

METHOD

Open Access



Accounting for *cis*-regulatory constraint prioritizes genes likely to affect species-specific traits

Alexander L. Starr¹, David Gokhman^{1,2} and Hunter B. Fraser^{1*} 

*Correspondence:
hbfraser@stanford.edu

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

² Present Address: Department
of Molecular Genetics, Weizmann
Institute of Science, Rehovot,
Israel

Abstract

Measuring allele-specific expression in interspecies hybrids is a powerful way to detect *cis*-regulatory changes underlying adaptation. However, it remains difficult to identify genes most likely to explain species-specific traits. Here, we outline a simple strategy that leverages population-scale allele-specific RNA-seq data to identify genes that show constrained *cis*-regulation within species yet show divergence between species. Applying this strategy to data from human-chimpanzee hybrid cortical organoids, we identify signatures of lineage-specific selection on genes related to saccharide metabolism, neurodegeneration, and primary cilia. We also highlight *cis*-regulatory divergence in *CUX1* and *EDNRB* that may shape the trajectory of human brain development.

Keywords: Hybrid, Gene expression, Evolution, Human, Chimpanzee, Primate, Brain, Development

Background

Changes in gene expression are thought to play a major role in the evolution of complex traits [1–4]. As a result, comparing gene expression between species can enable the identification of molecular changes underlying phenotypic divergence. However, obtaining accurate comparisons of gene expression between species is challenging due to confounding factors like age, environmental effects, differential cell type abundances, differences in developmental timing, and batch effects [5–7]. The use of interspecies hybrids overcomes these issues through the measurement of allele-specific expression (ASE) [8, 9]. In hybrids, the genomes of both species share the same nucleus and are exposed to identical environments, so there are no confounding environmental, batch, compositional, or developmental effects. This approach has been successfully applied in many species and advanced our understanding of the evolution of gene regulation and its role in establishing phenotypic variation [10–13]. Furthermore, the recent development of human-chimpanzee allotetraploid “hybrid” cells



© The Author(s) 2023. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

and organoids enables detailed, accurate quantification of differences in gene expression between humans and our closest living relatives [8, 9, 14].

Hybrids also enable the separation of *cis*- and *trans*-components of interspecies differences in gene expression [8, 9]. The *cis*-component is caused by differences in regulatory elements such as promoters or enhancers that only affect the expression of a nearby gene or genes on the same DNA molecule. The *trans*-component stems from changes in diffusible molecules such as transcription factors that can regulate gene expression throughout the genome. In hybrids, the genomes of the two species are exposed to the same *trans* factors. As a result, allele-specific differences in gene expression can only be explained by *cis*-regulatory differences. In addition, using ASE to identify differentially expressed genes (referred to as AS-DEGs) enables the elimination of many confounding factors (including the environmental, batch, compositional, and developmental timing effects mentioned above). Thus, ASE not only increases the signal-to-noise ratio, but also disentangles potentially evolutionarily significant gene-specific *cis*-regulation from broad *trans*-acting changes [4, 15, 16].

While the resulting list of AS-DEGs from hybrids is likely more accurate and isolates the *cis*-regulatory component, there are often thousands of AS-DEGs which make it difficult to prioritize candidate genes and pathways that may have played a major role in evolution. Differential expression *p*-values are commonly used to rank genes in comparative RNA-seq studies. However, genes with large and significant fold changes may often have low evolutionary constraint on expression level. These large fold changes in unconstrained genes (e.g., pseudogenes) could result in no or very limited phenotypic changes, since a lack of constraint implies a lack of phenotypic consequence of changes in expression. Therefore, a large and significant fold change alone is not sufficient to determine the importance of the gene in the evolution of the parental species. For example, consider a gene whose expression varies by two-fold between species. If this gene also varies by two-fold within each species individual members of the same species, it is unlikely to account for any species-specific phenotypes. In contrast, a gene that is under strong stabilizing selection—with little variation in expression within species but with a two-fold change between species—is more likely to have contributed to phenotypic divergence between species. Most studies of expression divergence between species do not include any comparison to within-species variation; the few exceptions have been limited by small sample sizes and the confounding factors inherent to any between-species comparison [17–20].

Here, we describe a method that leverages population-scale ASE data to approximate constraint on the *cis*-regulation of gene expression. This method ranks AS-DEGs identified from interspecies hybrids in a way that is likely more able to prioritize adaptive, functionally significant changes. We apply this method to ASE data from human-chimpanzee hybrid cortical organoids that recapitulate the gene expression patterns of the developing cerebral cortex. Using this dataset, we identify lineage-specific selection on the expression of genes related to neuronal homeostasis, glucose metabolism, primary cilium assembly, and glycan degradation [8, 9]. In addition, we highlight divergence in the expression of *CUX1* and *EDNRB* that may have played an important role in human brain evolution.

Results

In a typical ASE pipeline, AS-DEGs are identified by comparing the RNA-seq read counts from the allele from species 1 and the allele from species 2 and ranked using a p -value or false discovery rate (FDR) for differential expression (Fig. 1A). Various enrichment tests and knowledge from the literature can then be used to identify interesting trends and prioritize candidate genes. However, these previous methods do not consider within-species variation in gene expression levels (Fig. 1B). If some genes have highly variable expression even within a single species, then differences of a similar (or smaller) magnitude between species are unlikely to explain species-specific traits (e.g., *PDPR* in Fig. 1B). Conversely, differential expression of genes with highly constrained expression in at least one species are more likely to be responsible for differences in organismal phenotypes between species. For example, *ZNF331* and *RPS16* have similar fold change magnitudes in human-chimpanzee hybrid cortical organoids (Fig. 1C). However, the fold change for *ZNF331* lies well within the distribution of fold changes between alleles in the human population whereas the fold changes for *RPS16* are nearly outside the human population distribution (Fig. 1C). This indicates that the expression level of *RPS16* is much more constrained and that its differential *cis*-regulation between human and chimpanzee is more likely to have phenotypic consequences.

To systematically apply this concept, we compute the distribution of ASE in one population and use the Mann-Whitney U test to compare the population-level ASE distribution to the interspecies ASE distribution for that gene (see the “[Methods](#)” section) [21]. We use a large publicly available dataset of post-mortem human RNA sequencing data, GTEx, to estimate ASE variance for every gene. The Mann-Whitney p -value reflects how divergent the ASE of the gene is between species compared to its divergence within species and may better reflect the potential of a change in gene expression to alter phenotypes. The ranked list is then used in enrichment analyses and to identify top candidates [22, 23].

Establishing population-scale ASE distributions as a meaningful metric for constraint on gene expression

The GTEx v8 data includes RNA-seq from 838 individuals and 54 tissues for a total of 15,253 samples in which ASE values have been previously estimated [21]. To minimize any allelic bias in the population distribution, we normalize the median to 1. The resulting rankings are robust to choosing either the reference or alternative allele as the numerator for the population distribution (Spearman's $\rho = 0.999$, $p < 10^{-300}$). Notably, we minimize any batch effects between GTEx and the human-chimpanzee hybrid dataset by focusing on ASE since direct comparisons of expression levels between datasets are not involved. For example, if some technical factor (e.g., sequencing platform or RNA isolation method) caused a gene to show a spurious 2-fold higher expression in GTEx samples compared to the hybrid data, this would cancel out in the GTEx ASE calculation. In addition, biological factors (such as mutational target size) are also controlled for by comparing the expression of two different alleles.

First, we tested whether the variance in ASE in the human population is a reasonable proxy for constraint on gene expression. We compared the variance of the GTEx ASE

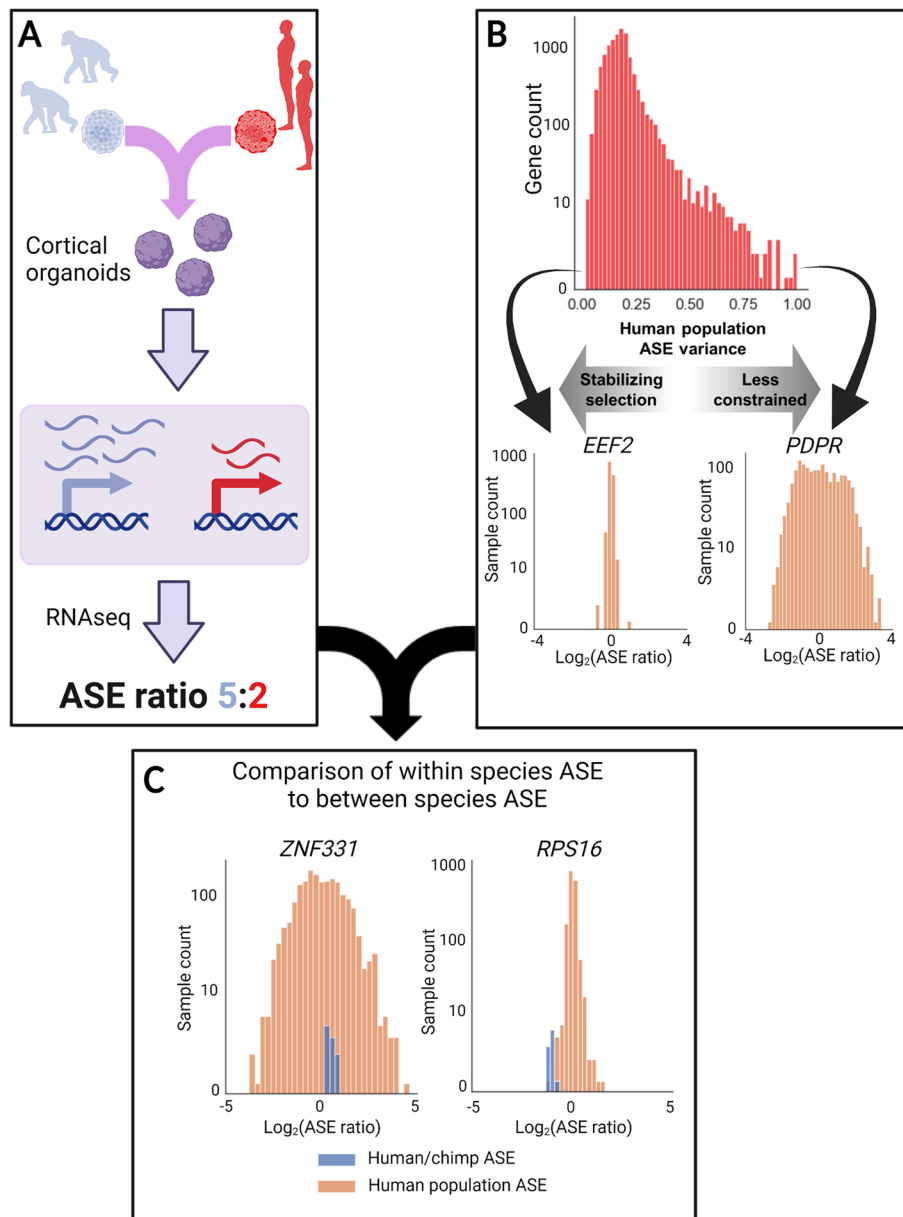


Fig. 1 Outline of the methodology. **A** Outline of a typical ASE pipeline. Hybrids are generated and RNA-seq is used to determine the relative expression of each allele. The ASE ratio is computed as the ratio of species-specific read counts between the two alleles. **B** The distribution of the variance in the ASE ratio for each gene in the GTEx data. Insets in orange show two genes at the extreme ends (*EEF2* with low variance suggesting strong stabilizing selection and *PDPR* with high variance suggesting less constraint on gene expression). For visualization purposes, we removed the few genes with GTEx ASE variance greater than 1 in this panel only. **C** Schematic of incorporation of the interspecies ASE and population level ASE. *ZNF331* has a wide range of ASE values, and the human-chimpanzee ASE is well within the population distribution whereas human-chimp ASE in *RPS16* is on the edge of the population distribution indicating the greater potential for functional relevance. For both **B** and **C**, only GTEx brain samples were used as this provided clearer illustrative examples, though results are similar using all GTEx samples (Fig. 2A)

distribution for each gene to its probability of haploinsufficiency score (pHI), a measure of sensitivity to a 50% reduction in gene dosage [24]. The pHI score was developed by using statistical learning approaches on a massive dataset of disease-associated copy

number variation in humans to predict the probability that each gene in the genome is haploinsufficient [24]. We observe a significant negative correlation (Spearman's $\rho = -0.28$, $p < 10^{-170}$) between the two measures indicating that the variance of the population-scale ASE distribution provides a reasonable proxy for constraint on expression levels (Additional file 1: Fig. S1A, B). This correlation remains significant when only five (as opposed to ten) reads from each allele for a gene in a sample are required for inclusion (Spearman's $\rho = -0.27$, $p < 10^{-160}$). Furthermore, genes with high pHI (> 0.75) value have lower ASE variance than genes with low pHI (< 0.25) when controlling for expression ($p < 10^{-150}$, paired t -test, Additional file 1: Fig. S2) [24]. While this indicates that variation in ASE is sufficient to approximate evolutionary constraint on gene expression, the strength of this relationship should not be interpreted as a quantitative estimate of how well ASE represents constraint. Overall, the correlation with pHI indicates that within-species ASE variance contains useful information about evolutionary constraint on gene expression.

In addition, it is important to determine the extent to which sample heterogeneity may affect ASE variance and thereby impact our results. For example, GTEx contains data from a wide range of adult tissues, and donors have variable sex and ancestry. To determine the extent to which these factors affect our results, we divided the GTEx dataset along lines of sex (male/female), ancestry (African descent/European descent), and brain/non-brain tissues. In all cases, the results obtained were very highly correlated (Spearman's $\rho > 0.97$, $p < 10^{-170}$, Fig. 2A–C). This indicates that sample heterogeneity

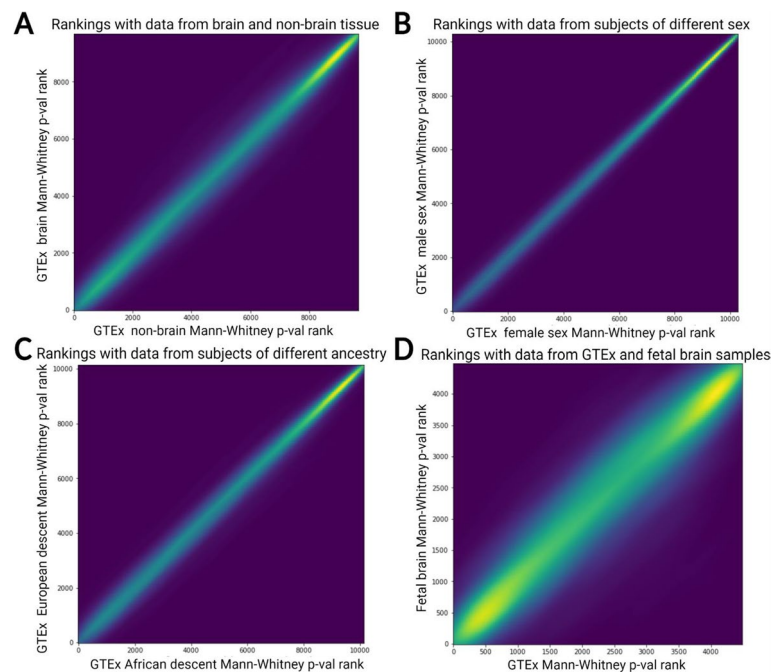


Fig. 2 Influence of population differences on gene rankings. **A** Comparison of ranks generated by our method derived from brain vs. non-brain samples. Spearman's $\rho = 0.98$, $p < 10^{-170}$. **B** Comparison of ranks using only female subjects vs. male GTEx subjects. Spearman's $\rho = 0.98$, $p < 10^{-170}$. **C** Comparison of ranks using only GTEx subjects of African descent vs. subjects of European descent. Spearman's $\rho = 0.99$, $p < 10^{-170}$. **D** Comparison of ranks derived from GTEx (across all tissues) vs. data from fetal cortex and primary fetal neurons/neural progenitors. Spearman's $\rho = 0.92$, $p < 10^{-170}$

is unlikely to have major effects on our results and that aggregating across all the GTEx samples provides a reasonable measure of constraint on gene expression within the human population.

The wide range of adult tissues in GTEx may or may not accurately reflect gene expression constraint in cortical organoids, which mimic fetal development [25, 26]. To test this, we used an ASE dataset generated from neural progenitors, neurons, and fetal cortical wall which resembles the developmental stage of cortical organoids more closely than GTEx samples [27]. There is a strong correlation between the rankings generated by comparing to the fetal cortex dataset and GTEx (Spearman's $\rho = 0.92$, $p < 10^{-170}$ for each of day 50, day 100, and day 150 after the initial differentiation of cortical organoids; Fig. 2D). While it is likely that the ASE variance for some genes changes over development, these genes appear to be somewhat rare. Overall, these results (Fig. 2) indicate that even though gene expression levels vary considerably across samples, ASE variance is robust to sample heterogeneity. This suggests that within- and between-species ASE values can be meaningfully compared even when they are not based on the same tissues or developmental time points. We therefore focused our analysis on the full GTEx data but verified results for specific genes in the fetal cortex data when appropriate.

An important factor in determining the applicability of our method to other organisms is the number of samples required for population-scale ASE variance to meaningfully reflect constraints on gene expression. To explore this, we restricted the GTEx data to only genes with greater than or equal to 5000 samples and down-sampled the GTEx data 100 times across a range of values and computed the Spearman correlation with pHI (the “Methods” section). Even when only 10 samples are used, the correlation with pHI is still highly significant, and the correlation is even stronger when we use all genes with 10 samples or more (Spearman's $\rho = -0.13$ and -0.17 , respectively, the former is shown in Additional file 1: Fig. S3). In addition, the Spearman correlation with 250 samples was nearly as strong as with 5000 samples (Spearman's $\rho = -0.225$ and -0.236 , respectively, Additional file 1: Fig. S3). While the number of samples required to accurately estimate ASE variance is likely a complex function of sequencing depth, library preparation methods, the genetic diversity present in a population, and other technical factors, our results indicate that meaningful estimates of ASE variance can be achieved for many species, including some non-model organisms.

Exploring the effects of incorporating population-scale ASE into interspecies comparisons of ASE

To visualize the differences in rankings produced by our method and the traditional method (ranking by either DESeq2 FDR or \log_2 fold change), we plotted the rankings on a scatterplot. This showed that our method appreciably alters the rankings (Additional file 1: Fig. S4A, B). Next, we sought to highlight the differences between our method and the traditional method for ranking genes (Fig. 1A vs. Fig. 1C). We did this by computing the difference in ranks between the two methods and used gene set enrichment analysis in Python (GSEAPY) to identify gene sets enriched near the top or bottom of the resulting sorted list (referred to as the difference in ranks list, Fig. 3A) [28]. A top-ranking gene would be one that is lowly ranked by the traditional method but highly ranked by the population comparison method. Many Gene Ontology (GO) categories related to

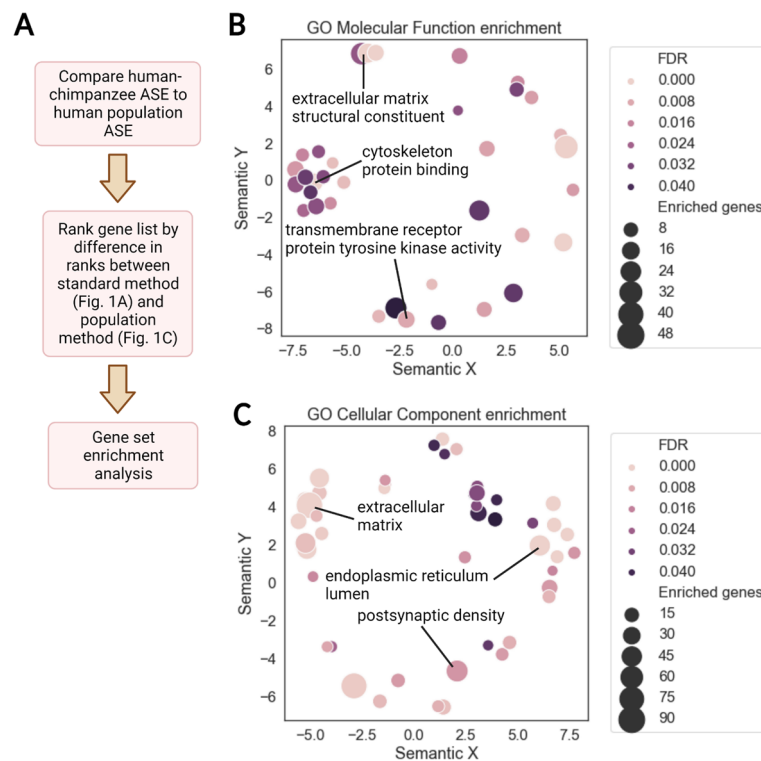


Fig. 3 Enrichment summary for the difference in ranks. **A** Pipeline for gene set enrichment analysis with GSEAPY preranked. Only genes with sufficient reads in the human-chimpanzee cortical organoid dataset and a sufficient number of samples in GTEx were included (see the “Methods” section), leaving approximately 10,000 genes in the list used for enrichment analysis. **B** Summary of GO molecular function enrichments across all time points with false discovery rate (FDR) < 0.05. REVIGO in conjunction with a custom python script was used to generate the plot [33]. The axes are derived from multidimensional scaling and measure semantic similarity, enabling the removal of redundant GO terms and visualization of the similarity between GO categories. Each circle represents a GO term and circles near each other contain similar genes in the corresponding gene set. Labeled gene sets are generally those with the lowest FDR in a cluster of terms on the plot or those highlighted in the text. The size of the circles indicates the number of genes that are driving the enrichment for that category. **C** Summary of GO cellular component enrichments across all time points with FDR < 0.05

cortical development were enriched near the top of the difference in ranks list including calcium and potassium channel activity, various transcription factor (TF) related terms, and neuroligand-receptor interaction (Fig. 3B, C). This is consistent with previous observations that TFs, ion channel subunits, and important players in canonical signaling pathways tend to have strongly constrained expression compared to other genes [24]. Many genes driving these enrichments are haploinsufficient and play key roles in neurodevelopment including *MEF2C*, *NEUROD2*, *CUX1*, and *EDNRB* (Additional file 1: Fig. S5A) [29–32].

Notably, differences in the expression of haploinsufficient genes are more likely to have phenotypic consequences than differences in the expression of genes for which loss of one copy has no clear phenotypic effect. For example, there is strong evidence that both increases and decreases in *CUX1* expression alter neurodevelopment in humans, which implies that the *CUX1* expression divergence between humans and chimpanzees is likely to have phenotypic consequences [29, 34]. *CUX1* is expressed at a lower level in

humans than chimpanzees across time points in both hybrid and parental cortical organoids (e.g., \log_2 fold change of -0.93 , $FDR < 0.005$ in parentals, mean \log_2 fold change of -0.74 , $FDR < 0.023$ in hybrids at day 150 of differentiation, Additional file 1: Fig. S5B, C). Furthermore, the per-sample ASE ratios are well outside the human fetal cortex population ASE distribution and only 2.5% of GTEx samples have a more extreme ASE value (Additional file 1: Fig. S5D, E).

Another haploinsufficient gene that is ranked highly by our method but not by the DESeq2 false discovery rate is *EDNRB*, which is one of the causal genes for Waardenburg-Hirschprung disease and plays an important role in neurodevelopment [32, 35, 36]. Human-chimpanzee ASE generally exceeds ASE found in human populations for *EDNRB* (Fig. 4A, B; Mann-Whitney U test $p = 0.00027$ comparing *EDNRB* distribution to human GTEx population). *EDNRB* is also one of the most strongly human-biased genes across all time points in both hybrid and parental cortical organoids (mean \log_2 fold-change = 4.46, $FDR < 0.005$ at day 150 in hybrids, \log_2 fold-change = 2.85, $FDR < 0.0005$ in parental samples, Fig. 4C).

Next, we analyzed published single-cell RNA-seq data from brain organoids to identify the cell types driving increased *EDNRB* expression [37]. We found that a previously identified radial glial cell (RGC) cluster was characterized by high *EDNRB* expression with non-zero expression in over 50% of cells [37]. Furthermore, this cluster had higher

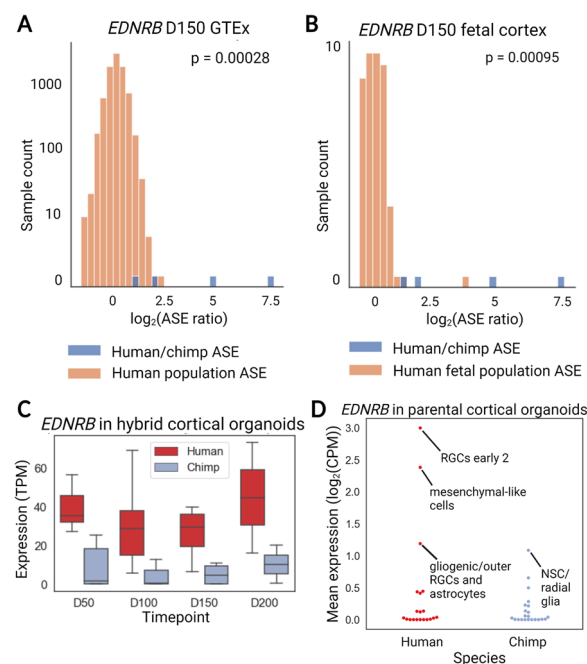


Fig. 4 Changes in the expression of *EDNRB* in great apes. **A** Comparison of human-chimpanzee *EDNRB* ASE to within-human ASE from GTEx. Raw ASE ratios (as opposed to the value derived from DESeq2) are indicated by “ASE Ratio.” The human-chimpanzee ASE is significantly outside of the human ASE distribution. p -values are from the Mann-Whitney U test comparing the distribution of human population ASE to human-chimpanzee ASE. **B** Comparison to the fetal brain population ASE distribution. **C** ASE of *EDNRB* across time points in cortical organoids. Expression from the human allele is consistently higher than expression from the chimpanzee allele. **D** Plot showing the mean expression of *EDNRB* across chimpanzee and human cell clusters [37]. Expression in the radial glia cluster “RGCs early 2” cluster is significantly higher than in all chimpanzee clusters by Mann-Whitney U test ($p < 10^{-16}$ for all comparisons). Data are from Kanton et al. [37]

expression than any chimpanzee cluster (Mann-Whitney U test, $p < 10^{-16}$, Fig. 4D). In addition, changes in *EDNRB* expression appear to be human-derived with respect to gorillas and macaques (although orangutans may have independently evolved similar expression to humans in early-stage brain organoids) (Additional file 1: Fig. S6A, B). Overall, our results suggest that a subpopulation of human radial glia has much higher *EDNRB* expression than chimpanzee radial glia.

Applying the sign test to constrained differentially expressed genes

While our method is designed to prioritize genes that are relevant to phenotypic differences between species, it does not imply selection on changes in gene expression on its own. On the other hand, the sign test is designed to detect lineage-specific selection based on systematic up- or downregulation of genes that deviates from neutral expectation [38, 39]. This is similar in principle to testing whether a coin is biased based on the outcome of multiple coin flips. As a result, an important requirement of the test is that the expression divergence of every gene being tested should be driven by independent genetic differences. For example, a single mutation in a *trans*-acting factor could cause a whole pathway to be downregulated, but this would only count as one genetic difference. For this reason, hybrids are ideal for applying the sign test to gene expression, since the *cis*-regulation of genes is typically independent (except in the case of some neighboring genes that share *cis*-elements, which can be accounted for by using a minimum distance threshold between genes). For example, in the human-chimpanzee hybrid cells, if a particular pathway contains significantly more genes with higher expression from the human allele than the chimpanzee allele (with significance determined by a binomial test) that would provide evidence of selection on that pathway in the human and/or chimpanzee lineage. Similar to all tests of selection, the sign test cannot discern what the cause of the selective pressure is. For example, many *cis*-regulatory changes could be compensating for a change in a single *trans*-acting factor. Alternatively, changes in gene expression might cause a phenotypic change that is being selected for. In either case, the sign test provides evidence of lineage-specific selection on *cis*-regulation.

Thus far, we have primarily focused on the differences between our new method and the traditional method for ranking genes. Having established their differences, we now turn to an analysis of results from our new approach (Fig. 1 and the “Methods” section). To examine human- and chimpanzee-biased genes separately, we sorted the list so that highly ranked genes with human-biased expression are at the top of the list and highly ranked genes with chimpanzee-biased expression are at the bottom of the list. In effect, this results in a test for directionally-biased *cis*-regulatory divergence that exceeds the *cis*-regulatory variation among most human alleles present in the GTEx population. More concretely, we first run GSEAPY preranked on this gene list and take all gene sets with GSEAPY FDR < 0.25 (the cutoff suggested by the GSEA authors) [28]. We then use the rank cutoff identified by GSEAPY to count the number of genes that are human-biased and the number that are chimpanzee-biased, excluding non-significant genes. We then apply the sign test (in the form of the binomial test) and consider any gene sets with binomial FDR less than 0.05 to be putatively under selection.

We elected to focus on the top ten most enriched terms ranked by the proportion of genes that show matching directionality (Additional file 2: Table S1). This strategy

identified several gene sets that are particularly relevant to brain development including *Other glycan degradation* (KEGG, human-biased, binomial $p = 0.03$, *Spastic dysarthria* (HPO, human-biased, binomial $p = 0.002$), *Gluconeogenesis* (REACTOME, chimpanzee-biased, binomial $p = 0.000061$), and *Nonmotile primary cilium assembly* (GO biological process, human-biased, binomial $p = 0.002$) (Fig. 5A–D, Additional file 2: Table S1). Encouragingly, genes related to carbohydrate metabolism and ciliary function have

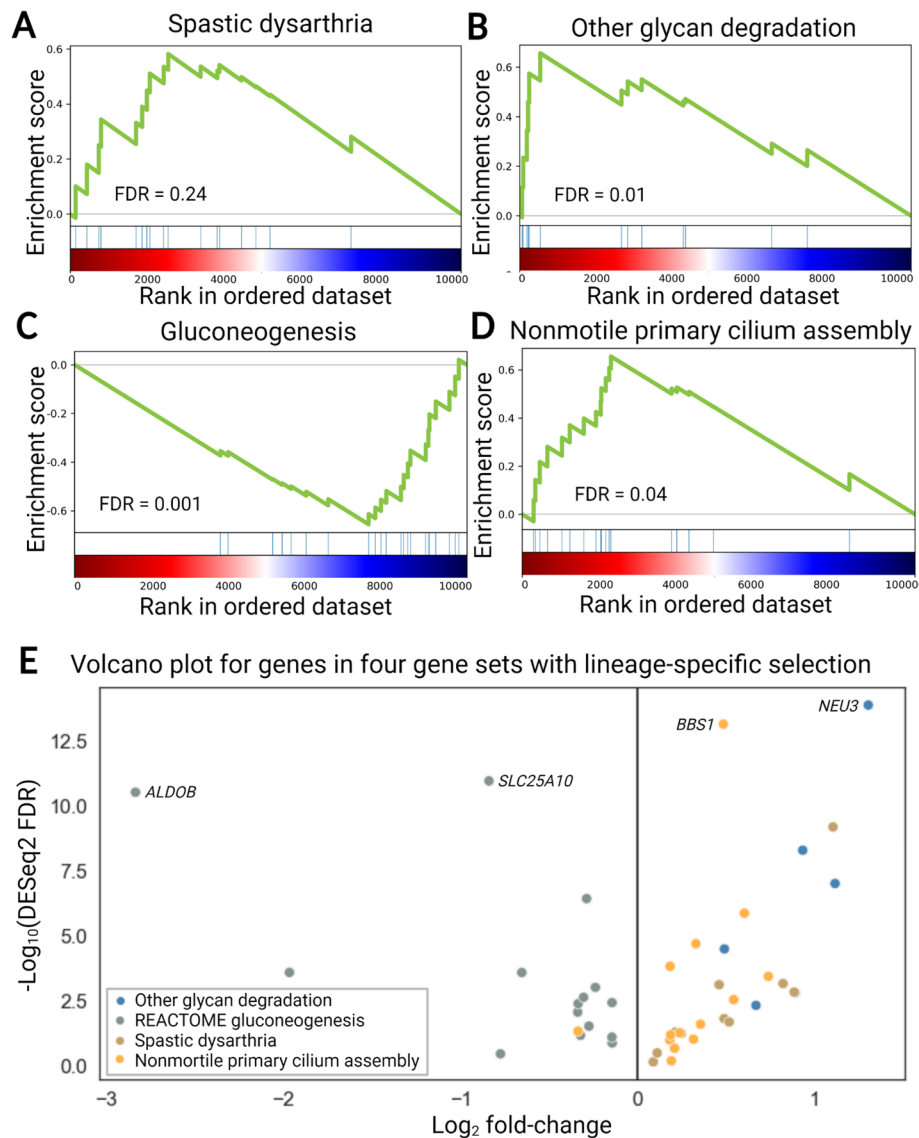


Fig. 5 Evidence of lineage-specific selection. **A** Summary of *Spastic dysarthria* enrichment (from human phenotype ontology). In **A–D**, each blue line represents a gene in the gene set, and the green curve is the cumulative enrichment score. **B** Summary of *other glycan degradation* enrichment (from KEGG). **C** Summary of *Gluconeogenesis* enrichment (from REACTOME). **D** Summary of *Nonmotile primary cilium assembly* enrichment (from GO Biological Process). **E** Volcano plot summarizing of \log_2 fold changes for genes driving the enrichments for *Spastic dysarthria*, *Other glycan degradation*, *Gluconeogenesis*, and *Nonmotile primary cilium assembly*. All genes are human-biased for *Other glycan degradation* and *Spastic dysarthria* whereas all genes are chimpanzee-biased for *gluconeogenesis*. All but one gene is human-biased for *Nonmotile primary cilium assembly*. The four genes with the lowest DESeq2 FDR are highlighted

previously been proposed to have played an important role in human evolution [8, 40, 41]. In addition, several other terms that are less relevant to the brain but may have undergone selection on gene expression in other tissues (e.g., *Regulation of bone mineralization* as well as *Metabolism of xenobiotics by cytochrome P450* and related terms) were significant using our method. In three out of the four brain-related gene sets we highlight, ASE for all genes was in the same direction at an identical rank cutoff (Fig. 5E, see Additional file 2: Table S1 for additional gene sets). The only exception was *Nonmotile primary cilium assembly* which has a single chimpanzee-biased gene. These strong asymmetries in ASE suggest lineage-specific selection acting on these gene sets [39]. Notably, the bias in ASE found in the hybrids for these genes generally matched the bias in expression found in the parental cortical organoids (Additional file 1: Fig. S7A–D).

All four of the gene sets we highlight could plausibly play an important role in evolutionary changes in brain development. Gluconeogenesis siphons oxaloacetate from the TCA cycle and eventually produces glucose [40]. Many of the gene expression changes driving the gluconeogenesis enrichment appear to be human-derived compared to other great apes (Additional file 1: Fig. S8). Decreased gluconeogenesis in the human lineage would likely enable increased flux through other anabolic pathways and the TCA cycle possibly increasing the availability of oxaloacetate for anabolic pathways that promote proliferation [40]. Six genes that were ranked very highly by our method and all have DESeq2 FDR < 0.1 drive the “Other glycan degradation” enrichment. Interestingly, the loss of function of three of the six human-biased glycan degradation genes (*MANBA*, *MAN2B2*, and *MAN2B1*) is associated with intellectual disability [42–44]. Spastic dysarthria is a condition in which patients speak in a characteristic slow, regular, monotone manner as the result of the degeneration of vulnerable upper motor neurons [45]. Spastic dysarthria typically co-occurs with other symptoms such as spastic ataxia so, rather than being connected to speech, this finding may be connected to selection for maintenance of neuronal homeostasis [46]. Finally, our findings add to the growing body of evidence that signaling at the primary cilium likely played an important role in human evolution [8, 41].

Discussion

Here, we presented a method incorporating population-scale ASE data as a proxy for constraint on expression. This ranking method helps reveal candidate genes and signatures of selection that may explain phenotypic differences between humans and chimpanzees. The test is based on the logic of comparing within- to between-species variation, similar in spirit to the Hudson-Kreitman-Aguade test although it is based on variation in ASE rather than protein-coding sequences [47]. Although the phenotypic consequences of these differences remain to be determined, our finding of probable polygenic lineage-specific selection on several gene sets suggests that these changes likely have some phenotypic effects. Collectively, these findings more than double the number of identified cases of lineage-specific polygenic selection on gene expression between humans and chimpanzees (the two previous examples being Hedgehog signaling and astrocyte-related genes) [8, 9]. Future work using outgroup species such as gorilla would be required to determine whether this selection occurred in the human or the chimpanzee lineage.

In addition to exploring gene sets prioritized by our method, we also highlighted two interesting genes. First, we showed that *EDNRB* is much more highly expressed in a subpopulation of human radial glia than in chimpanzee radial glia. *EDNRB* haploinsufficiency reduces the proliferation of cerebellar granule precursor cells and chemical inhibition of *EDNRB* signaling reduces the proliferation of mouse radial glia [35, 36]. Based on this, the change in *EDNRB* expression may have promoted human brain expansion by increasing the proliferation of the subpopulation of radial glia that express *EDNRB*. In addition, a recent study identified a population of caudal late interneuron progenitor (CLIP) cells marked by expression of *EDNRB* and *PTGDS* along with caudal ganglionic eminence markers [48]. As both *EDNRB* and *PTGDS* have strongly human-biased ASE it would be interesting to investigate if this population of cells exists in chimpanzee brain organoids and if it may have expanded in humans. The phenotypic implications of the higher *EDNRB* and *PTGDS* expression in humans will be an exciting area for further research.

We also investigated the decreased expression of the haploinsufficient transcription factor *CUX1* in human brain organoids compared to chimpanzee. Recent work linked a mutation in a human accelerated region (HAR) that likely increases *CUX1* expression to autism spectrum disorder, suggesting that human-derived changes in *CUX1* expression alter human behavior [34, 49]. The reduced expression from the human allele we highlight is surprising considering that a recent massively parallel reporter assay found that the HAR linked to *CUX1* should increase expression in humans [49]. As expression is lower in humans, haploinsufficiency of *CUX1* might provide a reasonable model of the phenotypic consequences of this change. *CUX1* haploinsufficiency in humans leads to delayed development of speech and motor skills [29]. One aspect of this condition is that individuals with one functional *CUX1* allele often close the developmental gap over time (i.e., cognitive impairments and delays disappear with age) [29]. This “catch-up” phenotype is very rare and may even be specific to *CUX1* haploinsufficiency [29]. Interestingly, humans develop more slowly than other great apes (known as neoteny) but eventually “catch up,” reminiscent of the *CUX1* phenotype. Overall, future investigation of the development of layer II-III cortical neurons and behavior of *CUX1* haploinsufficient mice will be required to explore its role in human evolution further.

While we have shown that our method can aid in the analysis of human-chimpanzee hybrid brain organoid data, its applicability to other tissues, cell types, and organisms underlies its potentially broad utility. Importantly, the strong correlation between the GTEx brain vs. non-brain rankings (Fig. 2A) and GTEx vs. fetal cortex rankings (Fig. 2D) suggests that the GTEx ASE distribution can be meaningfully compared with between-species ASE measured in diverse cell types and organoids. Our results also suggest that even with relatively small sample sizes per gene (a low as 10), our method still captures meaningful information about constraints on gene expression. In addition, while it is unclear if our method is applicable to hybrids between more diverged species than humans and chimpanzees, it is likely that ASE variance within one species will still be meaningful. The method can be applied to any species with sufficient gene expression data, e.g., comparing ASE in *Arabidopsis* interspecies hybrids to ASE within *A. thaliana* [50]. Overall, our method has the potential to aid in the identification of gene expression changes underlying species-specific traits across many different organisms.

Conclusions

We outlined a strategy that uses allele-specific expression data from interspecies hybrids and population-scale studies to prioritize genes that are more likely to impact species-specific traits and applied this method to data from human-chimpanzee cortical organoids. Our findings provide opportunities for targeted follow-up experiments and increase our understanding of how polygenic selection has shaped human and chimpanzee evolution. Overall, we anticipate that our method will become a useful tool for identifying functionally significant gene expression changes in many tissues and cell types across diverse species and will contribute to our understanding of how gene expression drives phenotypic diversification.

Methods

Read alignment and RNA-seq data processing

Data from hybrid cortical organoids was mapped as previously described [9]. Briefly, hornet, a rewritten version of wasp, was used in conjunction with a curated list of human-chimpanzee SNPs and indels to correct for mapping bias [51]. Reads for every sample were aligned to both the human and chimpanzee genomes and the \log_2 fold-change from both alignments was compared. Any genes with \log_2 fold change that differed by greater than 1 were removed. We used the ASE \log_2 fold change values available in the supplemental tables of Agoglia et al. Although this dataset is restricted to hybrids from two humans and two chimpanzees, previous work has shown that interspecies differences dominate over differences between populations within a species so we expect that our results generalize well. It also contains multiple independently derived hybrid lines and independent differentiations, reducing confounding by technical differences. Throughout, for hybrids the mean \log_2 fold change between human genome mapped and chimpanzee genome mapped reads is stated as well as the highest p -value. Additional data was downloaded from GSE106245, GSE153076, and phs000755.v2.p1 and mapped separately for each dataset to the respective species' genome (PanTro6 for chimpanzee, hg38 for human, mmul10 for rhesus macaque, Gorgor6 for gorilla, and PonAbe3 for orangutan) [7, 52, 53]. We used STAR v2.5.4 with arguments: `-outSAMattributes MD NH -outFilterMultimapNmax 1 -sjdbGTFfile -sjdbOverhang N` where N is 1 less than the read length used for each respective dataset [54]. For paired-end reads, we used Picard to remove duplicates with argument: `DUPLICATE_SCORING_STRATEGY = RANDOM`. We used HT-Seq with the following arguments: `-t exon -i gene_name -m intersection-strict -r pos` to count reads overlapping gene bodies [55]. Transcripts per million (TPM) was computed as previously described [56]. We used the likelihood ratio test in DESeq2 to test for differential expression in the downloaded datasets [23, 57].

Comparison of population and interspecies ASE distributions

GTEX data was downloaded from <https://www.gtexportal.org/home/datasets> and the fetal cortex ASE data was kindly provided by the Stein laboratory [21, 27]. GTEX contains data from 838 individuals, and the data from the Stein Laboratory was generated from approximately 90 individuals and includes neuron, neural progenitor, and cortical

wall data for a total of 235 samples. To preprocess the GTEx data, we split the file into two files each containing read counts from 1 of the alleles with an R script. As cortical organoids are a mixture of different cell types including neural progenitors and immature neurons, we pooled all counts per individual in the fetal cortex dataset. For each gene in each sample, we added one count to each gene (to prevent division by zero) and computed the ASE ratio as the ratio of counts from allele 1 (the reference) to counts from allele 2. The distribution for each gene was then normalized so that the median was 1. This normalization ensures that the Mann-Whitney U test p -values only take into account the variance in allelic expression in the human population and are not confounded by consistently higher/lower expression from a particular allele. Notably, flipping the sign of each value in the GTEx ASE distribution had minimal effect on the rankings (Spearman's $\rho = 0.999$, $p < 10^{-300}$ for day 50, day 100, and day 150 after the beginning of cortical organoid differentiation) supporting the efficacy of the correcting the median to 1 in isolating the variance of the human population expression distribution. For each gene, we removed any samples with fewer than 10 counts (not including the single added count) from each allele. For example, if a sample had greater than or equal to 10 counts from one allele but fewer than 10 from the other allele for gene A, this sample would be removed from the calculation for gene A. However, if that same sample had greater than or equal to 10 counts from each allele for gene B, that sample would be included for the calculation for gene B. Notably the rankings and our results are robust to requiring at least 5 counts from each allele instead of 10 (Spearman's $\rho = 0.996$, $p < 10^{-300}$). To filter out genes that are lowly expressed in cortical organoids, we removed genes with an average number of counts from the chimpanzee and human alleles less than 25 (i.e., mean of human read count less than 25 and mean of chimpanzee read count less than 25). In addition, we filtered out any genes showing mapping bias (listed in the supplemental tables of Agolia et al.) as well as genes on chromosomes 18 and 20 as parts of these chromosomes were duplicated in some cortical organoid samples. Previous work has shown that these structural changes have minimal effect on the computation of ASE values for genes outside the duplicated region [9]. After filtering, we computed the interspecies ASE distribution in a similar manner to the population ASE distribution (i.e., by taking the ratio of the counts from the human allele to the ratio of the counts of the chimpanzee allele). However, we did not require 10 counts from each allele and did not normalize the medians. We did not require 10 counts from each allele because we expect extreme differences in expression to be relatively common in between species comparisons. We compared the $\log_2(\text{ASE Ratio})$ interspecies distribution to the population distribution using the Mann-Whitney U test (a nonparametric test robust to the distribution of data) and used the resulting p -values to rank genes as described below.

Generation of gene rankings and enrichment analysis

To highlight the differences between our method and the more traditional method for ranking genes, we computed the difference in ranks by subtracting the DESeq2 FDR ranking from the MWU ranking. This gene ranking was then used in GSEAPY preranked with the rankings used as the score that GSEAPY uses to sort the list. We used REVIGO in conjunction with a custom python script to generate the plots shown in Fig. 2 [33]. In

this context, highly ranked genes are likely those that show relatively mild gene expression changes but have more constrained expression. For enrichment testing, we tested gene sets from the Gene Ontology cellular component, biological process, and molecular function categories, the human phenotype ontology, KEGG, and REACTOME using the same version as in Gokhman et al. [8, 58–62]. Regardless of which ranking was used, we used GSEAPY preranked with the following arguments: processes = 4, permutation_num = 1000, seed = 6, min_size = 10, and max_size = 300 to test for enrichment [28]. Following the authors' suggestion, we considered any category with an FDR below 0.25 to be nominally enriched [28]. Enrichment analysis was performed for each time point that cortical organoids were frozen for RNAseq (day 50, day 100, and day 150) separately, and significant terms were aggregated with only one copy of redundant terms that were significant in multiple time points included. Only genes with an average of 25 reads from at least one allele in the human-chimpanzee cortical organoid dataset and at least 50 individuals in GTEx were included, leaving approximately 10,000 genes in the list used for enrichment analysis.

We next ranked the genes using a signed version of the MWU ranking for use in the expression sign test. More specifically, genes were effectively ranked by the $\log_{10}(\text{MWU } p\text{-value})$ multiplied by the sign of mean DESeq2 \log_2 fold-change so that top-ranked genes with negative L2FC are at the bottom of the list and top-ranked genes with positive L2FC are at the top of the list. This ranking was then used in GSEAPY preranked with the above parameters. We then generated a list of all gene sets across all tested ontologies that were nominally enriched at an FDR of 0.25 (using the FDR from GSEAPY preranked) in at least one time point and that had greater than five genes driving the enrichment. To avoid testing the same gene set multiple times, we only tested each gene set at the time point that had the lowest GSEAPY FDR. We used the average of the DESeq2 \log_2 fold change from mapping to the human allele and from mapping to the chimpanzee allele as input for the binomial test to identify gene sets with significantly more human-biased or chimpanzee-biased changes than expected by chance. This \log_2 fold change was generated by comparing the reads from each species' allele in the cortical organoid data. As an example of the inputs to the binomial test, if a gene set had 8 human-biased and 3 chimpanzee-biased genes, then the binomial test was used with $k = 8$, $n = 11$, and $p = 0.5$. We considered any gene set with Benjamini-Hochberg corrected FDR < 0.1 to be significant. All statistical tests (Mann-Whitney U test, binomial test, correlations) were performed in python using the implementation in scipy.

Single-cell RNA-seq data processing and analysis

Single-cell data from human and chimpanzee organoids and associated metadata were downloaded from E-MTAB-7552 [37]. We used SCANPY to read in the counts matrix and filter the data so that only data from 1-month, 2-month, and 4-month-old organoids remained [63]. We used a two-sided Mann-Whitney U test to compare $EDNRB$ $\log_2(\text{counts per million})$ between the "RGC early 2" cluster and all chimpanzee clusters with and without cells with 0 $EDNRB$ counts included.

Down-sampling and controlling for gene expression in correlations with pHI

To investigate the dependency of our method on the sample sizes used to generate the population ASE distribution, we randomly down-sampled (without replacement) the full GTEx dataset to $n = 5000, 4000, 3000, 2000, 1000, 500, 250, 100, 50, 25,$ and 10 samples for each gene. This ensures that all genes have an equal number of samples in each condition. For example, if gene A had 8000 samples with quantifiable ASE, we would down-sample to 5000 samples for $n = 5000$ and 50 samples for $n = 50$. However, if gene B had 4000 samples, we would not include this gene in the $n = 5000$ condition but would include it for the $n = 50$ condition. We then restricted only to genes with 5000 or more samples in the full dataset and computed the Spearman correlation with pHI for each down-sampling and took the mean of the 100 Spearman correlations (shown in Additional file 1: Fig. S3). Restricting to only genes with 5000 or more samples is necessary to make a fair comparison. We also repeated this analysis for $n = 10$ and restricted to only genes with 10 or more samples.

To control for expression level in our comparison of the GTEx ASE variance to pHI, we first restricted only to the 2500 genes with the highest pHI (high pHI) and the 2500 genes with the lowest pHI (low pHI). We then separately ranked both gene sets by the mean of the within-tissue median TPM across all GTEx tissues and matched the high pHI gene and low pHI gene with the same mean TPM rank. For each matched pair of genes, we computed the difference in GTEx ASE variance and plotted a histogram of these values. We also compared the paired genes with a paired t -test.

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s13059-023-02846-8>.

Additional file 1: Figure S1. Relationship between the variance of the GTEx ASE distribution and the probability of Haploinsufficiency score. **Figure S2.** Relationship between population ASE variance and pHI when controlling for expression. **Figure S3.** Effect of population ASE sample size on correlation with pHI. **Figure S4.** Visualization of changes in gene ranking after incorporation of population ASE constraint. **Figure S5.** Exploration of changes in *CUX1* expression. **Figure S6.** Changes in *EDNRB* expression are human derived. **Figure S7.** Many gene expression changes have the same direction in hybrid and parental organoids. **Figure S8.** Changes in many gluconeogenesis genes are human-derived.

Additional file 2. The enriched terms that show evidence for lineage-specific selection based on the method outlined in the paper.

Additional file 3. Review history.

Acknowledgements

We thank Jason Stein and Nil Ayygun for providing processed fetal cortex, neuron, and neural progenitor ASE calls. We thank the members of the Fraser Lab for helpful discussions. Illustrations were created with BioRender.

Review history

The review history is available as Additional file 3.

Peer review information

Tim Sands and Anahita Bishop were the primary editors of this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

Authors' contributions

HB, DG, and ALS conceptualized the study. ALS performed all the analysis, validation, visualization, and writing of the software. ALS wrote the original manuscript. HB and DG reviewed and edited the manuscript. All authors approved the publication of the manuscript.

Funding

Funding for this work came from NIH R01HG012285. ALS was supported by an NDSEG fellowship.

Availability of data and materials

No new data nor materials were generated during this study. The datasets supporting the conclusions of this article are available in GSE, dbGaP, or the supplemental material of Agoglia et al. [64], GSE106245 [65], GSE153076 [66], and

phs000755.v2.p1 [67], phs002493.v1.p1 [68] (although the data was obtained directly from the Stein lab for this manuscript), and the GTEx portal [69]. The software used to perform the analysis in this paper is available on Zenodo and is open access under a Creative Commons Attribution 4.0 International license [70].

Declarations

Ethics approval and consent to participate

Access to restricted data was approved by dbGaP.

Consent for publication

Not relevant to our study.

Competing interests

The authors declare that they have no competing interests.

Received: 23 June 2022 Accepted: 4 January 2023

Published online: 19 January 2023

References

1. Reilly SK, Noonan JP. Evolution of gene regulation in humans. *Annu Rev Genomics Hum Genet.* 2016;17:45–67.
2. Romero IG, Ruvinsky I, Gilad Y. Comparative studies of gene expression and the evolution of gene regulation. *Nat Rev Genet.* 2012;13:505–16.
3. King M-C, Wilson AC. Evolution at two levels in humans and chimpanzees. *Science.* 1975;188:107–16.
4. Fraser HB. Gene expression drives local adaptation in humans. *Genome Res.* 2013;23:1089–96.
5. Kelley JL, Gilad Y. Effective study design for comparative functional genomics. *Nat Rev Genet.* 2020;21:385–6.
6. Housman G, Gilad Y. Prime time for primate functional genomics. *Curr Opin Genet Dev.* 2020;62:1–7.
7. Zhu Y, Sousa AMM, Gao T, Skarica M, Li M, Santpere G, et al. Spatiotemporal transcriptomic divergence across human and macaque brain development. *Science.* 2018;362:eaat8077.
8. Gokhman D, Agoglia RM, Kinnebrew M, Gordon W, Sun D, Bajpai VK, et al. Human–chimpanzee fused cells reveal cis-regulatory divergence underlying skeletal evolution. *Nat Genet.* 2021;53:467–76.
9. Agoglia RM, Sun D, Birey F, Yoon S-J, Miura Y, Sabatini K, et al. Primate cell fusion disentangles gene regulatory divergence in neurodevelopment. *Nature.* 2021;592:421–7.
10. Hu CK, York RA, Metz HC, Bedford NL, Fraser HB, Hoekstra HE. Cis-regulatory changes in locomotor genes are associated with the evolution of burrowing behavior. *Cell Rep.* 2022;38:110360.
11. Mack KL, Campbell P, Nachman MW. Gene regulation and speciation in house mice. *Genome Res.* 2016;26:451–61.
12. Combs PA, Krupp JJ, Khosla NM, Bua D, Petrov DA, Levine JD, et al. Tissue-specific cis-regulatory divergence implicates *eloF* in inhibiting interspecies mating in *Drosophila*. *Current Biology.* 2018;28:3969–3975.e3.
13. Zhang X, Borevitz JO. Global analysis of allele-specific expression in *Arabidopsis thaliana*. *Genetics.* 2009;182:943–54.
14. Song JHT, Grant RL, Behrens VC, Kučka M, Roberts Kingman GA, Soltys V, et al. Genetic studies of human–chimpanzee divergence using stem cell fusions. *Proc Natl Acad Sci.* 2021;118:e2117557118.
15. Prud'homme B, Gompel N, Carroll SB. Emerging principles of regulatory evolution. *Proc Natl Acad Sci.* 2007;104:8605–12.
16. Wittkopp PJ, Kalay G. Cis-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence. *Nat Rev Genet.* 2012;13:59–69.
17. Blekhman R, Oshlack A, Chabot AE, Smyth GK, Gilad Y. Gene regulation in primates evolves under tissue-specific selection pressures. *PLoS Genet.* 2008;4:e1000271.
18. Gilad Y, Oshlack A, Smyth GK, Speed TP, White KP. Expression profiling in primates reveals a rapid evolution of human transcription factors. *Nature.* 2006;440:242–5.
19. Roy S, Wapinski I, Pfiffner J, French C, Socha A, Konieczka J, et al. Arboretum: reconstruction and analysis of the evolutionary history of condition-specific transcriptional modules. *Genome Res.* 2013;23:1039–50.
20. Rohlf Rv, Nielsen R. Phylogenetic ANOVA: the expression variance and evolution model for quantitative trait evolution. *Syst Biol.* 2015;64:695–708.
21. Castel SE, Aguet F, Mohammadi P, Aguet F, Anand S, Ardlie KG, et al. A vast resource of allelic expression data spanning human tissues. *Genome Biol.* 2020;21:234.
22. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc Ser B Methodol.* 1995;57:289–300.
23. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* 2014;15:550.
24. Collins RL, Glessner JT, Porcu E, Niestroj L-M, Ulirsch J, Kellaris G, et al. A cross-disorder dosage sensitivity map of the human genome. *Genomic Med Inst.* 15:23–5. Available from: <https://doi.org/10.1101/2021.01.26.21250098> Lerner Research Institute
25. Paşca AM, Sloan SA, Clarke LE, Tian Y, Makinson CD, Huber N, et al. Functional cortical neurons and astrocytes from human pluripotent stem cells in 3D culture. *Nat Methods.* 2015;12:671–8.
26. Ferraro NM, Strober BJ, Einson J, Abell NS, Aguet F, Barbeira AN, et al. Transcriptomic signatures across human tissues identify functional rare genetic variation. *Science.* 2020;369:eaaz5900.
27. Aygün N, Elwell AL, Liang D, Lafferty MJ, Cheek KE, Courtney KP, et al. Brain-trait-associated variants impact cell-type-specific gene regulation during neurogenesis. *Am J Human Genet.* 2021;108:1647–68.

28. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci.* 2005;102:15545–50.
29. Platzer K, Cogné B, Hague J, Marcelis CL, Mitter D, Oberndorff K, et al. Haploinsufficiency of *CUX1* causes nonsyndromic global developmental delay with possible catch-up development. *Ann Neurol.* 2018;84:200–7.
30. Paciorkowski AR, Traylor RN, Rosenfeld JA, Hoover JM, Harris CJ, Winter S, et al. *MEF2C* Haploinsufficiency features consistent hyperkinesia, variable epilepsy, and has a role in dorsal and ventral neuronal developmental pathways. *Neurogenetics.* 2013;14:99–111.
31. Runge K, Mathieu R, Bugeon S, Lafi S, Beurrier C, Sahu S, et al. Disruption of *NEUROD2* causes a neurodevelopmental syndrome with autistic features via cell-autonomous defects in forebrain glutamatergic neurons. *Mol Psychiatry.* 2021;26:6125–48.
32. McCallion AS, Chakravarti A. *EDNRB/EDN3* and Hirschsprung disease type II. *Pigment Cell Res.* 2001;14:161–9.
33. Supek F, Bošnjak M, Škunca N, Šmuc T. REVIKO summarizes and visualizes long lists of Gene Ontology terms. *PLoS One.* 2011;6:e21800.
34. Doan RN, Bae B-I, Cubelos B, Chang C, Hossain AA, Al-Saad S, et al. Mutations in human accelerated regions disrupt cognition and social behavior. *Cell.* 2016;167:341–354.e12.
35. Vidovic M, Chen M-M, Lu Q-Y, Kalloniatis KF, Martin BM, Tan AHY, et al. Deficiency in endothelin receptor B reduces proliferation of neuronal progenitors and increases apoptosis in postnatal rat cerebellum. *Cell Mol Neurobiol.* 2008;28:1129–38.
36. Shinohara H, Udagawa J, Morishita R, Ueda H, Otani H, Semba R, et al. Gi2 signaling enhances proliferation of neural progenitor cells in the developing brain. *J Biol Chem.* 2004;279:41141–8.
37. Kanton S, Boyle MJ, He Z, Santel M, Weigert A, Sanchís-Calleja F, et al. Organoid single-cell genomic atlas uncovers human-specific features of brain development. *Nature.* 2019;574:418–22.
38. Kita R, Venkataram S, Zhou Y, Fraser HB. High-resolution mapping of *cis*-regulatory variation in budding yeast. *Proc Natl Acad Sci.* 2017;114:E10736–44.
39. Fraser HB. Genome-wide approaches to the study of adaptive gene expression evolution. *BioEssays.* 2011;33:469–77.
40. Gkini V, Namba T. Glutaminolysis and the control of neural progenitors in neocortical development and evolution. *Neuroscientist.* 2022;107385842110690.
41. Suresh H, Crow M, Jorstad N, Hodge R, Lein E, Dobin A, et al. Conserved coexpression at single cell resolution across primate brains. *bioRxiv.* 2022;2022.09.20.508736. Available from: <http://biorxiv.org/content/early/2022/09/22/2022.09.20.508736.abstract>
42. Alkhayat AH, Kraemer SA, Leipprandt JR, Macek M, Kleijer WJ, Friderici KH. Human-mannosidase cDNA characterization and first identification of a mutation associated with human-mannosidosis. *Hum Mol Genet.* 1998;7:75–83.
43. Cathey SS, Sarasua SM, Simensen R, Pietris K, Kimbrell G, Sillence D, et al. Intellectual functioning in alpha-mannosidosis. *JIMD Rep.* 2019;50:44–9.
44. Blomqvist M, Smeland MF, Lindgren J, Sikora P, Riise Stensland HMF, Asin-Cayuela J. β -Mannosidosis caused by a novel homozygous intragenic inverted duplication in *MANBA*. *Molecular Case Studies.* 2019;5:a003954.
45. Clark HM, Duffy JR, Whitwell JL, Ahlskog JE, Sorenson EJ, Josephs KA. Clinical and imaging characterization of progressive spastic dysarthria. *Eur J Neurol.* 2014;21:368–76.
46. Enderby P. Disorders of communication. 2013. p. 273–81.
47. Hudson RR, Kreitman M, Aguadé M. A test of neutral molecular evolution based on nucleotide data. *Genetics.* 1987;116:153–9.
48. Eichmüller OL, Corsini NS, Vértésy Á, Morassut I, Scholl T, Gruber V-E, et al. Amplification of human interneuron progenitors promotes brain tumors and neurological defects. *Science.* 2022;375:eabf5546.
49. Girsakis KM, Stergachis AB, DeGennaro EM, Doan RN, Qian X, Johnson MB, et al. Rewiring of human neurodevelopmental gene regulatory programs by human accelerated regions. *Neuron.* 2021;109:3239–3251.e7.
50. Zhang H, Zhang F, Yu Y, Feng L, Jia J, Liu B, et al. A comprehensive online database for exploring ~20,000 public Arabidopsis RNA-seq libraries. *Mol Plant.* 2020;13:1231–3.
51. van de Geijn B, McVicker G, Gilad Y, Pritchard JK. WASP: allele-specific software for robust molecular quantitative trait locus discovery. *Nat Methods.* 2015;12:1061–3.
52. Field AR, Jacobs FMJ, Fiddes IT, Phillips APR, Reyes-Ortiz AM, LaMontagne E, et al. Structurally conserved primate lncRNAs are transiently expressed during human cortical differentiation and influence cell-type-specific genes. *Stem Cell Reports.* 2019;12:245–57.
53. Benito-Kwiecinski S, Giandomenico SL, Sutcliffe M, Riis ES, Freire-Pritchett P, Kelava I, et al. An early cell shape transition drives evolutionary expansion of the human forebrain. *Cell.* 2021;184:2084–2102.e19.
54. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics.* 2013;29:15–21.
55. Givanna H Putri SAPTPJEPFZ. Analysing high-throughput sequencing data in Python with HTSeq 2.0. *ArXiv.* 2021.
56. Li B, Ruotti V, Stewart RM, Thomson JA, Dewey CN. RNA-Seq gene expression estimation with read mapping uncertainty. *Bioinformatics.* 2010;26:493–500.
57. Zhu A, Ibrahim JG, Love MI. Heavy-tailed prior distributions for sequence count data: removing the noise and preserving large differences. *Bioinformatics.* 2019;35:2084–92.
58. Carbon S, Ireland A, Mungall CJ, Shu S, Marshall B, Lewis S. AmiGO: online access to ontology and annotation data. *Bioinformatics.* 2009;25:288–9.
59. Jassal B, Matthews L, Viteri G, Gong C, Lorente P, Fabregat A, et al. The reactome pathway knowledgebase. *Nucleic Acids Res.* 2019.
60. Kanehisa M, Sato Y, Kawashima M, Furumichi M, Tanabe M. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res.* 2016;44:D457–62.
61. Köhler S, Doecken SC, Mungall CJ, Bauer S, Firth Hv, Bailleul-Forestier I, et al. The Human Phenotype Ontology project: linking molecular biology and disease through phenotype data. *Nucleic Acids Res.* 2014;42:D966–74.

62. Köhler S, Gargano M, Matentzoglou N, Carmody LC, Lewis-Smith D, Vasilevsky NA, et al. The Human Phenotype Ontology in 2021. *Nucleic Acids Res.* 2021;49:D1207-17.
63. Wolf FA, Angerer P, Theis FJ. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.* 2018;19:15.
64. Agoglia RM, Sun D, Birey F, Yoon S-J, Miura Y, Sabatini K, et al. Primate cell fusion disentangles gene regulatory divergence in neurodevelopment. *Nature.* 2021. <https://www.nature.com/articles/s41586-021-03343-3#Sec35>.
65. Field AR, Jacobs FMJ, Fiddes IT, Phillips APR, Reyes-Ortiz AM, LaMontagne E, et al. Structurally conserved primate LncRNAs are transiently expressed during human cortical differentiation and influence cell-type-specific genes. *Gene Expression Omnibus.* 2019. <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE106245>.
66. Benito-Kwiecinski S, Giandomenico SL, Sutcliffe M, Riis ES, Freire-Pritchett P, Kelava I, et al. An early cell shape transition drives evolutionary expansion of the human forebrain. *Gene Expression Omnibus.* 2021. <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE153076>.
67. Zhu Y, Sousa AMM, Gao T, Skarica M, Li M, Santpere G, et al. Spatiotemporal transcriptomic divergence across human and macaque brain development. *Database of Genotypes and Phenotypes.* 2018. https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000755.v2.p1.
68. Aygün N, Elwell AL, Liang D, Lafferty MJ, Cheek KE, Courtney KP, et al. Brain-trait-associated variants impact cell-type-specific gene regulation during neurogenesis. *Database of Genotypes and Phenotypes.* 2021. https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs002493.v1.p1.
69. Castel SE, Aguet F, Mohammadi P, Aguet F, Anand S, Ardlie KG, et al. A vast resource of allelic expression data spanning human tissues. *Genotype Expression project.* 2020. <https://www.gtexportal.org/home/datasets>. File name: phASER_WASP_GTEx_v8_matrix.txt.gz.
70. Starr A, Gokhman D, Fraser H. Accounting for cis-regulatory constraint prioritizes genes likely to affect species-specific traits. 2022. Zenodo. <https://doi.org/10.5281/zenodo.7312168>.

Publisher's Note

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

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions

