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

Ban Wang[1,*], Alexander L. Starr[1,*], Hunter B. Fraser[1,2]

*Indicates equal contribution

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

2. Correspondence: hbfraser@stanford.edu

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 tissues and 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.

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 determine whether genetic differences cause a gene to be differentially expressed. In post-mortem studies, genetic differences cannot be disentangled from sources of variation such as differences in diet, environment, cell type abundances, age, post-mortem interval, and other confounding factors. 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 on 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 et al., 2004). 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 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 differences between humans and chimpanzees are well-studied, there has been less focus on epigenetic changes, many of which are likely to underlie differences in 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 human-chimpanzee hybrid iPS cell-derived cell types 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 great apes as well as a resource that will inform functional dissection of human-specific molecular changes.

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 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)

144 (Burridge et al., 2014; Chal et al., 2016; Korytnikov & Nostro, 2016; Mallanna & Duncan, 2013; Maury et
145 al., 2015; Sharma et al., 2019).
146
147 Because our hybrid cells were grown concurrently with their human and chimpanzee diploid "parental"
148 cells, we performed an additional check for purity of the hybrid lines by quantifying genome-wide ASE.
149 We noticed that among our 25 RNA-seq samples, the two PP hybrid2 samples had a slight bias towards
150 higher expression from the chimpanzee alleles across all chromosomes. This is likely due to a small
151 fraction of contaminating chimpanzee cells in these samples. We corrected for this by reducing the
152 chimpanzee allele counts such that the number of reads assigned to the human and chimpanzee alleles
153 was equal. By simulating contamination of a hybrid sample with chimpanzee cells, we found that this
154 correction was conservative and that the log fold-change estimates were largely unaffected by
155 contamination after this correction (Methods, Supp Fig. 4).
156
157 We next investigated which genes were differentially expressed between the human and chimpanzee
158 allele in each cell type (Fig. 2a). We identified thousands of genes showing significantly biased ASE in
159 each cell type at a false discovery rate (FDR) cutoff of 0.05 (Fig. 2b). We detected a comparable number
160 of ASE genes in all cell types except SKM. As a result, we repeated all subsequent analyses both including
161 and excluding SKM and obtained qualitatively similar results regardless of whether SKM was included.
162
163 While a considerable number of genes had significant ASE in all cell types, many more genes only had
164 significant ASE in a single cell type, suggesting cell type-specific *cis*-regulatory divergence (Fig. 2b). A
165 notable family of developmentally important genes that exemplifies differences in ASE across cell types
166 is the neurotrophins and their receptors (Fig. 2c) (Caporali & Emanueli, 2009; Huang & Reichardt, 2001).
167 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 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 assigned many more genes to be cell type-specific, showed an even more significant ASE enrichment than the narrower definition (Fig. 2d). This result is not sensitive to the choice of FDR cutoff nor driven solely by a subgroup of highly expressed genes as it is well-preserved across FDR cutoffs (Supp Fig. 7) and different gene expression levels (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

191 when the same data are used to define two different quantities which are then compared (Supp Fig. 9)
192 (Fraser, 2019).
193
194 This enrichment (Fig. 2d) suggests that tissue-specific genes may have less constraint and/or more
195 frequent positive selection on their expression. We reasoned that if the trend was solely driven by
196 constraint, then controlling for constraint—even if imperfectly—would be expected to reduce the
197 strength of the relationship. To investigate this, we binned genes by their variance in ASE across a large
198 cohort of human samples which we have previously shown acts as a reasonable proxy for evolutionary
199 constraint on gene expression (Castel et al., 2020; Starr et al., 2023). Across cell types, we generally
200 observe significant enrichments in each bin and little difference in enrichment between bins, suggesting
201 that differences in constraint on expression of cell type-specific vs. ubiquitously expressed genes are not
202 solely responsible for our observations (Supp Fig. 10). Furthermore, we observe even stronger
203 enrichments using an alternative constraint metric, the probability of haploinsufficiency score (pHI)
204 likely due to the larger number of genes for which pHI can be calculated (Supp Fig. 11) (Collins et al.,
205 2022). Overall, our analysis suggests that differences in constraint are unlikely to fully explain these
206 trends, suggesting a potential role for positive selection.
207
208 **Lineage-specific selection has acted on tissue-specific gene expression divergence**
209
210 Next, we sought to use our RNA-seq data to identify instances of lineage-specific selection. We used a
211 previously published method that incorporates a tissue-specific measure of constraint on gene
212 expression to test whether groups of genes have likely been under selection (Starr et al., 2023). Using
213 this strategy, we detected several signals of putative lineage-specific selection, some of which were cell
214 type-specific (Additional file 2). Notably, the four most significant enrichments were found in 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 has been shown to play an important 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 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 play a major role in controlling 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-derived 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. Similarly, these genes have human-biased expression when comparing human, chimpanzee, and rhesus macaque glutamatergic cortical neurons in a previously published dataset (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.

239

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

241

242 While ASE provides insight into what gene expression changes might underlie phenotypic differences
243 between humans and chimpanzees, in the absence of additional data it is very difficult to prioritize
244 which specific mutations might cause expression divergence. To begin to fill this gap, we generated
245 ATAC-seq data from five of the six cell types (all except RPE), with each cell type sequenced to an
246 average depth of 184 million paired-end reads (Supp Fig. 14). ATAC-seq uses a hyperactive Tn5
247 transposase to cleave DNA that is not bound by nucleosomes to enrich for accessible chromatin (Fig. 3a),
248 a hallmark of active *cis*-regulatory elements (CREs) (Buenrostro et al., 2013; Corces et al., 2017). We
249 estimated ASCA for individual open chromatin peaks by mapping reads to both species' reference
250 genomes, correcting for mapping bias, generating a unified list of peaks across all samples, and then
251 counting reads supporting each allele in each peak (see Methods). Cell types clustered well using PC1
252 and PC2 of the ATAC-seq data, except for the HP sample which clustered closely with PP samples (Fig.
253 3b, Supp Fig. 15). However, performing PCA on just the HP and PP samples clearly separates the two cell
254 types (Supp Fig. 16). Within each cell type, species differences were clearly separated by PC1 or PC2
255 (Supp Fig. 17). As an example of ASCA, the accessibility of the promoter of *CTSF* was strongly
256 chimpanzee-biased, mirroring the chimpanzee-biased ASE for this gene (Fig. 3c).

257

258 As a first step in analyzing the ATAC-seq data, we intersected the peaks we identified with the genomic
259 annotations of chromatin states. These fifteen categories, predicted across many tissues and cell types
260 by the chromHMM model (Vu & Ernst, 2022), include terms such as "TSS" (transcription start site) and
261 "enhancer". We then plotted the median of the absolute human-chimpanzee ASCA log fold-changes for
262 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. 18a), as well as to removal of all peaks even slightly overlapping TSS or promoter-related annotations (Supp Fig. 18b). Given that heterochromatin/quiescent regions are highly divergent in ASCA (Fig. 3d), this suggests that changes in accessibility of these regions may be particularly important in the evolution of gene expression.

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, 19a), we performed down-sampling to equalize power to detect peaks across cell types (Supp Fig. 19b, 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 $\log_2$ fold-change or p-value cutoffs (Supp Figs. 20-21). 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 $\log_2$ 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. 22-23, Methods). Finally, we observe the same enrichments when controlling for a recently published metric for constraint on non-coding elements, suggesting that differences in evolutionary constraint may not be solely responsible for the observed trends (Supp Fig. 24) (S. Chen et al., 2022).

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. 25, Methods) and, analogously, dCAE (differential chromatin accessibility enrichment) measures how cell type-specific ASCA is for a *cis*-regulatory element (Supp Fig. 26, 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, and generalizes diffASE to an arbitrary number of cell types and any assay that produces log fold-changes (Combs et al., 2018; Hu et al., 2022; York et al., 2018). 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

311 accessibility may often impact tissue-specific gene expression, though this divergence is neither
312 necessary nor sufficient to do so.
313
314 **Identifying candidate causal *cis*-regulatory variants by integrating ASE and ASCA across cell types**
315
316 As high dEE and dCAE in a given cell type might be indicative of a causal link between the change in
317 chromatin accessibility and the change in expression, we focused on the 95 peak-gene pairs with
318 matching direction and used two different strategies to identify examples to investigate in detail. First,
319 we prioritized genes known to play important roles in development. For example, we found that the
320 promoter of *FABP7* has human-biased ASCA specifically in motor neurons (Fig. 4a-b) and that the *FABP7*
321 gene has human-biased ASE in motor neurons (Fig 4c). *FABP7* is used as a marker of glial cells and neural
322 progenitor cells (NPCs) and plays a key role in NPC proliferation and astrocyte function (Arai et al., 2005;
323 De Rosa et al., 2012; Ebrahimi et al., 2016; Watanabe et al., 2007). Using previously published single-
324 nucleus RNA-seq data from humans, chimpanzees, and rhesus macaques, we confirmed that *FABP7*
325 shows a human-derived up-regulation in several neuronal subtypes but not glial cells (Supp Fig. 27) (Ma
326 et al., 2022). To investigate the genetic basis of this cell type-specific divergence, we leveraged a
327 machine learning model, Sei (K. M. Chen et al., 2022), to nominate potentially causal variants in the
328 promoter of *FABP7* (see Methods). Sei is a deep neural network that takes DNA sequence as input and
329 predicts the probability that the sequence has a particular epigenetic state in a variety of cell types and
330 tissues (Fig. 4d) (K. M. Chen et al., 2022). While single-base substitutions differing between human and
331 chimpanzee had only minor impacts on predicted *cis*-regulatory activity, “chimpanizing” the human
332 sequence of the *FABP7* promoter at two small indels (by deleting one base at chr6: 122,779,291 and
333 inserting three bases at chr6: 122,779,115) was sufficient to make the Sei predictions for the
334 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. 28). 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. 29) (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. 30a), we investigated whether this reduced expression of human *GAD1*

359 also occurs in cortical organoids which contain GABAergic neurons together with other cell types in
360 which *GAD1* is not highly expressed. We analyzed our previously published data from human-
361 chimpanzee hybrid cortical organoids (Agoglia et al., 2021) and found that the expression of *GAD1* from
362 the chimpanzee allele spikes higher than that of the human allele around day 50 of hybrid cortical
363 organoid differentiation before dropping in expression over time to match the human expression level
364 (Supp Fig. 30b). Because ASE in the hybrid cells controls for any potential interspecies differences in
365 differentiation kinetics or cell type composition, this difference must be the result of *cis*-regulatory
366 divergence between humans and chimpanzees. This expression difference is also more pronounced in
367 comparisons of human and chimpanzee parental cortical organoids, with a higher absolute log fold-
368 change at day 50, day 100, and day 150, only returning to equal expression at day 200 (Supp Fig. 30c).
369 While this could be due to differences in cell type proportion between human and chimpanzee
370 organoids, it might also be due to a reinforcing *trans*-acting effect.
371
372 To test whether this difference in expression also occurs specifically during GABAergic neuron
373 differentiation we examined *GAD1* expression in single cell RNA-seq data from human and chimpanzee
374 cortical organoids (Kanton et al., 2019). We observed a spike in *GAD1* expression in less mature
375 chimpanzee GABAergic neurons that is absent in the corresponding part of the trajectory in human
376 neurons (Supp Fig. 31). Notably, a similar trend holds regardless of which GABAergic sub-trajectory (i.e.,
377 equivalent to GABAergic neurons from the caudal, lateral, or medial ganglionic eminences) is examined
378 suggesting this difference is not unique to a particular type of GABAergic neuron (Supp Fig. 31). Finally,
379 we examined the accessibility of the putative *GAD1* enhancer more closely. Consistent with a potential
380 role for this enhancer in the spike in *GAD1* expression during development, the accessibility of this
381 enhancer mirrors the expression of *GAD1* in human cortical and striatal organoids (Supp Fig. 32) with
382 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.

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, it is tempting to 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 potentially 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, we identified a variant that may account for the chimpanzee-biased ASCA in this region (K. M. Chen et al., 2022). 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.

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.

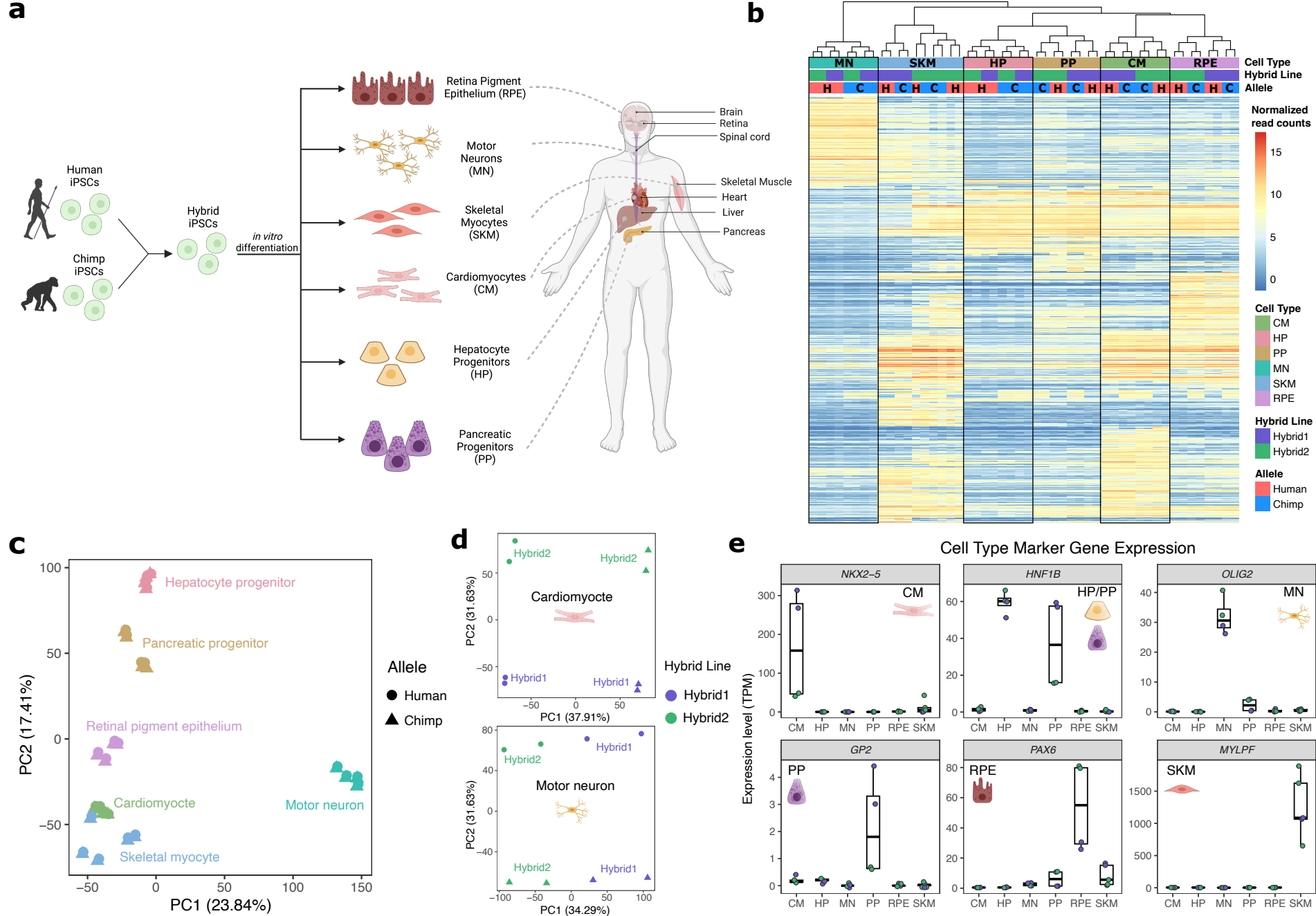

a
Human iPSCs
Chimp iPSCs
Hybrid iPSCs
in vitro differentiation
Retina Pigment Epithelium (RPE)
Motor Neurons (MN)
Skeletal Myocytes (SKM)
Cardiomyocytes (CM)
Hepatocyte Progenitors (HP)
Pancreatic Progenitors (PP)
Brain
Retina
Spinal cord
Skeletal Muscle
Heart
Liver
Pancreas
b
MN
SKM
HP
PP
CM
RPE
H
C
Cell Type
Hybrid Line
Allele
Normalized read counts
15
10
5
0
Cell Type
CM
HP
PP
MN
SKM
RPE
Hybrid Line
Hybrid1
Hybrid2
Allele
Human
Chimp
c
Hepatocyte progenitor
Pancreatic progenitor
Retinal pigment epithelium
Cardiomyocte
Motor neuron
Skeletal myocyte
PC1 (23.84%)
PC2 (17.41%)
Allele
Human
Chimp
d
Cardiomyocte
Motor neuron
Hybrid2
Hybrid1
PC1 (37.91%)
PC2 (31.63%)
PC1 (34.29%)
PC2 (31.63%)
Hybrid Line
Hybrid1
Hybrid2
e
Cell Type Marker Gene Expression
NKX2–5
CM
HNF1B
HP/PP
OLIG2
MN
GP2
PP
PAX6
RPE
MYLPF
SKM
Expression level (TPM)
CM
HP
MN
PP
RPE
SKM

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.

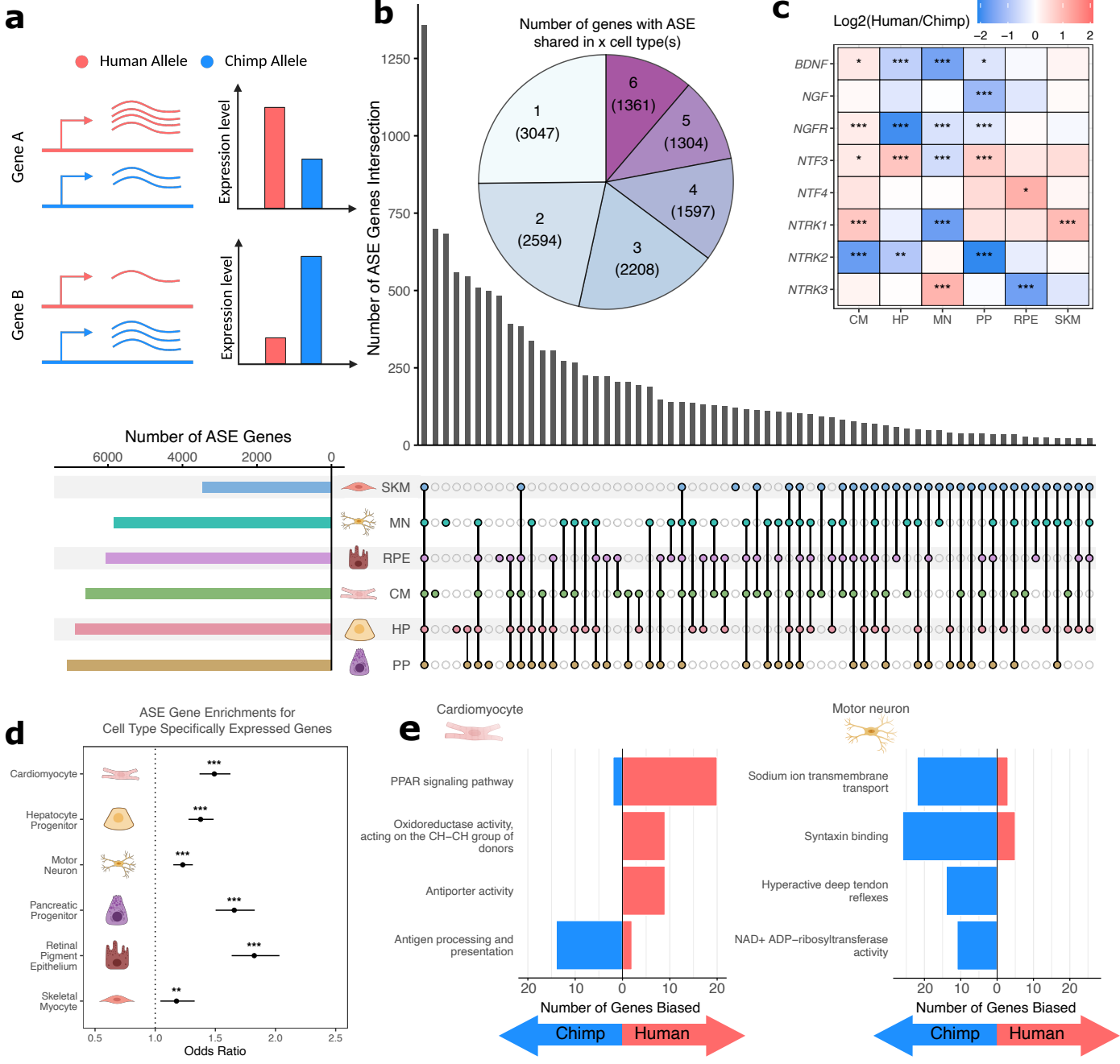
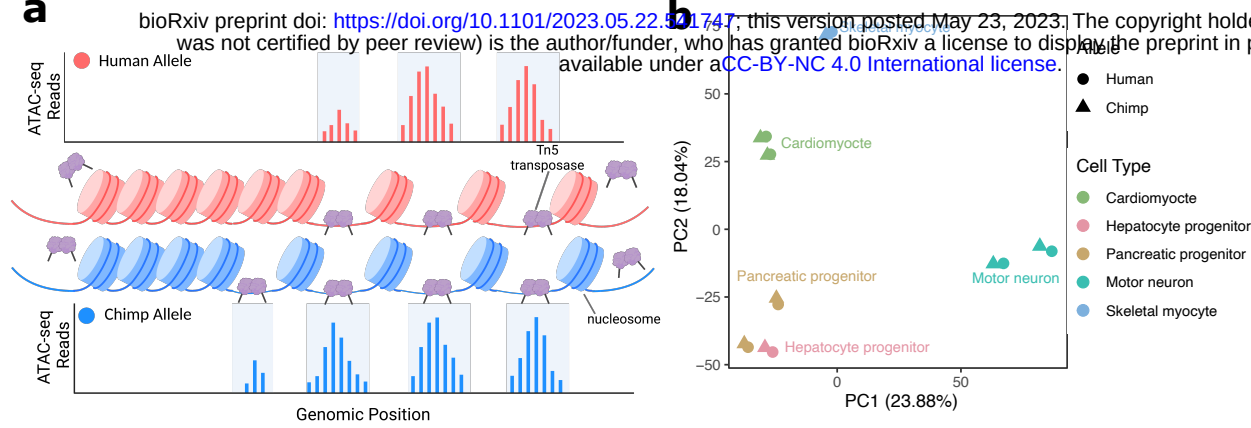
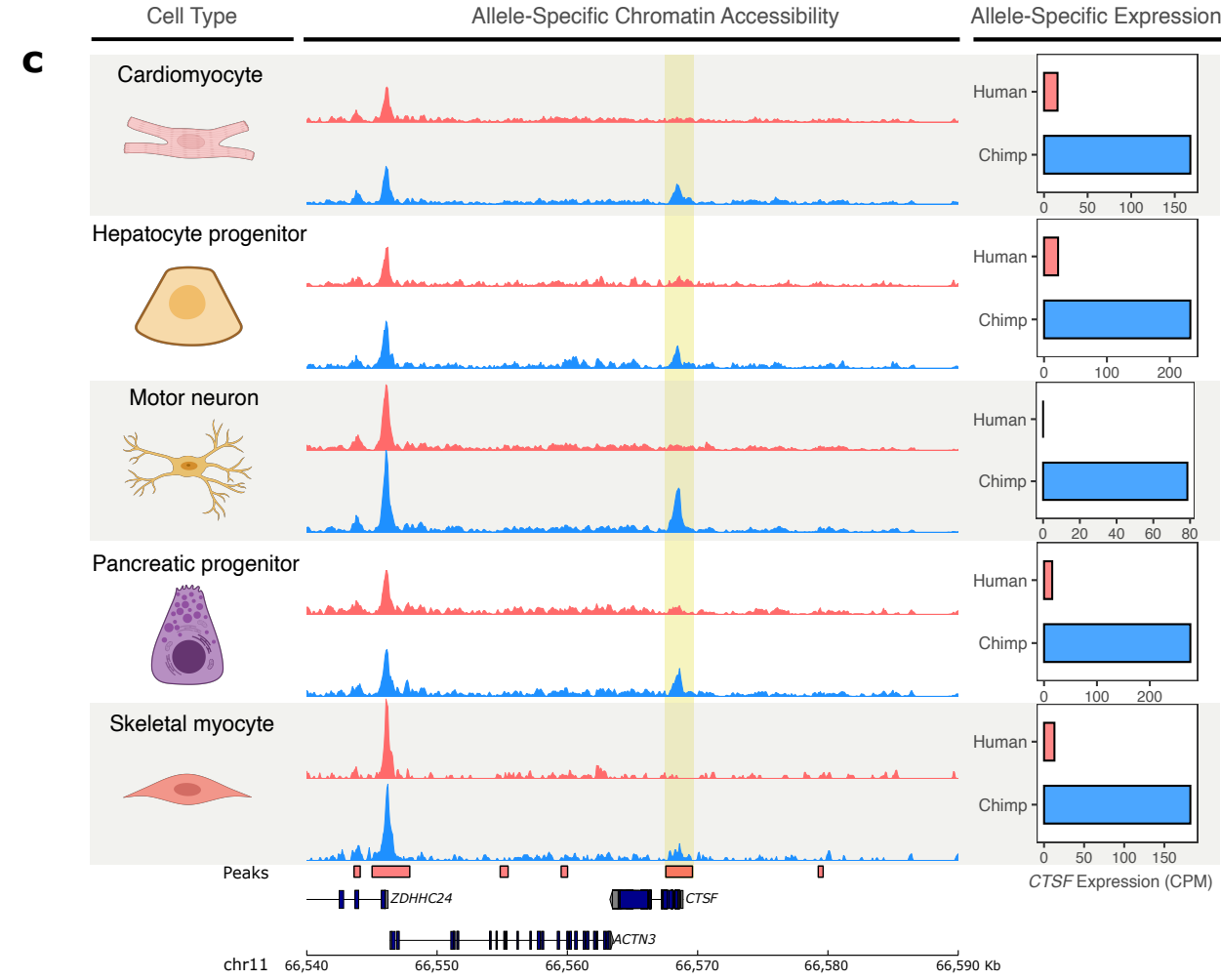

a
Human Allele
Chimp Allele
Gene A
Gene B
Expression level
Expression level
b
Number of genes with ASE shared in x cell type(s)
1 (3047)
2 (2594)
3 (2208)
4 (1597)
5 (1304)
6 (1361)
Number of ASE Genes Intersection
1250
1000
750
500
250
0
Number of ASE Genes
6000
4000
2000
0
SKM
MN
RPE
CM
HP
PP
c
Log2(Human/Chimp)
−2 −1 0 1 2
BDNF
NGF
NGFR
NTF3
NTF4
NTRK1
NTRK2
NTRK3
CM
HP
MN
PP
RPE
SKM
d
ASE Gene Enrichments for Cell Type Specifically Expressed Genes
Cardiomyocyte
Hepatocyte Progenitor
Motor Neuron
Pancreatic Progenitor
Retinal Pigment Epithelium
Skeletal Myocyte
0.5
1.0
1.5
2.0
2.5
Odds Ratio
e
Cardiomyocyte
PPAR signaling pathway
Oxidoreductase activity, acting on the CH−CH group of donors
Antiporter activity
Antigen processing and presentation
20 10 0 10 20
Number of Genes Biased
Chimp
Human
Motor neuron
Sodium ion transmembrane transport
Syntaxin binding
Hyperactive deep tendon reflexes
NAD+ ADP−ribosyltransferase activity
20 10 0 10 20
Number of Genes Biased
Chimp
Human

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 $\log_2$ 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.





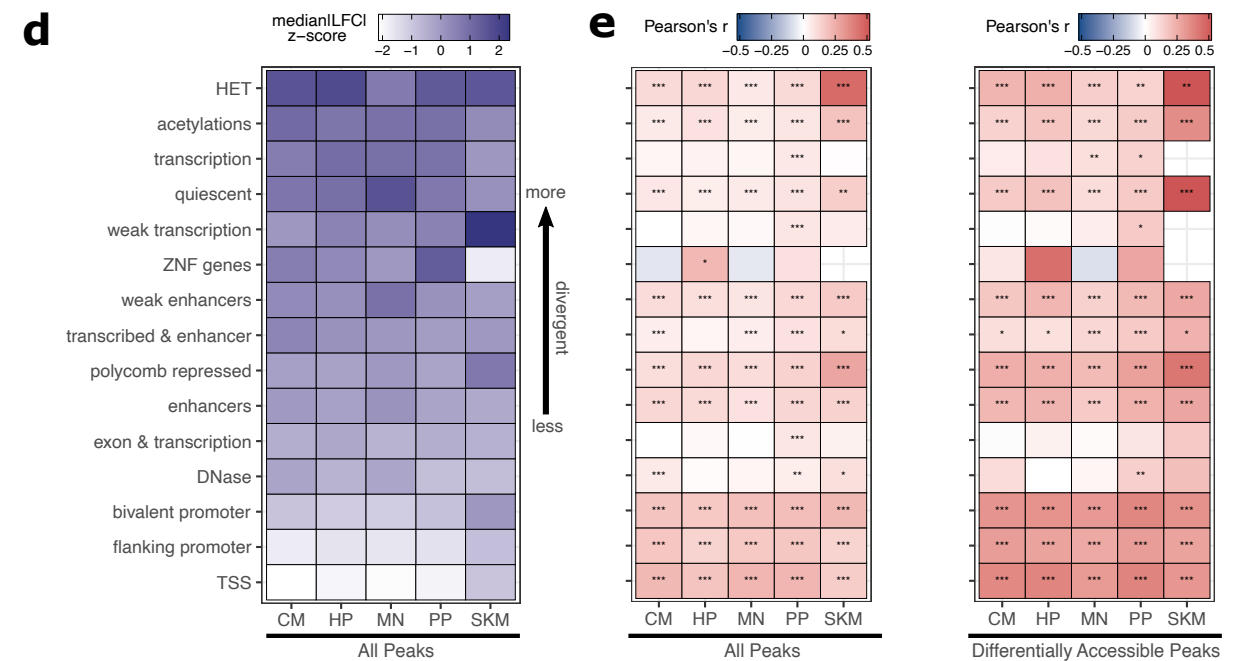


Figure 3: Allele-specific chromatin accessibility across diverse human-chimpanzee hybrid cell types. a) Schematic outlining the ATAC-seq protocol. A hyperactive transposase cleaves accessible DNA and adds adapters enabling measurement of chromatin accessibility. **b)** PCA on normalized allelic counts 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 $p > 0.05$.

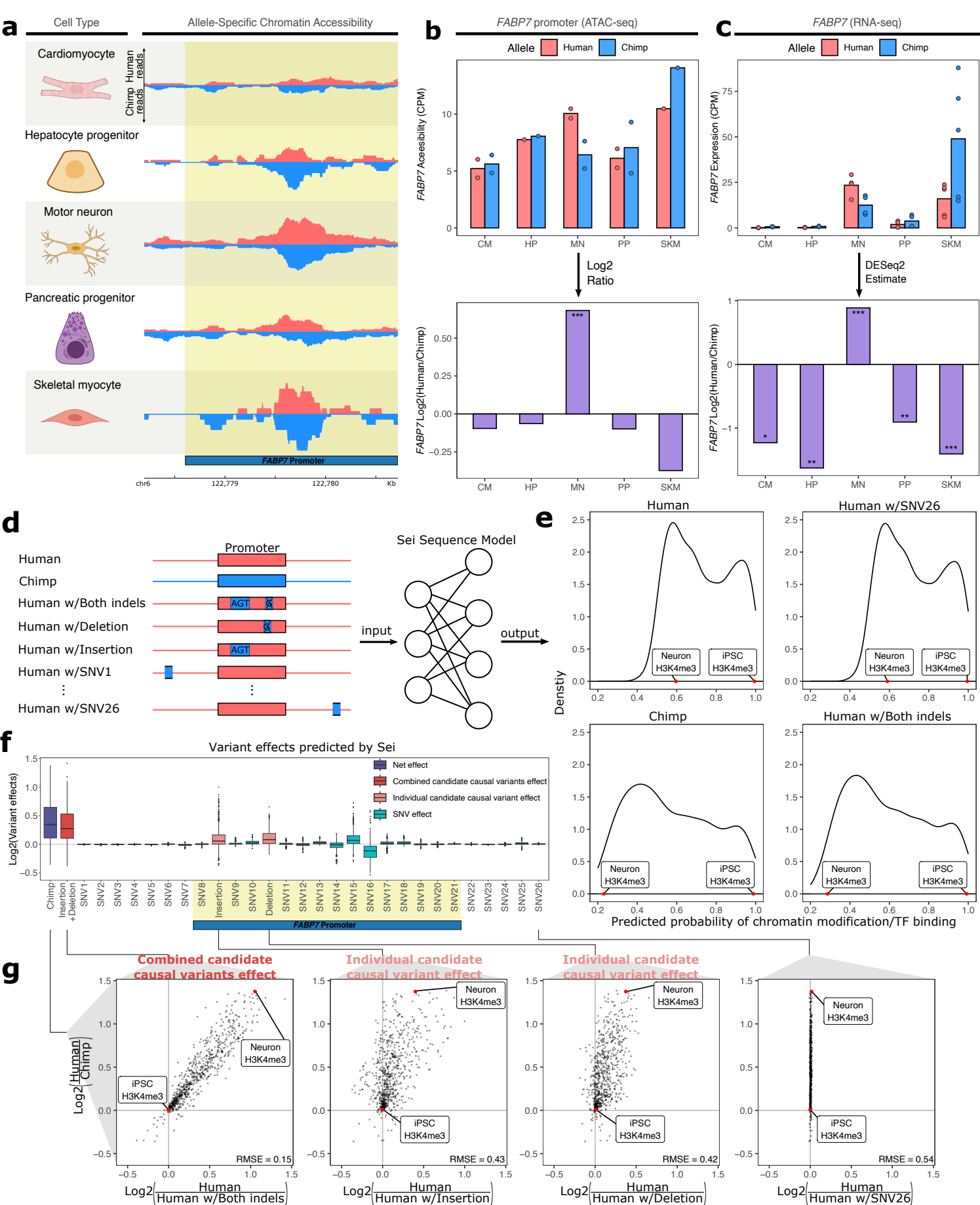

a
Cell Type
Allele-Specific Chromatin Accessibility
Cardiomyocyte
Hepatocyte progenitor
Motor neuron
Pancreatic progenitor
Skeletal myocyte
Human reads
Chimp reads
FABP7 Promoter
chr6
122,779
122,780
Kb
b
FABP7 promoter (ATAC-seq)
Allele Human Chimp
FABP7 Acessibility (CPM)
CM HP MN PP SKM
Log2 Ratio
FABP7 Log2(Human/Chimp)
c
FABP7 (RNA-seq)
FABP7 Expression (CPM)
DESeq2 Estimate
d
Promoter
Human
Chimp
Human w/Both indels
Human w/Deletion
Human w/Insertion
Human w/SNV1
Human w/SNV26
AGT
input
Sei Sequence Model
output
e
Human
Human w/SNV26
Chimp
Human w/Both indels
Denstiy
Neuron H3K4me3
iPSC H3K4me3
Predicted probability of chromatin modification/TF binding
f
Variant effects predicted by Sei
Log2(Variant effects)
Net effect
Combined candidate causal variants effect
Individual candidate causal variant effect
SNV effect
Chimp
Insertion
+Deletion
SNV1
SNV2
SNV3
SNV4
SNV5
SNV6
SNV7
SNV8
Insertion
SNV9
SNV10
Deletion
SNV11
SNV12
SNV13
SNV14
SNV15
SNV16
SNV17
SNV18
SNV19
SNV20
SNV21
SNV22
SNV23
SNV24
SNV25
SNV26
FABP7 Promoter
g
Combined candidate causal variants effect
Individual candidate causal variant effect
Individual candidate causal variant effect
Neuron H3K4me3
iPSC H3K4me3
RMSE = 0.15
RMSE = 0.43
RMSE = 0.42
RMSE = 0.54
Log2(Human/Chimp)
Log2(Human/Human w/Both indels)
Log2(Human/Human w/Insertion)
Log2(Human/Human w/Deletion)
Log2(Human/Human w/SNV26)

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.

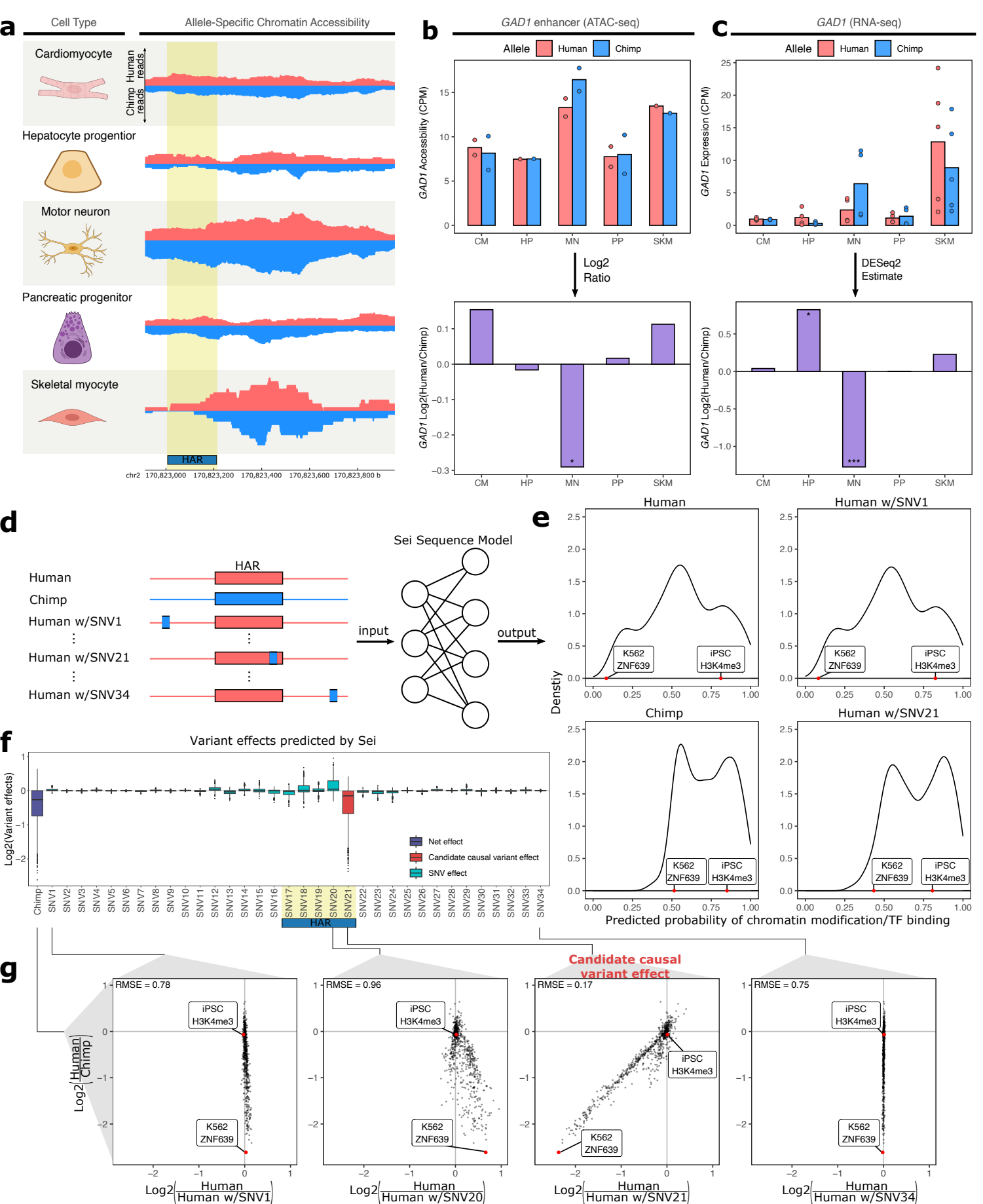

a
Cell Type
Allele-Specific Chromatin Accessibility
Cardiomyocyte
Hepatocyte progentior
Motor neuron
Pancreatic progenitor
Skeletal myocyte
Human reads
Chimp reads
HAR
chr2 170,823,000 170,823,200 170,823,400 170,823,600 170,823,800 b
b
GAD1 enhancer (ATAC-seq)
Allele Human Chimp
GAD1 Accessibility (CPM)
CM HP MN PP SKM
Log2 Ratio
GAD1 Log2(Human/Chimp)
c
GAD1 (RNA-seq)
GAD1 Expression (CPM)
DESeq2 Estimate
d
Human
Chimp
Human w/SNV1
Human w/SNV21
Human w/SNV34
HAR
input
Sei Sequence Model
output
e
Human
Human w/SNV1
Chimp
Human w/SNV21
Densty
K562 ZNF639
iPSC H3K4me3
Predicted probability of chromatin modification/TF binding
f
Variant effects predicted by Sei
Log2(Variant effects)
Net effect
Candidate causal variant effect
SNV effect
Chimp SNV1 SNV2 SNV3 SNV4 SNV5 SNV6 SNV7 SNV8 SNV9 SNV10 SNV11 SNV12 SNV13 SNV14 SNV15 SNV16 SNV17 SNV18 SNV19 SNV20 SNV21 SNV22 SNV23 SNV24 SNV25 SNV26 SNV27 SNV28 SNV29 SNV30 SNV31 SNV32 SNV33 SNV34
HAR
g
Candidate causal variant effect
RMSE = 0.78
RMSE = 0.96
RMSE = 0.17
RMSE = 0.75
Log2(Human/Chimp)
Log2(Human/Human w/SNV1)
Log2(Human/Human w/SNV20)
Log2(Human/Human w/SNV21)
Log2(Human/Human w/SNV34)

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.

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 3 motor neuron libraries (2 hybrid2 and 1 hybrid1) 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. The other three motor neuron libraries were prepared using Illumina Stranded mRNA Prep (Illumina, 20040532) and IDT for Illumina RNA UD Indexes Set A, Ligation (Illumina, 20040553) due to low yield of

548 total RNA. Notably, the four motor neuron libraries did not cluster by library preparation method. All
549 libraries were normalized, pooled at an equimolar ratio using Qubit measurements, and sequenced on
550 an Illumina HiSeq 4000 to generate 2x150bp paired-end reads.
551
552 **Identification of confident human-chimpanzee SNVs**
553 To identify a confident list of human-chimpanzee SNVs that could be used to quantify allele-specific
554 expression and chromatin accessibility, we first downloaded hg38-panTro6 MAF files from UCSC and
555 whole-genome sequencing data generated from the parental human and chimpanzee iPS cells (in the
556 form of bam files aligned to hg38 and panTro5, generously provided by the Gilad lab). We first
557 converted them back to fastq files and then mapped reads to panTro6 and hg38 (we mapped both
558 human and chimpanzee to both reference genomes) using bowtie2 with the flags —very-sensitive-local -
559 p 16 (Langmead & Salzberg, 2012). We then used a modified version of our previous approach to filter
560 out SNVs that could not be confidently identified as homozygous in the human and chimpanzee parental
561 lines (Agoglia et al., 2021). Briefly, we extracted SNVs and indels from both human and chimpanzee MAF
562 files, counted reads in the WGS data that supported the human, chimpanzee, or an alternative base at
563 that position, then filtered out any SNVs with < 2 reads or < 90% of reads supporting that species' base.
564 We then reformatted files, merged with indels for use in Hornet, and generated a modified bed file of
565 SNVs that includes the human and chimpanzee base at the SNV position (van de Geijn et al., 2015).
566
567 **Generation of allele-specific count tables**
568 An allele specific expression mapping and calling pipeline adapted from Agoglia et al. and our updated
569 high-confidence SNV list was used. The whole pipeline was carried out twice independently using hg38
570 or panTro6 as the reference genome. This approach was taken to eliminate genes showing strong
571 mapping bias, defined here as genes with an absolute difference in log 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 log 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 took hybrid and chimpanzee parental iPSC RNA-seq from Agoglia et al (Agoglia et al., 2021)

596 from the same iPS cell lines as used in this study, and then simulated chimpanzee parental
597 contamination by mixing chimpanzee parental data into hybrid 2 data to reach similar a similar degree
598 of chimpanzee bias level to that observed in the PP hybrid2 samples. We then identified genes showing
599 ASE (see "Identifying genes with ASE") using the counts from the original hybrid samples, simulated
600 contaminated samples, and corrected simulated contaminated samples respectively and compared the
601 outputs (Supp Fig. 4).
602
603 **PCA and hierarchical clustering**
604 Allelic counts were normalized by DESeq2 rlog and principal components analysis (PCA) was performed
605 on rlog normalized allelic counts with default centering and scaling (Love et al., 2014). The top 1,000
606 variable genes with the highest variance of normalized allelic counts across all cell types were used to
607 compute Euclidean distance matrices. The R package pheatmap was used to do hierarchical clustering
608 and heatmap plotting.
609
610 **Identifying genes with ASE**
611 DESeq2 was used to measure allele-specific expression (ASE) in each cell type (Love et al., 2014). All
612 reads from chromosome 20 were removed (as mentioned above). Two replicates per hybrid line per cell
613 type (plus one additional replicate for SKM hybrid2 for a total of 3 samples) were used by DEseq2 with
614 model ~hybLine+Species to measure differential expression level. A likelihood ratio test (test="LRT",
615 betaPrior=FALSE) was used to compute p-values. P-values were then false discovery rate adjusted using
616 an implementation of the Benjamini-Hochberg correction in the R package qvalue (Benjamini &
617 Hochberg, 1995; Storey Lab, n.d.). Log fold-changes were shrunk as recommended by the DESeq2
618 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

642 Odds ratios were calculated using the unconditional maximum likelihood estimate implemented in the R
643 package epitools function oddsratio.wald (), and 95% confidence intervals and p-values were calculated
644 using the normal approximation. A directly analogous procedure was performed to test for enrichment
645 of peaks with ASCA and cell type-specific peaks.
646
647 **Enrichment test stratified by expression level or evolutionary constraint**
648 Enrichment tests were carried out as in 'Enrichment test of cell type-specific expression patterns and
649 genes showing ASE' except that genes were split into five equal size bins depending on which factors
650 were used to stratify genes, and tests were done in each bin. When stratifying by expression level, genes
651 were ordered in ascending order based on expression level (TPM) and then split into 5 equal size bins
652 where genes in the 0-20% bin are the most lowly expressed genes and genes in the 80-100% bin are the
653 most highly expressed genes. When stratifying by constraint metrics such as ASE variance or pHI, genes
654 were ordered in ascending order based on ASE variance values and then split into five equal size bins
655 where the 0-20% bin contains genes with the lowest ASE variance (i.e. most evolutionarily constrained)
656 and the 80-100% bin contains genes with highest ASE variance (i.e. least evolutionarily constrained). To
657 stratify the ATAC data by constraint, we used the "QCed genomic constraint by 1kb regions" computed
658 by the gnomAD consortium (S. Chen et al., 2022). We further removed any regions that overlapped
659 protein coding exons from the human gtf file using bedtools subtract (Quinlan & Hall, 2010). If a peak
660 overlapped two or more 1 kilobase windows, it was assigned to the window with the highest constraint,
661 mirroring the procedure used by the gnomAD consortium to assign peaks to ENCODE regulatory
662 elements (S. Chen et al., 2022). Once the peaks were ranked by this metric, a procedure identical to that
663 for the gene expression constraint metrics outlined above was performed.
664
665 **Identification of lineage-specific selection on gene expression**

We used 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 (with the exceptions of RPE which was compared to all GTEx samples and MN which was compared to all brain and peripheral nerve samples as more closely matches tissues such as eyes for RPE are not available in GTEx). 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 then used 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). 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).

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

690 resulted in 2 CM libraries, 2 MN libraries, 2 PP libraries, 1 SKM library, and 1 HP library. After pooling,
691 libraries were sequenced on an Illumina Hiseq 4000 to produce 2x150 paired-end reads.
692
693 **Mapping the ATAC-seq data**
694 We trimmed reads using SeqPrep and then mapped them to the hg38 and panTro6 reference genomes
695 with bowtie2 in paired-end mode (John St. John, n.d.; Langmead & Salzberg, 2012). The following
696 parameters were used: -X 2000 --very-sensitive-local -p 16. After mapping, duplicates were removed via
697 Picard MarkDuplicates (Broad Institute, n.d.). We then removed multi-mapping reads with the command
698 samtools view -b -q 10 (Li et al., 2009). Due to the format of bowtie2's output, running Hornet on all
699 reads at once was excessively RAM intensive. Therefore, we split the bam files by chromosome and ran
700 Hornet on each of the chromosomes separately. We used the files of SNVs and indels generated as
701 described above as input to Hornet. After Hornet finished running, we used samtools merge to merge all
702 autosomes and sex chromosomes (we excluded the mitochondrial genome) to create a final bam file for
703 downstream analysis (Li et al., 2009).
704
705 **Peak calling and filtering**
706 As only one replicate was available for SKM and HP, we generated two pseudo-replicates by randomly
707 assigning reads to one of two files using Picard SplitSamByNumberOfReads (Broad Institute, n.d.). We
708 then called peaks on each file separately, as well as a merged file containing all the reads from a
709 particular cell type. For example, for MN, both replicates were pooled and peaks were called on that file
710 as well as the two replicates separately. Before peak calling, all bam files were converted to bed files.
711 We called peaks using MACS2 callpeak with the following arguments: -f BED -p 0.01 --nomodel --shift 75
712 --extsize 150 -B --SPMR --keep-dup all --call-summits (Y. Zhang et al., 2008, p. 2). We called peaks on
713 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 LiftOver (Kuhn et al., 2013).

Annotating the peak lists

To annotate the peaks, we used the list of TSS 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 outputted a list of non-promoter CREs (sometimes labeled as enhancers as enhancers are thought to be the most common type of CRE). We took this list and used bedtools closest to link them 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

738 altered all gene names in peaks to match those found in the GTF file. In some cases, the gene no longer
739 existed in the updated hg38 GTF in which case the gene name was replaced with NAN.
740
741 **Generating a unified peak list**
742 We next merged our cell type-specific peak list across all five cell types to create a unified peak list. To
743 do this, we iteratively intersected all the peaks with bedtools and then merged any overlapping peaks
744 (Quinlan & Hall, 2010). Finally, we added back any peaks that did not intersect a peak found in any other
745 cell types. We then took the chimpanzee and human-referenced versions of these peak lists and ran
746 them through the LiftOver-based non-homologous peak filtering pipeline described above to generate a
747 final file of all identified peaks as well as which cell type (s) they were called in (Kuhn et al., 2013). Then,
748 we reran the annotation pipeline described in 'Annotating the peak lists' on this new set of peaks.
749
750 **Counting reads in peaks and further peak filtering**
751 First, we split the bam files into reads that we could confidently assign to the chimpanzee genome and
752 reads we could assign to the human genome. We used our bed file of high-confidence SNVs and
753 required at least one SNV matching the human genome as well as no SNVs matching the chimpanzee
754 genome for a read to be assigned to human (and vice versa for chimp). We then used a custom Python
755 script to reformat the peak list bed files as GTF files and used HTSeq to count reads in peaks using the
756 following parameters: -s no -m union -r pos (Anders et al., 2015). We only kept peaks if they had a mean
757 read count across replicates within a cell type of at least 25 from either allele. For example, if a peak has
758 an average of 27 reads from the human allele and an average of 10 reads from the chimpanzee allele in
759 MN, that peak would be kept in MN. On the other hand, if the same peak had an average of 24 reads
760 from the human allele and an average of 10 reads from the chimpanzee allele in CM, that peak would be
761 discarded for CM.

762

763 We next filtered the reads to remove peaks that might be differentially accessible but show evidence of
764 mapping bias or do not agree between replicates. To do this, we removed any peaks with an absolute
765 log fold-change greater than one in one replicate but with a fold-change of any magnitude in the
766 opposite direction in the other. This was not done for SKM or HP as we had only one replicate. We then
767 removed any peaks that had a log fold-change in opposite directions with an absolute difference greater
768 than 1 in at least one replicate when comparing the human-referenced and chimpanzee-referenced
769 counts. Finally, as described in section ‘‘Detection of aneuploidy on chromosome 20 and slight
770 chimpanzee parental contamination in PP hybrid2 samples’ for RNA-seq data analysis, we removed any
771 peaks on chr20 and took this as our final list of peaks for downstream analyses. Allelic counts were
772 normalized as described in the RNA-seq data analysis. We tested for allele-specific chromatin
773 accessibility (ASCA) using the binomial test applied to the normalized allelic counts (summed by species
774 within a cell type). We considered any peaks with a binomial p-value less than 0.05 to be nominally
775 differentially accessible.

776

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

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

782

783 **Allelic chromatin accessibility tracks**

784 Allelic bam files with reads originating from the human allele and the chimpanzee allele (respectively)
785 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 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 log 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 log fold-change and the expression log 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.

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

For each target cell type, taking CM as an example, the log fold-change for gene A was fixed as target log fold-change, and the log fold-changes for gene A in the remaining cell types with an opposing sign (compared to the target log fold-change) were set to zero. Then, the dEE value was calculated as the proportion of the target log 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 log 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 log 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. 25b). 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 log 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, log 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 log 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 log 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 to integrate cells from different species (human and chimpanzee), and found nearest neighbors with the harmonized principal components (Korsunsky et al., 2019). 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 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 genomes used 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 by 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. BioRender was used to generate Figures 1a, 2a, and 3a.

938 Agoglia, R. M., Sun, D., Birey, F., Yoon, S.-J., Miura, Y., Sabatini, K., Pașca, S. P., & Fraser, H.
939 B. (2021). Primate cell fusion disentangles gene regulatory divergence in
940 neurodevelopment. *Nature*, *592* (7854), 421–427. https://doi.org/10.1038/s41586-021-
941 03343-3
942 Akbar, M., Calderon, F., Wen, Z., & Kim, H.-Y. (2005). Docosahexaenoic acid: A positive
943 modulator of Akt signaling in neuronal survival. *Proceedings of the National Academy of*
944 *Sciences*, *102* (31), 10858–10863. https://doi.org/10.1073/pnas.0502903102
945 Amemiya, H. M., Kundaje, A., & Boyle, A. P. (2019). The ENCODE Blacklist: Identification of
946 Problematic Regions of the Genome. *Scientific Reports*, *9* (1), 9354.
947 https://doi.org/10.1038/s41598-019-45839-z
948 Anders, S., Pyl, P. T., & Huber, W. (2015). HTSeq—A Python framework to work with high-
949 throughput sequencing data. *Bioinformatics (Oxford, England)*, *31* (2), 166–169.
950 https://doi.org/10.1093/bioinformatics/btu638
951 Arai, Y., Funatsu, N., Numayama-Tsuruta, K., Nomura, T., Nakamura, S., & Osumi, N. (2005).
952 Role of Fabp7, a downstream gene of Pax6, in the maintenance of neuroepithelial cells
953 during early embryonic development of the rat cortex. *The Journal of Neuroscience: The*
954 *Official Journal of the Society for Neuroscience*, *25* (42), 9752–9761.
955 https://doi.org/10.1523/JNEUROSCI.2512-05.2005
956 Ben-Ari, Y., Khalilov, I., Kahle, K. T., & Cherubini, E. (2012). The GABA Excitatory/Inhibitory
957 Shift in Brain Maturation and Neurological Disorders. *The Neuroscientist*, *18* (5), 467–
958 486. https://doi.org/10.1177/1073858412438697
959 Benito-Kwiecinski, S., Giandomenico, S. L., Sutcliffe, M., Riis, E. S., Freire-Pritchett, P., Kelava,
960 I., Wunderlich, S., Martin, U., Wray, G. A., McDole, K., & Lancaster, M. A. (2021). An
961 early cell shape transition drives evolutionary expansion of the human forebrain. *Cell*,
962 *184* (8), 2084-2102.e19. https://doi.org/10.1016/j.cell.2021.02.050

963 Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and
964 Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B*
965 *(Methodological)*, *57* (1), 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x
966 Blake, L. E., Roux, J., Hernando-Herraez, I., Banovich, N. E., Perez, R. G., Hsiao, C. J., Eres, I.,
967 Cuevas, C., Marques-Bonet, T., & Gilad, Y. (2020). A comparison of gene expression
968 and DNA methylation patterns across tissues and species. *Genome Research*, *30* (2),
969 250–262. https://doi.org/10.1101/gr.254904.119
970 Broad Institute. (n.d.). *Picard*.
971 Brose, N., Brunger, A., Cafiso, D., Chapman, E. R., Diao, J., Hughson, F. M., Jackson, M. B.,
972 Jahn, R., Lindau, M., Ma, C., Rizo, J., Shin, Y.-K., Söllner, T. H., Tamm, L., Yoon, T.-Y.,
973 & Zhang, Y. (2019). Synaptic vesicle fusion: Today and beyond. *Nature Structural &*
974 *Molecular Biology*, *26* (8), 663–668. https://doi.org/10.1038/s41594-019-0277-z
975 Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y., & Greenleaf, W. J. (2013).
976 Transposition of native chromatin for fast and sensitive epigenomic profiling of open
977 chromatin, DNA-binding proteins and nucleosome position. *Nature Methods*, *10* (12),
978 1213–1218. https://doi.org/10.1038/nmeth.2688
979 Burridge, P. W., Matsa, E., Shukla, P., Lin, Z. C., Churko, J. M., Ebert, A. D., Lan, F., Diecke,
980 S., Huber, B., Mordwinkin, N. M., Plews, J. R., Abilez, O. J., Cui, B., Gold, J. D., & Wu, J.
981 C. (2014). Chemically defined generation of human cardiomyocytes. *Nature Methods*, *11*
982 (8), 855–860. https://doi.org/10.1038/nmeth.2999
983 Calderari, S., Ria, M., Gérard, C., Nogueira, T. C., Villate, O., Collins, S. C., Neil, H., Gervasi,
984 N., Hue, C., Suarez-Zamorano, N., Prado, C., Cnop, M., Bihoreau, M.-T., Kaisaki, P. J.,
985 Cazier, J.-B., Julier, C., Lathrop, M., Werner, M., Eizirik, D. L., & Gauguier, D. (2018).
986 Molecular genetics of the transcription factor GLIS3 identifies its dual function in beta
987 cells and neurons. *Genomics*, *110* (2), 98–111.
988 https://doi.org/10.1016/j.ygeno.2017.09.001

Caporali, A., & Emanueli, C. (2009). Cardiovascular actions of neurotrophins. *Physiological Reviews*, *89* (1), 279–308. https://doi.org/10.1152/physrev.00007.2008

Castel, S. E., Aguet, F., Mohammadi, P., GTEx Consortium, Aguet, F., Anand, S., Ardlie, K. G., Gabriel, S., Getz, G. A., Graubert, A., Hadley, K., Handsaker, R. E., Huang, K. H., Kashin, S., Li, X., MacArthur, D. G., Meier, S. R., Nedzel, J. L., Nguyen, D. T., … Lappalainen, T. (2020). A vast resource of allelic expression data spanning human tissues. *Genome Biology*, *21* (1), 234. https://doi.org/10.1186/s13059-020-02122-z

Castro-Mondragon, J. A., Riudavets-Puig, R., Rauluseviciute, I., Lemma, R. B., Turchi, L., Blanc-Mathieu, R., Lucas, J., Boddie, P., Khan, A., Manosalva Pérez, N., Fornes, O., Leung, T. Y., Aguirre, A., Hammal, F., Schmelter, D., Baranasic, D., Ballester, B., Sandelin, A., Lenhard, B., … Mathelier, A. (2022). JASPAR 2022: The 9th release of the open-access database of transcription factor binding profiles. *Nucleic Acids Research*, *50* (D1), 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* (10), 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* (7), 940–949. https://doi.org/10.1038/s41588-022-01102-2

Chen, S., Francioli, L. C., Goodrich, J. K., Collins, R. L., Kanai, M., Wang, Q., Alföldi, J., Watts, N. A., Vittal, C., Gauthier, L. D., Poterba, T., Wilson, M. W., Tarasova, Y., Phu, W., Yohannes, M. T., Koenig, Z., Farjoun, Y., Banks, E., Donnelly, S., … Karczewski, K. J. (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

1014 Choi, W.-S., Xu, X., Goruk, S., Wang, Y., Patel, S., Chow, M., Field, C. J., & Godbout, R.
1015 (2021). FABP7 Facilitates Uptake of Docosahexaenoic Acid in Glioblastoma Neural
1016 Stem-like Cells. *Nutrients*, *13* (8), 2664. https://doi.org/10.3390/nu13082664
1017 Collins, R. L., Glessner, J. T., Porcu, E., Lepamets, M., Brandon, R., Lauricella, C., Han, L.,
1018 Morley, T., Niestroj, L.-M., Ulirsch, J., Everett, S., Howrigan, D. P., Boone, P. M., Fu, J.,
1019 Karczewski, K. J., Kellaris, G., Lowther, C., Lucente, D., Mohajeri, K., … Talkowski, M.
1020 E. (2022). A cross-disorder dosage sensitivity map of the human genome. *Cell*, *185* (16),
1021 3041-3055.e25. https://doi.org/10.1016/j.cell.2022.06.036
1022 Combs, P. A., Krupp, J. J., Khosla, N. M., Bua, D., Petrov, D. A., Levine, J. D., & Fraser, H. B.
1023 (2018). Tissue-Specific cis-Regulatory Divergence Implicates eloF in Inhibiting
1024 Interspecies Mating in Drosophila. *Current Biology*, *28* (24), 3969-3975.e3.
1025 https://doi.org/10.1016/j.cub.2018.10.036
1026 Conway, J. R., Lex, A., & Gehlenborg, N. (2017). UpSetR: An R package for the visualization of
1027 intersecting sets and their properties. *Bioinformatics*, *33* (18), 2938–2940.
1028 https://doi.org/10.1093/bioinformatics/btx364
1029 Corces, M. R., Trevino, A. E., Hamilton, E. G., Greenside, P. G., Sinnott-Armstrong, N. A.,
1030 Vesuna, S., Satpathy, A. T., Rubin, A. J., Montine, K. S., Wu, B., Kathiria, A., Cho, S.
1031 W., Mumbach, M. R., Carter, A. C., Kasowski, M., Orloff, L. A., Risca, V. I., Kundaje, A.,
1032 Khavari, P. A., … Chang, H. Y. (2017). An improved ATAC-seq protocol reduces
1033 background and enables interrogation of frozen tissues. *Nature Methods*, *14* (10), 959–
1034 962. https://doi.org/10.1038/nmeth.4396
1035 De Rosa, A., Pellegatta, S., Rossi, M., Tunici, P., Magnoni, L., Speranza, M. C., Malusa, F.,
1036 Miragliotta, V., Mori, E., Finocchiaro, G., & Bakker, A. (2012). A radial glia gene marker,
1037 fatty acid binding protein 7 (FABP7), is involved in proliferation and invasion of
1038 glioblastoma cells. *PloS One*, *7* (12), e52113.
1039 https://doi.org/10.1371/journal.pone.0052113

1040 Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson,
1041 M., & Gingeras, T. R. (2013). STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics*
1042 *(Oxford, England)*, *29* (1), 15–21. https://doi.org/10.1093/bioinformatics/bts635
1043 Ebrahimi, M., Yamamoto, Y., Sharifi, K., Kida, H., Kagawa, Y., Yasumoto, Y., Islam, A.,
1044 Miyazaki, H., Shimamoto, C., Maekawa, M., Mitsushima, D., Yoshikawa, T., & Owada,
1045 Y. (2016). Astrocyte-expressed FABP7 regulates dendritic morphology and excitatory
1046 synaptic function of cortical neurons. *Glia*, *64* (1), 48–62.
1047 https://doi.org/10.1002/glia.22902
1048 Ernst, J., & Kellis, M. (2017). Chromatin-state discovery and genome annotation with
1049 ChromHMM. *Nature Protocols*, *12* (12), 2478–2492.
1050 https://doi.org/10.1038/nprot.2017.124
1051 Feldblum, S., Erlander, M. G., & Tobin, A. J. (1993). Different distributions of GAD65 and
1052 GAD67 mRNAS suggest that the two glutamate decarboxylases play distinctive
1053 functional roles. *Journal of Neuroscience Research*, *34* (6), 689–706.
1054 https://doi.org/10.1002/jnr.490340612
1055 Field, A. R., Jacobs, F. M. J., Fiddes, I. T., Phillips, A. P. R., Reyes-Ortiz, A. M., LaMontagne,
1056 E., Whitehead, L., Meng, V., Rosenkrantz, J. L., Olsen, M., Hauessler, M., Katzman, S.,
1057 Salama, S. R., & Haussler, D. (2019). Structurally Conserved Primate LncRNAs Are
1058 Transiently Expressed during Human Cortical Differentiation and Influence Cell-Type-
1059 Specific Genes. *Stem Cell Reports*, *12* (2), 245–257.
1060 https://doi.org/10.1016/j.stemcr.2018.12.006
1061 Fraser, H. B. (2011). Genome-wide approaches to the study of adaptive gene expression
1062 evolution: Systematic studies of evolutionary adaptations involving gene expression will
1063 allow many fundamental questions in evolutionary biology to be addressed. *BioEssays*,
1064 *33* (6), 469–477. https://doi.org/10.1002/bies.201000094

1065 Fraser, H. B. (2013). Gene expression drives local adaptation in humans. *Genome Research*,
1066 *23* (7), 1089–1096. https://doi.org/10.1101/gr.152710.112
1067 Fraser, H. B. (2019). Improving Estimates of Compensatory cis–trans Regulatory Divergence.
1068 *Trends in Genetics*, *35* (1), 3–5. https://doi.org/10.1016/j.tig.2018.09.003
1069 García-Pérez, R., Esteller-Cucala, P., Mas, G., Lobón, I., Di Carlo, V., Riera, M., Kuhlwilm, M.,
1070 Navarro, A., Blancher, A., Di Croce, L., Gómez-Skarmeta, J. L., Juan, D., & Marquès-
1071 Bonet, T. (2021). Epigenomic profiling of primate lymphoblastoid cell lines reveals the
1072 evolutionary patterns of epigenetic activities in gene regulatory architectures. *Nature*
1073 *Communications*, *12* (1), 3116. https://doi.org/10.1038/s41467-021-23397-1
1074 Girskis, K. M., Stergachis, A. B., DeGennaro, E. M., Doan, R. N., Qian, X., Johnson, M. B.,
1075 Wang, P. P., Sejourne, G. M., Nagy, M. A., Pollina, E. A., Sousa, A. M. M., Shin, T.,
1076 Kenny, C. J., Scotellaro, J. L., Debo, B. M., Gonzalez, D. M., Rento, L. M., Yeh, R. C.,
1077 Song, J. H. T., … Walsh, C. A. (2021). Rewiring of human neurodevelopmental gene
1078 regulatory programs by human accelerated regions. *Neuron*, *109* (20), 3239-3251.e7.
1079 https://doi.org/10.1016/j.neuron.2021.08.005
1080 Gokhman, D., Agoglia, R. M., Kinnebrew, M., Gordon, W., Sun, D., Bajpai, V. K., Naqvi, S.,
1081 Chen, C., Chan, A., Chen, C., Petrov, D. A., Ahituv, N., Zhang, H., Mishina, Y.,
1082 Wysocka, J., Rohatgi, R., & Fraser, H. B. (2021). Human–chimpanzee fused cells reveal
1083 cis-regulatory divergence underlying skeletal evolution. *Nature Genetics*, *53* (4), 467–
1084 476. https://doi.org/10.1038/s41588-021-00804-3
1085 GTEx Consortium. (2017). Genetic effects on gene expression across human tissues. *Nature*,
1086 *550* (7675), 204–213. https://doi.org/10.1038/nature24277
1087 Horlbeck, M. A., Gilbert, L. A., Villalta, J. E., Adamson, B., Pak, R. A., Chen, Y., Fields, A. P.,
1088 Park, C. Y., Corn, J. E., Kampmann, M., & Weissman, J. S. (2016). Compact and highly
1089 active next-generation libraries for CRISPR-mediated gene repression and activation.
1090 *ELife*, *5*, e19760. https://doi.org/10.7554/eLife.19760

1091 Hu, C. K., York, R. A., Metz, H. C., Bedford, N. L., Fraser, H. B., & Hoekstra, H. E. (2022). Cis-
1092 Regulatory changes in locomotor genes are associated with the evolution of burrowing
1093 behavior. *Cell Reports*, *38* (7), 110360. https://doi.org/10.1016/j.celrep.2022.110360
1094 Huang, E. J., & Reichardt, L. F. (2001). Neurotrophins: Roles in neuronal development and
1095 function. *Annual Review of Neuroscience*, *24*, 677–736.
1096 https://doi.org/10.1146/annurev.neuro.24.1.677
1097 Hubisz, M. J., & Pollard, K. S. (2014). Exploring the genesis and functions of Human
1098 Accelerated Regions sheds light on their role in human evolution. *Current Opinion in*
1099 *Genetics & Development*, *29*, 15–21. https://doi.org/10.1016/j.gde.2014.07.005
1100 Ichim, G., Tauszig-Delamasure, S., & Mehlen, P. (2012). Neurotrophins and cell death.
1101 *Experimental Cell Research*, *318* (11), 1221–1228.
1102 https://doi.org/10.1016/j.yexcr.2012.03.006
1103 Jain, A., & Tuteja, G. (2019). TissueEnrich: Tissue-specific gene enrichment analysis.
1104 *Bioinformatics (Oxford, England)*, *35* (11), 1966–1967.
1105 https://doi.org/10.1093/bioinformatics/bty890
1106 John St. John. (n.d.). *SeqPrep*.
1107 Kanton, S., Boyle, M. J., He, Z., Santel, M., Weigert, A., Sanchís-Calleja, F., Guijarro, P., Sidow,
1108 L., Fleck, J. S., Han, D., Qian, Z., Heide, M., Huttner, W. B., Khaitovich, P., Pääbo, S.,
1109 Treutlein, B., & Camp, J. G. (2019). Organoid single-cell genomic atlas uncovers human-
1110 specific features of brain development. *Nature*, *574* (7778), 418–422.
1111 https://doi.org/10.1038/s41586-019-1654-9
1112 Ke, K., Song, Y., Shen, J., Niu, M., Zhang, H., Yuan, D., Ni, H., Zhang, Y., Liu, X., Dai, A., &
1113 Cao, M. (2015). Up-Regulation of Glis2 Involves in Neuronal Apoptosis After
1114 Intracerebral Hemorrhage in Adult Rats. *Cellular and Molecular Neurobiology*, *35* (3),
1115 345–354. https://doi.org/10.1007/s10571-014-0130-1

1116 Kelley, J. L., & Gilad, Y. (2020). Effective study design for comparative functional genomics.
1117 *Nature Reviews Genetics*, *21* (7), 385–386. https://doi.org/10.1038/s41576-020-0242-z
1118 King, M.-C., & Wilson, A. C. (1975). Evolution at Two Levels in Humans and Chimpanzees:
1119 Their macromolecules are so alike that regulatory mutations may account for their
1120 biological differences. *Science*, *188* (4184), 107–116.
1121 https://doi.org/10.1126/science.1090005
1122 Korsunsky, I., Millard, N., Fan, J., Slowikowski, K., Zhang, F., Wei, K., Baglaenko, Y., Brenner,
1123 M., Loh, P., & Raychaudhuri, S. (2019). Fast, sensitive and accurate integration of
1124 single-cell data with Harmony. *Nature Methods*, *16* (12), 1289–1296.
1125 https://doi.org/10.1038/s41592-019-0619-0
1126 Korytnikov, R., & Nostro, M. C. (2016). Generation of polyhormonal and multipotent pancreatic
1127 progenitor lineages from human pluripotent stem cells. *Methods*, *101*, 56–64.
1128 https://doi.org/10.1016/j.ymeth.2015.10.017
1129 Kozlenkov, A., Vermunt, M. W., Apontes, P., Li, J., Hao, K., Sherwood, C. C., Hof, P. R., Ely, J.
1130 J., Wegner, M., Mukamel, E. A., Creyghton, M. P., Koonin, E. V., & Dracheva, S. (2020).
1131 Evolution of regulatory signatures in primate cortical neurons at cell-type resolution.
1132 *Proceedings of the National Academy of Sciences*, *117* (45), 28422–28432.
1133 https://doi.org/10.1073/pnas.2011884117
1134 Kuhn, R. M., Haussler, D., & Kent, W. J. (2013). The UCSC genome browser and associated
1135 tools. *Briefings in Bioinformatics*, *14* (2), 144–161. https://doi.org/10.1093/bib/bbs038
1136 Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature*
1137 *Methods*, *9* (4), 357–359. https://doi.org/10.1038/nmeth.1923
1138 Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G.,
1139 Durbin, R., & 1000 Genome Project Data Processing Subgroup. (2009). The Sequence
1140 Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England)*, *25* (16), 2078–
1141 2079. https://doi.org/10.1093/bioinformatics/btp352

1142 Liang, D., Elwell, A. L., Aygün, N., Krupa, O., Wolter, J. M., Kyere, F. A., Lafferty, M. J., Cheek,
1143 K. E., Courtney, K. P., Yusupova, M., Garrett, M. E., Ashley-Koch, A., Crawford, G. E.,
1144 Love, M. I., de la Torre-Ubieta, L., Geschwind, D. H., & Stein, J. L. (2021). Cell-type-
1145 specific effects of genetic variation on chromatin accessibility during human neuronal
1146 differentiation. *Nature Neuroscience*, *24* (7), 941–953. https://doi.org/10.1038/s41593-
1147 021-00858-w
1148 Lopez-Delisle, L., Rabbani, L., Wolff, J., Bhardwaj, V., Backofen, R., Grüning, B., Ramírez, F., &
1149 Manke, T. (2021). pyGenomeTracks: Reproducible plots for multivariate genomic
1150 datasets. *Bioinformatics (Oxford, England)*, *37* (3), 422–423.
1151 https://doi.org/10.1093/bioinformatics/btaa692
1152 Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and
1153 dispersion for RNA-seq data with DESeq2. *Genome Biology*, *15* (12), 550.
1154 https://doi.org/10.1186/s13059-014-0550-8
1155 Ma, S., Skarica, M., Li, Q., Xu, C., Risgaard, R. D., Tebbenkamp, A. T. N., Mato-Blanco, X.,
1156 Kovner, R., Krsnik, Ž., de Martin, X., Luria, V., Martí-Pérez, X., Liang, D., Karger, A.,
1157 Schmidt, D. K., Gomez-Sanchez, Z., Qi, C., Gobeske, K. T., Pochareddy, S., … Sestan,
1158 N. (2022). Molecular and cellular evolution of the primate dorsolateral prefrontal cortex.
1159 *Science*, eabo7257. https://doi.org/10.1126/science.abo7257
1160 Mack, K. L., & Nachman, M. W. (2017). Gene Regulation and Speciation. *Trends in Genetics*,
1161 *33* (1), 68–80. https://doi.org/10.1016/j.tig.2016.11.003
1162 Mallanna, S. K., & Duncan, S. A. (2013). Differentiation of hepatocytes from pluripotent stem
1163 cells. *Current Protocols in Stem Cell Biology*, *26*, 1G.4.1-1G.4.13.
1164 https://doi.org/10.1002/9780470151808.sc01g04s26
1165 Maury, Y., Côme, J., Piskorowski, R. A., Salah-Mohellibi, N., Chevaleyre, V., Peschanski, M.,
1166 Martinat, C., & Nedelec, S. (2015). Combinatorial analysis of developmental cues

efficiently converts human pluripotent stem cells into multiple neuronal subtypes. *Nature Biotechnology*, *33* (1), 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* (3), 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*, 25474. 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* (6), 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* (12), 809–823. https://doi.org/10.1038/s41569-021-00569-6

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

Netherlands Brain Bank, Vermunt, M. W., Tan, S. C., Castelijns, B., Geeven, G., Reinink, P., de Bruijn, E., Kondova, I., Persengiev, S., Bontrop, R., Cuppen, E., de Laat, W., & Creyghton, M. P. (2016). Epigenomic annotation of gene regulatory alterations during evolution of the primate brain. *Nature Neuroscience*, *19* (3), 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

1193 Heart Tissues. *Scientific Reports*, *8* (1), 15312. https://doi.org/10.1038/s41598-018-
1194 33478-9
1195 Pollard, K. S., Salama, S. R., Lambert, N., Lambot, M.-A., Coppens, S., Pedersen, J. S.,
1196 Katzman, S., King, B., Onodera, C., Siepel, A., Kern, A. D., Dehay, C., Igel, H., Ares, M.,
1197 Vanderhaeghen, P., & Haussler, D. (2006). An RNA gene expressed during cortical
1198 development evolved rapidly in humans. *Nature*, *443* (7108), 167–172.
1199 https://doi.org/10.1038/nature05113
1200 Prud'homme, B., Gompel, N., & Carroll, S. B. (2007). Emerging principles of regulatory
1201 evolution. *Proceedings of the National Academy of Sciences*, *104* (suppl_1), 8605–8612.
1202 https://doi.org/10.1073/pnas.0700488104
1203 Quinlan, A. R., & Hall, I. M. (2010). BEDTools: A flexible suite of utilities for comparing genomic
1204 features. *Bioinformatics (Oxford, England)*, *26* (6), 841–842.
1205 https://doi.org/10.1093/bioinformatics/btq033
1206 Ramírez, F., Dündar, F., Diehl, S., Grüning, B. A., & Manke, T. (2014). deepTools: A flexible
1207 platform for exploring deep-sequencing data. *Nucleic Acids Research*, *42* (Web Server
1208 issue), W187-191. https://doi.org/10.1093/nar/gku365
1209 Reilly, S. K., & Noonan, J. P. (2016). Evolution of Gene Regulation in Humans. *Annual Review*
1210 *of Genomics and Human Genetics*, *17*, 45–67. https://doi.org/10.1146/annurev-genom-
1211 090314-045935
1212 Romero, I. G., Ruvinsky, I., & Gilad, Y. (2012). Comparative studies of gene expression and the
1213 evolution of gene regulation. *Nature Reviews Genetics*, *13* (7), 505–516.
1214 https://doi.org/10.1038/nrg3229
1215 Sharma, R., Khristov, V., Rising, A., Jha, B. S., Dejene, R., Hotaling, N., Li, Y., Stoddard, J.,
1216 Stankewicz, C., Wan, Q., Zhang, C., Campos, M. M., Miyagishima, K. J., McGaughey,
1217 D., Villasmil, R., Mattapallil, M., Stanzel, B., Qian, H., Wong, W., … Bharti, K. (2019).
1218 Clinical-grade stem cell-derived retinal pigment epithelium patch rescues retinal

1219 degeneration in rodents and pigs. *Science Translational Medicine*, *11* (475), eaat5580.
1220 https://doi.org/10.1126/scitranslmed.aat5580
1221 Shave, R. E., Lieberman, D. E., Drane, A. L., Brown, M. G., Batterham, A. M., Worthington, S.,
1222 Atencia, R., Feltrer, Y., Neary, J., Weiner, R. B., Wasfy, M. M., & Baggish, A. L. (2019).
1223 Selection of endurance capabilities and the trade-off between pressure and volume in
1224 the evolution of the human heart. *Proceedings of the National Academy of Sciences*,
1225 *116* (40), 19905–19910. https://doi.org/10.1073/pnas.1906902116
1226 Song, J. H. T., Grant, R. L., Behrens, V. C., Kučka, M., Roberts Kingman, G. A., Soltys, V.,
1227 Chan, Y. F., & Kingsley, D. M. (2021). Genetic studies of human-chimpanzee divergence
1228 using stem cell fusions. *Proceedings of the National Academy of Sciences of the United*
1229 *States of America*, *118* (51), e2117557118. https://doi.org/10.1073/pnas.2117557118
1230 Starr, A. L., Gokhman, D., & Fraser, H. B. (2023). Accounting for cis-regulatory constraint
1231 prioritizes genes likely to affect species-specific traits. *Genome Biology*, *24* (1), 11.
1232 https://doi.org/10.1186/s13059-023-02846-8
1233 Storey Lab. (n.d.). *Qvalue*.
1234 Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A.,
1235 Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., & Mesirov, J. P. (2005). Gene
1236 set enrichment analysis: A knowledge-based approach for interpreting genome-wide
1237 expression profiles. *Proceedings of the National Academy of Sciences*, *102* (43), 15545–
1238 15550. https://doi.org/10.1073/pnas.0506580102
1239 Su-Feher, L., Rubin, A. N., Silberberg, S. N., Catta-Preta, R., Lim, K. J., Ypsilanti, A. R., Zdilar,
1240 I., McGinnis, C. S., McKinsey, G. L., Rubino, T. E., Hawrylycz, M. J., Thompson, C.,
1241 Gartner, Z. J., Puelles, L., Zeng, H., Rubenstein, J. L. R., & Nord, A. S. (2022). Single
1242 cell enhancer activity distinguishes GABAergic and cholinergic lineages in embryonic
1243 mouse basal ganglia. *Proceedings of the National Academy of Sciences*, *119* (15),
1244 e2108760119. https://doi.org/10.1073/pnas.2108760119

Sullivan, P. F., Meadows, J. R. S., Gazal, S., Phan, B. N., Li, X., Genereux, D. P., Dong, M. X., Bianchi, M., Andrews, G., Sakthikumar, S., Nordin, J., Roy, A., Christmas, M. J., Marinescu, V. D., Wallerman, O., Xue, J. R., Li, Y., Yao, S., Sun, Q., … Lindblad-Toh, K. (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* (1), 5233. 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., & Pașca, S. P. (2020). Chromatin accessibility dynamics in a model of human forebrain development. *Science (New York, N.Y.)*, *367* (6476), eaay1645. 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* (10), 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* (11), 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* (4), 213–232. https://doi.org/10.1038/s41583-023-00675-z

Vierstra, J., Lazar, J., Sandstrom, R., Halow, J., Lee, K., Bates, D., Diegel, M., Dunn, D., Neri, F., Haugen, E., Rynes, E., Reynolds, A., Nelson, J., Johnson, A., Frerker, M., Buckley, M., Kaul, R., Meuleman, W., & Stamatoyannopoulos, J. A. (2020). Global reference

mapping of human transcription factor footprints. *Nature*, *583* (7818), 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* (1), 9. 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* (8), 452–463. https://doi.org/10.1038/s41574-018-0037-x

Watanabe, A., Toyota, T., Owada, Y., Hayashi, T., Iwayama, Y., Matsumata, M., Ishitsuka, Y., Nakaya, A., Maekawa, M., Ohnishi, T., Arai, R., Sakurai, K., Yamada, K., Kondo, H., Hashimoto, K., Osumi, N., & Yoshikawa, T. (2007). Fabp7 Maps to a Quantitative Trait Locus for a Schizophrenia Endophenotype. *PLoS Biology*, *5* (11), e297. https://doi.org/10.1371/journal.pbio.0050297

Wittkopp, P. J., Haerum, B. K., & Clark, A. G. (2004). Evolutionary changes in cis and trans gene regulation. *Nature*, *430* (6995), 85–88. https://doi.org/10.1038/nature02698

Wittkopp, P. J., & Kalay, G. (2012). Cis-regulatory elements: Molecular mechanisms and evolutionary processes underlying divergence. *Nature Reviews Genetics*, *13* (1), 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* (1), 15. 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* (1), 59. https://doi.org/10.1186/s13059-019-1663-x

1295 Yang, N., Chanda, S., Marro, S., Ng, Y.-H., Janas, J. A., Haag, D., Ang, C. E., Tang, Y., Flores,
1296 Q., Mall, M., Wapinski, O., Li, M., Ahlenius, H., Rubenstein, J. L., Chang, H. Y., Buylla,
1297 A. A., Südhof, T. C., & Wernig, M. (2017). Generation of pure GABAergic neurons by
1298 transcription factor programming. *Nature Methods*, *14* (6), 621–628.
1299 https://doi.org/10.1038/nmeth.4291
1300 Yao, Z., van Velthoven, C. T. J., Nguyen, T. N., Goldy, J., Sedeno-Cortes, A. E., Baftizadeh, F.,
1301 Bertagnolli, D., Casper, T., Chiang, M., Crichton, K., Ding, S.-L., Fong, O., Garren, E.,
1302 Glandon, A., Gouwens, N. W., Gray, J., Graybuck, L. T., Hawrylycz, M. J., Hirschstein,
1303 D., … Zeng, H. (2021). A taxonomy of transcriptomic cell types across the isocortex and
1304 hippocampal formation. *Cell*, *184* (12), 3222-3241.e26.
1305 https://doi.org/10.1016/j.cell.2021.04.021
1306 York, R. A., Patil, C., Abdilleh, K., Johnson, Z. V., Conte, M. A., Genner, M. J., McGrath, P. T.,
1307 Fraser, H. B., Fernald, R. D., & Streelman, J. T. (2018). Behavior-dependent *cis*
1308 regulation reveals genes and pathways associated with bower building in cichlid fishes.
1309 *Proceedings of the National Academy of Sciences*, *115* (47).
1310 https://doi.org/10.1073/pnas.1810140115
1311 Yu, X., Lin, J., Zack, D. J., & Qian, J. (2006). Computational analysis of tissue-specific
1312 combinatorial gene regulation: Predicting interaction between transcription factors in
1313 human tissues. *Nucleic Acids Research*, *34* (17), 4925–4936.
1314 https://doi.org/10.1093/nar/gkl595
1315 Zhang, S., Zhang, H., Zhou, Y., Qiao, M., Zhao, S., Kozlova, A., Shi, J., Sanders, A. R., Wang,
1316 G., Luo, K., Sengupta, S., West, S., Qian, S., Streit, M., Avramopoulos, D., Cowan, C.
1317 A., Chen, M., Pang, Z. P., Gejman, P. V., … Duan, J. (2020). Allele-specific open
1318 chromatin in human iPSC neurons elucidates functional disease variants. *Science*, *369*
1319 (6503), 561–565. https://doi.org/10.1126/science.aay3983

Zhang, Y., Liu, T., Meyer, C. A., Eeckhoute, J., Johnson, D. S., Bernstein, B. E., Nusbaum, C., Myers, R. M., Brown, M., Li, W., & Liu, X. S. (2008). Model-based analysis of ChIP-Seq (MACS). *Genome Biology*, *9* (9), R137. 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* (12), 2084–2092. https://doi.org/10.1093/bioinformatics/bty895