Genetics and Genomics
Evolutionary Biology

Cell type-specific *cis*-regulatory divergence in gene expression and chromatin accessibility revealed by human-chimpanzee hybrid cells

Ban Wang, Alexander L. Starr, Hunter B. Fraser

Department of Biology, Stanford University, Stanford, CA, USA

Reviewed Preprint

Revised by authors after peer review.

About eLife's process

Reviewed preprint version 2
January 10, 2024 (this version)

Reviewed preprint version 1
August 15, 2023

Sent for peer review
May 29, 2023

Posted to preprint server
May 23, 2023

Abstract

Although gene expression divergence has long been postulated to be the primary driver of human evolution, identifying the genes and genetic variants underlying uniquely human traits has proven to be quite challenging. Theory suggests that cell type-specific *cis*-regulatory variants may fuel evolutionary adaptation due to the specificity of their effects. These variants can precisely tune the expression of a single gene in a single cell type, avoiding the potentially deleterious consequences of *trans*-acting changes and non-cell type-specific changes that can impact many genes and cell types, respectively. It has recently become possible to quantify human-specific *cis*-acting regulatory divergence by measuring allele-specific expression in human-chimpanzee hybrid cells—the product of fusing induced pluripotent stem (iPS) cells of each species *in vitro*. However, these *cis*-regulatory changes have only been explored in a limited number of cell types. Here, we quantify human-chimpanzee *cis*-regulatory divergence in gene expression and chromatin accessibility across six cell types, enabling the identification of highly cell type-specific *cis*-regulatory changes. We find that cell type-specific genes and regulatory elements evolve faster than those shared across cell types, suggesting an important role for genes with cell type-specific expression in human evolution. Furthermore, we identify several instances of lineage-specific natural selection that may have played key roles in specific cell types, such as coordinated changes in the *cis*-regulation of dozens of genes involved in neuronal firing in motor neurons. Finally, using novel metrics and a machine learning model, we identify genetic variants that likely alter chromatin accessibility and transcription factor binding, leading to neuron-specific changes in the expression of the neurodevelopmentally important genes *FABP7* and *GAD1*. Overall, our results demonstrate that integrative analysis of *cis*-regulatory divergence in chromatin accessibility and gene expression across cell types is a promising approach to identify the specific genes and genetic variants that make us human.

eLife assessment

This is an **important** study that leverages a human-chimpanzee tetraploid iPSC model to test whether cis-regulatory divergence between species tends to be cell type-specific. The evidence supporting the study's primary conclusions together provide **convincing** evidence for enrichment of species differences in gene regulation in cell type-specific genes and regulatory elements, motivating future work with larger sample sizes of cell lines. This work will be of broad interest in evolutionary and functional genomics.

Introduction

In the past few million years, humans have evolved a multitude of unique phenotypes (Shave et al., 2019; Vanderhaeghen & Polleux, 2023). For example, our cardiovascular system has evolved to enable extended periods of physical exertion and the unique aspects of our nervous system enable human language and toolmaking (Shave et al., 2019; Vanderhaeghen & Polleux, 2023). Previous research suggests that much of human adaptation may be caused by changes in gene expression (Fraser, 2013; King & Wilson, 1975; Reilly & Noonan, 2016; Romero et al., 2012). To catalog these changes, studies have compared gene expression in post-mortem tissues of humans and our closest living relatives, chimpanzees (Blake et al., 2020; Kelley & Gilad, 2020; Ma et al., 2022). Although thousands of differentially expressed genes have been identified in post-mortem samples, it is generally not possible to disentangle the effects of genetic differences from the effects of confounding factors such as differences in diet, environment, cell type abundances, age, post-mortem interval, etc. In addition, for many traits the relevant gene expression differences may be specific to early development, but it is impossible to study fetal development *in vivo* in non-human great apes due to both ethical and technical difficulties. To circumvent these issues, several groups have used great ape iPS cells to study differences in gene expression in cell types present in early development (Benito-Kwiecinski et al., 2021; Field et al., 2019; Kanton et al., 2019; Pavlovic et al., 2018). While the use of iPS cells addresses many of the confounding factors present in post-mortem comparisons, they also introduce new issues such as interspecies differences in iPS cell differentiation kinetics, efficiency, and maturation. Overall, it remains tremendously challenging to identify human-specific changes in gene expression, which limits our ability to link expression differences to either phenotypic differences or natural selection in the human lineage.

One particularly powerful means of studying the evolution of gene expression is through the measurement of allele specific expression (ASE) in hybrids between two species (Combs et al., 2018; Fraser, 2011; Hu et al., 2022; Mack & Nachman, 2017; Wittkopp & Kalay, 2012). This approach has the advantage of eliminating many confounding factors inherent to interspecies comparisons, including differences in cell type composition, environmental factors, developmental stage, and response to differentiation protocols. Because the trans-acting environments of the two alleles in a hybrid are identical, ASE has the additional benefit of reflecting only *cis*-regulatory changes, which are thought to be less pleiotropic and more likely to drive evolutionary adaptation than broader *trans*-acting changes (Agoglia et al., 2021; Gokhman et al., 2021; Prud'homme et al., 2007; Wittkopp & Kalay, 2012). Furthermore, ASE enables the use of powerful methods that can detect lineage-specific natural selection and, as a result, contribute to our understanding of the selective pressures that have shaped the evolution of a wide variety of species (Fraser, 2011). Until recently, it has not been possible to disentangle *cis*- and *trans*-acting changes fixed in the human lineage since humans cannot hybridize with any other species. However, the development of human-chimpanzee hybrid iPS cells via *in vitro* cell

fusion enables measurement of ASE in a wide variety of tissues and developmental contexts (Agoglia et al., 2021; Gokhman et al., 2021; Song et al., 2021). This provides an effective platform to investigate general principles of hominid gene expression evolution, detect lineage-specific selection, and identify candidate gene expression changes underlying human-specific traits.

While gene expression divergence between humans and chimpanzees is well-studied, there has been less focus on epigenetic differences, many of which are likely to underlie divergent gene expression (García-Pérez et al., 2021; Kozlenkov et al., 2020; Netherlands Brain Bank et al., 2016; Trizzino et al., 2017). Furthermore, these studies, regardless of whether they utilize post-mortem tissues or cell lines, are subject to the same confounding factors mentioned above. Analogous to ASE, one can use the assay for transposase accessible chromatin using sequencing (ATAC-seq) in interspecies hybrids to measure allele-specific chromatin accessibility (ASCA) (Buenrostro et al., 2013; Corces et al., 2017; Liang et al., 2021; S. Zhang et al., 2020). As with ASE, ASCA is unaffected by many confounders inherent to between-species comparisons and only measures *cis*-regulatory divergence. Perhaps most importantly, ASCA can implicate specific regulatory elements that likely underlie gene expression differences. These regulatory elements can then be more closely studied to identify the likely causal genetic variants and the molecular mechanisms by which those variants alter gene expression.

Here, we generated RNA-seq and ATAC-seq data from six cell types, derived from human-chimpanzee hybrid iPS cells, and quantified ASE and ASCA. Using this dataset, we identified thousands of genes and *cis*-regulatory elements showing cell type-specific ASE and ASCA. We found that cell type-specific genes and *cis*-regulatory elements are more likely to have divergent expression and accessibility than their more broadly expressed/accessible counterparts. Furthermore, we provide evidence for polygenic selection on the expression level of genes associated with physiologically relevant gene sets including sodium channels and syntaxin-binding proteins in motor neurons. Finally, we use newly developed metrics and machine learning algorithms to link cell type-specific differences in chromatin accessibility and gene expression and identify putative causal mutations underlying these differences. Using this pipeline we identified motor neuron-specific increases in promoter chromatin accessibility and gene expression for *FABP7*, which plays a key role in neurodevelopment but is not well-studied in neurons. In addition, we focus on a human-accelerated region (HAR) near the promoter of *GAD1*. While this region is accessible in all cell types, both the accessibility of the HAR and the expression of *GAD1* are only chimpanzee-biased in motor neurons. Analysis of scRNA-seq from human and chimpanzee brain organoids showed that increased expression of *GAD1* also occurs in ventral forebrain inhibitory neurons. Overall, this study provides insight into the evolution of gene expression in hominids as well as a resource that will inform functional genomic dissection of human-specific traits.

Results

Cis-regulatory divergence of gene expression in six cell types is largely cell type-specific or shared across all cell types

To measure genome-wide cis-regulatory divergence in gene expression, we performed RNA-seq on six cell types derived from human-chimpanzee hybrid iPS cells (**Fig. 1a**). The cell types profiled were from six diverse developmental lineages including the motor neuron (MN), cardiomyocyte (CM), hepatocyte progenitor (HP), pancreatic progenitor (PP), skeletal myocyte (SKM), and retinal pigment epithelium (RPE) lineages. These represent all three germ layers and a variety of organs (**Fig. 1a**). It is worth noting that these differentiations do not necessarily lead to a pure population of cells, but rather a population of cells with different levels of maturity along a particular developmental lineage. For example, the SKM population likely contains fully

differentiated muscle fibers as well as a small population of proliferating satellite cells; for clarity we refer to this as the SKM cell type. As these different cell types are not shared between tissues, we use cell type-specific and tissue-specific interchangeably throughout the manuscript.

Two independently generated hybrid lines were differentiated for each cell type and at least two biological replicates per hybrid line per cell type were collected (see Methods). Each cell type was sequenced to an average depth of 134 million paired-end reads (Supp **Fig. 1**). We used a computational pipeline to quantify ASE adapted from the pipeline introduced by Agoglia et al (Agoglia et al., 2021). Briefly, we computed ASE by mapping reads to both the human and chimpanzee genomes, correcting for mapping bias, and assigning reads to the human or chimpanzee genome if a read contained one or more human-chimpanzee single nucleotide differences (see Methods).

As expected, the samples clustered predominantly by cell type (**Fig. 1b-c**). Within four of the six cell type clusters, individual samples clustered by line rather than species of origin, potentially indicating line to line variability in differentiation (**Fig. 1b**). This highlights the importance of measuring ASE which, by definition, is measured within each line and so is robust to variability between lines. Indeed, when performing PCA within cell types using allelic counts (i.e. counting reads from the human allele and chimpanzee allele separately), human and chimpanzee species differences were clearly separated by principal component (PC) 1 or PC2 in each cell type (**Fig. 1d**, Supp **Fig. 2**). To assess the success of our differentiations, we examined each cell type for the expression of known marker genes in our RNA-seq data (**Fig. 1e**, Supp **Fig. 3**). All cell types express canonical marker genes and do not express pluripotency markers, indicating that the differentiations were successful (**Fig. 1e**, Supp **Fig 3**, Additional file 1) (Burridge et al., 2014; Chal et al., 2016; Korytnikov & Nostro, 2016; Mallanna & Duncan, 2013; Maury et al., 2015; Sharma et al., 2019).

Because our hybrid cells were grown concurrently with their human and chimpanzee diploid "parental" cells, we performed an additional check for purity of the hybrid lines by quantifying genome-wide ASE. We noticed that among our 25 RNA-seq samples, the two PP hybrid2 samples had a slight bias towards higher expression from the chimpanzee alleles across all chromosomes. This is likely due to a small fraction of contaminating chimpanzee cells in these samples. We corrected for this by reducing the chimpanzee allele counts such that the number of reads assigned to the human and chimpanzee alleles was equal. By simulating contamination of a hybrid sample with chimpanzee cells, we found that this correction was conservative and that the log fold-change estimates were largely unaffected by contamination after this correction (Methods, Supp **Fig. 4**).

We next investigated which genes were differentially expressed between the human and chimpanzee allele in each cell type (**Fig. 2a**). We identified thousands of genes showing significantly biased ASE in each cell type at a false discovery rate (FDR) cutoff of 0.05 (**Fig. 2b**). We detected a comparable number of ASE genes in all cell types except SKM. As a result, we repeated all subsequent analyses both including and excluding SKM and obtained qualitatively similar results regardless of whether SKM was included.

While a considerable number of genes had significant ASE in all cell types, many more genes only had significant ASE in a single cell type, suggesting cell type-specific *cis*-regulatory divergence (**Fig. 2b**). A notable family of developmentally important genes that exemplifies differences in ASE across cell types is the neurotrophins and their receptors (**Fig. 2c**) (Caporali & Emanueli, 2009; Huang & Reichardt, 2001). For example, *NTRK3*, which plays a key role in the development of the nervous system, is only differentially expressed in RPE and MN but is chimpanzee-biased in RPE and human-biased in MN (Supp **Fig. 5**) (Ichim et al., 2012; Naito et al., 2017). In addition, the gene coding for its primary ligand (*NTF3*) is differentially expressed in a variety of cell types yet is human-biased in all cell types except MN in which it is chimpanzee-biased (Supp

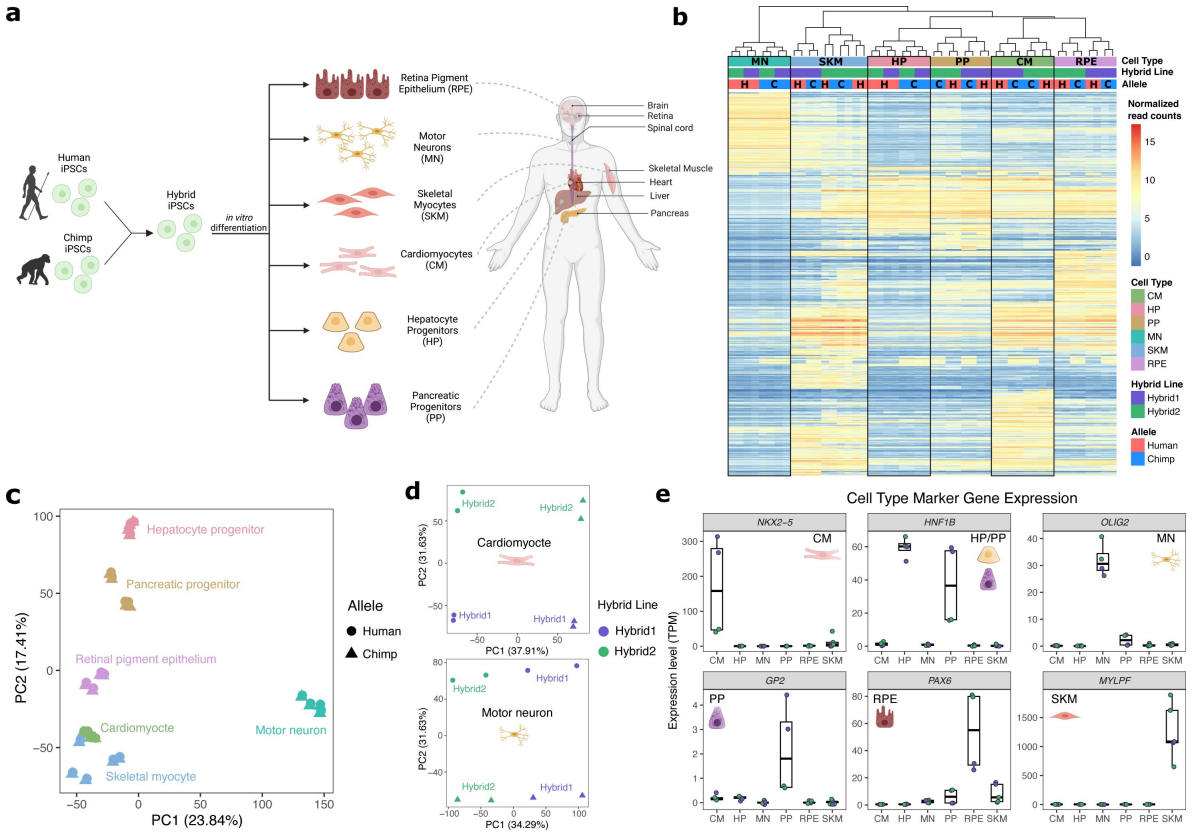


Figure 1

Allele-specific expression across diverse human-chimpanzee hybrid cell types.

a) Six cell types were differentiated from human-chimpanzee hybrid induced pluripotent stem cells. These six cell types represent diverse body systems, including motor neurons for the central nervous system, retinal pigment epithelium for eye, skeletal myocytes for skeletal muscle, cardiomyocytes for the heart, hepatocyte progenitors for the liver, and pancreatic progenitors for the pancreas. **b)** Heatmap showing the result of hierarchical clustering performed on genes with highly variable normalized allele counts. **c)** Result of running PCA on normalized allelic counts for all samples and cell types. **d)** Result of PCA performed on normalized allele counts for each individual cell type separately. Cardiomyocytes and motor neurons are shown here. **e)** Expression of marker genes for each cell type.

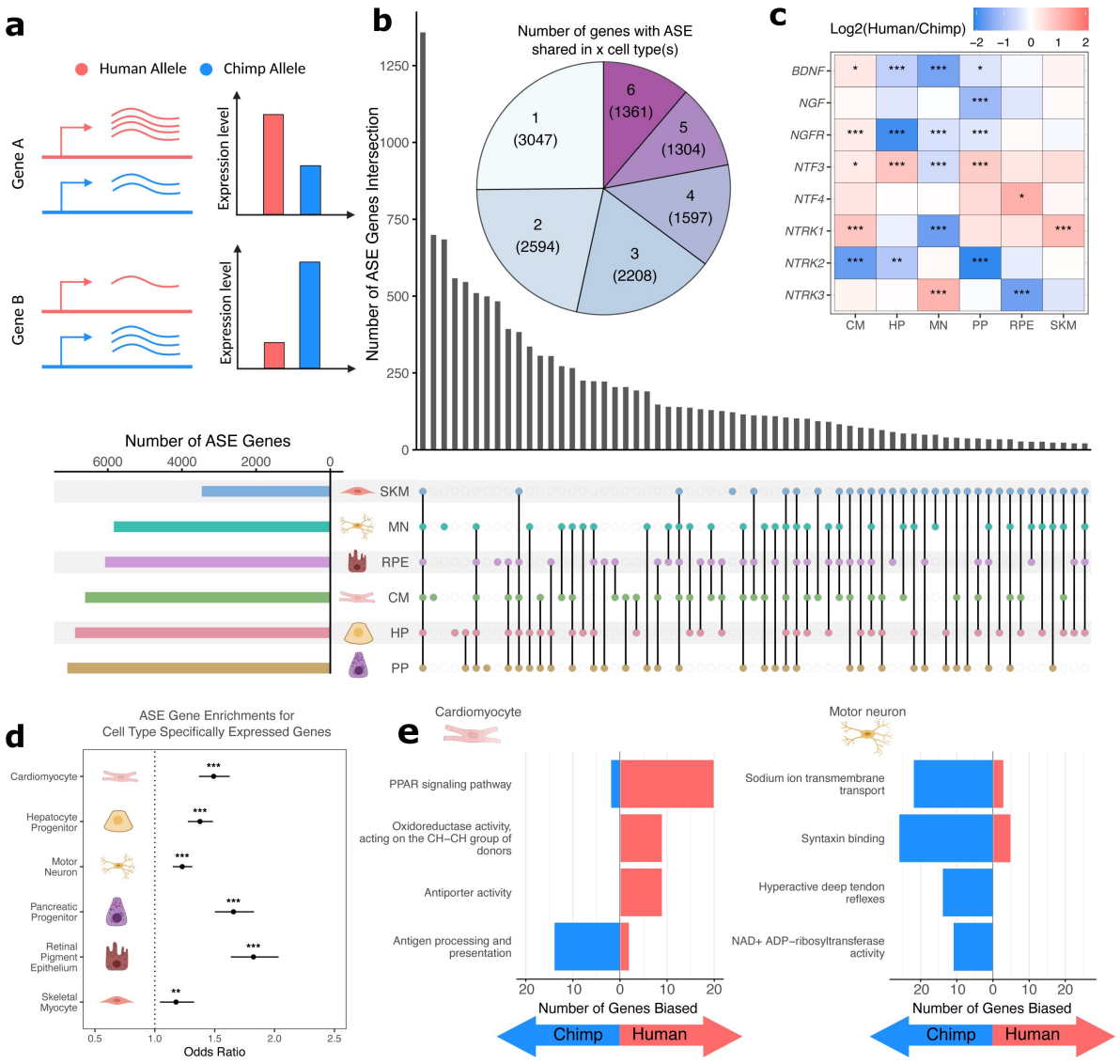


Figure 2

Human-chimpanzee ASE is largely cell type-specific.

a) Outline of measurement of allele-specific expression. Reads from the human and chimpanzee alleles are counted and differences in read counts identified. **b)** Thousands of genes with ASE were identified for each cell type, with many genes only showing ASE in a single cell type (Conway et al., 2017). **c)** The neurotrophins and their receptors as examples of genes showing cell type-specific ASE patterns. DESeq2 estimate of log2 fold-change (human/chimpanzee) are shown in the heatmap and significance is indicated by asterisks where *** indicates FDR < 0.005, ** indicates FDR < 0.01, and * indicates FDR < 0.05. Zero asterisks (i.e. a blank box) indicates FDR > 0.05. **d)** Plot showing that genes with ASE are enriched for genes showing cell type-specific expression patterns across all cell types. Asterisks indicate p-values rather than FDR using the same system as in 2c. **e)** Top gene sets with evidence for lineage-specific selection in cardiomyocytes and motor neurons are shown. The length of the bars indicates the number of genes in a category with biased expression in each cell type.

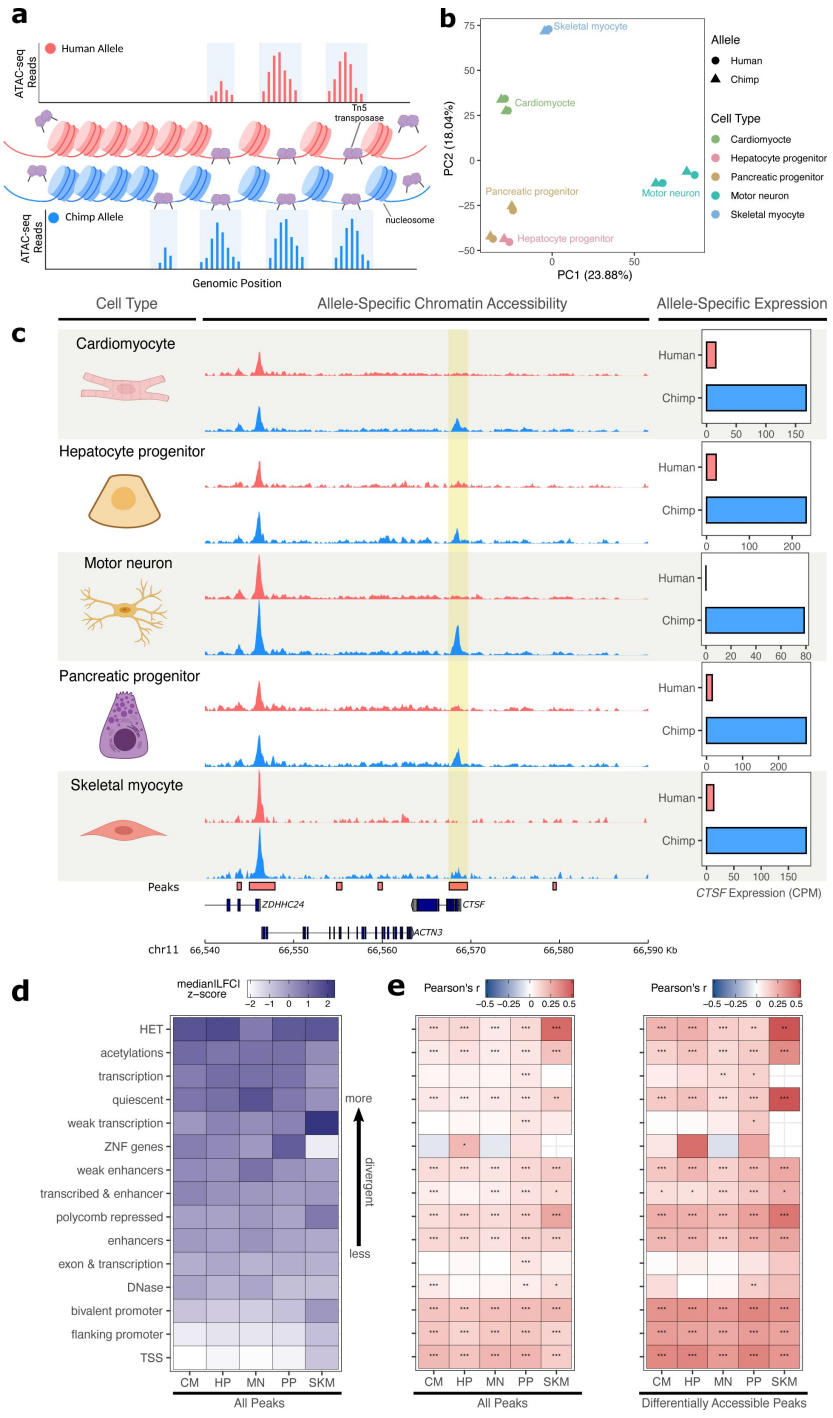


Figure 3

Allele-specific chromatin accessibility across diverse human-chimpanzee hybrid cell types.

a) Schematic outlining the ATAC-seq protocol. A hyperactive transposase cleaves accessible DNA and adds adapters enabling measurement of chromatin accessibility. **b)** PCA on normalized allelic counts from ATAC-seq. **c)** ASCA in the promoter of *CTSF,* and ASE for the *CTSF* gene. **d)** Differences in ASCA were quantified and plotted separately based on chromHMM annotation. The order is based on the median of z-score transformed absolute log fold-change between human and chimpanzee across all cell types, with higher z-scores indicating greater divergence in accessibility. **e)** Pearson correlation between ASE and ASCA for all cell types with all peaks (left) or only differentially accessible peaks (right, defined as peaks with nominal binomial p-value less than 0.05). Pearson's r values are shown in the heatmap and significance is indicated by asterisks where *** indicates $p < 0.005$, ** indicates $p < 0.01$, and * indicates $p < 0.05$. Zero asterisks (i.e. a blank box) indicates $p > 0.05$.

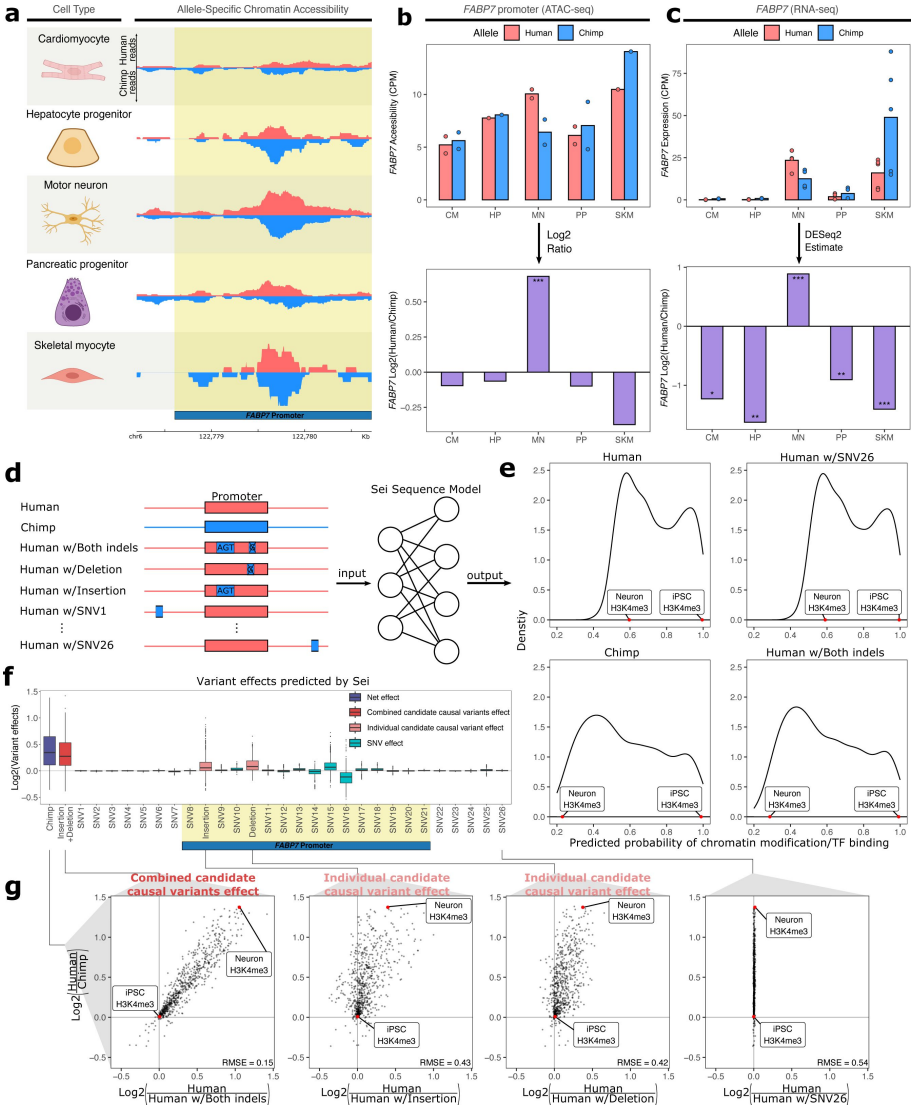


Figure 4

Motor neuron-specific human-biased ASE and ASCA for *FABP7* and the promoter of *FABP7*.

a) Allelic ATAC-seq tracks are shown in the peak containing the annotated *FABP7* promoter (highlighted in yellow). **b)** The top panel shows allelic CPM of the *FABP7* promoter across cell types and the bottom panel shows the log fold-change across cell types. **c)** The top panel shows allelic CPM of the *FABP7* gene across cell types and the bottom panel shows the log fold-change across cell types. **d)** Outline of the process for variant effect prediction with Sei for *FABP7*. All sequences input to Sei were centered at the *FABP7* promoter. The human sequence, chimpanzee sequence, and partially "chimpanized" human sequences (modified by systematically switching the human allele to the chimpanzee allele separately for each human-chimpanzee difference) were fed into Sei to predict the effects of these variants on chromatin state. **e)** Histogram of the probabilities of various chromatin states and transcription factor binding predicted by Sei was plotted for the human sequence, the chimpanzee sequence, and the human sequence with one human-chimpanzee difference swapped to match the chimpanzee sequence. The human sequence with SNV26 changed to the chimpanzee allele and the human sequence switched at both indels are shown as examples. A histone modification (H3K4me3) predicted in two cell types was labeled to illustrate how the predictions depend on both input sequence and cell type. **f)** Plot of the predicted effects of all single nucleotide differences and indels between the human and chimpanzee genomes in the Sei input window (the *FABP7* promoter is highlighted in yellow). **g)** Scatterplots showing the correlation of the effects of both indels (left panel), each individual indel (middle two panels), and a representative SNV (right panel) on Sei predictions with the difference in Sei predictions between the human and chimpanzee sequences. The root mean square error (RMSE) was computed and shown in each figure.

Fig. 5). *NTRK1* differential expression is similarly tissue-specific as it is strongly chimpanzee-biased in MN, but human-biased in CM and SKM (Supp **Fig. 5**). These results indicate that the regulatory landscape of these genes has undergone many complex *cis*-regulatory changes as the human and chimpanzee lineages have diverged.

To further investigate the relationship between tissue-specificity and ASE, we asked whether genes with variable expression across tissues are more likely to show ASE. Using a standard definition of cell type-specific genes—those with detectable expression in only one cell type in our study—we found that cell type-specific genes were typically enriched for ASE in the one cell type where they are expressed (Supp Fig. 6) (GTEx Consortium, 2017; Jain & Tuteja, 2019). However, other cell type-specific expression patterns such as uniquely low expression in a particular cell type may also indicate an important dosage-sensitive function in that cell type. We therefore focused on a broader definition of cell type-specificity in which genes that are differentially expressed between one cell type and all others in our study (FDR < 0.05 for each pairwise comparison) are considered cell type-specific for that cell type. We found that this more inclusive definition, which identified many more cell type-specific genes, showed an even more significant ASE enrichment than the narrower definition (**Fig. 2d**). This result is not sensitive to the choice of FDR cutoff (Supp Fig. 7) nor driven solely by a subgroup of highly expressed genes (Supp Fig. 8). This trend is also robust to separating samples into two groups and using one to define cell type-specific genes and the other to identify differentially expressed genes. This controls for spurious relationships that can result when the same data are used to define two different quantities which are then compared (Supp Fig. 9) (Fraser, 2019).

This enrichment (**Fig. 2d**) suggests that tissue-specific genes may have less constraint and/or more frequent positive selection on their expression. We reasoned that if the trend was solely driven by constraint, then controlling for constraint—even if imperfectly—would be expected to reduce the strength of the relationship. To investigate this, we binned genes by their variance in ASE across a large cohort of human samples which we have previously shown acts as a reasonable proxy for evolutionary constraint on gene expression (Castel et al., 2020; Starr et al., 2023). Across cell types, we generally observe significant enrichments in each bin and little difference in enrichment between bins, suggesting that differences in constraint on expression of cell type-specific vs. ubiquitously expressed genes are not solely responsible for our observations (Supp Fig. 10). Furthermore, we observe even stronger enrichments using an alternative constraint metric, the probability of haploinsufficiency score (pHI) likely due to the larger number of genes for which pHI can be calculated (Supp Fig. 11) (Collins et al., 2022). Overall, our analysis suggests that differences in constraint are unlikely to fully explain these trends, suggesting a potential role for positive selection.

Lineage-specific selection has acted on tissue-specific gene expression divergence

Next, we sought to use our RNA-seq data to identify instances of lineage-specific selection. In the absence of positive selection, one would expect that an approximately equal number of genes in a pathway would have human-biased vs. chimpanzee-biased ASE. Significant deviation from this expectation (as determined by the binomial test) rejects the null hypothesis of neutral evolution, instead providing evidence of lineage-specific selection on this pathway. Using our previously published modification of this test that incorporates a tissue-specific measure of constraint on gene expression, we detected several signals of lineage-specific selection, some of which were cell type-specific (Starr et al., 2023, Additional file 2). Notably, the four most significant enrichments were specific to motor neurons and cardiomyocytes and are highly relevant to those cell types (**Fig. 2e**; Additional file 2). In cardiomyocytes, the top pathway was “PPAR signaling pathway” which plays a key role in the regulation of heart morphology and lipid metabolism (Montaigne et al., 2021). For example, *NR1H3* (also known as *LXRA*) is strongly upregulated in human cardiomyocytes as well as all other cell types (Supp Fig. 12a). Furthermore, this upregulation

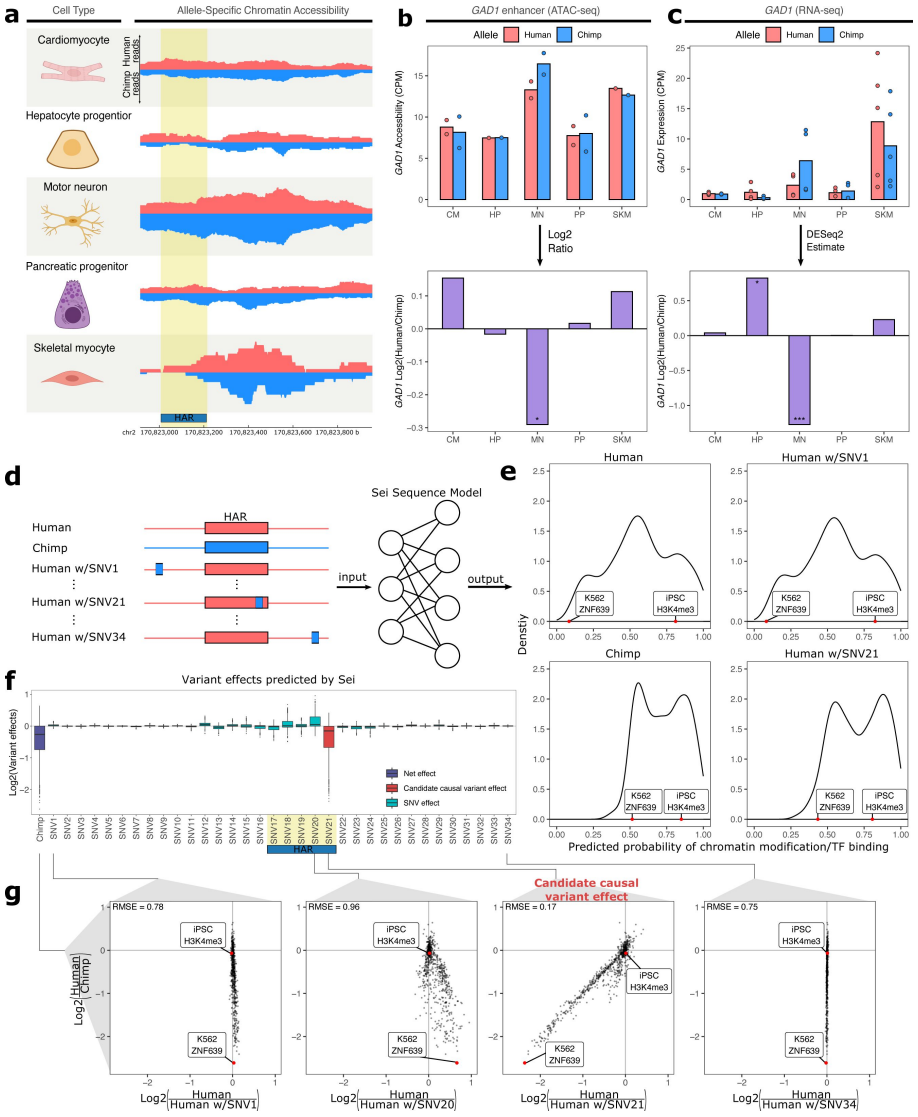


Figure 5

Motor neuron-specific chimpanzee-biased ASE for *GAD1* and ASCA for a HAR near the *GAD1* TSS.

a) Allelic ATAC-seq tracks are shown for the peak near the *GAD1* TSS that contains a HAR (highlighted in yellow). **b)** The top panel shows allelic CPM of the CRE near the *GAD1* TSS across cell types and the bottom panel shows the log fold-change across cell types. **c)** The top panel shows allelic CPM of the *GAD1* gene across cell types and the bottom panel shows the log fold-change across cell types. **d)** Outline of the process for variant effect prediction with Sei for *GAD1*. All input sequences to Sei were centered at the HAR. The human sequence, the chimpanzee sequence, and modified sequences with the human sequence altered at each substitution to match the chimpanzee sequence were fed into the Sei sequence model to predict the effects of these variants on the chromatin state. **e)** Histogram of the probabilities of various chromatin states and transcription factor binding predicted by Sei was plotted for the human sequence, the chimpanzee sequence, and two examples in which the human sequence with only one SNV "chimpanized" (human w/SNV) was input to Sei. The histogram of the probability of the sequence having a particular epigenomic annotation (predicted by Sei) was plotted for human, chimpanzee, human w/SNV1 changed to match the chimpanzee sequence, and human w/SNV21 changed to match the chimpanzee sequence. Two epigenomic annotations were labeled as examples that show the different values output by Sei with these two different sequence inputs. **f)** Plot of the predicted effects of all single nucleotide differences between human and chimpanzee in the Sei input window centered at the HAR (highlighted in yellow). Positions were switched to the chimpanzee allele individually. **g)** Scatterplots showing the correlation of the effects of four SNVs on Sei predictions with the difference in Sei predictions between the human and chimpanzee sequences. The root mean square error (RMSE) was computed and shown in each figure.

appears to have occurred in the human lineage based on data from non-hybrid cardiomyocytes as well as adult hearts (Supp Fig. 12b) (Blake et al., 2020; Pavlovic et al., 2018). Hybrid cells are essential in determining that the human-specific upregulation of *NR1H3* in cardiomyocytes has a strong genetic component as *NR1H3* expression is very responsive to diet and other environmental factors (Wang & Tontonoz, 2018).

In motor neurons, multiple categories showed a strong bias toward higher chimpanzee expression including "sodium ion transmembrane transport" and "syntaxin binding". The genes in these categories are of fundamental importance to the function of motor neurons as sodium ion transporters control excitability and syntaxin binding proteins control the release of neurotransmitters from synaptic vesicles (Brose et al., 2019; Meisler et al., 2021). Interestingly, several key genes in these sets appear to have human-chimpanzee differences in expression that extend beyond motor neurons to other neuronal types. For example, *SCN1B, SCN2B,* and *SYT2* have chimpanzee-biased ASE in our MN data. In contrast, these genes have been observed to have lower expression in human cortical glutamatergic neurons compared to neurons from chimpanzees and rhesus macaques (Supp Fig. 13) (Kozlenkov et al., 2020). We note that several other genes in these genes sets are not differentially expressed between humans and chimpanzees in this cortical neuron dataset, emphasizing the importance of studying individual neuron types (Kozlenkov et al., 2020). Overall, the strong bias in gene expression of sodium ion transporters and syntaxin binding proteins we observe suggests lineage-specific selection that may have altered the electrophysiological properties of human motor neurons.

Patterns of allele-specific chromatin accessibility reveal divergent *cis*-regulatory elements

While ASE provides insight into what gene expression changes might underlie phenotypic differences between humans and chimpanzees, in the absence of additional data it is very difficult to prioritize which specific mutations might cause expression divergence. To begin to fill this gap, we generated ATAC-seq data from five of the six cell types (all except RPE), with each cell type sequenced to an average depth of 184 million paired-end reads (Supp Fig. 14). ATAC-seq uses a hyperactive Tn5 transposase to cleave DNA that is not bound by nucleosomes to enrich for accessible chromatin (**Fig. 3a**), a hallmark of active *cis*-regulatory elements (CREs) (Buenrostro et al., 2013; Corces et al., 2017). We estimated ASCA for individual open chromatin peaks by mapping reads to both species' reference genomes, correcting for mapping bias, generating a unified list of peaks across all samples, and then counting reads supporting each allele in each peak (see Methods). After extensive filtering, we were left with 73,360 peaks across the five cell types with many fewer retained in SKM due to lower sequencing depth (Supp Fig. 15). As expected, most genes had only a few peaks assigned to them (Supp Fig. 16). Cell types clustered well using PC1 and PC2 of the ATAC-seq data, except for the HP sample which clustered closely with PP samples (**Fig. 3b**, Supp Fig. 17). However, performing PCA on just the HP and PP samples clearly separates the two cell types (Supp Fig. 18). Within each cell type, species differences were clearly separated by PC1 or PC2 (Supp Fig. 19). As an example of ASCA, the accessibility of the promoter of *CTSF* was strongly chimpanzee-biased, mirroring the chimpanzee-biased ASE for this gene (**Fig. 3c**).

As a first step in analyzing the ATAC-seq data, we intersected the peaks we identified with the genomic annotations of chromatin states. These fifteen categories, predicted across many tissues and cell types by the chromHMM model (Vu & Ernst, 2022), include terms such as "TSS" (transcription start site) and "enhancer". We then plotted the median of the absolute human-chimpanzee ASCA log fold-changes for each chromatin state and cell type (**Fig 3d**). The TSS and promoter annotations were the least divergent in their accessibility, whereas regions of heterochromatin were the most divergent (**Fig. 3d**). To explore the relationship between interspecies differences in ASCA and ASE, we assigned peaks to the nearest TSS and computed the Pearson correlation between the allelic log fold-change of chromatin accessibility and expression

within each cell type and chromatin state. As expected, TSS and promoter annotations showed the strongest correlation between ASCA and ASE, and correlations were stronger when including only differentially accessible peaks (**Fig. 3e**, Methods). Intriguingly however, ASCA of regions annotated as heterochromatin, polycomb repressed, or quiescent were as strongly correlated with ASE as elements identified as enhancers or DNase hypersensitivity sites (**Fig. 3e**). Notably this result is robust to how peaks and chromHMM annotations were intersected (Supp Fig. 20a), as well as to removal of all peaks even slightly overlapping TSS or promoter-related annotations (Supp Fig. 20b). It should be noted that chromHMM annotations such as heterochromatin do not imply that a region is constitutively heterochromatic, but instead reflect the most common chromatin state across a large compendium of cell types (Vu & Ernst, 2022). Indeed, the fact that we are focusing on ATAC-seq peaks indicates that the chromatin in these regions is accessible in at least one cell type in our study. This result suggests that CREs that are heterochromatic in some cell types may be more prone to large changes in accessibility during evolution (**Fig 3d**), with significant impacts on the cis-regulation of nearby genes (**Fig 3e**). Throughout, we refer to CREs assigned to the chromHMM annotation heterochromatin as “heterochromatin CREs”.

To further investigate the intriguing relationship between heterochromatin ASCA and ASE, we asked whether TSS proximity is an important factor. First, we removed all CREs annotated as promoters and recomputed correlations separately for CREs within 30 kilobases of a TSS (proximal CREs) and those greater than 30 kb away (distal CREs). Interestingly, differences in accessibility in proximal heterochromatin CREs have weak correlations with ASE compared to proximal enhancers (Supp Fig. 21a). However, differences in accessibility in distal heterochromatin CREs were roughly as strongly correlated with ASE as differences in accessibility in enhancer regions (Supp Fig. 21b).

Next, we partitioned genes into cell type-specifically and ubiquitously expressed and recomputed the correlations. While the results for ubiquitously expressed genes mirrored our initial finding of a relatively strong relationship between accessibility in heterochromatin CREs and gene expression, differences in accessibility in heterochromatin CREs were less correlated with ASE for cell type-specific genes than differences in accessibility in enhancers (Supp Fig. 22). Altogether, our analysis suggests that large changes in accessibility of distal heterochromatin CREs may be particularly important in the cis-regulatory evolution of more ubiquitously expressed genes.

Next, we investigated whether the analog of the relationship between cell type-specific gene expression and ASE (**Fig. 2d**) holds for chromatin accessibility. Since the number of called peaks is largely dependent on sequencing depth (Supp Fig. 14, 23a), we performed down-sampling to equalize power to detect peaks across cell types (Supp Fig. 23b, Methods). We then called peaks on the down-sampled data and identified peaks as cell type-specific if they were called as peaks in only one cell type. In agreement with the gene expression data, we observed that cell type-specific peaks are enriched for ASCA across all cell types and this enrichment generally holds when using varying log2 fold-change or p-value cutoffs (Supp Figs. 24-25). Analogous to our analysis of gene expression, we also applied a broader definition of cell type-specificity to the ATAC data, in which a peak was considered specific to a cell type if that peak had an absolute log2 fold-change greater than a chosen threshold (e.g. 0.5 or 1) across all pairwise comparisons with other cell types. We observe strong enrichment for ASCA in cell type-specific peaks using this definition except when using the most stringent cutoffs due to the very low number of peaks meeting these criteria (Supp Figs. 26-27, Methods). Notably, we observe the same enrichments when controlling for a recently published metric for constraint on non-coding elements that compares the observed vs. expected number of human polymorphisms (S. Chen et al., 2022), suggesting that differences in evolutionary constraint may not be solely responsible for the observed trends (Supp Fig. 28).

Finally, it is possible that CREs and genes that are less conserved will have more SNPs, and therefore more power to call ASCA and ASE, leading to systematically biased estimates. There is a weak positive correlation between the number of SNPs and the -log10(FDR) for ASE and a weak

negative or no correlation for ASCA (Supp Fig. 29). Similarly, we observe a weak relationship between the number of SNPs in CREs or genes and absolute log fold-change estimates (Supp Fig. 30). Although the relationship between the number of SNPs and ASE/ASCA is weak, we confirmed that cell type-specific genes and peaks are still strongly enriched for ASE and ASCA when stratifying by number of SNPs (Supp Fig. 31-32). Overall, our analysis suggests that the result that more cell type-specific genes and CREs are more evolutionarily diverged is robust to a variety of possible confounders.

We next explored the relationship between cell type-specific ASCA and ASE. To do this, we developed a novel metric called differential expression enrichment (dEE) to quantify how specific the log fold-change is to a particular cell type or tissue. Our method is based on expression enrichment (EE) (Yu et al., 2006), a metric that measures how specific gene expression is to a certain cell type/tissue. dEE estimates how cell type-specific ASE is for a gene (Supp Fig. 33, Methods) and, analogously, dCAE (differential chromatin accessibility enrichment) measures how cell type-specific ASCA is for a *cis*-regulatory element (Supp Fig. 34, Methods). dEE and dCAE are high in a cell type if there is a high absolute log fold-change in that cell type and much lower absolute log-fold changes or log fold-changes in the opposite direction in the other cell types. For example, dEE would be close to one for a gene in a cell type if the gene had strongly human-biased ASE in that cell type and very weakly human-biased or chimpanzee-biased ASE in the other cell types. On the other hand, if a gene did not have any strong allelic bias, that gene would have dEE close to zero. dEE is conceptually related to a metric we have previously introduced, diffASE (Combs et al., 2018; Hu et al., 2022; York et al., 2018), and generalizes diffASE to an arbitrary number of cell types and any assay that produces log fold-changes. Using these metrics, we identified 154 instances in which a gene with cell type-specific ASE (i.e., high dEE) had a peak with cell type-specific ASCA in the same cell type (i.e., high dCAE, see Additional file 3 for the full list). Of these, 95 showed ASCA and ASE in the same direction which is more than expected by chance (77 expected; $p < 0.005$, binomial test). These results suggest that tissue-specific *cis*-regulatory divergence in chromatin accessibility may often impact tissue-specific gene expression, though this divergence is neither necessary nor sufficient to do so.

Identifying candidate causal *cis*-regulatory variants by integrating ASE and ASCA across cell types

As high dEE and dCAE in a given cell type might be indicative of a causal link between the change in chromatin accessibility and the change in expression, we focused on the 95 peak-gene pairs with matching direction and used two different strategies to identify examples to investigate in detail. First, we prioritized genes known to play important roles in development. For example, we found that the promoter of *FABP7* has human-biased ASCA specifically in motor neurons (**Fig. 4a-b**) and that the *FABP7* gene has human-biased ASE in motor neurons (**Fig 4c**). *FABP7* is used as a marker of glial cells and neural progenitor cells (NPCs) and plays a key role in NPC proliferation and astrocyte function (Arai et al., 2005; De Rosa et al., 2012; Ebrahimi et al., 2016; Watanabe et al., 2007). Using previously published single-nucleus RNA-seq data from humans, chimpanzees, and rhesus macaques, we confirmed that *FABP7* shows a human-derived up-regulation in several neuronal subtypes but not glial cells (Supp Fig. 35) (Ma et al., 2022). To investigate the genetic basis of this cell type-specific divergence, we leveraged a machine learning model, Sei (K. M. Chen et al., 2022), to nominate potentially causal variants in the promoter of *FABP7* (see Methods). Sei is a deep neural network that takes DNA sequence as input and predicts the probability that the sequence has a particular epigenetic state in a variety of cell types and tissues (**Fig. 4d**) (K. M. Chen et al., 2022). While single-base substitutions differing between human and chimpanzee had only minor impacts on predicted *cis*-regulatory activity, "chimpanizing" the human sequence of the *FABP7* promoter at two small indels (by deleting one base at chr6: 122,779,291 and inserting three bases at chr6: 122,779,115) was sufficient to make the Sei predictions for the chimpanized human sequence closely match the predictions for the complete chimpanzee sequence (**Fig. 4e-g**). Making only one of these changes had substantial

but weaker effects in both cases, suggesting that both mutations might be functionally important (**Fig. 4f,g**). The 1-base insertion in the human lineage introduces a binding site for the neuronally expressed transcription factors GLIS2/3, suggesting a potential molecular mechanism (Calderari et al., 2018; Castro-Mondragon et al., 2022; Ke et al., 2015).

As another approach to ranking the 95 peaks, we searched for peaks containing human-chimpanzee sequence differences in otherwise highly conserved genomic positions, since these could reflect changes in selective pressure. Using PhyloP scores for 241 placental mammals (Sullivan et al., 2023) to assess conservation, one of the top-ranked peaks was a putative enhancer six kilobases away from the TSS of *GAD1*, which plays a key role in the synthesis of the neurotransmitter GABA (Feldblum et al., 1993). Notably, part of this peak has been classified as a human accelerated region (HAR) (Girskis et al., 2021; Hubisz & Pollard, 2014; Pollard et al., 2006)—a short sequence that is highly conserved in mammals yet contains an unusual number of human-specific mutations. Both the accessibility in the peak and *GAD1* expression are only chimpanzee-biased in motor neurons (**Fig. 4a-c**, Supp Fig. 36). Applying Sei to estimate the predicted effect of every variant in this region, we found that a single-nucleotide substitution within the HAR (chr2: 170,823,193) has by far the largest predicted cis-regulatory effect and most closely matches the differences in Sei predictions between the full human and chimpanzee haplotypes (**Fig. 4d-g**). Interestingly, this mutation is predicted to disrupt a binding site for several basic helix-loop-helix transcription factors that play essential roles in neuronal differentiation such as Ascl1 (Supp Fig. 37) (Castro-Mondragon et al., 2022; Mizuguchi et al., 2006; Yang et al., 2017).

As *GAD1* is only highly expressed in GABAergic neurons (and was therefore lowly expressed in the cell types studied here, Supp Fig. 38a), we investigated whether this reduced expression of human *GAD1* also occurs in cortical organoids which contain GABAergic neurons together with other cell types in which *GAD1* is not highly expressed. We analyzed our previously published data from human-chimpanzee hybrid cortical organoids (Agoglia et al., 2021) and found that the expression of *GAD1* from the chimpanzee allele spikes higher than that of the human allele around day 50 of hybrid cortical organoid differentiation before dropping in expression over time to match the human expression level (Supp Fig. 38b). Because ASE in the hybrid cells controls for any potential interspecies differences in differentiation kinetics or cell type composition, this difference must be the result of *cis*-regulatory divergence between humans and chimpanzees. This expression difference is also more pronounced in comparisons of human and chimpanzee parental cortical organoids, with a higher absolute log fold-change at day 50, day 100, and day 150, only returning to equal expression at day 200 (Supp Fig. 38c). While this could be due to differences in cell type proportion between human and chimpanzee organoids, it might also be due to a reinforcing *trans*-acting effect.

To test whether this difference in expression also occurs specifically during GABAergic neuron differentiation we examined *GAD1* expression in single cell RNA-seq data from human and chimpanzee cortical organoids (Kanton et al., 2019). Consistent with our cortical organoid results, we observed a peak in *GAD1* expression in less mature chimpanzee GABAergic neurons that is absent in the corresponding part of the trajectory in human neurons (Supp Fig. 39). Notably, a similar trend holds regardless of which GABAergic sub-trajectory (i.e., equivalent to GABAergic neurons from the caudal, lateral, or medial ganglionic eminences) is examined suggesting this difference is not unique to a particular type of GABAergic neuron (Supp Fig. 39). Finally, we examined the accessibility of the putative *GAD1* enhancer more closely. Consistent with a potential role for this enhancer in the spike in *GAD1* expression during development, the accessibility of this enhancer mirrors the expression of *GAD1* in human cortical and striatal organoids (Supp Fig. 40) with high accessibility between day 50 and day 100 before decreasing somewhat near day 150 (Trevino et al., 2020). Overall, our results demonstrate how the combination of RNA-seq, ATAC-seq, and machine learning models can nominate variants that may be responsible for cell type-specific changes in gene expression and chromatin accessibility.

Discussion

In this study, we quantified human-chimpanzee *cis*-regulatory divergence in gene expression and chromatin accessibility in six different cell types representing diverse developmental lineages. Across the thousands of genes with ASE, we found that most *cis*-regulatory divergence is specific to one or a few cell types. Furthermore, we found that divergent *cis*-regulation is linked to tissue-specificity, with tissue-specific genes being enriched for ASE and tissue-specific regulatory elements being enriched for ASCA. As this result was largely unchanged when stratifying by evolutionary constraint, our results suggest that changes in the expression of genes with more cell type-specific expression patterns may be less deleterious than changes in more broadly expressed genes, supporting the idea that cell-type specific divergence may be less pleiotropic (Wittkopp & Kalay, 2012). Overall, this suggests that broad changes in expression in cell type-specifically expressed genes may be an important substrate for evolution, but it remains unclear whether positive selection or lower constraint plays a larger role in driving the faster evolution of more cell type-specifically expressed genes. Future work will be required to more precisely quantify the relative roles of positive selection and evolutionary constraint in driving changes in gene expression.

We also identified several sets of genes evolving under lineage-specific selection that may have played a role in establishing unique facets of human physiology and behavior. Most interestingly, we found evidence for selection on sodium ion transporters and syntaxin binding proteins that may alter the electrophysiological properties of motor neurons, and potentially other types of neurons as well (Brose et al., 2019; Meisler et al., 2021). The complexity of the molecular machinery regulating neuronal excitability and synaptic vesicle release make it difficult to say what the effects of these gene expression changes are on the excitability of motor neurons without electrophysiology data from human and great ape neurons coupled with perturbation of candidate genes. However, given the divergence in locomotion and motor skills between humans and chimpanzees, one could speculate that these changes may have had some role in the evolution of motor control and learning in humans.

In this work, we developed two metrics—dEE and dCAE—to quantify the degree of cell type-specific differential expression and accessibility. These metrics are largely analogous to widely used metrics that quantify tissue- or cell type-specific expression level and applicable to any comparison of log fold-changes across conditions. They markedly improved our ability to identify matching cell type-specific ASE and ASCA and led to the identification of 95 peak-gene pairs that had highly cell type-specific concordant changes in accessibility and expression.

One such example is a human-derived increase in *FABP7* expression in several types of human neurons. As *FABP7* is not highly expressed in adult mouse neurons, the functional consequences of its higher expression in human neurons are difficult to predict (Yao et al., 2021). *FABP7* plays a role in the uptake of the fatty acid Docosahexaenoic Acid (DHA), an important component of neuronal membranes (Akbar et al., 2005; Choi et al., 2021). DHA promotes neuronal survival through phosphatidylserine accumulation, so it is possible that the human-specific *FABP7* expression increases neuronal DHA uptake leading to reduced apoptosis in human neurons during development and ultimately contributing to a larger number of neurons in humans (Akbar et al., 2005; Choi et al., 2021).

In addition, we identified a highly conserved developmentally dynamic enhancer near *GAD1* that may have partially lost activity in the human lineage resulting in a decrease in *GAD1* expression early in GABAergic neuron development. By integrating with the deep learning model Sei (K. M. Chen et al., 2022), we identified a variant that may account for the chimpanzee-biased ASCA in this region. Interestingly, the ASE of *GAD1* was coupled with a relatively small (though significant) magnitude of ASCA. This could potentially reflect divergence in transcription factor binding that leaves a “footprint” resulting in subtle ASCA (Vierstra et al., 2020). Overall, our data suggest that

this enhancer has lost activity in the human lineage, potentially altering the expression pattern of *GAD1* during neuronal development. *GAD1* is the rate-limiting enzyme for GABA synthesis so GABA levels are likely responsive to changes in *GAD1* expression (Feldblum et al., 1993). GABA release has complex context-specific effects on neurodevelopment, making it difficult to speculate as to what the phenotypic effects of reduced GABA synthesis during human neurodevelopment might be (Ben-Ari et al., 2012). However, the high conservation of this *cis*-regulatory element in placental mammals implies that its human-specific disruption is likely to have important neurodevelopmental effects. Careful perturbation of this enhancer and *GAD1* expression in mouse models will be required to explore this further.

In addition to following up on our findings on *GAD1* and *FABP7*, there are other exciting future directions for this work. First, additional bulk assays such as those that measure methylation, chromatin conformation, and translation rate could lead to a better understanding of what molecular features ultimately lead to cell type-specific changes in gene expression. Furthermore, the use of deep single cell profiling of hybrid lines derived from iPSCs from multiple individuals of each species during differentiation could enable the identification of many more highly context-specific changes in gene expression and chromatin accessibility such as the differences in *GAD1* we highlighted here. Finally, integration with data from massively parallel reporter assays and deep learning models will help us link specific variants to the molecular differences we identified in this study.

Overall, our study provides foundational data, insight, and computational tools that will improve our understanding of cell type-specific *cis*-regulatory evolution and the role it has played in the establishment of human-specific traits.

Methods

Generation of multiple human-chimpanzee hybrid cell types

We used two previously described human-chimpanzee hybrid iPS cell lines (hybrid1 and hybrid2, previously denoted Hy1-25 and Hy1-30 respectively) (Agoglia et al., 2021). Before differentiation, cells were routinely cultured on matrigel in mTeSR1 or mTeSR Plus (Stem Cell Technologies cat #85850 or cat #100-0276). Culture and *in vitro* differentiation of iPS cells into six cell types (motor neurons (MN), cardiomyocytes (CM), hepatocyte progenitors (HP), pancreatic progenitors (PP), skeletal myocytes (SKM), and retinal pigment epithelium (RPE)) was carried out by the Columbia Stem Cell Core Facility using published protocols (Burridge et al., 2014; Chal et al., 2016; Korytnikov & Nostro, 2016; Mallanna & Duncan, 2013; Maury et al., 2015; Sharma et al., 2019).

Preparation of RNA-seq libraries

All samples were cryopreserved in liquid nitrogen before RNA extraction (Milani et al., 2016). Cells were gently thawed and then washed with PBS and cell pellets were collected via centrifugation at 1,000 RPM for 5 minutes. Cell pellets were loosened by flicking the tube and an appropriate volume of Buffer RLT based on the cell count were added following the RNeasy Mini Kit (Qiagen, 74104) protocol. Total RNA extraction and on-column DNase digestion were performed using RNeasy Mini Kit (Qiagen, 74104) and RNase-Free DNase Set (Qiagen, 79254). RNA quality was assessed using the Agilent Bioanalyzer RNA Pico assay. Only samples with an RNA integrity number (RIN) greater than or equal to 7 were used to prepare cDNA libraries. All RNA-seq libraries except three motor neuron libraries were prepared using the TruSeq Stranded mRNA kit (Illumina, 20020594) and the TruSeq RNA CD Index Plate (Illumina, 20019792) from between 100 ng and 1 ug total RNA following the manufacturer's protocols. Due to low yield of total RNA, three motor neuron libraries (2 hybrid2 and 1 hybrid1) were prepared using Illumina Stranded mRNA Prep (Illumina, 20040532) and IDT for Illumina RNA UD Indexes Set A, Ligation (Illumina,

20040553). Notably, the four motor neuron libraries did not cluster by library preparation method. All libraries were normalized, pooled at an equimolar ratio using Qubit measurements, and sequenced on an Illumina HiSeq 4000 to generate 2x150bp paired-end reads.

Identification of confident human-chimpanzee SNVs

To identify a confident list of human-chimpanzee SNVs that could be used to quantify allele-specific expression and chromatin accessibility, we first downloaded hg38-panTro6 MAF files from UCSC and whole-genome sequencing data generated from the parental human and chimpanzee iPS cells (in the form of bam files aligned to hg38 and panTro5, generously provided by the Gilad lab). We first converted them back to fastq files and then mapped reads to panTro6 and hg38 (we mapped both human and chimpanzee to both reference genomes) using bowtie2 with the flags —very-sensitive-local - p 16 (Langmead & Salzberg, 2012). We then used a modified version of our previous approach to filter out SNVs that could not be confidently identified as homozygous in the human and chimpanzee parental lines (Agoglia et al., 2021). Briefly, we extracted SNVs and indels from both human and chimpanzee MAF files, counted reads in the WGS data that supported the human, chimpanzee, or an alternative base at that position, then filtered out any SNVs with < 2 reads or < 90% of reads supporting that species' base. We then reformatted files, merged with indels for use in Hornet, and generated a modified bed file of SNVs that includes the human and chimpanzee base at the SNV position (van de Geijn et al., 2015).

Generation of allele-specific count tables

An allele-specific expression pipeline adapted from Agoglia et al. and our updated high-confidence SNV list was used. The whole pipeline was carried out twice independently using hg38 or panTro6 as the reference genome. This approach was taken to eliminate genes showing strong mapping bias, defined here as genes with an absolute difference in log2 fold-change between the panTro6-referenced and hg38-referenced runs greater than one. All sequencing reads were trimmed with SeqPrep (adapters specified by the manufacturer for the different library preparation kits) and mapped using STAR with two passes and the following parameters: --outFilterMultimapNmax 1 (Dobin et al., 2013; John St. John, n.d.). Uniquely aligned reads were deduplicated with Picard and Hornet (an implementation of WASP which first removes reads overlapping indels) was used to correct for mapping bias (Broad Institute, n.d.; van de Geijn et al., 2015). Reads were assigned to either the human allele if they contained one or more human-chimpanzee single nucleotide differences that matched the human sequence and zero positions that matched the chimpanzee sequence (and vice versa for assigning reads to the chimpanzee allele) and counted per gene as previously described (Agoglia et al., 2021).

Detection of aneuploidy on chromosome 20 and slight chimpanzee parental contamination in PP hybrid2 samples

In our quality control process, we plotted the log2 fold-change for each gene along every chromosome and inspected the results. This revealed a clear bias toward the human allele on a part of chromosome 20 for hybrid2 samples, suggesting chromosome 20 aneuploidy which was also reported by Agoglia et al (Agoglia et al., 2021). As a result, we excluded chromosome 20 from all downstream analyses. In addition, we found that PP hybrid2 samples had a slight bias toward the chimpanzee allele across every chromosome which was most likely due to a small fraction of contaminating chimpanzee cells in these samples. Rather than removing these samples, we normalized the allele-specific count tables by subtracting a small number of reads from the chimpanzee allele counts calculated based on the biased ratio summarized from genome wide human and chimpanzee allele counts, to force a global log fold-change (across all autosomes except chromosome 20) of zero between the human and chimpanzee alleles. We applied this normalization to all other samples as well. To evaluate the success of this strategy, we simulated the effects of chimpanzee iPSC contamination and our subsequent correction. Specifically, using hybrid and chimpanzee parental iPSC RNA-seq from Agoglia et al (Agoglia et al., 2021) from the

same iPS cell lines as used in this study, we simulated chimpanzee iPSC contamination by mixing chimpanzee RNA-seq data into hybrid2 data to reach a similar degree of chimpanzee bias level to that observed in the PP hybrid2 samples. We then identified genes showing ASE (see "Identifying genes with ASE") using the counts from the original hybrid samples, simulated contaminated samples, and corrected simulated contaminated samples and compared the outputs (Supp **Fig. 4**).

PCA and hierarchical clustering

Allelic counts were normalized by DESeq2 rlog and principal components analysis (PCA) was performed on rlog normalized allelic counts with default centering and scaling (Love et al., 2014). The top 1,000 variable genes with the highest variance of normalized allelic counts across all cell types were used to compute Euclidean distance matrices. The R package pheatmap was used to do hierarchical clustering and heatmap plotting.

Identifying genes with ASE

DESeq2 was used to measure allele-specific expression (ASE) in each cell type (Love et al., 2014). All reads from chromosome 20 were removed (as mentioned above). Two replicates per hybrid line per cell type (plus one additional replicate for SKM hybrid2 for a total of 3 samples) were used by DEseq2 with model ~hybLine+Species to measure differential expression level. A likelihood ratio test (test="LRT", betaPrior=FALSE) was used to compute p-values. P-values were then false discovery rate adjusted using an implementation of the Benjamini-Hochberg correction in the R package qvalue (Benjamini & Hochberg, 1995; Storey Lab, n.d.). Log fold-changes were shrunk as recommended by the DESeq2 pipeline (Zhu et al., 2019). Differentially expressed genes were defined as those with FDR < 0.05 when aligned to hg38 and panTro6 as well as an absolute difference in log fold-change <= 1 when comparing the results from the two alignments.

Identifying cell type-specifically expressed genes

For the more traditional definition of cell type-specific genes, we required transcripts per million (TPM) < 1 for a gene in every cell type except one. In the cell type with TPM > 1, we varied how highly expressed the gene had to be in that cell type (again using a TPM cutoff, varying between one and five) to consider that gene to be specific to that cell type. A similar process to the one described in "Identifying differentially expressed genes" was used to identify cell type-specific genes based on the broader definition described in the main text. Rather than using allelic counts, total counts for each sample (i.e. all uniquely mapping deduplicated reads regardless of their allelic origin) were computed by summing all allelic and non-allelic counts. These counts were inputted to DESeq2 and the expression of each gene was compared pairwise between all cell types (Love et al., 2014). Genes were defined as cell type-specifically expressed in a cell type only if all pairwise comparisons between that cell type and other cell types resulted in an FDR < 0.05 using both hg38 and panTro6 aligned counts. Due to the markedly lower number of differentially expressed genes identified in SKM, results were computed both including and excluding SKM. An analogous procedure was used to identify more broadly defined cell type-specific peaks in the down-sampled ATAC-seq dataset. Peaks were defined as specific to a cell type if the absolute log fold-change was greater than 0.5 across all pairwise comparisons with the other cell types. We also tested an absolute log fold-change threshold of 1 to ensure that our results were not sensitive to the choice of cutoff.

Enrichment test for genes with cell type-specific expression patterns and genes showing ASE

Odds ratios were calculated using the unconditional maximum likelihood estimate implemented in the R package epitools function oddsratio.wald(), and 95% confidence intervals and p-values were calculated using the normal approximation. A directly analogous procedure was performed

to test for enrichment of peaks with ASCA and cell type-specific peaks.

Enrichment test stratified by expression level or evolutionary constraint

Enrichment tests were carried out as in 'Enrichment test of cell type-specific expression patterns and genes showing ASE' except that genes were split into five equal size bins depending on which factors were used to stratify genes, and tests were done in each bin. When stratifying by expression level, genes were ordered in ascending order based on expression level (TPM) and then split into five equally sized bins where genes in the 0-20% bin are the most lowly expressed genes and genes in the 80-100% bin are the most highly expressed genes. When stratifying by constraint metrics such as ASE variance (see "Identification of lineage-specific selection on gene expression" for details on how ASE variance was computed) or pHI, genes were ordered in ascending order based on ASE variance values and then split into five equal size bins where the 0-20% bin contains genes with the lowest ASE variance (i.e. most evolutionarily constrained) and the 80-100% bin contains genes with highest ASE variance (i.e. least evolutionarily constrained). To stratify the ATAC data by constraint, we used the "QCed genomic constraint by 1kb regions" computed by the gnomAD consortium (S. Chen et al., 2022). The gnomAD consortium computed the ratio of the observed and expected number of human polymorphisms in 1 kb regions tiling the human genome and then converted this metric to a z-score (see S. Chen et al., 2022 for additional details). We further removed any regions that overlapped protein-coding exons from the human gtf file using bedtools subtract (Quinlan & Hall, 2010). If a peak overlapped two or more 1 kilobase windows, it was assigned to the window with the highest constraint, mirroring the procedure used by the gnomAD consortium to assign peaks to ENCODE regulatory elements (S. Chen et al., 2022). Once the peaks were ranked by this metric, a procedure identical to that for the gene expression constraint metrics outlined above was performed.

Identification of lineage-specific selection on gene expression

We developed a modified version of our previously published pipeline, which uses ASE values from many individuals of a single species to estimate cis-regulatory constraint of each gene (Starr et al., 2023). We restricted these ASE values to GTEx samples from the tissue(s) of origin for each cell type (hepatocytes with liver, skeletal muscle with skeletal muscle, cardiomyocytes with the left ventricle of the heart, and pancreatic progenitors with the pancreas). As RPE and MN did not have clear matching tissues (e.g. GTEx does not include data from the eye), we compared RPE to all GTEx samples and MN to all brain and peripheral nerve samples. We then used the Mann-Whitney U test to compare the human population ASE distribution to the human-chimpanzee ASE distribution as previously described (Starr et al., 2023). We employed the previously described signed ranking by Mann-Whitney p-value that incorporates whether a gene has human or chimpanzee-biased ASE with GSEAPY and the binomial test to identify instances of lineage-specific selection (Starr et al., 2023). Positive selection on a gene set is only inferred if there is statistically significant human-or chimpanzee-biased ASE in that gene set (using an FDR-corrected p-value from the binomial test). Due to the focus on tissue-specificity, we did not filter redundant gene sets with GSEAPY FDR < 0.25 in multiple cell types (Starr et al., 2023; Subramanian et al., 2005). To compute ASE variance, we used the same tissue-of-origin-matched data from GTEx and computed the variance of the ASE ratios after filtering out samples with fewer than ten counts from the reference or alternate allele (see Starr et al., 2023 for additional details). For example, if a sample had 11 reference counts and 2 alternate counts for a gene, that sample would be excluded for that gene.

Preparation of ATAC-seq libraries

We used the OmniATAC protocol with the only modification being the use of 25,000 cells instead of 50,000 since the fused iPS cells are tetraploid (Corces et al., 2017). All samples for ATAC-seq prep were from the same vials used in RNA-seq library preparation except for the motor neuron

libraries due to the low yield of total RNA extracted from motor neurons. After library preparation and running samples on a Bioanalyzer, we noticed a considerable number of fragments greater than 1000 bases in length. To reduce these fragments, we size selected with Ampure beads using the protocol from the Kaestner lab available here: *https://www.med.upenn.edu/kaestnerlab/assets/user-content/documents/ATAC-seq-Protocol-(Omni)-Kaestner-Lab.pdf*. After size selection and rerunning on the Bioanalzyer, we pooled the libraries together and sequenced them to compute quality control metrics. We used the R package ChrAccR to compute TSS enrichment scores (Mueller, Fabian, n.d.). We pooled all libraries with TSS enrichment score greater than 3.5. This resulted in 2 CM libraries, 2 MN libraries, 2 PP libraries, 1 SKM library, and 1 HP library. After pooling, libraries were sequenced on an Illumina Hiseq 4000 to produce 2x150 paired-end reads.

Mapping the ATAC-seq data

We trimmed reads using SeqPrep and then mapped them to the hg38 and panTro6 reference genomes with bowtie2 in paired-end mode (John St. John, n.d.; Langmead & Salzberg, 2012). The following parameters were used: -X 2000 --very-sensitive-local -p 16. After mapping, duplicates were removed via Picard MarkDuplicates (Broad Institute, n.d.). We then removed multi-mapping reads with the command samtools view -b -q 10 (Li et al., 2009). Due to the format of bowtie2's output, running Hornet on all reads at once was excessively RAM intensive. Therefore, we split the bam files by chromosome and ran Hornet on each of the chromosomes separately. We used the files of SNVs and indels generated as described above as input to Hornet. After Hornet finished running, we used samtools merge to merge all autosomes and sex chromosomes (we excluded the mitochondrial genome) to create a final bam file for downstream analysis (Li et al., 2009).

Peak calling and filtering

As only one replicate was available for SKM and HP, we generated two pseudo-replicates by randomly assigning reads to one of two files using Picard SplitSamByNumberOfReads (Broad Institute, n.d.). We then called peaks on each file separately, as well as a merged file containing all the reads from a particular cell type. For example, for MN, both replicates were pooled and peaks were called on that file as well as the two replicates separately. Before peak calling, all bam files were converted to bed files. We called peaks using MACS2 callpeak with the following arguments: -f BED -p 0.01 --nomodel --shift 75--extsize 150 -B --SPMR --keep-dup all --call-summits (Y. Zhang et al., 2008, p. 2). We called peaks on both the chimpanzee-referenced and human-referenced bam files. After peak calling, we sought to filter peaks using a modified version of the ENCODE pipeline designed to eliminate peaks that lack a one-to-one ortholog between humans and chimpanzees. The following pipeline was run on each cell type separately. We first filtered peaks that were not called in both replicates as well as the pooled file using code from the ENCODE pipeline based on bedtools and awk (Quinlan & Hall, 2010). We then used a custom Python script to merge overlapping peaks and used UCSC LiftOver to lift the peaks from hg38 to the panTro6 and back to hg38 as well as from panTro6 to hg38 (Kuhn et al., 2013). We then used bedtools to intersect the resulting human referenced files and filtered out any peaks that did not have at least 25% overlap with a peak in the other file (Quinlan & Hall, 2010). After filtering out peaks overlapping ENCODE blacklisted regions and merging overlapping peaks again, we lifted the file that was originally chimpanzee-referenced back to the chimpanzee genome (Amemiya et al., 2019). Finally, we removed human-referenced peaks if their chimpanzee-referenced counterpart failed to lift over (Kuhn et al., 2013). As only a relatively small number of peaks failed to lift over (e.g. because the region was split in the new genome), any peaks that failed to lift over were excluded.

Annotating the peak lists

To annotate the peaks, we used the list of TSSs defined by Horlbeck et al. to annotate peaks (Horlbeck et al., 2016). We lifted over each TSS to hg38, expanded 1000 bases on either side of the midpoint of each TSS to generate promoters, and merged any promoters that overlapped while retaining all unique gene names associated with the promoter (Kuhn et al., 2013). We then used

reciprocal LiftOver with panTro6 to filter out non-orthologous promoters and used bedtools intersect to link peaks to promoters and expanded the peak to include the entirety of the promoter if necessary (Quinlan & Hall, 2010). Through this process, we also compiled a list of non-promoter CREs (sometimes labeled as enhancers as enhancers are thought to be the most common type of CRE). We used bedtools closest to link these non-promoter CREs to the two closest protein coding genes (Quinlan & Hall, 2010). Notably, the gene naming conventions differ for the Horlbeck et al. TSS list and the GTF file used for RNA-seq processing. We altered all gene names in peaks to match those found in the GTF file. In some cases, the gene no longer existed in the updated hg38 GTF in which case the gene name was replaced with NAN.

Generating a unified peak list

We next merged our cell type-specific peak list across all five cell types to create a unified peak list. To do this, we iteratively intersected all the peaks with bedtools and then merged any overlapping peaks (Quinlan & Hall, 2010). Finally, we added back any peaks that did not intersect a peak found in any other cell types. We then took the chimpanzee and human-referenced versions of these peak lists and ran them through the LiftOver-based non-homologous peak filtering pipeline described above to generate a final file of all identified peaks as well as which cell type (s) they were called in (Kuhn et al., 2013). Then, we reran the annotation pipeline described in 'Annotating the peak lists' on this new set of peaks. In total, this process resulted in 251,669 ATAC peaks.

Counting reads in peaks and further peak filtering

First, we split the bam files into reads that we could confidently assign to the chimpanzee genome and reads we could assign to the human genome. We used our bed file of high-confidence SNVs and required at least one SNV matching the human genome as well as no SNVs matching the chimpanzee genome for a read to be assigned to human (and vice versa for chimp). We then used a custom Python script to reformat the peak list bed files as GTF files and used HTSeq to count reads in peaks using the following parameters: -s no -m union -r pos (Anders et al., 2015). We only kept peaks if they had a mean read count across replicates within a cell type of at least 25 from either allele. For example, if a peak has an average of 27 reads from the human allele and an average of 10 reads from the chimpanzee allele in MN, that peak would be kept in MN. On the other hand, if the same peak had an average of 24 reads from the human allele and an average of 10 reads from the chimpanzee allele in CM, that peak would be discarded for CM.

We next filtered the reads to remove peaks that might be differentially accessible but show evidence of mapping bias or do not agree between replicates. To do this, we removed any peaks with an absolute log2 fold-change greater than one in one replicate but with a fold-change of any magnitude in the opposite direction in the other. This was not done for SKM or HP as we had only one replicate. We then removed any peaks that had a log fold-change in opposite directions with an absolute difference greater than 1 in at least one replicate when comparing the human-referenced and chimpanzee-referenced counts. Finally, as described in section "Detection of aneuploidy on chromosome 20 and slight chimpanzee parental contamination in PP hybrid2 samples' for RNA-seq data analysis, we removed any peaks on chr20 and took this as our final list of peaks for downstream analyses. Allelic counts were normalized as described in the RNA-seq data analysis. We tested for allele-specific chromatin accessibility (ASCA) using the binomial test applied to the normalized allelic counts (summed by species within a cell type). We considered any peaks with a binomial p-value less than 0.05 to be nominally differentially accessible. After this filtering, we retained 76,360 peaks for additional analysis. Using the set of peaks that passed filtering for each cell type, we plotted the total number of promoter and non-promoter peaks that passed filtering in each cell type as well as the number of peaks per gene.

Down-sampling to identify cell type-specific ATAC-seq peaks

As the number of peaks detected by ATAC-seq is generally a function of read depth and our read depth varied widely across cell types, we restricted to one replicate (always hybrid1 if two replicates were available) and down-sampled reads to match the SKM sample with lowest sequencing depth. We then called peaks for cell types with a single ATAC replicate as described above.

Allelic chromatin accessibility tracks

Allelic bam files with reads originating from the human allele and the chimpanzee allele (respectively) were obtained as described in 'Counting reads in peaks and further peak filtering'. Two replicates in CM, MN, and PP were pooled by cell type. Bam files were converted into bigWig files using the python package deepTools bamCoverage with options: --binSize 1 --normalizeUsing CPM --effectiveGenomeSize 2862010578 --ignoreForNormalization chr20 --extendReads (Ramírez et al., 2014). Tracks were visualized and plotted using the python package pyGenomeTracks (Lopez-Delisle et al., 2021). When comparing human and chimpanzee log fold-change track differences in each cell type, deepTools bigwigCompare was used to compare between human bigWig and chimpanzee bigWig with options: -- pseudocount 1 --skipZeroOverZero --operation log2 -bs 1 (Ramírez et al., 2014).

ChromHMM annotation and correlation with ASE

A universal chromHMM annotation was obtained for each peak based on overlap with any of the 15 categories in chromHMM (excluding the blacklist category, for which peaks had already been removed) (Ernst & Kellis, 2017; Vu & Ernst, 2022). Divergence was measured as the z-score transformed median of the absolute log2 fold-change of human and chimpanzee normalized counts in each peak. Each peak was assigned to the closest gene and then Pearson correlation was computed between the chromatin accessibility log2 fold-change and the expression log2 fold-change for each peak and its nearest gene. Pearson correlation was computed only on categories including at least 15 peaks. When showing results for differentially accessible peaks, only peaks with binomial p-values less than 0.05 were kept and used in computing the Pearson correlation. When assigning a unique chromHMM to each peak, the chromHMM category that covered the largest portion of each peak was used. When filtering out promoter-related annotations, peaks covering any promoter-related chromHMM categories ("TSS", "flanking promoter" and "bivalent promoter") were filtered out and the analysis described above was repeated. Of the 76,630 peaks retained, 76,221 overlapped at least one chromHMM annotation. We recomputed these correlations (with promoters excluded) after splitting peaks into those less than 30 kilobases from the nearest TSS and peaks greater than or equal to 30 kilobases from the nearest TSS. We also recomputed the correlations for cell type-specifically and ubiquitously expressed genes separately.

Testing for influence of number of SNVs on p-values and log fold-change

To assign SNVs to genes, we converted all exons in the human gtf file to bed format and intersected the exons with our list of confident human-chimpanzee SNVs. We then used a custom python script to count the number of unique SNVs assigned to each gene and plotted the relationship between DESeq2 p-value and log fold-change estimate. For the ATAC peaks, we performed an identical procedure using the bed file of peaks rather than a bed file of exons. We then plotted the number of SNVs in a peak with the binomial test p-value and raw log fold-change estimate.

Computation of differential expression enrichment (dEE) and differential chromatin accessibility enrichment (dCAE)

For each target cell type, taking CM as an example, the log2 fold-change for gene A was fixed as target log2 fold-change, and the log2 fold-changes for gene A in the remaining cell types with an opposing sign (compared to the target log2 fold-change) were set to zero. Then, the dEE value was calculated as the proportion of the target log2 fold-change in the sum of the zeroed log fold-changes across all cell types. For example, the dEE for gene A in CM would be abs (target LFC)/sum (abs(LFC after zeroing)). dEE ranges from zero to one and low dEE value indicates differential expression with similar magnitude and direction across cell types, and/or the gene does not have any strong allelic bias, whereas a high dEE value indicates that this gene is only strongly differentially expressed (with the sign the log2 fold-change has in that cell type) in a particular cell type. dCAE uses the same procedure as dEE except the table is populated with the log2 fold-changes derived from chromatin accessibility measurements. dEE and dCAE are sensitive to the inclusion or exclusion of cell types (by definition), so we excluded RPE when integrative analysis combining results from dEE and dCAE was performed (to match the cell types for which dCAE could be computed, Supp Fig. 33b). After restricting to genes defined as having significant ASE or significant ASCA, we defined genes with dEE >= 0.75 in a particular cell type as showing cell type-specific ASE and peaks with dCAE >= 0.75 in one cell type as showing cell type-specific ASCA. We used bedtools intersect to intersect the peaks with our list of human-chimpanzee single nucleotide differences and the 241-way placental mammal PhyloP scores (Quinlan & Hall, 2010; Sullivan et al., 2023). We also checked whether peaks that contained human-chimpanzee differences in sites with high PhyloP scores were in the list of HARs described in Girskis et al (Girskis et al., 2021).

Predicting regulatory activity with single-variant resolution

We used sequences in fasta format as input to the deep neural network model Sei (K. M. Chen et al., 2022). Sei requires a 4096 base pair input sequence, so we put the center of our region of interest at the center of the input window and expanded equally on either side to contain 4096 base pairs. The human sequence was retrieved from hg38 and the corresponding chimpanzee sequence was retrieved from panTro6. The effect size when comparing the probabilities of each sequence having a particular chromatin state was computed as the log of the human sequence probability divided by the chimpanzee sequence probability. Only annotations for which either the chimpanzee sequence or the human sequence had a probability value greater than or equal to 0.5 were kept for downstream analysis. All SNVs between human and chimpanzee in this input window were identified and ordered based on coordinates. For each SNV position, the human sequence was changed to the chimpanzee allele at that position to generate a new sequence that was input to Sei. The log2 fold-change for each chromatin annotation was computed for each input sequence as described above and used as a measure of the effect of this change on the sequence. Similarly, an indel can be introduced to modify the human sequence and input to Sei. With indels, the center of the regions of interest (promoter or HAR) were always at the center of the input window and the start or end of the sequence inputted to Sei could possibly lose or gain base pairs. However, we found that for the small indels shown here this had essentially no effect on the Sei output.

Processing of publicly available datasets

The data from Blake et al. and Pavlovic et al. were processed as previously described (Blake et al., 2020; Pavlovic et al., 2018; Starr et al., 2023). For the Pavlovic et al. data, log2 fold-changes were computed in DESeq2 with the scaled proportion of cardiomyocytes present in each sample (available in the supplemental materials of Pavlovic et al.), sex, and whether cardiomyocytes were treated with T3 as covariates (i.e. using the model ~sex+scaled_proportion_cardiomyocytes+T3_Treatment+species) (Pavlovic et al., 2018). No

covariates were included for Blake et al. as they had little impact on the data (Blake et al., 2020). The log2 fold-changes and FDR corrected p-values were directly downloaded from the supplemental materials of Kozlenkov et al. (Kozlenkov et al., 2020).

The processed data from Ma et al. (Ma et al., 2022) were downloaded from *http://resources.sestanlab.org/PFC/*. We pseudobulked the data by cell type by summing counts within each individual. We then separately input each pairwise comparison of two species (human to chimpanzee or human to rhesus macaque) into DESeq2 with no covariates to test for differential expression and compute log2 fold-changes.

The counts tables from Kanton et al.[12] were downloaded from *https://www.ebi.ac.uk/biostudies/arrayexpress/studies/E-MTAB-7552* and processed with SCANPY (Kanton et al., 2019, p. 29409532; Wolf et al., 2018). The data were filtered by removing cells with n_genes_by_counts > 2500 and >5% mitochondrial reads. We also removed cells with fewer than 200 unique genes and genes that had non-zero counts in fewer than 3 cells. After filtering, any chimpanzee cells not falling in the category (defined by Kanton et al. (Kanton et al., 2019)) "ventral forebrain progenitors and neurons" were eliminated and human cells not in the categories "ventral progenitors and neurons 1", "ventral progenitors and neurons 2", or "ventral progenitors and neurons 3" were similarly eliminated. We then merged the two counts tables, normalized/logarithmized the counts, computed PCA, used harmony (Korsunsky et al., 2019) to integrate cells from different species (human and chimpanzee), and found nearest neighbors with the harmonized principal components. We then ran Leiden clustering with resolution = 0.5 to identify 7 subclusters (one of which appeared to be a technical artifact with very low counts that was removed) (Traag et al., 2019).

We identified cell types and lineages using canonical marker genes (*MKI67* and *HES5* for progenitors, *NKX2-1* and *LHX6* for the medial ganglionic eminence or MGE, *MEIS2* and *ZFHX3* for the lateral ganglionic eminence or LGE, and *SCGN* and *NR2F1* for the caudal ganglionic eminence or CGE) (Su-Feher et al., 2022). We then used the implementation of PAGA in SCANPY to compute pseudotime using the first cell in the progenitor subcluster as the root (Wolf et al., 2019). We binned cells into five equal bins along pseudotime and compared the expression of cells with non-zero counts for *GAD1* in each pseudotime bin. Within each bin, we used a Wilcoxon test to test for higher expression of *GAD1* in chimpanzee cells compared to human cells. We repeated the pseudotime analysis, binning, and comparing of *GAD1* gene expression for each subtrajectory (MGE, LGE, and CGE).

Description of Additional file 1

This file contains an extended evaluation of the success of the differentiations in generating the desired cell type and other cell types likely to be present in each sample.

Description of Additional file 2

This file contains the results of running the test for selection described by Starr et al (2023) on each cell type.

Description of Additional file 3: This file contains the peak-gene pairs that had concordant high dCAE and high dEE in each cell type.

Declarations

The authors have nothing to declare.

Ethics approval and consent to participate:

Not relevant to our study.

Consent for publication

Not relevant to our study.

Availability of data and materials

Raw and processed data generated by this study are publicly available through the Gene Expression Omnibus under accession GSE232949: *https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE232949*. The snRNAseq data from Ma et al. are available here: *http://resources.sestanlab.org/PFC/*. The scRNAseq data from Kanton et al. are available here: *https://www.ebi.ac.uk/biostudies/arrayexpress/studies/E-MTAB-7552*. The bulk RNA-seq data from Blake et al. and Pavlovic et al. are available here: *https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE112356* and here: *https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE110471* respectively. The log fold-changes and associated statistics were used directly from the supplemental materials of Kozlenkov et al. The human-chimp pairwise alignment used to identify SNVs and indels is available here: *https://hgdownload.soe.ucsc.edu/goldenPath/hg38/vsPanTro6/*. The human and chimpanzee genomesused are available here: *https://hgdownload.soe.ucsc.edu/goldenPath/hg38/bigZips/* and here: *https://hgdownload.soe.ucsc.edu/goldenPath/panTro6/bigZips/* respectively. The 241-way PhyloP scores were downloaded from here: *https://hgdownload.soe.ucsc.edu/goldenPath/hg38/cactus241way/*. The "QCed genomic constraint by 1kb regions" from the gnomAD consortium are available here: *https://gnomad.broadinstitute.org/downloads#v3*. All scripts for performing analyses and making figures in this manuscript is publicly available at *https://github.com/banwang27/multi-celltypes*.

Competing interests: The authors have no competing interests to declare.

Funding: Funding for this work came from NIH R01HG012285. A.L.S was supported by an NDSEG fellowship.

Author contributions: HBF conceived of the study. ALS and BW performed all analysis, visualization, validation, and writing of software. ALS wrote the manuscript with input from BW and HBF. BW made all figures with input from ALS and HBF. All authors approved the publication of the manuscript.

Acknowledgements

The authors wish to acknowledge the Columbia Stem Cell Core for their hard work in differentiating these cells. We also acknowledge members of the Fraser lab past and present for helpful discussion. **Figure 1a**, 2a, and 3a were created with *https://www.biorender.com/*.

References

Agoglia R. M., Sun D., Birey F., Yoon S.-J., Miura Y., Sabatini K., Pașca S. P., Fraser H.B. (2021) **Primate cell fusion disentangles gene regulatory divergence in neurodevelopment** *Nature* **592**:421–427 https://doi.org/10.1038/s41586-021-03343-3

Akbar M., Calderon F., Wen Z., Kim H.-Y (2005) **Docosahexaenoic acid: A positive modulator of Akt signaling in neuronal survival** *Proceedings of the National Academy of Sciences* **102**:10858–10863 https://doi.org/10.1073/pnas.0502903102

Amemiya H. M., Kundaje A., Boyle A. P (2019) **The ENCODE Blacklist: Identification of Problematic Regions of the Genome** *Scientific Reports* **9** https://doi.org/10.1038/s41598-019-45839-z

Anders S., Pyl P. T., Huber W (2015) **HTSeq—A Python framework to work with high-throughput sequencing data. Bioinformatics (Oxford** *England* **31**:166–169 https://doi.org/10.1093/bioinformatics/btu638

Arai Y., Funatsu N., Numayama-Tsuruta K., Nomura T., Nakamura S., Osumi N (2005) **Role of Fabp7, a downstream gene of Pax6, in the maintenance of neuroepithelial cells during early embryonic development of the rat cortex** *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience* **25**:9752–9761 https://doi.org/10.1523/JNEUROSCI.2512-05.2005

Ben-Ari Y., Khalilov I., Kahle K. T., Cherubini E (2012) **The GABA Excitatory/Inhibitory Shift in Brain Maturation and Neurological Disorders** *The Neuroscientist* **18**:467–486 https://doi.org/10.1177/1073858412438697

Benito-Kwiecinski S. *et al.* (2021) **An early cell shape transition drives evolutionary expansion of the human forebrain** *Cell* **184**:2084–2102 https://doi.org/10.1016/j.cell.2021.02.050

Benjamini Y., Hochberg Y (1995) **Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing** *Journal of the Royal Statistical Society: Series B (Methodological* **57**:289–300 https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Blake L. E., Roux J., Hernando-Herraez I., Banovich N. E., Perez R. G., Hsiao C. J., Eres I., Cuevas C., Marques-Bonet T., Gilad Y (2020) **A comparison of gene expression and DNA methylation patterns across tissues and species** *Genome Research* **30**:250–262 https://doi.org/10.1101/gr.254904.119

Institute Broad **Broad Institute. (n.d.). Picard.** *Picard*

Brose N. *et al.* (2019) **Synaptic vesicle fusion: Today and beyond** *Nature Structural & Molecular Biology* **26**:663–668 https://doi.org/10.1038/s41594-019-0277-z

Buenrostro J. D., Giresi P. G., Zaba L. C., Chang H. Y., Greenleaf W. J (2013) **Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position** *Nature Methods* **10**:1213–1218 https://doi.org/10.1038/nmeth.2688

Burridge P. W. *et al.* (2014) **Chemically defined generation of human cardiomyocytes** *Nature Methods* **11**:855–860 https://doi.org/10.1038/nmeth.2999

Calderari S. *et al.* (2018) **Molecular genetics of the transcription factor GLIS3 identifies its dual function in beta cells and neurons** *Genomics* **110**:98–111 https://doi.org/10.1016/j.ygeno.2017.09.001

Caporali A., Emanueli C (2009) **Cardiovascular actions of neurotrophins** *Physiological Reviews* **89**:279–308 https://doi.org/10.1152/physrev.00007.2008

Castel S. E. *et al.* (2020) **A vast resource of allelic expression data spanning human tissues** *Genome Biology* **21** https://doi.org/10.1186/s13059-020-02122-z

Castro-Mondragon J. A. *et al.* (2022) **JASPAR 2022: The 9th release of the open-access database of transcription factor binding profiles** *Nucleic Acids Research* **50**:D165–D173 https://doi.org/10.1093/nar/gkab1113

Chal J., Al Tanoury Z., Hestin M., Gobert B., Aivio S., Hick A., Cherrier T., Nesmith A. P., Parker K. K., Pourquié O. (2016) **Generation of human muscle fibers and satellite-like cells from human pluripotent stem cells in vitro** *Nature Protocols* **11**:1833–1850 https://doi.org/10.1038/nprot.2016.110

Chen K. M., Wong A. K., Troyanskaya O. G., Zhou J (2022) **A sequence-based global map of regulatory activity for deciphering human genetics** *Nature Genetics* **54**:940–949 https://doi.org/10.1038/s41588-022-01102-2

Chen S. *et al.* (2022) **A genome-wide mutational constraint map quantified from variation in 76,156 human genomes [Preprint]** *Genetics* https://doi.org/10.1101/2022.03.20.485034

Choi W.-S., Xu X., Goruk S., Wang Y., Patel S., Chow M., Field C. J., Godbout R (2021) **FABP7 Facilitates Uptake of Docosahexaenoic Acid in Glioblastoma Neural Stem-like Cells** *Nutrients* **13** https://doi.org/10.3390/nu13082664

Collins R. L. *et al.* (2022) **A cross-disorder dosage sensitivity map of the human genome** *Cell* **185**:3041–3055 https://doi.org/10.1016/j.cell.2022.06.036

Combs P. A., Krupp J. J., Khosla N. M., Bua D., Petrov D. A., Levine J. D., Fraser H. B (2018) **Tissue-Specific cis-Regulatory Divergence Implicates eloF in Inhibiting Interspecies Mating in Drosophila** *Current Biology* **28**:3969–3975 https://doi.org/10.1016/j.cub.2018.10.036

Conway J. R., Lex A., Gehlenborg N (2017) **UpSetR: An R package for the visualization of intersecting sets and their properties** *Bioinformatics* **33**:2938–2940 https://doi.org/10.1093/bioinformatics/btx364

Corces M. R. *et al.* (2017) **An improved ATAC-seq protocol reduces background and enables interrogation of frozen tissues** *Nature Methods* **14**:959–962 https://doi.org/10.1038/nmeth.4396

De Rosa A. *et al.* (2012) **A radial glia gene marker, fatty acid binding protein 7 (FABP7), is involved in proliferation and invasion of glioblastoma cells** *PloS One* **7** https://doi.org/10.1371/journal.pone.0052113

Dobin A., Davis C. A., Schlesinger F., Drenkow J., Zaleski C., Jha S., Batut P., Chaisson M., Gingeras T. R (2013) **STAR: Ultrafast universal RNA-seq aligner. Bioinformatics (Oxford *England* 29**:15–21 https://doi.org/10.1093/bioinformatics/bts635

Ebrahimi M. *et al.* (2016) **Astrocyte-expressed FABP7 regulates dendritic morphology and excitatory synaptic function of cortical neurons** *Glia* **64**:48–62 https://doi.org/10.1002/glia.22902

Ernst J., Kellis M (2017) **Chromatin-state discovery and genome annotation with ChromHMM** *Nature Protocols* **12**:2478–2492 https://doi.org/10.1038/nprot.2017.124

Feldblum S., Erlander M. G., Tobin A. J (1993) **Different distributions of GAD65 and GAD67 mRNAS suggest that the two glutamate decarboxylases play distinctive functional roles** *Journal of Neuroscience Research* **34**:689–706 https://doi.org/10.1002/jnr.490340612

Field A. R. *et al.* (2019) **Structurally Conserved Primate LncRNAs Are Transiently Expressed during Human Cortical Differentiation and Influence Cell-Type-Specific Genes** *Stem Cell Reports* **12**:245–257 https://doi.org/10.1016/j.stemcr.2018.12.006

Fraser H. B (2011) **Genome-wide approaches to the study of adaptive gene expression evolution: Systematic studies of evolutionary adaptations involving gene expression will allow many fundamental questions in evolutionary biology to be addressed** *BioEssays* **33**:469–477 https://doi.org/10.1002/bies.201000094

Fraser H. B (2013) **Gene expression drives local adaptation in humans** *Genome Research* **23**:1089–1096 https://doi.org/10.1101/gr.152710.112

Fraser H. B (2019) **Improving Estimates of Compensatory cis–trans Regulatory Divergence** *Trends in Genetics* **35**:3–5 https://doi.org/10.1016/j.tig.2018.09.003

García-Pérez R. *et al.* (2021) **Epigenomic profiling of primate lymphoblastoid cell lines reveals the evolutionary patterns of epigenetic activities in gene regulatory architectures** *Nature Communications* **12** https://doi.org/10.1038/s41467-021-23397-1

Girskis K. M. *et al.* (2021) **Rewiring of human neurodevelopmental gene regulatory programs by human accelerated regions** *Neuron* **109**:3239–3251 https://doi.org/10.1016/j.neuron.2021.08.005

Gokhman D. *et al.* (2021) **Human–chimpanzee fused cells reveal cis-regulatory divergence underlying skeletal evolution** *Nature Genetics* **53**:467–476 https://doi.org/10.1038/s41588-021-00804-3

GTEx Consortium (2017) **Genetic effects on gene expression across human tissues** *Nature* **550**:204–213 https://doi.org/10.1038/nature24277

Horlbeck M. A. *et al.* (2016) **Compact and highly active next-generation libraries for CRISPR-mediated gene repression and activation** *ELife* **5** https://doi.org/10.7554/eLife.19760

Hu C. K., York R. A., Metz H. C., Bedford N. L., Fraser H. B., Hoekstra H. E (2022) **Cis-Regulatory changes in locomotor genes are associated with the evolution of burrowing behavior** *Cell Reports* **38** https://doi.org/10.1016/j.celrep.2022.110360

Huang E. J., Reichardt L. F (2001) **Neurotrophins: Roles in neuronal development and function** *Annual Review of Neuroscience* **24**:677–736 https://doi.org/10.1146/annurev.neuro.24.1.677

Hubisz M. J., Pollard K. S (2014) **Exploring the genesis and functions of Human Accelerated Regions sheds light on their role in human evolution** *Current Opinion in Genetics & Development* **29**:15–21 https://doi.org/10.1016/j.gde.2014.07.005

Ichim G., Tauszig-Delamasure S., Mehlen P (2012) **Neurotrophins and cell death** *Experimental Cell Research* **318**:1221–1228 https://doi.org/10.1016/j.yexcr.2012.03.006

Jain A., Tuteja G (2019) **TissueEnrich: Tissue-specific gene enrichment analysis. Bioinformatics (Oxford** *England* **35**:1966–1967 https://doi.org/10.1093/bioinformatics/bty890

St. John John **John St. John. (n.d.). SeqPrep.**

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

Ke K. *et al.* (2015) **Up-Regulation of Glis2 Involves in Neuronal Apoptosis After Intracerebral Hemorrhage in Adult Rats** *Cellular and Molecular Neurobiology* **35**:345–354 https://doi.org/10.1007/s10571-014-0130-1

Kelley J. L., Gilad Y (2020) **Effective study design for comparative functional genomics** *Nature Reviews Genetics* **21**:385–386 https://doi.org/10.1038/s41576-020-0242-z

King M.-C., Wilson A. C. (1975) **Evolution at Two Levels in Humans and Chimpanzees: Their macromolecules are so alike that regulatory mutations may account for their biological differences** *Science* **188**:107–116 https://doi.org/10.1126/science.1090005

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

Korytnikov R., Nostro M. C (2016) **Generation of polyhormonal and multipotent pancreatic progenitor lineages from human pluripotent stem cells** *Methods* **101**:56–64 https://doi.org/10.1016/j.ymeth.2015.10.017

Kozlenkov A. *et al.* (2020) **Evolution of regulatory signatures in primate cortical neurons at cell-type resolution** *Proceedings of the National Academy of Sciences* **117**:28422–28432 https://doi.org/10.1073/pnas.2011884117

Kuhn R. M., Haussler D., Kent W. J (2013) **The UCSC genome browser and associated tools** *Briefings in Bioinformatics* **14**:144–161 https://doi.org/10.1093/bib/bbs038

Langmead B., Salzberg S. L (2012) **Fast gapped-read alignment with Bowtie 2** *Nature Methods* **9**:357–359 https://doi.org/10.1038/nmeth.1923

Li H., Handsaker B., Wysoker A., Fennell T., Ruan J., Homer N., Marth G., Abecasis G., Durbin R., 1000 Genome Project Data Processing Subgroup. (2009) **The Sequence Alignment/Map format and SAMtools** *Bioinformatics* :2078–2079 https://doi.org/10.1093/bioinformatics/btp352

Liang D. *et al.* (2021) **Cell-type-specific effects of genetic variation on chromatin accessibility during human neuronal differentiation** *Nature Neuroscience* **24**:941–953 https://doi.org/10.1038/s41593-021-00858-w

Lopez-Delisle L., Rabbani L., Wolff J., Bhardwaj V., Backofen R., Grüning B., Ramírez F., Manke T (2021) **pyGenomeTracks: Reproducible plots for multivariate genomic datasets. Bioinformatics (Oxford** *England* **37**:422–423 https://doi.org/10.1093/bioinformatics/btaa692

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

Ma S. *et al.* (2022) **Molecular and cellular evolution of the primate dorsolateral prefrontal cortex. Science** *eabo* **7257** https://doi.org/10.1126/science.abo7257

Mack K. L., Nachman M. W (2017) **Gene Regulation and Speciation** *Trends in Genetics* **33**:68–80 https://doi.org/10.1016/j.tig.2016.11.003

Mallanna S. K., Duncan S. A (2013) **Differentiation of hepatocytes from pluripotent stem cells** *Current Protocols in Stem Cell Biology* **26**:1–1 https://doi.org/10.1002/9780470151808.sc01g04s26

Maury Y., Côme J., Piskorowski R. A., Salah-Mohellibi N., Chevaleyre V., Peschanski M., Martinat C., Nedelec S (2015) **Combinatorial analysis of developmental cues efficiently converts human pluripotent stem cells into multiple neuronal subtypes** *Nature Biotechnology* **33**:89–96 https://doi.org/10.1038/nbt.3049

Meisler M. H., Hill S. F., Yu W (2021) **Sodium channelopathies in neurodevelopmental disorders** *Nature Reviews Neuroscience* **22**:152–166 https://doi.org/10.1038/s41583-020-00418-4

Milani P., Escalante-Chong R., Shelley B. C., Patel-Murray N. L., Xin X., Adam M., Mandefro B., Sareen D., Svendsen C. N., Fraenkel E (2016) **Cell freezing protocol suitable for ATAC-Seq on motor neurons derived from human induced pluripotent stem cells** *Scientific Reports* **6** https://doi.org/10.1038/srep25474

Mizuguchi R., Kriks S., Cordes R., Gossler A., Ma Q., Goulding M (2006) **Ascl1 and Gsh1/2 control inhibitory and excitatory cell fate in spinal sensory interneurons** *Nature Neuroscience* **9**:770–778 https://doi.org/10.1038/nn1706

Montaigne D., Butruille L., Staels B (2021) **PPAR control of metabolism and cardiovascular functions** *Nature Reviews Cardiology* **18**:809–823 https://doi.org/10.1038/s41569-021-00569-6

Mueller Fabian **Mueller, Fabian. (n.d.). ChrAccR.**

Naito Y., Lee A. K., Takahashi H (2017) **Emerging roles of the neurotrophin receptor TrkC in synapse organization** *Neuroscience Research* **116**:10–17 https://doi.org/10.1016/j.neures.2016.09.009

Bank Netherlands Brain *et al.* (2016) **Epigenomic annotation of gene regulatory alterations during evolution of the primate brain** *Nature Neuroscience* **19**:494–503 https://doi.org/10.1038/nn.4229

Pavlovic B. J., Blake L. E., Roux J., Chavarria C., Gilad Y (2018) **A Comparative Assessment of Human and Chimpanzee iPSC-derived Cardiomyocytes with Primary Heart Tissues** *Scientific Reports* **8** https://doi.org/10.1038/s41598-018-33478-9

Pollard K. S. *et al.* (2006) **An RNA gene expressed during cortical development evolved rapidly in humans** *Nature* **443**:167–172 https://doi.org/10.1038/nature05113

Prud'homme B., Gompel N., Carroll S. B (2007) **Emerging principles of regulatory evolution** *Proceedings of the National Academy of Sciences* **104** https://doi.org/10.1073/pnas.0700488104

Quinlan A. R., Hall I. M (2010) **BEDTools: A flexible suite of utilities for comparing genomic features. Bioinformatics (Oxford** *England* **26**:841–842 https://doi.org/10.1093/bioinformatics/btq033

Ramírez F., Dündar F., Diehl S., Grüning B. A., Manke T (2014) **deepTools: A flexible platform for exploring deep-sequencing data** *Nucleic Acids Research* **42** https://doi.org/10.1093/nar/gku365

Reilly S. K., Noonan J. P (2016) **Evolution of Gene Regulation in Humans** *Annual Review of Genomics and Human Genetics* **17**:45–67 https://doi.org/10.1146/annurev-genom-090314-045935

Romero I. G., Ruvinsky I., Gilad Y (2012) **Comparative studies of gene expression and the evolution of gene regulation** *Nature Reviews Genetics* **13**:505–516 https://doi.org/10.1038/nrg3229

Sharma R. *et al.* (2019) **Clinical-grade stem cell-derived retinal pigment epithelium patch rescues retinal degeneration in rodents and pigs** *Science Translational Medicine* **11** https://doi.org/10.1126/scitranslmed.aat5580

Shave R. E. *et al.* (2019) **Selection of endurance capabilities and the trade-off between pressure and volume in the evolution of the human heart** *Proceedings of the National Academy of Sciences* **116**:19905–19910 https://doi.org/10.1073/pnas.1906902116

Song J. H. T., Grant R. L., Behrens V. C., Kučka M., Roberts Kingman G. A., Soltys V., Chan Y. F., Kingsley D. M (2021) **Genetic studies of human-chimpanzee divergence using stem cell fusions** *Proceedings of the National Academy of Sciences of the United States of America* **118** https://doi.org/10.1073/pnas.2117557118

Starr A. L., Gokhman D., Fraser H. B (2023) **Accounting for cis-regulatory constraint prioritizes genes likely to affect species-specific traits** *Genome Biology* **24** https://doi.org/10.1186/s13059-023-02846-8

Lab Storey **Storey Lab. (n.d.). Qvalue.**

Subramanian A. *et al.* (2005) **Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles** *Proceedings of the National Academy of Sciences* **102**:15545–15550 https://doi.org/10.1073/pnas.0506580102

Su-Feher L. *et al.* (2022) **Single cell enhancer activity distinguishes GABAergic and cholinergic lineages in embryonic mouse basal ganglia** *Proceedings of the National Academy of Sciences* **119** https://doi.org/10.1073/pnas.2108760119

Sullivan P. F. *et al.* (2023) **Leveraging Base Pair Mammalian Constraint to Understand Genetic Variation and Human Disease [Preprint]** *Genomics* https://doi.org/10.1101/2023.03.10.531987

Traag V. A., Waltman L., van Eck N. J. (2019) **From Louvain to Leiden: Guaranteeing well-connected communities** *Scientific Reports* **9** https://doi.org/10.1038/s41598-019-41695-z

Trevino A. E., Sinnott-Armstrong N., Andersen J., Yoon S.-J., Huber N., Pritchard J. K., Chang H. Y., Greenleaf W. J. (2020) **Chromatin accessibility dynamics in a model of human forebrain development. Science (New York** *N.Y* **367** https://doi.org/10.1126/science.aay1645

Trizzino M., Park Y., Holsbach-Beltrame M., Aracena K., Mika K., Caliskan M., Perry G. H., Lynch V. J., Brown C. D (2017) **Transposable elements are the primary source of novelty in primate gene regulation** *Genome Research* **27**:1623–1633 https://doi.org/10.1101/gr.218149.116

van de Geijn B., McVicker G., Gilad Y., Pritchard J. K. (2015) **WASP: Allele-specific software for robust molecular quantitative trait locus discovery** *Nature Methods* **12**:1061–1063 https://doi.org/10.1038/nmeth.3582

Vanderhaeghen P., Polleux F (2023) **Developmental mechanisms underlying the evolution of human cortical circuits** *Nature Reviews Neuroscience* **24**:213–232 https://doi.org/10.1038/s41583-023-00675-z

Vierstra J. *et al.* (2020) **Global reference mapping of human transcription factor footprints** *Nature* **583**:729–736 https://doi.org/10.1038/s41586-020-2528-x

Vu H., Ernst J (2022) **Universal annotation of the human genome through integration of over a thousand epigenomic datasets** *Genome Biology* **23** https://doi.org/10.1186/s13059-021-02572-z

Wang B., Tontonoz P (2018) **Liver X receptors in lipid signalling and membrane homeostasis** *Nature Reviews Endocrinology* **14**:452–463 https://doi.org/10.1038/s41574-018-0037-x

Watanabe A. *et al.* (2007) **Fabp7 Maps to a Quantitative Trait Locus for a Schizophrenia Endophenotype** *PLoS Biology* **5** https://doi.org/10.1371/journal.pbio.0050297

Wittkopp P. J., Kalay G (2012) **Cis-regulatory elements: Molecular mechanisms and evolutionary processes underlying divergence** *Nature Reviews Genetics* **13**:59–69 https://doi.org/10.1038/nrg3095

Wolf F. A., Angerer P., Theis F. J (2018) **SCANPY: Large-scale single-cell gene expression data analysis** *Genome Biology* **19** https://doi.org/10.1186/s13059-017-1382-0

Wolf F. A., Hamey F. K., Plass M., Solana J., Dahlin J. S., Göttgens B., Rajewsky N., Simon L., Theis F. J (2019) **PAGA: Graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells** *Genome Biology* **20** https://doi.org/10.1186/s13059-019-1663-x

Yang N. *et al.* (2017) **Generation of pure GABAergic neurons by transcription factor programming** *Nature Methods* **14**:621–628 https://doi.org/10.1038/nmeth.4291

Yao Z. *et al.* (2021) **A taxonomy of transcriptomic cell types across the isocortex and hippocampal formation** *Cell* **184**:3222–3241 https://doi.org/10.1016/j.cell.2021.04.021

York R. A., Patil C., Abdilleh K., Johnson Z. V., Conte M. A., Genner M. J., McGrath P. T., Fraser H. B., Fernald R. D., Streelman J. T (2018) **Behavior-dependent cis regulation reveals genes and pathways associated with bower building in cichlid fishes** *Proceedings of the National Academy of Sciences* **115** https://doi.org/10.1073/pnas.1810140115

Yu X., Lin J., Zack D. J., Qian J (2006) **Computational analysis of tissue-specific combinatorial gene regulation: Predicting interaction between transcription factors in human tissues** *Nucleic Acids Research* **34**:4925–4936 https://doi.org/10.1093/nar/gkl595

Zhang S. *et al.* (2020) **Allele-specific open chromatin in human iPSC neurons elucidates functional disease variants** *Science* **369**:561–565 https://doi.org/10.1126/science.aay3983

Zhang Y. *et al.* (2008) **Model-based analysis of ChIP-Seq (MACS)** *Genome Biology* **9** https://doi.org/10.1186/gb-2008-9-9-r137

Zhu A., Ibrahim J. G., Love M. I (2019) **Heavy-tailed prior distributions for sequence count data: Removing the noise and preserving large differences** *Bioinformatics* **35**:2084–2092 https://doi.org/10.1093/bioinformatics/bty895

Article and author information

Ban Wang
Department of Biology, Stanford University, Stanford, CA, USA

Alexander L. Starr
Department of Biology, Stanford University, Stanford, CA, USA

Hunter B. Fraser
Department of Biology, Stanford University, Stanford, CA, USA
For correspondence: hbfraser@stanford.edu

Copyright

Editors

Reviewing Editor
Jenny Tung
Duke University, Durham, United States of America

Senior Editor
Christian Landry
Université Laval, Québec, Canada

Reviewer #1 (Public Review):

This study aims to identify gene expression differences exclusively caused by cis-regulatory genetic changes by utilizing hybrid cell lines derived from human and chimpanzee. While previous attempts have focused on specific tissues, this study expands the comparison to six different tissues to investigate tissue specificity and derive insights into the evolution of gene expression.

One notable strength of this work lies in the use of composite cell lines, enabling a comparison of gene expression between human and chimpanzee within the same nucleus and shared trans factors environment. However, a potential weakness of the methodology is the use of bulk RNA-seq in diverse tissues, which limits the ability to determine cell-type-specific gene expression and chromatin accessibility regions. Their approach, using hybrid lines, naturally accounts for cell type heterogeneity avoiding the risk of false positives introduced by the otherwise confounding differences in cell type abundances between species, albeit the challenge of false negatives remains an issue. The authors now dully acknowledge this limitation in the manuscript.

Another concern is the use of two replicates derived from the same pair of individuals. While the authors produced cell lines from two pairs of individuals in a previous study (Agloglia et al., 2021). The reason for this experimental design is cost limitations. The authors now acknowledge that the use of replicates could enhance the ability to detect "more" species-specific changes in expression and chromatin accessibility. I would emphasize that replicates would increase robustness to the present findings, given that they are derived from a single pair of individuals.

Furthermore, the study offers the opportunity to relate inter-species differences to trends in molecular evolution. The authors discovered that expression variance and haploinsufficiency score do not fully account for the enrichment of divergence in cell-type-specific genes. The reviewer suggested exploring this further by incorporating external datasets that bin genes based on interindividual transcriptomics variation as a measure of extant transcriptomics constraint (e.g., GTEx reanalysis by Garcia-Perez et al., 2023 - PMID: 36777183). The authors considered this question to be out of the scope of the paper, yet in my opinion this would enhance one of the main findings of this study.

Additionally, stratifying sequence conservation on ASCA regions, which exhibit similar enrichment of cell-type-specific features, using the Zoonomia data mentioned also in the text (Andrews et al., 2023 -- PMID: 37104580) could provide valuable insights. While the author did not find Zoonomia Phastcons values available, they used PhastCons derived from a 470-way alignment of mammals. I commend the authors for their diligent efforts, which undoubtedly bolster their findings that an enrichment in ASCA is evident across all levels of sequence conservation. However, this recent analysis indicates the presence of a potential relationship between sequence conservation and ASCA. It may be advantageous to consider evaluating more quantile subdivisions of maxZ values and pPhastCons values, with the inclusion of these results in the supplementary materials. This approach would be preferable, even if the precise reasons behind the observed discrepancy are not fully elucidated.

Another potential strength of this study is the identification of specific cases of paired allele-specific expression (ASE) and allele-specific chromatin accessibility (ASCA) with biological significance. Prioritizing specific variants remains a challenge, and the authors apply a machine learning approach to identify potential causative variants that disrupt binding sites in two examples (FABP7 and GAD1 in motor neurons). However, additional work is needed to convincingly demonstrate the functionality of these selected variants. Strengthening this section with additional validation of ASE, ASCA, and the specific putative causal variants

identified would enhance the overall robustness of the paper. The authors have opted to defer these validations to future studies.

Additionally, the authors support the selected ASE-ASCA pairs by examining external datasets of adult brain comparative genomics (Ma et al., 2022) and organoids (Kanton et al., 2019). While these resources are valuable for comparing observed species biases, the analysis is not systematic, even for the two selected genes. For example, it would be beneficial to investigate if FABP7 exhibits species bias in any cell type in Kanton et al.'s organoids or if GAD1 is species-biased in adult primate brains from Ma et al. Comparing these datasets with the present study, along with the Agoglia et al. reference, would provide a more comprehensive perspective. In the revised version of the manuscript the authors have evaluated the expression of GAD1 in Ma et al, and FABP7 in Sousa et al 2017. For instance, GAD1 show cell type specific species biases in the later. The authors opted for not showing this in the manuscript, However, it remains unclear why certain datasets were favored over others, or why FABP7 should not be evaluated in Kanton et al.

The use of the term "human-derived" in ASE and ASCA has now been avoided.

Finally, throughout the paper, the authors refer to "hybrid cell lines." It has been suggested to use the term "composite cell lines" instead to address potential societal concerns associated with the term "hybrid," which some may associate with reproductive relationships (Pavlovic et al., 2022 -- PMID: 35082442). The authors have presented an eloquent and persuasive explanation that I found to be highly informative.

- https://doi.org/10.7554/eLife.89594.2.sa1

Reviewer #3 (Public Review):

The authors utilize chimpanzee-human hybrid cell lines to assess cis-regulatory evolution. These hybrid cell lines offer a well-controlled environment, enabling clear differentiation between cis-regulatory effects and environmental or other trans effects.
In their research, Wang et al. expand the range of chimpanzee-human hybrid cell lines to encompass six new developmental cell types derived from all three germ layers. This expansion allows them to discern cell type-specific cis-regulatory changes between species from more pleiotropic ones. Although the study investigates only two iPSC clones, the RNA- and ATAC-seq data produced for this paper is a valuable resource.

The authors begin their analysis by examining the relationship between allele-specific expression (ASE) as a measure of species divergence and cell type specificity. They find that cell-type-specific genes exhibit more divergent expression. By integrating this data with measures of constraint within human populations, the authors conclude that the increased divergence of tissue-specific genes is, at least in part, attributable to positive selection. A similar pattern emerges when assessing allele-specific chromatin accessibility (ASCA) as a measure of divergence of cis-regulatory elements (CREs) in the same cell lines.

By correlating these two measures, the authors identify 95 CRE-gene pairs where tissue-specific ASE aligns with tissue-specific ASCA. Among these pairs, the authors select two genes of interest for further investigation. Notably, the authors employ an intriguing machine learning approach in which they compare the inferred chromatin state of the human sequence with that of the chimpanzee sequence to pinpoint putatively causal variants.

Overall, this study delves into the examination of gene expression and chromatin accessibility within hybrid cell lines, showcasing how this data can be leveraged to identify potential causal sequence differences underlying between-species expression changes.

All in all most conclusions appear solid, with the exception of the interpretation of a cell type/state identification machine learning model to pinpoint putatively causal variants. The described variants lack any functional validation and there is no data that measure the certainty of the results.

- https://doi.org/10.7554/eLife.89594.2.sa0

Author Response

The following is the authors' response to the original reviews.

> ***eLife assessment***
>
> *This is an important study that leverages a human-chimpanzee tetraploid iPSC model to test whether cis-regulatory divergence between species tends to be cell type-specific. The evidence supporting the study's primary conclusion--that species differences in gene regulation are enriched in cell type-specific genes and regulatory elements--is compelling, although attention to biases introduced by sequence conservation is merited, and the case that is made for cell type-specific changes reflecting adaptive evolution is incomplete. This work will be of broad interest in evolutionary and functional genomics.*
>
> ***Public Reviews:***
>
> ***Reviewer #1 (Public Review):***
>
> *This study aims to identify gene expression differences exclusively caused by cis-regulatory genetic changes by utilizing hybrid cell lines derived from human and chimpanzee. While previous attempts have focused on specific tissues, this study expands the comparison to six different tissues to investigate tissue specificity and derive insights into the evolution of gene expression.*
>
> *One notable strength of this work lies in the use of composite cell lines, enabling a comparison of gene expression between human and chimpanzee within the same nucleus and shared trans factors environment. However, a potential weakness of the methodology is the use of bulk RNA-seq in diverse tissues, which limits the ability to determine cell-type-specific gene expression and chromatin accessibility regions.*

We agree that profiling single cells could lead to additional exciting discoveries. Although heterogeneity in cell types within samples will indeed reduce our power to detect cell-type-specific divergence, thankfully any heterogeneity will not introduce false positives, since our use of interspecies hybrids controls for differences in cell-type abundance. As a result, we think that the molecular differences we identified in this study represent a subset of the true cell-type specific cis-regulatory differences that would be identified with deep single-cell profiling. We have included a new paragraph in the discussion on future directions, highlighting the utility of single-cell profiling as an exciting future direction (lines 482-490): "In addition to following up on our findings on GAD1 and FABP7, there are other exciting future directions for this work. First, additional bulk assays such as those that measure methylation, chromatin conformation, and translation rate could lead to a better understanding of what molecular features ultimately lead to cell type-specific changes in gene expression. Furthermore, the use of deep single cell profiling of hybrid lines derived from iPSCs from multiple individuals of each species during differentiation could enable the identification of many more highly context-specific changes in gene expression and chromatin accessibility such as the differences in GAD1 we highlighted here. Finally, integration with data from massively parallel reporter assays and deep learning models will help us link specific variants to the molecular differences we identified in this study."

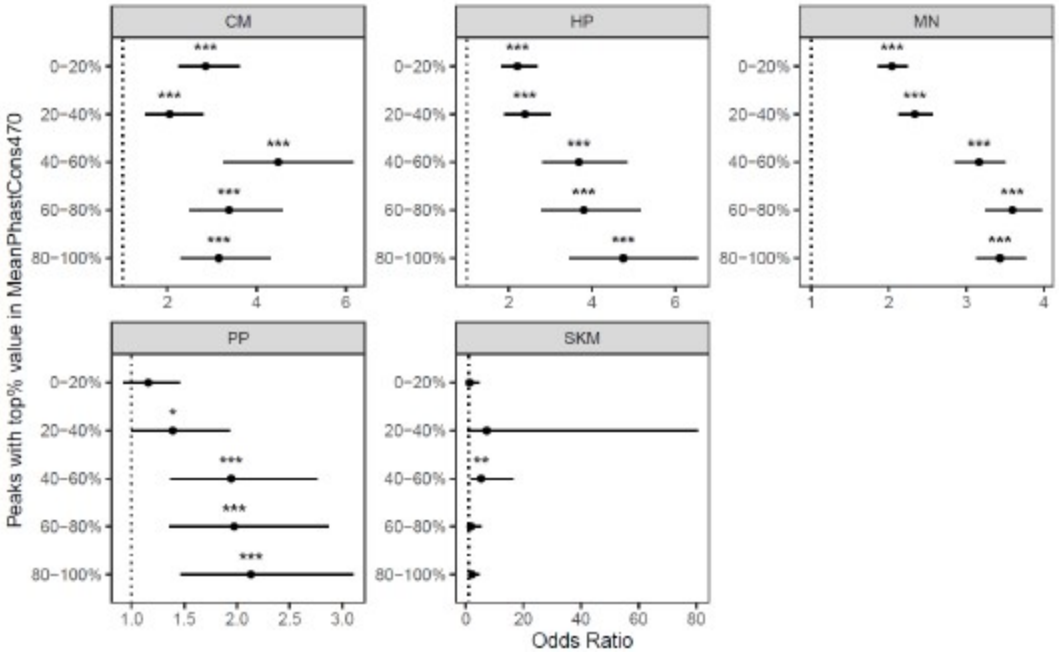
> *Another concern is the use of two replicates derived from the same pair of individuals. While the authors produced cell lines from two pairs of individuals in a previous study (Agloglia et al., 2021), I wonder why only one pair was used in this study. Incorporating interindividual variation would enhance the robustness of the species differences identified here.*

We agree that additional replicates, especially from lines from other individuals, would have improved the robustness of the species differences we identified. In our experience with these hybrid cells (as well as related work from many other labs), inter-species differences typically have much larger magnitudes than intra-species differences, so we expect that the vast majority of differences we identified would be validated with data from additional individuals. Unfortunately, differentiating additional cells and generating these data for this study would be cost-prohibitive. We now mention the use of additional replicates in lines 485-488 of the discussion: “Furthermore, the use of deep single cell profiling of hybrid lines derived from iPSCs from multiple individuals of each species during differentiation could enable the identification of many more highly context-specific changes in gene expression and chromatin accessibility such as the differences in GAD1 we highlighted here.”

> *Furthermore, the study offers the opportunity to relate inter-species differences to trends in molecular evolution. The authors discovered that expression variance and haploinsufficiency score do not fully account for the enrichment of divergence in cell-type-specific genes. The reviewer suggests exploring this further by incorporating external datasets that bin genes based on interindividual transcriptomics variation as a measure of extant transcriptomics constraint (e.g., GTEx reanalysis by Garcia-Perez et al., 2023 - PMID: 36777183). Additionally, stratifying sequence conservation on ASCA regions, which exhibit similar enrichment of cell-type-specific features, using the Zoonomia data mentioned also in the text (Andrews et al., 2023 -- PMID: 37104580) could provide valuable insights.*

To address this, we used PhastCons scores computed from a 470-way alignment of mammals as we could not find publicly available PhastCons data from Zoonomia. When stratifying by the median PhastCons score of all sites in a peak, we observe very similar results to those obtained when stratifying by the constraint metric from the gnomAD consortium (see below). The one potential difference is that peaks in the top two bins have slightly weaker enrichment relative to the other bins when using PhastCons, but this is not the case when using gnomAD’s metric. We have elected to include this in the public review but not the manuscript as we are reluctant to add to the complexity of what is already complex analysis.

Author response image 1.



Finally, we think that comparisons of the properties of gene expression variance computed from ASE (as done by Starr et al.) and total expression (as done by Garcia-Perez et al.) is a very interesting, potentially complex question that is beyond the scope of this paper but an exciting direction for future work.

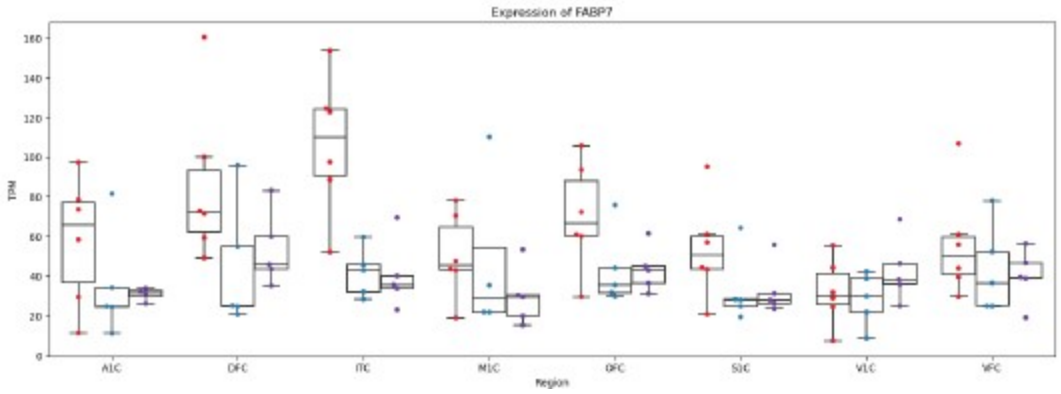
> *Another potential strength of this study is the identification of specific cases of paired allele-specific expression (ASE) and allele-specific chromatin accessibility (ASCA) with biological significance.*
>
> *Prioritizing specific variants remains a challenge, and the authors apply a machine-learning approach to identify potential causative variants that disrupt binding sites in two examples (FABP7 and GAD1 in motor neurons). However, additional work is needed to convincingly demonstrate the functionality of these selected variants. Strengthening this section with additional validation of ASE, ASCA, and the specific putative causal variants identified would enhance the overall robustness of the paper.*

We strongly agree with the reviewer that additional work validating our results would be of considerable interest. We hope to perform follow-up experiments in the future. For now, we have been careful to present these variants only as candidate causal variants.

> *Additionally, the authors support the selected ASE-ASCA pairs by examining external datasets of adult brain comparative genomics (Ma et al., 2022) and organoids (Kanton et al., 2019). While these resources are valuable for comparing observed species biases, the analysis is not systematic, even for the two selected genes. For example, it would be beneficial to investigate if FABP7 exhibits species bias in any cell type in Kanton et al.'s organoids or if GAD1 is species-biased in adult primate brains from Ma et al. Comparing these datasets with the present study, along with the Agoglia et al. reference, would provide a more comprehensive perspective.*

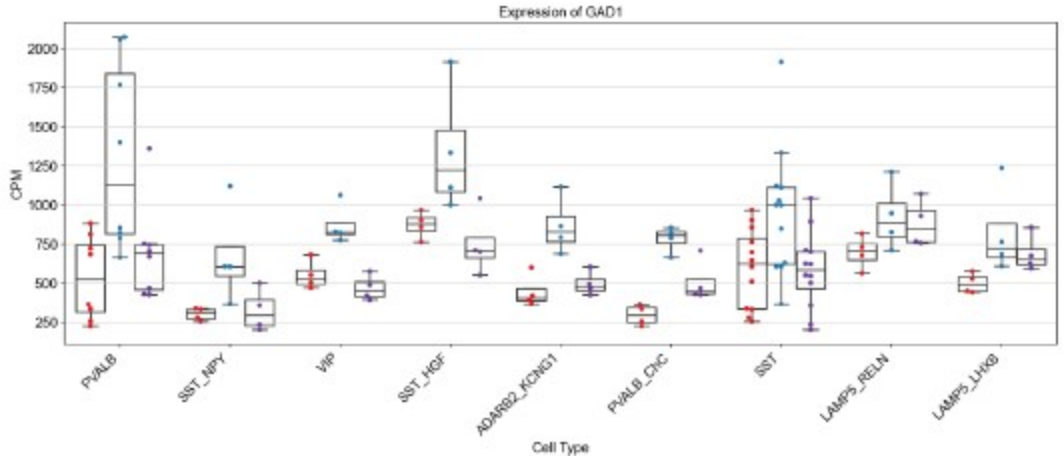
We agree with the reviewer's suggestion that investigating GAD1 and FABP7 expression in other datasets is worthwhile. Unfortunately, the difference in human vs. chimpanzee organoid maturation rates and effects of culture conditions in Kanton et al. makes it unsuitable for plotting the expression of FABP7 as its expression is highly dependent on neuronal maturation. We therefore plotted bulk RNAseq data from multiple cortical regions from Sousa et al. 2017 (see below). This corroborates our claim that FABP7 has human-biased expression in adult humans compared to chimpanzees and rhesus macaques. We also investigated expression of GAD1 in the Ma et al. data as the reviewer suggested.

Author response image 2.



While there are differences in GAD1 expression between adult humans and chimpanzees, they are unlikely to be linked to the HAR we highlight as it is likely a transiently active cis-regulatory element (see below). In addition, some cell types seem to have chimpanzee-derived changes in GAD1 expression (e.g. SST positive neurons) whereas others seem to have human-derived changes in GAD1 expression (e.g. LAMP5 positive neurons).

Author response image 3.



While these are potentially interesting observations, we think that their inclusion in the manuscript might distract from our emphasis on the cell type-specific and developmental stage-specific of the changes in FABP7 and GAD1 expression we observe so we have not included them in the manuscript.

> *The use of the term "human-derived" in ASE and ASCA should be avoided since there is no outgroup in the analysis to provide a reference for the observed changes.*

We agree with the reviewer that the term human-derived should be used with care and have changed the phrasing of line 230 to “human-chimpanzee differences in expression”. With regard to FABP7 we think that our analysis of the Ma et al. data—which includes data from rhesus macaques as an outgroup—justifies our use of “human-derived” in lines 360 and 457. As chimpanzee and macaque expression of FABP7 are similar but human expression is quite different, the most parsimonious explanation for our observations is that FABP7 upregulation occurred in the human lineage.

> *Finally, throughout the paper, the authors refer to "hybrid cell lines." It has been suggested to use the term "composite cell lines" instead to address potential societal concerns associated with the term "hybrid," which some may associate with reproductive relationships (Pavlovic et al., 2022 -- PMID: 35082442). It would be interesting to know the authors' perspective on these concerns and recommendations presented in Pavlovic et al., given their position as pioneers in this field.*

We appreciate this question. Whether to refer to our fused cells as “hybrids” or not was indeed a question we considered at great length, starting from the very beginning of this project in 2015. From consultations with multiple bioethicists-- both formal and informal-- we have long been aware of the possibility of misunderstanding based on the word “hybrid”. However, we felt this possibility was outweighed by the long and well-established history of other scientists referring to interspecies fused cells as hybrids. This convention-- which is based on hundreds of papers about heterokaryons, somatic cell hybrids, and radiation hybrids-- goes back over 50 years (e.g. Bolund et al, Exp Cell Res 1969). Soon after the establishment of this nomenclature, cell fusion became widespread and ever since then it has become commonplace to generate interspecies hybrid cells from animals, plants and fungi.

It is also important to note that in over two years since we published the first two papers on humanchimpanzee fused cells, we have been unable to find any misunderstanding of our use of the term “hybrid”. We have searched blogs, media articles, and social media, all with no evidence of misunderstanding. Therefore, in the current manuscript, rather than creating confusion by renaming a well-established approach, we have opted to clearly and prominently define hybrid cells: in the abstract of our paper we introduce the hybrid cells as “the product of fusing induced pluripotent stem (iPS) cells of each species in vitro.”

> ***Reviewer #2 (Public Review):***
>
> *In this paper, Wang and colleagues build on previous technical and analytical achievements in establishing tetraploid human-chimpanzee hybrid iPSCs to investigate the cell type-specificity of allelespecific expression and allele-specific chromatin accessibility across six differentiated cell types (here, "allele-specific" indicates species differences with a cis-regulatory basis). The combined body of work is remarkable in its creativity and ambition and has real potential for overcoming major challenges in understanding the evolutionary genetics of between-species differences. The present paper contributes to these efforts by showing how differentiated cells can be used to test a long-standing hypothesis in evolutionary genetics: that cis-regulatory changes may be particularly important in divergence because of their potential for modularity.*
>
> *In my view, the paper succeeds in making this case: allele (species)-specific expression (ASE) and allelespecific chromatin accessibility (ASCA) are enriched in genes asymmetrically expressed in one cell type, and many cases of ASE/ASCA are cell type-specific. The authors do an excellent job showing that these results are robust across a*

> *set of possible analysis decisions. It is somewhat less clear whether these enrichments are primarily a product of relaxed constraint on cell type-specific genes or primarily result from positive selection in the human or chimp lineage. While the authors attempt to control for constraint using several variables (variance in ASE in humans and the sequence-based probability of haploinsufficiency score, pHI), these are imperfect proxies for constraint. For the pHI scores, enrichments for ASE also appear to be strongest in the least constrained genes. Overall, the relative role of relaxation of constraint versus positive selection is unresolved, although the manuscript's language leans in favor of an important role for selection.*

We agree with the reviewer and apologize for the wording that indeed focused more on positive selection than relaxed constraint. We have added language clarifying that our stance is that our analyses suggest some role for positive selection, but that we do not claim that positive selection plays a larger role than reduced constraint (lines 432-437): “Overall, this suggests that broad changes in expression in cell type-specifically expressed genes may be an important substrate for evolution but it remains unclear whether positive selection or lower constraint plays a larger role in driving the faster evolution of more cell type-specifically expressed genes. Future work will be required to more precisely quantify the relative roles of positive selection and evolutionary constraint in driving changes in gene expression.”

> *The remainder of the manuscript draws on the cell type-specific ASE/ASCA data to nominate candidate genes and pathways that may have been important in differentiating humans and chimpanzees. Several approaches are used here, including comparing human-chimp ASE to the distribution of ASE observed in humans and investigating biases in the direction of ASE for genes in the same pathway. The authors also identify interesting candidate genes based on their role in development or their proximity to human accelerated regions (where many changes have arisen on the human lineage in otherwise deeply conserved sequence) and use a deep neural network to identify sequence changes that might be causally responsible for ASE/ASCA. These analyses have value and highlight potential strategies for using ASE/ASCA and hybrid cell line data as a hypothesis-generating tool. Of course, the functional follow-up that experimentally tested these hypotheses or linked sequence/expression changes in the candidate pathways to organismal phenotype would have strengthened the paper further- but this is a lot to ask in an already technically and analytically challenging piece of work.*

We thank the reviewer for the kind words and strongly agree that follow-up experiments and orthogonal analyses will be key in validating our results and establishing links to human-specific phenotypes.

> *As a minor critique, the present paper is very closely integrated with other manuscripts that have used the hybrid human-chimp cell lines for biological insight or methods development. Although its contributions make it a strong stand-alone contribution, some aspects of the methods are not described in sufficient detail for readers to understand (even on a general conceptual level) without referencing that work, which may somewhat limit reader understanding.*

We agree with the points the reviewer raises regarding the clarity of our methods. We have amended several sections to provide more conceptual information while pointing the reader to other publications for the technical details. For convenience, we include the text here as well as in the new draft.

Lines 207-214 now provide more intuition for the method used to detect lineage-specific selection: “Next, we sought to use our RNA-seq data to identify instances of lineage-specific

selection. In the absence of positive selection, one would expect that an approximately equal number of genes in a pathway would have human-biased vs. chimpanzee-biased ASE. Significant deviation from this expectation (as determined by the binomial test) rejects the null hypothesis of neutral evolution, instead providing evidence of lineage-specific selection on this pathway. Using our previously published modification of this test that incorporates a tissue-specific measure of constraint on gene expression, we detected several signals of lineage-specific selection, some of which were cell type-specific (Starr et al., 2023, Additional file 2)." This is also reflected in the Methods in lines 729-731: "Positive selection on a gene set is only inferred if there is statistically significant human- or chimpanzee-biased ASE in that gene set (using an FDR-corrected p-value from the binomial test)."

> ***Reviewer #3 (Public Review):***
>
> *The authors utilize chimpanzee-human hybrid cell lines to assess cis-regulatory evolution. These hybrid cell lines offer a well-controlled environment, enabling clear differentiation between cis-regulatory effects and environmental or other trans effects.*
>
> *In their research, Wang et al. expand the range of chimpanzee-human hybrid cell lines to encompass six new developmental cell types derived from all three germ layers. This expansion allows them to discern cell type-specific cis-regulatory changes between species from more pleiotropic ones. Although the study investigates only two iPSC clones, the RNA- and ATAC-seq data produced for this paper is a valuable resource.*
>
> *The authors begin their analysis by examining the relationship between allele-specific expression (ASE) as a measure of species divergence and cell type specificity. They find that cell-type-specific genes exhibit more divergent expression. By integrating this data with measures of constraint within human populations, the authors conclude that the increased divergence of tissue-specific genes is, at least in part, attributable to positive selection. A similar pattern emerges when assessing allele-specific chromatin accessibility (ASCA) as a measure of divergence of cis-regulatory elements (CREs) in the same cell lines.*
>
> *By correlating these two measures, the authors identify 95 CRE-gene pairs where tissue-specific ASE aligns with tissue-specific ASCA. Among these pairs, the authors select two genes of interest for further investigation. Notably, the authors employ an intriguing machine-learning approach in which they compare the inferred chromatin state of the human sequence with that of the chimpanzee sequence to pinpoint putatively causal variants.*
>
> *Overall, this study delves into the examination of gene expression and chromatin accessibility within hybrid cell lines, showcasing how this data can be leveraged to identify potential causal sequence differences underlying between-species expression changes.*

We appreciate this assessment.

> *I have three major concerns regarding this study:*
>
> 1. *The only evidence that the cells are indeed differentiated in the right direction is the expression of one prominent marker gene per cell type. Especially for the comparison of conservation between the differentiated cell types, it would be beneficial to describe the cell type diversity and the differentiation success in more detail.*

We appreciate this assessment. We agree that evidence beyond a single marker gene is necessary to demonstrate that the differentiations were successful and that a discussion of the limitations of these differentiations in the manuscript is worthwhile. We included figures

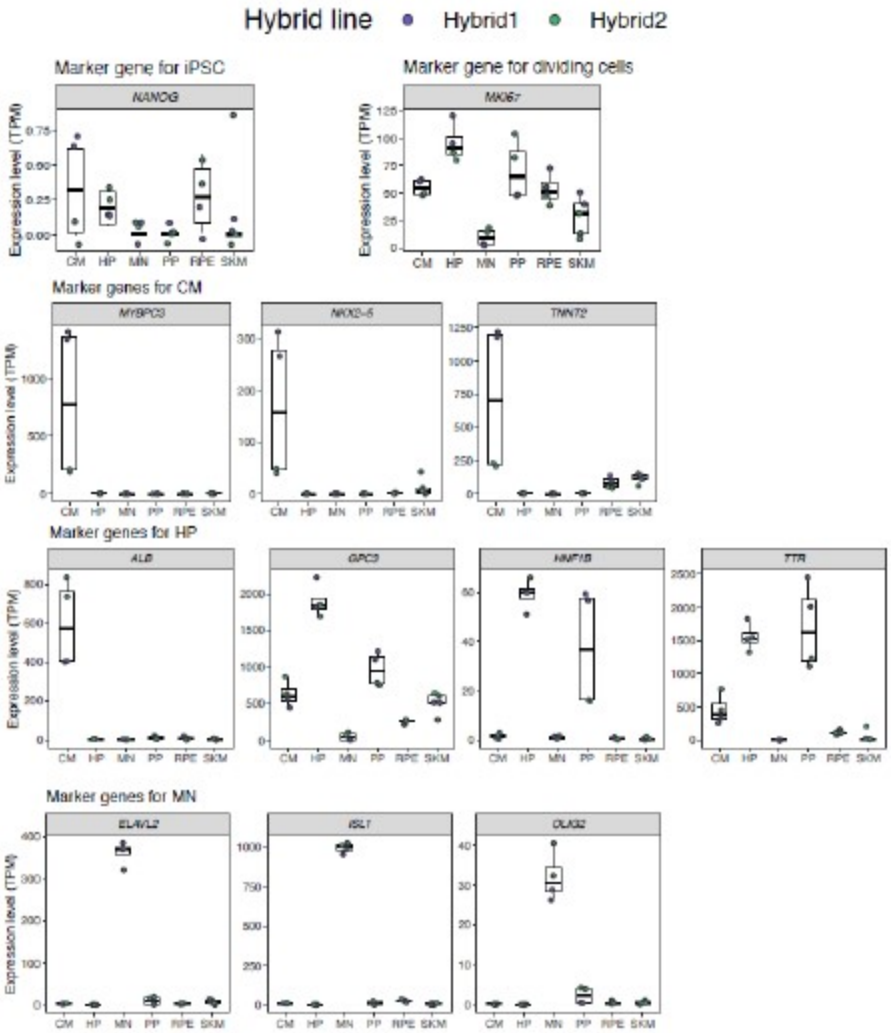
showing additional marker genes and a thorough discussion of the differentiations in the supplement. For convenience, we have copied the supplemental figure and text here:

“Before continuing with the analysis, we tested whether the differentiations were successful and contained primarily our target cell types. The very low expression of NANOG, a marker for pluripotency, across all differentiations indicates that the samples contain very few iPSCs (Agoglia et al., 2021). For cardiomyocytes (CM), NKX2-5, MYBPC3, and TNNT2 definitively distinguish CM from other heart cell types and their high expression indicates successful differentiations (Burridge et al., 2014). For motor neurons, the high expression of ELAVL2, a pan-neuronal marker, indicates a high abundance of neurons in the sample (Mickelsen et al., 2019). The expression of ISL1 and OLIG2 further demonstrates that these are motor neurons and not other types of neurons (Maury et al., 2015). For retinal pigment epithelium (RPE), the combined expression of MITF, PAX6, and TYRP1 provides strong evidence that the differentiations were successful in producing RPE cells (Sharma et al., 2019). For skeletal muscle, the very high expression of MYL1, MYLPF, and MYOG indicates that these samples contain a high proportion of skeletal muscle cells (Chal et al., 2016). In general, all these populations of cells contain some proportion of progenitors as there is detectable expression of MKI67 in all samples.

The low expression of ALB (a marker for mature hepatocytes) and the high expression of TTR and GPC3 (markers for hepatocyte progenitors) combined with the high expression of HNF1B indicate that the bulk of the cells in the HP samples are hepatocyte progenitors rather than mature hepatocytes or endoderm cells, although there are likely some endoderm cells and immature hepatocytes in the sample (Hay et al., 2008; Mallanna & Duncan, 2013). Similarly, the combined expression of PDX1 and NKX6-1 and the low expression of NEUROG3 (a marker of endocrine progenitors which differentiate from pancreatic progenitors) in the PP samples indicates that these primarily contain pancreatic progenitors but likely contain some endocrine progenitors and endoderm cells (Cogger et al., 2017; Korytnikov & Nostro, 2016).

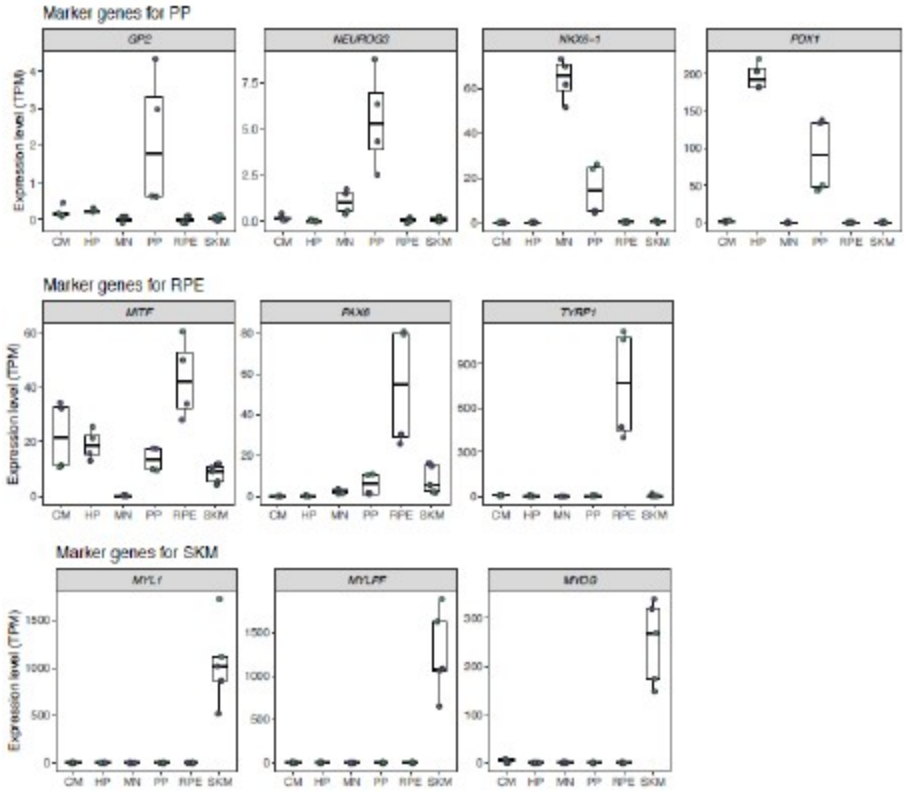
Notably, HP and PP are closely related cell types that are derived from the same lineage. Indeed, heterogeneous multipotent progenitors can contribute to both the adult liver and adult pancreas in mice (Willnow et al., 2021). Progenitors that express PDX1 (often used as a marker for the pancreatic lineage) can differentiate into hepatocytes (Willnow et al., 2021). As a result, some overlap in the transcriptomic signature of both cell types is expected and we cannot rule out that the HP samples contain cells that could differentiate into pancreatic cells or that the PP samples contain cells that could differentiate into hepatocytes. However, the expression of NKX6-1 and GP2, markers for pancreatic progenitors, in the PP samples but not the HP samples indicates that these two populations of cells are distinct. Overall, the similarity of PP and HP likely explains the lower number of cell type-specific genes and genes showing cell type-specific ASE for these cell types. This similarity does not alter the conclusions presented in the main text.”

Author response image 4.



Author response image 5.

Marker gene expression in different cell types. In order, the panels show: a marker for pluripotency, a marker gene for dividing cells, marker genes for cardiomyocytes, marker genes for hepatocytes and hepatocyte progenitors, marker genes for motor neurons, marker genes for pancreatic progenitors and more mature pancreatic cell types, marker genes for retinal pigment epithelial cells, and marker genes for skeletal myocytes. Hepatocyte progenitors and pancreatic progenitors generally show similar gene expression profiles. TPM: transcript per million.



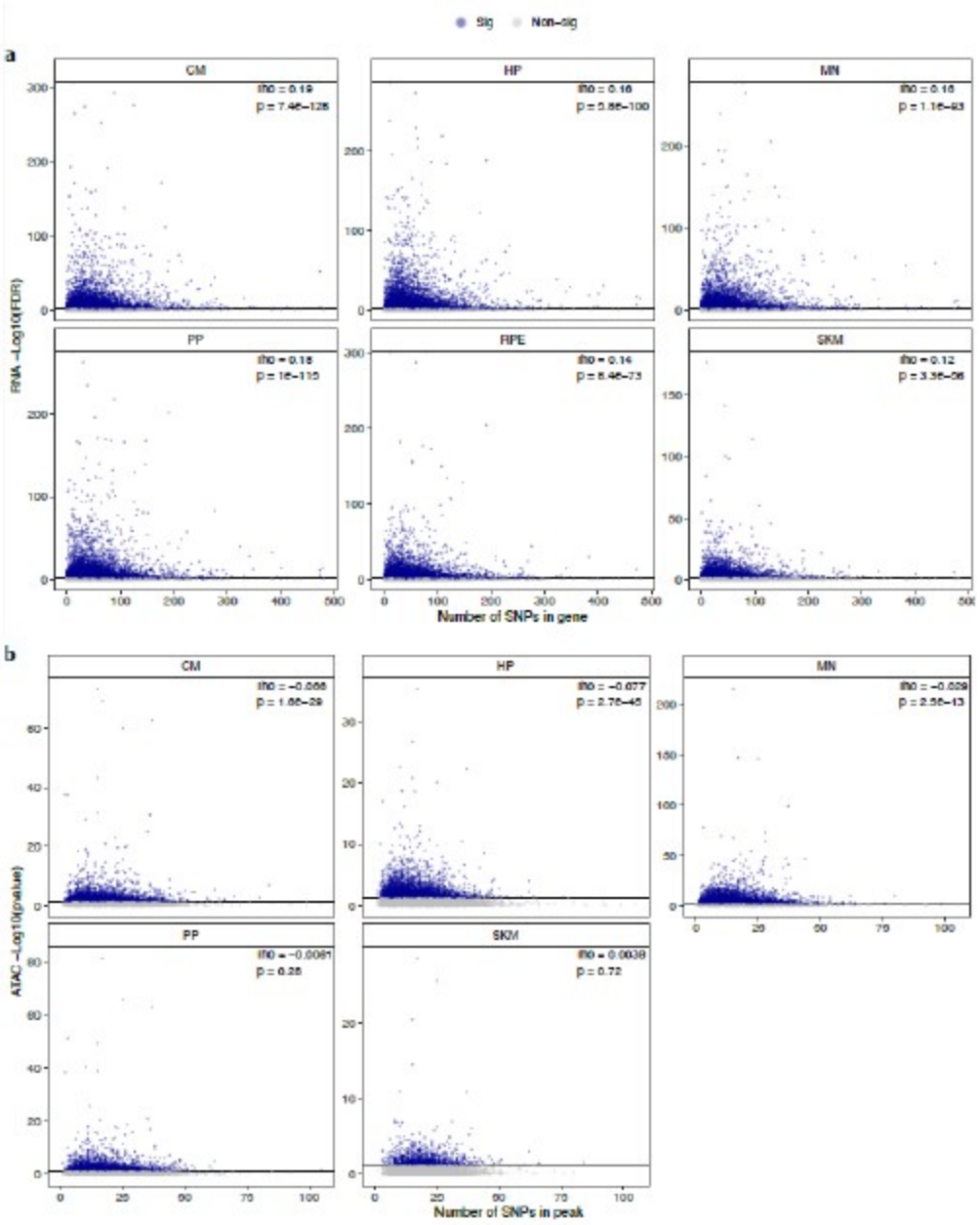
> *1. Check for a potential confounding effect of sequence similarity on the power to detect ASE or ASCA.*

We agree that checking for confounding by power to detect ASE or ASCA would increase confidence in our results. We have added supplementary figures 29-33 to show the results as well as a discussion of these figures in the text (lines 318-326):

“Finally, it is possible that CREs and genes that are less conserved will have more SNPs, and therefore more power to call ASCA and ASE, leading to systematically biased estimates. There is a weak positive correlation between the number of SNPs and the -log10(FDR) for ASE and a weak negative or no correlation for ASCA (Supp Fig. 29). Similarly, we observe a weak relationship between the number of SNPs in CREs or genes and absolute log fold-change estimates (Supp Fig. 30). Although the relationship between the number of SNPs and ASE/ASCA is weak, we confirmed that cell type-specific genes and peaks are still strongly enriched for ASE and ASCA when stratifying by number of SNPs (Supp Fig. 31-32). Overall, our analysis suggests that the result that more cell type-specific genes and CREs are more evolutionarily diverged is robust to a variety of possible confounders.”

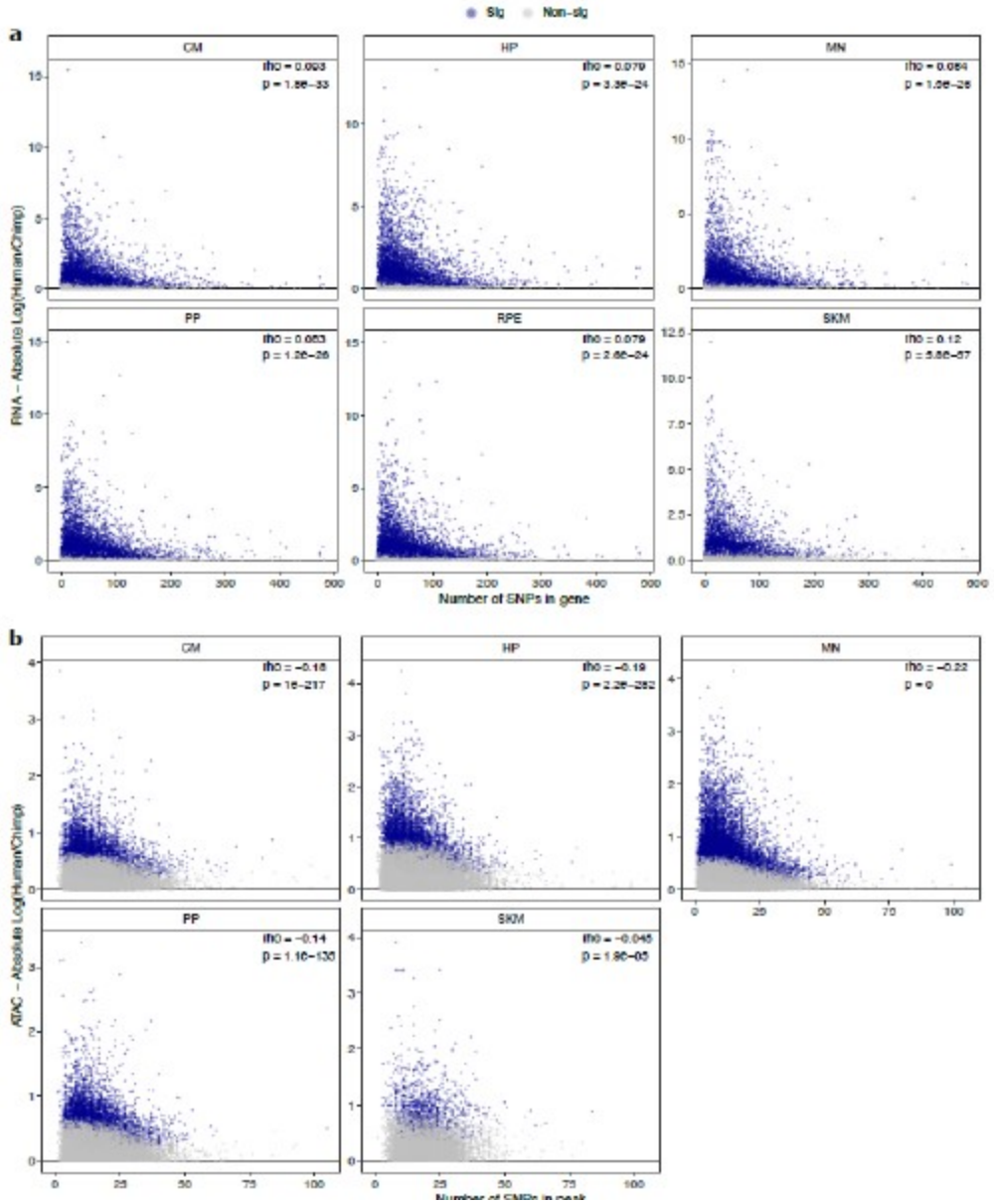
Author response image 6.

Relationship between number of SNPs and -log10(FDR) in a) ASE and -log10(pvalue) b) ASCA. These scatter plots show the relationship between the number of SNPs in a gene or peak and the -log10(FDR) for ASE or ASCA. Genes with significant ASE (FDR < 0.05) and peaks with significant ASCA (binomial p-value < 0.05) were annotated as blue dots, and all other genes and peaks were annotated as gray dots. All genes in each cell type in RNA-seq are shown. For clarity, the few outlier peaks with more than 200 SNPs are excluded from these plots.



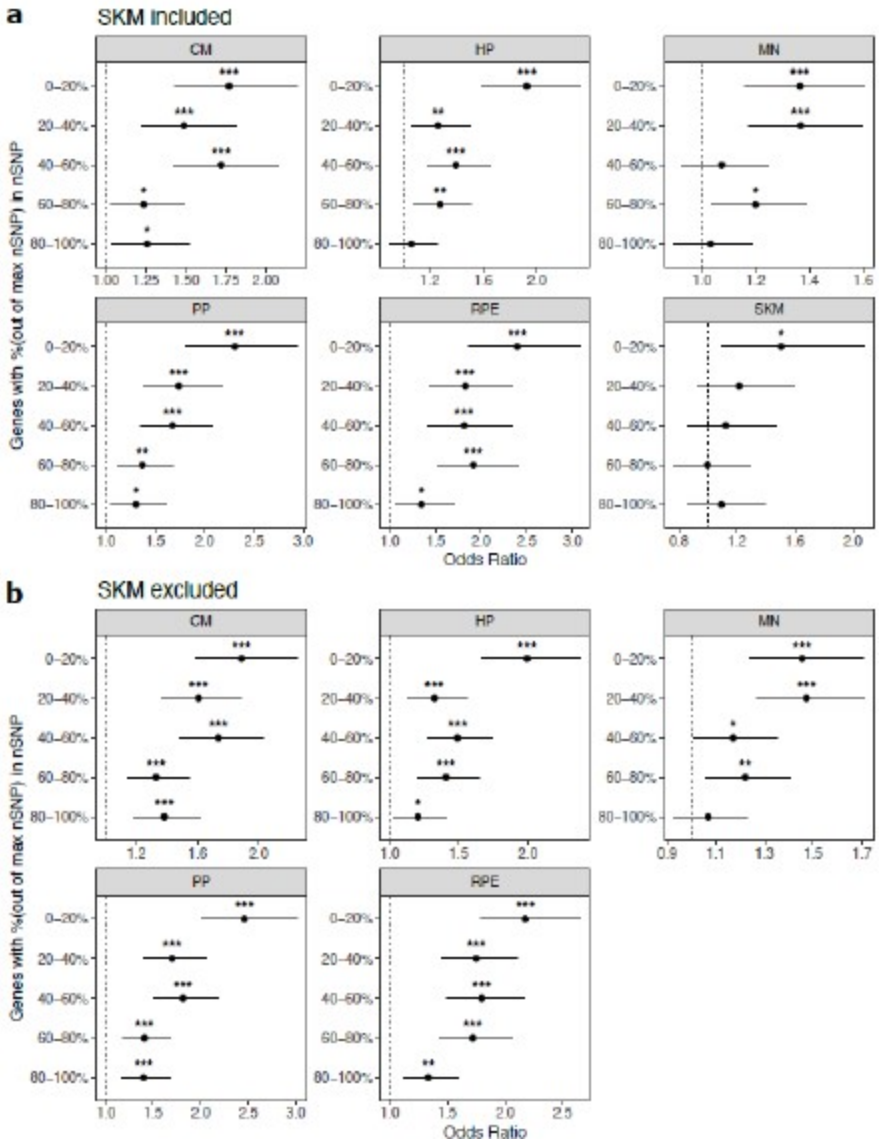
Author response image 7.

Relationship between number of SNPs and absolute log2 fold-change in a) ASE and b) ASCA. These scatter plots show the relationship between the number of SNPs in a gene or peak and the estimated absolute log2 fold-change for ASE or ASCA. Genes with significant ASE (FDR < 0.05) and peaks with significant ASCA (binomial p-value < 0.05) were annotated as blue dots, and all other genes and peaks were annotated as gray dots. All genes in each cell type in RNA-seq are shown. For clarity, the few outlier peaks with more than 200 SNPs are excluded from these plots.



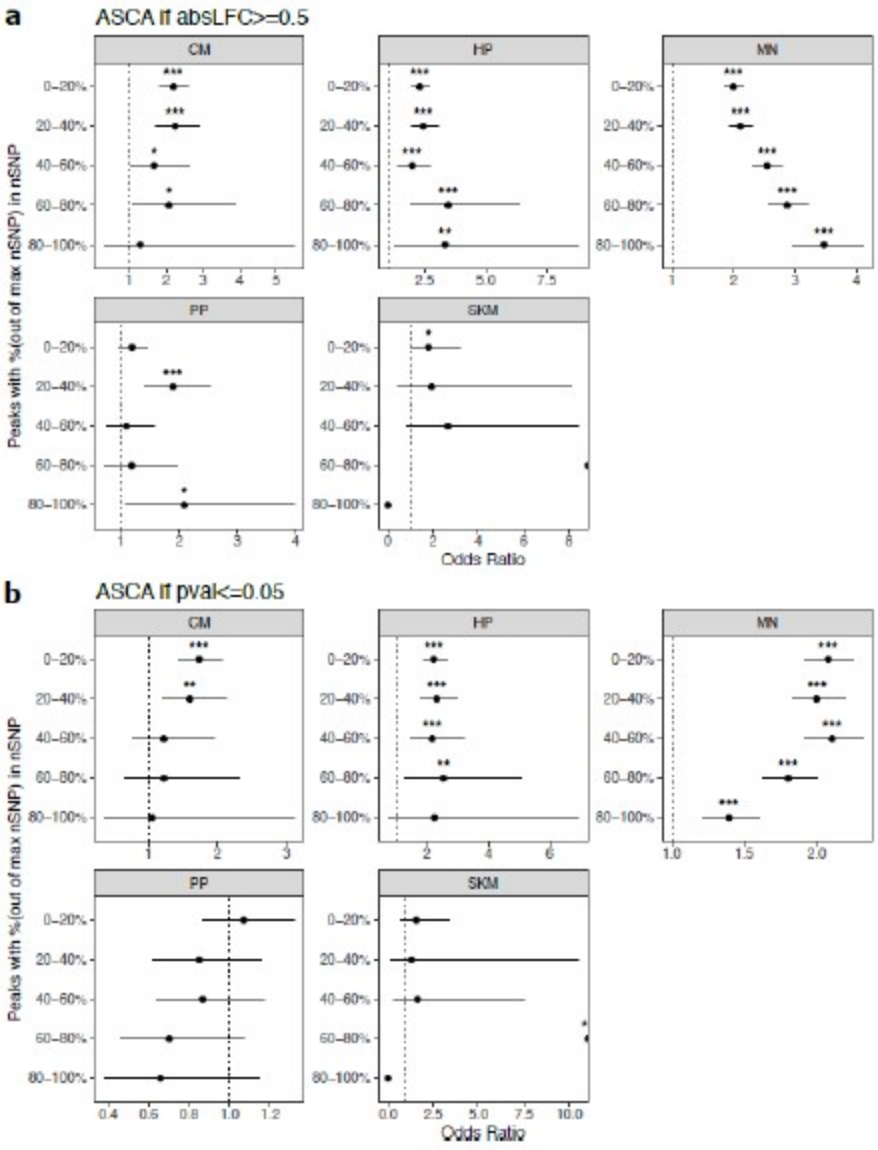
Author response image 8.

Cell type-specifically expressed genes are enriched for genes with ASE when stratifying by the number of SNPs per gene. a) Results when SKM is included. Genes were put into five bins with an equal number of genes in each bin. Genes with the fewest SNPs are in the 0-20% bin and genes with the most SNPs are in the 80-100% bin. Significance (using the Wald test) is indicated by asterisks where *** indicates $p < 0.005$, ** indicates $p < 0.01$, and * indicates $p < 0.05$. b) The same as in (a) but excluding SKM.



Author response image 9.

Cell type-specific peaks are enriched for ASCA when stratifying by the number of SNPs per peak. a) Peaks with an absolute log2 fold-change greater than or equal to 0.5 were called as having ASCA. Peaks were put into five bins with an equal number of peaks in each bin. Peaks with the fewest SNPs are in the 0-20% bin and genes with the most SNPs are in the 80-100% bin. Significance (using the Wald test) is indicated by asterisks where *** indicates $p < 0.005$, ** indicates $p < 0.01$, and * indicates $p < 0.05$. b) The same as in (a) but peaks with a binomial p-value less than or equal to 0.05 were called as having ASCA.



1. *In the last part the authors showcase 2 examples for which the log2 fold changes in chromatin state scores as inferred by the machine learning model Sei are used. This is an interesting and creative approach, however, more sanity checks on this application are necessary.*

We agree with the reviewer about the importance of sanity checks and apologize for omitting these from the manuscript. Below we highlight several such checks from previous publications:

In the original Sei paper (Chen et al. 2022), the authors included several tests of their model's ability to predict the effects on individual genetic variants. Using eQTL data from GTEx, they found that variants predicted to increase enhancer activity were more likely to be up-regulating eQTLs, and those predicted to increase polycomb repression had the expected repressive effect. These relationships became stronger when restricting the analysis only to fine-mapped eQTLs with >95% posterior probabilities of causality. Chen et al. also found that previously known disease-causing noncoding variants from the Human Gene Mutation

Database were far more likely to reduce predicted enhancer/promoter activity than matched variants not linked to any disease.

In addition, we note that a similar approach to ours was recently used to analyze all HARs and included considerable efforts to validate the utility of the Sei predictions in identifying causal variants (Whalen et al. 2023 in Neuron). For example, Whalen et al. found that the Sei output correlated with the effects of genetic variants on expression in a massively parallel reporter assay. They also found that the effect sizes predicted by Sei were much higher for variants in HARs than polymorphic variants in the human population, which is consistent with the idea that variants in HARs lie in highly conserved bases that are more likely to disrupt cis-regulatory elements. Finally, Whalen et al. found that effects on chromatin state predicted by Sei were generally highly correlated across tissues, supporting our approach that leverages all Sei outputs regardless of which cell type or tissue they correspond to. Overall, we think that Sei is a potentially powerful way to prioritize causal variants and that improved machine learning models trained on more extensive and context-specific data will be even more powerful.