here, a suspension containing dissociated cells (or nuclei) is distributed onto multi-well plates (typically 96-well) where each well contains a different barcode, which is then attached to for example cDNA molecules. Then, these cells are pooled and distributed across more multi-well plates, again containing and attaching a different barcode within each well. This approach scales exponentially and after three rounds of barcoding over 880.000 (96^3) combinations are possible. Therefore, the chance that two cells share the same combination of barcodes is slim, effectively achieving single-cell resolution. Exploiting this principle, it became possible to assay hundreds of thousands of cells in one experiment at a fraction of commercial costs[58,59], which allowed the single-cell profiling of entire embryos or developmental trajectories[60–63]. Other important applications are in the medical field, for example the profiling of cancer cells and discovery of their mutations, which can differ from cell to cell[64]. Thus, single-cell technologies are of major importance in both basic and applied research and likely will continue to be so for the foreseeable future. However, their widespread application also showed that these technologies are not able to resolve many open questions. More specifically, to understand processes such as gene regulation or cell fate decisions, multiple layers

of information (e.g. gene expression and chromatin accessibility) need to be measured simultaneously in the same cell to directly query intermolecular dynamics between the transcriptome and epigenome and how genetic variation impacts cellular function and phenotype[65]. For example, measuring chromatin accessibility and gene expression within the same cell might allow to directly link a CRE to their target gene[6]. Alternatively, directly measuring how shifts in histone modifications or TF binding impact gene expression can advance our knowledge about CREs and how they function[66]. In short, measuring multiple layers of information in single cells is poised to advance and transform our understanding of the genome in both health and disease.
These technologies are commonly known as 'multiomic' single-cell technologies and only few are currently available. As these are early days, they still suffer from several drawbacks. For one, commercial solutions are incredibly expensive, placing a limit on sample size, throughput and study design. The few custom protocols that exist produce data of suboptimal quality, have high cost or suboptimal throughput[67,68] and are thus not frequently used. The most advanced custom technology to date is called SHARE-seq[69], which is able to quantify gene expression and chromatin accessibility in hundreds of thousands of cells simultaneously with reasonable data quality. Only, its framework is very inflexible placing clear limitations on study design and resulting in prohibitive costs. For example, analyzing allele-specific expression is very challenging as the sequencing read has a maximum length of 100bp and thus less frequently captures genetic variation needed to separate allelic signals. To conclude, advancing these technologies by making them cheaper, more suited for a variety of study designs and improving upon data quality would enable the scientific community to harness their full power to address questions that have been asked for decades.

Investigating the Evolution of Gene Regulation

How can one quantify and characterize the relative contribution of *cis*- or *trans*-regulatory changes to the evolution of gene regulation? There are two popular approaches to investigate this between species, subspecies or populations.
The first makes use of F1 hybrids in diploid organisms. First, expression changes between the parental species are measured for each gene. Next, gene expression between the alleles in the F1 hybrids is measured and these two measurements are then compared. The F1 hybrid context is important here: as both alleles are present in

the same cell, any indirect effect from diffusible factors, such as the presence or absence of a TF, lncRNA, etc. would affect both alleles. This is often referred to as a "common *trans* environment" and typically removes a portion of transcript level differences between the parental species. However, if a difference in expression still persists between F1 hybrid alleles ("allelic imbalance"), these remaining differences must be due to changes in *cis*. This can be due to the presence or change of a CRE on one parental allele, or differences in TADs due to specific diseases or (induced) mutations[70,71]. This approach therefore allows to investigate each gene, determine if evolved expression change is due to *cis*- or *trans*-regulatory changes as well as directly quantify their net effect. It was first used in 2004[72] and has since been a frequent choice to investigate regulatory evolution genome-wide[73–76]. However, this approach also comes with several limitations. For one, since it can only measure net expression change, it can't determine how many regulatory changes have occurred. A second major shortcoming is that it is not possible to identify potential loci or genetic variants that might be responsible for observed expression changes.

The second approach in turn makes use of available genetic variation in populations and is known as expression quantitative loci (eQTL) mapping. By measuring gene expression across many individuals in a diverse population, one can then correlate gene expression with their genetic variants and identify new functional loci mediating expression change (eQTL) as well as estimate their effect sizes[77]. These loci are then separated into *cis*- or *trans*-eQTL. However, the definition of *cis* in this approach is often different compared to the first approach for several reasons: first, *cis* is solely defined by distance of the eQTL to the gene, especially among the earliest studies[78]. Second, since gene expression is typically measured at the total expression level, as opposed to the allele-specific expression level, eQTL studies typically cannot directly determine *cis* regulation. Furthermore, this approach has less statistical power in identifying *trans*-eQTL because of their higher multiple testing burden. Another major drawback is that since eQTL mapping relies on populations with natural genetic variation, it cannot be performed across as large evolutionary divergences as in the F1 hybrid approach, limiting its potential for deciphering general patterns of regulatory evolution.

In summary, both approaches are frequently used and come with their own set of advantages and disadvantages. An ideal approach would combine the per-gene

measure of the first approach with the identification of candidate genetic loci of the second.

The *cis*-regulatory Hypothesis

What causes phenotypic variation that selection can act upon? What is the genetic basis for adaptation? These are some of the oldest questions in evolutionary biology and already have been asked by Darwin (only the former) and Fisher[79,80]. Heritable phenotypic variation must arise through genetic variation, which ultimately leads to different phenotypes by altering e.g., developmental processes or behavior. Studying how gene regulation and gene expression evolves is thus one approach to address these questions, as gene expression is assumed to immensely influence a phenotype. How *cis*- and *trans*-regulatory changes of gene regulation shape the evolution of gene regulation is therefore an important question and extensive field of study and over the decades, several hypotheses have been put forward. Today, one of the most prominent, well-studied and well-supported idea on how gene regulation evolves to produce phenotypic variation is called the '*cis*-regulatory hypothesis'. This states that genetic variation in CREs plays a more important role in producing phenotypic variation (and thus adaptation) than mutations in coding sequences[81,82]. Phrased more broadly, the differential regulation of proteins is more important than changes in the proteins themselves. While first formulated in the 2000s, the roots of the *cis*-regulatory hypothesis lie in the 1960s and 1970s[83,84], with several important discoveries throughout the next decades leading up to it. First, King and Wilson discovered in their landmark paper that differences in proteins between chimpanzees and humans cannot be sufficient for the observed differences between the species but noted that a change in expression may account for the major organismal differences[85]. Second, the discovery of *Hox* genes showed that a set of highly conserved genes makes up the majority of known body plans in animals[86,87]. Previously, it was thought that this differed between different species[88]. Finally, the Human Genome Project at the time revealed that the human genome only encodes around 24,500 genes, which was notably less than expected and only around twice as much as a worm or a fly[89]. In addition, this made clear that while the number of proteins does not increase according to organismal complexity, the amount of non-coding DNA in the genome does.

Altogether, these and further discoveries[90,91] culminated in the formulation of the '*cis* regulatory hypothesis', which states several foundational principles[81,82]. First, because CREs act in a modular, independent and additive fashion, they can precisely facilitate gene expression change in a single cell type or developmental timepoint[16,92,93]. In comparison, a TF is usually implicated in several cell types or processes, so that a change in its coding sequences lead to more pleiotropic, potentially negative effects. For example, it could disrupt essential protein-protein interactions or abolish TF binding to a promotor altogether, which can have catastrophic consequences[94,95]. Second, coding sequences are often highly conserved between different species and taxa[96,97]. Already in 2002, Aparicio et al. reported that three quarters of predicted proteins in humans have a homolog in the pufferfish, despite around 450 million years of evolutionary divergence[98]. This 'deep homology' [82] again emphasizes that proteins do not arise newly at a rate comparable to the diversity of organismal complexity. Lastly, the mutational target size for CREs is greater than for coding sequences[99], meaning that the chance of beneficial *cis*-regulatory mutations is higher. However, this does depend upon the specific organism. For example, the baker's yeast *Saccharomyces cerevisiae* has a highly condensed genome with 68% of the genome consisting of coding sequences, potentially resulting in a higher target size for *trans*-regulatory factors[100–102].

When this hypothesis was explicitly formulated, a limited number of case studies provided experimental support for it, causing some initial pushback[103]. However, over the following years many striking examples of major adaptive changes caused by *cis*-regulatory changes were published. For example, one study examined progressive limb loss in snakes. It identified snake-specific sequence changes in a long-range limb enhancer of *Shh*, which otherwise is conserved across a wide range of vertebrates. When they substituted the murine counterpart with the snake ortholog, this resulted in severe limb reduction whereas substitution with the human or fish ortholog resulted in normal limb development[104]. Further probing revealed that the loss of a single TF binding site in the enhancer resulted in this major change of body plan. Similarly, stickleback fish (*Gasterosteus aculeatus*) repeatedly evolved pelvic reduction when adapting to freshwater environments. The gene *Pitx1* has been hypothesized to mediate this effect, but its coding sequence remained unchanged in freshwater stickleback fish[105]. A further study could show that the repeated deletion of a tissue-specific enhancer of *Pitx1* is responsible for pelvic reduction[106]. These and other

studies showcased that even small regulatory changes can result in strong phenotypic variation and lead to new adaptive alleles. More evidence came from studies that concentrated on more genome-wide patterns of regulatory evolution[107]. For example, two studies showed that between yeast species or sister species of *Drosophila* the majority of expression differences are due to *cis*-regulatory changes[108,109].

Now, 15 years later, where does the *cis*-regulatory hypothesis stand? In general, it is accepted by many that *cis*-regulatory changes play a major, if not the major role in regulatory evolution and adaptation. Several principles or trends, that are however not without prominent exceptions, seem to have emerged. For example, one recurring conclusion has been that gene expression generally is under stabilizing selection. This is suggested because it has been repeatedly shown that *cis*- and *trans*-regulatory changes act frequently upon the same gene in opposing directions, effectively compensating each other[72,78,110–112]. Next, when comparing evolution of gene expression within and between species, it has often been found that *trans*-regulatory changes contribute more to changes within species[101,113,114]. In contrast to this, with increasing evolutionary divergence, *cis*-regulatory changes then seem to become pervasive, especially between species [74,75,111,115].

However, opposing theories have also been put forward, most prominently the so-called 'Omnigenic Model' [116]. This partitions genes influencing a trait into 'core genes' and 'peripheral genes' and postulates that core genes have a large direct effect on a trait but peripheral genes explain the majority of heritability via indirect *trans*-effects[112]. This theory is mostly the consequence of Genome-Wide Association Studies (GWAS) repeatedly showing that the majority of heritability of a trait or disease is explained by large numbers of small-effect variants[117,118].

Lastly, I want to note two open questions which when addressed, might progress our understanding of regulatory evolution and adaptation and open up new avenues for future studies. For one, the overwhelming majority of available studies investigated regulatory evolution on a tissue-level by averaging expression data over all cell types. However, regulatory changes are likely adaptive in a distinct cell type, therefore expanding our knowledge how individual cell types evolve regulatory changes might reveal previously undetectable evolutionary patterns. Second, it is curious to note that the majority of all above-described studies have been performed in either plants, yeast, flies or fish and only very few are investigating how evolution of gene regulation proceeds in mammals or to what extent they follow the described trends. In fact, even

in the house mouse, one of the most studied animals to date, our knowledge is fragmentary. Given the differences between other taxa, it is entirely possible that mammals might show unique evolutionary dynamics[119,120] in their regulatory evolution.

The genus *Mus* and its Application in Evolutionary Biology

Because of their phylogenetic history and subsequent widespread use in biological sciences, the genus *Mus* (more specifically its subgenus *Mus*) is an excellent system for studying evolution. The most frequently used mouse strains in these types of study originate from the house mouse, *Mus musculus*. House mice are spread across all inhabited continents; however, they have evolved into several main subspecies, namely *Mus musculus musculus*, *Mus musculus domesticus* and *Mus musculus castaneus*. These three subspecies likely started to diverge around 0.5 million years ago in the Indo-valley in current-day India[121–123] and from there colonized Eurasia and evolved into their respective subspecies[124,125]. In the present day, *M. m. castaneus* is mainly found in India and Southeast Asia, *M. m. musculus* is spread across most of north Eurasia and *M. m. domesticus* is found in western Europe, Africa and by means of human travel, throughout all of North and South America. Molecular data indicates that *M. m. castaneus* and *M. m. musculus* are more closely related than either to *M. m. domesticus*[119,120,123,126], though these relationships are not entirely clear. Indicative of how distinct the subspecies are, a stable hybrid zone between *M. m. domesticus* (east) and *M. m. musculus* (west) has been identified which runs through Central Europe, from Scandinavia through Germany up to the black sea[127,128]. Additionally, several other subspecies of *M. musculus* are identified or hypothesized[129], for example *M. m. molossinus* in Japan[130]. However, these will not be described further in this thesis. Other frequently used mouse strains in evolutionary research stem from different *Mus* species, the most prominent of which from *Mus spretus*. Also known as the 'Algerian Mouse', this is the closest related species to *Mus musculus*, with estimates of these species diverging around 1-3 million years ago[126,131]. It inhabits south-western Europe as well as the mediterranean coast of Africa[132] and thus lives sympatrically with *M. m. domesticus* [133]. Again, many more species in the subgenus *Mus* exist, some of which like *Mus caroli* are also used in evolutionary research. However, these will not be described further.

In the 20th century, mice as laboratory organisms soared in popularity and the first inbred mouse strains were created from *Mus musculus*[134]. However, partially because it wasn't known at the time, partially because the original mice stock before was crossed and selected for its appearance, these early mouse lines upon which all 'classical' strains today are based, did not come from a single subspecies of *Mus musculus*[125] but rather were a mix. The most widely-used mouse strain in all of biological and medical science, C57BL/6 (BL6), originated at this time and genetic studies showed that the majority of its genome derives from *M. m. domesticus* with smaller contributions from *M. m. musculus* and *M. m. castaneus*[120,135].

These classical mouse strains were used in many important scientific and medical discoveries and continue to be incredibly useful today. However, they offered little opportunity in researching evolutionary biology as the phenotypic variation between them is minimal. Therefore, over several decades new strains from pure *M. musculus* subspecies as well as several other *Mus* species were derived, which are now known as so-called 'wild-derived' strains[136]. For all these strains, several mice were caught in the wild and over many generations inbred mouse strains were derived. For example, the mouse strains PWD/Ph (PWD) or PWK/Ph (PWK) were derived from the *Mus musculus musculus* subspecies by trapping pairs of wild mice in 1972 in the central part of the Czech Republic[137]. Similarly, the strain CAST/Ei (CAST) was derived from wild-caught *M. m. castaneus* mice in Thailand and the strain SPRET/Ei (SPRET) from wild-caught *M. spretus* mice. Together, these strains form an excellent basis for the investigation of evolutionary processes for multiple reasons. First, they have ample genetic and phenotypic variation between them. Thus, they are commonly used to study speciation, the molecular basis of phenotypic variation or the genetics of complex traits[138]. Second, all these wild-derived strains can form viable interspecific F1 hybrid crosses with BL6 mice, which makes them an invaluable tool for e.g. genetic mapping or investigating regulatory evolution[139–141]. Perhaps most extraordinary, it was possible to derive F1 hybrid embryonic stem cells from these crosses[142], opening up the possibility to assess evolutionary changes in developmental processes.

Regarding research into regulatory evolution across *Mus*, and more specifically into how *cis*- and *trans*-regulatory changes contribute to the evolution of gene regulation, very few studies have been published. For example, when using F1 hybrids to assess *cis*- or *trans*-regulatory changes in whole testis between PWK/PhJ and LEWES/EiJ (a wild-derived strain of *M. m. domesticus*), it was found that *cis*-changes contribute more

to expression divergence than changes in *trans*[143]. A second study assessing liver concluded that only 2% of differentially expressed genes changed due to *trans*-regulatory changes alone compared to 43% of genes with *cis*-acting changes[110]. Lastly, the most thorough study so far over several tissues and strains suggested that over 80% of genes have *cis*-regulatory variation[140]. However, what is missing is an overarching thorough description how gene regulation evolved between the different species and subspecies. As described, singular data points between few comparisons do exist, but these are not enough to extrapolate a more global view or to compare it with known patterns of regulatory evolution in other taxa.

Altogether, the unique phylogenetic history of *Mus* paired with extensive phenotypic as well as genetic variation[120,134], the availability of inbred strains from different species and subspecies and the viability of interspecific crosses makes the genus *Mus* an excellent system for probing evolutionary processes.

Objectives

The study of regulatory evolution addresses some of the oldest questions in evolutionary biology. In this thesis, I have combined the latest technologies in transcriptomics and epigenomics with a classical F1 hybrid design to investigate how gene expression evolves across *Mus* and what role *cis*-regulatory elements play in facilitating it.

In Chapter One, I developed a single-cell multiomic technique called easySHARE-seq measuring both gene expression (RNA-seq) and chromatin accessibility (ATAC-seq) within single cells[144]. This method provides an incredibly flexible framework and eliminates cost-prohibitive barriers, opening up multiomic single-cell techniques to a wide variety of study designs and applications. I showcase the utility and quality of this method by profiling murine liver nuclei. I then connect CREs to their target genes using the simultaneous measurements and lastly identify novel marker genes and CREs displaying a morphogen-dependent dosage effect (zonation).

In Chapter Two, I investigate cell-type specific regulatory across *Mus* by applying this method to liver nuclei from four different subspecies and species as well as their F1 hybrids. I show that with increasing evolutionary divergence, *cis*-regulatory changes become pervasive but that between closer related species, *trans*-regulatory changes account for the majority of expression differences. I then describe how patterns of

regulatory evolution can differ strongly between cell types and also suggest that some cell types might follow a common evolutionary trajectory, even across different species. Lastly, I connect CREs to their target genes within each cell type and species and find that those connected to genes differentially regulated in *cis* show signatures of adaptive evolution.

Additionally, I was involved in a collaboration investigating regulatory evolution between chimpanzee and human by fusing embryonic stem cells from both species to create allotetraploid (4n) stem cells and measuring their gene expression[145]. This study can be found in the Appendix III.

Lastly, I recapitulate the developments and findings from Chapter One and Two in the Discussion where I describe their impacts, especially in context of our current understanding of regulatory evolution, to what degree they agree with the *cis*-regulatory hypothesis and what this contributes to our understanding of the genetic basis of phenotypic variation. I also propose future directions based on these results that might help shed further light onto underlying evolutionary principles.

Chapter One

Flexible and high-throughput simultaneous profiling of gene expression and chromatin accessibility in single cells

Volker Soltys, Moritz Peters, Dingwen Su, Marek Kučka, Yingguang Frank Chan
bioRxiv 2024.02.26.581705; doi: https://doi.org/10.1101/2024.02.26.581705.
Published as pre-print in bioRxiv, under review in Nature Communications.
see Thesis Appendix I

Abstract

Gene expression and chromatin accessibility are highly interconnected processes. Disentangling one without the other provides an incomplete picture of gene regulation. However, simultaneous measurements of RNA and accessible chromatin are technically challenging, especially when studying complex organs with rare cell-types. Here, we present easySHARE-seq, an elaboration of SHARE-seq, providing simultaneous measurements of ATAC- and RNA-seq within single cells, enabling identification of cell-type specific *cis*- regulatory elements (CREs). easySHARE-seq retains high scalability, improves RNA-seq data quality while also allowing for flexible study design. Using 19,664 joint profiles from murine liver nuclei, we linked CREs to their target genes and uncovered complex regulation of key genes such as *Gata4*. We further identify *de novo* genes and *cis*-regulatory elements displaying zonation in Liver sinusoidal epithelial cells (LSECs), a challenging cell type with low mRNA levels, demonstrating the power of multimodal measurements. EasySHARE-seq therefore provides a flexible platform for investigating gene regulation across cell types and scale.

Contributions

V.S. and Y.F.C. designed the experiments. V.S. and M.P. developed the barcoding framework for easySHAREseq. V.S. developed the rest of the protocol and performed experiments. V.S. performed the computational analyses advised by Y.F.C. V.S. drafted the manuscript. M.P., D.S., M.K. and Y.F.C. helped with experimental or computational support. All authors reviewed the manuscript. Y.F.C. directed the study with input from all authors.

Chapter Two

The evolutionary dynamics of cell-type specific regulatory evolution in *Mus*

Volker Soltys, Moritz Peters, Dingwen Su, Yingguang Frank Chan

Manuscript ready to submit.

see Thesis Appendix II

Abstract

Evolution of gene regulation plays a critical role in adaptation and can occur through gene regulatory changes acting in *cis* or *trans*, yet how this differs between individual cell types is poorly understood. Here, we applied single-cell multiomics to 63,551 primary liver nuclei in a set of four closely-related mouse species and their F1 hybrids and profile both gene expression and chromatin accessibility simultaneously at single-cell resolution to investigate cell-type specific regulatory changes as well as linking 118,344 putative *cis*-regulatory elements (pCREs) to their target genes. Between the closest related species, 31.8% of regulatory changes occurred solely in *trans* compared to only 14.8% in *cis*, but the proportion of *cis*-regulated genes increases with both increasing evolutionary divergence and expression difference. However, we find considerable differences in the patterns of regulatory evolution between cell types and that some show consistent regulatory changes independent of species. Lastly, we show that linked pCREs are under purifying selection yet those linked to *cis*-regulated genes show increased genetic divergence, consistent with adaptive evolution. This approach therefore dissects regulatory evolution between cell types and not only allows identification of *cis*-regulated genes but also of possible pCREs facilitating the regulatory change.

Contributions

V.S. and Y.F.C. designed the experiments. V.S. performed the experiments. V.S. performed computational analyses with input from Y.F.C, M.P. and D.S. V.S. wrote the manuscript. M.P., D.S., M.K. and Y.F.C. helped with experimental or computational support. All authors reviewed the manuscript. Y.F.C. directed the study with input from all authors.

Discussion

The assessment of regulatory evolution in order to understand how genetic variation translates into phenotypic variation is a long-standing research topic with the ultimate aim to understand adaptive evolution. Efforts to disentangle *cis*- from *trans*-regulatory changes using F1 hybrids at a genomic scale have started with the advent of Next-Generation Sequencing in 2004[72] and since proved an informative and popular study design. However, as much as was discovered using this approach, little progress has been made in refining or advancing this design. As a result, while there are plenty of case studies, they all suffer from similar limitations and therefore restrict the potential for scientific discovery.

In contrast to that, the breadth of available methodologies to survey general or specific aspects of gene regulation is immense. Most likely, there are more techniques available and established today than at any time before in molecular biology and importantly, these are incredibly diverse. Perhaps most influential have been single-cell techniques, allowing for readouts of single cells and assessment of cell types. They promise enormous advancements in both health and disease and thus are probably one of the most frequently used assays in academic and corporate science.

This thesis aimed to combine the classical F1 hybrid design with new cutting-edge technology to generate new insights when investigating regulatory evolution. By first developing easySHARE-seq, we had a single-cell assay measuring gene expression and surveying the regulatory landscape whose framework suited the requirements of this study. By applying easySHARE-seq to multiple mouse species and their F1 hybrids, we were able to conduct the first study that not only used this design to investigate regulatory evolution in mammals and do so on a cell-type level but also provided an approach to identify *cis*-regulatory elements that possibly facilitated differential gene expression on a per-gene basis. As this immensely expands the possibilities of this design, we expect that similar studies in diverse systems will be conducted following this thesis and its publications.

The impact of easySHARE-seq

The future of single-cell techniques is bright. Already a staple and cornerstone of various fields, including basic research, drug discovery or cancer biology[146,147], continuous drop of costs and increasing ease of use will make these techniques even more widespread. The same applies to multiomic single-cell techniques, which promise to transform our insight into interconnected processes such as gene regulation. However, these come with additional challenges such as the need for more refined cell isolation (and potentially fixation) protocols, more complicated molecular biology and most of all, not yet matured methodologies, showing a clear need for improvement in these areas. The original SHARE-seq[69] methodology is a major step toward making multiomic techniques widely available but it still suffered from drawbacks making it impracticable and economically not feasible for many study designs. Many of these drawbacks have been resolved by the development of easySHARE-seq. For one, it is compatible with standard Illumina sequencing. In contrast, SHARE-seq needs 99bp of sequencing within Index 1, which almost always requires a private and expensive sequencing run. EasySHARE-seq libraries however can be sequenced in concert with other Illumina libraries, even at commercial sequencing services, decreasing the expenses immensely. This comes with the additional advantage of sequencing up to 300bp of the insert (read), which allows for identification and discovery of genetic variants. In this thesis, this was particularly important since many analyses in Chapter Two rely on allele-specific expression. Other fields where variant discovery is vital include cancer biology since cancer cells potentially harbor private variants. Paired with a more flexible framework, we expect easySHARE-seq to be the prominent choice for many, but not all study designs.

Global trends of regulatory evolution in *Mus*

We found several general trends when investigating regulatory evolution in *Mus.* With increasing evolutionary divergence, the fraction of genes whose expression changes are mediated in *cis* increased. This observation was consistent across all cell types. This trend is well documented across other genera such as *Drosophila*, *Saccharomyces* or *Gasterosteus* [74,75,111] but has not been shown before in mammals. Additionally, we could convincingly show that the stronger an evolved expression difference is, the more likely it is regulated in *cis*. However, between the closest related

species (BL6 & CAST/PWD), the majority of expression differences were mediated in *trans*. Here, it is important to keep in mind that CAST, PWD and BL6 are subspecies (I refer to them as separate species for the sake of readability) whereas SPRET is a fully separate species. Here again, many previous studies (but not all) described that more genes are differentially regulated in *trans* within species[74,108,148–151] compared to between species, where *cis*-regulatory changes are the primary source of regulatory variation[74,108,109]. However, all these studies were either performed in yeast or flies and it was not known if mammals, who have a vastly higher fraction of non-coding DNA in their genome[152], would follow these trends. Interestingly, as shown in *Appendix III*, the majority of regulatory variation between chimp and human is also mediated in *cis*[145], underlining the generality of these trends. These observations lead us to conclude that while during initial divergence most expression differences are due to *trans*-regulatory factors, *cis*-regulatory changes are the dominant driver of regulatory evolution in *Mus.* Therefore, CREs are of major importance during regulatory evolution and presumably, adaptation.

A further observation in Chapter Two is that *cis*- and *trans*-regulatory changes often have opposing effects onto a given gene, effectively resulting in little or no expression change. The fraction of genes with opposing *cis* and *trans* effects were the highest in CAST (compared to BL6), followed by PWD and then SPRET. First, the general observation of frequent opposing *cis* and *trans* effects has been well documented, even within mice[74,110,111,143]. This has led to the hypothesis that gene expression is generally under stabilizing selection, resulting in the maintenance of a mean, non-extreme phenotype[78]. However, here it is again important to distinguish within and between species comparisons. Between species, the genetic variants leading to opposing *cis*- and *trans*-regulatory effects are co-inherited. Within species however, there is no guarantee that this is the case, yet the fraction of genes displaying these effects are high. As of yet, no mechanism has been proposed that explains these observations. It is possible that the genetic variants conferring the opposing *cis*- and *trans*-regulatory effects are frequently closely linked, though it seems unlikely that this can explain these widespread effects in within-species comparisons. One study suggested that the high fraction of opposing *cis*- and *trans*-effects is overestimated due to experimental bias and could therefore be an artifact[153]. Lastly, the high frequency of opposing *cis*- and *trans*-effects also showcase how diverged the *Mus musculus* subspecies have become. These contradictory regulatory changes can

cause misexpression in the hybrid, which reduces interbreeding and thus gene flow between the subspecies, ultimately causing each subspecies to evolve into fully separate species.

As mentioned before, studies investigating regulatory evolution in *Mus* species are few but some have been conducted. First, Goncalves et al.[110] assessed regulatory evolution in liver between BL6 and CAST using F1 hybrids. They reported that only 0.6% of genes show expression differences in *trans* and 14% in *cis*. This stands in stark contrast to the results of this study where 31.7% of genes had *trans*-regulatory changes and 14.8% in *cis*. There are several possibilities that might explain the discrepancy. Goncalves et al. was published in 2012 and thus limited by the sequencing capabilities of its time, resulting in lower total transcript number and less statistical power. Potentially then, their results are enriched for highly expressed and highly diverged genes, which in turn are enriched for *cis*-regulated genes (*see above*). Additionally, as this study design relies on identification of genetic variants to analyze allele-specific expression, the set of identified genetic variation between BL6 and CAST was far less complete in 2012. A second study investigated regulatory evolution between PWK and LEWES mice (a wild-derived strain from *Mus musculus domesticus*) in whole testis and also concluded that *cis*-regulatory changes are more frequent than *trans*-regulatory changes (24% compared to 9%). This is most likely a tissue-specific effect. Testes show the highest rate of transcriptome change of any organ or tissue in mammals when comparing between species[154,155] and as described before, higher expression changes are more likely to be mediated in *cis*. In addition, multiple other studies indicate that a high frequency of *cis*-regulatory changes is a testes-specific effect[141,156]. Altogether, our results and previous studies highlight that while general trends do exist and CREs are the dominant driver of regulatory evolution, patterns of evolution of gene regulation are highly species and tissue specific and thus dependent on evolutionary and demographic history of the species assayed. This is exemplified by the differing patterns of regulatory evolution in CAST and PWD, despite sharing similar evolutionary divergence and number of genetic variants.

Cell-type specific evolutionary dynamics

We also analyzed evolution of gene regulation on a cell-type level. While several studies before analyzed regulatory evolution in cell types using eQTL analysis, this study is the first to use the F1 hybrid design and therefore assess larger evolutionary divergences. Assessing regulatory evolution on a cell-type level is essentially a trade-off between cellular resolution and statistical power[157]. The less frequent a cell-type is, the less statistical power and thus genes can be analyzed simply because of lower transcript numbers. This can potentially bias rare cell types toward showing increased *cis*-regulatory changes since higher expression changes are more likely to be detected and as describe above, more frequently regulated in *cis*.

Our analysis revealed that cell types generally follow similar evolutionary trajectories within a species. For example, between BL6 and CAST, all cell types had a higher proportion of *trans*-regulated genes compared to *cis*. However, there are numerous cell-type specific differences within a species. Kupffer Cells in CAST for example had a substantially higher frequency of genes with compensatory gene regulation and thus no net expression change. This could potentially reflect increased stabilizing selection due to conserved and important immunological functions of this cell type[158].

More intriguing however were cell types that showed consistent patterns across species in their regulatory evolution. We show that hepatocytes consistently had the highest proportion of differentially expressed genes of any cell type and those genes also consistently evolved the highest expression changes of any cell type. This was true across all species and perhaps not surprising given that hepatocytes fulfill the main metabolic functions in the liver. In addition though, we found that hepatocyte-specific ATAC-seq peaks, which are highly enriched for hepatocyte-specific CREs, harbored significantly increased genetic variation compared to other cell-type specific peaks. Again, this was consistent across all species. Since the fixation of genetic variants is ultimately the result of selection (and chance), this can hardly be explained by biases in methodology or differences in statistical power. We speculate on two possible, non-exclusive explanations for this observation. First, it is possible that some cell types are predisposed toward certain types of regulatory change by their function, hierarchy or interaction within a tissue. More intriguingly though, it could be that similar strengths and durations of selective pressures can cause similar regulatory responses and that in this case, hepatocytes simply experienced similar selective pressures

across all species. This has already been predicted by Stern & Orgogozo[99] in a landmark paper describing predictions and consequences of the *cis*-regulatory hypothesis. In short, they argue that strong or weak selection and differing durations of selection result in the selection of different types of mutations. They make their case comparing domesticated populations, which they presume to have experienced strong selective pressures, to wild populations and show that the types of regulatory changes that are more likely fixed varies between these two categories. I propose that our observations might fit into this theoretical framework. Since this study is the first to assess cell-type specific regulatory evolution, previous studies were not able to detect these evolutionary dynamics. This might also explain substantial differences between organs and tissues, since those likely experience differing selective pressures. However, I also want to note that more studies across diverse taxa and organs are needed before any confidence can be placed in this proposition.

Linking CREs to their target gene

In this study, we showcase the use of multiomic measurements to link CREs to their target gene. Connecting CREs to target genes is usually laborious and time-intensive and this approach potentially streamlines this effort. More intriguingly though, this provides a defined starting point for the identification of genetic variants underlying expression change, especially for genes who are differentially regulated in *cis*. This therefore directly addresses some long-standing questions in evolutionary biology and to a degree, incorporates one of the major advantages of eQTL mapping into the F1 hybrid design. However, it is limited by several aspects. For one, the identification of links depends on data quality and amount, rendering it difficult to link CREs in rare cell types. Second, these links are only correlations. To definitively connect a CRE to a gene, additional assays are needed. Lastly, the statistical framework for these links is novel and likely to improve in the future. Nevertheless, we argue that this approach identifies true links between CREs and their target gene, mainly because of three datapoints. For one, we found that links to *cis*-regulated genes are the least conserved between species. This is expected since *cis*-regulated genes are defined by a change in *cis*, including the functional loss of a CRE, whereas links to *trans*-regulated genes should be more conserved. Second, links between differentially regulated genes and differentially accessible ATAC-seq peaks scale according to the fraction of genes

regulated in *cis*. This again follows expectations since a change in expression of a *cis*-regulated gene should be accompanied by a change in a CRE. Third, pCREs linked to *cis*-regulated genes have a higher rate of genetic variation than pCREs linked to genes with different regulatory changes, indicating that we might also capture pCREs with functional, potentially adaptive genetic variation. To summarize, these described datapoints provide evidence that using multiomic measurements to link CREs to target genes identifies some true biological connections, providing a defined starting point to identify genetic variants underlying expression change and thus, phenotypic variation.

An updated *cis*-regulatory hypothesis?

We could show that CREs play a major role during regulatory evolution in *Mus* and thus, in their adaptation. This can be seen by an increasing number of *cis*-regulatory changes with increasing evolutionary divergence and also by the observation that stronger expression changes are more likely to be mediated in *cis*. We could also show that CREs linked to *cis*-regulated genes have increased genetic variation compared to those linked to other genes, which is consistent with adaptive evolution. But how does the high frequency of *trans*-regulatory changes within species fit into this picture?
One aspect of the *cis*-regulatory hypothesis that has been rethought is the higher target size for *cis*-regulatory elements. While it is true that the fraction of non-coding DNA is far larger than coding DNA, the effective target size for *cis*-regulatory changes at a gene is likely smaller. To evolve expression change in a gene, many other genes (more precisely, their proteins and other *trans*-factors) which influence this focal gene can be modified or e.g. increased in concentration. In contrast, in a given cellular context only few CREs can be modified to differentially regulate the focal gene. Thus, the effective target size is often much higher for *trans*-regulatory changes[101], which may result in an initial increase of *trans*-regulated genes within species as mutations resulting in *trans*-regulatory changes are more likely to occur. Over time, subspecies become fully different species and more beneficial *cis*-regulatory mutations arise and can be fixed, leading to CREs becoming the dominant driver of regulatory evolution. As described above, this study was not the first to observe this trend and these dynamics have been proposed before[111].
Curiously though, this might also provide the groundwork for unifying seemingly opposite models of regulatory evolution, namely the *cis*-regulatory hypothesis and the

omnigenic model. As a reminder, the omnigenic model partitions genes influencing a phenotype into ‘core genes’, which are few and have a large effect, and ‘peripheral genes’, which explain the majority of the phenotype through small *trans*-effects[116]. This was mainly the result of GWAS studies, which repeatedly identified a few large effect loci and many small effect loci influencing a trait. However, what exactly do GWAS studies measure? They statistically correlate a genetic variant with a trait, relying on large natural populations. In other words, they only compare genetic variation within species, not between. As described above, studies using the F1 hybrid design repeatedly find more *trans*-regulatory changes within species and *cis*-regulatory changes to be dominant between species. Therefore, both approaches might identify similar trends, but across different evolutionary scales since GWAS is not applicable to investigate evolutionary divergences as large as the F1 hybrid design. A supporting aspect is that even within species, the largest expression changes are usually mediated in *cis*, which is also consistent with the ‘core genes’ in the omnigenic model. Thus, these two seemingly opposing models of regulatory evolution might be unifiable. However, there are other theoretical aspects of these models that are still in opposition, therefore caution is advised against overinterpreting this speculation.

Closing Remarks

This study comprehensively surveys regulatory evolution across *Mus* and provides further evidence that *cis*-regulatory elements are the major contributor to evolution of gene regulation, also in mammals. It also describes regulatory differences between cell types and an approach to link CREs to their target gene.

In the future, we hope that similar designs will be used to survey regulatory evolution across a diverse array of organs, species, taxa and kingdoms to expand our knowledge about regulatory evolution and how the underlying dynamics depend on selective pressures. We also hope that further studies might expand, improve or modify this design to suit their needs. Ultimately, this will shed light onto some of the oldest questions of evolutionary biology and how life generated the vast diversity of past and present forms.

Glossary

e.g.:	from Latin exempli gratia or “for example”
bp:	base pair
Mb:	megabase bair
mRNA	messenger- ribonucleic acid
cDNA:	complementary deoxyribonucleic acid
DNA:	deoxyribonucleic acid
polyA:	poly-adenylated
CRE:	cis-regulatory element
eQTL:	expression quantitative trait loci
GWAS:	genome-wide association study
CRISPR:	clustered regularly interspaced short palindromic repeats
ATAC-seq:	Assay for Transposase-Accessible Chromatin using sequencing
ChIP-seq:	chromatin immunoprecipitation sequencing
ChRO-seq:	chromatin run-on and sequencing
CUT&Tag:	cleavage under targets and tagmentation
RNA-seq:	ribonucleic acid sequencing
SHARE-seq:	simultaneous high-throughput ATAC and RNA expression sequencing
TAD:	topologically associated domain
TF:	transcription factor

Acknowledgements

There are many people to thank without whom this thesis would not be as it is or not have happened at all. I want to sincerely thank all of you.

I want to thank my supervisor Prof. Dr. Yingguang Frank Chan for providing me with the opportunity and the freedom to pursue my research interests.

I further thank my TAC members Prof. Dr. Detlef Weigel and Prof. Dr. Boris Maček, for their scientific support and advice as well as Prof. Dr. Alfred Nordheim for agreeing to review this thesis on short notice.

I want to thank Dingwen Su and Moritz Peters for scientific discussion, general and experimental support. Going through this thesis together made it much easier.

I also want to thank all my past or present colleagues who offered support in many ways: Insa Hirschberg, Marek Kučka, Julia Hagauer, Felicity Jones, Elena Avdievich, Cholpon Zhakshylikova, Melanie Kirch, João Castro, Stanley Neufeld, Enni Harjunmaa, Layla Hiramatsu and Sebastian Kick. Further, I thank everyone on the FML floor, especially past and present Weir Lab members, for the good atmosphere.

I also want to thank everyone who provided additional support: Sinja Mattes, Cemal Yilmaz, Rebecca Schwab, the whole IT staff, the Genome Center staff at the MPI Tübingen as well as the Researcher Support Team.

I especially want to thank my friends and family for constant support, distractions and encouragement.

Lastly, the biggest thank you goes to my partner Anne. Thank you so much for your never-ending support, encouragement, scientific discussions, incredibly tasty food, good and bad jokes and of course, your love.

References

1. Kohler, R. E. *Lords of the Fly: Drosophila Genetics and the Experimental Life*. (University of Chicago Press, Chicago, IL, 1994).

2. Huxley, J. *Evolution: The Modern Synthesis*. (Allen & Unwin, London, 1942).

3. Falconer, D. S. & Mackay. *Introduction To Quantitative Genetics 4th Edition*. (1996).

4. Crick, F. H. On protein synthesis. *Symp Soc Exp Biol* **12**, 138–163 (1958).

5. Watson, J. D. *Molecular Biology of the Gene*. (Pearson Education, 2004).

6. Gasperini, M., Tome, J. M. & Shendure, J. Towards a comprehensive catalogue of validated and target-linked human enhancers. *Nat Rev Genet* **21**, 292–310 (2020).

7. Spitz, F. & Furlong, E. E. M. Transcription factors: from enhancer binding to developmental control. *Nat Rev Genet* **13**, 613–626 (2012).

8. Lettice, L. A. *et al.* A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. *Human Molecular Genetics* **12**, 1725–1735 (2003).

9. Dawson, S. J., Morris, P. J. & Latchman, D. S. A Single Amino Acid Change Converts an Inhibitory Transcription Factor into an Activator (∗). *Journal of Biological Chemistry* **271**, 11631–11633 (1996).

10. Landini, A. *et al.* Genetic regulation of post-translational modification of two distinct proteins. *Nat Commun* **13**, 1586 (2022).

11. Siddiqui-Jain, A., Grand, C. L., Bearss, D. J. & Hurley, L. H. Direct evidence for a G-quadruplex in a promoter region and its targeting with a small molecule to repress c-MYC transcription. *Proceedings of the National Academy of Sciences* **99**, 11593–11598 (2002).

12. Vance, K. W. & Ponting, C. P. Transcriptional regulatory functions of nuclear long noncoding RNAs. *Trends Genet* **30**, 348–355 (2014).

13. Workman, J. L. Nucleosome displacement in transcription. *Genes Dev.* **20**, 2009–2017 (2006).

14. Iwafuchi-Doi, M. & Zaret, K. S. Cell fate control by pioneer transcription factors. *Development* **143**, 1833–1837 (2016).

15. Lambert, S. A. *et al.* The Human Transcription Factors. *Cell* **172**, 650–665 (2018).

16. Kim, S. & Wysocka, J. Deciphering the multi-scale, quantitative *cis*-regulatory code. *Molecular Cell* **83**, 373–392 (2023).

17. Jindal, G. A. & Farley, E. K. Enhancer grammar in development, evolution, and disease: dependencies and interplay. *Developmental Cell* **56**, 575–587 (2021).

18. Kvon, E. Z. Using transgenic reporter assays to functionally characterize enhancers in animals. *Genomics* **106**, 185–192 (2015).

19. Farley, E. K. *et al.* Suboptimization of developmental enhancers. *Science* **350**, 325–328 (2015).

20. Farley, E. K. *et al.* Suboptimization of developmental enhancers. *Science* **350**, 325–328 (2015).

21. Clapier, C. R., Iwasa, J., Cairns, B. R. & Peterson, C. L. Mechanisms of action and regulation of ATP-dependent chromatin-remodelling complexes. *Nat Rev Mol Cell Biol* **18**, 407–422 (2017).

22. Morgan, M. A. J. & Shilatifard, A. Reevaluating the roles of histone-modifying enzymes and their associated chromatin modifications in transcriptional regulation. *Nat Genet* **52**, 1271–1281 (2020).

23. Richter, W. F., Nayak, S., Iwasa, J. & Taatjes, D. J. The Mediator complex as a master regulator of transcription by RNA polymerase II. *Nat Rev Mol Cell Biol* **23**, 732–749 (2022).

24. Matkar, S., Thiel, A. & Hua, X. Menin: a scaffold protein that controls gene expression and cell signaling. *Trends in Biochemical Sciences* **38**, 394–402 (2013).

25. Li, B., Carey, M. & Workman, J. L. The Role of Chromatin during Transcription. *Cell* **128**, 707–719 (2007).

26. Stampfel, G. *et al.* Transcriptional regulators form diverse groups with context-dependent regulatory functions. *Nature* **528**, 147–151 (2015).

27. Dixon, J. R. *et al.* Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**, 376–380 (2012).

28. Nora, E. P. *et al.* Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature* **485**, 381–385 (2012).

29. Benayoun, B. A. & Veitia, R. A. A post-translational modification code for transcription factors: sorting through a sea of signals. *Trends in Cell Biology* **19**, 189–197 (2009).

30. Zaret, K. S. & Carroll, J. S. Pioneer transcription factors: establishing competence for gene expression. *Genes Dev.* **25**, 2227–2241 (2011).

31. Chivu, A. G. *et al.* Evolution of promoter-proximal pausing enabled a new layer of transcription control. 2023.02.19.529146 Preprint at https://doi.org/10.1101/2023.02.19.529146 (2023).

32. Kornblihtt, A. R. *et al.* Alternative splicing: a pivotal step between eukaryotic transcription and translation. *Nat Rev Mol Cell Biol* **14**, 153–165 (2013).

33. Hargrove, J. L. & Schmidt, F. H. The role of mRNA and protein stability in gene expression. *The FASEB Journal* **3**, 2360–2370 (1989).

34. Catalanotto, C., Cogoni, C. & Zardo, G. MicroRNA in Control of Gene Expression: An Overview of Nuclear Functions. *Int J Mol Sci* **17**, 1712 (2016).

35. Montavon, T. *et al.* A Regulatory Archipelago Controls *Hox* Genes Transcription in Digits. *Cell* **147**, 1132–1145 (2011).

36. Kwak, H., Fuda, N. J., Core, L. J. & Lis, J. T. Precise Maps of RNA Polymerase Reveal How Promoters Direct Initiation and Pausing. *Science* **339**, 950–953 (2013).

37. Wang, Z.-Y. *et al.* Transcriptome and translatome co-evolution in mammals. *Nature* **588**, 642–647 (2020).

38. Gross, D. S. & Garrard, W. T. Nuclease hypersensitive sites in chromatin. *Annu Rev Biochem* **57**, 159–197 (1988).

39. Johnson, S. M., Tan, F. J., McCullough, H. L., Riordan, D. P. & Fire, A. Z. Flexibility and constraint in the nucleosome core landscape of Caenorhabditis elegans chromatin. *Genome Res* **16**, 1505–1516 (2006).

40. Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat Methods* **10**, 1213–1218 (2013).

41. Prescott, S. L. *et al.* Enhancer divergence and cis-regulatory evolution in the human and chimp neural crest. *Cell* **163**, 68–83 (2015).

42. Liu, C. *et al.* An ATAC-seq atlas of chromatin accessibility in mouse tissues. *Sci Data* **6**, 65 (2019).

43. Bozek, M. *et al.* ATAC-seq reveals regional differences in enhancer accessibility during the establishment of spatial coordinates in the Drosophila blastoderm. *Genome Res.* **29**, 771–783 (2019).

44. Daugherty, A. C. *et al.* Chromatin accessibility dynamics reveal novel functional enhancers in C. elegans. *Genome Res* **27**, 2096–2107 (2017).

45. Heintzman, N. D. *et al.* Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat Genet* **39**, 311–318 (2007).

46. Stefflova, K. *et al.* Cooperativity and Rapid Evolution of Cobound Transcription Factors in Closely Related Mammals. *Cell* **154**, 530–540 (2013).

47. Dekker, J., Marti-Renom, M. A. & Mirny, L. A. Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. *Nat Rev Genet* **14**, 390–403 (2013).

48. Hughes, J. R. *et al.* Analysis of hundreds of cis-regulatory landscapes at high resolution in a single, high-throughput experiment. *Nat Genet* **46**, 205–212 (2014).

49. Lieberman-Aiden, E. *et al.* Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome. *Science* **326**, 289–293 (2009).

50. Lonfat, N. & Duboule, D. Structure, function and evolution of topologically associating domains (TADs) at HOX loci. *FEBS Lett* **589**, 2869–2876 (2015).

51. Okhovat, M. *et al.* TAD evolutionary and functional characterization reveals diversity in mammalian TAD boundary properties and function. *Nat Commun* **14**, 8111 (2023).

52. Canver, M. C. *et al.* BCL11A enhancer dissection by Cas9-mediated in situ saturating mutagenesis. *Nature* **527**, 192–197 (2015).

53. Trapnell, C. Defining cell types and states with single-cell genomics. *Genome Res* **25**, 1491–1498 (2015).

54. Pinho, S. & Frenette, P. S. Haematopoietic stem cell activity and interactions with the niche. *Nat Rev Mol Cell Biol* **20**, 303–320 (2019).

55. Shetty, S., Lalor, P. F. & Adams, D. H. Liver sinusoidal endothelial cells — gatekeepers of hepatic immunity. *Nat Rev Gastroenterol Hepatol* **15**, 555–567 (2018).

56. Pijuan-Sala, B. *et al.* A single-cell molecular map of mouse gastrulation and early organogenesis. *Nature* **566**, 490–495 (2019).

57. Kashima, Y. *et al.* Single-cell sequencing techniques from individual to multiomics analyses. *Exp Mol Med* **52**, 1419–1427 (2020).

58. Cusanovich, D. A. *et al.* A Single-Cell Atlas of In Vivo Mammalian Chromatin Accessibility. *Cell* **174**, 1309-1324.e18 (2018).

59. Martin, B. K. *et al.* Optimized single-nucleus transcriptional profiling by combinatorial indexing. *Nat Protoc* **18**, 188–207 (2023).

60. Saunders, L. M. *et al.* Embryo-scale reverse genetics at single-cell resolution. *Nature* **623**, 782–791 (2023).

61. Huang, X. *et al.* Single-cell, whole-embryo phenotyping of mammalian developmental disorders. *Nature* **623**, 772–781 (2023).

62. Buenrostro, J. D. *et al.* Integrated Single-Cell Analysis Maps the Continuous Regulatory Landscape of Human Hematopoietic Differentiation. *Cell* **173**, 1535-1548.e16 (2018).

63. Rosenberg, A. B. *et al.* Single-cell profiling of the developing mouse brain and spinal cord with split-pool barcoding. *Science* **360**, 176–182 (2018).

64. Wu, F. *et al.* Single-cell profiling of tumor heterogeneity and the microenvironment in advanced non-small cell lung cancer. *Nat Commun* **12**, 2540 (2021).

65. Vandereyken, K., Sifrim, A., Thienpont, B. & Voet, T. Methods and applications for single-cell and spatial multi-omics. *Nat Rev Genet* **24**, 494–515 (2023).

66. Zhu, C. *et al.* Joint profiling of histone modifications and transcriptome in single cells from mouse brain. *Nat Methods* **18**, 283–292 (2021).

67. Chen, S., Lake, B. B. & Zhang, K. High-throughput sequencing of the transcriptome and chromatin accessibility in the same cell. *Nat Biotechnol* **37**, 1452–1457 (2019).

68. Cao, J. *et al.* Joint profiling of chromatin accessibility and gene expression in thousands of single cells. *Science* **361**, 1380–1385 (2018).

69. Ma, S. *et al.* Chromatin Potential Identified by Shared Single-Cell Profiling of RNA and Chromatin. *Cell* **183**, 1103-1116.e20 (2020).

70. Spielmann, M., Lupiáñez, D. G. & Mundlos, S. Structural variation in the 3D genome. *Nat Rev Genet* **19**, 453–467 (2018).

71. Lupiáñez, D. G. *et al.* Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. *Cell* **161**, 1012–1025 (2015).

72. Wittkopp, P. J., Haerum, B. K. & Clark, A. G. Evolutionary changes in cis and trans gene regulation. *Nature* **430**, 85–88 (2004).

73. Wang, X., Werren, J. H. & Clark, A. G. Allele-Specific Transcriptome and Methylome Analysis Reveals Stable Inheritance and Cis-Regulation of DNA Methylation in Nasonia. *PLOS Biology* **14**, e1002500 (2016).

74. Coolon, J. D., McManus, C. J., Stevenson, K. R., Graveley, B. R. & Wittkopp, P. J. Tempo and mode of regulatory evolution in Drosophila. *Genome Res.* **24**, 797–808 (2014).

75. Verta, J.-P. & Jones, F. C. Predominance of cis-regulatory changes in parallel expression divergence of sticklebacks. *eLife* **8**, e43785 (2019).

76. Shi, X. *et al.* Cis- and trans-regulatory divergence between progenitor species determines gene-expression novelty in Arabidopsis allopolyploids. *Nat Commun* **3**, 950 (2012).

77. Nica, A. C. & Dermitzakis, E. T. Expression quantitative trait loci: present and future. *Philos Trans R Soc Lond B Biol Sci* **368**, 20120362 (2013).

78. Signor, S. A. & Nuzhdin, S. V. The Evolution of Gene Expression in cis and trans. *Trends in Genetics* **34**, 532–544 (2018).

79. Fisher, R. A. *The Genetical Theory of Natural Selection*. xiv, 272 (Clarendon Press, Oxford, England, 1930). doi:10.5962/bhl.title.27468.

80. Darwin, C. *On the Origin of Species*. (J. Murray, London, 1859).

81. Stern, D. L. Perspective: Evolutionary Developmental Biology and the Problem of Variation. *Evolution* **54**, 1079–1091 (2000).

82. Carroll, S. B. Evo-Devo and an Expanding Evolutionary Synthesis: A Genetic Theory of Morphological Evolution. *Cell* **134**, 25–36 (2008).

83. Zuckerkandl, E. & Pauling, L. Evolutionary Divergence and Convergence in Proteins. in *Evolving Genes and Proteins* (eds. Bryson, V. & Vogel, H. J.) 97–166 (Academic Press, 1965). doi:10.1016/B978-1-4832-2734-4.50017-6.

84. Jacob, F. Evolution and Tinkering. *Science* **196**, 1161–1166 (1977).

85. King, M. C. & Wilson, A. C. Evolution at two levels in humans and chimpanzees. *Science* **188**, 107–116 (1975).

86. Scott, M. P. & Weiner, A. J. Structural relationships among genes that control development: sequence homology between the Antennapedia, Ultrabithorax, and fushi tarazu loci of Drosophila. *Proceedings of the National Academy of Sciences* **81**, 4115–4119 (1984).

87. McGinnis, W., Garber, R. L., Wirz, J., Kuroiwa, A. & Gehring, W. J. A homologous protein-coding sequence in drosophila homeotic genes and its conservation in other metazoans. *Cell* **37**, 403–408 (1984).

88. Stent, G. S. From probability to molecular biology: From egg to embryo. By J. M. W. Slack. New York: Cambridge University Press. (1983). 241 pp. $49.50. *Cell* **36**, 567–570 (1984).

89. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).

90. Gehring, W. J. & Ikeo, K. Pax 6: mastering eye morphogenesis and eye evolution. *Trends in Genetics* **15**, 371–377 (1999).

91. Halder, G., Callaerts, P. & Gehring, W. J. Induction of Ectopic Eyes by Targeted Expression of the eyeless Gene in Drosophila. *Science* **267**, 1788–1792 (1995).

92. Adachi, Y. *et al.* Conserved cis-regulatory modules mediate complex neural expression patterns of the eyeless gene in the Drosophila brain. *Mechanisms of Development* **120**, 1113–1126 (2003).

93. Wray, G. A. *et al.* The Evolution of Transcriptional Regulation in Eukaryotes. *Mol Biol Evol* **20**, 1377–1419 (2003).

94. Lee, T. I. & Young, R. A. Transcriptional Regulation and its Misregulation in Disease. *Cell* **152**, 1237–1251 (2013).

95. He, X. & Zhang, J. Why Do Hubs Tend to Be Essential in Protein Networks? *PLOS Genetics* **2**, e88 (2006).

96. Kusserow, A. *et al.* Unexpected complexity of the Wnt gene family in a sea anemone. *Nature* **433**, 156–160 (2005).

97. Mikkelsen, T. S. *et al.* Genome of the marsupial Monodelphis domestica reveals innovation in non-coding sequences. *Nature* **447**, 167–177 (2007).

98. Aparicio, S. *et al.* Whole-Genome Shotgun Assembly and Analysis of the Genome of Fugu rubripes. *Science* **297**, 1301–1310 (2002).

99. Stern, D. L. & Orgogozo, V. The Loci of Evolution: How Predictable Is Genetic Evolution? *Evolution* **62**, 2155–2177 (2008).

100. Gruber, J. D., Vogel, K., Kalay, G. & Wittkopp, P. J. Contrasting Properties of Gene-Specific Regulatory, Coding, and Copy Number Mutations in Saccharomyces cerevisiae: Frequency, Effects, and Dominance. *PLOS Genetics* **8**, e1002497 (2012).

101. Metzger, B. P. H. *et al.* Contrasting Frequencies and Effects of cis- and trans-Regulatory Mutations Affecting Gene Expression. *Molecular Biology and Evolution* **33**, 1131–1146 (2016).

102. Dujon, B. The yeast genome project: what did we learn? *Trends in Genetics* **12**, 263–270 (1996).

103. Hoekstra, H. E. & Coyne, J. A. The Locus of Evolution: Evo Devo and the Genetics of Adaptation. *Evolution* **61**, 995–1016 (2007).

104. Kvon, E. Z. *et al.* Progressive Loss of Function in a Limb Enhancer During Snake Evolution. *Cell* **167**, 633-642.e11 (2016).

105. Shapiro, M. D. *et al.* Genetic and developmental basis of evolutionary pelvic reduction in threespine sticklebacks. *Nature* **428**, 717–723 (2004).

106. Chan, Y. F. *et al.* Adaptive Evolution of Pelvic Reduction in Sticklebacks by Recurrent Deletion of a Pitx1 Enhancer. *Science* **327**, 302–305 (2010).

107. Jones, F. C. *et al.* The genomic basis of adaptive evolution in threespine sticklebacks. *Nature* **484**, 55–61 (2012).

108. Wittkopp, P. J., Haerum, B. K. & Clark, A. G. Regulatory changes underlying expression differences within and between Drosophila species. *Nat Genet* **40**, 346–350 (2008).

109. Tirosh, I., Reikhav, S., Levy, A. A. & Barkai, N. A Yeast Hybrid Provides Insight into the Evolution of Gene Expression Regulation. *Science* **324**, 659–662 (2009).

110. Goncalves, A. *et al.* Extensive compensatory cis-trans regulation in the evolution of mouse gene expression. *Genome Res.* **22**, 2376–2384 (2012).

111. Metzger, B. P. H., Wittkopp, P. J. & Coolon, Joseph. D. Evolutionary Dynamics of Regulatory Changes Underlying Gene Expression Divergence among Saccharomyces Species. *Genome Biology and Evolution* **9**, 843–854 (2017).

112. Liu, X., Li, Y. I. & Pritchard, J. K. Trans Effects on Gene Expression Can Drive Omnigenic Inheritance. *Cell* **177**, 1022-1034.e6 (2019).

113. Rhoné, B. *et al.* No excess of *cis* -regulatory variation associated with intra-specific selection in wild pearl millet (*Cenchrus americanus*). *Genome Biol Evol* evx004 (2017) doi:10.1093/gbe/evx004.

114. Chen, J., Nolte, V. & Schlötterer, C. Temperature Stress Mediates Decanalization and Dominance of Gene Expression in Drosophila melanogaster. *PLOS Genetics* **11**, e1004883 (2015).

115. Agoglia, R. M. *et al.* Primate cell fusion disentangles gene regulatory divergence in neurodevelopment. *Nature* **592**, 421–427 (2021).

116. Boyle, E. A., Li, Y. I. & Pritchard, J. K. An expanded view of complex traits: from polygenic to omnigenic. *Cell* **169**, 1177–1186 (2017).

117. Loh, P.-R. *et al.* Contrasting genetic architectures of schizophrenia and other complex diseases using fast variance-components analysis. *Nat Genet* **47**, 1385–1392 (2015).

118. International Schizophrenia Consortium *et al.* Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* **460**, 748–752 (2009).

119. White, M. A. Fine-Scale Phylogenetic Discordance across the House Mouse Genome. *PLoS Genetics* **5**, (2009).

120. Keane, T. M. *et al.* Mouse genomic variation and its effect on phenotypes and gene regulation. *Nature* **477**, 289–294 (2011).

121. Boursot, P., Auffray, J.-C., Britton-Davidian, J. & Bonhomme, F. The Evolution of House Mice. *Annual Review of Ecology, Evolution and Systematics* **24**, 119–152 (1993).

122. Din, W. *et al.* Origin and radiation of the house mouse: clues from nuclear genes. *Journal of Evolutionary Biology* **9**, 519–539 (1996).

123. Phifer-Rixey, M. & Nachman, M. W. Insights into mammalian biology from the wild house mouse Mus musculus. *eLife* **4**, e05959 (2015).

124. Duveau, F. *et al.* Mutational sources of trans-regulatory variation affecting gene expression in Saccharomyces cerevisiae. *eLife* **10**, e67806 (2021).

125. Didion, J. P. & de Villena, F. P.-M. Deconstructing Mus gemischus: advances in understanding ancestry, structure, and variation in the genome of the laboratory mouse. *Mamm Genome* **24**, 1–20 (2013).

126. Suzuki, H., Shimada, T., Terashima, M., Tsuchiya, K. & Aplin, K. Temporal, spatial, and ecological modes of evolution of Eurasian *Mus* based on mitochondrial and nuclear gene sequences. *Molecular Phylogenetics and Evolution* **33**, 626–646 (2004).

127. Kraft, R. & Kraft, R. Merkmale und Verbreitung der Hausmause Mus musculus musculus L., 1758 und Mus musculus domesticus Rutty, 1772 (Rodentia, Muridae) in Bayern. *Säugetierkundliche Mitteilungen* **32**, 1–12 (1985).

128. Ďureje, Ľ., Macholán, M., Baird, S. J. E. & Piálek, J. The mouse hybrid zone in Central Europe: from morphology to molecules. *fozo* **61**, 308–318 (2012).

129. Lawal, R. A. *et al.* Taxonomic assessment of two wild house mouse subspecies using whole-genome sequencing. *Sci Rep* **12**, 20866 (2022).

130. Yang, H. *et al.* Subspecific origin and haplotype diversity in the laboratory mouse. *Nat Genet* **43**, 648–655 (2011).

131. She, J. X., Bonhomme, F., Boursot, P., Thaler, L. & Catzeflis, F. Molecular phylogenies in the genus Mus: Comparative analysis of electrophoretic, scnDNA hybridization, and mtDNA RFLP data. *Biol J Linn Soc* **41**, 83–103 (1990).

132. Mammifères), S. A. (Société F. pour l' É. et la P. des. IUCN Red List of Threatened Species: Mus spretus. *IUCN Red List of Threatened Species* (2016).

133. Silver, L. M. *Mouse Genetics: Concepts and Applications*. (Oxford University Press, 1995).

134. Beck, J. A. *et al.* Genealogies of mouse inbred strains. *Nat Genet* **24**, 23–25 (2000).

135. Frazer, K. A. *et al.* A sequence-based variation map of 8.27 million SNPs in inbred mouse strains. *Nature* **448**, 1050–1053 (2007).

136. Guénet, J.-L. & Bonhomme, F. Wild mice: an ever-increasing contribution to a popular mammalian model. *Trends in Genetics* **19**, 24–31 (2003).

137. Gregorová, S. & Forejt, J. PWD/Ph and PWK/Ph Inbred Mouse Strains of Mus т. musculus Subspecies-a Valuable Resource of Phenotypic Variations and Genomic Polymorphisms. *Folia biologica* **46**, 31–41 (2000).

138. Mashimo, T. *et al.* A nonsense mutation in the gene encoding 2′-5′-oligoadenylate synthetase/L1 isoform is associated with West Nile virus susceptibility in laboratory mice. *Proceedings of the National Academy of Sciences* **99**, 11311–11316 (2002).

139. Robert, B. *et al.* Investigation of genetic linkage between myosin and actin genes using an interspecific mouse back-cross. *Nature* **314**, 181–183 (1985).

140. Crowley, J. J. *et al.* Analyses of allele-specific gene expression in highly divergent mouse crosses identifies pervasive allelic imbalance. *Nat Genet* **47**, 353–360 (2015).

141. Panten, J. *et al.* The dynamic genetic determinants of increased transcriptional divergence in spermatids. *Nat Commun* **15**, 1272 (2024).

142. Hochepied, T. *et al.* Breaking the Species Barrier: Derivation of Germline-Competent Embryonic Stem Cells from *Mus spretus* × C57BL/6 Hybrids. *STEM CELLS* **22**, 441–447 (2004).

143. Mack, K. L., Campbell, P. & Nachman, M. W. Gene regulation and speciation in house mice. *Genome Res.* **26**, 451–461 (2016).

144. Soltys, V., Peters, M., Su, D., Kučka, M. & Chan, Y. F. Flexible and high-throughput simultaneous profiling of gene expression and chromatin accessibility in single cells. 2024.02.26.581705 Preprint at https://doi.org/10.1101/2024.02.26.581705 (2024).

145. Song, J. H. T. *et al.* Genetic studies of human–chimpanzee divergence using stem cell fusions. *Proceedings of the National Academy of Sciences* **118**, e2117557118 (2021).

146. Zhang, Y. *et al.* Single-cell RNA sequencing in cancer research. *Journal of Experimental & Clinical Cancer Research* **40**, 81 (2021).

147. Van de Sande, B. *et al.* Applications of single-cell RNA sequencing in drug discovery and development. *Nat Rev Drug Discov* **22**, 496–520 (2023).

148. Schaefke, B. *et al.* Inheritance of Gene Expression Level and Selective Constraints on Trans- and Cis-Regulatory Changes in Yeast. *Molecular Biology and Evolution* **30**, 2121–2133 (2013).

149. Lemos, B., Araripe, L. O., Fontanillas, P. & Hartl, D. L. Dominance and the evolutionary accumulation of cis- and trans-effects on gene expression. *Proceedings of the National Academy of Sciences* **105**, 14471–14476 (2008).

150. Coolon, J. D. *et al.* Molecular Mechanisms and Evolutionary Processes Contributing to Accelerated Divergence of Gene Expression on the Drosophila X Chromosome. *Molecular Biology and Evolution* **32**, 2605–2615 (2015).

151. Emerson, J. J. *et al.* Natural selection on cis and trans regulation in yeasts. *Genome Res.* **20**, 826–836 (2010).

152. Shabalina, S. A. & Spiridonov, N. A. The mammalian transcriptome and the function of non-coding DNA sequences. *Genome Biology* (2004).

153. Fraser, H. B. Improving Estimates of Compensatory *cis–trans* Regulatory Divergence. *Trends in Genetics* **35**, 3–5 (2019).

154. Brawand, D. *et al.* The evolution of gene expression levels in mammalian organs. *Nature* **478**, 343–348 (2011).

155. Murat, F. *et al.* The molecular evolution of spermatogenesis across mammals. *Nature* **613**, 308–316 (2023).

156. Kopania, E. E. K., Larson, E. L., Callahan, C., Keeble, S. & Good, J. M. Molecular Evolution across Mouse Spermatogenesis. *Molecular Biology and Evolution* **39**, msac023 (2022).

157. Cuomo, A. S. E. *et al.* Single-cell RNA-sequencing of differentiating iPS cells reveals dynamic genetic effects on gene expression. *Nat Commun* **11**, 810 (2020).

158. Nguyen-Lefebvre, A. T. & Horuzsko, A. Kupffer Cell Metabolism and Function. *J Enzymol Metab* **1**, 101 (2015).

Appendix

Appendix I: Flexible and high-throughput simultaneous profiling of gene expression and chromatin accessibility in single cells

Flexible and high-throughput simultaneous profiling of gene expression and chromatin accessibility in single cells

Volker Soltys[1*], Moritz Peters[1], Dingwen Su[1], Marek Kučka[1,2,], Yingguang Frank Chan[1,3]

1 Friedrich Miescher Laboratory of the Max Planck Society, 72076 Tübingen, Germany
2 Department of Translational Genomics, University of Cologne, 50931 Cologne, Germany
3 University of Groningen, Groningen Institute of Evolutionary Life Sciences, 9747 AG Groningen, Netherlands

* Corresponding authors
volker.soltys@tue.mpg.de; frank.chan@rug.nl

Abstract

Gene expression and chromatin accessibility are highly interconnected processes. Disentangling one without the other provides an incomplete picture of gene regulation. However, simultaneous measurements of RNA and accessible chromatin are technically challenging, especially when studying complex organs with rare cell-types. Here, we present easySHARE-seq, an elaboration of SHARE-seq, providing simultaneous measurements of ATAC- and RNA-seq within single cells, enabling identification of cell-type specific *cis*-regulatory elements (CREs). easySHARE-seq retains high scalability, improves RNA-seq data quality while also allowing for flexible study design. Using 19,664 joint profiles from murine liver nuclei, we linked CREs to their target genes and uncovered complex regulation of key genes such as *Gata4*. We further identify *de novo* genes and *cis*-regulatory elements displaying zonation in Liver sinusoidal epithelial cells (LSECs), a challenging cell type with low mRNA levels, demonstrating the power of multimodal measurements. EasySHARE-seq therefore provides a flexible platform for investigating gene regulation across cell types and scale.

Introduction

Gene expression and chromatin state together influence fundamental processes such as gene regulation or cell fate decisions [1–3]. A better understanding of these mechanisms and their interactions will be a major step toward decoding developmental trajectories or reconstructing cellular taxonomies in both health and disease. However, to fully capture these complex relationships, multiple information layers need to be measured simultaneously. For example, prior studies have argued that chromatin state is often predictive of gene expression and can also prime cells toward certain lineage decisions or even induce tissue regeneration[4–6]. However, these studies depend on the computational integration of separately measured modalities. By assuming a shared biological state, this restricts the discovery of novel and potentially fine-scale differences and renders it challenging to identify the root cause of erroneous cell states[7].

The last decade has seen an explosive growth in single-cell methodologies, with new assays, increasing throughput and a suite of computational tools[8]. Most non-commercial high-throughput methodologies rely on combinatorial indexing for single-cell barcoding, where sequential rounds of barcodes combine to create unique cellular barcode combinations [9,10]. Compared to single-modality assays, multi-omic technologies, which capture two or more information layers, are relatively new. Therefore, they are still limited in sensitivity and throughput and commercial kits can be expensive such that multi-omic studies tend to have limited sample sizes [11,12].

To address these problems, we built upon the previously published protocol called SHARE-seq[13] and developed easySHARE-seq, a protocol for simultaneously measuring gene expression and chromatin accessibility within single cells using combinatorial indexing. Major improvements include easySHARE-seq's barcoding framework, which allows for expanded and flexible study design, all while being compatible with standard Illumina sequencing, thereby removing economic hurdles. Importantly, easySHARE-seq retains the scalability and improves upon RNA-seq sensitivity of the original SHARE-seq protocol. Here, we used easySHARE-seq to profile 19,664 murine liver nuclei and show that we can recover high quality data in both RNA-seq and ATAC-seq channels, which are highly congruent and share equal power in classifying cell types. We then surveyed the *cis*-regulatory landscape of Liver Sinusoidal Endothelial Cells (LSECs), leveraging the simultaneous measurements of gene expression and chromatin accessibility and identified 40,957 links between expressed genes and nearby ATAC-seq peaks. Notably, genes with the highest number of links were enriched for transcription factors and regulators known to control important functions within LSECs. Lastly, we show that easySHARE-seq can be used to investigate micro-scale changes in accessibility and gene expression by identifying novel markers and open chromatin regions displaying zonation in LSECs. This technology improves our toolkit of multi-omic protocols needed for advancing our knowledge about gene regulation and cell fate decisions.

Results

easySHARE-seq reliably labels both transcriptome and accessible chromatin in individual cells

To develop a multi-omic single-cell (sc) RNA and scATAC-seq protocol that allows for flexible study design while being highly scalable, we built upon SHARE-seq[13] to create easySHARE-seq, which uses two rounds of ligation to simultaneously label cDNA and DNA fragments in the same cell (**Fig. 1A**). Due to a much more streamlined barcoding structure, easySHARE-seq allows 300bp sequencing of the insert. This longer read-length leads to a higher recovery of DNA variants, thus increasing the power to detect allele-specific signals or cell-specific variation, e.g., in hybrids or cancer cells[14].

To generate libraries, fixed and permeabilized cells or nuclei (we will use "cells" afterwards to refer to both) are transposed by Tn5 transposase carrying a custom adapter with a single-stranded overhang (**Fig. 1B**). Next, mRNA is reverse transcribed (RT) using a biotinylated poly(T) primer with an identical overhang. Subsequently, the cells are individually barcoded in two rounds of combinatorial indexing with 192 barcodes in each round, creating a total of 36,864 possible barcode combinations.The first barcode is ligated onto the already present overhang and itself contains a second single-stranded overhang, onto which the second barcode can be ligated. Importantly, in the easySHARE-seq design, we have kept the total length of the barcode within 17nt ("Index 1" read; **Fig. 1B, Suppl. Fig. 1A**), allowing for multiplexing of easySHARE-seq libraries with standard Illumina libraries. In contrast, in the original publication, SHARE-seq libraries required Index 1 lengths of 99nt, a highly custom configuration which would require a costly private sequencing.

After barcoding, the cells are aliquoted into sub-libraries of approximately 3,500 cells each and reverse crosslinked. A streptavidin pull-down of the biotinylated RT-primer is performed to separate the cDNA molecules from the chromatin ("fragments"). Each sub-library is then prepared for sequencing and amplified using matched indexing primers to allow identification of paired cellular scRNA- and scATAC-seq profiles. By scaling up the numbers of sub-libraries, this barcoding strategy therefore allows for high-throughput experiments of hundreds of thousands of cells, only limited by the availability of indexing primers. For a detailed description of the flexibility of easySHARE-seq, instructions on how to modify and incorporate the framework into new designs as well as critical steps to assess when planning to use easySHARE-seq see Supplementary Notes.

To evaluate the accuracy and cell-specificity of the barcoding, we first performed easySHARE-seq on a mixed pool between human and murine cell lines (HEK and OP-9 respecitvely). This design allows us to identify two or more cells sharing the same barcode ('doublets'; **Fig. 1C**, left). After sequencing, we recovered a total of 3,808 cells. Both chromatin and transcriptome profiles separated well within each cell (**Fig. 1C**, middle), with cDNA showing a lower accuracy with increasing transcript counts, likely due to less precise read mapping. We identified a total of 124 doublets (**Fig, 1C**, right), which gives a final doublet rate of 6.34% factoring in the undetectable intra-species doublets. For comparison, a 10X Chromium Next GEM experiment with 10,000 cells has a doublet rate of ~7.9% (www.10xgenomics.com). Importantly, easySHARE-seq doublet rates can be lowered further by aliquoting fewer cells within each sub-library. To summarise, easySHARE-seq provides a high-throughput and flexibility framework for accurately measuring chromatin accessibility and gene expression in single cells.

Simultaneous scATAC-seq and scRNA-seq profiling in murine primary liver cells

To assess data quality and investigate the relationship between gene expression and chromatin accessibility, we focused on murine liver. The liver consists of a diverse set of defined primary cell types, ranging from large and potentially multinucleated hepatocytes to small non-parenchymal cell types such as Liver Sinusoidal Endothelial Cells[15] (LSECs).

We generated matched high-quality chromatin and gene expression profiles for 19,664 adult liver cells across four age-matched mice (2 male, 2 female), amounting to a recovery rate of 70.2% (28,000 input cells). Each nuclei had on average 3,629 UMIs and 2,213 fragments (74% of all RNA-seq reads were cDNA, 55.9% mean ATAC-seq fragments in peaks; **Suppl. Fig. 1B & D**). In terms of UMIs per cell, easySHARE-seq therefore out-performed other previously published multi-omic and representative single channel assays (**Fig. 2B**; see figure legend for tissue type and study). Consistent with nuclei as input material, the majority of cDNA molecules were intronic (69.6%, **Suppl. Fig. 1C & H**). Regarding DNA fragments per cell, easySHARE-seq performed similarly to other published multi-omic assays (**Fig. 2C**) and scATAC-seq libraries displayed the characteristic banding pattern with reads being highly enriched at transcription start sites (TSS; **Suppl. Fig. 1 E, F, H**).

To visualise and identify cell types, we first projected the ATAC- and RNA-seq modalities separately into 2D Space and then used Weighted Nearest Neighbor[16] (WNN) integration to combine both modalities into a single UMAP visualisation (**Fig. 2A)**. Importantly, the same cells independently clustered together in the scRNA- and scATAC-seq UMAPs, showcasing high congruence between the two modalities (**Suppl. Fig. 2A&B**). We then annotated previously published cell types based on gene expression of previously established marker genes [17,18]. Marker gene expression was highly specific to the clusters (**Fig. 2D, Suppl. Fig. 2F**) and we recovered all expected cell types (**Suppl. Fig.2C**). Importantly, the same cell types were identified using each modality independently, showcasing high congruence between the scATAC- and scRNA-seq modalities (**Fig. 2E**). Altogether, our results show that easySHARE-seq generates high quality joint cellular profiles of chromatin accessibility and gene expression within primary tissue, expanding our toolkit of multi-omic protocols.

Uncovering the cis-regulatory landscape of key regulators through peak-gene associations

As easySHARE-seq simultaneously measures chromatin accessibility and gene expression, it allows to direct investigation of the relationship between them to potentially connect *cis*-regulatory elements (CREs) to their target genes. To do so, we adopted the analytical framework from Ma et al.[13], which queries if an increased expression within a cell is significantly correlated with chromatin accessibility at a peak while controlling for GC content and accessibility strength. Focusing on LSECs (1,501 cells), we calculated associations between putative CREs (pCREs, defined as peaks with a significant peak–gene association) and each expressed gene, considering all peaks within ± 500kb of the TSS. We identified 40,957 significant peak–gene associations (45% of total peaks, $P < 0.05$, FDR = 0.1) with 15,061 genes having at least one association (76.8% of all expressed genes, **Suppl. Fig. 3A,C**). Conversely, some rare pCREs (2.9%) were associated with five or more genes (0.03% when considering only pCREs within ± 50 kb of a TSS (**Suppl. Fig. 3B,D**)). These pCREs tended to cluster to regions of higher expressed gene density (2.15 mean expressed genes within 50kbp vs 0.93 for all global peaks) and their associated genes were enriched for biological processes such as mRNA processing, histone modifications and splicing (**Suppl. Fig. 3H**), possibly reflecting loci with increased regulatory activity.

Focusing on genes, we ranked them based on their number of associated pCREs (**Fig. 2F**). Within the top 1% genes with the most pCRE associations were many key regulators and transcription factors. Examples include *Taf5,* which directly binds the TATA-box[19] and is required for initiation of transcription, or *Gata4,* which has been identified as the master regulator for LSEC specification during development as well as controlling regeneration and metabolic maturation of liver tissue in adult mice [20,21]. As such, it incorporates a variety of signals and its expression needs to be strictly regulated, which is reflected in its many pCREs associations (**Fig. 2H**). Similarly, *Igf1* also integrates signals from many different pCREs[22] (**Suppl. Fig. 3G**). Notably, pCRES are significantly enriched at transcription start sites (TSS), even relative to background enrichment (**Fig. 2G**).
To summarise, easySHARE-seq allows the direct investigation of the relationship between chromatin accessibility and gene expression and identify putative *cis*-regulatory elements at genomic scale, even in small cell types with relatively low mRNA contents (**Suppl. Fig. 2D)**.

De novo identification of open chromatin regions and genes displaying zonation in LSECs

We next investigated the process of zonation in LSECs. The liver consists of hexagonal units called lobules where blood flows from the portal vein and arteries toward a central vein [23,24] (**Fig. 3A)**. The central–portal (CP) axis is characterised by a morphogen gradient, e.g. *Wnt2*, secreted by central vein LSECs, with the resulting micro-environment giving rise to spatial division of labour among hepatocytes [25–27]. Studying zonation in non-parenchymal cells such as LSECs is challenging as these are small cells with low mRNA content (**Suppl. Fig. 2D,E**), lying below the detection limit of current spatial transcriptomic techniques. As a result, only very few studies assess zonation in LSECs on a genomic level [28]. However, LSECs are critical to liver function as they line the artery walls, clear and process endotoxins, play a critical role in liver regeneration and secrete morphogens themselves to regulate hepatocyte gene expression [29–31], rendering their understanding a prerequisite for tackling many diseases.
We therefore asked if we can recover known zonation gradients and potentially identify novel marker genes and open chromatin regions displaying zonation. We noticed that LSECs clustered in a distinct linear pattern in our UMAP projection and therefore divided them into equal bins along UMAP2 coordinates (**Suppl. Fig. 4A**, number of cells per bin 80-260, median: 128). We then calculated mean normalised expression and mean normalised accessibility within each bin. This recovered gene expression and chromatin accessibility gradients for major known zonation marker genes[28] (**Fig. 3B,C**). For example, *Wnt2* expression decreased strongly along the CP axis as did chromatin accessibility of all three peaks at the *Wnt2* locus (**Fig. 3B**). We also recovered the zonation profiles for the majority of known pericentral (increasing along the CP-axis), periportal (decrease along the CP-axis) and non-monotonic markers (decrease toward both ends) as well as their associated chromatin regions (**Fig. 3C**). Gene expression zonation profiles can also be recovered by ordering LSECs along pseudotime (**Suppl. Fig. 4C,D**). In contrast, simply subclustering LSECs and comparing expression between these clusters was too broad for the assessment of zonation (**Suppl. Fig. 4A,B**).
Next, we sought to identify novel marker genes and open chromatin regions displaying zonation in LSECs based on the decrease or increase of mean expression or accessibility along the previously established bins. In total, we classified 153 genes and 381 open chromatin regions as pericentral and 209 genes and 465 open chromatin regions showed periportal zonation profiles (**Fig. 3D**). The list of markers contained many genes regulating epithelial growth and angiogenesis (e.g. *Efna1, Nrg2, Zfpm1, Zfpm2, Bmpr2*)[32–34], related to

regulating hepatocyte functions and communication (e.g. *Dll4, Foxo1, Sp1, Snx3*)[35–37] as well as immunological functions (e.g. *Sirt2, Cd59a)*[38,39]*,* suggesting that these processes show variation along the PC axis. As dysregulation of LSEC zonation is implicated in multiple illnesses such as liver cirrhosis or non-alcoholic fatty liver disease [40,41], these genes are potential new biomarkers for their identification and the open chromatin regions starting points for investigating the role of gene regulation in their emergence.

Discussion

Understanding complex processes such as gene regulation or disease states requires the integration of multiple layers of information. Here, we show that easySHARE-seq provides a high-quality, high-throughput and flexible platform for joint profiling of chromatin accessibility and gene expression within single cells. We show that both modalities are highly congruent with one another and we leverage their simultaneous measurements to identify peak–gene interactions and survey the *cis*-regulatory landscape of LSECs. We also show that easySHARE-seq can be used to assess micro-scale changes such as zonation in LSECs across both gene expression and chromatin accessibility. These cells have low mRNA content and we recovered zonation profiles of many transcription factors, which are often lowly expressed, further demonstrating the power of easySHARE-seq.

Besides improving upon RNA-seq data quality, we argue that easySHARE-seq has many advantages, especially in terms of the sequencing flexibility due to the barcode design, which can help remove hurdles for incorporating multi-omic single-cell assays into study designs. Combined with shorter experimental times (~12h total), easySHARE-seq might be particularly suited for studies where higher sample sizes are required or ones that rely on identification of genomic variants, e.g., in diverse, non-inbred individuals or in cancer. In terms of costs per cell, easySHARE-seq performs similarly to standard SHARE-seq with ~0.056 cents/cell, a fraction of the costs (<25%) of commercially available platforms, even before factoring in the specialized instrument costs. A comparison between technologies can be found in **Table 1**.

We envision easySHARE-seq as another technological step toward ultimately understanding gene regulation in health and disease, surveying *cis*-regulatory landscapes during differentiation and lineage commitment and determining genetic variants affecting those processes.

Figure 1

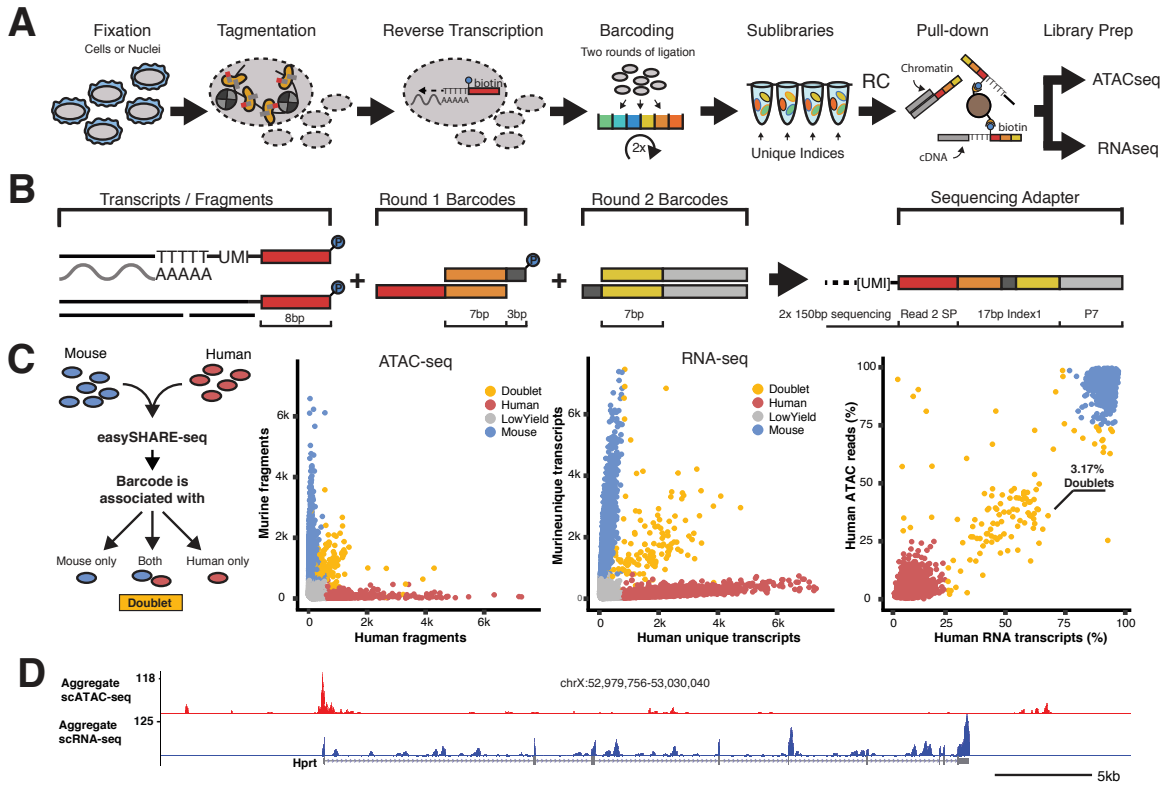


Figure 1: easySHARE-seq enables highly-accurate simultaneous scATAC-seq and scRNA-seq profiling

(A) Schematic workflow of easySHARE-seq.
(B) Generation and structure of the single-cell barcoding within Index 1.
(C) Principle of a species-mixing experiment. Cells are mixed prior to easySHARE-seq and sequences associated with each cell barcode are assessed for genome of origin (left panel). Unique ATAC fragments per cell aligning to the mouse or human genome (middle left). Cells are coloured according to their assigned origin (red: human; blue: mouse; orange: doublet). Middle right: Same plot but with RNA UMIs. Right: Percentage of ATAC fragments or RNA UMIs per cell relative to total sequencing reads mapping uniquely to the human genome. 3.17% of all observed cells classified as doublets. Accounting for same-species doublets, this results in a doublet rate of 6.34%.
(D) Aggregate chromatin accessibility (red) and expression-seq (blue) profile of OP-9 cells at the *Hprt* locus.

Figure 2

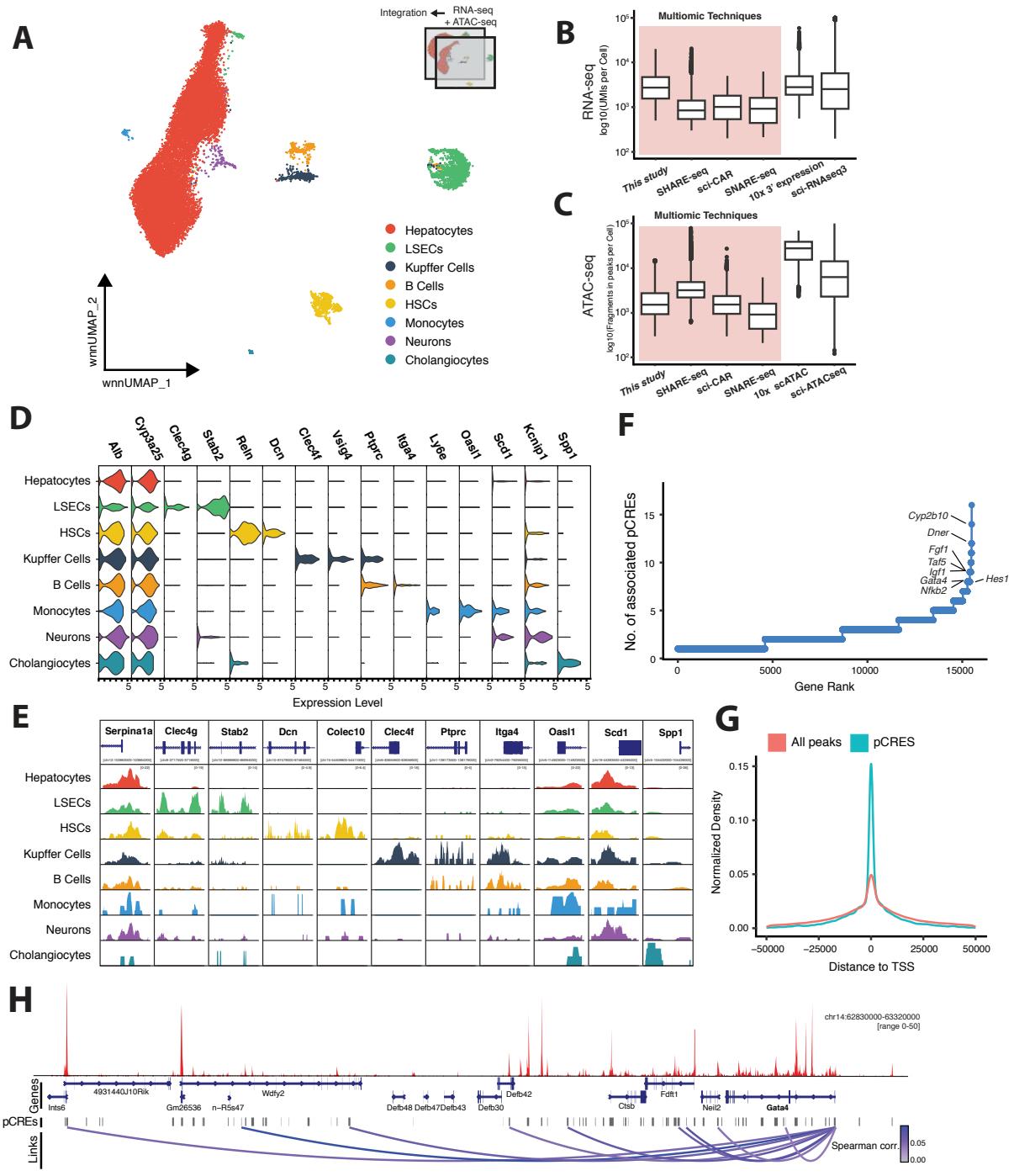

A
Integration
RNA-seq
+ ATAC-seq
Hepatocytes
LSECs
Kupffer Cells
B Cells
HSCs
Monocytes
Neurons
Cholangiocytes
wnnUMAP_2
wnnUMAP_1
B
Multiomic Techniques
RNA-seq
log10(UMIs per Cell)
This study
SHARE-seq
sci-CAR
SNARE-seq
10x 3' expression
sci-RNAseq3
C
Multiomic Techniques
ATAC-seq
log10(Fragments in peaks per Cell)
This study
SHARE-seq
sci-CAR
SNARE-seq
10x scATAC
sci-ATACseq
D
Alb
Cyp3a25
Clec4g
Stab2
Reln
Dcn
Clec4f
Vsig4
Ptprc
Itga4
Ly6e
Oasl1
Scd1
Kcnip1
Spp1
Hepatocytes
LSECs
HSCs
Kupffer Cells
B Cells
Monocytes
Neurons
Cholangiocytes
Expression Level
E
Serpina1a
Clec4g
Stab2
Dcn
Colec10
Clec4f
Ptprc
Itga4
Oasl1
Scd1
Spp1
Hepatocytes
LSECs
HSCs
Kupffer Cells
B Cells
Monocytes
Neurons
Cholangiocytes
F
No. of associated pCREs
Gene Rank
Cyp2b10
Dner
Fgf1
Taf5
Igf1
Gata4
Nfkb2
Hes1
G
All peaks
pCRES
Normalized Density
Distance to TSS
H
chr14:62830000-63320000
[range 0-50]
Genes
pCREs
Links
4931440J10Rik
Wdfy2
Ints6
Gm26536
n-R5s47
Defb48
Defb47
Defb43
Defb30
Defb42
Ctsb
Fdft1
Neil2
Gata4
Spearman corr.

Figure 2: Joint expression and chromatin accessibility profiling in primary liver nuclei

(A) UMAP visualisation of WNN-integrated scRNAseq and scATACseq modalities of 19,664 liver nuclei. Nuclei are coloured by cell types.
(B) Comparison of UMIs/cell across different single-cell technologies. Red shading denotes all multi-omic technologies. Datasets are this study, SHARE-seq[13] (murine skin cells), sci-CAR[11] (murine kidney nuclei), SNARE-seq[12] (adult & neonatal mouse cerebral cortex nuclei), 10x 3' Expression[17] (murine liver nuclei) and sci-RNAseq3 [9](E16.5 mouse embryo nuclei).
(C) Comparison of unique fragments per cell across different single-cell technologies. Colouring as in (B). Datasets differing to (B) are 10x 3'scATAC[42] (murine liver nuclei) and sciATAC-seq[43] (murine liver nuclei).
(D) Normalised gene expression of representative marker genes per cell type.
(E) Aggregate ATAC-seq tracks at marker accessibility peaks per cell type.
(F) Genes ranked by number of significantly correlated pCREs ($P < 0.05$, FDR = 0.1) per gene (±500kbp from TSS) in LSECs. Marked are transcription factors & regulators within the top 1% of genes with a critical role in LSECs.
(G) Significantly correlated pCREs are enriched for TSS proximity. Normalised density of all peaks versus pCREs within ±50kbp of nearest TSS.
(H) Aggregate scATAC-seq track of LSECs at the *Gata4* locus and 500kbp upstream region. Loops denote pCREs significantly correlated with *Gata4* and are coloured by Spearman correlation of respective pCRE–*Gata4* comparison

Figure 3

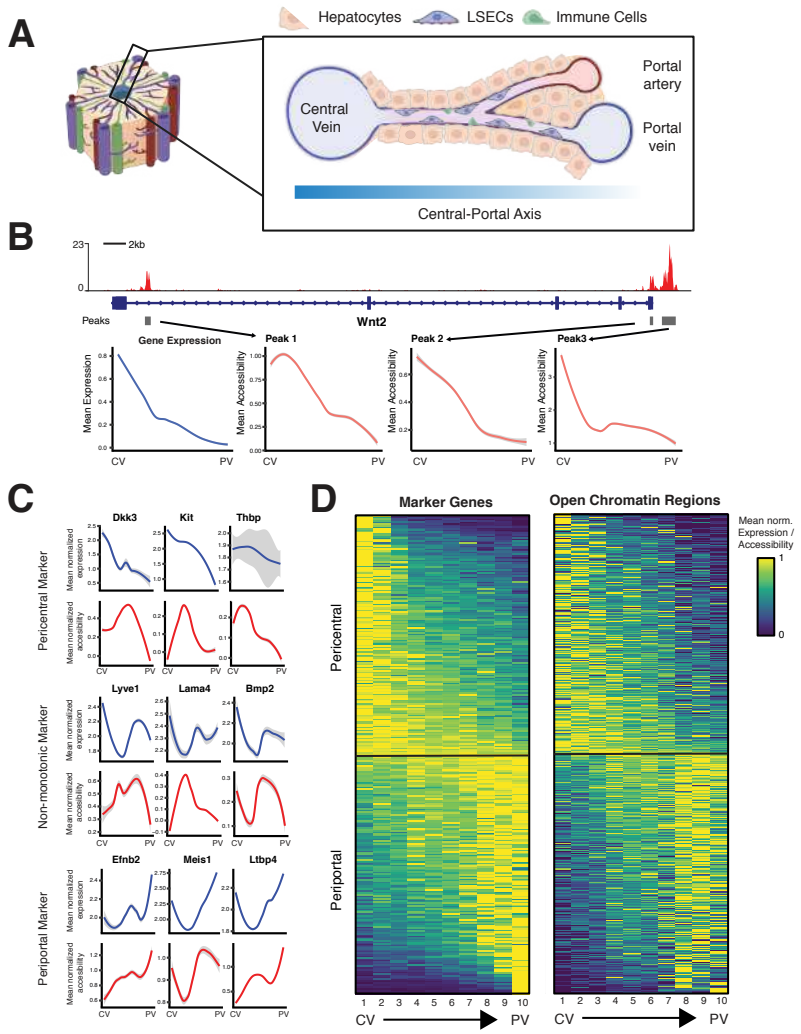


Figure 3: Zonation profiles in LSECs across gene expression and chromatin accessibility

(A) Schematic of a liver lobule. A liver lobule has a 'Central–Portal Axis' starting from the central vein to the portal vein and portal artery. The sinusoidal capillary channels are lined with LSECs.

(B) Changes along the Central–Portal Axis at the *Wnt2* locus. Top: Aggregate scATACseq profile (red) of LSECs at *Wnt2* locus. Grey bars denote identified peaks. Bottom: In blue, loess trend line of mean normalised *Wnt2* gene expression along the Central–Portal-Axis (central vein, CV; portal vein, PV; split into equal 10 bins). In red, loess trend line of mean normalised chromatin accessibility in peaks at the *Wnt2* locus along the CP-axis.

(C) Loess trend line of mean normalised expression (blue) and mean normalised accessibility along the Central–Portal axis for pericentral markers (top, increased toward the central vein, *Dkk3*, *Kit* and *Thbp*), non-monotonic markers (middle, increased between the veins, *Lyve1*, *Lama4* and *Bmp2*) and periportal markers (increased toward the portal vein, *Efnb2*, *Meis1* & *Ltbp4*)

(D) Left: Zonation profiles of 362 genes along the Central–Portal axis. Right: Zonation profiles of 846 open chromatin regions along the Central–Portal axis. All profiles are normalised by their maximum.

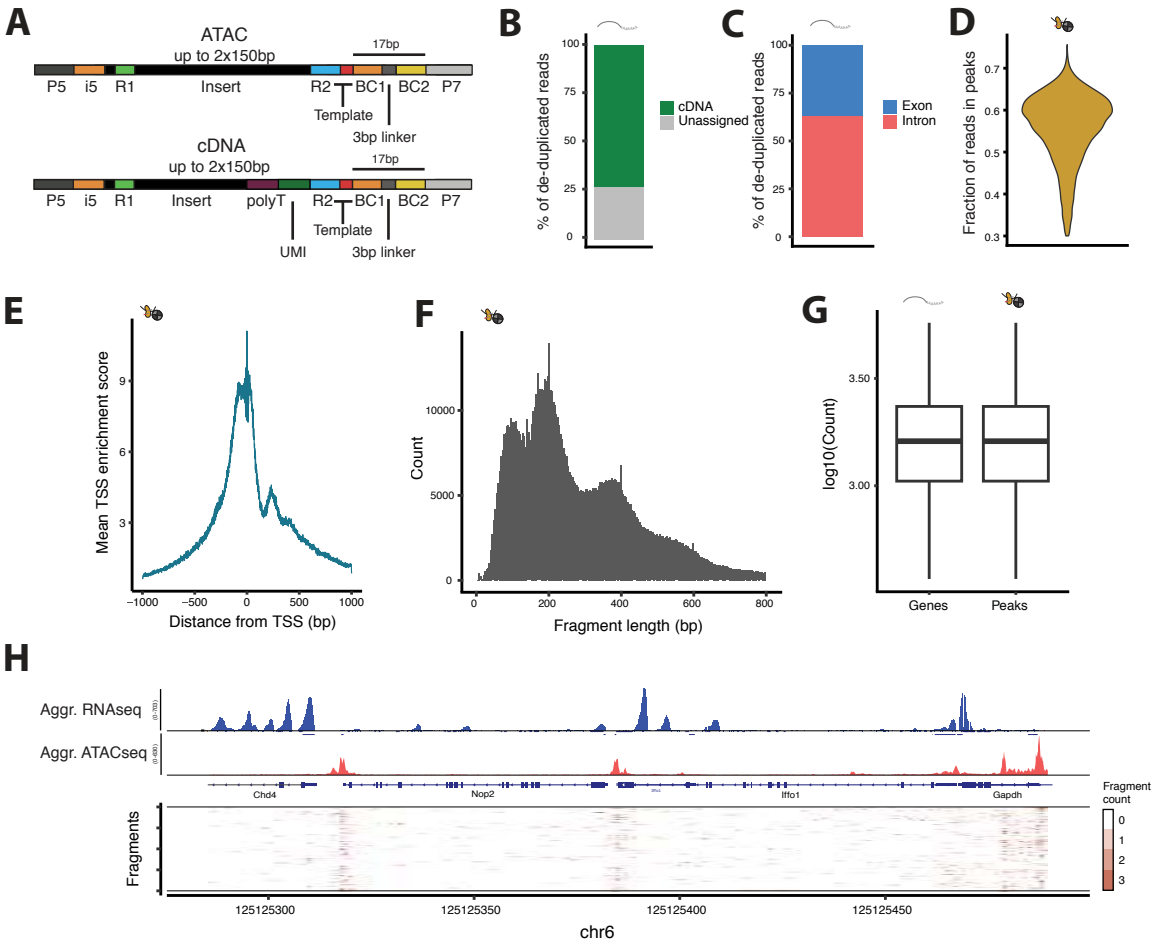
Table 1

Comparison of single-cell techniques

	Cost / Cell	Throughput	Multiomic?	Special equipment?	Std. sequencing?	Potential insert length?
This study	5.6 ct	> 200.000	Yes	No	Yes	> 200bp
SHARE-seq	4.33 ct	> 200.000	Yes	No	No	100bp
10x Multiome	25.8 ct	80.000	Yes	Yes	No	100bp
sci-RNA-seq3	1 ct	> 200.000	No	No	Yes	> 200bp

Table 1: Comparison between different single-cell technologies

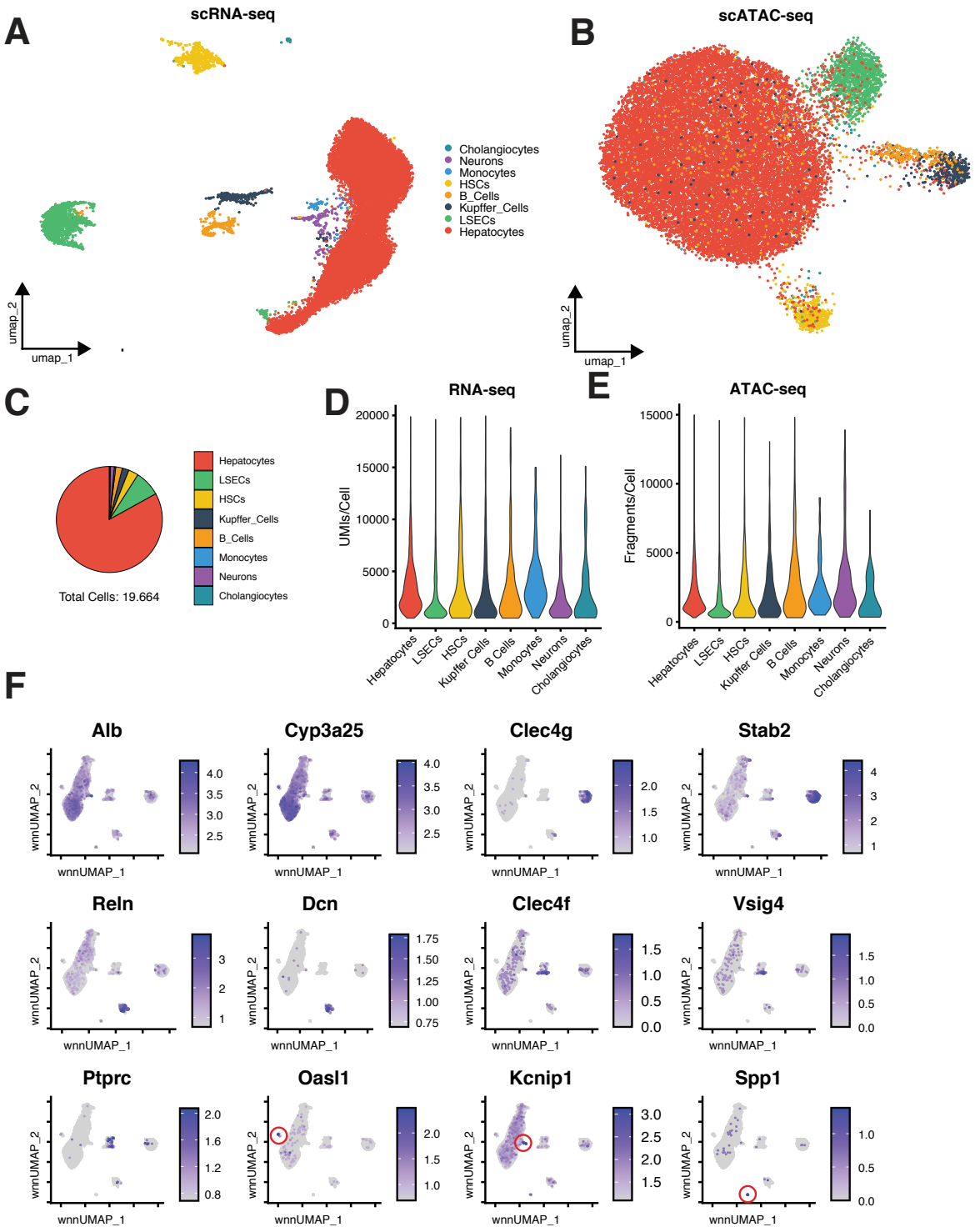
Supplementary Figure 1



Supplementary Figure 1: Barcode structure and summary of quality control measures in liver nuclei

(A) Structure of a scATAC-seq and scRNA-seq sequencing read. Created with Biorender.com
(B) Percentage of total scRNAseq sequencing reads containing cDNA fragments.
(C) Percentage of de-duplicated scRNAseq sequencing reads overlapping an exon or intron.
(D) Distribution of fraction of reads in peaks (FRiP) per cell in the scATAC-seq data (mean: 0.55).
(E) Mean TSS enrichment score per cell in relation to distance from nearest TSS in the scATACseq data.
(F) Histogram of fragment length in scATAC sequencing reads
(G) Expressed genes and accessible peaks per cell (mean expressed genes: 1,798; mean accessible peaks: 1,983)
(H) Top: Aggregate scRNA-seq (blue) and scATAC-seq (red) of all liver nuclei at *Nop2/Iffo2/Gapdh* locus. Bottom: Chromatin accessibility profiles of 100 individual cells.

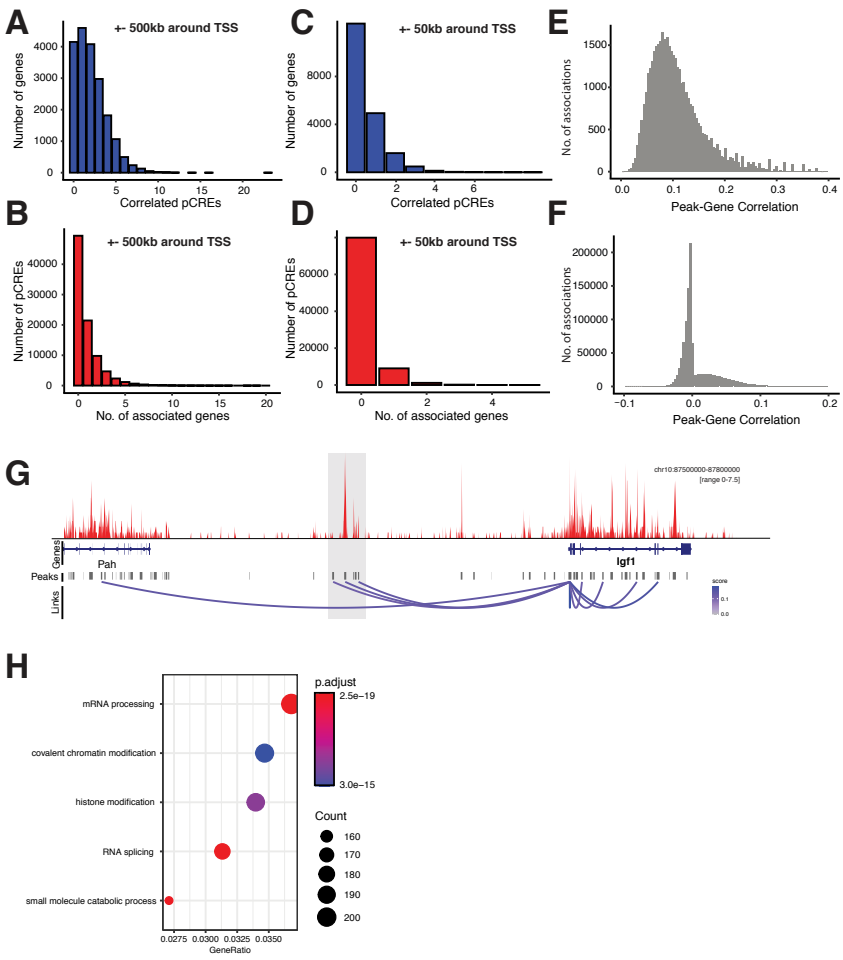
Supplementary Figure 2



Supplementary Figure 2: easySHAREseq robustly separates cell types

(A) UMAP visualisation of merged and integrated scRNA-seq data. Nuclei are coloured according to their cell type.
(B) UMAP visualisation of merged and integrated scATAC-seq data. Nuclei are coloured according to their cell type.
(C) Fraction of cell types recovered relative to total cells
(D) Distribution of UMIs per cell split by cell type. Some cell types (e.g. LSECs) consistently yield less UMIs.
(E) Distribution of unique fragments per cell split by cell types. Some cell types (e.g. LSECs) consistently yield less fragments.
(F) WNN-UMAPs with cells coloured according to the mean expression strength of a given marker gene. Red circles indicate the position of the cell population showing elevated expression for this marker gene.

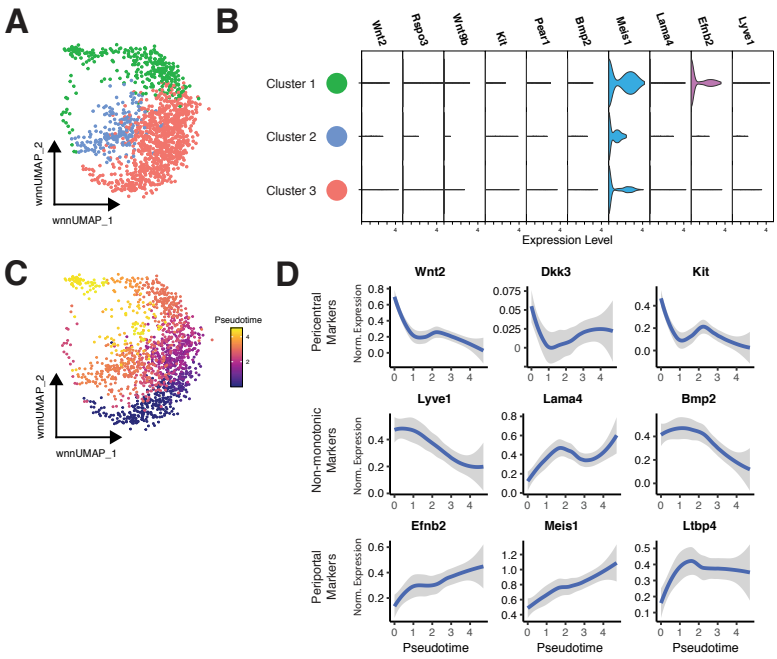
Supplementary Figure 3



Supplementary Figure 3: Summary of peak–gene correlations

(A) Number of significantly correlated pCREs ($P < 0.05$, FDR = 0.1) per gene, considering all peaks ±500kbp of the TSS
(B) Number of genes a given pCREs is significantly correlated with ($P < 0.05$, FDR = 0.1), considering all peaks ±500kbp of the TSS
(C) Number of significantly correlated pCREs ($P < 0.05$, FDR = 0.1) per gene, considering all peaks ±50kbp of the TSS
(D) Number of genes a given pCREs is significantly correlated with ($P < 0.05$, FDR = 0.1), considering all peaks ±50kbp of the TSS
(E) Histogram of Spearman correlations of all significant peak–gene correlations ($P < 0.05$)
(F) Histogram of Spearman correlations of all non-significant peak–gene correlations ($P > 0.05$)
(G) Aggregate scATAC-seq track of LSECs at the *Igf1* locus and its upstream region. Loops denote significantly correlated pCREs with *Igf1* and are coloured by their respective Spearman correlation. Shaded grey area denotes potentially LSEC-specific *cis*-regulatory element regulating *Igf1* expression.
(H) Gene Ontology enrichment analysis of genes whose associated pCREs are associated with five or more genes.

Supplementary Figure 4



Supplementary Figure 4: Investigation of LSEC zonation

(A) Subclustering LSECs reveals three distinct clusters.
(B) Comparison of marker gene expression across the three identified LSEC subclusters does not allow for fine-scale cell-type assignments.
(C) Subclustered LSECs coloured by pseudotime.
(D) Loess-Curve of marker gene expression of pericentral, non-monotonic and periportal marker genes along pseudotime.

Methods

Animal Model & Tissue preparation

Mice

All animal experimental procedures were carried out under the licence number EB 01-21M at Friedrich Miescher Laboratory of the Max Planck Society in Tübingen, Germany. The procedures were reviewed and approved by the Regierungspräsidium Tübingen, Germany. Liver was collected from both male and female wild-type C57BL/6 and PWD/PhJ mice aged between 9 to 11 weeks.

Study design

From each strain, we generated easySHARE-seq libraries for one male and one female mice from each strain (four total). For each individual, we sequenced two sub-libraries, resulting in 8 easySHAREseq libraries.

Cell Culture

For the species-mixing experiment, HEK Cells were cultured in media containing DMEM/F-12 with GlutaMAX™ Supplement, 10% FBS and 1% Penicillin-Streptomycin (PenStrep) at 37°C and 5% CO_2. Cells were harvested on the day of the experiment by simply pipetting them off the plate and were then spun down for 5 min at 250G.
For the second cell line, murine OP9-DL4 cells were cultured in alpha-MEM medium containing 5% FBS and 1% PenStrep. On the day of the experiment, the cells were harvested by aspirating the media and adding 4 ml of Trypsin, followed by an incubation at 37°C for 5 min. Then, 5ml of media was added and cells were spun down for 5 min at 250G.
After counting both cell lines using TrypanBlue and the Evos Countess II, equal cell numbers were mixed.

Liver Nuclei

The liver was extracted, rinsed in HBSS, cut into small pieces, frozen in liquid nitrogen and stored in the freezer at -80 °C for a maximum of two weeks. On the day of the experiment, 1 ml of ice cold Lysis Solution (0.1% Triton-X 100, 1mM DTT, 10mMM Tris-HCl pH8, 0.1mM EDTA, 3mM $Mg(Ac)_2$, 3mM $CaCl_2$ and 0.32M sucrose) was added to the tube. The cell suspension was transferred to a pre-cooled Douncer and dounced 10x using Pestle A (loose) and 15x using Pestle B (tight). The solution was added to a thick wall ultracentrifuge tube on ice and topped up with 4ml ice cold Lysis Solution. Then 9 ml of Sucrose solution (10mM Tris-HCl pH8.0, 3mM $Mg(Ac)_2$, 3mM DTT, 1.8M sucrose) was carefully pipetted to the bottom of the tube to create a sucrose cushion. Samples were spun in a pre-cooled ultracentrifuge with a SW-28 rotor at 24,400rpm for 1.5 hours at 4 °C. Afterwards, all supernatant was carefully aspirated so as not to dislodge the pellet at the bottom and 1 ml ice cold DEPC-treated water supplemented with 10µl SUPERase & 15µl Recombinant RNase Inhibitor was added. Without resuspending, the tube was kept on ice for 20 min. The pellet was then resuspended by pipetting ~15 times slowly up and down followed by a 40 µm cell straining step. Counting of the nuclei using DAPI and the Evos Countess II was immediately followed up by fixation.

easySHARE-seq protocol

Preparing the barcoding oligonucleotides

There are two barcoding rounds in easySHARE-seq with 192 unique barcodes distributed across two 96-well plates in each round (see **Suppl. Table 1** for a full list of oligonucleotide sequences). Each barcode (BC) is pre-annealed as a DNA duplex for improved stability. The first round of barcodes contains two single-stranded linker sequences at its ends as well as a 5' phosphate group to ligate the different barcodes together. The first single-stranded overhang links the barcode to a complementary overhang at the 5' end of the cDNA molecule or transposed DNA molecule, which originates either from the RT primer or the Tn5 adapter. The second overhang (3bp) is used to ligate it to the second round of barcodes (**Fig.1B**). Each duplex needs to be annealed prior to cellular barcoding, preferably on the day of the experiment. No blocking oligos are needed.

The Round1 BC plates contain 10µl of 4µM duplexes in each well and Round2 BC plates contain 10µl of 6µM barcode duplexes in each well, all in Annealing Buffer (10mM Tris pH8.0, 1mM EDTA, 30mM KCl). Pre-aliquoted barcoding plates can be stored at -20 °C for at least three months. On the day of the experiment, the oligo plates were thawed and annealed by heating plates to 95 °C for 2 min, followed by cooling down the plates to 20 °C at a rate of -2 °C per minute. Finally, the plates were spun down. Until the annealed barcoding plates are needed, they should be kept on ice or in the fridge.

This barcoding scheme is very flexible and currently supports a throughput of ~350,000 cells (assuming 96 indexing primers) per experiment, limited only by sequencing cost and availability of indexing primer. The barcodes were designed to have at least a Hamming distance of 2. See Supplementary Notes for further details on the barcoding system and flexibility.

Tn5 *preparation*

Tn5 was expressed in-house as previously described [44]. Two differently loaded Tn5 are needed for easySHARE-seq, one for the tagmentation, loaded with an adapter for attaching the first barcodes (termed Tn5-B2S), and one for library preparation with a standard illumina sequencing adapter (termed Tn5-A-only). See Supplementary Table 1 for all sequences.

To assemble Tn5-B2S, two DNA duplexes were annealed: 20 µM Tn5-A oligo with 22 µM Tn5-reverse and 20 µM Tn5-B2S with 22 µM Tn5-reverse, all in 50 mM NaCl and 10mM Tris pH8.0. Oligos were annealed by heating the solution to 95 °C for 30 s and cooling it down to 20 °C at a rate of 2 °C/min. An equal volume of duplexes was pooled and then 200 µl of unassembled Tn5 was mixed with 16.5 µl of duplex mix. The Tn5 was then incubated at 37 °C for 1 hour, followed by 4 °C overnight. The Tn5 can then be stored at -20 °C. In our hands, Tn5 did not show a decrease in activity after 10 months of storage.

To assemble Tn5-A-only, 10 µM of Tn5-A and 10.5 µM Tn5-reverse was annealed using the same conditions as described above. Again, 200 µl of unassembled Tn5 was mixed with 16.5 µl of Tn5-A duplex and incubated at 37 °C for 1 hour, followed by 4 °C overnight. The Tn5 can then be stored for later and repeated use for more than 10 months at -20 °C.

We observed an increase in all Tn5 activity during the first months of storage, possibly due to continued transposome assembly in storage.

Fixation

One million liver nuclei ("cells" for short) were added to ice-cold PBS for 4 ml total. After mixing, 87 µl 16% formaldehyde solution (0.35%; for liver nuclei) or 25 µl 16% formaldehyde solution

(0.1%; for HEK and OP9 cells) was added and the suspension was mixed by pipetting up and down exactly 3 times with a P1000 pipette set to 700 µl. The suspension was incubated at room temperature for 10 min. Fixation was stopped by adding ice-cold Stop-Mix (224 µl 2.5M glycine, 200 µl 1M Tris-HCl pH8.0, 53 µl 7.5% BSA in PBS). The suspension was mixed exactly 3 times with a P1000 pipette set to 850 µl and incubated on ice for 3 min followed by a centrifugation at 500G for 5 min at 4°C. Supernatant was removed and the pellet was resuspended in 1 ml Nuclei Isolation Buffer (NIB; 10mM Tris pH8.0, 10mM NaCl, 2mM $MgCl_2$, 0.1% NP-40) and kept on ice for 3 min followed by straining the suspension with a 40 µm cell strainer. It was then spun down at 500G for 3 min at 4°C and re-suspended in ~100-200µl PBSi (1x PBS + 0.4 U/µl Recombinant RNaseInhibitor, 0.04% BSA, 0.2 U/µl SUPERase, freshly added), depending on the amount of input cells. Cells were then counted using DAPI and the Countess II and concentration was adjusted to 2M cells/ml using PBSi.

Tagmentation

In a typical easySHARE-seq experiment for this study, 8 tagmentation reactions with 10,000 cells each followed by 3 RT reactions were performed. This results in sequencing libraries for around 30,000 cells. To increase throughput, simply increase the amount of tagmentation and RT reactions accordingly. No adjustment is needed to the barcoding. Each tube and PCR strip until the step of Reverse Crosslinking was coated before use by rinsing it with PBS+0.5% BSA.

For each tagmentation reaction, 5 µl of 5X TAPS-Buffer, 0.25µl 10% Tween, 0.25µl 1% Digitonin, 3 µl PBS, 1 µl Recombinant RNaseInhibitor and 9µl of H2O was mixed. TAPS Buffer was made by first making a 1M TAPS stock solution in H_2O, followed by adjustment of the pH to 8.5 by titrating 10M NaOH. Then, 4.25ml H_2O, 500µl 1M TAPS pH8.5, 250µl 1M $MgCl_2$ and 5ml N-N-Di-Methyl-Formamide (DMF) was mixed on ice and in order. When adding DMF, the buffer heats up so it is important to be kept on ice. The resulting 5X TAPS-Buffer can then be stored at 4°C for short term use (1-2 months) or for long-term storage at -20°C (> 6 months).

Then, 5 µl of cell suspension at 2M cells/ml in PBSi was added to the tagmentation mix for each reaction, mixed thoroughly and finally 1.5µl of Tn5-B2S was added. The reaction was incubated on a shaker at 37°C for 30 min at 850 rpm. Afterwards, all reactions were pooled on ice into a pre-cooled 15ml tube. The reaction wells were washed with ~30 µl PBSi which was then added to the pooled suspension in order to maximize cell recovery. The suspension was then spun down at 500G for 3 min at 4°C. Supernatant was aspirated and the cells were washed with 200µl NIB followed by another centrifugation at 500G for 3 min at 4°C.

We only observed cell pellets when centrifuging after fixation and only when using cell lines as input material. Therefore, when aspirating supernatant at any step it is preferable to leave around 20-30µl liquid in the tube. Additionally, it is recommended to pipette gently at any step as to not damage and fracture the cells.

Reverse Transcription

As stated above, three tagmentation reactions were combined into one RT reaction. When increasing cells to more than 30,000 per RT reaction, we observed a steep drop in reaction efficiency.

The Master Mix for one RT reaction contained 3µl 100µM RT-primer, 2µl 10mM dNTPs, 6µl 5X MaximaH RT Buffer, 4.5µl 50% PEG6000, 1.5 µl H_2O, 1.5µl SUPERase and 1.66µl MaximaH RT. The RT primer contains a polyT tail, a 10bp UMI sequence, a biotin molecule and an adapter sequence used for ligating onto the first round of barcoding oligos.

The cell suspension was resuspended in 10µl NIB per RT reaction and added to the Master Mix for a total of 30µl. As PEG is present, it is necessary to pipette ~30 times up and down to ensure proper mixing. The RT reaction was performed in a PCR cycler with the following protocol: 52°C for 12 min; then 2 cycles of 8°C for 12s, 15°C for 45s, 20°C for 45s, 30°C for 30s, 42°C for 2min and 50°C for 3 min. Finally, the reaction was incubated at 52°C for 5 more minutes. All reactions were then pooled on ice into a pre-cooled and coated 15ml tube and the reaction wells were washed with ~40µl NIB, which was then added to the pooled cell suspension in order to maximise cell recovery. The suspension was then spun down at 500G for 3 min at 4°C. Supernatant was aspirated and the cells were washed in 150µl NIB and spun down again at 500G for 3min at 4°C. This washing step was repeated once more, followed by resuspension of the cells in 2ml Ligation Mix (400µl 10x T4-Buffer, 40µl 10% Tween-20, 1460µl Annealing Buffer and 100µl T4 DNA Ligase, added last).

Single-cell barcoding

Using a P20 pipette, 10µl of cell suspension in the ligation mix was added to each well of the two annealed Round1 BC plates, taking care as to not touch the liquid at the bottom of each well. The plates were then sealed, shaken gently by hand and quickly spun down (~ 8s) followed by an incubation on a shaker at 25°C for 30 min at 350 rpm. After 30 min, the cells from each well were pooled into a coated PCR strip using a P200 multichannel pipette set to 30µl. In order to pool, each row was pipetted up and down three times before adding the liquid to the PCR strip. After 8 columns were pooled into the strip, the suspension was transferred into a coated 5ml tube on ice. This process was repeated until both plates were pooled, taking care to aspirate most liquid from the plates. The cell suspension was then spun down for 3min at 500G at 4°C. Supernatant was aspirated and the cells were resuspended thoroughly in 2 ml new Ligation Mix. Now, 10µl of cell suspension was added into each well of the annealed Round2 barcoding plates using a P20 pipette, taking care as to not touch the liquid within each well. The plates were sealed, shaken gently by hand and spun down quickly followed by incubating them on a shaker at 25°C for 45 min at 350 rpm. The cells were then pooled again using the above described procedure into a new coated 15ml Tube. The cells were spun down at 500G for 3 min at 4°C. Supernatant was aspirated, the cells were washed with 150µl NIB and spun down again. Finally, the cells were resuspended in ~60µl NIB and counted. For counting, 5µl of cells were mixed with 5µl of NIB and 1x DAPI and counted on the Evos Countess II, taking the dilution into account. Sub-libraries of 3,500 cells were made and the volume was adjusted to 25µl by addition of NIB.

Using 3,500 cells results in a doublet rate of ~6.3%. The recovery rate of cells after sequencing depends on the input material (and QC thresholds), with cell lines recovering around 80% of input cells (~2,800-3,000 cells) and liver nuclei around 70% (~2,300-2,500 cells).

Reverse-Crosslinking

To each sub-library of 3,500 cells, 30µl 2x Reverse Crosslinking (RC) Buffer (0.4% SDS, 100mM NaCl, 100mM Tris pH8.0) as well as 5µl ProteinaseK was added. The sub-libraries were mixed and incubated on a shaker at 62°C for one hour at 800 rpm. Afterwards, they were transferred to a PCR cycler into a deep well module set to 62°C (lid to 80°C) for an additional hour. Afterwards, each sub-library was incubated at 80°C for 10 min and finally 5µl of 10% Tween-20 to quench the SDS and 35µl of NIB was added for a total volume of 100µl.

The lysates can be stored at this point at -20°C for at least two days, which greatly simplifies handling many sub-libraries at once. Longer storage has not been extensively tested.

Streptavidin Pull-Down

Each transcript contains a biotin molecule as the RT primers are biotinylated which is used to separate the scATAC-seq libraries from the scRNA-seq libraries. For each sublibrary, 50µl M280 Streptavidin beads were washed three times with 100µl B&W Buffer (5mM Tris pH8.0, 1M NaCl, 0.5mM EDTA) supplemented with 0.05% Tween-20, using a magnetic stand. Afterwards, the beads were resuspended in 100µl 2x B&W Buffer and added to the sublibrary, which were then shaken at 25°C for one hour at 900 rpm. Now all cDNA molecules are attached to the beads whereas transposed molecules are within the supernatant. The lysate was put on a magnetic stand to separate supernatant and beads.

It likely is possible to reduce the number of M280 beads in this step, significantly reducing overall costs. However, this has not been extensively tested.

scATAC-seq library preparation

The supernatant from each sub-library was cleaned up with a Qiagen MinElute Kit and eluted twice into 30µl 10mM Tris pH8.0 total. PCR Mix containing 10µl 5X Q5 Reaction Buffer, 1µl 10mM dNTPs, 2µl 10µM i7-TruSeq-long primer, 2µl 10µM Nextera N5XX Indexing primer, 4.5µl H2O and 0.5µl Q5 Polymerase was added (All Oligo sequences in **Suppl. Table 1**). Importantly, in order to distinguish the samples, each sub-library needs to be indexed with a different N5XX Indexing primer. The fragments were amplified with the following protocol: 72°C for 6 min, 98°C for 1 min, then cycles of 98°C for 10s, 66°C for 20s and 72°C for 45s followed by a final incubation at 72°C for 2 min. The number of PCR cycles strongly depends on input material (Liver: 17 PCR cycles, Cell Lines: 15 PCR cycles). The reactions were then cleaned up with custom size selection beads with 0.55X as upper cutoff and 1.4X as lower cutoff and eluted into 25µl 10mM Tris pH8.0. Libraries were quantified using the Qubit HS dsDNA Quantification Kit and run on the Agilent 2100 bioanalyzer with a High Sensitivity DNA Kit.

cDNA library preparation

The beads containing the cDNA molecules were washed three times with 200µl B&W Buffer supplemented with 0.05% Tween-20 before being resuspended in 100µl 10mM Tris ph8.0 and transferred into a new PCR strip. The strip was put on a magnet and the supernatant was aspirated. The beads were then resuspended in 50µl Template Switch Reaction Mix: 10µl 5X MaximaH RT Buffer, 2µl 100µM TS-oligo, 5µl 10mM dNTPs, 3µl Enzymatics RNaseIn, 15µl 50% PEG6000, 14µl H2O and 1.25µl MaximaH RT. The sample was mixed well and incubated at 25°C for 30 min followed by an incubation at 42°C for 90 min. The beads were then washed with 100µl 10mM Tris while the strip was on a magnet and resuspended in 60µl H_2O. To each well, 40µl PCR Mix was added containing 20µl 5X Q5 Reaction Buffer, 4µl 10µM i7-Tru-Seq-long primer, 4µl 10µM Nextera N5XX Indexing primer, 2µl 10mM dNTPs, 9µl H2O and 2µl Q5 Polymerase. The resulting mix can be split into two 50µl PCR reactions or run in one 100µl reaction. The PCR involved initial incubation at 98°C for 1 min followed by PCR cycles of 98°C for 10s, 66°C for 20s and 72°C for 3 min with a final incubation at 72°C for 5 min. Importantly, in order to distinguish the samples, each sub-library needs to be indexed with a different N5XX Indexing primer. The number of PCR cycles strongly depends on input material (Liver: 15 cycles, Cell lines: 13 cycles).

The PCR reactions were cleaned up with custom size selection beads using 0.7X as a lower cutoff (70µl) and eluted into 25µl 10mM Tris pH8.0. The cDNA libraries were quantified using the Qubit HS dsDNA Quantification Kit.

scRNA-seq library preparation

As the cDNA molecules are too long for sequencing (mean length > 700bp), they need to be shortened on one side. To achieve this, 25ng of each cDNA library was transferred to a new strip and volume was adjusted to 20µl using H_2O. Then 5µl 5X TAPS Buffer and 0.8µl Tn5-A-only was added and the sample was incubated at 55°C for 10 min. To stop the reaction, 25µl 1% SDS was added followed by another incubation at 55°C for 10 min. The sample was then cleaned up with custom size selection beads using a ratio of 1.3X and eluted into 30µl. Then 20µl PCR mix was added containing 10µl 5X Q5 reaction buffer, 1µl 10mM dNTPs, 2µl 10µM i7-Tru-Seq-long primer, 2µl 10µM Nextera N5XX Indexing primer (note: each sample needs to receive the **same** index primer as was used in the cDNA library preparation), 4.5µl H_2O and 0.5µl Q5 Polymerase. The PCR reaction was carried out with the following protocol: 72°C for 6 min, 98°C for 1 min, followed by 5 cycles of 98°C for 10s, 66°C for 20s and 72°C for 45s with a final incubation at 72°C for 2 min. Libraries were purified using custom size selection beads with a ratio of 0.5X as an upper cutoff and 0.8X as a lower cutoff. The final scRNA-seq libraries were quantified using the Qubit HS dsDNA Quantification Kit and run on the Agilent 2100 bioanalyzer with a High Sensitivity DNA Kit.

Sequencing

Both scATAC-seq and scRNA-seq libraries were sequenced simultaneously as they were indexed with different Index2 indices (N5XX). All libraries were sequenced on the Nova-seq 6000 platform (Illumina) using S4 2x150bp v1.5 kits (Read 1: 150 cycles, Index 1: 17 cycles, Index 2: 8 cycles, Read 2: 150 cycles). Libraries were partly multiplexed with standard Illumina sequencing libraries.

Custom Size selection beads

To make custom size selection beads, we washed 1ml of SpeedBeads on a magnetic stand in 1ml of 10mM Tris-HCl pH8.0 and re-suspended them in 50ml Bead Buffer (9g PEG8000, 7.3g NaCl, 500ul 1M Tris HCl pH8.0, 100ul 0.5M EDTA, add water to 50ml). The beads don't differ in their functionality from other commercially available ready-to-use size selection beads. They can be stored at 4°C for > 3 months.

Analysis

Gene annotations and Genomic variants

The reference genome and the Ensembl gene annotation of the C57BL/6J genome (mm10) were downloaded from Ensembl (Version GRCm38, release 102). Gene annotations for PWD/PhJ mice were downloaded from Ensembl. A consensus gene annotation set in mm10 coordinates was constructed by filtering for genes present in both gene annotations.

easySHARE-RNA-seq pre-processing

Fastq files were demultiplexed using a custom C-script, allowing one mismatch within each barcode segment. The reads were trimmed using cutadapt[48]. UMIs were then extracted from bases 1-10 in Read 2 using UMI-Tools[45] and added to the read name. Only reads with TTTTT at the bases 11-15 of Read 2 were kept (> 96%), allowing one mismatch. Lastly, the barcode was also moved to the read name.

Species-Mixing Experiments

RNA-seq reads were aligned to a composite hg38-mm10 genome using STAR[46]. The resulting bamfile was then filtered for uniquely mapping reads and reads mapping to chrM, chrY or unmapped scaffolds or containing unplaced barcodes were removed. Finally, the reads were deduplicated using UMItools[45]. ATAC-seq reads were also aligned to a composite genome using bwa[47]. Duplicates were removed with Picard tools and reads mapping to chrM, chrY or unmapped scaffolds were filtered out. Additionally, reads that were improperly paired or had an alignment quality < 30 were also removed.

The reads were then split depending on which genome they mapped to and reads per barcode were counted. Barcodes needed to be associated with at least 700 fragments and 500 UMIs in order to be considered a cell for the analysis. A barcode was considered a doublet when either the proportion of UMIs or fragments assigned to a species was less than 75%. This cutoff was chosen to mitigate possible mapping bias within the data.

easySHARE-RNA-seq processing and read alignment

We only used Read 1 for all our RNA-seq analyses as sequencing quality tends to drop after a polyT tail is sequenced in R2. Each sample was mapped to mm10 using the twopass mode in STAR[46] with the parameters --outFilterMultimapNmax 20 --outFilterMismatchNmax 15. We then processed the bamfiles further by moving the UMI and barcode from the read name to a bam flag, filtering out multimapping reads and reads without a definitive barcode. To determine if a read overlapped a transcript, we used featureCounts from the subread package[48]. UMI-Tools was used to collapse the UMIs of aligned reads, allowing for one mismatch and de-duplication of the reads. Finally, (single-cell) count matrices were created also using UMI-Tools.

easySHARE-ATAC-seq pre-processing and read alignment

Fastq files were demultiplexed using a custom C-script, allowing one mismatch within each barcode segment. The paired reads were trimmed using cutadapt[49] and the resulting reads were mapped to the mm10 genome using bwa mem[47]. Reads with alignment quality < Q30, unmapped, undetermined barcode, or mapped to mtDNA were discarded. Duplicates were removed using Picard tools. Open chromatin regions were called by subsampling the bamfiles from all samples to a common depth, merging them into a pooled bamfile and using the peak caller MACS2[50] with the parameters -nomodel -keep-dup -min-length 100. The count matrices as well as the FRiP score was generated using featureCounts from the Subread package[48] together with the tissue-specific peak set.

Filtering, Integration & Dimensional reduction of scRNAseq data

The count matrices were loaded into Seurat[51] and cells were then filtered for >200 detected genes, >500 UMIs and < 20.000 UMIs. The sub-libraries coming from the same experiment were then merged together and normalised. Merged experiments from the same species (one from male mouse, one from female mouse) were then integrated by first using SCTransform[52] to normalise the data, then finding common features between the two experiments using FindIntegrationAnchors() and finally integrated using IntegrateData(). Lastly, the integrated datasets from C57BL/6 and PWD/PhJ were again integrated using IntegrateData(). To visualise the data, we projected the cells into 2D space by UMAP using the first 30 principal components and identified clusters using FindClusters().

Filtering, Integration & Dimensional reduction of scATACseq data

Fragments per cell were counted using sinto and the resulting fragment file was loaded into Signac[53] alongside the count matrices and the peakset. We calculated basic QC statistics using base Signac and cells were then filtered for a FRiP score of at least 0.3, > 300 fragments, < 15.000 fragments, a TSS enrichment > 2 and a nucleosome signal < 4. Again, sublibraries coming from the same experiment were merged. We then integrated all four experiments (C57BL/6 & PWD/PhJ, one male & one female mouse each) by finding common features across datasets using FindIntegrationAnchors() using PCs 2:30 and then integrating the data using IntegrateEmbeddings(). To visualise the data, we projected the cells into 2D space by UMAP.

Weighted-Nearest-Neigbor (WNN) Analysis & Cell type identification

In order to use data from both modalities simultaneously, we created a multimodal Seurat object and used WNN[16] clustering to visualise and leverage both modalities for downstream analysis. Afterwards, we assigned cell cycle scores and excluded clusters consisting of nuclei solely in the G2M-phase (2 clusters, 121 nuclei total). Cell types were assigned via expression of previously known marker genes, which allows subsetting the data into cell types.

Calculating Peak–Gene Associations

Peak–gene associations were calculated following the framework described by Ma et al[13]. In short, Spearman correlation was calculated for every peak–gene pair within a +-500kb window around the TSS of the expressed gene. To obtain a background estimation, we used chromVAR[54] (*getBackgroundPeaks*()) to generate 100 background peaks matched in GC bias and chromatin accessibility but randomly distributed throughout the genome. We calculated the Spearman correlation between every background-gene comparison, resulting in a null distribution with known population mean and standard deviation. We then calculated the z-score for the peak–gene pair in question ((correlation - population mean)/ standard deviation) and used a one-sided z-test to determine the p-value. This functionality is also implemented in Signac under the function *LinkPeaks*(). Increasing the number of background peaks to 200, 350 or 500 for each peak–gene pair does not impact the results (*data not shown*).

Analysis of LSEC zonation markers

To analyse gene expression and chromatin accessibility along LSEC zonation, we subsetted our data for LSECs only, extracted expression values and wnnUMAP coordinates and binned the data along the wnnUMAP_2 axis into 10 equal sized bins. We then calculated the mean expression/accessibility for each gene/peak in each bin, excluding cells that contained a zero count. To identify novel marker genes, we excluded genes with low expression and calculated the moving average (for three bins) across the bins. We then required the moving average to continuously decrease (for pericentral marker genes) or increase (for periportal marker genes), allowing two exceptions. Lastly, we divided the means for each gene by their maximum to normalise the values. Identification of *cis*-regulatory elements displaying zonation effects had equal requirements.

Imputation of pseudotime was performed in Monocle3[55] with standard parameters. Gene expression was smoothed over both bins and pseudotime (separately) with local polynomial regression fitting (loess).

Gene Ontology Analysis

Gene Ontology Analysis was done using the R package clusterProfiler[56] with standard parameters.

Data Availability

All data can be accessed using the accession number GSE256434. All code used in data analysis is available at https://github.com/vosoltys/easySHARE_seq.git.

Acknowledgements

We thank members of the Chan and Jones lab for helpful discussions and critical reading of the manuscript. We are very grateful to Arnar Breevoort and Alex Pollen for sharing tissue preparation protocols and a very helpful research visit. We thank Sinja Mattes and all animal care takers at the Friedrich Miescher Laboratory for their work. We also thank the Genome Center in the Max Planck Institute for Biology Tübingen for providing support. The OP9-DL4 cells were a kind gift from Juan Carlos Zúñiga-Pflücker. M.P. is supported by an International Max Planck Research School fellowship. M.K. and Y.F.C. were supported European Research Council Starting Grant 639096 “HybridMiX" and Proof-of-Concept Grant 101069216 “Haplotagging”. The research was supported by the Max Planck Society.

Author Contributions

V.S. and Y.F.C. designed the experiments. V.S. and M.P. developed the barcoding framework for easySHAREseq. V.S. developed the rest of the protocol and performed experiments. V.S. performed the computational analyses advised by Y.F.C. V.S. drafted the manuscript. M.P., D.S., M.K. and Y.F.C. helped with experimental or computational support. All authors reviewed the manuscript. Y.F.C. directed the study with input from all authors.

Declaration of Interest

The authors declare no competing interests.

References

1. Anderson, E., Devenney, P. S., Hill, R. E. & Lettice, L. A. Mapping the Shh long-range regulatory domain. *Development* **141**, 3934–3943 (2014).
2. Nord, A. S. *et al.* Rapid and Pervasive Changes in Genome-wide Enhancer Usage during Mammalian Development. *Cell* **155**, 1521–1531 (2013).
3. Spitz, F. & Furlong, E. E. M. Transcription factors: from enhancer binding to developmental control. *Nat Rev Genet* **13**, 613–626 (2012).
4. Buenrostro, J. D. *et al.* Integrated Single-Cell Analysis Maps the Continuous Regulatory Landscape of Human Hematopoietic Differentiation. *Cell* **173**, 1535-1548.e16 (2018).
5. Zhang, C., Macchi, F., Magnani, E. & Sadler, K. C. Chromatin states shaped by an epigenetic code confer regenerative potential to the mouse liver. *Nat Commun* **12**, 4110 (2021).
6. Lara-Astiaso, D. *et al.* Chromatin state dynamics during blood formation. *Science* **345**, 943–949 (2014).
7. Stuart, T. *et al.* Comprehensive Integration of Single-Cell Data. *Cell* **177**, 1888-1902.e21 (2019).
8. Kashima, Y. *et al.* Single-cell sequencing techniques from individual to multiomics analyses. *Exp Mol Med* **52**, 1419–1427 (2020).
9. Martin, B. K. *et al.* Optimized single-nucleus transcriptional profiling by combinatorial indexing. *Nat Protoc* **18**, 188–207 (2023).
10. Rosenberg, A. B. *et al.* Single-cell profiling of the developing mouse brain and spinal cord with split-pool barcoding. *Science* **360**, 176–182 (2018).
11. Cao, J. *et al.* Joint profiling of chromatin accessibility and gene expression in thousands of single cells. *Science* **361**, 1380–1385 (2018).
12. Chen, S., Lake, B. B. & Zhang, K. High-throughput sequencing of the transcriptome and chromatin accessibility in the same cell. *Nat Biotechnol* **37**, 1452–1457 (2019).
13. Ma, S. *et al.* Chromatin Potential Identified by Shared Single-Cell Profiling of RNA and Chromatin. *Cell* **183**, 1103-1116.e20 (2020).
14. Lähnemann, D. *et al.* Eleven grand challenges in single-cell data science. *Genome Biology* **21**, 31 (2020).
15. Aizarani, N. *et al.* A Human Liver Cell Atlas reveals Heterogeneity and Epithelial Progenitors. *Nature* **572**, 199–204 (2019).
16. Hao, Y. *et al.* Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573-3587.e29 (2021).
17. Su, Q. *et al.* Single-cell RNA transcriptome landscape of hepatocytes and non-parenchymal cells in healthy and NAFLD mouse liver. *iScience* **24**, 103233 (2021).
18. Han, X. *et al.* Mapping the Mouse Cell Atlas by Microwell-Seq. *Cell* **172**, 1091-1107.e17 (2018).
19. Chen, X. *et al.* Structural insights into preinitiation complex assembly on core promoters. *Science* **372**, eaba8490 (2021).
20. Winkler, M. *et al.* Endothelial GATA4 controls liver fibrosis and regeneration by preventing a pathogenic switch in angiocrine signaling. *J Hepatol* **74**, 380–393 (2021).
21. Géraud, C. *et al.* GATA4-dependent organ-specific endothelial differentiation controls liver development and embryonic hematopoiesis. *J Clin Invest* **127**, 1099–1114.
22. Lara-Diaz, V. *et al.* IGF-1 modulates gene expression of proteins involved in inflammation, cytoskeleton, and liver architecture. *J Physiol Biochem* **73**, 245–258 (2017).

23. Baratta, J. L. *et al.* Cellular Organization of Normal Mouse Liver: A Histological, Quantitative Immunocytochemical, and Fine Structural Analysis. *Histochem Cell Biol* **131**, 713–726 (2009).
24. Jungermann, K. & Kietzmann, T. Zonation of parenchymal and nonparenchymal metabolism in liver. *Annu Rev Nutr* **16**, 179–203 (1996).
25. Braeuning, A. *et al.* Differential gene expression in periportal and perivenous mouse hepatocytes. *The FEBS Journal* **273**, 5051–5061 (2006).
26. Planas-Paz, L. *et al.* The RSPO–LGR4/5–ZNRF3/RNF43 module controls liver zonation and size. *Nat Cell Biol* **18**, 467–479 (2016).
27. Wang, B., Zhao, L., Fish, M., Logan, C. Y. & Nusse, R. Self-renewing diploid Axin2+ cells fuel homeostatic renewal of the liver. *Nature* **524**, 180–185 (2015).
28. Halpern, K. B. *et al.* Paired-cell sequencing enables spatial gene expression mapping of liver endothelial cells. *Nat Biotechnol* **36**, 962–970 (2018).
29. Knolle, P. A. & Wohlleber, D. Immunological functions of liver sinusoidal endothelial cells. *Cell Mol Immunol* **13**, 347–353 (2016).
30. Smedsrød, B. Clearance function of scavenger endothelial cells. *Comparative Hepatology* **3**, S22 (2004).
31. Rafii, S., Butler, J. M. & Ding, B.-S. Angiocrine functions of organ-specific endothelial cells. *Nature* **529**, 316–325 (2016).
32. Theilmann, A. L. *et al.* Endothelial BMPR2 Loss Drives a Proliferative Response to BMP (Bone Morphogenetic Protein) 9 via Prolonged Canonical Signaling. *Arteriosclerosis, Thrombosis, and Vascular Biology* **40**, 2605–2618 (2020).
33. Russell, K. S., Stern, D. F., Polverini, P. J. & Bender, J. R. Neuregulin activation of ErbB receptors in vascular endothelium leads to angiogenesis. *American Journal of Physiology-Heart and Circulatory Physiology* **277**, H2205–H2211 (1999).
34. Vihanto, M. M. *et al.* Hypoxia up-regulates expression of Eph receptors and ephrins in mouse skin. *FASEB J* **19**, 1689–1691 (2005).
35. Shen, Z. *et al.* Delta-Like Ligand 4 Modulates Liver Damage by Down-Regulating Chemokine Expression. *Am J Pathol* **186**, 1874–1889 (2016).
36. Zellmer, S. *et al.* Transcription factors ETF, E2F, and SP-1 are involved in cytokine-independent proliferation of murine hepatocytes. *Hepatology* **52**, 2127–2136 (2010).
37. Dong, X. C. *et al.* Inactivation of Hepatic Foxo1 by Insulin Signaling Is Required for Adaptive Nutrient Homeostasis and Endocrine Growth Regulation. *Cell Metabolism* **8**, 65–76 (2008).
38. Wang, X., Yu, Y., Xie, H.-B., Shen, T. & Zhu, Q.-X. Complement regulatory protein CD59a plays a protective role in immune liver injury of trichloroethylene-sensitized BALB/c mice. *Ecotoxicology and Environmental Safety* **172**, 105–113 (2019).
39. Ren, H. *et al.* Sirtuin 2 Prevents Liver Steatosis and Metabolic Disorders by Deacetylation of Hepatocyte Nuclear Factor 4α. *Hepatology* **74**, 723 (2021).
40. Miyao, M. *et al.* Pivotal role of liver sinusoidal endothelial cells in NAFLD/NASH progression. *Laboratory Investigation* **95**, 1130–1144 (2015).
41. Su, T. *et al.* Single-Cell Transcriptomics Reveals Zone-Specific Alterations of Liver Sinusoidal Endothelial Cells in Cirrhosis. *Cellular and Molecular Gastroenterology and Hepatology* **11**, 1139–1161 (2021).
42. Nikopoulou, C. *et al.* Spatial and single-cell profiling of the metabolome, transcriptome and epigenome of the aging mouse liver. *Nat Aging* **3**, 1430–1445 (2023).
43. Cusanovich, D. A. *et al.* A Single-Cell Atlas of In Vivo Mammalian Chromatin Accessibility. *Cell* **174**, 1309-1324.e18 (2018).

44. Picelli, S. *et al.* Tn5 transposase and tagmentation procedures for massively scaled sequencing projects. *Genome Res.* **24**, 2033–2040 (2014).
45. Smith, T., Heger, A. & Sudbery, I. UMI-tools: modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy. *Genome Res* **27**, 491–499 (2017).
46. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
47. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
48. Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930 (2014).
49. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17**, 10–12 (2011).
50. Zhang, Y. *et al.* Model-based Analysis of ChIP-Seq (MACS). *Genome Biology* **9**, R137 (2008).
51. Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat Biotechnol* **36**, 411–420 (2018).
52. Hafemeister, C. & Satija, R. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. *Genome Biology* **20**, 296 (2019).
53. Stuart, T., Srivastava, A., Madad, S., Lareau, C. A. & Satija, R. Single-cell chromatin state analysis with Signac. *Nat Methods* **18**, 1333–1341 (2021).
54. Schep, A. N., Wu, B., Buenrostro, J. D. & Greenleaf, W. J. chromVAR: inferring transcription-factor-associated accessibility from single-cell epigenomic data. *Nat Methods* **14**, 975–978 (2017).
55. Trapnell, C. *et al.* The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat Biotechnol* **32**, 381–386 (2014).
56. Yu, G., Wang, L.-G., Han, Y. & He, Q.-Y. clusterProfiler: an R Package for Comparing Biological Themes Among Gene Clusters. *OMICS: A Journal of Integrative Biology* **16**, 284–287 (2012).

Supplementary Notes

Flexible and high-throughput simultaneous profiling of gene expression and chromatin accessibility in single cells

Volker Soltys[1*#], Moritz Peters[1], Dingwen Su[1], Marek Kučka[1,2,], Yingguang Frank Chan[1,3#]

1 Friedrich Miescher Laboratory of the Max Planck Society, 72076 Tübingen, Germany
2 Department of Translational Genomics, University of Cologne, 50931 Cologne, Germany
3 University of Groningen, Groningen Institute of Evolutionary Life Sciences, 9747 AG Groningen, Netherlands

* Corresponding authors
volker.soltys@tue.mpg.de; frank.chan@rug.nl

Flexibility and applicability of the easySHARE-seq framework

EasySHARE-seq uses a flexible barcoding framework that can be tailored to various experimental designs. As mentioned in the main text, it allows for sequencing of fragment lengths of > 200bp, which can be critical in e.g. studies investigating patterns of allele-specific expression or profiling of individual cancer cells and their mutations. However, in study designs not dependent on SNP Coverage, sequencing costs can be cut with no downside by only sequencing 100bp per fragment.
The entire barcoding can also be easily adapted into other protocols, such as scTCR-seq (CITR-seq; unpublished), allowing for paired investigation of T-cell receptor chains in millions of cells. It is also straightforward to adapt easySHARE-seq to a scRNA-seq only protocol with equal or even increased throughput as well as sample indexing, allowing to run a single experiments for e.g. multiple replicates. To achieve this, the tagmentation step can simply be skipped and the RT-primer can be switched out for /5Phos/GGGCTCGGAGATGTGTATAAGAGACAGNNNNNNNNNN-[8bp-Sample-Index]-/biodT/TTTTTTTTTTTTTTTTTTTTTTTTTVN. Before the first 10bp UMI in R2 an individual sample is now represented with an 8bp barcode. Alternatively, this 8bp barcode can function as an additional cell barcode, which allows for increasing sub-library cell numbers and thereby increasing throughput. All other protocol steps remain. Switching to a scATAC-seq only protocol is done by simply excluding the RT step. This will significantly cut experimental cost as no RT, RNase Inhibitors or Streptavidin beads are required, bringing the cost per cell down to ~2.5 cents/cell (in a 100.000 cell experiment). However, sample indexing is not possible in the current framework.
Throughput of easySHARE-seq is only limited by the availability of Nextera N5XX Indexing Primers, theoretically enabling the simultaneous profiling of up to a million cells.
Lastly, to cut further costs on easySHARE-seq, it is possible to perform only a single ligation step. Leaving out the first ligation (in the BC plates 1) still produces easySHARE-seq libraries

as the initial overhang is 8bp long and therefore can theoretically form a stable hybridization at room temperature.

Critical optimization steps for using easySHARE-seq efficiently

The general molecular steps of easySHARE-seq are quite robust. However, in order to use easySHARE-seq efficiently, some prior optimizations should be performed.
As with most scRNA-seq experiments, sample preparation and fixation have the highest impact on success and quality of the experiment. As sample preparation can be quite different between tissues, general good practice is including a sufficient amount of RNase Inhibitor, especially when input material is concentrated in small volumes.
The strength of fixation has a direct impact on data quality of both the scATAC-seq and scRNA-seq. In general, higher fixation leads to an increase of data quality in the scRNA-seq but makes the tagmentation in the scATAC-seq less efficient. Therefore, fixation parameters can to some extent be adjusted based on the requirements and importance of the respective output modality. Fixation strength should also be optimized in a tissue-specific manner. For example, fixing cell lines in 0.15% PFA was generally sufficient for data quality and maintaining cell integrity throughout the protocol. Liver nuclei needed a higher fixation of 0.35% and Bone Marrow Cells (*not shown*) needed to be fixed in 1% PFA as they are both fragile and contain low amounts of mRNA molecules. Another critical factor is the fixation volume, e.g. fixation with 1% PFA in 1 ml leads to a different outcome than fixation in 4 ml. Generally, it is advisable to fix input material in higher volumes and with a low concentration of cells (~1M/ml) as this leads to more consistent results and less clumping. For initial experiments to devise fixation strength, we advise to simply skip the barcoding step. This can be done by using a standard Tn5 for ATAC-seq and the following RT-primer: GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG/biodT/TTTTTTTTTTTTTTTTTTTTTTTTTVN.
Using these modifications ensures that both the scATAC-seq and scRNA-seq can be amplified with standard Nextera N7XX and N5XX primers, allowing for cost-efficient testing of easySHARE-seq parameters. Additionally, cell integrity should be periodically checked to detect cell clumps and assess cell integrity.
Another critical aspect is minimizing freeze-thaw cycles for barcoding oligos, especially for oligos containing phosphorylation modifications. Repeated freezing and thawing leads to a strong decline in protocol efficiency. Please feel free to contact the First Author for further questions.

Example workflow of easySHARE-seq library generation of 200.000 cells

To perform an experiment with a yield of ~200.000 cells, one needs to perform ~48 tagmentation reactions with 10.000 cells per reaction. After tagmentation, those get distributed into 16 RT reactions. Barcoding is then performed as described in one reaction. Afterwards, 96 sublibraries of ~3.500 cells are aliquoted and can be further processed. After Reverse Crosslinking, the samples can be stored at -20C until the next day.
To simplify the cleanup of the lysate for the scATAC-seq library preparation, they can be cleaned-up with size selection beads by adding 150ul per well.

Appendix II: The evolutionary dynamics of cell-type specific regulatory evolution in *Mus*

The evolutionary dynamics of cell-type specific regulatory evolution in *Mus*

Volker Soltys[1], Moritz Peters[1], Dingwen Su[1], Yingguang Frank Chan[1,2]

1 Friedrich Miescher Laboratory of the Max Planck Society, 72076 Tübingen, Germany
2 University of Groningen, Groningen Institute of Evolutionary Life Sciences, 9747 AG Groningen, Netherlands

Evolution of gene regulation plays a critical role in adaptation and can occur through gene regulatory changes acting in *cis* or *trans*, yet how this differs between individual cell types is poorly understood. Here, we applied single-cell multiomics to 63,551 primary liver nuclei in a set of four closely-related mouse species and their F1 hybrids and profile both gene expression and chromatin accessibility simultaneously at single-cell resolution to investigate cell-type specific regulatory changes as well as linking 118,344 putative *cis*-regulatory elements (pCREs) to their target genes. Between the closest related species, 31.8% of regulatory changes occurred solely in *trans* compared to only 14.8% in *cis*, but the proportion of *cis*-regulated genes increases with both increasing evolutionary divergence and expression difference. However, we find considerable differences in the patterns of regulatory evolution between cell types and that some show consistent regulatory changes independent of species. Lastly, we show that linked pCREs are under purifying selection yet those linked to *cis*-regulated genes show increased genetic divergence, consistent with adaptive evolution. This approach therefore dissects regulatory evolution between cell types and not only allows identification of *cis*-regulated genes but also of possible pCREs facilitating the regulatory change.

Introduction

Heritable changes in gene expression contribute to phenotypic differences and adaptation, and can be in either *cis*-regulatory elements or *trans*-regulatory factors[1,2]. Mutations in *cis*-regulatory elements are thought to be an important substrate for adaptive evolution as they tend to be specific to the gene, cell type or timepoint and therefore might be able to precisely alter gene expression in cell types or during developmental processes[2–4]. In contrast, changes in *trans*-regulatory factors[3,5] might lead to pleiotropic, potentially deleterious effects since they usually affect many genes across several cell types. A common approach to assess the contribution of *cis*- and *trans*-regulatory effects on expression changes is by comparing differential expression between parental species to allele-specific expression in their F1 Hybrids[6,7]. Studies employing this design surveyed regulatory evolution across several taxa within and between species[8–11]. Across *Saccharomyces*, *Drosophila* and *Arabidopsis*, these studies repeatedly showed that with increasing evolutionary divergence, *cis*-regulatory changes become pervasive though in closer related species the majority of regulatory differences are mediated in *trans*[8,9,12]. Surprisingly, although mice are one of the most widely used laboratory animals, how regulatory evolution proceeded in this genus has not been investigated comprehensively. Additionally, a common shortcoming of these studies is the assessment of regulatory changes on a tissue level, yet adaptive gene expression changes likely take effect in individual cell types[13,14]. Thus, how regulatory evolution is influenced by individual cell types that may need to adapt in different ways and potentially experience different selective pressures is currently unknown.

Another challenge in understanding regulatory evolution is identifying regulatory elements or even particular genetic variants underlying expression change as linking those together is challenging. While approaches such as expression quantitative locus (eQTL) mapping can correlate sets of variants with gene expression change, with recent studies doing so even within cell types[13,15,16], distinguishing between *cis*- and *trans*-regulatory divergence is defined solely by proximity and they fall short of implicating specific regulatory elements[17]. Additionally, they have limited power in detecting *trans*-eQTLs due to the need to correct for a larger number of statistical tests.

Here we combine a F1 hybrid system in four strains and species across *Mus* with a single-cell multiomic assay measuring both gene expression and chromatin accessibility simultaneously to investigate cell-type specific regulatory evolution in mammals using liver. We find that while increasing *cis*-regulation with increasing evolutionary divergence is a common trend, regulatory changes differ substantially between cell types. Additionally, we argue that some cell types might be biased toward certain regulatory changes regardless of species. We then link putative *cis*-regulatory elements (pCREs) to their target genes by correlating the simultaneous single-cell measurements of gene expression and chromatin accessibility. This approach therefore not only allows to investigate if a gene is differentially regulated in a cell type and by what mechanism (*cis, trans*,..) but also identifies candidate pCREs causing potentially adaptive regulatory change.

Results

Global differences in gene expression and regulatory landscape reflect evolutionary divergence

In order to investigate cell-type specific regulatory evolution, we profiled liver nuclei of several subspecies and species with increasing evolutionary divergence across *Mus* as well as their F1 hybrids using easySHARE-seq (**Fig. 1A**), which we previously showed measures gene expression and chromatin accessibility simultaneously in single cells[18]. Specifically, we used C57BL/6 mice (BL6; mostly *Mus musculus domesticus*) in combination with the wild-derived CAST/EiJ (CAST; *Mus musculus castaneus*), PWD/PhJ (PWD; *Mus musculus musculus*) and SPRET/EiJ strains (SPRET; *Mus spretus*). The CAST, PWD and SPRET strains represent an increasing evolutionary divergence to BL6 mice (**Suppl. Fig. 1A**), with BL6, CAST and PWD originating from different subspecies of *Mus musculus*. For convenience, we will refer to all mice as separate species from now on. Additionally, we included BL6xCAST, BL6xPWD and BL6xSPRET F1 hybrids in the design (BL6 was the dam in all cases), allowing us to disentangle *cis*- from *trans*-regulatory effects as described above.

In total, we recovered 63,551 liver nuclei (after quality control, using the expression data; **Suppl. Fig. 2A-E**). We clustered the nuclei and annotated cell types using expression of previously identified marker genes[19,20] (**Fig. 1B**) and identified a total of 8 distinct cell types with hepatocytes representing the majority of the nuclei (~77%, **Suppl. Fig. 2I**).

We then sought to assess global differences in gene expression and chromatin accessibility between the species. Using principal component analysis (PCA) on the aggregated scRNA- and scATAC-seq data, we found that global transcriptome and chromatin accessibility profiles were primarily separated by species, with F1 hybrids consistently clustering in between the focal species and BL6 (**Fig. 1C**). Next, we identified differentially expressed (DE) genes and differentially accessible (DA) peaks in the aggregated datasets between BL6 and each wild-derived strain (**Fig. 1D**). The proportion of DE genes scaled with evolutionary divergence (25.12% in CAST, 33.4% in PWD and 40.2% in SPRET), as did the proportion of DA peaks (27.1% in CAST, 33.7% in PWD and 44% in SPRET). The same trend was observed when comparing DE genes and DA peaks between F1 hybrid alleles (**Suppl. Fig. 1E**). Notably, PWD mice differed substantially stronger from BL6 than CAST mice, even though they harbour comparable genetic variation[21], perhaps reflecting an increased rate of adaptation in liver of PWD mice.

We next identified DE genes between BL6 and each wild-derived strain within cell types (**Suppl. Fig. 1F**). In each cell type, SPRET had the highest proportion of DE genes and across cell types, hepatocytes were the most diverged. However, because they represent ~77% of all nuclei, increased power in detecting DE genes might exaggerate differences. In all cell types except hepatocytes, the differences between CAST and PWD were far less pronounced compared to the global difference and in some cases (e.g., LSECs & Neurons), CAST mice had a higher proportion of DE genes than PWD mice.

To summarise, single-cell methodologies reveal that differences in both gene expression and regulatory landscape increase with evolutionary divergence across several *Mus* species.

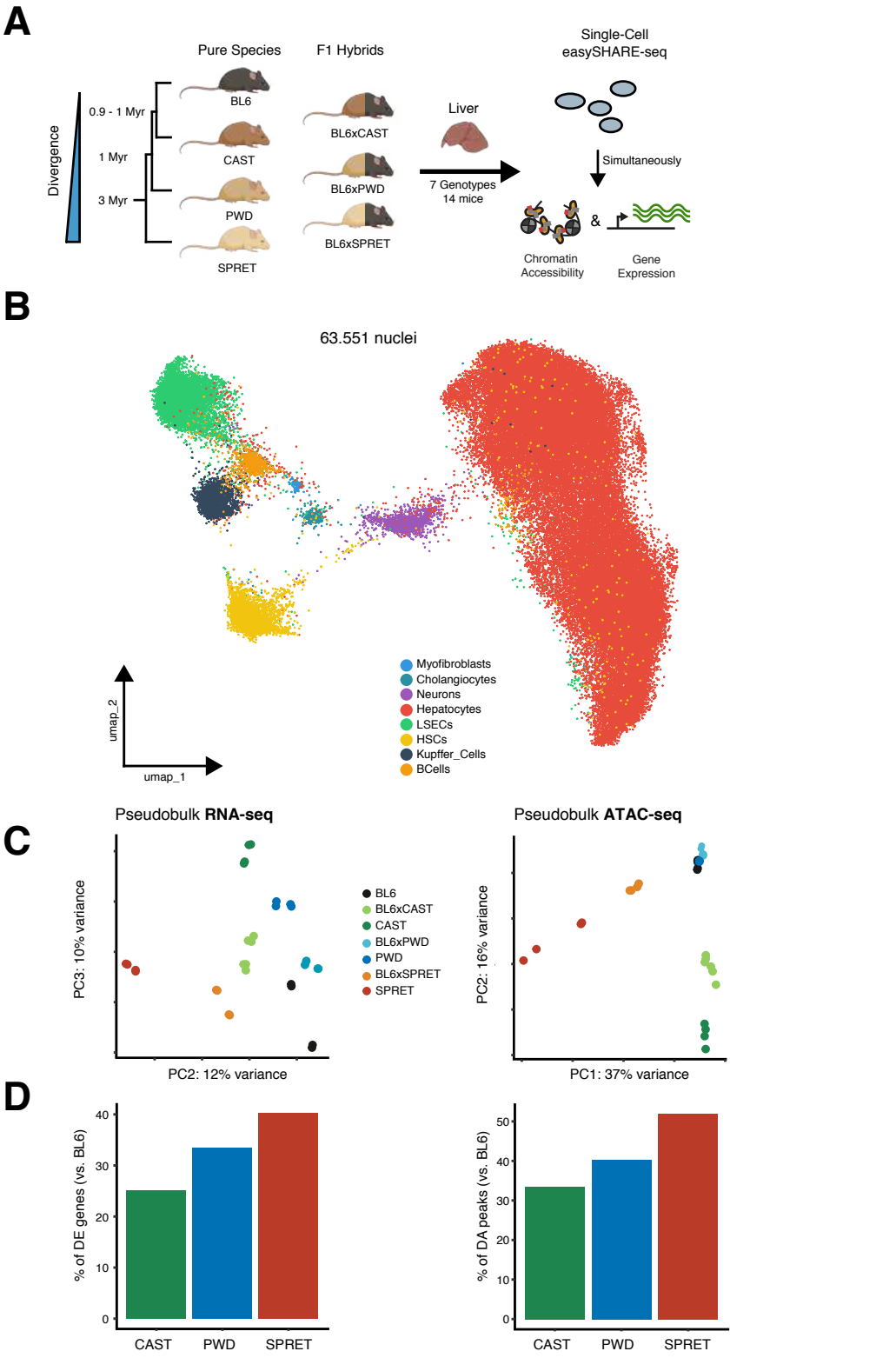


Fig. 1: Single-cell multiomics recovers evolutionary divergences across *Mus*

- **(A)** Overview of the study design and data.
- **(B)** UMAP visualisation of 63,551 liver nuclei using scRNA-seq data. Nuclei are coloured by cell types.
- **(C)** Principal components analysis of pseudobulk scRNA-seq or scATAC-seq modality. Samples are coloured by genotype. Both modalities separate genotypes.
- **(D)** Percentage of differentially expressed genes or differentially accessible peaks between BL6 and CAST, PWD or SPRET. Differences scale with evolutionary divergence.

Cell-type specific regulatory evolution reveals a shift from *trans*- to *cis*- dominance with increasing divergence

To investigate how gene regulation evolves across our species and cell types, we classified genes into regulatory categories by comparing expression differences between the different species (CAST/PWD/SPRET to BL6) and the F1 hybrid alleles. In the F1 hybrid, both alleles are subject to the same *trans* environment and thus are regulated by a common set of *trans* factors (e.g., transcription factors). Therefore, if a difference that was detected between the parental species is not detectable between the F1 hybrid alleles anymore, *trans*-acting changes cause this gene expression difference[7]. However, if any allelic imbalance persists, this must be due to a *cis*-acting variant, which can only interact with loci on its own allele.

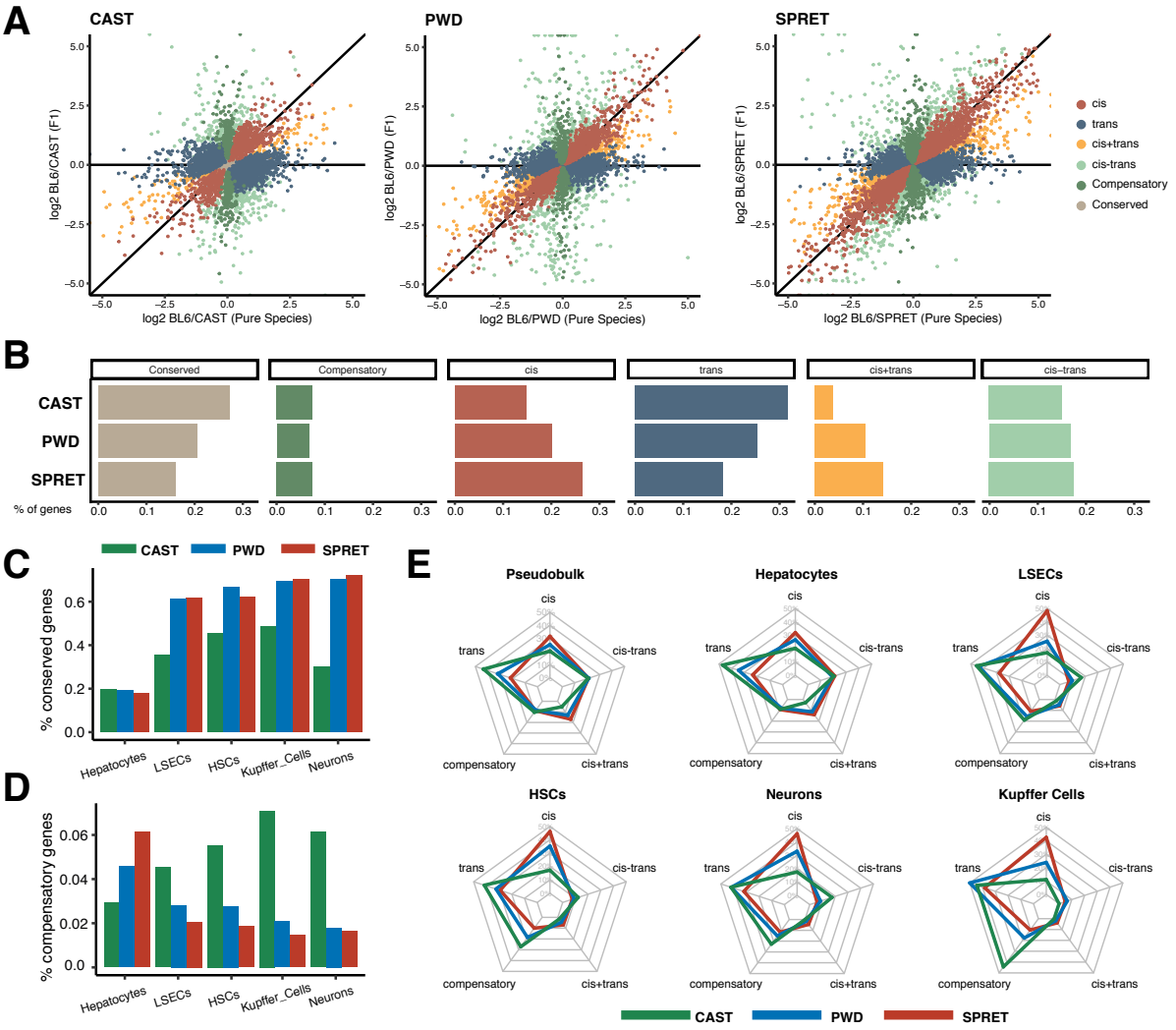


Fig. 2: Cell-type specific regulatory evolution across *Mus*

(A) Scatterplot of pseudobulk log2 fold-change (FC) in expression between the parental species vs. F1 hybrid alleles for BL6 vs. CAST (left), PWD (middle) and SPRET (right). Each dot is a gene. Genes are coloured based on the regulatory category they were assigned to (see Methods). Horizontal line represents 100% *trans* effects, diagonal line 100% *cis* effects.

(B) Frequency of regulatory changes across the different *Mus* species from (**A**). With increasing evolutionary divergence, *cis*-regulatory changes become more frequent whereas *trans*-regulatory changes decrease in frequency.
(C) Percentage of genes whose gene regulation is conserved split by species and cell type. Despite CAST having the least differentially expressed genes (**Fig. 1D**), it has the least conserved gene regulation in all cell types but hepatocytes.
(D) Percentage of genes who have been classified as compensatory split by species and cell type. CAST have the highest frequency of compensatory genes in all cell types but hepatocytes.
(E) Frequency of regulatory changes as in (**B)** calculated by excluding conserved genes for each cell type. Each coloured line corresponds to a different species contrast. Grey lines correspond to different percentages.

We first focused on global trends along increasing evolutionary divergence and aggregated all expression data on a species level. First, we found that the number of genes without evidence for regulatory changes decreased with increasing evolutionary divergence (27.1% in CAST, 20.5% in PWD and 16.1% in SPRET, **Fig. 2A,B**) whereas comparable fractions were classified as compensatory, where opposing *cis*- and *trans*-effects lead to no net difference in gene expression between the species (7.44% in CAST, 6.72% in PWD and 7.45% in SPRET). Between BL6 and CAST, 31.7% of genes purely had *trans*-regulatory changes whereas only 14.8% had regulatory changes solely due to *cis*-acting variation. However, with increasing divergence, the fraction of genes with *trans*-regulatory changes decreased and simultaneously, *cis*-regulatory changes became more dominant (25.5% *trans* & 20.1% *cis* in PWD; 18.2% *trans* & 26.5% *cis* in SPRET). Lastly, both *cis* and *trans*-regulatory changes can act simultaneously on the same gene, either in the same direction (*cis*+*trans*) or in opposing directions (*cis*-*trans*). *Cis*-*trans* effects did not differ substantially between the contrasts (15.15, 16.8% & 17.6% in CAST, PWD & SPRET) whereas *cis*+*trans* changes increased in frequency (3.8%, 10.4% & 14.2% in CAST, PWD & SPRET). Importantly, our results are independent of data filtering or P-value cut off as we obtained similar results when varying those parameters (**Suppl. Fig. 3A,B**). In both PWD and SPRET, genes with *cis*-regulatory changes had a larger effect size (absolute log2 fold-change of expression) than those with *trans*-regulatory changes (PWD: 0.95 mean effect size for *cis*-regulated genes, 0.72 for *trans*; SPRET: 1.11 mean effect size for *cis*-regulated genes, 0.79 in *trans*). Surprisingly, the opposite is true in CAST (0.88 mean effect size for *cis*-regulated genes, 0.96 in *trans,* **Suppl. Fig. 3C**). Lastly, we asked how inheritance patterns change with increasing evolutionary divergence and found that with increasing divergence, genes that are inherited additively increased in frequency (CAST 9.8% of genes, PWD: 18%, SPRET 23.3%; **Suppl. Fig. 3D**).

When species adapt to a new environment, selection pressure can differ between cell types and together with chance might cause each cell type to adapt differently. To therefore investigate how regulatory evolution differs on a cell type level, we leveraged our previously identified cell types (**Fig. 1B**). We excluded cell types with low cell counts and insufficient data (**Suppl. Fig. 4B**) and split genes again into regulatory categories. In all cell types but hepatocytes, CAST had the lowest fraction of genes with conserved gene regulation (**Fig. 2C**) despite having the overall least number of DE genes (**Fig. 1D**). However, two opposing regulatory changes can effectively compensate for one another, leading to detectable regulatory changes without expression changes. In agreement with this, in all cell types but hepatocytes, CAST had the highest fraction of genes with compensatory regulatory changes (**Fig. 2D**).

In order to compare evenly across cell types, we then focused on genes for which we identified regulatory changes (**Fig. 2E**). The fraction of regulatory categories in hepatocytes resembled the aggregated data ('pseudobulk') the most (pearson's r: 0.99, mean fold-change: 1.07), consistent with them being the most abundant cell type. Next, we found that the overall global trends could still be observed (**Suppl. Fig. 4C**). For one, in every cell type the proportion of *cis*-regulatory changes increased with increasing evolutionary divergence and second, CAST had a higher frequency of *trans*-regulatory changes compared to *cis* in all cell types. However, cell types also showed both cell type and genotype specific differences. For example, in Liver Sinusoidal Endothelial Cells (LSECs), SPRET had the highest proportion of *cis*-regulatory changes across all cell types (48.35% of genes with regulatory change, 1.37-fold increase compared to pseudobulk). Also, CAST and PWD nearly had a similar frequency of *trans*-regulatory changes (CAST: 45.6% of gene with regulatory changes, PWD: 43.9%). Notably, in all cell types other than Hepatocytes, the fraction of genes in the '*cis*-*trans*' category was strongly decreased (mean 20.2% of genes in Hepatocytes, mean 9.6% of genes across all other cell types). In general, the fraction of genes regulated in *cis* tended to have the strongest change in SPRET compared to pseudobulk, with a 1.41-fold increase on average (PWD: 1.18-fold, CAST: 0.81-fold) as did *trans*-regulated genes (CAST: 1.03-fold, PWD: 1.32-fold, SPRET: 1.44-fold).
To summarise, between the closest related species regulatory changes occurred mostly in *trans,* but with increasing evolutionary divergence, *cis*-regulation becomes more dominant. Additionally, patterns of regulatory evolution not only differ by species but also by cell type.

The proportion of *cis*-regulated genes increases with expression divergence

We next asked what characteristics of a gene shape how its regulation evolves. We first wondered about a relationship between expression divergence (evolved expression change between the species) and regulatory mode since we found that *cis*-regulatory changes had higher effect sizes than *trans* (absolute log2 fold-change in expression, see above & **Suppl. Fig. 3C**). To do so, we split genes into percentiles of increasing expression divergence and calculated the frequency of types of regulatory changes in each percentile. We found that with increasing expression divergence between the species, genes are more likely to be differentially regulated in *cis* compared to *trans* (**Fig. 3A**). Indeed, the fraction of *trans*-regulated genes dropped steeply among the top 5 percentiles of genes with the highest expression divergence (mean decrease of 13.9% compared to previous 50 percentiles). Both these trends are observed across every species and cell type (**Suppl. Fig. 5A**). This indicates that during regulatory evolution in *Mus*, stronger expression changes are more likely to be mediated by *cis*-regulatory elements, possibly due to increased pleiotropic consequences caused by a *trans* factor.

We also asked if there is a relationship between the regulatory mode and gene expression variance since *cis*- and *trans*-regulatory changes might alter expression with different degrees of precision. To eliminate a possible relationship between total level of expression of a gene and gene expression variance, we applied a variance stabilising transformation before calculating gene expression variance and then split genes into increasing percentiles along it. We found that *cis* and *trans*-regulated genes do not consistently differ in their extent

of gene expression variance (**Suppl. Fig. 5B**). Across species, there is also no relationship between magnitude of expression variance and patterns of regulatory evolution. However, with increasing evolutionary divergence, genes with higher expression variance were less likely to evolve expression changes between the species (**Suppl. Fig. 5C**).

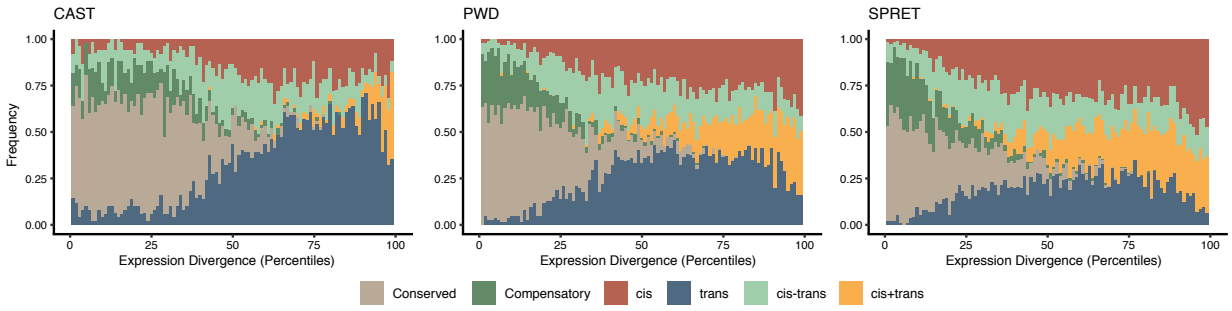


Fig. 3: The proportion of *cis*-regulated genes scales with expression divergence

(A) Frequency of regulatory categories along absolute expression divergence between the species (BL6 vs. CAST/PWD/SPRET). Genes are grouped into percentiles of increasing expression divergence.

Transcription factors are more frequently regulated in *trans*

We next asked if transcription factors (TF) showed different patterns of regulatory evolution compared to all other genes. These regulators tend to be placed more centrally in gene regulatory networks and therefore directly influence a large number of genes[22]. In consequence, a change in *trans* could potentially lead to stronger pleiotropic effects. We define TFs using the gene ontology annotation GO:0003700 ("DNA-binding transcription factor activity"), which defines 2,669 genes as TF, 775 of which are expressed in each species on average (in the pseudobulk data). We then compared the ratio of *cis*- and *trans*-regulatory changes in expressed TFs to all other expressed genes by combining numbers across all cell types. We found that with increasing evolutionary divergence, TFs remain more frequently regulated in *trans* than all other genes (CAST: not significant, PWD: $P < 0.001$, SPRET: $P< 0.001$, Fisher's exact test; **Suppl. Fig. 6A**). This pattern is consistent across all but one cell type (**Suppl. Fig. 6B**). When using different criteria to subset transcription factors, we could confirm this pattern (GO:0010468 ("regulation of gene expression") or curated lists from Zhou et al.[23] or from Hammelman et al.[24]**; Suppl. Fig. 6D**). When subsetting genes into other, unrelated categories (GO:0046907 "Intracellular transport" or GO:0051246 "regulation of protein metabolic process"), we did not find any enrichment in *trans*-regulation (**Suppl. Fig. 6C**).

Altogether, this shows that between our closest related species, TFs are equally likely being *trans*-regulated as other genes. However, when *cis*-regulation gets more dominant with greater evolutionary distances, regulation of TFs changes at a slower rate.

Stronger gene expression changes correlate with increased genetic variation in cell-type specific *cis*-regulatory elements

So far, we showed that with increasing evolutionary divergence and magnitude of expression difference, the proportion of *cis*-regulated genes increases in all cell types. However, we wondered if cell types also show specific evolutionary patterns independent of species and evolutionary divergence. As each cell type fulfils distinct roles within the liver but highly similar roles across species, we reasoned that if a cell type shows bias toward certain patterns of regulatory evolution, they should do so across all species.

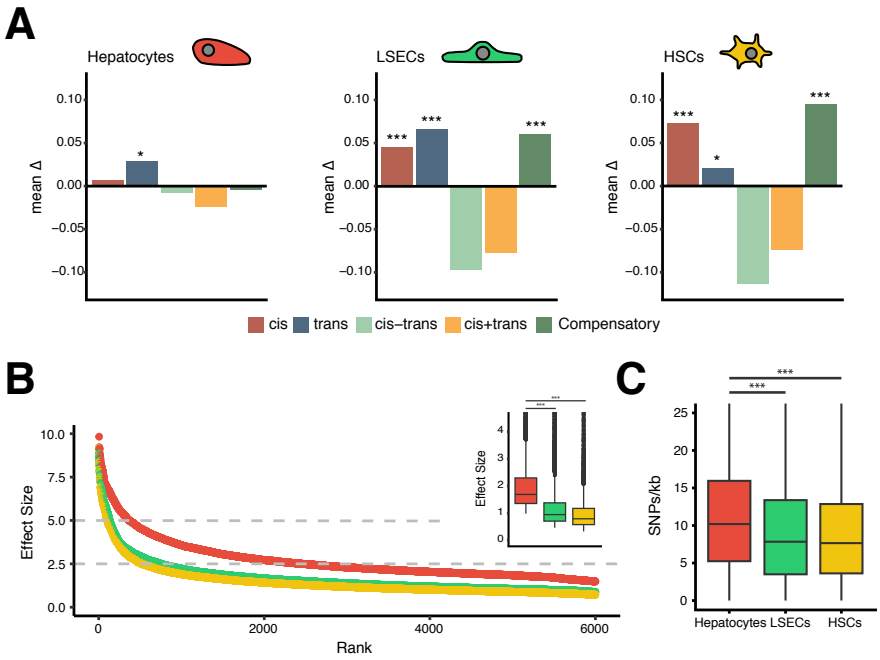


Fig. 4: High gene expression changes correlate with increased genetic variation in cell-type specific *cis*-regulatory elements

- **(A)** Mean percentage difference of regulatory categories within a cell type compared to pseudobulk, averaged across the three species contrasts. Proportion of genes in each regulatory category were compared to pseudbulk using Fisher's exact test (*: $P < 0.05$, ***: $P < 0.001$).
- **(B)** Ranked effect size of top 1,800 genes (absolute log2FC in expression compared to BL6) for each cell type. For each cell type, we selected the top 2,000 genes with the highest effect size in each species contrast (BL6 vs. CAST/PWD/SPRET). Top right: Boxplot of plotted effect size. Tested for significant differences in effect size using Welch's t-test (***: $P < 0.001$). Mean effect sizes: 2.05 (Hepatocytes), 1.21 (LSECs), 1.03 (HSCs).
- **(C)** SNPs/kb in cell type specific ATAC-seq peaks (Hepatocytes 36,660 peaks, LSECs 13,833 peaks, HSCs 7,563 peaks). Mean SNPs/kb: 11.24 (Hepatocytes), 9.19 (LSECs), 8.93 (HSCs), median: 10.19 (Hepatocytes), 7.83 (LSECs), 7.65 (HSCs). Tested for differences in using Welch's t-test (***: $P < 0.001$).

First, we asked if we can detect consistent differences in how gene regulation evolves for each cell type (**Fig. 4A**). We found that both LSECs and Hepatic Stellate Cells (HSCs) consistently have a higher proportion of *cis*-regulated genes (+4.6% & +7.2%, Fisher's exact test: $P < 0.001$) as well as compensatory genes (+6.1% & 9.4%, Fisher's exact test: $P < 0.001$) compared to pseudobulk. Hepatocytes did only differ slightly from pseudobulk, in line with being the most abundant cell type, with the exception of a subtle increase of *trans*-regulated

genes (+2.8%, $P < 0.05$), which is also detected in the other two cell types (LSECs +6.7%, HSCs +2%, $P < 0.001$). To summarise, cell types exhibit specific and consistent patterns of regulatory evolution, even across differing evolutionary divergence.

Next, we wondered to what extent cell-type identity can shape specific regulatory changes independent of species. More precisely, given the higher frequency of *cis*-regulatory changes with higher magnitudes of expression divergence (**Fig. 3**), we reasoned that cell types which consistently evolve stronger gene expression changes might show higher rates of genetic variation in their *cis*-regulatory elements.

To examine this, we calculated the effect size for differentially expressed genes in each cell type in each species (absolute log2 fold-change in expression between BL6 and CAST/PWD/SPRET). We then combined the top 2,000 most differentially expressed genes in each species per cell type. We found that differentially expressed genes in hepatocytes had a substantially higher effect size compared to the other cell types ($P > 0.001$, two-tailed t-test; mean effect size 2.05 ∓ 1.21 standard deviation in Hepatocytes vs. 1.21 ∓ 0.86 in LSECs & 1.03 ∓ 0.81 in HSCs; **Fig. 4B**). This finding was consistent independently for each species and independent of total expression level of each cell type (**Suppl. Fig. 7A,C**). We then identified ATAC-seq peaks that are consistently cell-type specific in all three species and assessed their rate of genetic variation (in SNPs/kb, **Fig. 4C**). Surprisingly, we found that hepatocyte-specific peaks have a strongly increased rate of genetic variation compared to the other cell types (+24.1%, t-test: $P < 0.001$). This observation was again confirmed independently for each species (**Suppl. Fig. 7B**). This shows that hepatocytes consistently evolved a higher fraction of DE genes (**Suppl. Fig. 1G**) and stronger expression changes than other cell types, which in turn correlates with consistently higher rates of genetic variation in their *cis*-regulatory elements.

Taken together, this could point to hepatocytes having experienced increased selective pressure compared to the other cell types, which resulted in a higher frequency of possibly adaptive genetic variation in *cis*.

Linking putative *cis*-regulatory elements to their target genes

We next sought to leverage our simultaneous measurements of gene expression and chromatin accessibility and link putative *cis*-regulatory elements (pCREs) to their target gene. While using F1 hybrids is effective in identifying if a *cis*-acting variant differentially regulates a gene, determining which regulatory element most likely functionally evolved is difficult. However, as we measure gene expression and chromatin accessibility simultaneously within the same cell, this enables us to directly test for correlations between increased gene expression and chromatin accessibility at a focal peak (**Fig. 5A**) and potentially identify a set of pCREs that changed during *cis*-regulatory evolution of each species and investigate their properties.

To achieve this, we first integrated scATAC- and scRNA-seq modalities using Weighted-Nearest-Neighbor Analysis[25]. This resulted in matched profiles of gene expression and chromatin accessibility for 53,257 nuclei (**Suppl. Fig. 2H**). In order to calculate the peak-gene links, we followed the analytical framework from Ma et al.[26], which additionally controls for GC content and accessibility strength, and calculated links within +- 500kb of each transcription start site (TSS) for each species in each cell type. We identified a total of 118,344

links (34.9% of total ATAC-seq peaks are linked, P < 0.05, FDR = 0.1. From here on, 'pCRE' refers to a linked ATAC-seq peak), ranging from 11,760 in CAST to 41,971 in BL6, both summed across all cell types (**Suppl. Fig. 8A**). When comparing between cell types, we found that LSECs had on average the most links per gene (P < 0.001 using two-tailed t-test; 4.39 vs. 3.89 in Hepatocytes, **Fig. 5B**), possibly indicating a more active regulatory landscape. We then ranked genes based on their number of linked pCREs by combining links across all species and cell types (**Fig. 5C**). Among the top 100 genes with the most links were many housekeeping genes (e.g. *Atp5b*, *Atp5d, Rnh1, Rexo1* or *Rps15)*, general regulators of transcription (e.g. *Ell*, *Eef2* or *Atf5*) or genes involved in other core cell functions (e.g. *Map2k2*), since most of these genes are important to a cells' function independent of its cell type identity. When assessing links per gene separately for each cell type (**Suppl. Fig. 8B**), we additionally found many cell-type specific genes and regulators (e.g. *Cyp2e1*, *Alb* (**Fig. 5D**), *Mst1*[27] in Hepatocytes, *Stat3* in LSECs[28]).

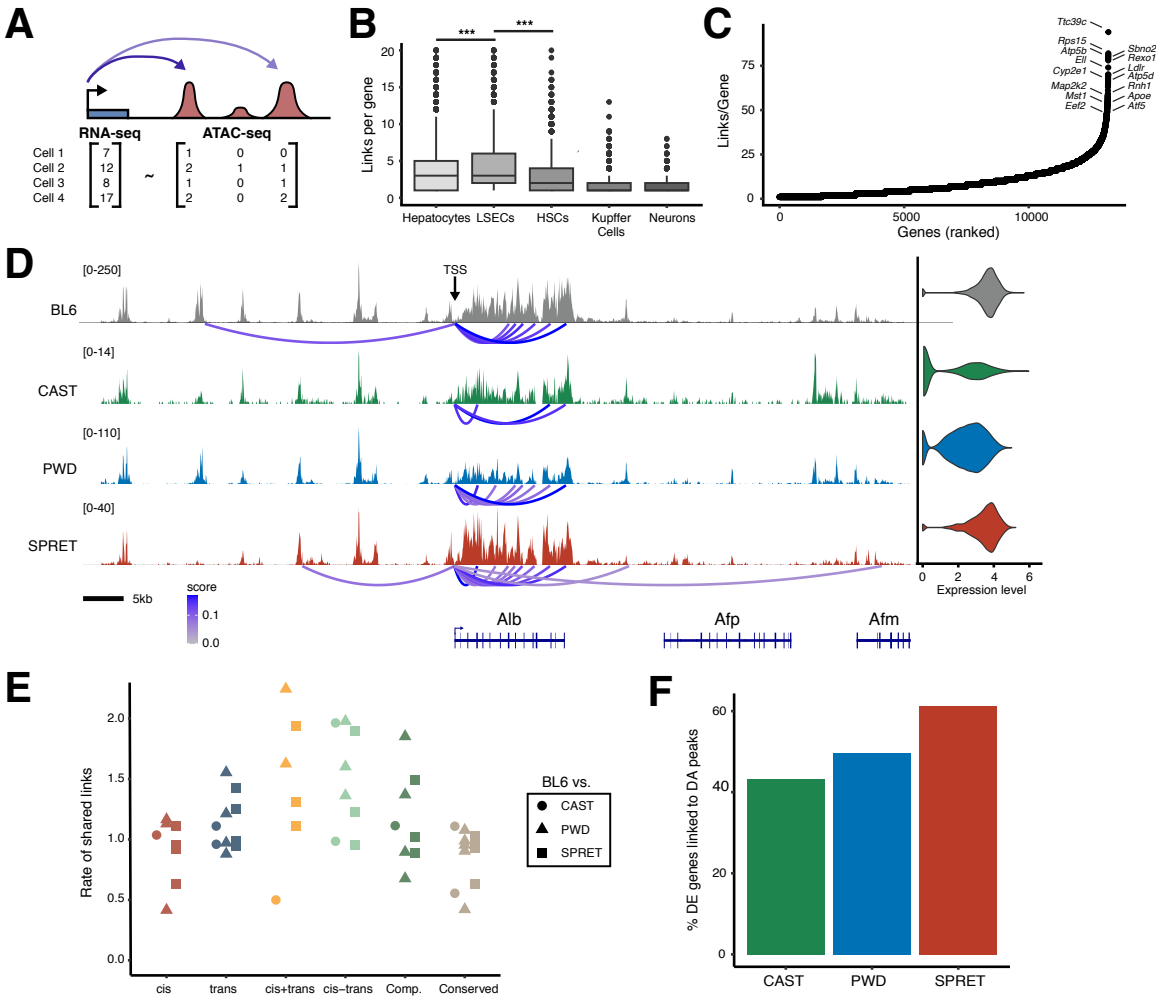


Fig. 5: Leveraging multiomic measurements to link pCREs to their target genes at scale

(A) Schematic depicting the conceptual framework for linking pCREs to their target genes
(B) Mean links per gene for each cell type. In LSECs, genes have significantly more links on average than in the other cell types (two-tailed Welch's t.test, P < 0.001).

(C) Genes ranked by combined links per gene across all species and cell types. Several important regulators or housekeeping genes are highlighted.
(D) Aggregated ATAC-seq data for each species from hepatocytes at the *Alb* locus. Loops denote identified links to *Alb*, their colour signifies z-score (see Methods). Right: Violin plot of *Alb* expression per hepatocyte.
(E) Rate of shared links across species and cell types, split by the regulatory mode of the genes they are linked to. Each data point denotes a cell type, shape signifies the different species. Only cell types with more than 200 identified links were plotted. For even comparison, rates are normalised by their background expectation (see Methods or main text).
(F) Percentage of DE genes linked to DA peaks for each species.

To summarise, we leveraged our single-cell multiomic measurements to identify links between pCREs and target genes at scale in each species within each cell type. Genes with the most links were either essential to a cell's function or key cell-type specific regulators. As such, their expression needs to be strictly regulated and having a high number of links likely reflects increased regulatory activity.

Links to *cis*-regulated genes are the least shared across species

Having identified links between regulatory elements and genes, we next asked if these links might capture pCREs that facilitated expression change during the regulatory evolution of each species.
First, if we capture true biological connections to 'causal' CREs, links to *cis*-regulated genes should be less shared between species as a CRE for example becomes non-functional in one species. To assess this, we compared per cell type how many links identified in BL6 are still identified in the other species. We then separated the links based on the regulatory category of their target genes, calculated their frequency and normalised it by the initial frequency (for example, if 50% of links in BL6 are to genes that are *trans*-regulated between BL6 and CAST, the expectation would be that 50% of shared links between these species have a *trans*-regulated target gene). We found that links to *cis*-regulated genes are less shared between BL6 and the other species (**Fig. 5E**). For example, in LSECs links to *cis*-regulated genes were on average 22% less likely to be shared than those to *trans*-regulated genes (1.06 *cis* vs. 1.36 *trans*). We found the same trend in hepatocytes (0.88 *cis* vs. 0.93 *trans)*, HSCs (0.79 *cis* vs. 0.97 *trans*) and Kupffer Cells (1.11 *cis* vs. 1.23 *trans*). Next, we tested how often DA peaks are linked to DE genes and found that with increasing evolutionary divergence and thus increasing *cis*-regulation (**Fig. 2**), the percentage of DE genes being linked to DA peaks increases (**Fig. 5F**, CAST: 40.8%, PWD:45.7%, SPRET: 55.9%, links from all cell types combined).
Altogether, this suggests that the identified links between pCREs and target genes to some extent capture pCREs that facilitated expression change during *cis*-regulatory evolution of each species.

pCREs linked to *cis*-regulated genes show signatures of adaptive evolution

Lastly, we reasoned that if we identify true biological connections, we might detect signatures of selection in our identified pCREs. We first compared the overall rate of genetic variation

(SNPs per kb) in different genomic features (**Fig. 6A,** compared to BL6). We found that compared to the genomic background, ATAC-seq peaks have lower rates of genetic variation in all three species ($P < 0.001$ in all comparisons, two-tailed t-test, mean decrease: CAST -5.5%, PWD -7.4%, SPRET-5.1%). However, pCREs have even further decreased genetic variation ($P < 0.001$, two-tailed t-test; mean decrease: CAST 14.5%, PWD 18.1%, SPRET 11.9%). This indicates for one, that even compared to ATAC-seq peaks, pCREs might be highly enriched for functional elements as sequence evolution is more constrained. Second, pCREs are likely under purifying selection.

Next, we reasoned that if pCREs are highly enriched for functional elements that facilitated expression changes during *cis*-regulatory evolution, we might be able to detect signatures of adaptive evolution in their genomic sequence. To do so, we separated pCREs based on the regulatory category of their target genes and again assessed their rates of genetic variation (**Fig. 6B**). This revealed that pCREs linked to *cis*-regulated genes have higher rates of genetic variation than those linked to *trans*-regulated genes (CAST: + 9.65%, $P < 0.05$, 8.02 SNPs/kb for *cis* vs. 7.32 SNPs/kb for *trans*-regulated genes; PWD: +3.37%, $P = 0.06$, 6.74 SNPs/kb vs.6.52 SNPs/kb, SPRET: +2.1%, $P < 0.05$, 14.3 SNPs/kb vs. 14.0 SNPs/kb, two-tailed t-test in all cases). This is consistent across all but one cell type (**Suppl. Fig. 9A**).

Taken together, exploiting simultaneous measurements of gene expression and chromatin accessibility in single cells to link regulatory elements to their target genes most likely captures functional relationships as well as pCREs that facilitated expression changes during regulatory evolution. In general, these pCREs are under purifying selection. However, those linked to *cis*-regulated genes show an increased rate of genetic variation, a signature of adaptive evolution. This shows that the combination of study design and methodology not only allows to determine if a gene is differentially regulated in *cis* but also captures which regulatory elements are likely candidates causing the differential gene expression.

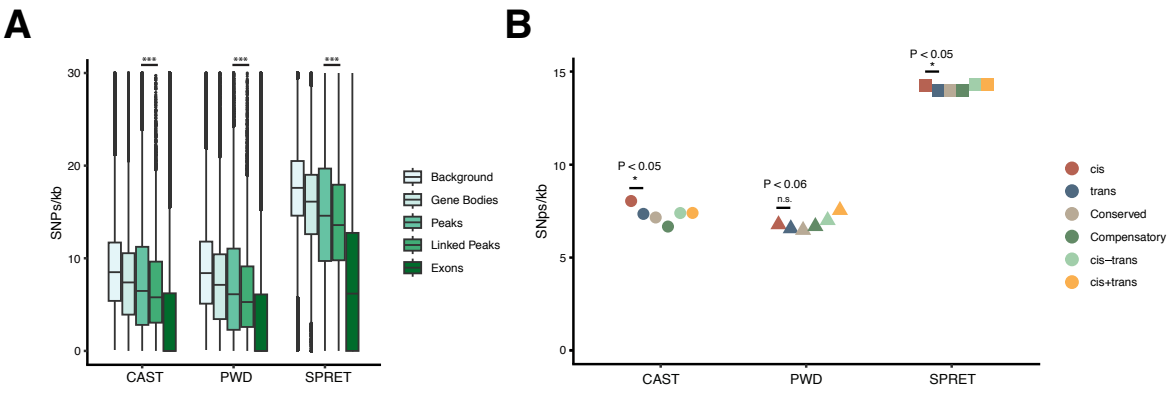


Fig. 6: Genetic variation is higher in pCREs linked to *cis*-regulated genes compared to *trans*-linked pCREs

(A) Rates of genetic variation in CAST/PWD/SPRET compared to BL6 in either the genomic background, gene bodies, ATAC-seq peaks, pCREs or exons. pCREs consistently have lower rates of genomic variation than non-linked ATAC-seq peaks (two-tailed Welch's t.test, $P < 0.001$). All pCREs across different cell types were combined per species.

(B) Genetic variation in pCREs split by regulatory mode of their target gene. pCREs linked to *cis*-regulated genes have a higher rate of variation than those linked to *trans*-regulated genes (CAST & SPRET: $P < 0.05$, PWD: $P = 0.06$; one-tailed Welch's t.test).

Discussion

When species adapt to new environments, gene regulation changes individually for each cell type. To understand cell-type specific evolution of gene regulation across *Mus*, we combine single-cell transcriptomics and epigenomics with a classical F1 hybrid design across several species. We find that with increasing evolutionary divergence, *cis*-acting changes become more dominant but we also uncovered cell-type specific regulatory patterns. Our experimental strategy further enabled us to link regulatory elements to their target genes for each cell type and species, finding that regulatory elements linked to *cis*-regulated genes show signatures of adaptive evolution.

Our study is the first to survey evolution of gene regulation across several species in a mammalian genus. We identify pervasive *cis*-regulation with increasing evolutionary divergence, a pattern consistent across several previous studies[8–10,29,30] . However, between closer related species, the majority of expression differences are mediated by changes in *trans*. This stands in direct contrast with previous results involving similar or identical mouse strains[31,32]. For example, a 2012 study between BL6 and CAST in the liver concluded that only around 2% of genes are differentially regulated in *trans*. Yet, studies across similar evolutionary divergences in different organisms are consistent with our results[8,9]. We additionally found that gene expression and chromatin accessibility in PWD was substantially more diverged from BL6 than CAST, even though they harbour comparable genetic variation[21]. They also differed in how differential gene regulation evolved.
Altogether, this leads us to conclude, like others[8,33,34], that while *cis*-regulatory differences will become pervasive with increasing divergence, factors such as tissue combined with evolutionary as well as demographic history of a species can have strong individual effects and result in differences in how gene regulation evolves.

Our study also is the first to investigate evolution of gene regulation on a cell type level using F1 hybrids. This allowed us to confirm that increasing *cis*-regulation with both increasing divergence and expression difference is a common trend throughout all cell types that we investigated. We also found that cell-type specific differences in how gene regulation evolves are common and mostly influenced by species. However, we also identified some cell types that show consistent patterns in their regulatory evolution, regardless of species. In our case, hepatocytes which are fulfilling the main metabolic function of the liver, consistently had the highest proportion of DE genes and highest gene expression changes. These were possibly facilitated by a higher frequency of *cis*-regulatory adaptive changes, as evidenced by an increased rate of genetic variation in their potential regulatory elements compared to other cell types and suggested by the trends described above.
This could imply two possible explanations. First, some cell types are predisposed toward certain types of regulatory change by their function and hierarchy within a tissue. For example, LSECs tend to have a more controlling function by maintaining immune homeostasis or by maintaining hepatic stellate cell quiescence[35,36]. Thus, they might disproportionally express genes toward the top of signalling cascades. Hepatocytes in turn fulfil the metabolic functions and thus express many metabolic enzymes, which might be toward the bottom of signalling cascades[37].

Second, differing strengths and durations of selective pressures might cause different regulatory responses and in these particular species, hepatocytes simply happened to experience similar selective pressures. These dynamics have already been predicted in Stern & Orgogozo[2] in their landmark paper on *cis*-regulatory evolution. Cell types of the liver might be particularly subject to differing selective pressures. The liver is responsible for metabolic homeostasis as well as processing toxins, processes which change significantly when changing the environment or diet (and thus are likely under selection)[38], yet not all cell types are involved in these processes. Exemplifying the strong selection pressures that can act upon the liver, a study showed *Mus musculus* acquired an adaptive allele conferring rodenticide resistance from *Mus spretus* via introgression[39], This allele was adaptive due to polymorphisms in *Vkorc1*, which encodes for an enzyme that is a crucial part of the vitamin K cycle which is needed for coagulation factors, which are produced in the liver[40].
However, without further analysis of more tissues across a wider range of organisms, we caution against overinterpreting these results.

Using simultaneous single-cell transcriptomics and epigenomics to link regulatory elements is a powerful approach in the context of evolutionary studies. Our results demonstrate that links between putative regulatory elements and their target genes are capturing *cis*-regulatory elements that confer differential gene expression across species. Combining this approach with a F1 hybrid design, this thus allows not only to determine which genes are differentially regulated in *cis* but also provides a narrow starting point for the identification of causal regulatory elements and variants, which traditionally is challenging. However, it is important to note several limitations.
First, while single-cell sequencing provides cellular resolution, here the trade-off is statistical power, leading to the need for high sample sizes, especially in rare cell types. Second, the list of pCREs for a gene is likely not exhaustive and external factors such as data quality and relative abundance of a cell type do influence the number of identified links. Lastly, with this approach it is currently difficult to detect pCREs mediating weaker expression changes as these links do not become non-functional between species. In the future though, we expect this approach to become even more powerful with further increases in precision, higher sample sizes and more advanced linking frameworks.

Collectively, our results provide a comprehensive survey of cell-type specific regulatory evolution across *Mus*. We identify *cis*-regulatory changes as the dominant driver of gene expression changes and elucidate how cell-type specific adaptations are driven by both cell type identity and underlying species.

Methods

Mice

All animal experimental procedures were carried out under the licence number EB 01-21M at Friedrich Miescher Laboratory of the Max Planck Society in Tübingen, Germany. The procedures were reviewed and approved by the Regierungspräsidium Tübingen, Germany. Liver was collected from both male and female wild-type C57BL/6NCrl, CAST/EiJ, SPRET/EiJ and PWD/PhJ mice as well as from C57BL/6NCrlxCAST/EiJ, C57BL/6NCrlxSPRET/EiJ, and C57BL/6NCrlxPWD/PhJ F1 Hybrids, all aged between 9 to 11 weeks. In the case of F1 Hybrids, the dam was always a C57BL/6NCrl mouse. All samples were collected between 6:30-7AM on the day of the experiment to minimise the influence of circadian effects.

Study design

We generated easySHARE-seq libraries for one male and one female mouse from each genotype (seven genotypes, fourteen mice). From each individual, we sequenced two sub-libraries, the only exception being C57BL/6NCrlxCAST/EiJ F1 Hybrids, from which we sequenced three, resulting in 30 total easySHAREseq libraries,

Liver Nuclei

The liver was extracted, rinsed in HBSS, cut into small pieces, frozen in liquid nitrogen and stored in the freezer at -80 °C for a maximum of two weeks. On the day of the experiment, 1 ml of ice cold Lysis Solution (0.1% Triton-X 100, 1mM DTT, 10mMM Tris-HCl pH8, 0.1mM EDTA, 3mM $Mg(Ac)_2$, 3mM $CaCl_2$ and 0.32M sucrose) was added to the tube. The cell suspension was transferred to a pre-cooled Douncer and dounced 10x using Pestle A (loose) and 15x using Pestle B (tight). The solution was added to a thick wall ultracentrifuge tube on ice and topped up with 4ml ice cold Lysis Solution. Then 9 ml of Sucrose solution (10mM Tris-HCl pH8.0, 3mM $Mg(Ac)_2$, 3mM DTT, 1.8M sucrose) was carefully pipetted to the bottom of the tube to create a sucrose cushion. Samples were spun in a pre-cooled ultracentrifuge with a SW-28 rotor at 24,400rpm for 1.5 hours at 4 °C. Afterwards, all supernatant was carefully aspirated so as not to dislodge the pellet at the bottom and 1 ml ice cold DEPC-treated water supplemented with 10µl SUPERase & 15µl Recombinant RNase Inhibitor was added. Without resuspending, the tube was kept on ice for 20 min. The pellet was then resuspended by pipetting ~15 times slowly up and down followed by a 40 µm straining step. Counting of the nuclei using DAPI and the Evos Countess II was immediately followed up by fixation.

Fixation

One million liver nuclei were added to ice-cold PBS for 4 ml total. After mixing, 87 µl 16% formaldehyde solution (0.35%) was added and the suspension was mixed by pipetting up and down exactly 3 times with a P1000 pipette set to 700 µl. The suspension was incubated at room temperature for 10 min. Fixation was stopped by adding ice-cold Stop-Mix (224 µl 2.5M glycine, 200 µl 1M Tris-HCl pH8.0, 53 µl 7.5% BSA in PBS). The suspension was mixed exactly 3 times with a P1000 pipette set to 850 µl and incubated on ice for 3 min followed by a centrifugation at 500G for 5 min at 4°C. Supernatant was removed and the pellet was resuspended in 1 ml Nuclei Isolation Buffer (NIB; 10mM Tris pH8.0, 10mM NaCl, 2mM $MgCl_2$, 0.1% NP-40) and kept on ice for 3 min followed by straining the suspension with a 40 µm

strainer. It was then spun down at 500G for 3 min at 4°C and re-suspended in ~100-200µl PBSi (1x PBS + 0.4 U/µl Recombinant RNaseInhibitor, 0.04% BSA, 0.2 U/µl SUPERase, freshly added), depending on the amount of input nuclei. Nuclei were then counted using DAPI and the Countess II and concentration was adjusted to 2M nuclei/ml using PBSi.

easySHARE-seq

EasySHARE-seq was performed as previously described [REF]. In short, per mouse 9 reactions of 10,000 nuclei each were tagmented. For each tagmentation reaction, 5 µl of 5X TAPS-Buffer, 0.25µl 10% Tween, 0.25µl 1% Digitonin, 3 µl PBS, 1 µl Recombinant RNaseInhibitor and 9µl of H2O was mixed. TAPS Buffer was made by first making a 1M TAPS stock solution in H_2O, followed by adjustment of the pH to 8.5 by titrating 10M NaOH. Then, 4.25ml H_2O, 500µl 1M TAPS pH8.5, 250µl 1M $MgCl_2$ and 5ml N-N-Di-Methyl-Formamide (DMF) was mixed on ice and in order. Then, 5 µl of nuclei suspension at 2M nuclei/ml in PBSi was added to the tagmentation mix for each reaction, mixed thoroughly and finally 1.5µl of Tn5 (produced in-house as previously described [PICELLI]) loaded with a custom adapter was added (for all oligo and adapter sequences, see **Suppl. Table 1**). The reactions were incubated on a shaker at 37°C for 30 min at 850 rpm. Afterwards, all reactions were pooled on ice. The suspension was then spun down at 500G for 3 min at 4°C. Supernatant was aspirated and the nuclei were washed with 200µl NIB followed by another centrifugation at 500G for 3 min at 4°C.

Three tagmentation reactions were then combined into one Reverse Transcription (RT) reaction for a total of three Rt reactions. The Master Mix for one RT reaction contained 3µl 100µM RT-primer (custom), 2µl 10mM dNTPs, 6µl 5X MaximaH RT Buffer, 4.5µl 50% PEG6000, 1.5 µl H_2O, 1.5µl SUPERase and 1.66µl MaximaH RT. The nuclei suspension was resuspended in 10µl NIB per RT reaction and added to the Master Mix for a total of 30µl and pipetted ~30 times up and down to ensure proper mixing. The RT reaction was performed in a PCR cycler with the following protocol: 52°C for 12min; then 2 cycles of 8°C for 12s, 15°C for 45s, 20°C for 45s, 30°C for 30s, 42°C for 2min and 50°C for 3 min. Finally, the reaction was incubated at 52°C for 5 more minutes. All reactions were then pooled on ice. The suspension was spun down at 500G for 3 min at 4°C, supernatant was aspirated and the nuclei were washed in 150µl NIB and spun down again at 500G for 3min at 4°C. This washing step was repeated once more, followed by resuspension of the nuclei in 2ml Ligation Mix (400µl 10x T4-Buffer, 40µl 10% Tween-20, 1460µl Annealing Buffer (10mM Tris pH8.0, 1mM EDTA, 30mM KCl) and 100µl T4 DNA Ligase, added last).

We then performed single-cell barcoding, which consists of two sequential rounds of ligation with 192 pre-aliquoted barcodes (BC; 2x 96-well plates) in each round (For a detailed description, see [REF]).

10µl of nuclei suspension in the ligation mix was added to each well of the two annealed Round1 BC plates. The plates were then sealed and incubated on a shaker at 25°C for 30 min at 350 rpm. Afterwards, all nuclei were pooled into a 5ml tube on ice. The nuclei suspension was then spun down for 3min at 500G at 4°C. Supernatant was aspirated and the nuclei were resuspended thoroughly in 2ml new Ligation Mix. Now, 10µl of nuclei suspension was added into each well of the Round2 BC plates and incubated on a shaker at 25°C for 45 min at 350 rpm. The nuclei were then pooled into a 15ml Tube and spun down at 500G for 3 min at 4°C. Supernatant was aspirated, the nuclei were washed with 150µl NIB

and spun down again. Finally, the nuclei were resuspended in ~60µl NIB and counted. Sub-libraries of 3,500 nuclei were made and the volume was adjusted to 25µl by addition of NIB. To each sub-library of 3,500 nuclei, 30µl 2x Reverse Crosslinking (RC) Buffer (0.4% SDS, 100mM NaCl, 100mM Tris pH8.0) as well as 5µl ProteinaseK was added. The sub-libraries were mixed and incubated on a shaker at 62°C for one hour at 800 rpm. Afterwards, they were transferred to a PCR cycler into a deep well module set to 62°C (lid to 80°C) for an additional hour. Lastly, each sub-library was incubated at 80°C for 10 min and 5µl of 10% Tween-20 to quench the SDS and 35µl of NIB was added for a total volume of 100µl.
Each transcript contains a biotin molecule which is now used to separate the scATAC-seq libraries from the scRNA-seq libraries. For each sublibrary, 50µl M280 Streptavidin beads were washed three times with 100µl B&W Buffer (5mM Tris pH8.0, 1M NaCl, 0.5mM EDTA) supplemented with 0.05% Tween-20, using a magnetic stand. Afterwards, the beads were resuspended in 100µl 2x B&W Buffer and added to the sublibrary, which were then shaken at 25°C for one hour at 900 rpm.

scATAC-seq library preparation

The supernatant from each sub-library was cleaned up with a Qiagen MinElute Kit and eluted twice into 30µl 10mM Tris pH8.0 total. PCR Mix containing 10µl 5X Q5 Reaction Buffer, 1µl 10mM dNTPs, 2µl 10µM i7-TruSeq-long primer, 2µl 10µM Nextera N5XX Indexing primer, 4.5µl H2O and 0.5µl Q5 Polymerase was added. Importantly, in order to distinguish the samples, each sub-library needs to be indexed with a different N5XX Indexing primer. The fragments were amplified with the following protocol: 72°C for 6 min, 98°C for 1 min, then cycles of 98°C for 10s, 66°C for 20s and 72°C for 45s followed by a final incubation at 72°C for 2 min. The reactions were then cleaned up with custom size selection beads with 0.55X as upper cutoff and 1.4X as lower cutoff and eluted into 25µl 10mM Tris pH8.0. Libraries were quantified using the Qubit HS dsDNA Quantification Kit and run on the Agilent 2100 bioanalyzer with a High Sensitivity DNA Kit.

cDNA & scRNA-seq library preparation

The beads containing the cDNA molecules were washed three times with 200µl B&W Buffer supplemented with 0.05% Tween-20 before being resuspended in 100µl 10mM Tris pH8.0 and transferred into a new PCR strip. The beads were then resuspended in 50µl Template Switch Reaction Mix: 10µl 5X MaximaH RT Buffer, 2µl 100µM TS-oligo, 5µl 10mM dNTPs, 3µl Enzymatics RNaseIn, 15µl 50% PEG6000, 14µl H2O and 1.25µl MaximaH RT. The sample was mixed well and incubated at 25°C for 30 min followed by an incubation at 42°C for 90 min. The beads were then washed with 100µl 10mM Tris while the strip was on a magnet and resuspended in 60µl H_2O. To each well, 40µl PCR Mix was added containing 20µl 5X Q5 Reaction Buffer, 4µl 10µM i7-Tru-Seq-long primer, 4µl 10µM Nextera N5XX Indexing primer, 2µl 10mM dNTPs, 9µl H2O and 2µl Q5 Polymerase. The PCR involved initial incubation at 98°C for 1 min followed by PCR cycles of 98°C for 10s, 66°C for 20s and 72°C for 3 min with a final incubation at 72°C for 5 min. Importantly, in order to distinguish the samples, each sub-library needs to be indexed with a different N5XX Indexing primer.
The PCR reactions were cleaned up with custom size selection beads using 0.7X as a lower cutoff (70µl) and eluted into 25µl 10mM Tris pH8.0. The cDNA libraries were quantified using the Qubit HS dsDNA Quantification Kit.

As the cDNA molecules are too long for sequencing (mean length > 700bp), they need to be shortened on one side. To achieve this, 25ng of each cDNA library was transferred to a new strip and volume was adjusted to 20µl using H_2O. Then, 5µl 5X TAPS Buffer and 0.8µl Tn5 loaded with exclusively one sequencing adapter was added and the sample was incubated at 55°C for 10 min. To stop the reaction, 25µl 1% SDS was added followed by another incubation at 55°C for 10 min. The sample was then cleaned up with custom size selection beads using a ratio of 1.3X and eluted into 30µl. Then 20µl PCR mix was added containing 10µl 5X Q5 reaction buffer, 1µl 10mM dNTPs, 2µl 10µM i7-Tru-Seq-long primer, 2µl 10µM Nextera N5XX Indexing primer (note: each sample needs to receive the **same** index primer as was used in the cDNA library preparation), 4.5µl H_2O and 0.5µl Q5 Polymerase. The PCR reaction was carried out with the following protocol: 72°C for 6 min, 98°C for 1 min, followed by 5 cycles of 98°C for 10s, 66°C for 20s and 72°C for 45s with a final incubation at 72°C for 2 min. Libraries were purified using custom size selection beads with a ratio of 0.5X as an upper cutoff and 0.8X as a lower cutoff. The final scRNA-seq libraries were quantified using the Qubit HS dsDNA Quantification Kit and run on the Agilent 2100 bioanalyzer with a High Sensitivity DNA Kit.

Sequencing

ScATAC-seq and scRNA-seq libraries were sequenced simultaneously as they were indexed with different Index2 indices (N5XX). All libraries were sequenced on the Nova-seq 6000 platform (Illumina) using S4 2x150bp v1.5 kits (Read 1: 150 cycles, Index 1: 17 cycles, Index 2: 8 cycles, Read 2: 150 cycles).

Analysis

Gene annotations and Genomic variants

The reference genome and the Ensembl gene annotation of the C57BL/6J genome (mm10) were downloaded from Ensembl (Version GRCm38, release 102). Gene annotations for PWK/PhJ, SPRET/EiJ and CAST/EiJ mice were downloaded from Ensembl. VCF files containing SNPs and InDels of PWK/PhJ, SPRET/EiJ and CAST/EiJ mice compared to mm10 were downloaded from the Mouse Genomes Project website (www.mousegenomes.org). A consensus GTF in mm10 coordinates was constructed by filtering for genes present across all gene annotations.

Since there is no available PWD/PhJ SNP or InDel set, we started with the PWK/PhJ variant files. PWD/PhJ and PWK/PhJ mice are highly similar wild-derived mouse strains from the *Mus musculus musculus* subspecies and originated in 1972 from a pair of wild caught mice trapped in the central part of Czechia[41]. We pooled all our available PWD/PhJ data (scATAC- & scRNA-seq) and filtered the PWK/PhJ VCF file for variants also detected in our dataset, to which we will from hereon refer to as PWD/PhJ variants. Over 85% of variants with coverage were also detectable in our dataset, highlighting the similarity between the two mouse strains.

easySHARE-RNA-seq pre-processing

Fastq files were demultiplexed using a custom C-script, allowing one mismatch within each barcode segment. The reads were trimmed using cutadapt[42]. UMIs were then extracted from bases 1-10 in Read 2 using UMI-Tools[43] and added to the read name. Only reads with TTTTT

at the bases 11-15 of Read 2 were kept (> 96% of all reads), allowing one mismatch. Lastly, the barcode was also moved to the read name.

easySHARE-ATAC-seq pre-processing

Fastq files were demultiplexed using a custom C-script, allowing one mismatch within each barcode segment. The paired reads were trimmed using cutadapt.

easySHARE-RNA-seq read alignment

We only used Read 1 for all our (sc)RNA-seq analyses as sequencing quality tends to drop after a polyT tail is sequenced in R2. In order to mitigate the possible effects of mapping bias (some species mapping better or worse to the mm10 genome) and since there is no publicly available PWD/PhJ reference genome, we modified the approach from Gao et al and crowley et al. [44,45]. In short, the *vcf2diploid*[46] tool was used to construct 'Artificial Genomes' (AG) for CAST/EiJ, PWD/PhJ and SPRET/EiJ by incorporating the respective SNPs and InDels into the mm10 genome. Additionally, *vcf2diploid* reports a chain file as output. scRNA-seq data from each species or F1 hybrids was mapped to both mm10 and the respective AG using the two-pass mode in STAR[47] with the parameters --outFilterMultimapNmax 20 --outFilterMismatchNmax 15. Mapping Quality (MAPQ) of each read was compared across the two alignments and the better mapping location was kept. In case of equal MAPQ, we kept the mm10-mapped read. We then generated UMI-collapsed count matrices for each mapped genome using featureCounts[48] from the Subread package and UMItools. A detailed description and verification of this approach can be found in the **Supplementary Notes**.

easySHARE-ATAC-seq read alignment

All reads were mapped to the mm10 genome using bwa mem[49]. Reads with alignment quality < Q30, unmapped, undetermined barcode, or mapped to mtDNA were discarded. Duplicates were removed using Picard tools. Open chromatin regions were called by subsampling the bamfiles from all samples to a common depth, merging them into a pooled bamfile and using the peakcaller MACS2[50] with the parameters -nomodel -keep-dup -min-length 100. All peaks on the chromosome X or Y were removed. The count matrices as well as the FRiP score was generated using featureCounts. For any single-cell analysis involving the ATAC-seq data, we did not correct for potential mapping bias since we observed highly similar mapping efficiencies across all species (between 98.03%-99.45% of reads mapped). Additionally, any potential bias would only decrease our power for analyses such as linking peaks to genes.

Assigning reads to allele of origin in F1 Hybrids

In order to assign each read from the F1 Hybrids to the respective allele of origin, we first used samtools[51] mpileup inputting the sample bamfiles and our variant (VCF) files containing SNP data with the parameters -A -B -C 0 -Q 0 -R --output-extra QNAME. The resulting output file contained a list of readnames overlapping each SNP as well as information about having either the REF or ALT variant. Using a custom python script, we then filtered for reads overlapping either only REF or only ALT position. Reads overlapping no variant or containing REF and ALT variants were discarded.

Principal Component Analysis

For PC analysis, we used a count matrix generated from pseudobulk sample bamfiles for RNA- and ATAC-seq. The matrix was loaded into DESeq2[52] and prefiltered by requiring at least 10 counts per gene across all samples combined. We then performed variance stabilizing transformation and did PC analysis using the *stats* R package.

Filtering, Integration & Dimensional reduction of scRNAseq data

The count matrices were loaded into Seurat[53] and cells were then filtered for >100 detected genes, >250 UMIs and < 20.000 UMIs. The sub-libraries coming from the same experiment were then merged together. Merged experiments from the same genotype (one from male mouse, one from female mouse) were then integrated by first using SCTransform then finding common features between the two experiments using FindIntegrationAnchors() and finally integrated using IntegrateData(). Lastly, the integrated datasets from all genotypes were sequentially integrated onto one another. To visualise the data, we projected the cells into 2D space by UMAP using the first 30 principal components and identified clusters using FindClusters(). Afterwards, we assigned cell cycle scores and excluded clusters consisting of nuclei solely in the G2M-phase. Cell types were assigned via expression of previously known marker genes. We extracted the barcodes for each cell type and subsetted the bamfiles from each genotype into separate bamfiles for each cell type.

Filtering, Integration & Dimensional reduction of scATACseq data

Fragments per cell were counted using sinto and the resulting fragment file was loaded into Signac[54] alongside the count matrices and the peakset. We calculated basic QC statistics using base Signac and cells were then filtered for a FRiP score of at least 0.3, > 150 fragments, < 15.000 fragments. Again, sub-libraries coming from the same experiment were merged. We then integrated all experiments at once by finding common features across datasets using FindIntegrationAnchors() using PCs 2:30 and then integrating the data using IntegrateEmbeddings(). To visualise the data, we projected the cells into 2D space by UMAP.

Weighted-Nearest-Neigbor (WNN) Analysis & Cell type identification

In order to use data from both modalities simultaneously, we created a multimodal Seurat object and used WNN[25] clustering to visualise and leverage both modalities for downstream analysis such as calculating Peak-Gene Associations.

Differential Gene Expression

DIfferentially expressed (DE) genes between pseudobulk samples, cell types or alleles were analysed using the *edgeR*[55] package. Genes were required to have at least 30 total counts across all samples. DE Genes were calculated by fitting a negative binomial generalised linear model and performing a Quasi likelihood (QL) F-test. We corrected for multiple testing by using a Benjamini-Hochberg correction with an FDR of 0.05.

Differential Chromatin Accessibility

DIfferentially chromatin accessibility was calculated between pseudobulk samples or alleles using the *DiffBind*[56] package. Peaks were required to have at least 20 total counts across all samples to be considered. We applied a mapping bias correction for all species except C57BL/6 in the form of species- and peak-specific correction factors, calculated by directly measuring the extent of mapping bias on each peak. Lastly, we normalised by full library size

and used the default settings for calculating differentially accessible peaks (DESeq2 workflow, Wald test, FDR=0.1). For a detailed description of our mapping bias correction, see *Supplementary Notes*.

Categorization of genes into regulatory modes

To assign the mode of gene regulation, we used the approach from Metzger et al[8]. In each contrast, we combined the raw read counts from all four experiments for both parental species (e.g. 4x C57BL/6 vs 4x SPRET/EiJ) as well as the allele-specific reads from the F1 Hybrids (e.g. 4x C57BL/6 allele vs 4x SPRET/EiJ allele, from a total of four F1 Hybrids). Genes on the X chromosome or showing sex biased gene expression across all species (52 genes) were removed. Then, total counts for both comparisons (parentals and alleles) were downsampled to equal read counts using Fisher's noncentral hypergeometric distribution, implemented in the BiasedUrn R package {CITATION}. For analysis with pseudobulk counts, genes with less than 20 total reads were removed. For cell-type specific analyses, genes with less than 10 total reads were removed. We then tested for differences in total expression (DE) between the parental species per gene using the binomial exact test. To test for significant *cis*-regulatory differences ('*cis*-effects'), allele-specific counts from F_1 hybrids were compared using a binomial exact test for each gene. Lastly, to detect significant *trans*-regulatory differences ('*trans*-effects'), Fisher's exact test was used to compare the ratio of allele-specific counts in the parental species with the ratio of allele-specific counts in the F_1 hybrids for each gene. We then categorised each gene into regulatory modes using the following criteria:

- **Conserved**: No significant *cis*-effect, no significant DE, no significant *trans*-effect
- **Compensatory**: Significant *cis*-effect, no significant DE, significant *trans*-effect
- ***cis***: Significant *cis*-effect, significant DE, no significant trans-effect
- ***trans***: No significant *cis*-effect, significant DE, significant *trans*-effect
- ***cis+trans***: significant *cis*-effect, significant DE, significant *trans*-effect and log2(Species1/Species2) / log2(Allele1/Allele2) > 1
- ***cis-trans***: significant *cis*-effect, significant DE, significant *trans*-effect and log2(Species1/Species2) / log2(Allele1/Allele2) < 1
- **Ambiguous**: All other patterns

For all tests, a false discovery rate (FDR) corrected *p*-value of 0.05 was used. Effect sizes for each gene were calculated using the absolute log2 fold change between the parental species.

Categorization of genes into mode of inheritance

To assign the mode of inheritance, we again used the approach from Metzger et al[8]. In each contrast, we combined the raw read counts from all four experiments for both parental species (e.g. 4x C57BL/6 vs 4x SPRET/EiJ) as well as the total reads from the F1 Hybrids (e.g. 4x C57BL/6xSPRET/EiJ) and down-sampled each dataset to a common read count using Fisher's noncentral hypergeometric distribution within the R package biasedUrn. Genes with a total read count of less than 20 were removed in each contrast. We then used a binomial exact test to compare the total expression between the parental species as well as

between each parental species and the F1 hybrid. We then categorised each gene into modes of inheritance using the following criteria:

- ***Dominance***: Significant difference between the parental species and between *one* parental species and the F1 hybrid.
- **Additive**: Significant difference between the parental species and between each parental species and the F1 hybrid. Additionally, the F1 hybrid read count is in between the read counts of the parental species.
- ***Conserved***: No significant difference between the parental species or between any parental species and the F1 hybrid.
- ***Over/Underdominance***: Significant difference between the parental species and between each parental species and the F1 hybrid. Additionally, the F1 hybrid read count is either greater or lesser than both read counts from the parental species.

For all tests, a false discovery rate (FDR) corrected *p*-value of 0.01 was used.

Calculating Gene Expression Variance

In order to calculate gene expression variance, we aggregated gene expression data per sub-library per cell type for each species contrast and filtered out genes with less than 10 total reads across all samples. To ensure there is no relationship between total expression of a gene and its expression variance, we then performed a variance stabilising transformation using the *DESeq2* R package and extracted the variance values.

Defining transcription factors

For Suppl. Fig. 6, we defined TFs using the gene ontology annotation GO:0003700 ("DNA-binding transcription factor activity"). For alternative definitions in Suppl. Fig. 6D we used the term GO:0010468 ("regulation of gene expression") and the curated lists of TFs from Zhou et al.[23] and from Hammelman et al[24].

Calculating effect sizes per gene per cell type

For calculating the effect sizes per gene in each cell type, made use of the already identified differentially expressed genes per cell type and species contrast (see above, 'Differential Gene Expression'). We then ranked the genes by the absolute log2FC in expression (effect size) in each contrast. In order to have equal contribution from each species, we then combined and plotted the top 2,000 genes from each contrast per cell type. In order to compare total expression level across cell types, we normalised read counts across experiments within each strain using the DESeq2 standard workflow.

Identifying cell-type specific ATAC-seq peaks

Cell-type specific ATAC-seq peaks were identified in Seurat / Signac by comparing each cell type against all other cell types utilising a logistic regression framework combined with a likelihood ratio test (standard workflow, FindMarkers()). We then subset for peaks with an adjusted p-value lower than 0.05 and a average log2 fold-change higher than 1.

Gene Ontology Analysis

Gene Ontology Analysis was done using the R package clusterProfiler[57] with standard parameters.

Calculating Links (Peak–Gene Associations)

Peak–gene associations were calculated following the framework described by Ma et al[26]. In short, Spearman correlation was calculated for every peak–gene pair within a +-500kb window around the TSS of the expressed gene. To obtain a background estimation, we used chromVAR (*getBackgroundPeaks*()) to generate 100 background peaks matched in GC bias and chromatin accessibility but randomly distributed throughout the genome. We calculated the Spearman correlation between every background-gene comparison, resulting in a null distribution with known population mean and standard deviation. We then calculated the z-score for the peak–gene pair in question ((correlation - population mean)/ standard deviation) and used a one-sided z-test to determine the p-value. This functionality is also implemented in Signac under the function *LinkPeaks*().

When calculating the rate of shared links between BL6 and CAST/PWD/SPRET for each regulatory category, we first extracted all shared links and calculated the percentages of linked genes that fall into each regulatory category (e.g. *cis*, *trans*,..). We then normalised them by their initial frequency. For example, 9.44% of all shared links between BL6 and SPRET LSECs are to *trans*-regulated genes. However, in total only 6.48% of all links in BL6 are to *trans*-regulated genes. This results in a normalised rate of shared links of 1.42.

Acknowledgements

We thank members of the Chan and Jones lab for helpful discussions and critical reading of the manuscript. We are very grateful to Arnar Breevoort and Alex Pollen for sharing tissue preparation protocols and a very helpful research visit. We thank the Genome Center in the Max Planck Institute for Biology for providing support. We also thank Sinja Mattes and all animal caretakers at the Max Planck Institute for Biology Tübingen. M.P. and D.S. are supported by an International Max Planck Research School fellowship. M.K. and Y.F.C. were supported European Research Council Starting Grant 639096 “HybridMiX" and Proof-of-Concept Grant 101069216 “Haplotagging”. The research was supported by the Max Planck Society.

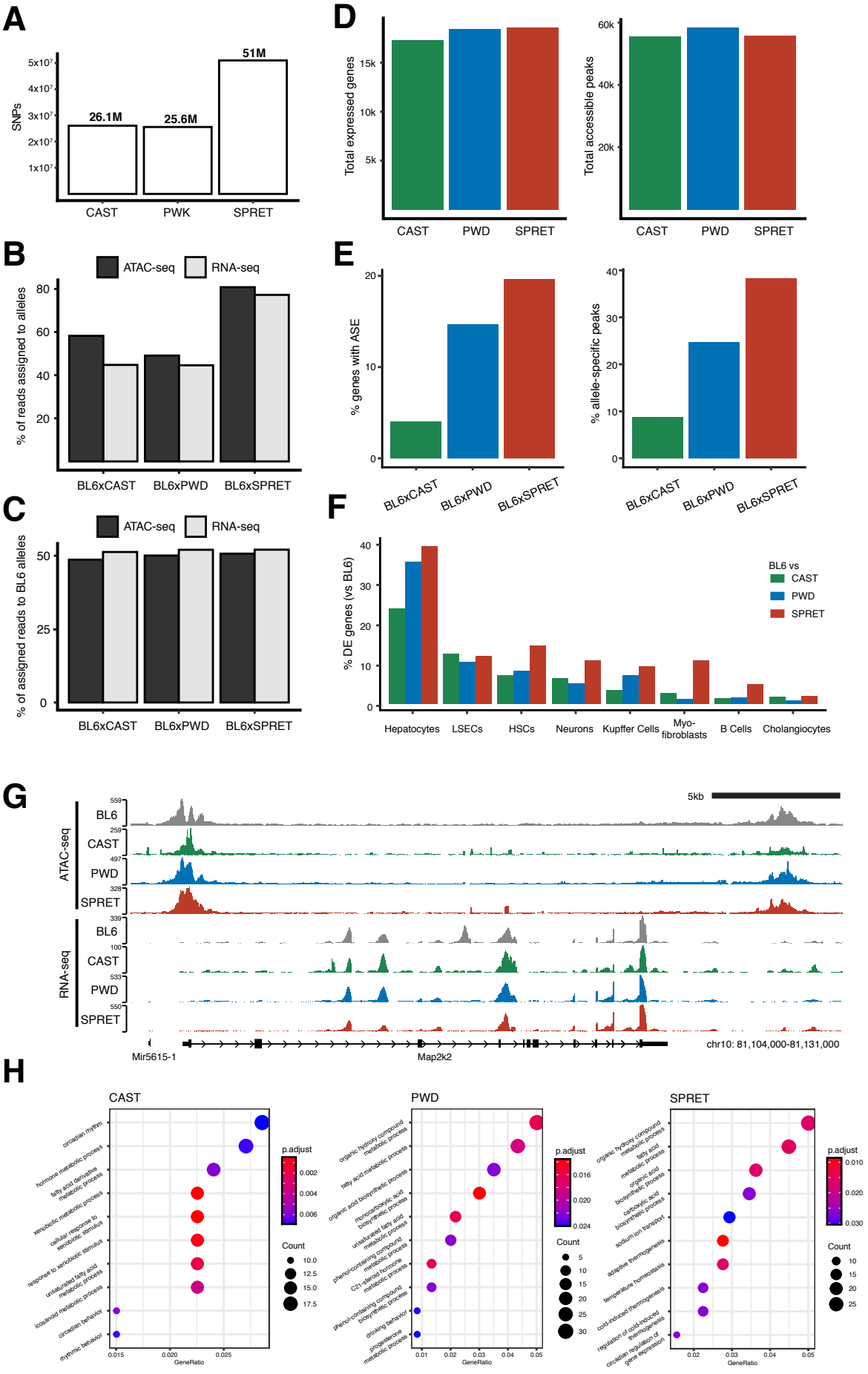
Author Contributions

V.S. and Y.F.C. designed the experiments. V.S. performed the experiments. V.S. performed computational analyses with input from Y.F.C, M.P. and D.S. V.S. wrote the manuscript. M.P., D.S., M.K. and Y.F.C. helped with experimental or computational support. All authors reviewed the manuscript. Y.F.C. directed the study with input from all authors.

Declaration of Interest

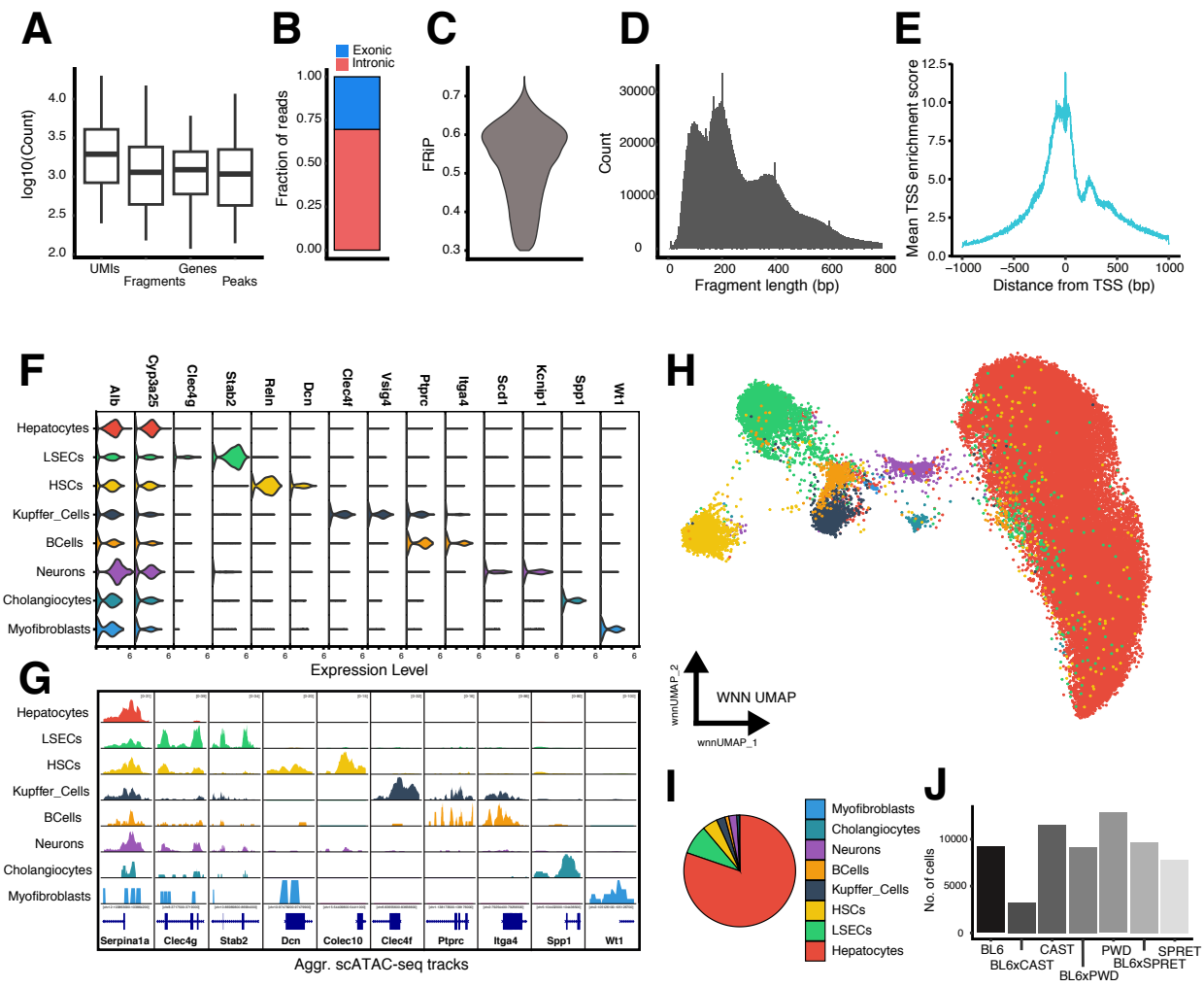
The authors declare no competing interests.

Supplemental Figures


A
SNPs
5x10⁷
4x10⁷
3x10⁷
2x10⁷
1x10⁷
26.1M
25.6M
51M
CAST
PWK
SPRET
B
ATAC-seq
RNA-seq
% of reads assigned to alleles
80
60
40
20
BL6xCAST
BL6xPWD
BL6xSPRET
C
ATAC-seq
RNA-seq
% of assigned reads to BL6 alleles
50
25
0
BL6xCAST
BL6xPWD
BL6xSPRET
D
Total expressed genes
15k
10k
5k
CAST
PWD
SPRET
Total accessible peaks
60k
40k
20k
CAST
PWD
SPRET
E
% genes with ASE
20
10
0
BL6xCAST
BL6xPWD
BL6xSPRET
% allele-specific peaks
40
30
20
10
0
BL6xCAST
BL6xPWD
BL6xSPRET
F
% DE genes (vs BL6)
40
30
20
10
0
BL6 vs
CAST
PWD
SPRET
Hepatocytes
LSECs
HSCs
Neurons
Kupffer Cells
Myo-fibroblasts
B Cells
Cholangiocytes
G
5kb
ATAC-seq
BL6
559
CAST
259
PWD
497
SPRET
328
RNA-seq
BL6
339
CAST
100
PWD
533
SPRET
550
Mir5615-1
Map2k2
chr10: 81,104,000-81,131,000
H
CAST
PWD
SPRET
p.adjust
Count
GeneRatio
circadian rhythm
hormone metabolic process
fatty acid derivative metabolic process
xenobiotic metabolic process
cellular response to xenobiotic stimulus
response to xenobiotic stimulus
unsaturated fatty acid metabolic process
icosanoid metabolic process
circadian behavior
rhythmic behavior
organic hydroxy compound metabolic process
fatty acid metabolic process
organic acid biosynthetic process
monocarboxylic acid biosynthetic process
unsaturated fatty acid metabolic process
phenol-containing compound metabolic process
C21-steroid hormone metabolic process
phenol-containing compound biosynthetic process
drinking behavior
progesterone metabolic process
organic hydroxy compound metabolic process
fatty acid metabolic process
organic acid biosynthetic process
carboxylic acid biosynthetic process
sodium ion transport
adaptive thermogenesis
temperature homeostasis
cold-induced thermogenesis
regulation of cold-induced thermogenesis
circadian regulation of gene expression

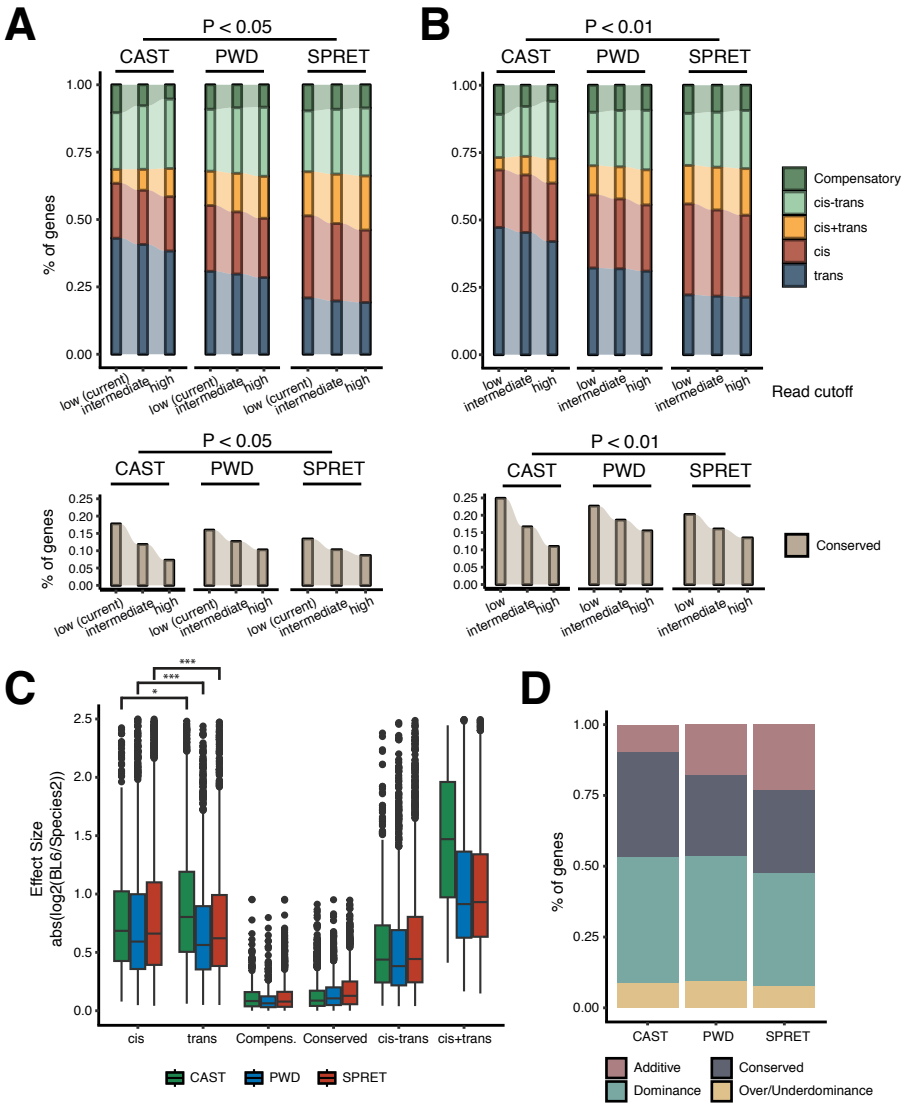
Suppl. Fig. 1: General overview and trends of the data

(A) Number of genomic variants of each species compared to mm10.
(B) Percentage of total reads assignable to alleles in the F1 Hybrids.
(C) Percentage of reads assigned to the BL6 allele in the F1 Hybrids.
(D) Total number of expressed genes and accessible peaks in each species contrast.
(E) Percentage of genes with allele-specific expression and allele-specific chromatin accessibility between the F1 Hybrid alleles.
(F) Percentage of DE genes in each species contrast within each cell type.
(G) Aggregated scATAC- and scRNA-seq track at the *Map2k2* locus. Tracks are coloured by species. The first four tracks are depicting ATAC-seq, the last four tracks are expression data.
(H) Gene ontology analysis of the top 1000 most differentially expressed genes in each species contrast.



Suppl. Fig. 2: Single-cell data quality control and cell type markers

(A) Boxplots depicting detected numbers of UMIs (transcripts; mean: 3.056, median: 1.948) and Fragments (ATAC-seq; mean: 1.890, median:1.321) per cell recovered from the respective modalities as well as the number of expressed genes (mean: 1.488, median :1.231) and accessible peaks (mean: 1.720, median:1.249) per cell.
(B) Fraction of cDNA reads overlapping exons (30.36%) or introns (69.64%).
(C) Violin plot of Fraction of Reads in Peaks (FRiP) per cell in the scATAC-seq (mean: 0.53)
(D) Histogram of fragment length in scATAC-seq sequencing reads.
(E) Mean TSS enrichment score per cell in relation to distance from nearest TSS in the scATAC-seq data.
(F) Normalised gene expression of representative marker genes per cell type
(G) Aggregate scATAC-seq tracks at marker accessibility peaks per cell type.
(H) UMAP-visualisation of Weighted-Nearest-Neigbour (WNN) integrated scRNA-seq and scATAC-seq modalities of 53.257 liver nuclei. Nuclei are coloured by cell type.
(I) Fraction of cell types recovered relative to total cells.
(J) Total number of cells recovered for each genotype.



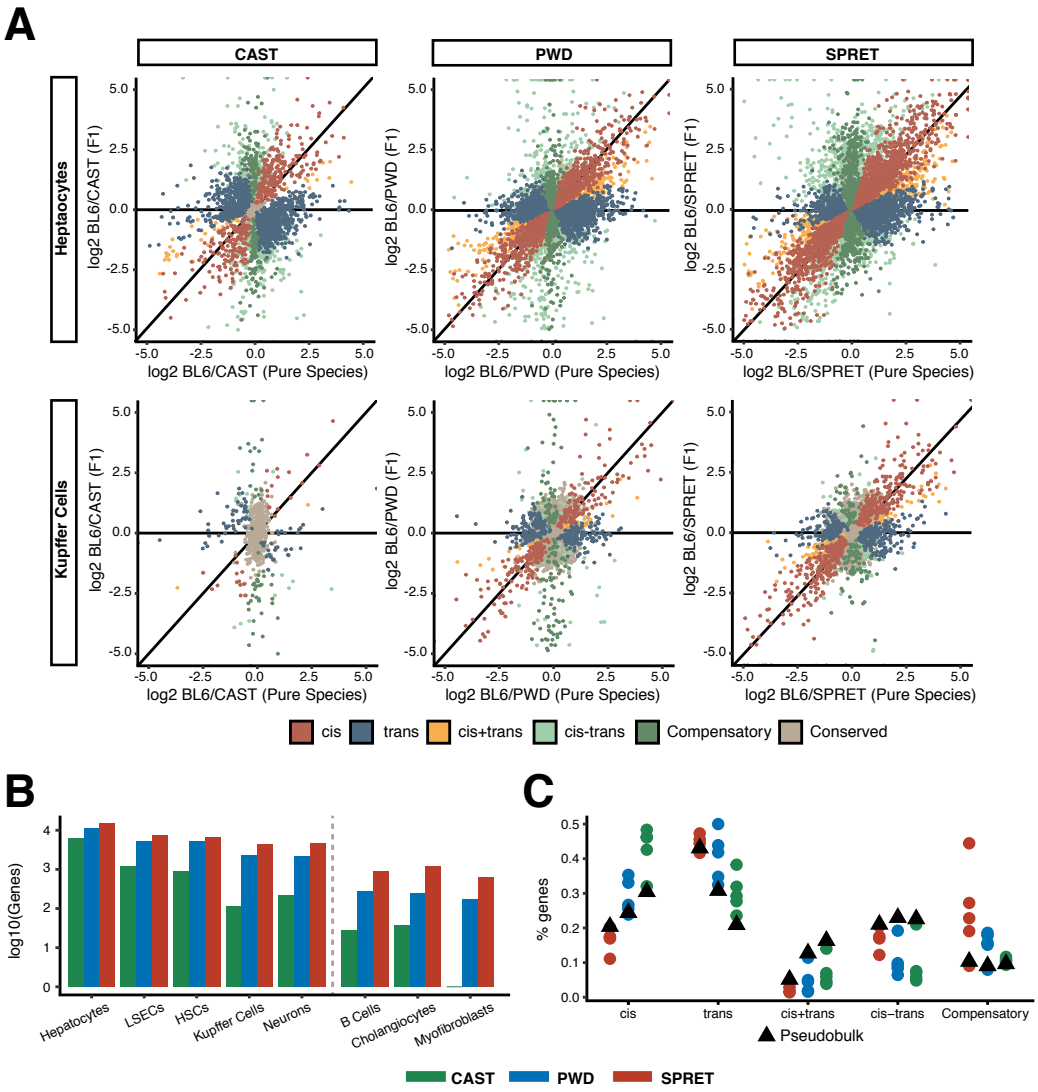
Suppl. Fig. 3: Frequency of regulatory categories is independent of filtering cutoffs

(A) Top: Fraction of genes assigned to regulatory categories with conserved genes excluded in all three species contrasts when using different filtering cutoffs per gene after normalisation (low (cutoff used in this study): 20 total reads across all samples, intermediate: 50 total reads, high; 100 total reads) and using an adjusted p-value cutoff of 0.05 in all statistical tests (Fisher's exact & 2x Binomial Exact Test). Bottom: The fraction of genes classified as conserved with different filtering cutoffs.

(B) Same as (A) but using an adjusted P-value cutoff of 0.01 in all statistical tests.

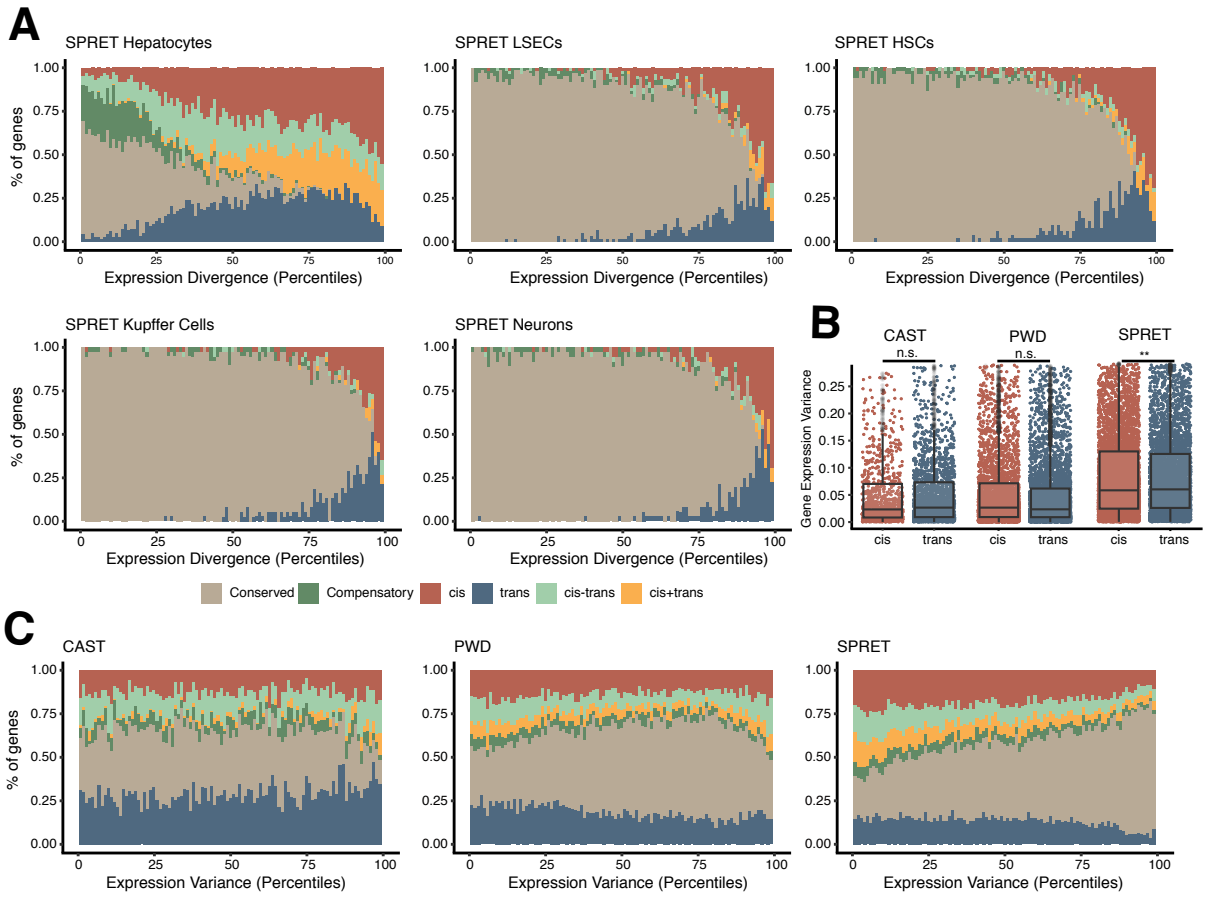
(C) Boxplots of effect sizes per gene (abs. log2(BL6 count / Species2 count)) split by regulatory mode and coloured by species contrast. Difference in effect sizes between *cis* and *trans* regulatory categories within each species were tested with a two-tailed Welch's t-test (***: $P < 0.001$, *: $P < 0.05$).

(D) The fraction of genes classified by their mode of inheritance within each pseudobulk species contrast.



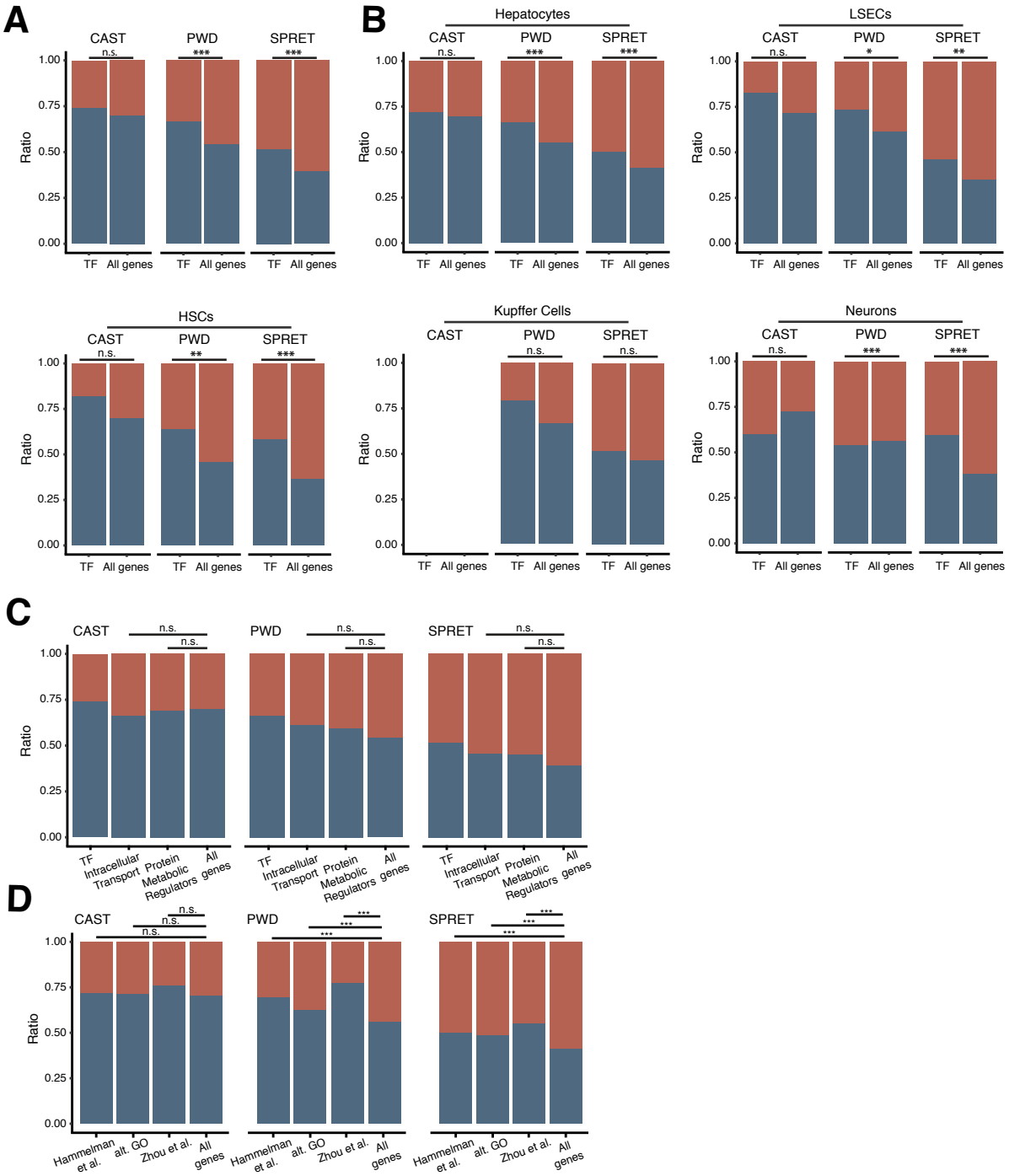
Suppl. Fig. 4: Overview of cell type specific regulatory patterns

(A) Scatterplot of log2FC in expression between the parental species vs. F1 hybrid alleles for CAST (left), PWD (middle) and SPRET (right) compared to BL6. Each dot is a gene. Genes are coloured based on the regulatory category they were assigned to (see Methods). Top row are the scatterplots for the most abundant cell type (Hepatocytes) and bottom row for the cell type with the least amount of cells recovered (Kupffer Cells). Horizontal line represents 100% *trans* effects, diagonal line 100% *cis* effects.
(B) Number of genes passing filtering and considered for analysis of regulatory mode within each cell type and contrast. Dotted line depicts cutoff for cell types considered.
(C) Overview of percentages of genes classified into regulatory modes. Dots are representing individual cell types and are coloured by species contrast. Triangle represents pseudobulk fractions from **Fig. 2B.**



Suppl. Fig. 5: Relationship of regulatory categories with expression divergence and expression variance

(A) Frequency of regulatory categories along genes grouped into percentiles of absolute expression divergence between BL6 and SPRET for each cell type.
Comparison of gene expression variance between *cis*- and *trans*-regulated genes for each species contrast. Each dot is a gene. Expression variance for a gene was calculated within each cell type separately and all *cis*- or *trans*-regulated genes across all
(B) cell types are shown. We tested for significant differences using a two-tailed Welch's t-test (**: $P < 0.01$).
(C) Frequency of regulatory categories along genes grouped into percentiles of increasing expression variance between all three species contrasts.



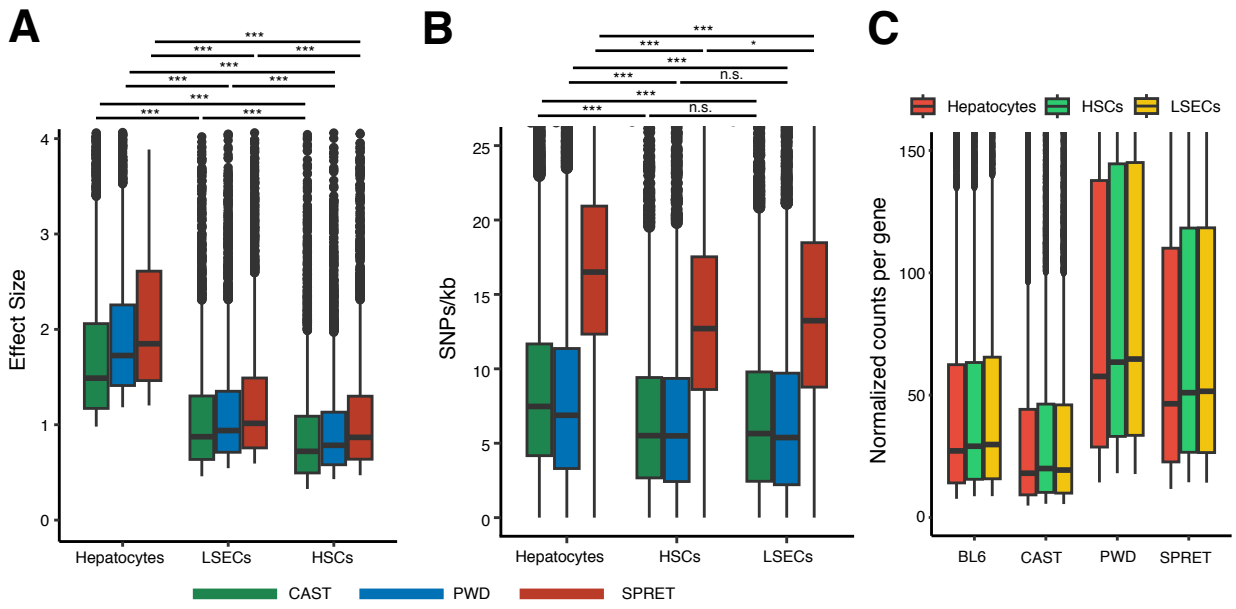
Suppl. Fig. 6: Transcription factors are more frequently regulated in *trans*

(A) The ratio of *cis*- and *trans*-regulated genes combined across all cell types for either transcription factors (TF) or all genes. We tested for significant differences in the ratios between the different groups using Fisher's exact test (n.s.: $P > 0.05$, *: $P < 0.05$, **: $P < 0.01$, ***: $P < 0.001$).

(B) The ratio of *cis*- and *trans*-regulated genes in different cell types for either transcription factors (TF) or all genes. Statistical tests as in (A).

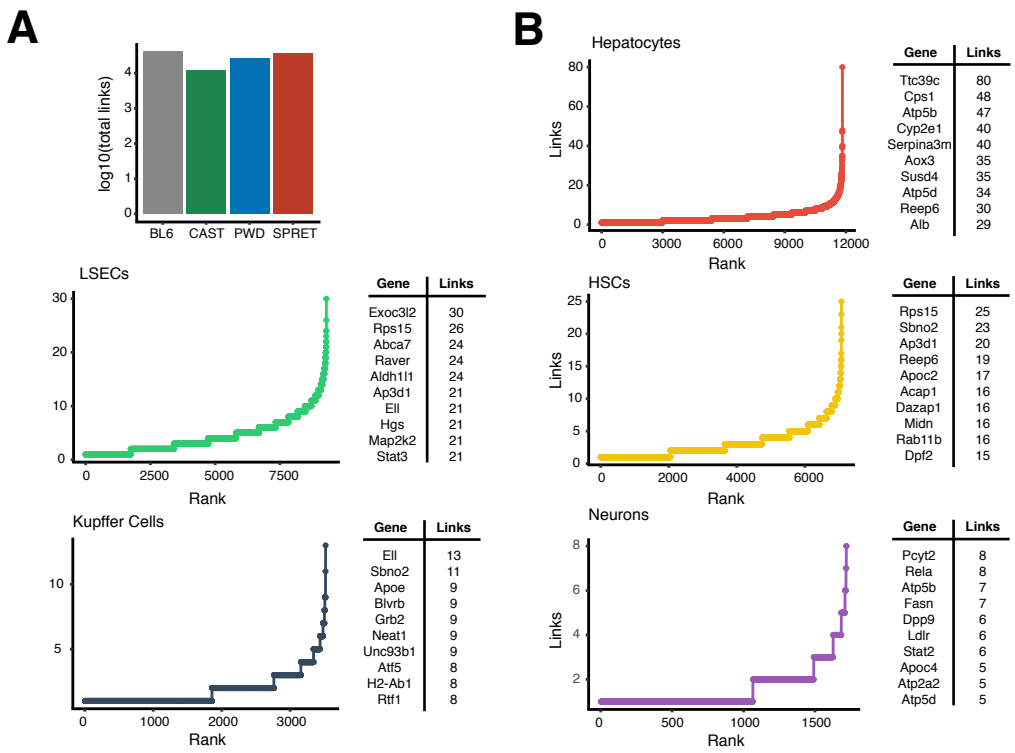
(C) The ratio of *cis*- and *trans*-regulated genes in each cell type for groups of genes based on other GO annotations (see axis labels). Statistical tests as in (A).

(D) The ratio of *cis*- and *trans*-regulated genes in each cell type using different definitions for defining TFs (see Methods). Statistical tests as in (A).



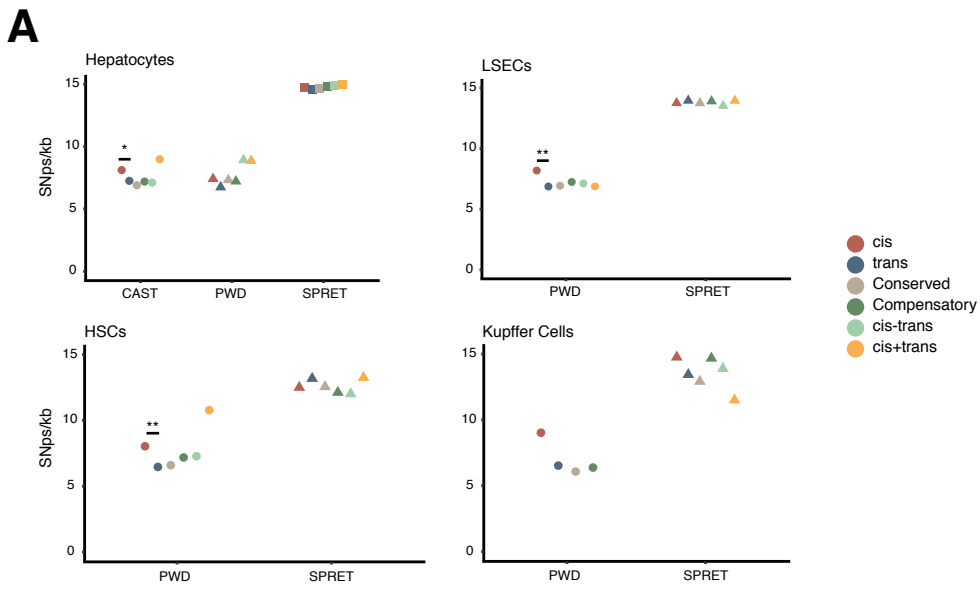
Suppl. Fig. 7: Cell type specific effect sizes and genetic variation

(A) Boxplot of gene expression effect sizes per cell type split by species contrast. Tested for significant differences using a two-tailed Welch's t-test (***: $P < 0.001$).
(B) SNPs/kb in cell type specific ATAC-seq peaks split by species. Tested for differences using a two-tailed Welch's t-test (***: $P < 0.001$, *: $P < 0.05$).
(C) Total expression level of cell types within each species of genes plotted in (A). Reads are normalized using DESeq2.



Suppl. Fig. 8: Cell-type specific gene-pCRE links

(A) Total number of links recovered per species, summed across all cell types.
(B) Genes are ranked by their number of linked pCREs for each cell type. List of genes on the right side of each plot displays 10 relevant genes out of the top20 with the most links.



Suppl. Fig. 9: Rates of genomic variation

(A) SNPs/kb in pCREs split by regulatory mode of their target gene in different cell types. Tested for differences in mean SNPs per kb using a one-tailed Welch's t.test (**: $P < 0.01$, *: $P < 0.05$)

References

1. Carroll, S. B. Evo-Devo and an Expanding Evolutionary Synthesis: A Genetic Theory of Morphological Evolution. *Cell* **134**, 25–36 (2008).
2. Stern, D. L. & Orgogozo, V. The Loci of Evolution: How Predictable Is Genetic Evolution? *Evolution* **62**, 2155–2177 (2008).
3. Signor, S. A. & Nuzhdin, S. V. The Evolution of Gene Expression in cis and trans. *Trends in Genetics* **34**, 532–544 (2018).
4. Chan, Y. F. *et al.* Adaptive Evolution of Pelvic Reduction in Sticklebacks by Recurrent Deletion of a Pitx1 Enhancer. *Science* **327**, 302–305 (2010).
5. Stern, D. L. & Orgogozo, V. Is Genetic Evolution Predictable? *Science* **323**, 746–751 (2009).
6. Cowles, C. R., Hirschhorn, J. N., Altshuler, D. & Lander, E. S. Detection of regulatory variation in mouse genes. *Nat Genet* **32**, 432–437 (2002).
7. Wittkopp, P. J., Haerum, B. K. & Clark, A. G. Evolutionary changes in cis and trans gene regulation. *Nature* **430**, 85–88 (2004).
8. Metzger, B. P. H., Wittkopp, P. J. & Coolon, Joseph. D. Evolutionary Dynamics of Regulatory Changes Underlying Gene Expression Divergence among Saccharomyces Species. *Genome Biology and Evolution* **9**, 843–854 (2017).
9. Coolon, J. D., McManus, C. J., Stevenson, K. R., Graveley, B. R. & Wittkopp, P. J. Tempo and mode of regulatory evolution in Drosophila. *Genome Res.* **24**, 797–808 (2014).
10. Verta, J.-P. & Jones, F. C. Predominance of cis-regulatory changes in parallel expression divergence of sticklebacks. *eLife* **8**, e43785 (2019).
11. Agoglia, R. M. *et al.* Primate cell fusion disentangles gene regulatory divergence in neurodevelopment. *Nature* **592**, 421–427 (2021).
12. Shi, X. *et al.* Cis- and trans-regulatory divergence between progenitor species determines gene-expression novelty in Arabidopsis allopolyploids. *Nat Commun* **3**, 950 (2012).
13. Panten, J. *et al.* The dynamic genetic determinants of increased transcriptional divergence in spermatids. *Nat Commun* **15**, 1272 (2024).
14. Elorbany, R. *et al.* Single-cell sequencing reveals lineage-specific dynamic genetic regulation of gene expression during human cardiomyocyte differentiation. *PLOS Genetics* **18**, e1009666 (2022).
15. Ben-David, E. *et al.* Whole-organism eQTL mapping at cellular resolution with single-cell sequencing. *eLife* **10**, e65857 (2021).
16. van der Wijst, M. *et al.* The single-cell eQTLGen consortium. *eLife* **9**, e52155 (2020).
17. Knowles, D. A. *et al.* Allele-specific expression reveals interactions between genetic variation and environment. *Nat Methods* **14**, 699–702 (2017).
18. Soltys, V., Peters, M., Su, D., Kučka, M. & Chan, Y. F. Flexible and high-throughput simultaneous profiling of gene expression and chromatin accessibility in single cells. 2024.02.26.581705 Preprint at https://doi.org/10.1101/2024.02.26.581705 (2024).
19. Su, Q. *et al.* Single-cell RNA transcriptome landscape of hepatocytes and non-parenchymal cells in healthy and NAFLD mouse liver. *iScience* **24**, 103233 (2021).
20. Han, X. *et al.* Mapping the Mouse Cell Atlas by Microwell-Seq. *Cell* **172**, 1091-1107.e17

(2018).
21. Keane, T. M. *et al.* Mouse genomic variation and its effect on phenotypes and gene regulation. *Nature* **477**, 289–294 (2011).
22. Madan Babu, M. & Teichmann, S. A. Evolution of transcription factors and the gene regulatory network in Escherichia coli. *Nucleic Acids Res* **31**, 1234–1244 (2003).
23. Zhou, Q. *et al.* A mouse tissue transcription factor atlas. *Nat Commun* **8**, 15089 (2017).
24. Hammelman, J., Patel, T., Closser, M., Wichterle, H. & Gifford, D. Ranking Reprogramming Factors for Directed Differentiation. 2021.05.14.444080 Preprint at https://doi.org/10.1101/2021.05.14.444080 (2021).
25. Hao, Y. *et al.* Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573-3587.e29 (2021).
26. Ma, S. *et al.* Chromatin Potential Identified by Shared Single-Cell Profiling of RNA and Chromatin. *Cell* **183**, 1103-1116.e20 (2020).
27. Zhou, D. *et al.* Mst1 and Mst2 maintain hepatocyte quiescence and suppress the development of hepatocellular carcinoma through inactivation of the Yap1 oncogene. *Cancer Cell* **16**, 425–438 (2009).
28. Kano, A. *et al.* Endothelial Cells Require STAT3 for Protection against Endotoxin-induced Inf lammation. *J Exp Med* **198**, 1517–1525 (2003).
29. Nourmohammad, A. *et al.* Adaptive Evolution of Gene Expression in Drosophila. *Cell Reports* **20**, 1385–1395 (2017).
30. Gordon, K. L. & Ruvinsky, I. Tempo and Mode in Evolution of Transcriptional Regulation. *PLOS Genetics* **8**, e1002432 (2012).
31. Mack, K. L., Campbell, P. & Nachman, M. W. Gene regulation and speciation in house mice. *Genome Res.* **26**, 451–461 (2016).
32. Goncalves, A. *et al.* Extensive compensatory cis-trans regulation in the evolution of mouse gene expression. *Genome Res.* **22**, 2376–2384 (2012).
33. Lemmon, Z. H., Bukowski, R., Sun, Q. & Doebley, J. F. The Role of cis Regulatory Evolution in Maize Domestication. *PLOS Genetics* **10**, e1004745 (2014).
34. Mack, K. L., Square, T. A., Zhao, B., Miller, C. T. & Fraser, H. B. Evolution of Spatial and Temporal cis-Regulatory Divergence in Sticklebacks. *Molecular Biology and Evolution* **40**, msad034 (2023).
35. Poisson, J. *et al.* Liver sinusoidal endothelial cells: Physiology and role in liver diseases. *J Hepatol* **66**, 212–227 (2017).
36. Shetty, S., Lalor, P. F. & Adams, D. H. Liver sinusoidal endothelial cells — gatekeepers of hepatic immunity. *Nat Rev Gastroenterol Hepatol* **15**, 555–567 (2018).
37. Schulze, R. J., Schott, M. B., Casey, C. A., Tuma, P. L. & McNiven, M. A. The cell biology of the hepatocyte: A membrane trafficking machine. *J Cell Biol* **218**, 2096–2112 (2019).
38. Karimkhanloo, H. *et al.* Mouse strain-dependent variation in metabolic associated fatty liver disease (MAFLD): a comprehensive resource tool for pre-clinical studies. *Sci Rep* **13**, 4711 (2023).
39. Song, Y. *et al.* Adaptive introgression of anticoagulant rodent poison resistance by hybridization between old world mice. *Curr Biol* **21**, 1296–1301 (2011).
40. Olson, J. P., Miller, L. L. & Troup, S. B. Synthesis of clotting factors by the isolated perfused rat liver. *J Clin Invest* **45**, 690–701 (1966).

41. Gregorová, S. & Forejt, J. PWD/Ph and PWK/Ph Inbred Mouse Strains of Mus т. musculus Subspecies-a Valuable Resource of Phenotypic Variations and Genomic Polymorphisms. *Folia biologica* **46**, 31–41 (2000).
42. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17**, 10–12 (2011).
43. Smith, T., Heger, A. & Sudbery, I. UMI-tools: modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy. *Genome Res* **27**, 491–499 (2017).
44. Gao, Q., Sun, W., Ballegeer, M., Libert, C. & Chen, W. Predominant contribution of cis-regulatory divergence in the evolution of mouse alternative splicing. *Molecular Systems Biology* **11**, 816 (2015).
45. Crowley, J. J. *et al.* Analyses of allele-specific gene expression in highly divergent mouse crosses identifies pervasive allelic imbalance. *Nat Genet* **47**, 353–360 (2015).
46. Rozowsky, J. *et al.* AlleleSeq: analysis of allele-specific expression and binding in a network framework. *Mol Syst Biol* **7**, 522 (2011).
47. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
48. Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930 (2014).
49. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
50. Zhang, Y. *et al.* Model-based Analysis of ChIP-Seq (MACS). *Genome Biology* **9**, R137 (2008).
51. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
52. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology* **15**, 550 (2014).
53. Hao, Y. *et al.* Dictionary learning for integrative, multimodal and scalable single-cell analysis. *Nat Biotechnol* **42**, 293–304 (2024).
54. Stuart, T., Srivastava, A., Madad, S., Lareau, C. A. & Satija, R. Single-cell chromatin state analysis with Signac. *Nat Methods* **18**, 1333–1341 (2021).
55. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).
56. Stark, R. & Brown, G. DiffBind: differential binding analysis of ChIP-Seq peak data. *R package version* **100**, (2011).
57. Wu, T. *et al.* clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. *The Innovation* **2**, 100141 (2021).

Supplementary Notes

Mitigating mapping bias in the scRNA-seq

In order to minimise the potential effect of mapping bias onto our data and to circumvent the lack of reference genome for the PWD/PhJ strain, we modified the approach from Crowley et al.[1] and Gao et al [2]. For each species (except C57BL/6, since mm10 is the proper reference genome), we constructed "Artificial Genomes" (AG) using the *vcf2diploid* tool[3] (**Suppl. Fig. 1A)**. *Vcf2diploid* incorporates SNPs and InDels of each species into the mm10 genome and additionally outputs a chain file for lifting genome coordinates over from mm10. This allows us to map reads from each species to both genomes (mm10 and the appropriate AG), compare mapping quality across them and keep the better mapping read, which is especially useful in the case of F1 hybrids. Even though reference genomes are available for CAST/EiJ and SPRET/EiJ mice, we used this approach for all three species in order to generate consistent results. Additionally, this allowed us to use mm10 gene annotations.

After mapping each sample using STAR[4], we compare the mapping quality (MAPQ) for each read in both alignments and discard the alignment with lower MAPQ. In case of an equal MAPQ, we kept the mm10 mapping read. We then discard multimapping reads, reads not overlapping a gene and non-primary alignments. Next, we merge corresponding bamfiles and use UMI-tools[6] to remove transcript duplicates. Importantly, since easySHARE-seq relies on fragmentation *after* a first PCR, transcript duplicates do not necessarily have the same genomic sequence. However, by de-duplicating on a gene level using the *–per-gene* flag, UMI-tools still accurately identifies and removes transcript duplicates. Lastly, we use UMI-tools to generate a (single-cell) count matrix for each sample.

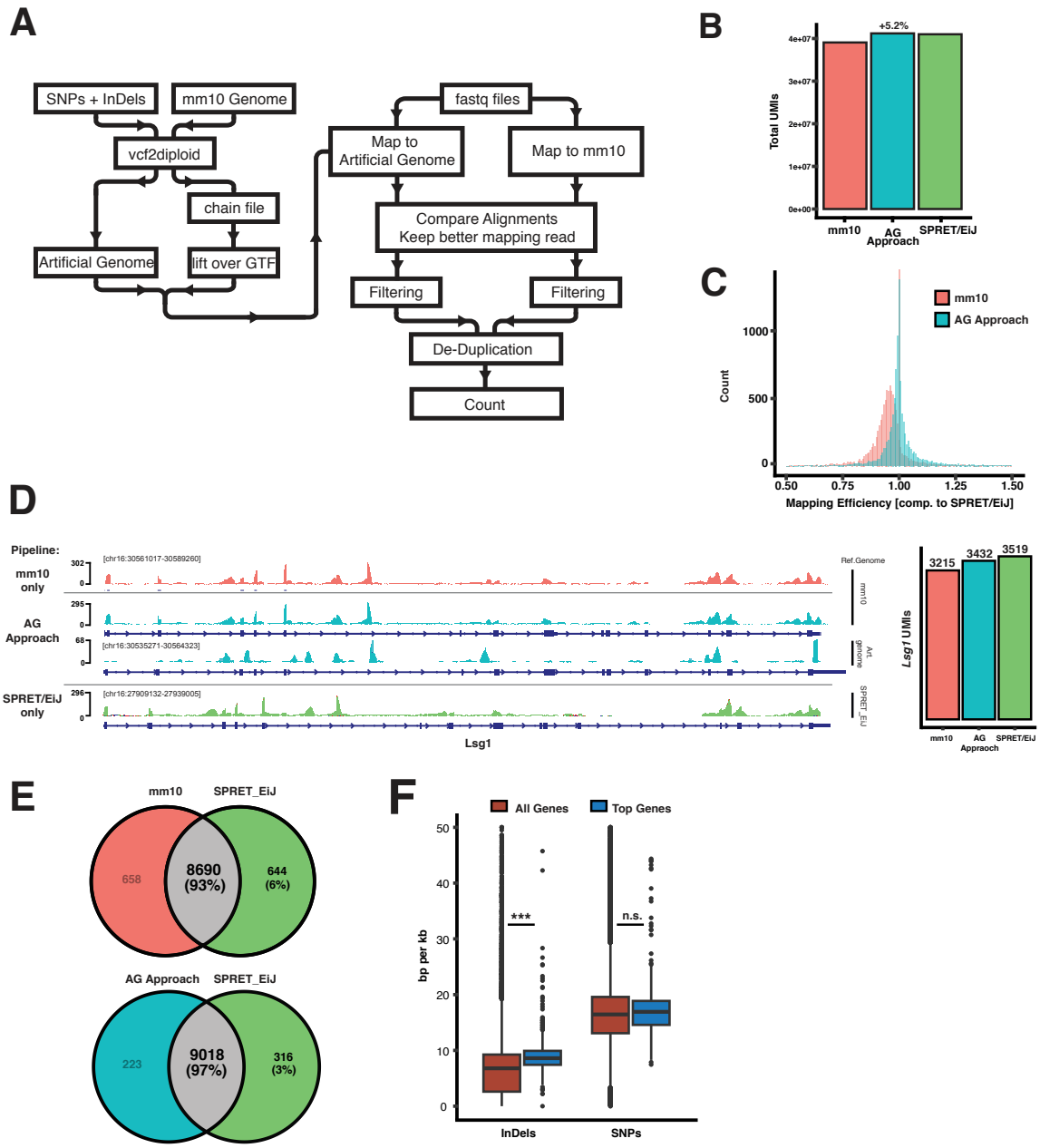
To verify our approach, we made use of our scRNA-seq SPRET/EiJ data since it is the most evolutionarily distant species from C57BL/6 and thus mapping bias should have the highest effects. Additionally, a reference genome SPRET/EiJ is available at www.mousegenomes.org. We mapped this data three different ways: to mm10 only, using the AG approach or to the SPRET/EiJ reference genome. Using the AG approach, we recovered 5.2% more UMIs than compared to mm10 only (**Suppl. Fig. 2B**) and nearly identical UMI counts compared to the SPRET/EiJ reference genome (< 0.5% difference). Interestingly, 69% of reads additionally mapping to the AG (compared to mm10) are classified as unplaced when mapping to mm10. We then compared mapping efficiency for the top 10.000 expressed genes and calculated how many UMIs per gene are recovered when mapping to mm10 or using the AG approach, both compared to the SPRET/EiJ reference genome (**Suppl. Fig. 2C**). Mapping using the AG approach outperforms mapping to mm10 and is highly similar compared to mapping to the SPRET/EiJ reference genome (median mapping efficiency: 0.9516 (mm10) vs. 0.9976 (AG) vs. 1 (SPRET/EiJ)). For example, mapping using the AG approach recovers 97.5% of UMIs for *Lsg1* compared to mapping to SPRET/EiJ, compared to 91.36% in mm10 (**Suppl. Fig. 1D**). Next, we calculated differentially expressed (DE) genes between all three differently mapped SPRET/EiJ datasets and the pseudobulked C57BL/6 scRNA-seq dataset generated for this study (**Suppl. Fig. 1E**). Mapping to mm10 only recovers 93% of DE genes whereas mapping using the AG approach recovers 96.6% of DE genes. Additionally, using the AG approach falsely classifies only 223 genes as DE (1.2% of expressed genes).

Lastly, we investigated the drivers of mapping bias. We ordered genes by total amount of recovered UMIs (mapped with the AG approach vs mm10) and subsetted for the top genes

that collectively have 50% of the recovered UMIs (1.193 total genes). We then calculated SNPs/kilobase (kb) and base pairs (bp) in InDels/kb for these genes and compared them to all genes (**Suppl. Fig. 1F**). While the frequency of SNPs did not differ, the genes with 50% of recovered UMIs were highly enriched for bp in InDels ($p < 0.001$, Welch's t-test), indicating that these are the main driver of mapping bias.
To summarise, we show that using the AG approach provides accurate mapping and mitigates the majority of mapping bias effects. While we cannot exclude the possibility that underlying mapping bias still influences some of our analysis, we presume it to play a very minor role, especially when comparing across species.

Correcting for Mapping Bias in the ATAC-seq

In order to minimise the impact of mapping bias onto comparative analyses involving the ATAC-seq data (e.g. differentially accessible peaks), we directly measured mapping bias for each species at each peak, allowing us to correct for it. To do so, we made use of three publicly available reference genomes for CAST/EiJ, SPRET/EiJ and PWK/PhJ, using PWK/PhJ as a proxy for PWD/PhJ (see main article). All genomes as well as gDNA sequencing data used to construct them and C57BL/6 gDNA sequencing data were downloaded from www.mousegenomes.org and mapped to the appropriate reference genome (e.g. CAST/EiJ gDNA data to the CAST/EiJ reference genome).
We then calculated per species peak-specific correction factors (**Suppl. Fig. 2A**). First, we transferred peaks from mm10 coordinates to e.g. the CAST/EiJ reference genome using blat. We filtered out blat hits with more than 80 mismatches, lower than 90% of the sequence aligned or hits on a different chromosome. If multiple hits were retained, we chose the one with the highest score, resulting in 85.2% of peaks confidently transferred to CAST/EiJ coordinates (85.1% or 82.8% for PWK/PhJ and SPRET/EiJ, respectively). Next, we extracted an equal number of reads overlapping each peak from C57BL/6 and gDNA sequencing data of the second species (e.g. 100 CAST/EiJ and 100 C57BL/6 reads for the same peak). Peaks with less than 25 reads extracted were excluded from further analysis. We then remapped these reads to mm10. If a focal peak has no mapping bias, the ratio of mapped C57BL/6 and e.g. CAST/EiJ gDNA reads will be 1. If it differs from 1, it is due to mapping bias. Therefore, we use this ratio directly as peak-specific correction factors and multiply e.g. CAST/EiJ peak counts before analysing differentially accessible peaks.
The majority of peak-specific correction factors were between 0.95 and 1.05, with SPRET/EiJ having slightly higher average correction factors, consistent with being the most evolutionarily distant species (**Suppl. Fig. 2B**, mean correction factors: 1.0136 (CAST/EiJ), 1.012933 (PWD/PhJ) and 1.0221 (SPRET/EiJ)). We then analysed differentially accessible peaks between CAST/EiJ and C57BL/6 samples and compared the results with and without applying the mapping bias correction. Applying the correction leads to balanced log2 fold changes (**Suppl. Fig. 2C**). Without correcting for mapping bias, 51.7% of differentially accessible peaks have higher accessibility in C57BL/6. However, after correcting for mapping bias, differentially accessible peaks are evenly balanced between the species with only 50.4% of peaks having higher accessibility in C57BL/6 (**Suppl. Fig. 2D**), showcasing the validity and effectiveness of our approach to correct for mapping bias.
To summarise, we developed a simple approach to directly measure mapping bias and show that it is appropriate when correcting mapping bias.



Supplementary Figure 1: Using the AG Approach for scRNA-seq mapping mitigates the majority of mapping bias effects

(A) Schematic overview of the major steps in the analytical pipeline.

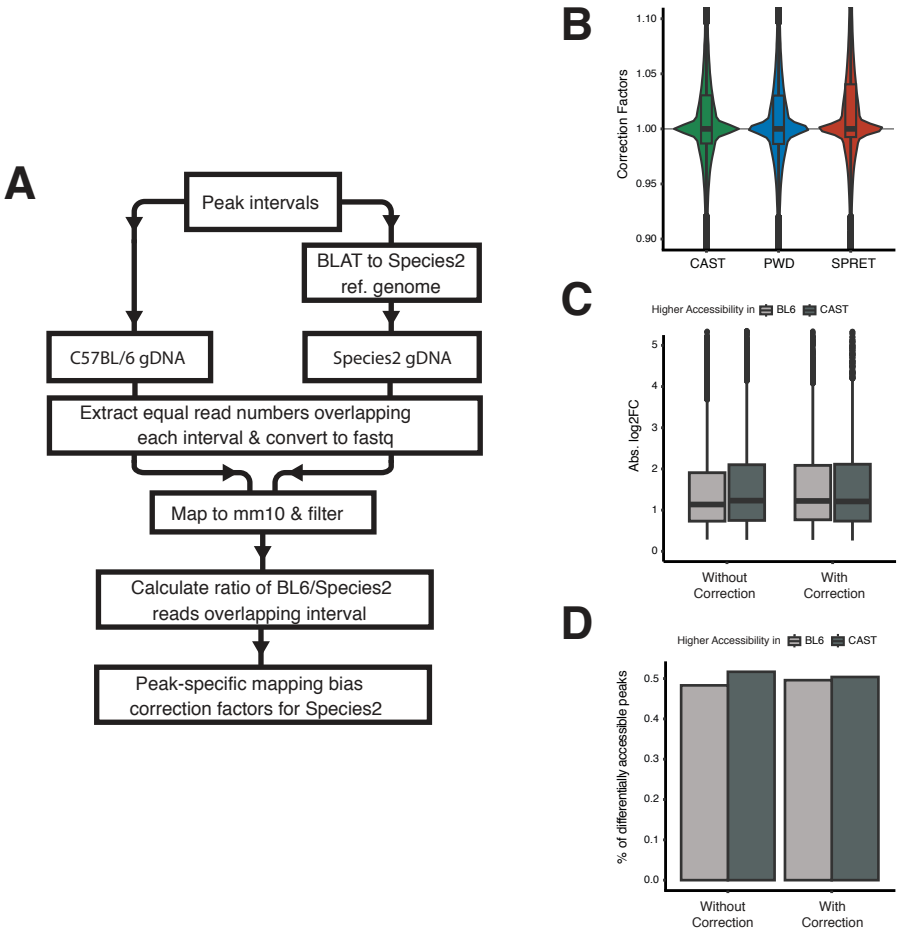
(B) Total UMIs recovered when SPRET/EiJ scRNA-seq data is mapped to the mm10 genome (red), using the AG Approach (blue) or to the SPRET/EiJ genome (green). Using the AG approach recoveres nearly identical to UMIs compared to SPRET/EiJ (< 0.5% difference).

(C) Relative mapping efficiency per gene compared to SPRET/EiJ mapped data. Number of UMIs recovered per gene in relation to number of UMIs recovered when the data is mapped to the SPRET/EiJ genome. Red: mapped to mm10. Blue: using the AG approach.

(D) Representative tracks of the differently mapped scRNA-seq data at the *Lsg1* locus. Red: mapped to mm10. Blue: using the AG approach. Top track is the mm10 track, bottom is the Artificial Genome track. Green: mapped to SPRET/EiJ genome. Right: Number of UMIs recovered for *Lsg1* per mapping approach.

(E) Results of differential expression (DE) analysis between C57BL/6 data and differently mapped SPRET/EiJ data. Colours indicate how the SPRET/EiJ data has been mapped. Top: DE analysis with mm10-mapped SPRET/EiJ data against C57BL/6 data only recovers 93% of DE genes compared to SPRET/EiJ-mapped SPRET/EiJ data. Bottom: Mapping the SPRET/EiJ data using the AG approach recovers 97% of DE genes.

(F) Number of bp/kb in InDels or SNPs. Red: All Genes. Blue: 1.193 genes which cumulatively recover 50% of total UMIs when using the AG approach compared to mm10 mapping. Top genes are significantly enriched for InDels but not for SNPs compared to all genes (t-test, $p < 0.001$).



Supplementary Figure 2: Mapping Bias Correction in ATAC-seq analysis

(A) Schematic overview of the major steps in the analytical pipeline
(B) Distribution of correction factors for each species
(C) Absolute log2 fold-changes with and without correction when calculating differentially accessible peaks between BL6 and CAST. Split by directionality of higher accessibility.
(D) Percentage of differentially accessible with higher accessibility in BL6 vs CAST, either with or without mapping bias correction.

References

1. Crowley, J. J. *et al.* Analyses of allele-specific gene expression in highly divergent mouse crosses identifies pervasive allelic imbalance. *Nat Genet* **47**, 353–360 (2015).
2. Gao, Q., Sun, W., Ballegeer, M., Libert, C. & Chen, W. Predominant contribution of cis-regulatory divergence in the evolution of mouse alternative splicing. *Molecular Systems Biology* **11**, 816 (2015).
3. Rozowsky, J. *et al.* AlleleSeq: analysis of allele-specific expression and binding in a network framework. *Mol Syst Biol* **7**, 522 (2011).
4. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
5. Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930 (2014).
6. Smith, T., Heger, A. & Sudbery, I. UMI-tools: modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy. *Genome Res* **27**, 491–499 (2017).

Appendix III: Genetic studies of human–chimpanzee divergence using stem cell fusions

Genetic studies of human–chimpanzee divergence using stem cell fusions

Janet H. T. Song[a,b,1], Rachel L. Grant[a,1], Veronica C. Behrens[a,1], Marek Kučka[c], Garrett A. Roberts Kingman[a], Volker Soltys[c], Yingguang Frank Chan[c], and David M. Kingsley[a,d,2]

[a]Department of Developmental Biology, Stanford University School of Medicine, Stanford, CA 94305; [b]Department of Genetics, Stanford University School of Medicine, Stanford, CA 94305; [c]Friedrich Miescher Laboratory of the Max Planck Society, 72076 Tübingen, Germany; and [d]HHMI, Stanford University School of Medicine, Stanford, CA 94305

Contributed by David M. Kingsley; received September 24, 2021; accepted November 10, 2021; reviewed by Philip Reno and Clifford Tabin

Complete genome sequencing has identified millions of DNA changes that differ between humans and chimpanzees. Although a subset of these changes likely underlies important phenotypic differences between humans and chimpanzees, it is currently difficult to distinguish causal from incidental changes and to map specific phenotypes to particular genome locations. To facilitate further genetic study of human–chimpanzee divergence, we have generated human and chimpanzee autotetraploids and allotetraploids by fusing induced pluripotent stem cells (iPSCs) of each species. The resulting tetraploid iPSCs can be stably maintained and retain the ability to differentiate along ectoderm, mesoderm, and endoderm lineages. RNA sequencing identifies thousands of genes whose expression differs between humans and chimpanzees when assessed in single-species diploid or autotetraploid iPSCs. Analysis of gene expression patterns in interspecific allotetraploid iPSCs shows that human–chimpanzee expression differences arise from substantial contributions of both *cis*-acting changes linked to the genes themselves and *trans*-acting changes elsewhere in the genome. To enable further genetic mapping of species differences, we tested chemical treatments for stimulating genome-wide mitotic recombination between human and chimpanzee chromosomes, and CRISPR methods for inducing species-specific changes on particular chromosomes in allotetraploid cells. We successfully generated derivative cells with nested deletions or interspecific recombination on the X chromosome. These studies confirm an important role for the X chromosome in *trans* regulation of expression differences between species and illustrate the potential of this system for more detailed *cis* and *trans* mapping of the molecular basis of human and chimpanzee evolution.

human–chimpanzee evolution | tetraploid | *cis*/*trans* gene regulation | genetic mapping

Humans have had a long-standing interest in the features that distinguish our species from other animals (1, 2). Comparative studies have characterized many morphological, physiological, and behavioral similarities and differences among great apes (3). Paleontological studies have traced the origin and timing of the appearance of various human features in the fossil record (4). More recently, advances in sequencing technologies have allowed for the comparative genomic analysis of humans, chimpanzees, other nonhuman primates, and even extinct archaic human lineages such as Neanderthals and Denisovans (5).

Whole-genome comparisons indicate that ∼4% of the base pairs in the human genome differ from those in chimpanzees. Sifting through this set of ∼125 million DNA changes to separate the causal mutations contributing to phenotypic differences between humans and chimpanzees from inconsequential or neutral changes is a daunting problem, and has been compared to searching for needles in a haystack (3).

In evolutionary studies of other organisms, genetic crosses between different lineages have helped localize and prioritize chromosome regions that influence different traits. The formation of F1 hybrids, followed by chromosome recombination during meiosis, can be used to produce F2 offspring that inherit different combinations of alleles from the parental lineages. By comparing different genotypes and phenotypes across a large panel of meiotic mapping progeny, it has now been possible to map some evolutionary traits to particular chromosome regions in yeast, fruit flies, butterflies, sticklebacks, mice, and other organisms (6).

Traditional meiotic mapping approaches are limited to organisms that can be crossed to produce viable and fertile offspring. However, related approaches have also been developed for comparing genotypes and phenotypes in somatic cells without meiosis, when traditional crosses are not possible. Cells of even distantly related organisms can be fused in vitro to produce somatic cell hybrids that contain the genetic information from both lineages. The fused cells sometimes lose chromosomes of one or the other starting species, producing progeny cell lines that can be used to assign genes or cellular phenotypes to particular chromosomes (7). Hybrids can also be irradiated to fragment chromosomes and stimulate additional segregation of genetic information, an approach that has been used for fine mapping of genomic linkage relationships (8). Mitotic recombination within cultured cells can also be stimulated by mutations in DNA pathways, by chemicals that damage DNA, or by targeted breaks

Significance

Comparative studies of humans and chimpanzees have revealed many anatomical, physiological, behavioral, and molecular differences. However, it has been challenging to map these differences to particular chromosome regions. Here, we develop a genetic approach in fused stem cell lines that makes it possible to map human–chimpanzee molecular and cellular differences to specific regions of the genome. We illustrate this approach by mapping chromosome regions responsible for species-specific gene expression differences in fused tetraploid cells. This approach is general, and could be used in the future to map the genomic changes that control many other human–chimpanzee differences in various cell types or organoids in vitro.

Author contributions: J.H.T.S., R.L.G., V.C.B., G.A.R.K., Y.F.C., and D.M.K. designed research; J.H.T.S., R.L.G., V.C.B., M.K., G.A.R.K., and V.S. performed research; J.H.T.S., R.L.G., V.C.B., M.K., G.A.R.K., V.S., Y.F.C., and D.M.K. contributed new reagents/analytic tools; J.H.T.S., R.L.G., V.C.B., M.K., and V.S. analyzed data; and J.H.T.S., R.L.G., V.C.B., and D.M.K. wrote the paper with input from all authors.

Reviewers: P.R., Philadelphia College of Osteopathic Medicine; and C.T., Harvard Medical School.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

[1]J.H.T.S., R.L.G., and V.C.B. contributed equally to this work.

[2]To whom correspondence may be addressed. Email: kingsley@stanford.edu.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2117557118/-/DCSupplemental.

Published December 17, 2021.

induced by Cas9 and guide RNAs (gRNAs) designed to alter particular locations in the genome. Mutations and chemical inhibitors of the Bloom Syndrome helicase gene (*BLM*) have been used to recover homozygous mutants in somatic cell gene screens (9, 10) or to induce recombination between chromosomes of distantly related mouse strains for studies of the genomic basis of evolutionary differences (11). The ability to induce breaks at particular loci with CRISPR-Cas9 has also made it possible to choose both the location and the direction of recombination between genomes in nonmeiotic cells, enabling high-resolution mapping without traditional crosses in yeast (12).

Development of similar approaches for human and chimpanzee cells would be very useful for studying the genomic basis of evolutionary differences that have evolved in hominids. Many molecular and cellular phenotypes that can be assayed and scored under cell culture conditions are known to differ between humans and chimpanzees. Recent studies have generated well-matched sets of human and chimpanzee induced pluripotent stem cell (iPSC) lines (13), and have shown that human and chimpanzee iPSCs can be fused to produce hybrids useful for comparing species-specific expression in cortical spheroids and neural crest cells (14, 15). Here we generate both autotetraploid (same species) and allotetraploid (different species) fusion lines from human and chimpanzee iPSCs, and use them to identify whether gene expression differences are due to *cis*- or *trans*-acting differences between species. We also test both random and targeted methods for stimulating DNA breaks and chromosome exchanges in allotetraploid iPSCs, providing a general method for further localizing the specific genomic changes that underlie human and chimpanzee differences in vitro.

Results

Generation and Initial Characterization of Autotetraploid and Allotetraploid iPSC Lines. To generate autotetraploids and allotetraploids, we labeled human and chimpanzee iPSC lines (13) with diffusible fluorescent dyes and fused them using electrofusion (Fig. 1*A* and *Materials and Methods*). Tetraploid cells were enriched by either fluorescence-activated cell sorting (FACS) or manual inspection and grown clonally. Successful fusion in expanded clones was confirmed by FACS analysis for DNA content using propidium iodide and by karyotyping. In total, we generated two human autotetraploid lines ("H1H1" lines, from human iPSC line H23555 [H1]); five chimpanzee autotetraploid lines ("C1C1" lines from chimpanzee iPSC line C3649 [C1]); and 22 human–chimpanzee allotetraploid lines from different fusion events including 12 "H1C1" lines derived from H1 and C1 and 10 "H2C2" lines derived from human iPSC line H20961 (H2) and chimpanzee iPSC line C8861 (C2) (Dataset S1).

Tetraploid iPSCs were larger than diploid cells but had normal morphology and could be routinely propagated under the same conditions as diploid iPSCs (*SI Appendix*, Fig. S1). We performed G-banded karyotyping on the initial diploid parental lines, as well as the newly generated autotetraploid and allotetraploid lines to examine their genome stability (Dataset S1). Fusion lines showed the tetraploid karyotypes expected from fusing their originating diploid lines. However, some of the tetraploid lines contained additional chromosomal abnormalities, including aneuploidies common to diploid human iPSC cultures (16) such as deletion of human chr18q (asterisk in Fig. 1*B*).

To assess the pluripotency and differentiation potential of the tetraploid iPSC lines, we differentiated representative diploid (H1, H2, C1, C2), autotetraploid (H1H1a, H1H1b, C1C1a, C1C1c), and allotetraploid (H1C1a, H1C1b, H2C2a, H2C2b) lines into ectoderm, mesoderm, and endoderm (*Materials and Methods*). Quantitative PCR (qPCR) for the expression of pluripotency (*NANOG*, *DNMT3B*), ectoderm (*PAX6*, *RAX*), mesoderm (*TBXT*, *HAND1*), and endoderm (*FOXA2*, *SOX17*) markers showed specific differentiation of tetraploid lines into all three lineages (Fig. 1*C*, Dataset S3, and *SI Appendix*, *Supplemental Materials and Methods*). For endoderm differentiation, a subset of lines (H1, C1C1c, H1C1b) showed lower expression of endoderm markers compared to all other cell lines, as well as persistent expression of pluripotency marker genes. Tetraploid cells thus retain broad differentiation abilities, but conditions may need to be optimized for particular cell lines or differentiation endpoints.

Diploid and Autotetraploid iPSC Lines Have Similar Gene Expression Profiles. To examine whether tetraploidization altered normal gene expression patterns, we used RNA sequencing (RNAseq) to characterize transcriptional differences due to ploidy, but not to species differences (i.e., H1 vs. H1H1 and C1 vs. C1C1). At a false discovery rate (FDR) of 5%, we detected 189 differentially expressed genes between H1 and H1H1, and 181 differentially expressed genes between C1 and C1C1, with at least a twofold change in expression (Dataset S4 and *SI Appendix*, *Supplemental Materials and Methods*). Neither set of differentially expressed genes was enriched for gene ontology categories (*SI Appendix*, *Supplemental Materials and Methods*), and only 13 genes were differentially expressed in both H1 compared to H1H1 and C1 compared to C1C1. We conclude that the creation of tetraploid cells alone does not activate a coordinated set of gene expression changes.

To assess gene expression variability between different cell lines from the same species, we also profiled global RNA patterns from a second set of human and chimpanzee diploid iPSC lines. We detected 410 differentially expressed genes between H1 and H2 and 181 differentially expressed genes between C1 and C2 at an FDR of 5% with at least a twofold change in expression (Dataset S4). Using principal component analysis, we found that global transcriptional profiles grouped by species, with human-derived lines clustering separately from chimpanzee-derived lines, and that diploid lines clustered more closely with their derived autotetraploid line than another diploid line of the same species (Fig. 2*A*). These results indicated that the transcriptional profiles of the diploid lines and their derived autotetraploid lines were at least as similar as the transcriptional profiles of two diploid lines from the same species. Taken together, our data suggest that tetraploid iPSCs behave similarly to diploid iPSCs at the level of gene expression.

Differential Gene Expression and Allele-Specific Gene Expression Reveal Human- and Chimpanzee-Specific Gene Expression Profiles. We next used our RNAseq data to identify gene expression differences between human and chimpanzee iPSCs (Dataset S5). Differential gene expression (DE) analysis between human-only and chimpanzee-only iPSC lines identified 5,984 genes differentially expressed between species. There were no significant gene ontology enrichments for DE genes with at least a twofold change in expression (*SI Appendix*, *Supplemental Materials and Methods*). Allele-specific expression (ASE) comparisons between the human allele and the chimpanzee allele in allotetraploid iPSC lines identified 4,540 allele-specific expressed genes. ASE results from this study and the ASE results from a previous study (14) that independently generated human–chimpanzee allotetraploid fusions from similar diploid iPSC lines were highly concordant (Pearson's $r = 0.72$; *SI Appendix*, Fig. S2), suggesting that human–chimpanzee gene expression differences are robust and reproducible across laboratories.

***cis*- and *trans*-Acting Regulatory Changes Are Both Important Contributors to Human–Chimpanzee Gene Expression Differences.** Determining whether gene expression differences between two species are due to *cis*-acting or *trans*-acting regulatory changes is possible when gene expression can be compared between each single species and a hybrid (17). We therefore leveraged

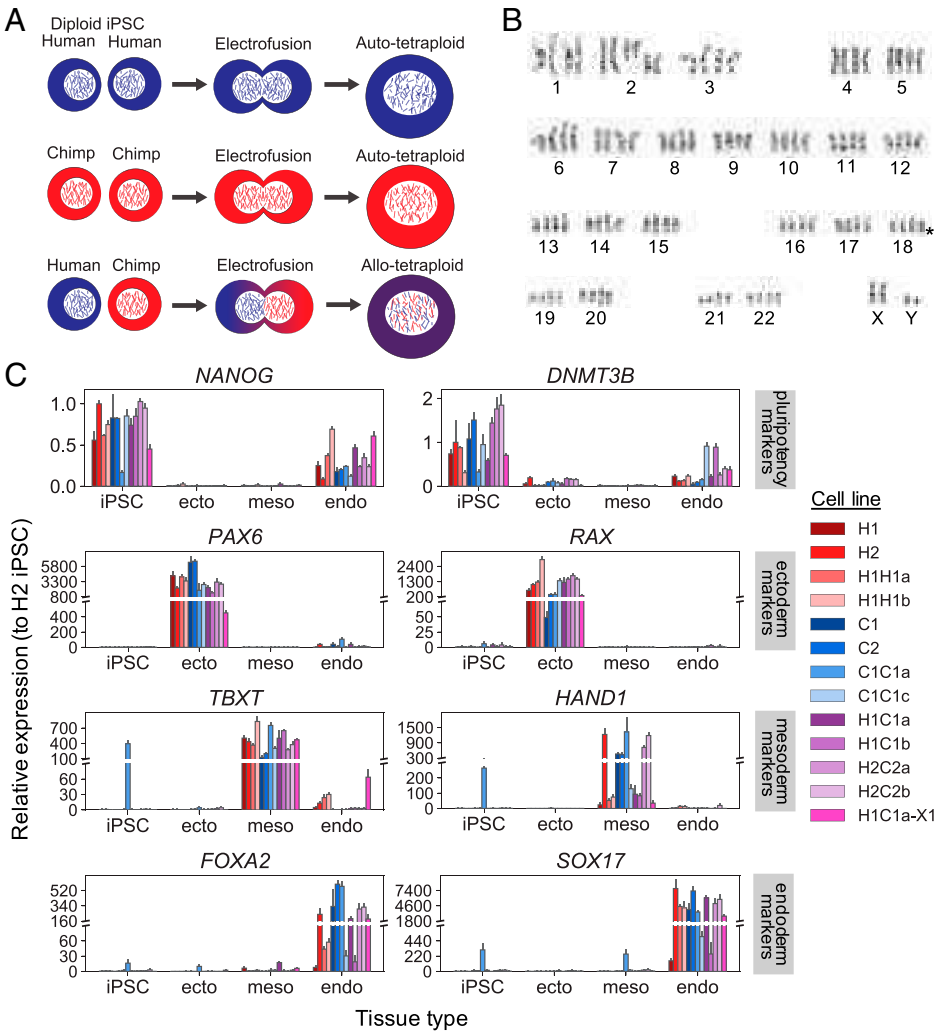


Fig. 1. Generation and differentiation of autotetraploid and allotetraploid iPSC lines. Autotetraploid and allotetraploid cells contain the expected number of chromosomes and express expected marker genes after trilineage differentiation. (*A*) Human and chimpanzee diploid iPSC lines were labeled with diffusible dyes and subjected to electrofusion to generate autotetraploid and allotetraploid iPSC lines. (*B*) Tetraploid lines (H2C2a shown) exhibit karyotypes with four copies of each chromosome. Asterisk denotes location of a common iPSC human chr18q deletion (16), present in a subset of our cell lines. See Dataset S1 for detailed karyotype description of all lines. (*C*) Relative expression of pluripotency (*NANOG*, *DNMT3B*), ectoderm (*PAX6*, *RAX*), mesoderm (*TBXT*, *HAND1*), and endoderm (*FOXA2*, *SOX17*) marker genes tested via qRT-PCR after incubating cell lines under trilineage differentiation conditions. Cell lines tested are two human diploid lines (H1, H2), two human autotetraploid lines (H1H1a, H1H1b), two chimpanzee diploid lines (C1, C2), two chimpanzee autotetraploid lines (C1C1a, C1C1c), four allotetraploid lines (H1C1a, H1C1b, H2C2a, H2C2b), and one fluorescently marked allotetraploid line (H1C1a-X1). Gene expression is plotted relative to a human diploid undifferentiated iPSC line (H2). Error bars represent the SD of $N = 3$ cell culture replicates maintained as iPSCs or differentiated independently; 146 of 156 gene expression differences between undifferentiated cells and the tissue type in which a marker is expected to be expressed are significant by two-tailed Student's *t* test at 5% FDR (see Dataset S3 for complete *P* value list).

the RNAseq data from human-only, chimpanzee-only, and human–chimpanzee allotetraploid iPSC lines to determine the regulatory type for genes that were differentially expressed between human-only and chimpanzee-only iPSCs (*Materials and Methods*). Specifically, when a *cis*-acting regulatory change causes a gene to be differentially expressed, the expression difference should be maintained in allotetraploid cells where both human and chimpanzee alleles are in the same *trans*-acting environment. Conversely, when a *trans*-acting regulatory change causes a gene to be differentially expressed, the expression difference should disappear in allotetraploid cell lines.

Our regulatory type classifications identified 5,956 genes with no net regulatory changes between our human-only and chimpanzee-only iPSC lines. Of these, 92.6% (5,515 genes) were classified as conserved between human and chimpanzee, and 7.4% (441 genes) were classified as compensatory (*cis*- and *trans*-regulatory differences acting in opposite directions resulting in no net expression difference between species) (Fig. 2*C*).

Of 4,671 genes with regulatory changes between human-only and chimpanzee-only iPSC lines, 44.4% (2,073 genes) were regulated primarily in *cis*, 31.4% (1,465 genes) were regulated primarily in *trans*, and the remaining 1,133 genes were regulated both in *cis* and in *trans* (Fig. 2*C*). This final category was further broken down into a *cis*+*trans* category (*cis*- and *trans*-regulatory changes acting in the same direction) and a *cis*−*trans* category (*cis*- and *trans*-regulatory changes acting in opposite directions). This yielded 20.6% (961 genes) and 3.7% (172 genes) regulated in *cis*+*trans* and *cis*−*trans*, respectively. Other genes that did not satisfy the conditions for any category (3,515 genes) were classified as ambiguous.

Genes with primarily *cis*-regulatory changes had a larger median effect size than genes with primarily *trans*-regulatory changes (median $|log_2(FC)|$ of 1.09 vs. 0.64, $P < 10^{-56}$ by two-tailed Mann–Whitney *U* test; Fig. 2*D*). Genes classified as *cis*+*trans* had the highest effect size of any regulatory type category (median $|log_2(FC)|$ of 1.21).

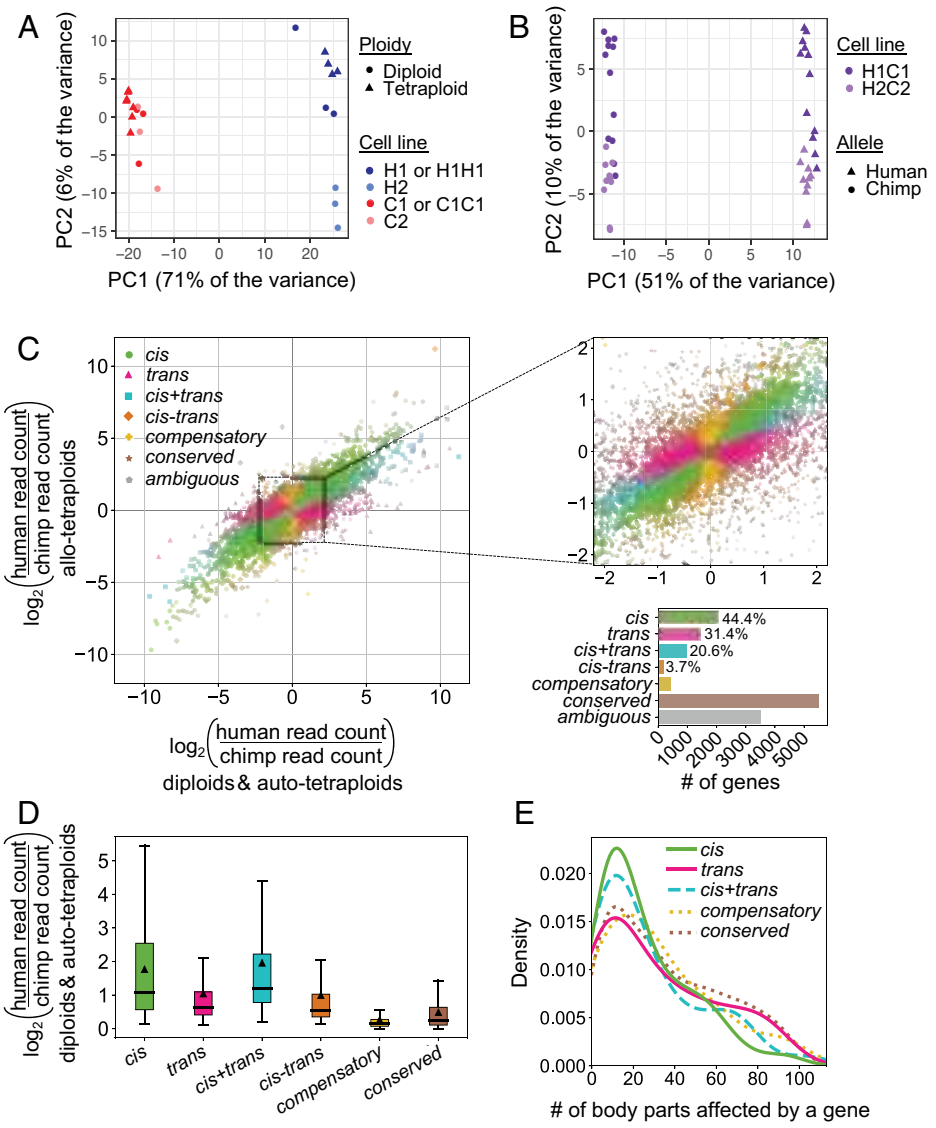


Fig. 2. Gene expression profiling of human and chimpanzee diploid, autotetraploid, and allotetraploid iPSC lines. Tetraploidization does not result in coordinated gene expression changes, but thousands of genes are expressed differently between human and chimpanzee iPSCs due to a mixture of *cis*- and *trans*-regulatory changes. (*A*) Principal component analysis (PCA) of RNAseq of H1, H2, C1, C2, H1H1, and C1C1 diploid and autotetraploid iPSC lines. The cell lines cluster by species along PC1 and by cell line along PC2. Autotetraploid lines cluster with their cognate diploid line. (*B*) PCA of RNAseq of H1C1 and H2C2 allotetraploid lines. Allotetraploid lines are each represented by two dots, one for reads mapping to the human transcriptome and one for reads mapping to the chimpanzee transcriptome (*Materials and Methods*). Expression from human alleles (triangles) cluster separately from chimpanzee alleles (circles) in allotetraploid lines along PC1. PC2 separates the two sets of allotetraploid cell lines. (*C*) Each gene's expression pattern was classified by regulatory type (*cis*, *trans*, *cis+trans*, *cis–trans*, compensatory, conserved, or ambiguous) by comparing DE between human- and chimpanzee-only iPSCs (*x* axis) and allele-specific gene expression between human and chimpanzee alleles within allotetraploid iPSCs (*y* axis). (*Left*) Data for all genes. (*Upper Right*) Zoom-in of dense center region. (*Lower Right*) Bar graph indicating number of genes per category and relative contribution (percentage) of each category to genes with human–chimpanzee regulatory differences. (*D*) Box plot showing distribution of effect sizes for gene expression changes in each regulatory category. Median effect size is indicated by thick horizontal lines, and mean effect size is indicated by triangles. All pairwise comparisons are statistically significant (adjusted $P < 0.012$ by two-tailed Mann–Whitney *U* test). (*E*) Density plot (smoothed histogram) showing the distribution of body parts influenced by genes [according to the Gene ORGANizer database (20)] in each regulatory category. For genes classified as *cis*, *trans*, and *cis+trans*, only genes with $|log_2(FC)| \geq 1$ are plotted. The *cis–trans* category is not included because only five genes have $|log_2(FC)| \geq 1$. Note that genes classified as *cis* or *cis+trans* tend to influence fewer body parts than conserved genes (median 18 body parts for both *cis* and *cis+trans* genes compared to median 30 body parts for conserved genes, adjusted $P = 0.00028$ and $P = 0.0035$ by two-tailed Mann–Whitney *U* test after FDR correction). This trend is not observed for *trans* and compensatory regulatory types (median 24 and 27 body parts, adjusted $P = 0.11$ and $P = 0.21$, respectively).

Gene ontology enrichments for genes classified as *trans* included processes related to the skeletal, cartilage, and muscular systems (Dataset S6). Although we did not assess gene expression differences in skeletal, cartilage, or muscle cells, previous studies that assessed regulatory differences between human and chimpanzee embryonic stem cells similarly found gene ontology enrichments associated with differentiated tissues, including the vocal tract (18). The enrichments seen in the current experiments suggest that some of the dramatic skeletal and muscular differences between humans and chimpanzees may be driven by *trans*-acting regulatory changes. Additionally, genes classified as conserved had gene ontology enrichments related to voltage-gated ion channels (Dataset S6), which are important for maintaining critical features of iPSCs, including

proliferation capacity and differentiation potential (19). Finally, genes classified as compensatory had enrichments related to ligase activity, neurexin protein binding, and phosphatidylserine binding, while all other regulatory type classifications had no significant gene ontology enrichments.

We also used the Gene ORGANizer database (20), which links genes to the body parts they affect based on phenotypes associated with Mendelian disorders, to test whether genes that were differentially expressed between humans and chimpanzees tend to influence more or fewer biological systems than conserved genes. We found that genes with primarily *cis*-regulatory changes and at least a twofold change in expression influenced a median of 18 body parts compared to a median of 30 body parts influenced by conserved genes (adjusted $P = 2.8 \times 10^{-4}$ by two-tailed Mann–Whitney U test; Fig. 2*E*). Interestingly, the greater the expression differences between human and chimpanzee as indicated by higher $|log_2(FC)|$, the fewer body parts a gene with primarily *cis*-regulatory changes tended to influence (*SI Appendix*, Fig. S3). Similar trends were observed for genes classified as *cis*+*trans* but not other regulatory categories.

Removing reads mapping to genes on chromosomes that were karyotypically abnormal in any of our iPSC lines did not significantly change our regulatory type classification or effect size results (*SI Appendix*, Fig. S4). Together, our results indicate that both *cis*- and *trans*-acting regulatory changes are important contributors to the widespread gene expression differences between humans and chimpanzees in iPSCs, with *cis*-regulatory changes tending to be larger and to act on genes affecting fewer biological systems (Fig. 2 *C*–*E*).

Prospects for Genetic Mapping. Further localization of both *cis*- and *trans*-regulatory differences would be greatly aided if it were possible to generate mapping panels that carry different dosages of human and chimpanzee alleles at known locations throughout the genome. Previous studies in yeast, *Drosophila*, and cultured mammalian cell lines have used mitotic recombination to generate useful mapping panels from somatic cells (11, 12, 21, 22). To boost the rate of mitotic recombination, common strategies have been to treat cells with small molecules that promote DNA damage (23, 24), or to induce targeted recombination at specific loci using CRISPR-Cas9 (12, 21).

To assess whether small molecules could stimulate mitotic cross-overs in iPSCs, we performed sister chromatid exchange (SCE) assays by incubating cells with BrdU for two cell cycles (25). Chromosomes where both strands incorporate BrdU stain lighter than chromosomes where only one strand has incorporated BrdU, making it possible to visualize SCE events in mitotic chromosome spreads (Fig. 3*A*). We tested camptothecin, a topoisomerase inhibitor previously found to induce SCE events in iPSCs (23). Consistent with prior findings, treatment of 100 nM camptothecin for 1 h induced a 4.5-fold increase in SCE events in both autotetraploid and allotetraploid iPSCs ($P < 10^{-8}$ by one-tailed Student's t test; Fig. 3*B*).

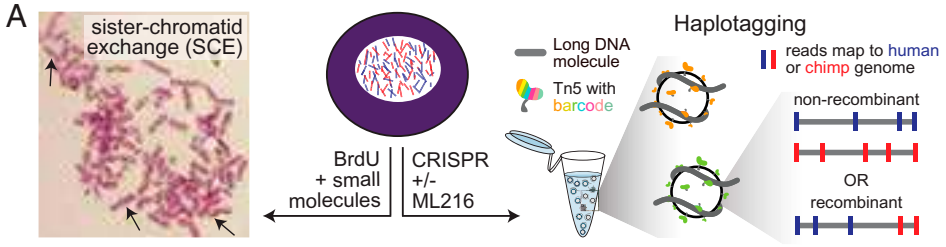
We also tested ML216, a BLM inhibitor, which has been found to induce SCE events in cultured human cells (9, 10, 24). However, we found that treatment with ML216 over a range of concentrations from 12.5 μM to 150 μM did not increase the rate of SCE events in iPSCs. We additionally tested mitomycin C, which cross-links DNA and is known to induce SCE events in yeast and fungi (26). Treatment of 4 ng/mL mitomycin C for 24 h in tetraploid iPSCs increased the rate of SCE events twofold ($P < 10^{-5}$ by one-tailed Student's t test; Fig. 3*B*). Although SCE assays can only reliably assess intraspecific cross-over events, these results suggest that the application of camptothecin or mitomycin C to allotetraploid iPSCs has the potential to similarly increase the rate of interspecific mitotic recombination.

An alternate approach is to induce targeted cross-overs using CRISPR-Cas9. This strategy has previously been used to induce recombination in yeast and *Drosophila* (12, 21). To determine the rate of interspecific recombination events at target loci, we used a recently developed technique called haplotagging to directly detect recombinant junctions by barcoding DNA molecules prior to sequencing (27, 28). Following sequencing, reads were aligned to a composite human–chimpanzee genome and comparatively assigned to their species of origin. Reads derived from the same DNA molecule were tagged with the same barcode, enabling molecule reconstruction (*Materials and Methods*). Barcoded molecules that mapped to orthologous intervals in human and chimpanzee and showed switched runs of variants from one species to the other (human to chimpanzee, or vice versa) were scored as likely interspecific recombination events within the corresponding genomic interval.

In the allotetraploid line H1C1a, we targeted genomic loci on chr20q13.33, chr21q22.3, and chrXq28 with CRISPR gRNAs and then performed haplotagging to over 200× molecular coverage (*SI Appendix*, Figs. S5 and S6). Based on a recent study suggesting that ML216 acts synergistically with CRISPR-Cas9 to induce loss of heterozygosity at targeted loci in human iPSCs (22), we also assessed whether the addition of 25 μM of the BLM inhibitor ML216 starting 12 h before gRNA targeting and ending 48 h posttargeting would affect the rate of interspecific recombination. We did not observe an enrichment in interspecific recombination events at any of the target loci with or without ML216 treatment (*SI Appendix*, Fig. S6). However, genome-wide interspecific recombination events trended 1.35-fold higher when comparing ML216-treated samples against samples that were only treated with CRISPR-Cas9 ($P = 0.052$ by one-tailed paired Student's t test; Fig. 3*B*). After ML216 treatment for 60 h (approximately three cell divisions), we detected a total of 878 interspecific recombination events in ∼83 million analyzed molecules. This translates to an endogenous rate of ∼0.8 recombination events per cell per generation and an increased rate of approximately one recombination event per cell per generation after ML216 treatment. These apparent rates in allotetraploid cells are substantially higher than previously reported mitotic recombination rates in diploid mouse embryonic stem cells [0.01 to 0.04 recombination events per cell per generation after ML216 treatment; *SI Appendix*, *Supplemental Materials and Methods* (10, 11, 29)]. Further investigation will be required to assess whether ML216 significantly increases the rate of interspecific recombination and whether other small molecules such as camptothecin can also increase the rate of recombination events between human and chimpanzee chromosomes in allotetraploid iPSCs.

Targeted *cis*- and *trans*-Mapping on the X Chromosome. To further enrich for cells that may contain interspecific mitotic recombination events at specific loci, we fluorescently marked allotetraploid cells at distal chromosome ends. This allowed us to use FACS to isolate cells with expected signatures of recombination (Fig. 4*A*). The gene density and enrichment of disease-related genes on distal chrX, particularly chrXq28, made the distal region of chrX a particularly attractive target for further study (30). Through two rounds of CRISPR-Cas9–mediated homologous recombination (HR), we generated five allotetraploid lines derived from H1C1a and one from H2C2b, each carrying *GFP* on the human chrX and *mCherry* on the chimpanzee chrX (*SI Appendix*, Fig. S7). Some allotetraploid cells that underwent two rounds of CRISPR-Cas9 HR insertion maintained largely normal karyotypes, while others showed more extensive aneuploidies (Dataset S1).

We then targeted the double fluorescently marked lines with species-specific gRNAs to induce interspecific recombination events on the X chromosome. Because the allotetraploids were derived from fusions of male cells, only one X chromosome was present from each species. In the double fluorescently marked lines, cells with no recombination events on chrX should



B

Treatment	Assay	Range	Best	Recomb. prop.	Ratio to ctrl	N
BrdU only	SCE	---	---	0.106 ± 0.008	---	16
+ camptothecin	SCE	25-500 uM	100 uM	0.486 ± 0.017	4.58 ± 0.15	7
+ ML216	SCE	12.5-150 uM	---	0.061 ± 0.005	0.58 ± 0.04	10
+ mitomycin C	SCE	1-8 ng/ml	4 ng/ml	0.213 ± 0.022	2.01 ± 0.20	7
CRISPR only	Haplotagging	---	---	$6.89 \pm 3.14 \times 10^{-6}$	---	4
+ ML216	Haplotagging	25 uM	---	$9.17 \pm 4.14 \times 10^{-6}$	1.35 ± 0.02	4

Fig. 3. Effect of small molecules on chromosome recombination frequencies in allotetraploid cells. The small molecules camptothecin, ML216, and mitomycin C were assessed for their effect on intraspecific and interspecific recombination. (*A*) Allotetraploid cells (*Center*) were treated with BrdU and small molecules to measure intraspecific SCE events by microscopy (*Left*), or treated with Cas9 and gRNAs with or without ML216 followed by haplotagging to identify interspecific recombinant molecules by sequencing (*Right*) (*Materials and Methods*). (*B*) The proportion of chromosomes that had SCE events after treatment with the indicated concentrations of each small molecule is listed (recomb. prop.). N denotes number of replicate experiments where the recomb. prop. was quantified for all concentrations or the best concentration (when indicated). Note that BrdU was added to all cells to visualize SCE, and all BrdU+drug treatments were compared to the BrdU-only condition (ratio to ctrl). For haplotagging, the effect of CRISPR guides targeting specific loci was assessed with or without ML216. Compared to CRISPR alone, ML216 may elevate the proportion of interspecific recombinant molecules genome wide (ratio to ctrl). Values are mean ± SEM.

carry both human and chimpanzee fluorescent markers, while recombinant cells should carry two copies of a single marker from either human or chimpanzee chrX. We treated the double fluorescently marked allotetraploid line H1C1a-X1 with either a chimpanzee-specific gRNA targeting chrXq28 or a human-specific gRNA targeting chrXq22.1, in combination with 25 μM ML216 (*SI Appendix*, Fig. S8 and *Materials and Methods*).

Cells targeted with the chimpanzee-specific gRNA were sorted for the absence of mCherry, which marks the chimpanzee chrX, and increased intensity of GFP, to select for likely recombination events that result in two human alleles on the distal end of chrXq. Because we observed higher fluorescence intensity of the human chrX marker GFP in untreated cells during the G2/M cell cycle phase, we also used Hoechst DNA staining to sort specifically from G1 cells in the experiments with the chimpanzee-specific gRNA (*SI Appendix*, Fig. S9 and *Materials and Methods*). Similarly, cells targeted with the human-specific gRNA were sorted for the absence of GFP, which marks the human chrX, and increased intensity of mCherry. For the human-specific gRNA sorts, we also incorporated an additional marker by staining for a linked cell-surface protein, TSPAN6. Located on chrXq, *TSPAN6* has 1.4-fold higher *cis*-regulated expression from the chimpanzee allele compared to the human allele (adjusted $P = 1.4 \times 10^{-4}$ by Welch's t test); protein staining of TSPAN6 showed a similar difference (*SI Appendix*, Fig. S9 and *Materials and Methods*). Cells targeted with human-specific gRNA were thus sorted for absence of human marker GFP and increased intensity of both chimpanzee marker mCherry and of TSPAN6.

A total of 951 allotetraploid candidate colonies were grown from single cells after FACS. Additional genotyping confirmed that 172 colonies carried distal chrXq from a single species (Dataset S7). As expected, in 172/172 (100%) of these cases, the missing chrXq corresponded to the species targeted by the gRNA (79/79 for the human gRNA, and 93/93 for the chimpanzee gRNA). To distinguish between deletion and recombination events, we determined the relative dosage of chrXq in these colonies by performing qPCR assays on genomic DNA at chr6p, chrXp, and chrXq (Dataset S2 and *SI Appendix*, *Supplemental Materials and Methods*). We found that 171/172 (99.4%) colonies had lost the distal end of chrX of one species without altering the chrX dosage of the other species, as expected if targeting of the X chromosome had produced a species-specific deletion in these colonies.

We also identified one colony (0.6%) that had not only lost the distal end of the chimpanzee chrX but also doubled the dosage of the distal end of human chrX, consistent with a possible recombination event. Whole-genome sequencing of this putative recombinant line confirmed that it was an interspecific recombinant, with the first 140.1 Mb of human chrX fused to the distal 27.6 Mb of chimpanzee chrX (Xrec1 in Fig. 4 *B* and *C* and *SI Appendix*, Fig. S10). Sequence reads that span the precise junction between the human and chimpanzee sequences show that recombination did not occur in a region of large-scale homology between the two X chromosomes. Instead, a 4 bp microhomology occurs directly at the junction site, suggesting that the recombination event was likely produced by microhomology-mediated end joining and not by HR (31). No other human–chimpanzee recombinant chromosomes were found in the sequenced Xrec1 line when compared to an untreated control line, suggesting that recombination events elsewhere in the genome are rare in cells that survive chrX targeting, FACS, and plating and growth of colonies from single cells.

We next leveraged the species-specific targeted lines as a panel of deletion lines for fine-mapping studies. We performed bulk RNAseq on seven lines with partial chimpanzee chrX deletions, eight lines with partial human chrX deletions, and nine control lines without chrX deletions (Dataset S1). We identified the approximate breakpoint of each deletion by examining the ratio of reads that uniquely map to either the human or the chimpanzee genome along chrX (Fig. 4*B*, *SI Appendix*, Fig. S11, and *Materials and Methods*). In every case, mapped breakpoints were consistent with our results from genomic PCR and qPCR assays (Dataset S7). Three of the seven chimpanzee chrX deletions mapped within 1 Mb of the chimpanzee-specific gRNA target site, and two of the eight human chrX deletions mapped within 1 Mb of the human-specific gRNA target site (Fig. 4*B*). The remaining lines had a range of breakpoints that were up to 80 Mb away from the species-specific guide targeting sites, and one cell line appeared to have lost the targeted X chromosome completely (Fig. 4*B*).

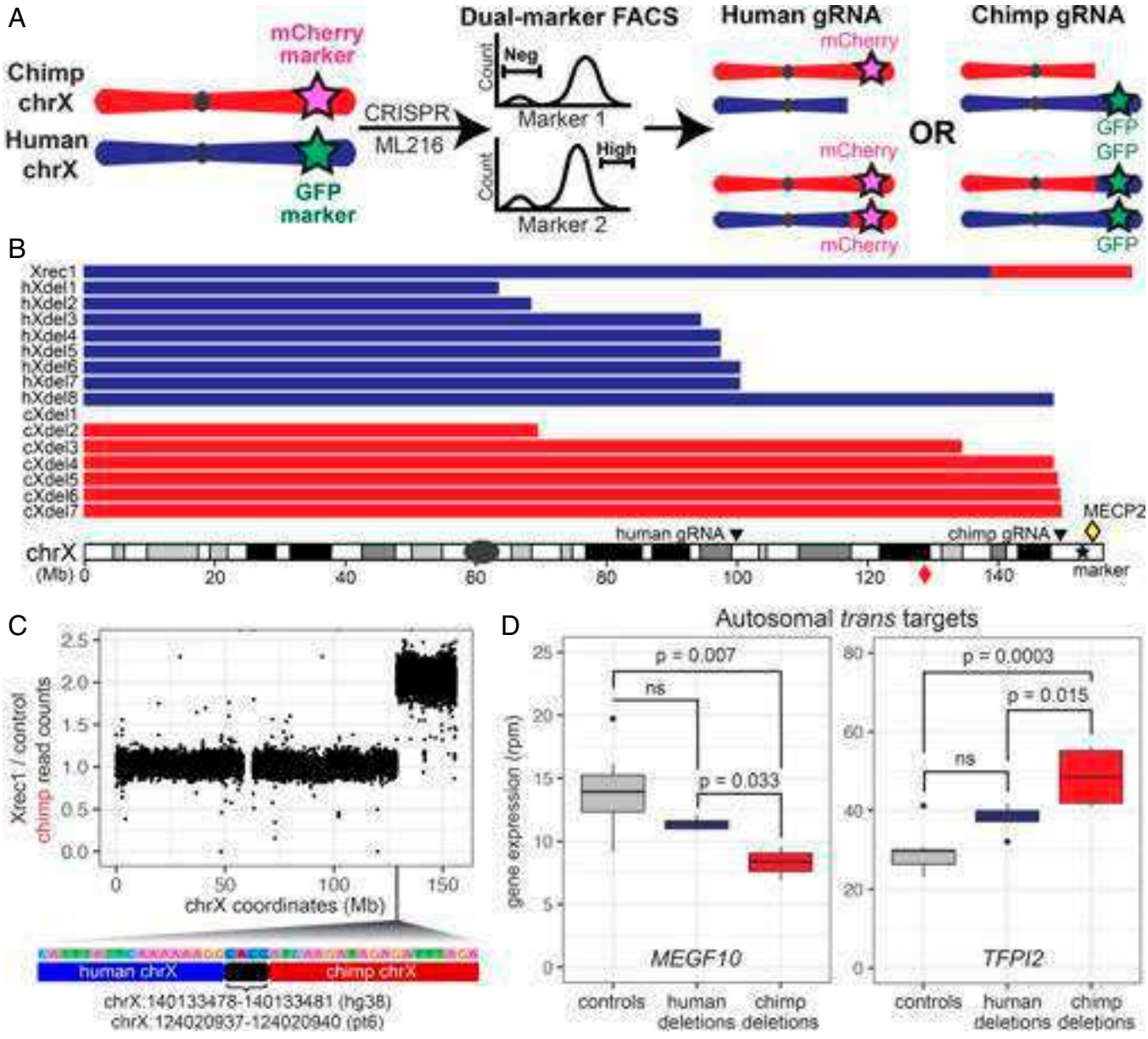


Fig. 4. X chromosome targeting generates recombinant and deletion lines for genetic mapping. Cell lines with recombination and deletion events on the X chromosome were isolated by FACS and used for further mapping of regulatory sequences. (*A*) Allotetraploid cells marked with GFP on the distal human chrX and mCherry on the distal chimpanzee chrX were treated with ML216, Cas9, and species-specific chrX gRNAs, followed by FACS for expected signatures of deletion or recombination (loss of fluorescent marker on targeted chrX, or retention or gain of fluorescent marker from homologous, untargeted chrX). (*B*) Cell lines recovered from sorting contained either a recombinant chrX or species-specific distal chrX deletions. Breakpoint locations for the recombinant (Xrec1), human deletions (hXdel#), and chimpanzee deletions (cXdel#) are shown relative to human chrX. Positions of species-specific gRNAs are shown. Red diamond symbol indicates the distal portion of chimpanzee chrX that is recombined onto the first 140.1 Mb of human chrX in Xrec1. Yellow diamond symbol indicates the position of *MECP2* on human chrX. (*C*) Whole-genome DNA sequencing of Xrec1 and a control sample (X1-S; *Materials and Methods*) showed an increase in chimpanzee read depth ratio along the X chromosome and identified human–chimpanzee spanning sequence reads at the point of transition, locating the precise point of cross-over for a human–chimpanzee recombinant X chromosome (position along chrX shown in hg38 coordinates; bracket indicates 4 bp of microhomology found in both human and chimpanzee chrX at the indicated coordinates). (*D*) Expression of autosomal genes *MEGF10* and *TFPI2* was significantly different in four lines that have lost distal chimpanzee chrX sequences (cXdel4–cXdel7) when compared to nine control lines without deletions or to five lines that have lost distal human chrX sequences (hXdel3–hXdel7), as expected if a *trans*-regulatory factor that differs between humans and chimpanzees maps to distal chrX (*SI Appendix*, *Supplemental Materials and Methods*).

To further characterize the breakpoints in X chromosome deletion lines, we performed whole-genome DNA sequencing of cXdel5 and cXdel6, which were plated and expanded from single cells after FACS selection. DNA sequencing indicated that cXdel5 cells had lost chimpanzee sequences distal to 147 Mb, retained chimpanzee sequences proximal to 140 Mb, and likely contained a subclonal mixture of deletion breakpoints in between. The cXdel6 cells had lost chimpanzee sequences distal to 148 Mb, retained sequences proximal to 147 Mb, and likely contained a subclonal mixture of insertions in between (*SI Appendix*, Fig. S12*A*). CRISPR targeting can thus induce terminal chromosome deletions, with staggered endpoints forming in the region around the breakpoints.

The X chromosome breakpoints in cXdel5 and cXdel6 cells were located near the genes *FMR1* and *AFF2*. To test whether species-specific chromosome deletions cause species-specific

changes in gene expression near the breakpoints, we examined the level of expression of the human and chimpanzee alleles of *FMR1* and *AFF2* in cXdel4, cXdel5, cXdel6, cXdel7, and control cells. The human alleles of *FMR1* and *AFF2* showed normal expression in the chimpanzee chrX deletion lines compared to control cells. In contrast, the chimpanzee alleles of *FMR1* and/or *AFF2* were not expressed when the genes were located distal to the deletion breakpoint in cXdel4, showed reduced or absent expression when the genes were located in the region of staggered deletions in cXdel5, and showed normal expression when the genes were located proximal to the terminal deletions in cXdel6 and cXdel7 (*SI Appendix*, Fig. S12*B*). Chimpanzee-specific chrX deletions thus can disrupt gene expression in *cis* without resulting in compensatory up-regulation of the corresponding human allele on the remaining X chromosome.

If the X chromosome encodes *trans* regulators of autosomal gene targets, partial deletions of either the human or chimpanzee X chromosome could result in significant gene expression changes for genes located on autosomes. Indeed, 42 autosomal genes showed significant changes in expression in the four deletion lines that removed regions on the chimpanzee X chromosome distal to breakpoints around 148 Mb when compared to control lines without chrX deletions, and even more autosomal genes (147) showed significant changes in expression in the five cell lines that removed regions on the human X chromosome distal to breakpoints around 95 Mb (Dataset S8). Interestingly, seven of the genes altered by loss of distal chimpanzee chrX regions were not significantly different in the cell lines that had lost even larger regions of the human X chromosome. These genes also showed the expected signatures of a species-specific *trans* effect when comparing gene expression levels among the different deletion lines (*SI Appendix*, *Supplemental Materials and Methods*). The autosomal genes included *MEGF10* and *TFPI2*, which were both classified as having a significant *trans* component in our studies of intact diploid, autotetraploid, and allotetraploid cells (Fig. 4*D* and Dataset S5). The magnitude of differential expression seen after species-specific removal of the distal X chromosome ranged from 60 to 80% of the overall expected *trans* component. Thus, *trans* regulators encoded on the X chromosome may contribute to a fraction of the species-specific *trans*-expression differences observed in these autosomal genes. Extensions of this approach could be used to further localize the responsible *trans* factors on the X chromosome, as well as *trans* factors on other chromosomes.

Discussion

Understanding the molecular basis of human evolution is a grand and ambitious challenge in biological research. At the molecular level, researchers have cataloged the DNA sequence changes between humans and nonhuman primates (5) and identified many RNA expression differences between humans and chimpanzees across multiple tissues and developmental stages (13, 32, 33). However, it has been difficult to map the exact sequence changes that cause particular gene expression differences or other species-specific traits. Here, we have used intraspecific and interspecific iPSC fusions to determine whether human–chimpanzee gene expression changes are controlled in *cis* or *trans*, and have developed genetic methods for further mapping both *cis* and *trans* effects to particular locations in the genome.

Regulatory changes appear to be a key driver of evolution in humans and other systems (34), and we and others have worked to determine the relative contribution of *cis*- and *trans*-acting regulatory changes to gene expression differences between species. As in previous studies with human and chimpanzee iPSCs (13, 14), we found thousands of genes with species-specific expression differences. By comparing DE in single-species and cross-species fusions, we found that 1) both *cis*- and *trans*-regulatory changes are key contributors to human–chimpanzee differences, and 2) genes with *cis*-regulatory changes had, on average, more divergent expression than genes with *trans*-regulatory changes. Both of these findings are consistent with previous genome-wide studies of human–chimpanzee tetraploid cortical spheroids, human–chimpanzee tetraploid cranial neural crest cells, and interspecific hybrids of mice, maize, *Arabidopsis*, and yeast (14, 15, 17, 35).

We also found that genes with *cis*-regulatory changes tended to influence fewer body parts than genes conserved between human and chimpanzee iPSCs. Furthermore, the number of body parts affected by a gene declines as the expression difference between humans and chimpanzees increases. *cis*-regulatory changes are often thought to be favored in evolution because of their ability to avoid negative pleiotropy and restrict changes to particular tissues (34, 36). Our data suggest that genes that influence fewer biological processes are also more likely to evolve large expression differences as species diverge during evolution.

To facilitate further genetic mapping of human–chimpanzee differences, we examined multiple strategies to induce recombination events in allotetraploid iPSCs, including both genome-wide and targeted approaches. The BLM inhibitor ML216 has been successfully used to induce interchromosomal recombination in other systems (11, 22). In our experiments, ML216 treatments did not cause a measurable increase in intrachromosomal exchange events scored by SCE assays, but may have stimulated a modest ∼35% increase in the number of human–chimpanzee recombinant molecules identified by haplotagging (Fig. 3). These differences could be nonsignificant, or might result from molecular differences between intraspecific and interspecific recombination events; confounding effects of BrdU, which promotes DNA damage and differentiation (*SI Appendix*, Fig. S13), in the SCE assay (37); or synergistic effects with CRISPR-Cas9 targeting, as previously reported for other mitotic recombination assays in human iPSCs (22). Further varying BLM activity by either pharmacological or genetic strategies (9, 10) or by treating with ML216 for multiple passages could be tested for larger effects on the overall rate of recombination. Camptothecin and mitomycin C treatments are also promising candidates for further study, given their strong promotion of SCE events in allotetraploid iPSCs (Fig. 3*B*).

We also tested the ability of Cas9 and gRNAs to stimulate interspecific chromosome exchange events at particular locations in the genome. In contrast to prior work in yeast, *Drosophila*, and human iPSCs (12, 21, 22), we did not observe an enrichment in interspecific recombination events at the site of targeting with or without ML216 by analyzing bulk populations with haplotagging. We also did not recover recombination events at the site of CRISPR targeting in the fluorescently marked lines that we sorted to enrich for signatures of rare recombination events on the X chromosome. We did recover many lines that carried species-specific X chromosome deletions with breakpoints near the site of CRISPR targeting. We further recovered a single line carrying a confirmed human–chimpanzee recombinant X chromosome. However, the cross-over junction in the recombinant line was located tens of megabases away from the CRISPR targeting site and may be the result of a spontaneous or ML216-induced, rather than a CRISPR-induced, breakpoint on the X chromosome.

Our overall rates of recovering targeted X chromosome changes in allotetraploid lines were low (from ∼78 million input cells, 951 colonies survived FACS selection and plating, of which a single colony contained a recombinant chromosome and 171 colonies contained deletion chromosomes). We note that estimated rates of interspecific recombination appeared orders of magnitude higher when bulk cells were analyzed by haplotagging shortly after ML216 treatment (approximately one genome-wide interspecific recombination event per cell per generation). Interspecific recombination rates in allotetraploid

cells may be overestimated by haplotagging, due to barcode sharing between DNA molecules or errors in assigning reads to the correct species when comparing allotetraploid cells with reference genomes. Alternatively, high rates of interspecific recombination may be incompatible with long-term growth and survival in allotetraploid cells such that only cells with low numbers of recombinant chromosomes survive FACS selection, plating, and growth at clonal density after treatments. Despite the low overall rate of recovering useful cells in the targeting experiments, our experiments show that informative panels can be successfully generated by treating large numbers of cells and selecting for changes on particular chromosomes.

A variety of strategies may make it possible to increase the rate of homology-directed interspecific recombination. Following induction of double-strand breaks by small molecules or Cas9, homology-directed repair (HDR) pathways compete with several other pathways, including nonhomologous end joining (NHEJ) (38). Studies in other systems have shown that the HDR pathway can be stimulated by expressing a plasmid with *RAD18*, a gene involved in the DNA damage response, or by treating cells with the small-molecule RS-1 which increases the activity of the HDR-promoting protein RAD51 (39, 40). Conversely, the competing NHEJ pathway can be suppressed using the small-molecule Scr7 to inhibit DNA Ligase IV, a key component of NHEJ (39). Studies in yeast show that tethering Cas9 to Spo11, a DSB-inducing protein with a key role in initiating meiotic recombination, can stimulate cross-overs in naturally recombination-cold regions (41). These and other approaches can now be tested for their ability to stimulate targeted recombination between human and chimpanzee chromosomes in allotetraploid cells.

Like genetic mapping using recombinants, deletion mapping has also been used to map phenotypes to specific genomic regions in many organisms (42, 43). Our targeting and sorting strategies have already successfully produced a panel of deletion lines useful for further mapping of *cis* effects and *trans*-acting factors on the X chromosome. The fraction of the genome removed by the induced chrX deletions is similar to the fraction of the genome removed by typical deficiency mapping chromosomes in *Drosophila* [0.2% of the genome deleted, on average (43)]. The staggered deletions that form after chromosome targeting, both in different colonies and within the same colony (e.g., cXdel5) after FACS selection, could be harnessed for further fine mapping of *cis*-regulatory effects in a chromosomal region of interest.

Panels of chromosome deletion lines can also be used to map species-specific *trans* regulators. *trans* effects appear to contribute to more than 50% of the gene expression differences identified between humans and chimpanzees in iPSCs (Fig. 2*C*), and are similarly pervasive in other systems (14, 15, 17, 35, 44). Our targeted X chromosome deletion lines suggest that human–chimpanzee differences in the autosomal genes *MEGF10* and *TFPI2* are controlled, in part, by species-specific *trans* effects that map to the most distal ∼8 Mb of the X chromosome. One of the genes located in this distal X chromosome region is *MECP2*, which encodes a methyl DNA-binding protein that can activate or repress expression of target genes (45). Loss-of-function mutations in *MECP2* lead to Rett syndrome, a severe neurodevelopmental disorder. Intriguingly, prior research has identified both *MEGF10* and *TFPI2* as genes regulated by MeCP2 in human cells (46, 47). Further, both MEGF10 and MeCP2 have been linked to the pruning of neural synapses by astrocytes (48, 49), a cell type that has undergone changes in number, spatial organization, and function during human evolution (50, 51). Given that gene regulation in iPSCs cells has been shown to be similar to that in somatic tissues in some contexts (52), it is tempting to speculate that this potential *trans* regulation might contribute to human–chimpanzee astrocyte differences or changes in neural processes and circuits pruned by astrocytes. Future experiments to selectively knock out either the human or chimpanzee *MECP2* allele could test whether MeCP2 indeed regulates the species-specific expression of *MEGF10* and *TFPI2* in iPSCs, as well as potentially identify other species-specific *trans* targets for this key transcriptional regulator.

We have focused most of our current studies on gene expression differences that are detectable in undifferentiated iPSCs. It is possible that tetraploidization will disrupt gene expression or limit the differentiation potential of autotetraploid and allotetraploid iPSC lines. However, previous studies have shown that tetraploid mouse embryos can form most major organs, and rare humans with tetraploid karyotypes have been reported to survive for up to 2 y after birth (53, 54). In addition, our global RNA profiling experiments showed no large-scale gene expression disruptions between diploid and autotetraploid lines. We also find that diploid lines are more similar to their cognate autotetraploid lines than to other diploid lines of the same species. Thus, diploid and tetraploid iPSCs appear remarkably similar at the gene expression level. Future studies will be needed to determine whether this similarity is maintained under a variety of differentiation conditions. Our initial experiments show that diploid, autotetraploid, and allotetraploid cells can all express characteristic gene markers of ectoderm, mesoderm, or endoderm under appropriate differentiation conditions (Fig. 1*C* and Dataset S3), and other tetraploid fusion lines have recently been differentiated into cortical spheroids or neural crest cells in vitro (14, 15). We caution that some of the lines in our own experiments showed incomplete endoderm differentiation, and previously reported allotetraploid lines showed substantial expression of mesenchymal markers when incubated under conditions that stimulate cortical spheroid formation in diploids (14). In vitro differentiation protocols may thus need to be altered or optimized for tetraploid iPSCs to find conditions suitable for formation of particular cell types of interest.

We have found that tetraploid iPSCs can be grown, repeatedly passaged, tagged with fluorescent markers, and subcloned while maintaining grossly normal karyotypes. Whole-genome sequencing of Xrec1 after CRISPR-Cas9 targeting of chrX shows that induced changes also occur specifically on the targeted chromosome of interest. However, we have also found karyotypic abnormalities in some cell lines when multiple subclones are expanded from a particular cell fusion or treatment (Dataset S1). DNA sequencing further shows that heterogeneity may exist within a colony grown from single cells, such as the staggered breakpoints occurring on the X chromosome in cXdel5 and cXdel6 (*SI Appendix*, Fig. S12). At times, it may be possible to put such heterogeneity to experimental advantage. For example, the staggered deletions occurring within cXdel5 cells may make it possible to establish a larger panel of subclones that could be used for additional fine mapping of *cis* and *trans* factors on the X chromosome, all derived from a single initial round of targeting. However, further study of karyotypic and chromosomal stability in tetraploid iPSC lines is clearly warranted, and we recommend that interested researchers continue to monitor key cell lines and derivatives using periodic karyotyping and whole-genome sequencing approaches.

Beyond mapping the *cis* and *trans* regulators of species-specific gene expression differences, we envision that allotetraploid iPSC lines will also be useful for mapping cellular and tissue differences between humans and chimpanzees. For example, many metabolic differences have evolved alongside major changes in diet between humans and chimpanzees (55). These changes are likely accompanied by cellular changes in enzyme levels and metabolite production that could be scored under appropriate in vitro conditions. In addition, neural progenitors in humans have been shown to have a longer prometaphase and longer metaphase compared to those in chimpanzees (56). These and other cellular traits can be assessed in culture and are compelling

candidates for allotetraploid genetic mapping approaches. Recent advances in organoid technology also make it possible to study organ-level phenotypes that differ between humans and chimpanzees, including differences in organ size, connectivity, and cell type composition (33). Just as meiotic mapping panels have propelled our understanding of evolution in other organisms, further development of mapping methods in human–chimpanzee allotetraploids should provide powerful new genetic approaches for our quest to understand what makes us human.

Materials and Methods

Generation and Maintenance of Tetraploid iPSC Lines. Human and chimpanzee diploid iPSC lines were labeled with diffusible fluorescent dyes and fused on an Eppendorf Multiporator at 4-V AC for 80 s, 16-V DC for 20 μs, and 6-V post-AC for 95 s (*SI Appendix, Supplemental Materials and Methods*). Tetraploid lines were confirmed by propidium iodide staining and karyotyping (Dataset S1). Diploid and tetraploid iPSC lines were routinely propagated feeder-free (*SI Appendix, Supplemental Materials and Methods*).

Trilineage Differentiation. Diploid and tetraploid iPSC lines were differentiated with the STEMdiff Trilineage Differentiation Kit according to the manufacturer's instructions (STEMCELL Technologies, catalog #05230). Differentiation was assessed using qRT-PCR for pluripotency, ectoderm, mesoderm, and endoderm gene markers (Dataset S2 and *SI Appendix, Supplemental Materials and Methods*).

RNAseq Analysis. Sequencing reads were aligned to a composite human–chimpanzee genome (hg38 and pt6), and the number of uniquely mapped reads that overlap each gene was determined using a curated exon annotation (*SI Appendix, Supplemental Materials and Methods*). DE analysis between diploids and autotetraploid iPSCs was performed with DESeq2 (57), and genes with an adjusted $P < 0.05$ and at least a twofold change in expression were called as significant (*SI Appendix, Supplemental Materials and Methods*).

DE between single-species iPSCs, ASE in allotetraploids, and regulatory type classifications were carried out as a combination of previously described methods (35, 44). Genes were classified as *cis*, *trans*, *cis+trans*, *cis–trans*, compensatory, conserved, or ambiguous based on different combinations of significant DE, significant ASE, significant $log_2(FC)$ difference between DE and ASE ("*trans* effects"), and direction of *cis* contribution and *trans* contribution to the DE $log_2(FC)$ (*SI Appendix, Supplemental Materials and Methods*).

SCE Assay. Camptothecin (Sigma Aldrich, catalog #C9911-100MG), ML216 (Cayman Chemical, catalog #15186), and mitomycin C (Sigma Aldrich, catalog #M4287-2MG) were applied to iPSCs with 10 μM BrdU (*SI Appendix, Supplemental Materials and Methods*). The SCE assay was then performed as previously described (25).

Haplotagging. Haplotagging was performed as previously described (28). Reads were aligned to a composite human–chimpanzee genome (hg38 and pt6) and assigned to their molecule of origin by barcode. Variants between hg38 and pt6 were identified for each read and filtered by multiple criteria (*SI Appendix, Supplemental Materials and Methods*). Molecules were scored as recombinant if they contained one interspecific event and approximately five supporting variants per species.

FACS of Fluorescently Marked Allotetraploid Lines. Using two rounds of HR, we inserted an EF1a-EGFP-IRES-PuroR cassette at human chrXq28 and an EF1a-mCherry-IRES-NeoR cassette at chimpanzee chrXq28 in allotetraploid iPSCs (*SI Appendix,* Fig. S7 and *Supplemental Materials and Methods*). Double-marked iPSCs were treated with 25 μM ML216 starting 12 h before nucleofection of CRISPR-Cas9 and gRNA and continuing until 48 h postnucleofection. We then employed multiple sorting strategies to enrich for chrX recombination or deletion events (*SI Appendix*, Fig. S9 and *Supplemental Materials and Methods*).

DNA Sequencing Analysis of chrX Recombinant and Deletion Lines. DNA sequencing reads from the recombinant allotetraploid cell line, two chimpanzee chrX deletion lines, and a control allotetraploid line were aligned to a composite human–chimpanzee (hg38–pt6) reference genome (*SI Appendix, Supplemental Materials and Methods*). The read counts in the recombinant or deletion lines were normalized to read counts in the control line (*SI Appendix*, Figs. S10 and S12).

Data Availability. Data supporting the findings of this study are included in the main text and *SI Appendix* or deposited in publicly available databases. RNAseq data generated in this study are available at Gene Expression Omnibus (GSE184768) (58), and the DNA sequence containing the recombination site for H1C1a-X1-Xrec1 is available at GenBank (OK283040) (59). Additional materials will be made available upon request.

ACKNOWLEDGMENTS. We thank Yoav Gilad for human and chimpanzee iPSC lines, and the Stanford Shared FACS Facility, WiCell Research Institute, Kyle Loh, Alyssa Benjamin, and members of the D.M.K. and Y.F.C. labs for useful discussions. This work was supported, in part, by predoctoral fellowships from the NSF (J.H.T.S., R.L.G., G.A.R.K.), by Stanford Graduate Fellowships (J.H.T.S., R.L.G., G.A.R.K.), by a fellowship from the Center for Computational, Evolutionary and Human Genomics at Stanford (J.H.T.S.), by NIH Grant T32GM007790 (V.C.B.), and by European Research Council Grant "HybridMiX" 639096 (Y.F.C). Y.F.C. is supported by the Max Planck Society, and D.M.K. is an investigator of the Howard Hughes Medical Institute.

1. C. Linnaeus, *Systema Naturae* (Johan Willem Groot, ed. 1, 1735).
2. T. H. Huxley, *Evidence as to Man's Place in Nature* (Williams and Northgate, 1863).
3. A. Varki, T. K. Altheide, Comparing the human and chimpanzee genomes: Searching for needles in a haystack. *Genome Res.* **15**, 1746–1758 (2005).
4. S. L. Robson, B. Wood, Hominin life history: Reconstruction and evolution. *J. Anat.* **212**, 394–425 (2008).
5. A. Levchenko, A. Kanapin, A. Samsonova, R. R. Gainetdinov, Human accelerated regions and other human-specific sequence variations in the context of evolution and their relevance for brain development. *Genome Biol. Evol.* **10**, 166–188 (2018).
6. H. A. Orr, The genetic theory of adaptation: A brief history. *Nat. Rev. Genet.* **6**, 119–127 (2005).
7. F. H. Ruddle, R. S. Kucherlapati, Hybrid cells and human genes. *Sci. Am.* **231**, 36–44 (1974).
8. D. R. Cox, M. Burmeister, E. R. Price, S. Kim, R. M. Myers, Radiation hybrid mapping: A somatic cell genetic method for constructing high-resolution maps of mammalian chromosomes. *Science* **250**, 245–250 (1990).
9. G. Guo, W. Wang, A. Bradley, Mismatch repair genes identified using genetic screens in Blm-deficient embryonic stem cells. *Nature* **429**, 891–895 (2004).
10. K. Yusa *et al.*, Genome-wide phenotype analysis in ES cells by regulated disruption of Bloom's syndrome gene. *Nature* **429**, 896–899 (2004).
11. S. Lazzarano *et al.*, Genetic mapping of species differences via in vitro crosses in mouse embryonic stem cells. *Proc. Natl. Acad. Sci. U.S.A.* **115**, 3680–3685 (2018).
12. M. J. Sadhu, J. S. Bloom, L. Day, L. Kruglyak, CRISPR-directed mitotic recombination enables genetic mapping without crosses. *Science* **352**, 1113–1116 (2016).
13. I. Gallego Romero *et al.*, A panel of induced pluripotent stem cells from chimpanzees: A resource for comparative functional genomics. *eLife* **4**, e07103 (2015).
14. R. M. Agoglia *et al.*, Primate cell fusion disentangles gene regulatory divergence in neurodevelopment. *Nature* **592**, 421–427 (2021).
15. D. Gokhman *et al.*, Human-chimpanzee fused cells reveal *cis*-regulatory divergence underlying skeletal evolution. *Nat. Genet.* **53**, 467–476 (2021).
16. D. Baker *et al.*, Detecting genetic mosaicism in cultures of human pluripotent stem cells. *Stem Cell Rep.* **7**, 998–1012 (2016).
17. S. A. Signor, S. V. Nuzhdin, The evolution of gene expression in *cis* and *trans*. *Trends Genet.* **34**, 532–544 (2018).
18. C. V. Weiss *et al.*, The *cis*-regulatory effects of modern human-specific variants. *eLife* **10**, e63713 (2021).
19. J. Zhang *et al.*, Regulation of cell proliferation of human induced pluripotent stem cell-derived mesenchymal stem cells via ether-à-go-go 1 (hEAG1) potassium channel. *Am. J. Physiol. Cell Physiol.* **303**, C115–C125 (2012).
20. D. Gokhman *et al.*, Gene ORGANizer: Linking genes to the organs they affect. *Nucleic Acids Res.* **45**, W138–W145 (2017).
21. E. Brunner *et al.*, CRISPR-induced double-strand breaks trigger recombination between homologous chromosome arms. *Life Sci. Alliance* **2**, e201800267 (2019).
22. Y. Yoshimura, A. Yamanishi, T. Kamitani, J. S. Kim, J. Takeda, Generation of targeted homozygosity in the genome of human induced pluripotent stem cells. *PLoS One* **14**, e0225740 (2019).
23. O. Momcilovic *et al.*, DNA damage responses in human induced pluripotent stem cells and embryonic stem cells. *PLoS One* **5**, e13410 (2010).
24. G. H. Nguyen *et al.*, A small molecule inhibitor of the BLM helicase modulates chromosome stability in human cells. *Chem. Biol.* **20**, 55–62 (2013).
25. D. M. Stults, M. W. Killen, A. J. Pierce, The sister chromatid exchange (SCE) assay. *Methods Mol. Biol.* **1105**, 439–455 (2014).
26. D. D. Hurst, S. Fogel, Mitotic recombination and heteroallelic repair in *Saccharomyces cerevisiae*. *Genetics* **50**, 435–458 (1964).
27. A. Dréau, V. Venu, E. Avdievich, L. Gaspar, F. C. Jones, Genome-wide recombination map construction from single individuals using linked-read sequencing. *Nat. Commun.* **10**, 4309 (2019).
28. J. I. Meier *et al.*, Haplotype tagging reveals parallel formation of hybrid races in two butterfly species. *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2015005118 (2021).
29. G. Luo *et al.*, Cancer predisposition caused by elevated mitotic recombination in Bloom mice. *Nat. Genet.* **26**, 424–429 (2000).
30. A. Kolb-Kokocinski *et al.*, The systematic functional characterisation of Xq28 genes prioritises candidate disease genes. *BMC Genomics* **7**, 29 (2006).
31. M. McVey, S. E. Lee, MMEJ repair of double-strand breaks (director's cut): Deleted sequences and alternative endings. *Trends Genet.* **24**, 529–538 (2008).
32. A. M. M. Sousa *et al.*, Molecular and cellular reorganization of neural circuits in the human lineage. *Science* **358**, 1027–1032 (2017).
33. S. Kanton *et al.*, Organoid single-cell genomic atlas uncovers human-specific features of brain development. *Nature* **574**, 418–422 (2019).

34. S. B. Carroll, Evo-devo and an expanding evolutionary synthesis: A genetic theory of morphological evolution. *Cell* **134**, 25–36 (2008).
35. X. Shi *et al.*, *Cis*- and *trans*-regulatory divergence between progenitor species determines gene-expression novelty in *Arabidopsis* allopolyploids. *Nat. Commun.* **3**, 950 (2012).
36. G. A. Wray, The evolutionary significance of *cis*-regulatory mutations. *Nat. Rev. Genet.* **8**, 206–216 (2007).
37. M. H. Sherman, C. H. Bassing, M. A. Teitell, Regulation of cell differentiation by the DNA damage response. *Trends Cell Biol.* **21**, 312–319 (2011).
38. H. Yang *et al.*, Methods favoring homology-directed repair choice in response to CRISPR/Cas9 induced-double strand breaks. *Int. J. Mol. Sci.* **21**, 6461 (2020).
39. J. Pinder, J. Salsman, G. Dellaire, Nuclear domain 'knock-in' screen for the evaluation and identification of small molecule enhancers of CRISPR-based genome editing. *Nucleic Acids Res.* **43**, 9379–9392 (2015).
40. T. S. Nambiar *et al.*, Stimulation of CRISPR-mediated homology-directed repair by an engineered RAD18 variant. *Nat. Commun.* **10**, 3395 (2019).
41. R. Sarno *et al.*, Programming sites of meiotic crossovers using Spo11 fusion proteins. *Nucleic Acids Res.* **45**, e164 (2017).
42. R. A. Bergstrom, Y. You, L. C. Erway, M. F. Lyon, J. C. Schimenti, Deletion mapping of the head tilt (*het*) gene in mice: A vestibular mutation causing specific absence of otoliths. *Genetics* **150**, 815–822 (1998).
43. R. K. Cook *et al.*, The generation of chromosomal deletions to provide extensive coverage and subdivision of the *Drosophila melanogaster* genome. *Genome Biol.* **13**, R21 (2012).
44. C. J. McManus *et al.*, Regulatory divergence in *Drosophila* revealed by mRNA-seq. *Genome Res.* **20**, 816–825 (2010).
45. R. Tillotson, A. Bird, The molecular basis of MeCP2 function in the brain. *J. Mol. Biol.* **43**, 1602–1623 (2019).
46. S. D. Konduri *et al.*, Promoter methylation and silencing of the tissue factor pathway inhibitor-2 (*TFPI-2*), a gene encoding an inhibitor of matrix metalloproteinases in human glioma cells. *Oncogene* **22**, 4509–4516 (2003).
47. E. Landucci *et al.*, iPSC-derived neurons profiling reveals GABAergic circuit disruption and acetylated α-tubulin defect which improves after iHDAC6 treatment in Rett syndrome. Exp. Cell Res. **368**, 225–235 (2018).
48. D. T. Lioy *et al.*, A role for glia in the progression of Rett's syndrome. *Nature* **475**, 497–500 (2011).
49. W. S. Chung *et al.*, Astrocytes mediate synapse elimination through MEGF10 and MERTK pathways. *Nature* **504**, 394–400 (2013).
50. N. A. Oberheim *et al.*, Uniquely hominid features of adult human astrocytes. *J. Neurosci.* **29**, 3276–3287 (2009).
51. Y. Zhang *et al.*, Purification and characterization of progenitor and mature human astrocytes reveals transcriptional and functional differences with mouse. *Neuron* **89**, 37–53 (2016).
52. H. Kilpinen *et al.*, Common genetic variation drives molecular heterogeneity in human iPSCs. *Nature* **546**, 370–375 (2017).
53. M. S. Golbus, R. Bachman, S. Wiltse, B. D. Hall, Tetraploidy in a liveborn infant. *J. Med. Genet.* **13**, 329–332 (1976).
54. M. Guc-Scekic, J. Milasin, M. Stevanovic, L. J. Stojanov, M. Djordjevic, Tetraploidy in a 26-month-old girl (cytogenetic and molecular studies). *Clin. Genet.* **61**, 62–65 (2002).
55. R. Blekhman *et al.*, Comparative metabolomics in primates reveals the effects of diet and gene regulatory variation on metabolic divergence. *Sci. Rep.* **4**, 5809 (2014).
56. F. Mora-Bermúdez *et al.*, Differences and similarities between human and chimpanzee neural progenitors during cerebral cortex development. *eLife* **5**, e18683 (2016).
57. M. I. Love, W. Huber, S. Anders, Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
58. J. H. Song *et al.*, Genetic studies of human-chimpanzee divergence using stem cell fusions. Gene Expression Omnibus. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE184768. Deposited 24 September 2021.
59. J. H. T. Song *et al.*, Homo sapiens x Pan troglodytes tetraploid cell line cell line H1C1a-X1-Xrec1 genomic sequence. GenBank. https://www.ncbi.nlm.nih.gov/nuccore/OK283040. Deposited 23 September 2021.

PNAS
www.pnas.org

Supplementary Information for

Genetic studies of human-chimpanzee divergence using stem cell fusions

J.H.T. Song, R.L. Grant, V.C. Behrens, M. Kucka, G.A. Roberts Kingman, V. Soltys, Y.F. Chan, D.M. Kingsley

David M. Kingsley
E-mail: kingsley@stanford.edu

This PDF file includes:

Supplementary text
Figs. S1 to S13
Legends for Dataset S1 to S8
SI References

Other supplementary materials for this manuscript include the following:

Datasets S1 to S8

Supporting Information Text

Supplemental Materials and Methods

Cell culture maintenance. The induced pluripotent stem cell lines H23555 (H1), H20961 (H2), C3649 (C1), and C8861 (C2) were provided by the Gilad laboratory (1). Cultures were tested for and maintained mycoplasma free. Diploid and tetraploid lines were routinely propagated feeder-free in mTeSR1 or mTeSR Plus media (STEMCELL Technologies, cat #85850 and cat #100-0276) on cell culture plastics coated with Geltrex basement membrane matrix (Gibco, cat #A1413302). When confluent, cells were passaged using Accutase (Millipore, cat #SCR005) with 1μM thiazovivin (Tocris, cat #3845) or using 0.5mM EDTA as previously described (2, 3). Cells were imaged on the EVOS FL microscope (Thermo Fisher) at 4X magnification, unless otherwise noted.

Generation of tetraploid iPSC lines. One diploid iPSC line was labeled with CellTracker Green CMFDA Dye (1:667) (Thermo Fisher, cat #C7025) and the other diploid iPSC line was labeled with CellTracker Blue CMAC dye (1:500) (Thermo Fisher, cat #C2110) or CellTracker Red CMTPX Dye (1:1000) (Thermo Fisher, cat #C34552) per the manufacturer's instructions. Cells were washed multiple times to remove excess dye and allowed to recover after labeling in mTeSR1 + 1μM thiazovivin for at least 1 hour prior to fusion. $7x10^5$ cells from each line were combined, washed twice in 1ml fusion buffer, resuspended in 350μl fusion buffer, and fused in the helix fusion chamber of an Eppendorf Multiporator at 4V AC for 80s, 16V DC for 20μs, 6V post-AC for 95s. After 10 minutes at room temperature, 1ml of mTeSR1 + 1μM thiazovivin was added to the helix fusion chamber. Fusion buffer was hypoosmolar electrofusion buffer (Eppendorf, cat #940002150) diluted in water (normally, 60% hypoosmolar electrofusion buffer and 40% water).

Tetraploid clones were then selected in one of two ways. In the first method, 250-350μl of the resulting suspension after fusion was immediately plated in a 10-cm plate. Double-labeled cells were screened the following day under a fluorescent microscope and marked on the plate. The diffusible dyes were only visible for 2 days after fusion. Surrounding diploid cells were removed on each subsequent day by manual scraping. When the originally identified double-labeled cells grew into colonies, they were picked into a 96-well plate and screened as described below.

In the second method, 7-15 fusions were performed on the same day using the helix fusion chamber as described above and collected in a large volume of mTeSR1 + 1μM thiazovivin for fluorescence-activated cell sorting (FACS). For increased viability in the second method, we used only CellTracker Green CMFDA and CellTracker Blue CMAC dyes. Cells were gently pipetted every hour to avoid CellTracker dye diffusion and undesired labeling of nearby non-fused cells. Prior to FACS, cells were resuspended in 500μl FACS buffer (0.5% BSA fraction V, 5mM EDTA, 1% Penicillin/Streptavidin, and 1μM thiazovivin in 1X PBS) and strained through a 30μm filter. Double-labeled cells were collected into 1 well of a 96-well plate by FACS. After 3 days, cells were collected and seeded at 1-5$x10^3$ cells in 10-cm plates. Individual colonies were then picked into 96-well plates.

To identify tetraploid colonies, we performed propidium iodide staining (Invitrogen, cat #P3566) on fixed cells to examine ploidy via FACS analysis. Briefly, cells were fixed in 80% EtOH overnight at 4°C. The next day, cells were washed twice in 1X PBS and stained at 37°C for 10 minutes in 20μg/ml RNase A, 40μg/ml of propidium iodide, and 0.1% Triton X-100 in 1X PBS. Cells with both 4N and 8N DNA content by FACS, suggesting that they contain tetraploid cells, were expanded. Expanded colonies were further screened for DNA content by karyotyping as described previously (4). Colonies that contained only tetraploid cells were maintained as stocks. G-banded karyotyping was also performed by WiCell (Table S1).

Trilineage differentiation. Diploid and tetraploid iPSC lines were seeded in 12-well plates and differentiated with the STEMdiff™ Trilineage Differentiation Kit according to the manufacturer's instructions (STEMCELL Technologies, cat #05230). Additionally, untreated cells were collected two days after seeding. Three replicate wells of cells per cell line were collected per condition. Differentiation was assessed using reverse transcription quantitative PCR (RT-qPCR) for pluripotency, ectoderm, mesoderm, and endoderm gene markers (Quantitative PCR SI Methods).

Quantitative PCR (qPCR). RT-qPCR was used to assess differentiation potential for trilineage differentiation samples, and qPCR was used to study DNA marker dosage in chrX targeted cell lines. All reactions were performed using Brilliant II SYBR Green Low ROX qPCR Master Mix (Agilent, cat #600830) on a QuantStudio 5 Real-Time PCR System (Thermo Fisher).

For trilineage differentiation, three replicate wells of cells per condition were collected in Trizol and applied to the Direct-Zol RNA Miniprep kit (Zymo Research, cat #R2051) for RNA extraction per the manufacturer's instructions. cDNA was synthesized from RNA with the SuperScript™ VILO™ cDNA Synthesis Kit (Thermo Fisher, cat #11754250). To assess differentiation potential of trilineage differentiation samples, qPCR was performed in triplicate on all samples with two marker genes each for pluripotency (*NANOG*, *DNMT3B*), ectoderm (*PAX6*, *RAX*), mesoderm (*TBXT*, *HAND1*), endoderm (*FOXA2*, *SOX17*), and housekeeping (*GAPDH*, *YWHAZ*) (5–11). All primer pairs span a large intron, have efficiencies between 90-110%, bind identical sequences in humans and chimpanzees, produce PCR products of identical length in both species, and were chosen from the literature or designed in-house (Table S2). For each sample, the quantity of each marker gene was calculated by comparing to a standard curve of pooled samples. This quantity was normalized by dividing by the geometric mean of the quantities of the two housekeeping genes (*GAPDH*, *YWHAZ*) in the same sample and then divided by the normalized quantity of the marker gene in undifferentiated iPSCs from the H2 human diploid line. Two-tailed Student's t-tests were used to determine statistically significant differences in marker gene expression between differentiated and undifferentiated iPSCs at 5% Benjamini-Hochberg FDR.

For determination of chrXq dosage relative to other chromosomes, cells were harvested from 96-well plates using Accutase (Millipore, cat #SCR005), and DNA was extracted using the DNeasy 96 Blood & Tissue Kit (Qiagen). Reactions were performed either in duplicate or in triplicate with primers for chromosomes 6p, Xp and Xq (Table S2).

Library preparation for RNA sequencing. Samples were flash frozen and stored at -80°C as a pellet. RNA extraction, library preparation, and sequencing for the chrX deletion samples were performed by Genewiz. All other RNA sequencing samples were prepared in-house before sequencing on the Illumina HiSeq 4000 with Novogene. Briefly, samples were resuspended in Trizol and directly applied to the Direct-Zol RNA Miniprep kit (Zymo Research, cat #R2051) for RNA extraction per the manufacturer's instructions. Technical replicates for each line were collected from thaws of different frozen vials. Only samples with $RIN > 9$ were used for RNA sequencing.

1μg of RNA was used for library preparation. RNA sequencing libraries were prepared with the TruSeq Stranded mRNA Library Prep (Illumina, cat #20020595) using the IDT for Illumina – TruSeq RNA UD Indexes (96 Indexes, 96 Samples) (Illumina, cat #20022371) according to the manufacturer's instructions with one modification. Prior to PCR amplification, 10% of a subset of samples were run under the recommended PCR conditions with SybrGreen on the QuantStudio 5 Real-Time PCR System. We identified the number of PCR cycles required to reach the crossing point by qPCR and used that number of cycles for PCR amplification on the entire set of samples. We ran 8 PCR cycles. Libraries were pooled and sequenced to around 10 million reads per sample for the diploid and auto-tetraploid lines and around 20 million reads per sample for the allo-tetraploid lines. Five independently-derived C1C1 auto-tetraploid lines, two independently-derived H1H1 auto-tetraploid lines (two technical replicates each), twelve independently-derived H1C1 allo-tetraploid lines, and ten independently-derived H2C2 allo-tetraploid lines were sequenced. Three technical replicates were sequenced for each diploid line.

Alignment of RNA sequencing to composite human-chimpanzee genome. Sequencing reads were trimmed for adapter sequences using cutadapt v1.8.1 (12), and read quality was confirmed using fastqc v0.11.9 (13). Reads were aligned using STAR v2.7.1a with two-pass mapping (14). Samples were mapped to a composite human-chimpanzee genome (hg38 and pt6). The number of uniquely-mapped reads ($MAPQ = 255$) that overlap each gene was counted using featureCounts from the subread v1.6.0 package (15).

To generate the gene annotations used in featureCounts, GRCh38.94 human exon annotations from Ensembl (16) were mapped from hg38 to pt6 using pslMap (17). After removing mappings where the number of bases that map is less than half of the query exon size, we retained only exons that uniquely mapped from humans to chimpanzees. We then removed genes for which exons map to opposite DNA strands, different scaffolded chromosomes, or where consecutive exons map more than 800kb apart. We further filtered out exons where more than 10% of reads from diploid or auto-tetraploid lines map to the incorrect species when mapped to the composite genome. This resulted in 48,735 annotated genes that contain at least 1 exon (byexon-gene). We also used a second set of annotations. We identified SNPs that differed between the human and chimpanzee cell lines using the GATK RNA variant pipeline (18, 19) and assigned SNPs to genes annotated in humans. We also filtered out SNPs where more than 10% of reads from diploid or auto-tetraploid lines map to the incorrect species when mapped to the composite genome. This resulted in 14,333 annotated genes with at least 1 SNP (bysnp-gene). Read counts for the byexon-gene annotation were also adjusted for feature length to account for differences between feature length in the human and chimpanzee genomes. Results were very similar across both annotations, and results are reported for the byexon-gene annotation in the current study.

Gene expression analysis of diploids and auto-tetraploid iPSC lines. After sequencing reads were aligned to the composite human-chimpanzee genome as described above, differential gene expression analysis was performed with DESeq2 (20) using default parameters. We called genes as significant if they had an adjusted $p < 0.05$ after Benjamini-Hochberg FDR correction and at least a 2-fold change in expression.

Differential gene expression, allele-specific gene expression, and *cis*/*trans* analysis between humans and chimpanzees in diploid, auto-tetraploid, and allo-tetraploid iPSC lines. Differential expression (DE) between single-species iPSCs, allele-specific expression (ASE) in allo-tetraploids, and regulatory type classifications were carried out as a combination of previously described methods (21, 22).

After RNA sequencing reads were aligned to the composite human-chimpanzee genome as described above, reads mapping to genes on human chromosome 18 (and the orthologous chimpanzee genes) were removed. Next, each sample was downsampled to 9,711,244 reads (for DE) or 11,490,119 reads (for ASE), and genes with fewer than 10 reads assigned to both the human and the chimpanzee orthologs were excluded. For each gene, $log_2(FC)$ was calculated between each human-only and each chimpanzee-only sample (for DE) or between the human allele and the chimp allele in each allo-tetraploid cell line (for ASE). Genes with significantly different $log_2(FC)$ between human and chimpanzee were determined to be DE or ASE. Each gene was tested for significant "*trans*-effects" by testing for a significant $log_2(FC)$ difference between single-species iPSCs and allo-tetraploid iPSCs. Significance for all $log_2(FC)$ differences was determined by Welch's t-test at 5% Benjamini-Hochberg FDR . Importantly, only half of the allo-tetraploid samples were used to determine whether a gene is significantly ASE, and the other half were used to determine significant "*trans*-effects" since this has been reported to reduce false classification as *compensatory* (23).

Finally, the *cis*-contribution (C) and *trans*-contribution (T) to the observed DE $log_2(FC)$ (D) was calculated for each gene. Specifically, the *cis*-contribution (C) was equal to the ASE $log_2(FC)$, and the *trans*-contribution was calculated as $T = D - C$.

Genes were classified by regulatory type based on the following criteria:

cis: significant DE, significant ASE, no significant "*trans*-effects," *cis*-contribution and *trans*-contribution to DE $log_2(FC)$ in the same direction

trans: significant DE, not significant ASE, significant "*trans*-effects"

cis+trans: significant DE, significant ASE, significant "*trans*-effects," *cis*-contribution and *trans*-contribution to DE $log_2(FC)$ in the same direction

cis-trans: significant DE, significant ASE, significant "*trans*-effects," *cis*-contribution and *trans*-contribution to DE $log_2(FC)$ in opposite directions

compensatory: not significant DE, significant ASE, significant "*trans*-effects"

conserved: not significant DE, not significant ASE, no significant "*trans*-effects"

ambiguous: all other patterns

All results reported in this paper used the "by-exon-gene" annotation as described in the "Alignment of RNA sequencing to composite human-chimpanzee genome" section above (except for *TSPAN6*, which was not included in the "byexon-gene" annotation and was assessed using the "bysnp-gene" annotation).

Gene ontology enrichments. Significant gene ontology enrichments (adjusted $p < 0.05$ after Benjamini-Hochberg FDR correction) were determined using the R package clusterProfiler's enrichGO function (24) for the annotation data sets "Biological Process," "Molecular Function," and "Cellular Component." The set of analyzed genes was used as the background reference list.

Gene expression analysis of X chromosome deletion lines. RNA sequencing reads were aligned to the composite human-chimpanzee genome as described above. To identify the approximate location of X chromosome deletions, we computed the ratio of human read counts to chimpanzee read counts for each deletion line normalized to control (non-deletion) lines. A count of 1 was added to any sample with allele counts of zero, and ratios were calculated for genes with more than 10 counts on average and where at least half of the samples had at least 5 counts. Approximate deletion breakpoints were then determined by visual inspection.

To identify autosomal genes whose expression may be affected by *trans*-regulators on the X chromosome, we carried out differential gene expression analysis of control and human and chimpanzee chrX targeted deletion lines using DESeq2 (20) with the Wald test at 5% Benjamini-Hochberg FDR. *Trans*-regulated candidates were identified by the following five criteria: (1) Genes on autosomes that showed significant expression changes when comparing the four lines with deletion breakpoints of the chimpanzee chrX around 148Mb (cXdel4-cXdel7) to the nine control lines that lack deletions; (2) Genes on autosomes that did not show significant expression changes when comparing the five lines with deletion breakpoints of the human chrX around 95Mb (hXdel3-hXdel7) to the nine control lines that lack deletions; (3) Genes that met the first two criteria whose expression level was also significantly different in comparisons between the chimpanzee (cXdel4-cXdel7) and human (hXdel3-hXdel7) terminal deletions; (4) Genes that also showed the same direction of change in cell lines carrying shorter (cXdel4-cXdel7) and larger chimpanzee chrX deletions (cXdel1-cXdel3) compared to control lines; and (5) Genes where the hXdel8 line which has a human deletion breakpoint near the distal chimpanzee chrX deletion lines maintained expression within the range of the control lines.

Sister chromatid exchange (SCE) assay. Cells were passaged the day before testing. For camptothecin (Sigma Aldrich, cat #C9911-100MG), camptothecin and 10μM BrdU were applied to cells for 1 hour before being replaced with fresh media containing 10μM BrdU overnight. For ML216 (Cayman Chemical, cat #15186) and mitomycin C (Sigma Aldrich, cat #M4287-2MG), cells were incubated with the small molecule and 10μM BrdU for 24-48 hours with a media change every 24 hours. Cells were then moved to fresh media containing 10μM BrdU and 0.1μg/ml colcemid for 4 hours and subsequently collected for sister chromatid exchange (SCE) assay as previously described (25). Cells were alternatively first collected into a 1.5ml tube before adding new media containing 10μM BrdU and 0.1μg/ml colcemid, with no obvious change in results. Multiple metaphase spreads were imaged at 100X, and recombination events were counted using the ImageJ Cell Counter function. *P*-values were calculated using the 1-tailed Student's t-test.

Haplotagging. Haplotagging was performed as previously described (26). Briefly, genomic DNA from each sample was mixed with individually barcoded magnetic beads containing bead-immobilized active Tn5 transposase for tagmentation with up to 21 million barcode diversity. Tagged DNA was then PCR amplified, size selected, and sequenced on a NovaSeq 6000 instrument (Illumina).

Reads were aligned to a composite human-chimpanzee genome (hg38 and pt6) using EMA, a barcode-first variant of the bwa aligner (27). For the analysis, we focused on regions that reciprocally and uniquely mapped between the two species assemblies, with the mapping based on the hg38 to pt6 chain files from the UCSC Genome Browser (28) and pslMap (17). 500bp orthologous regions with greater than 2-fold difference in read coverage were excluded from further analysis. Each read

was also assigned to a molecule based on its barcode (retained as the BX beadTag). For each read, we identified variants between hg38 and pt6 (SNPs and indels). The variant annotation file was generated by first parsing the maf file between hg38 and pt6 from the UCSC Genome Browser (28). We also included variants identified by running the GATK variant pipeline (18, 19) on reads that map uniquely to either hg38 or pt6 and where all reads assigned to a given barcode map to only one species. If no variants in our resulting annotation file were identified in a read but the read uniquely mapped to either hg38 or pt6, the read itself was considered as a variant.

Along each molecule, we coded the species assignment (e.g *H-H-H-H-H-C-C-C* where *H* is a human variant and *C* is a chimpanzee variant). We then applied the following strict filters to identify a high-confidence set of recombinant molecules: (1) Identified SNPs must have a phred quality score of at least 30; (2) Given the low rate of mitotic recombination, multiple "switch" events (e.g. *H-H-H-C-H-H-H*) are likely artifacts and such variants were removed; (3) We also excluded possible mapping artifacts where particular variants were found at the boundary of multiple recombination events; (4) Variants contained in 500bp regions with greater than 2-fold difference in the directionality of switch events (e.g. switch events were predominantly $H \rightarrow C$ instead of $C \rightarrow H$) were removed; (5) We included only paired recombinant molecules that could be "reciprocal events" to further account for biases in the directionality of switch events; (6) We excluded any variants that are in ENCODE blacklist regions (29); (7) All recombinant molecules must contain only 1 switch event and > 5 supporting variants per species.

To calculate the genome-wide recombination rate, we divided the number of recombination events by the approximate number of analyzed human-chimpanzee tetraploid genomes (molecular coverage of molecules that passed the above filters). To compare the inter-specific recombination rate from haplotagging to previously reported recombination rates in the literature, we examined previous reports that selected for recombination events near single-locus, drug-selectable markers following ML216 treatment in mouse embryonic stem cells (30–32). To extrapolate single-locus marker rates to genome-wide estimates, we calculated the genomic distance between the centromere and the drug-selectable marker, and estimated the genome-wide recombination rate as (size of diploid genome / genomic distance studied) * reported recombination rate.

To assess the effect of CRISPR targeting to specific loci, we examined the recombination rate in the 250kb interval surrounding the target loci with and without filters, with no difference in the relative enrichment at the target loci. Data visualizations were generated with ggplot2 (33) and karyoploteR (34). In Fig. S6, samples were plotted with their experimental batch due to differences in read and molecular coverage.

Generation of fluorescently-tagged allo-tetraploid lines. We cloned two plasmids, one with homology arms (chrX:153,850,316-153,851,493, hg38) flanking a EF1a-EGFP-IRES-PuroR cassette to target human chrX and the second with homology arms (chrX:149,205,726-149,208,867, pt6) flanking a EF1a-mCherry-IRES-NeoR cassette to target chimpanzee chrX (Fig. S7), into the pMAXGFP plasmid backbone (Lonza). Guide RNAs (gRNAs) were designed to linearize the plasmid containing the insertion cassette and cut the target insertion site (HR_X_gRNA_1 and HR_X_gRNA_2 in Table S2). gRNAs were then *in vitro* transcribed as described above.

2.5μl of 40μM Cas9-NLS purified protein (QB3, UC Berkeley) was mixed with 2.5μg each of both gRNAs for 10 minutes at room temperature. This complex and 1.875μg of the plasmid targeting human chrX were nucleofected into $3x10^6$ cells using the Nucleofection Stem Cell Kit 2 (Lonza, cat #VPH-5022) and program A-33 on the Nucleofector 2b Device (Lonza). Immediately after nucleofection, 1ml of pre-warmed media (mTeSR1 + 1μM thiazovivin) was added to the reaction. The reaction was allowed to recover for 20 minutes at room temperature, and 5 separate reactions were pooled and plated on one 10-cm plate. We also nucleofected pMAXGFP separately as a positive control for nucleofection efficiency.

After cells recovered and expanded (∼5 days post-nucleofection), cells with insertion events were selected by multiple days of puromycin treatment. We examined selection efficacy via fluorescence under an EVOS FL microscope. After multi-day selection, we picked colonies into 96-well plates. When colonies reached confluency, they were split and screened for proper insertion events by PCR, using primer pairs where one primer targets nearby genomic DNA and a second primer targets the insertion construct. We verified target-site insertion events using primer sets at both the 5' and 3' ends and separately with species-specific primers (Table S2). We confirmed the insertion sequence via PCR followed by Sanger sequencing with primers chrX-F2 and chrX-R2 (Table S2). To confirm that the insertion was inserted into the target locus and nowhere else in the genome, we expanded promising colonies for Southern blot analysis. Colonies verified by both PCR and Southern blot were then subject to a second round of nucleofection to insert the mCherry cassette into chimpanzee chrX. Double-marked colonies were selected for using Geneticin (Thermo Fisher, cat #10131035) and puromycin (Sigma-Aldrich cat #P8833). Double-marked lines were confirmed by PCR, Southern blot, and visual inspection of fluorescence.

CRISPR/Cas9 treatment of iPSC lines. Guide RNAs were designed (Table S2) and *in vitro* transcribed (IVT) as previously described (35). Briefly, CRISPR IVT target oligos containing the gRNA and the CRISPR IVT scaffold oligo (HPLC-purified) were synthesized by Integrated DNA Technologies. 40 cycles of PCR were performed between the CRISPR IVT scaffold oligo using Phusion DNA polymerase (Thermo Scientific, cat #F530L) and the CRISPR IVT target oligo, and the PCR product was purified using the QIAquick PCR Purification Kit (Qiagen, cat #28104). The PCR product was then *in vitro* transcribed using the MEGAscript T7 transcription kit (Thermo Fisher, cat #AM1334) for 16 hours at 37°C. The reaction was treated with DNase, and transcribed gRNA was extracted with phenol/chloroform and precipitated with isopropanol. Transcribed gRNA was resuspended to approximately 2μg/ul.

To select the highest efficiency guides, we tested performance in 96-well plate format. For each guide, 1μl of 40μM Cas9-NLS purified protein (QB3, UC Berkeley) was complexed with 2μg of gRNA for 10 minutes at room temperature. We performed nucleofection of $2x10^5$ cells per reaction using the P3 Primary Cell 96-well Nucleofector Kit (Lonza, cat #V4SP-3096) on

the Amaxa 96-well Shuttle Device (Lonza) with program CA-137. Two days post-nucleofection, cells were collected for DNA extraction using phenol/chloroform. We used primers bracketing the target cut site (Table S2) and Sanger sequenced the products for analysis using TIDE (36) to determine guide efficiency. For a subset of guides, we further confirmed cutting events by cloning the gel-extracted PCR product into the TOPO TA vector (Life Technologies, cat #450641) and performing colony PCR followed by Sanger sequencing to identify lesions at the target cut site.

For targeted recombination, we nucleofected cells with CRISPR/Cas9 and gRNA using the same nucleofection conditions as described above in the "Generation of fluorescently-tagged allo-tetraploid lines" section. For CRISPR+ML216 conditions, we treated cells with 25μM ML216 starting 12 hours before nucleofection, as previously described (37). After recovering for 1 hour, nucleofected cells were plated directly into media with ML216, and ML216 media was replaced again after 24 hours. At 48 hours post-nucleofection, cells were collected for FACS or haplotagging experiments.

Fluorescence activated cell sorting (FACS). Allo-tetraploid cells with fluorescently-marked chrX were subjected to CRISPR+ ML216 treatment as described above.

For cells treated with chimpanzee-specific gRNA, we selected for loss of mCherry (from the chimpanzee chrX insertion) and possible duplication of GFP (from the human chrX insertion) by sorting for mCherry-negative, high-intensity-GFP cells. To eliminate cells with high GFP due to duplicated DNA content at the G2/M phase of the cell cycle, we stained cells with Hoechst 33342 (Thermo Fisher, cat #62249) for 30 minutes at 37°C to sort only from cells in the G1 cell cycle phase.

For cells treated with human-specific gRNA, we selected for loss of GFP (from the human chrX insertion) and possible duplication of mCherry (from the chimpanzee chrX insertion) by sorting for GFP-negative, high-intensity-mCherry cells. As a second marker of two copies of chimpanzee chrX downstream of the gRNA cutsite, we chose a cell-surface protein, TSPAN6, that has *cis*-regulatory changes with 1.4-fold higher expression in chimpanzee relative to human (adjusted $p = 1.4x10^{-4}$ by Welch's t-test after Benjamini-Hochberg FDR correction). Because this human-chimpanzee gene expression difference can also be observed at the level of protein expression by antibody staining, sorting for high TSPAN6 protein acted as a second marker to potentially sort for two copies of chimpanzee chrX downstream of the gRNA cutsite (Fig. S9).

Treated cells were stained with either Hoechst 33342 at 10μg/mL for 30 minutes at 37°C, or with TSPAN6 primary antibody (1:10; LS Bio, cat #LS-C160272-400) for 1 hour at 4°C followed by 30 minutes at 4°C with a goat anti-rabbit secondary antibody conjugated with APC fluorophore (1:500; Thermo Fisher, cat #A-10931). Cells were sorted single-cell into 96-well plates on a BD Influx cell sorter at the Stanford Shared FACS Facility. Representative sorting gating schemes are shown in Fig. S9.

DNA sequencing analysis of chrX recombinant and deletion lines. DNA from the recombinant allo-tetraploid cell line (H1C1a-X1-Xrec1), two chimpanzee chrX deletion lines (H1C1a-X1-cXdel5 and H1C1a-X1-cXdel6), and a control allo-tetraploid line (H1C1a-X1-S) (Table S1) were extracted and sequenced to 30X coverage with 150bp paired-end reads by GeneWiz. Illumina adapters were removed using Picard Tools (http://broadinstitute.github.io/picard/), and reads were then aligned to a composite human-chimpanzee (hg38-pt6) reference genome using BWA-MEM with the -M flag (38). Duplicate reads were marked with Picard Tools and removed using samtools (39). We filtered out reads with $MAPQ < 30$ and reads that did not lift over between hg38 and pt6 using pslMap (17). For the recombinant line (Fig. S10) or the deletion lines (Fig. S12), we normalized observed read counts to the read counts in the control H1C1a-X1-S over 10kb sliding windows to account for any sequencing or mapping bias and visualized this ratio along human chrX coordinates. For the recombinant line, inspection of reads at the likely recombination site revealed the exact junction site as a 4bp microhomology (CACC) found at both human chrX:140133478-140133481 (hg38) and chimpanzee chrX:124020937-124020940 (pt6).

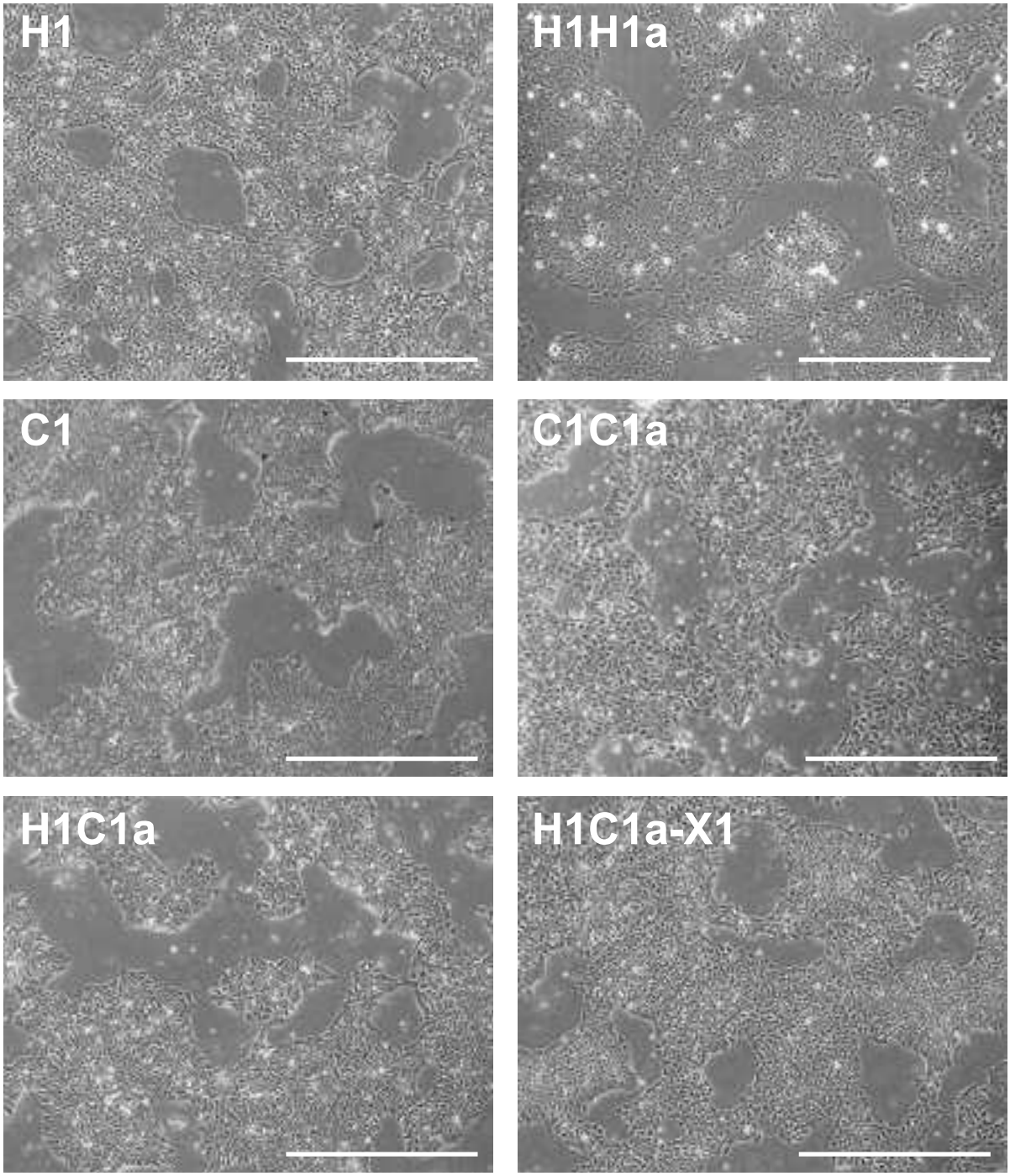


Fig. S1. Morphologies of auto- and allo-tetraploid iPSC lines are similar to those of diploid iPSC lines. Representative brightfield images of human diploid (H1), human auto-tetraploid (H1H1a), chimpanzee diploid (C1), chimpanzee auto-tetraploid (C1C1a), allo-tetraploid (H1C1a), and chrX-marked allo-tetraploid (H1C1a-X1) lines are shown. Scale bars are 1mm.

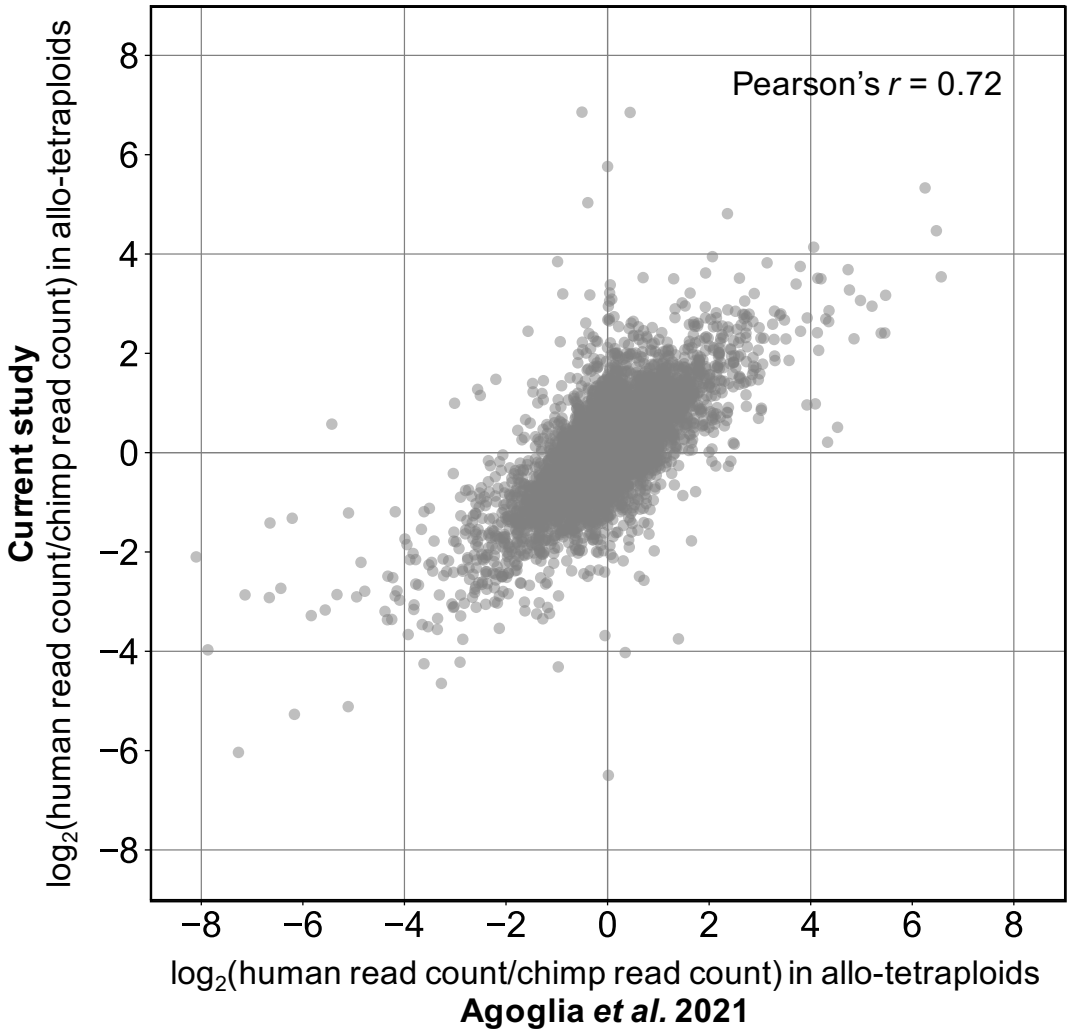


Fig. S2. Allele-specific expression in human-chimpanzee allo-tetraploid iPSCs is reproducible across studies. Allele-specific expression (ASE) $log_2(FC)$ values from RNAseq generated by Agoglia *et al.* 2021 (40) (x-axis) and ASE $log_2(FC)$ values from the RNAseq data reported in this study (y-axis) are highly concordant (Pearson's $r = 0.72$). Allo-tetraploid cells were derived from independent human-chimpanzee iPSC fusion events in the two studies, and different pipelines were used for mapping reads, assigning reads to the human or chimpanzee version of a gene, and calling genes with significant ASE. ASE differences in human-chimpanzee allo-tetraploid iPSCs are thus highly reproducible and robust to different analysis methods.

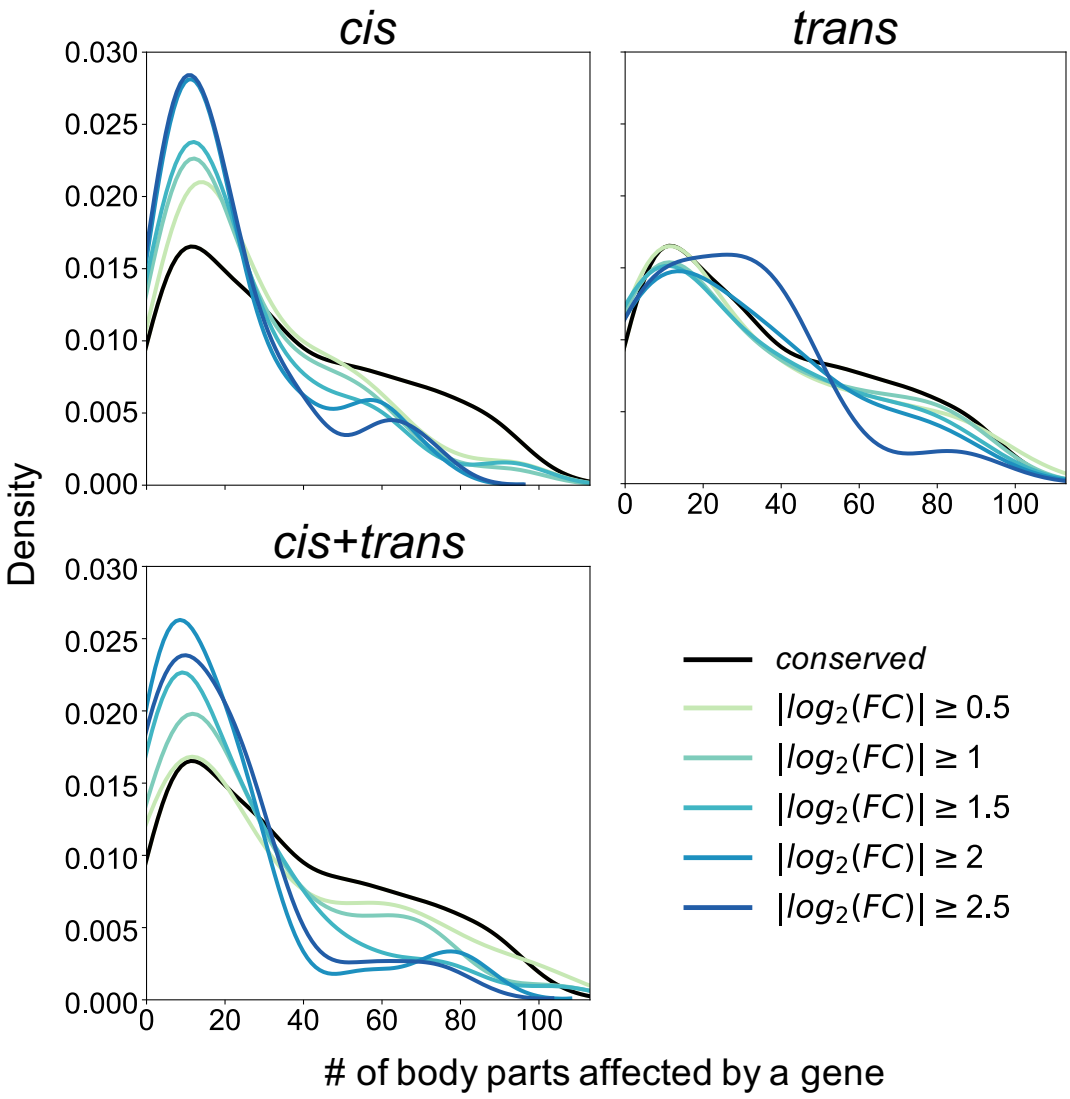


Fig. S3. Genes with increasingly divergent expression between human and chimpanzee iPSCs influence fewer body parts for *cis* and *cis+trans* regulatory types. Density plots (smoothed histograms) showing the distribution of body parts influenced by genes (according to the Gene ORGANizer database (41)) with human-chimpanzee expression differences due to *cis* (upper left), *trans* (upper right), and *cis+trans* (lower left) regulatory changes at increasing $|log_2(FC)|$ cutoffs. The *cis-trans* category is not included because only 5 genes have $|log_2(FC)\geq 1|$. For genes classified as *cis* and *cis+trans*, the median number of body parts influenced decreases with higher $|log_2(FC)|$ cutoffs (22, 18, 17, 15, 15 body parts and 22.5, 18, 16, 14, 14 body parts, respectively, for $|log_2(FC)|\geq 0.5, 1, 1.5, 2, 2.5$). All comparisons between the median number of body parts influenced by *conserved* genes (median of 30 body parts influenced) and by *cis* or *cis+trans* genes at the various $|log_2(FC)|$ cutoffs are statistically significant (adjusted $p < 0.04$ by two-tailed Mann-Whitney U test after FDR correction). This trend does not hold for gene expression differences due to *trans*-regulatory changes (adjusted $p > 0.19$ for all comparisons between *conserved* genes and *trans* genes at the various $|log_2(FC)|$ cutoffs).

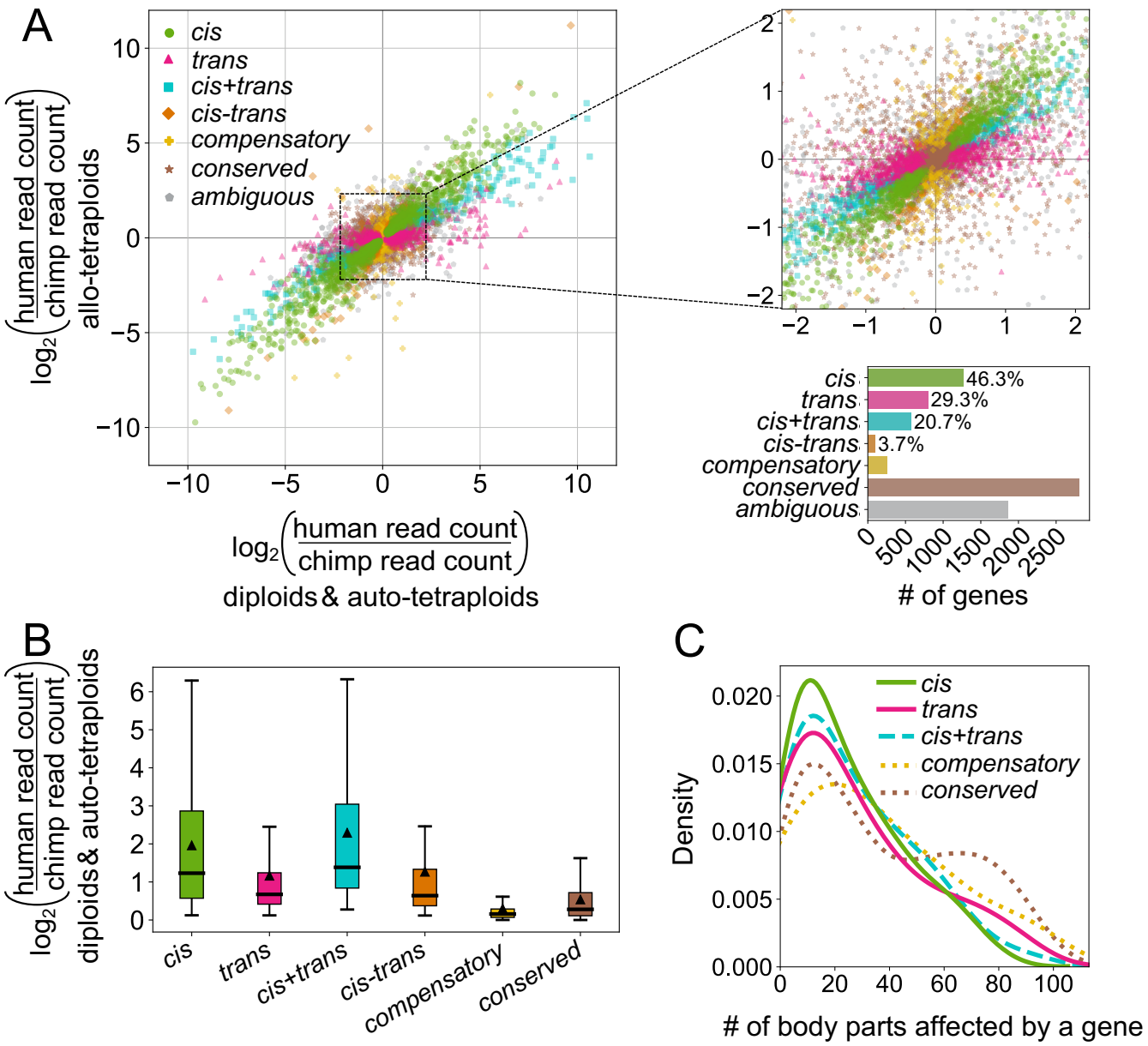


Fig. S4. *Cis* and *trans* analysis results are robust to aneuploidies. Removing chromosomes with aneuploidies or abnormalities in any of the cell lines used for RNAseq does not meaningfully change the observed *cis* and *trans* trends demonstrated in Fig. 2C-D. In addition to genes on human chromosome 18 and their orthologous genes in chimpanzee (deletion of one copy of human chr18q is shared by a subset of cell lines and was removed for the *cis* and *trans* analysis shown in Fig. 2C-E), genes on human chromosomes 7, 12, 20, and Y (and orthologous genes in chimpanzee) and chimpanzee chromosomes 1, 2A, 2B, 11, 13, 14, 17, 19, 20, and Y (and orthologous genes in human) were removed prior to analysis. **(A)** See Fig. 2C legend. **(B)** See Fig. 2D legend. All pairwise comparisons are statistically significant by two-tailed Mann-Whitney U test after FDR correction with adjusted $p < 10^{-5}$ except *trans* compared to *cis-trans* ($p = 0.28$). **(C)** See Fig. 2E legend. Genes classified as *cis*, *trans*, or *cis+trans* tend to influence fewer body parts than *conserved* genes (median 19, 20, 19.5 body parts, respectively, compared to median 29 body parts for *conserved* genes, adjusted $p = 0.0029, 0.024, 0.024$ by two-tailed Mann-Whitney U test after FDR correction). Note that the comparison between the *trans* and *conserved* categories is not statistically significant in Fig. 2E.

Fig. S5. Distribution of genome-wide inter-specific recombination events identified by haplotagging. Representative chromosome plots showing locations of inter-specific molecules detected by haplotagging after genome-wide sequencing and filtering. Green: cells treated with gRNA-chr20 (g20). Orange: cells treated with gRNA-chr20 and ML216 (g20+ML216).

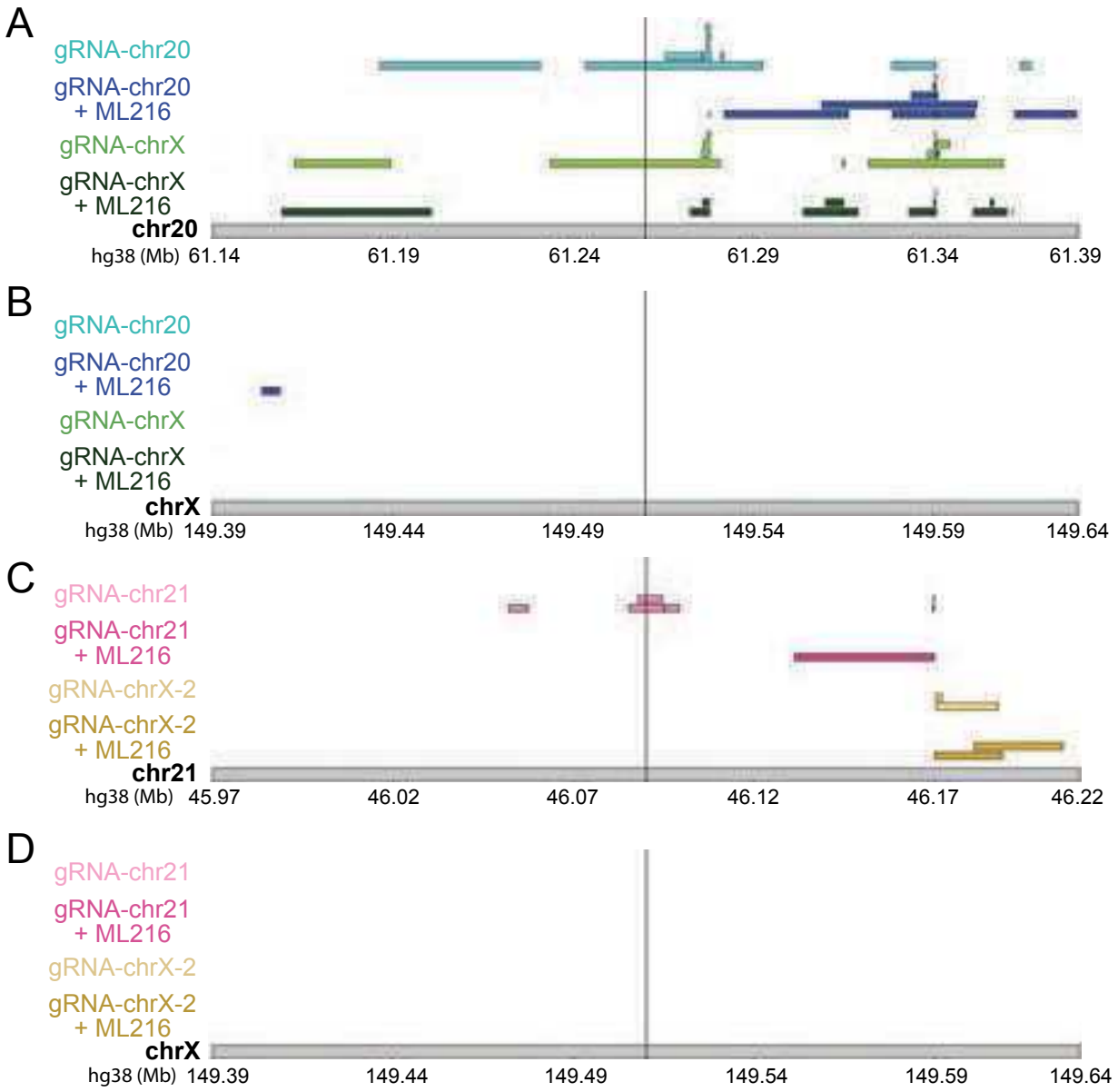


Fig. S6. CRISPR targeting does not elevate inter-specific recombination rates at target loci. Plot of inter-specific recombination events in the 250 kb window surrounding CRISPR target loci on chr20 **(A)**, chrX **(B)**, chr21 **(C)**, or a second guide location on chrX **(D)**. Each horizontal rectangle represents the boundaries of an inter-specific recombination event detected by haplotagging. Vertical lines indicate the gRNA target site. Events are filtered for molecules that contain only 1 inter-specific event and have > 5 supporting variants per species but are otherwise pre-filtering. The lack of enrichment at the target sites does not change with different filters (SI Methods).

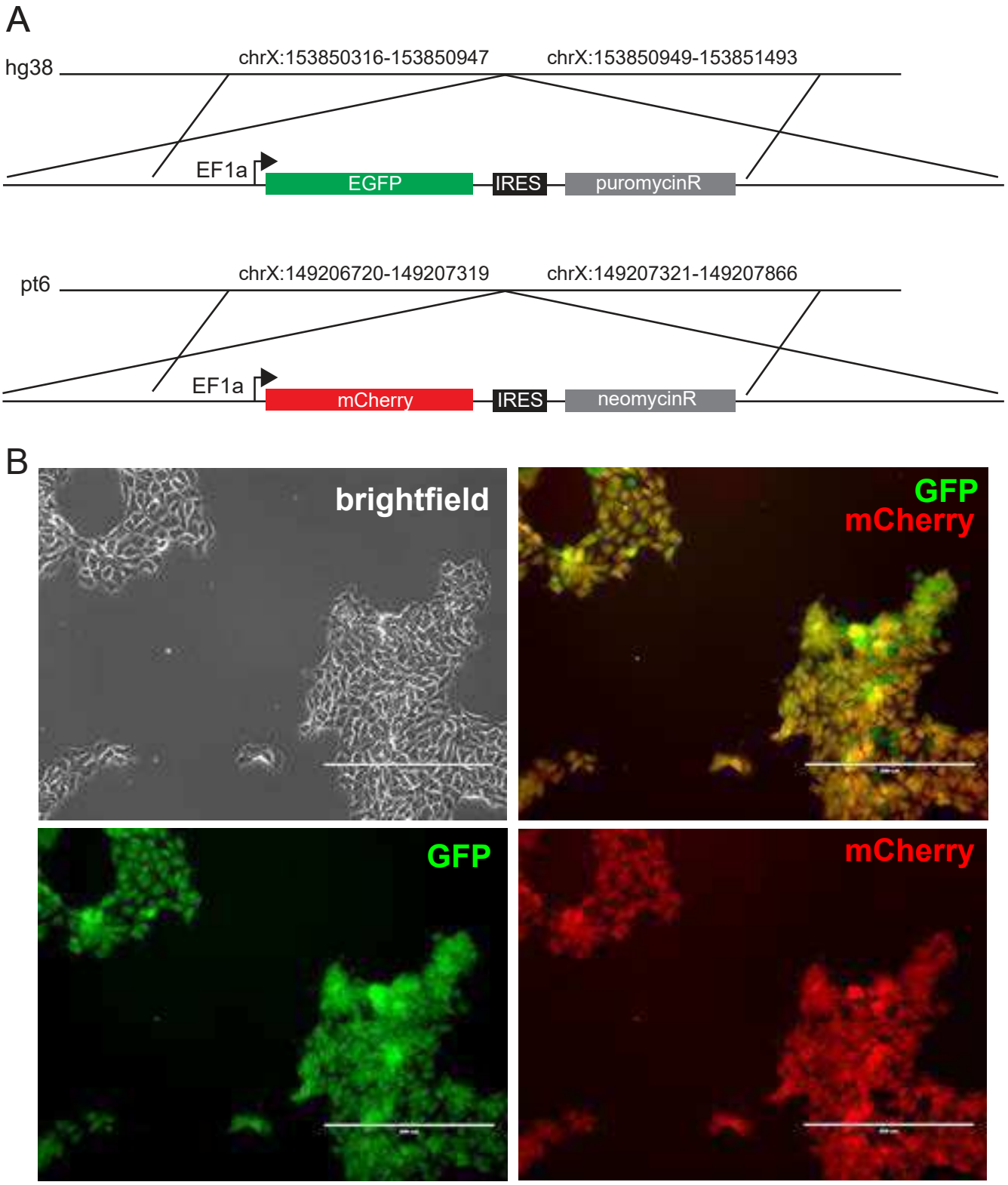


Fig. S7. Generation of fluorescently-marked allo-tetraploid lines. Construct diagram and microscopy images for fluorescently-marked line. **(A)** Constructs containing EGFP or mCherry were inserted onto the human or chimpanzee chrX, respectively, using CRISPR-guided homologous recombination (Materials and Methods). Coordinates show locations of human and chimpanzee homology arms used in the constructs. **(B)** Allo-tetraploid H1C1a-X1 shown in brightfield, GFP, and mCherry. Cells marked with both GFP and mCherry appear yellow.

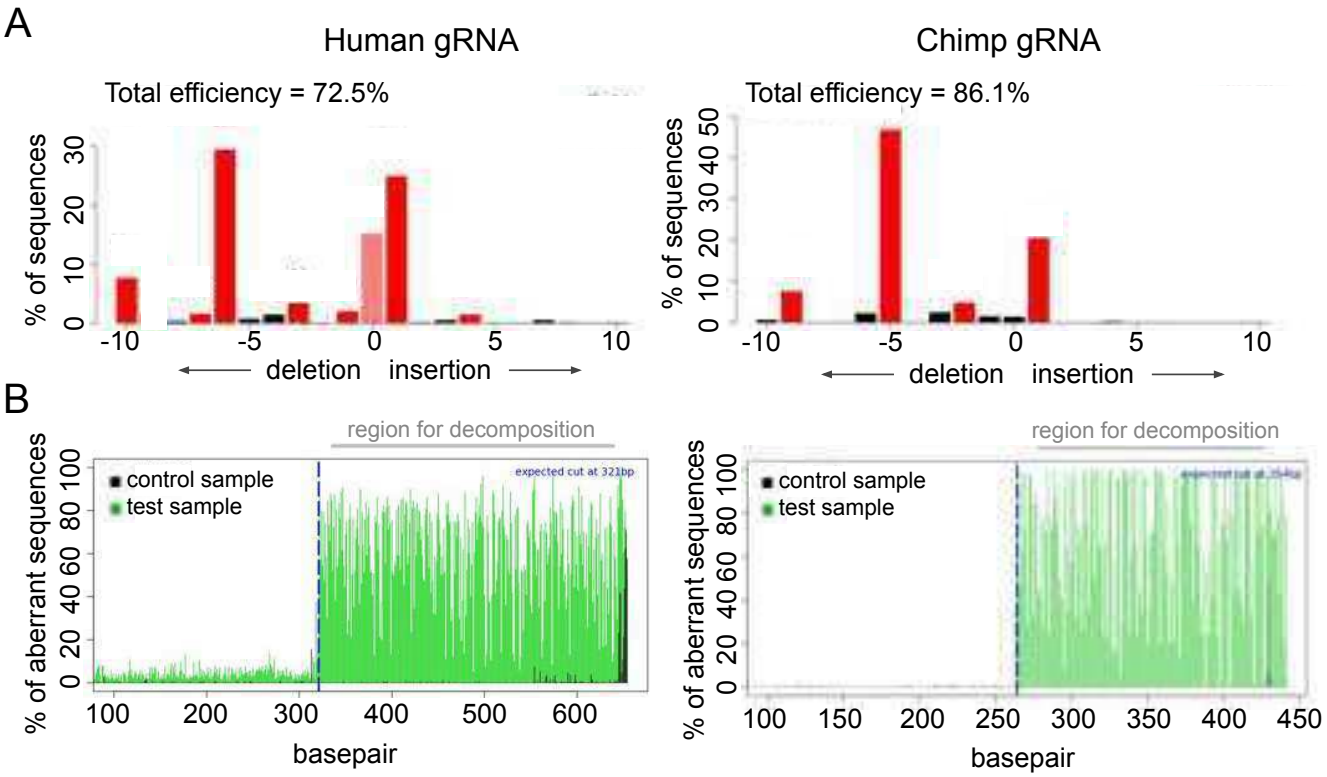


Fig. S8. CRISPR/Cas9 gRNA editing efficiency and indel spectra for human and chimpanzee chrX guides. (A) For human- and chimpanzee-specific gRNAs, the spectrum and frequency of small insertions and deletions, gRNA efficiency, and **(B)** aberrant sequence signal plots are shown. Plots generated with Sanger sequence data in TIDE (Tracking of Indels by DEcomposition) (36).

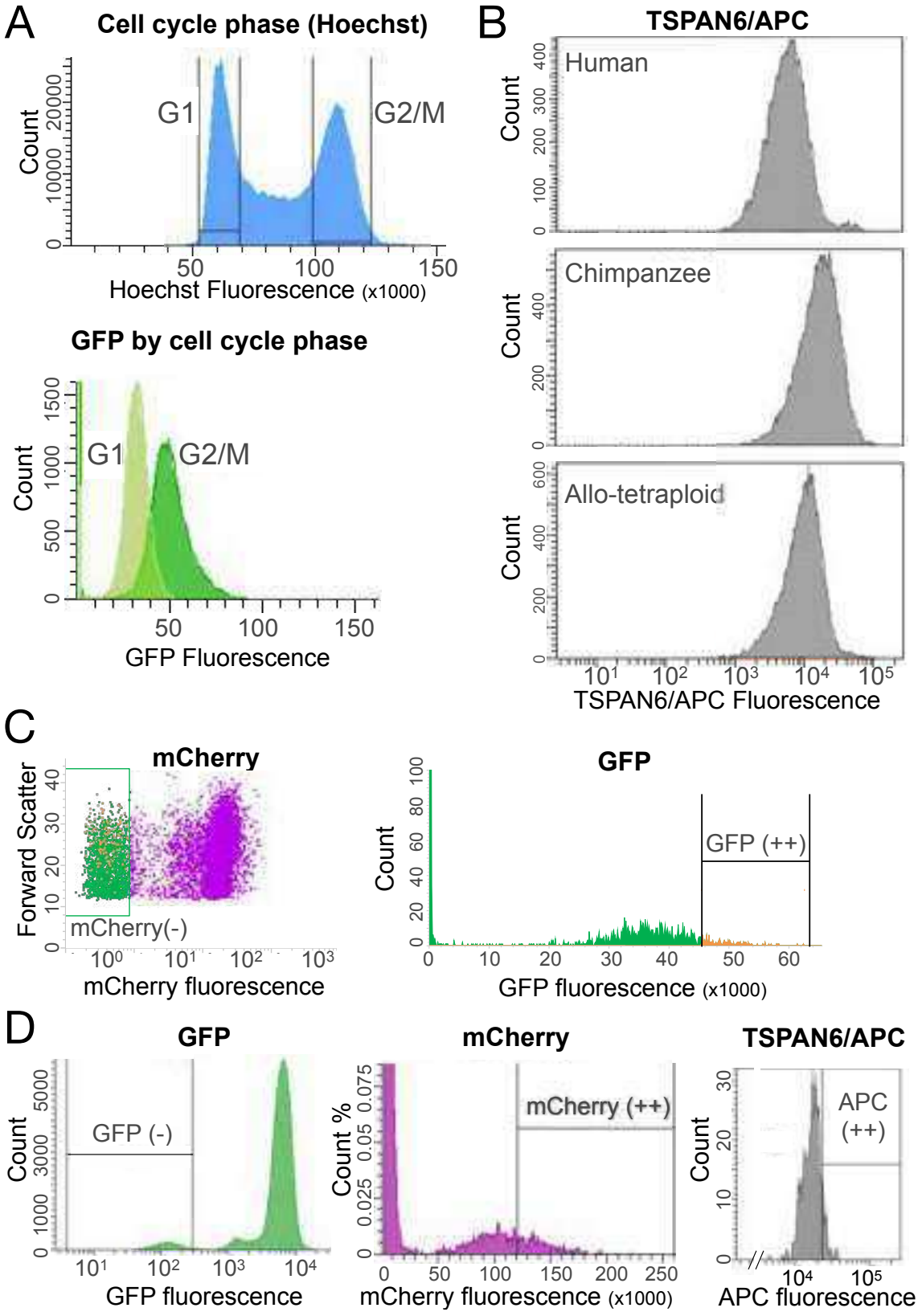


Fig. S9. Fluorescence activated cell sorting (FACS) plots for chrX targeting. (A) Cell cycle phase determined by Hoechst peaks shows that G2/M cells exhibit higher GFP fluorescence than G1 cells. **(B)** Staining for TSPAN6 cell-surface protein with APC secondary antibody shows that chimpanzee TSPAN6-APC fluorescence intensity is higher than human TSPAN6-APC fluorescence intensity, with allo-tetraploid cells intermediate between human and chimpanzee values. **(C)** After G1 gating, cells treated with chimpanzee-specific gRNA are sorted for negative mCherry and high GFP fluorescence. **(D)** Cells treated with human-specific gRNA are sorted for negative GFP, high mCherry, and high TSPAN6-APC.

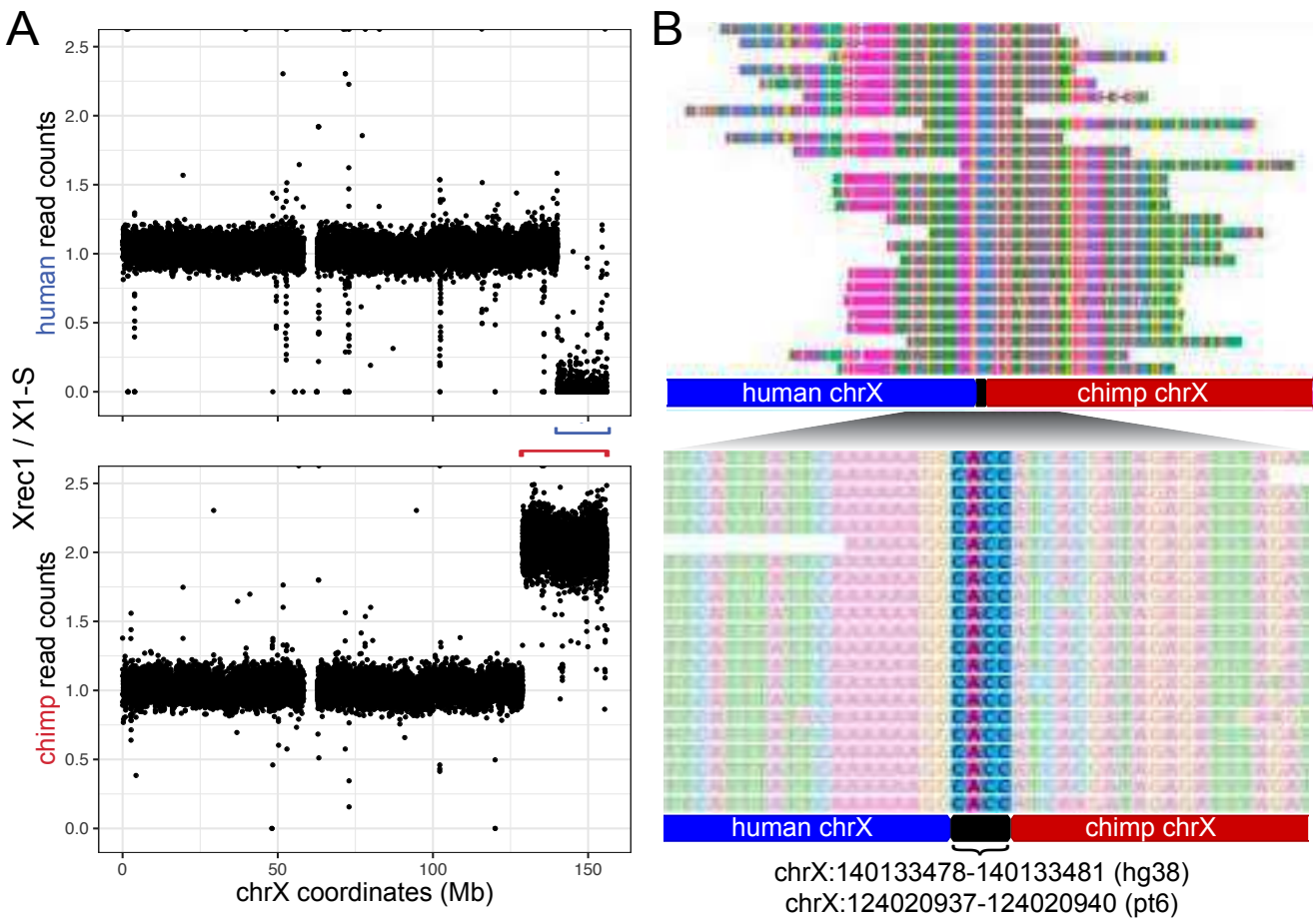


Fig. S10. DNA sequencing identifies the site of recombination between human and chimpanzee X chromosomes. Whole-genome DNA sequencing data from the recombinant allo-tetraploid line H1C1a-X1-Xrec1 (Xrec1). **(A)** Read counts that align to either the human or chimpanzee allele along chrX were normalized to read counts for H1C1a-X1-S (X1-S), a control sample also sequenced in parallel (see Materials and Methods). This ratio was plotted along the X chromosome in hg38 coordinates. Blue bracket: region with no human read counts in Xrec1. Red bracket: larger region with twice as many chimpanzee read counts in Xrec1. **(B)** Reads that span the inter-specific recombination site in Xrec1 align to the appropriate locations in human chrX and chimpanzee chrX. The recombination site is a 4bp microhomology (highlighted region in close-up) that is found in both human chrX and chimpanzee chrX at the indicated coordinates.

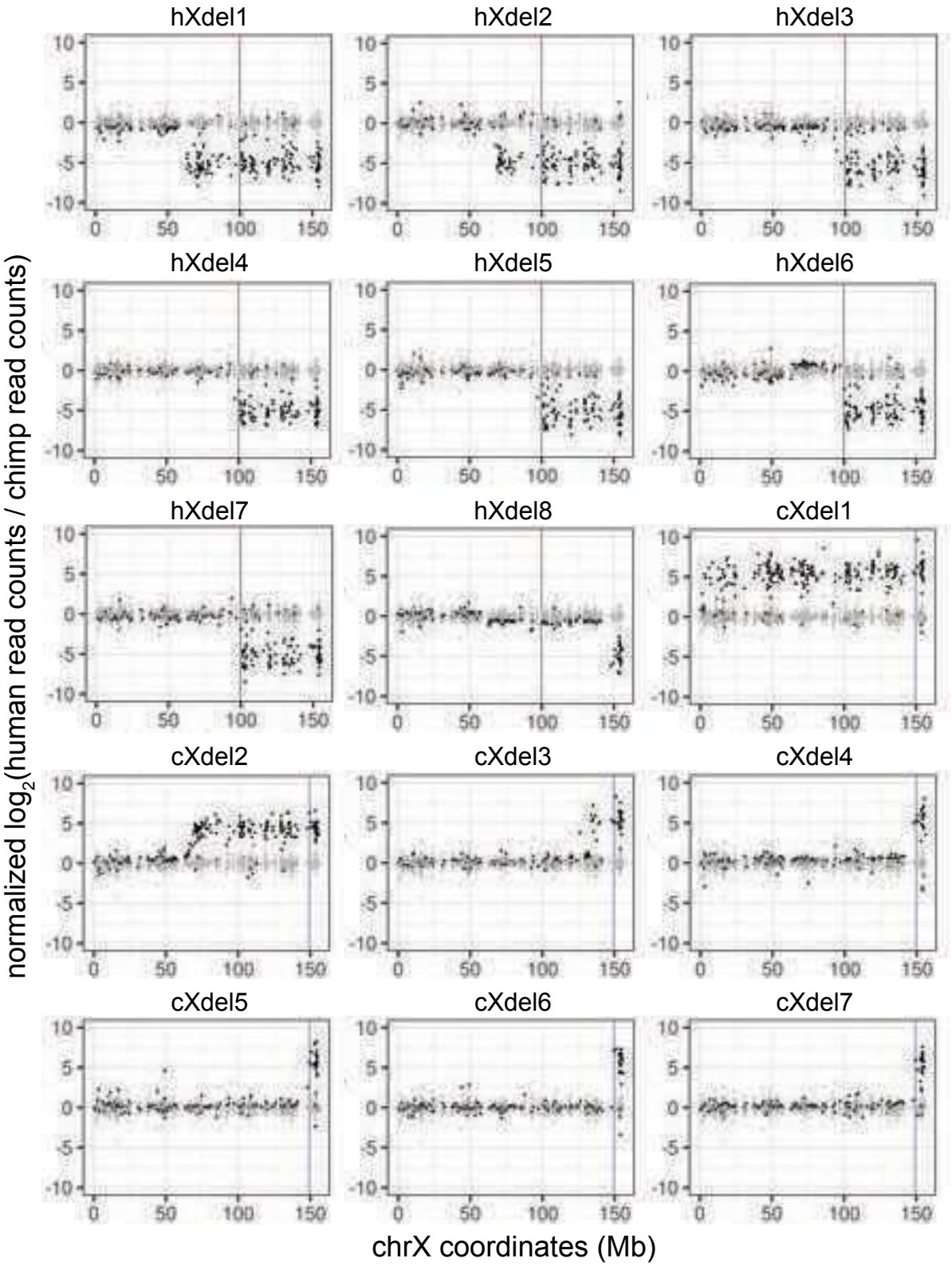


Fig. S11. RNAseq of chrX lines localizes terminal deletion breakpoints. The relative allelic expression of genes along chromosome X is plotted for each of the chrX deletion lines (Table S1). The y-axis is the ratio of reads that map to the human or chimpanzee allele in the deletion line (black) normalized to the ratio of reads that map to the human or chimpanzee allele in the control (non-deletion) lines (gray) (see SI Methods). Each dot represents a gene on chromosome X plotted along the x-axis at its hg38 coordinate. The vertical line is the species-specific gRNA target site used to generate each deletion line.

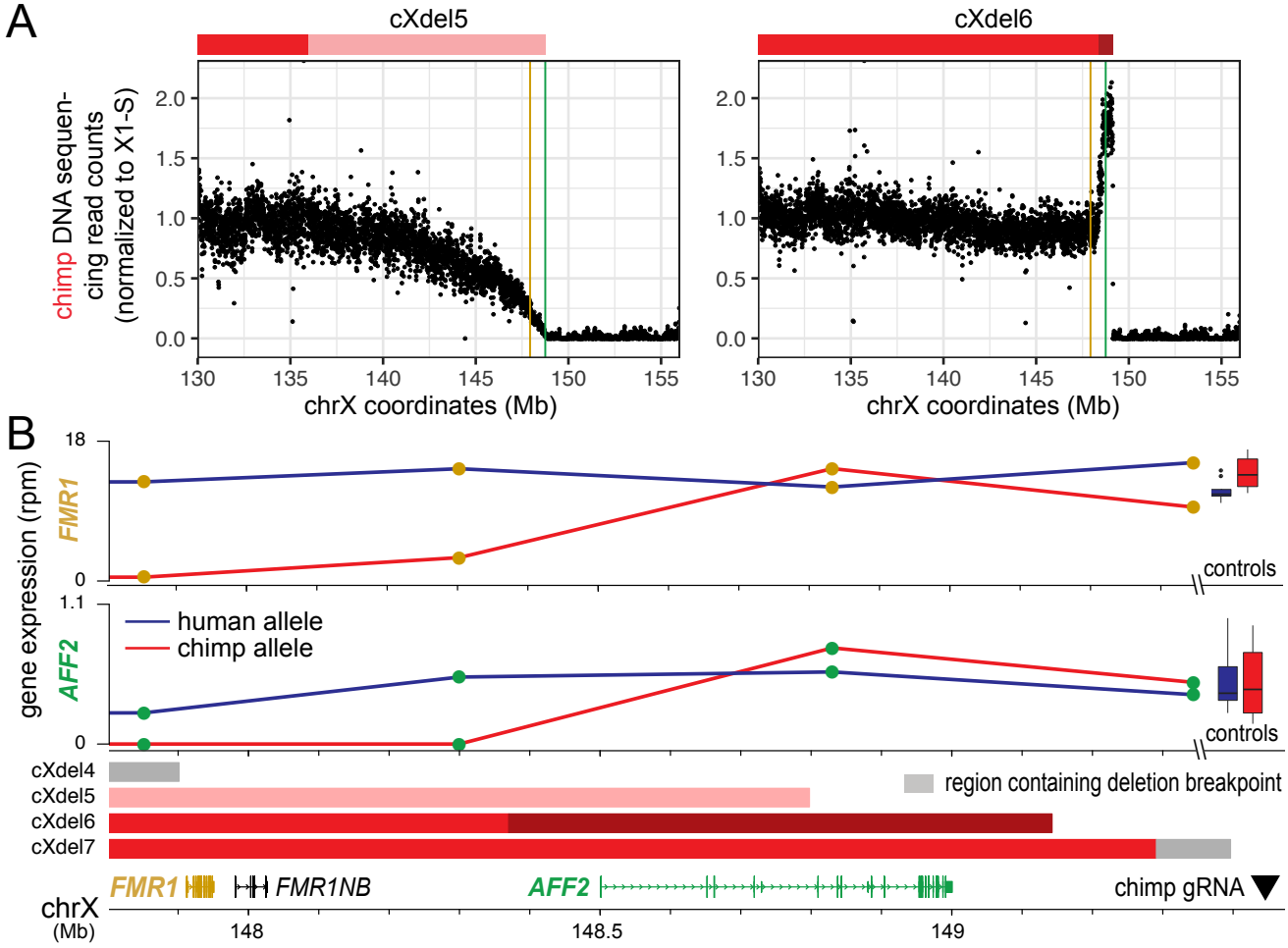


Fig. S12. Mapping chromosome breakpoints and gene expression levels in chrX deletion lines. (A) Whole-genome DNA sequencing was performed for two chimpanzee chrX deletion lines, cXdel5 and cXdel6. The ratio of the read counts that align to the chimpanzee allele for each deletion line was normalized to a control line, X1-S, and plotted along the X chromosome in hg38 coordinates (SI Methods). Colored bars above each plot indicate regions showing evidence of staggered deletions in cXdel5 (pink) or staggered insertions in cXdel6 (dark red), likely arising from a mixture of endpoints within the cell lines. Yellow line: location of *FMR1*. Green line: location of *AFF2*. **(B)** Expression of human alleles of *FMR1* or *AFF2* in the four chimpanzee chrX deletion lines is similar to control lines, as expected. Expression of the chimpanzee alleles of *FMR1* and *AFF2* is missing in cXdel4, whose terminal deletion includes both genes. Expression of chimpanzee *FMR1* is lower in cXdel5, likely corresponding to heterogeneous deletion of the gene in approximately ~75% of cells (panel A above). Expression of chimpanzee *FMR1* and *AFF2* appears normal in cXdel6 and cXdel7, whose terminal deletions do not include these loci. Red: regions of chimpanzee chrX present in line. Pink: region in cXdel5 containing non-clonal deletions. Dark red: region in cXdel6 containing non-clonal insertions. Gray: regions containing deletion breakpoints based on PCR assays and gene expression profiling (Materials and Methods).

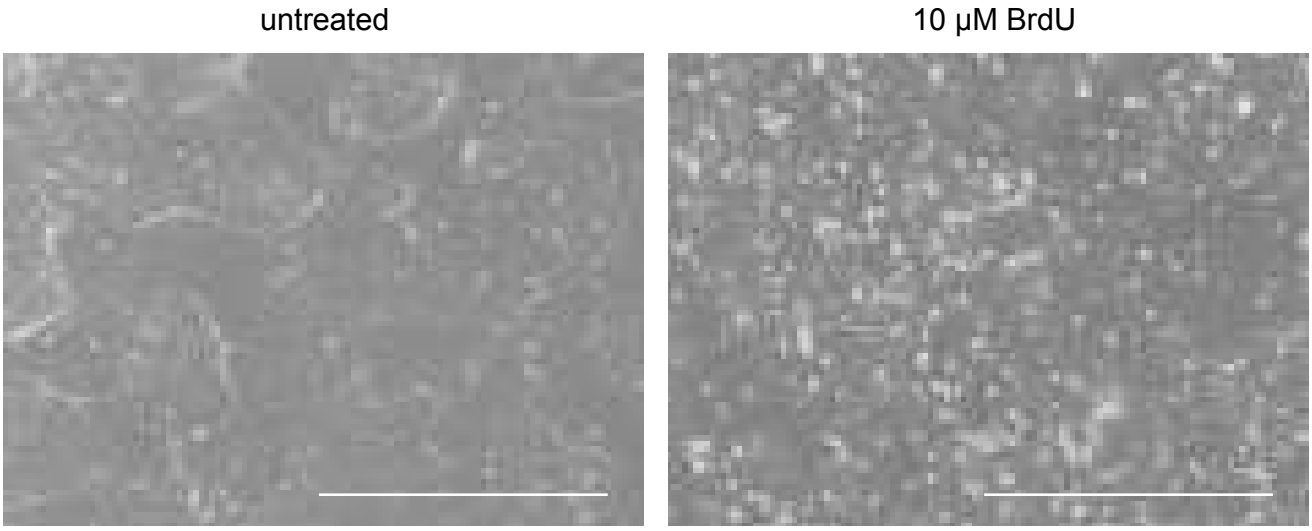


Fig. S13. BrdU induces differentiation of iPSCs. After passaging, iPSCs treated with 10μM of BrdU (right panel) are flatter and more spread out compared to untreated iPSCs (left panel). BrdU-treated cells also do not form the colonies typical of iPSCs and fail to divide, suggesting that they have terminally differentiated. Scale bars are 1mm.

SI Dataset S1 (Table S1)
iPSC lines used and generated in the current study.

SI Dataset S2 (Table S2)
Primers and gRNAs.

SI Dataset S3 (Table S3)
Trilineage differentiation results.

SI Dataset S4 (Table S4)
Differential gene expression analysis of diploid and auto-tetraploid iPSC lines.

SI Dataset S5 (Table S5)
Differential gene expression, allele-specific gene expression, and regulatory type (*cis/trans*) analysis between humans and chimpanzees in diploid, auto-tetraploid, and allo-tetraploid iPSC lines.

SI Dataset S6 (Table S6)
Gene ontology enrichments for regulatory type (*cis/trans*) categories.

SI Dataset S7 (Table S7)
qPCR and PCR results on chrX for sorted colonies treated with CRISPR+ML216.

SI Dataset S8 (Table S8)
Differential gene expression analysis of chrX deletion iPSC lines.

References

1. I Gallego Romero, et al., A panel of induced pluripotent stem cells from chimpanzees: A resource for comparative functional genomics. *eLife* **4**, e07103 (2015).
2. J Beers, et al., Passaging and colony expansion of human pluripotent stem cells by enzyme-free dissociation in chemically defined culture conditions. *Nat. Protoc.* **7**, 2029–2040 (2012).
3. KM Loh, et al., Mapping the pairwise choices leading from pluripotency to human bone, heart, and other mesoderm cell types. *Cell* **166**, 451–467 (2016).
4. B Howe, A Umrigar, F Tsien, Chromosome preparation from cultured cells. *J. Vis. Exp.*, e50203 (2014).
5. O Fedrigo, et al., A pipeline to determine RT-qPCR control genes for evolutionary studies: Application to primate gene expression across multiple tissues. *PLoS ONE* **5**, e12545 (2010).
6. C Bock, et al., Reference maps of human ES and iPS cell variation enable high-throughput characterization of pluripotent cell lines. *Cell* **144**, 439–452 (2011).
7. CE Filby, et al., Stimulation of Activin A/Nodal signaling is insufficient to induce definitive endoderm formation of cord blood-derived unrestricted somatic stem cells. *Stem Cell Res. & Ther.* **2**, 16 (2011).
8. Y Panina, A Germond, S Masui, TM Watanabe, Validation of common housekeeping genes as reference for qPCR gene expression analysis during iPS reprogramming process. *Sci. Reports* **8**, 8716 (2018).
9. YL Kuang, et al., Evaluation of commonly used ectoderm markers in iPSC trilineage differentiation. *Stem Cell Res.* **37**, 101434 (2019).
10. A Gunne-Braden, et al., *GATA3* mediates a fast, irreversible commitment to *BMP4*-driven differentiation in human embryonic stem cells. *Cell Stem Cell* **26**, 693–706.e9 (2020).
11. Y Li, et al., Generation of an induced pluripotent stem cell line SDUBMSi005-A from a patient with double primary gastric and colon carcinoma. *Stem Cell Res.* **53**, 102253 (2021).
12. M Martin, Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J.* **17**, 10–12 (2011).
13. S Andrews, et al., *FastQC*. (2010) Babraham Institute.
14. A Dobin, et al., STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
15. Y Liao, GK Smyth, W Shi, featureCounts: An efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930 (2014).
16. AD Yates, et al., Ensembl 2020. *Nucleic Acids Res.* **48**, D682–D688 (2020).
17. J Zhu, et al., Comparative genomics search for losses of long-established genes on the human lineage. *PLoS Comput. Biol.* **3**, e247 (2007).
18. A McKenna, et al., The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
19. GA Van der Auwera, et al., From FastQ data to high confidence variant calls: The Genome Analysis Toolkit best practices pipeline. *Curr. Protoc. Bioinforma.* **43**, 1–11 (2013).
20. MI Love, W Huber, S Anders, Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
21. CJ McManus, et al., Regulatory divergence in *Drosophila* revealed by mRNA-seq. *Genome Res.* **20**, 816–825 (2010).
22. X Shi, et al., *Cis*- and *trans*-regulatory divergence between progenitor species determines gene-expression novelty in *Arabidopsis* allopolyploids. *Nat. Commun.* **3**, 950 (2012).
23. HB Fraser, Improving estimates of compensatory *cis–trans* regulatory divergence. *Trends Genet.* **35**, 88 (2019).
24. G Yu, LG Wang, Y Han, QY He, clusterProfiler: An R package for comparing biological themes among gene clusters. *OMICS* **16**, 284–287 (2012).
25. DM Stults, MW Killen, AJ Pierce, The sister chromatid exchange (SCE) assay. *Methods Mol. Biol.* **1105**, 439–455 (2014).
26. JI Meier, et al., Haplotype tagging reveals parallel formation of hybrid races in two butterfly species. *Proc. Natl. Acad. Sci. USA* **118**, e2015005118 (2021).
27. A Shajii, I Numanagić, B Berger, Latent variable model for aligning barcoded short-reads improves downstream analyses. *Res. Comput. Mol. Biol.* **10812**, 280–282 (2018).
28. J Navarro Gonzalez, et al., The UCSC Genome Browser database: 2021 update. *Nucleic Acids Res.* **49**, D1046–D1057 (2021).
29. HM Amemiya, A Kundaje, AP Boyle, The ENCODE blacklist: Identification of problematic regions of the genome. *Sci. Reports* **9**, 9354 (2019).
30. G Luo, et al., Cancer predisposition caused by elevated mitotic recombination in Bloom mice. *Nat. Genet.* **26**, 424–429 (2000).
31. K Yusa, et al., Genome-wide phenotype analysis in ES cells by regulated disruption of Bloom's syndrome gene. *Nature* **429**, 896–899 (2004).
32. S Lazzarano, et al., Genetic mapping of species differences via *in vitro* crosses in mouse embryonic stem cells. *Proc. Natl. Acad. Sci. USA* **115**, 3680–3685 (2018).
33. H Wickham, *ggplot2: Elegant graphics for data analysis.* (Springer-Verlag New York), (2016).
34. B Gel, E Serra, karyoploteR: An R/Bioconductor package to plot customizable genomes displaying arbitrary data. *Bioinformatics* **33**, 3088–3090 (2017).
35. JI Wucherpfennig, CT Miller, DM Kingsley, Efficient CRISPR-Cas9 editing of major evolutionary loci in sticklebacks.

Evol. Ecol. Res. **20**, 107–132 (2019).
36. EK Brinkman, T Chen, M Amendola, B van Steensel, Easy quantitative assessment of genome editing by sequence trace decomposition. *Nucleic Acids Res.* **42**, e168–e168 (2014).
37. Y Yoshimura, A Yamanishi, T Kamitani, JS Kim, J Takeda, Generation of targeted homozygosity in the genome of human induced pluripotent stem cells. *PLoS ONE* **14**, e0225740 (2019).
38. H Li, R Durbin, Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
39. H Li, et al., The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
40. RM Agoglia, et al., Primate cell fusion disentangles gene regulatory divergence in neurodevelopment. *Nature* **592**, 421–427 (2021).
41. D Gokhman, et al., Gene ORGANizer: Linking genes to the organs they affect. *Nucleic Acids Res.* **45**, W138–W145 (2017).
42. L Yu, et al., Core pluripotency factors promote glycolysis of human embryonic stem cells by activating *GLUT1* enhancer. *Protein Cell* **10**, 668–680 (2019).
43. SP Bharathan, et al., Systematic evaluation of markers used for the identification of human induced pluripotent stem cells. *Biol. Open* **6**, 100–108 (2017).