Influence of *cis*-regulatory elements on expression divergence in human segmental duplications

Colin J. Shew[1,2†], Gulhan Kaya[1], Sean P. McGinty[1], and Megan Y. Dennis[1-3†]

[1]Genome Center, [2]MIND Institute, [3]Department of Biochemistry & Molecular Medicine, University of California, Davis, CA, USA

[†]Co-corresponding authors:
Colin J. Shew, Ph.D.
University of California, Davis
One Shields Avenue
Genome Center, 4303 GBSF
Davis, CA 95616
Email: colinshew@gmail.com

Megan Y. Dennis, Ph.D.
University of California, Davis, School of Medicine
One Shields Avenue
Genome Center, 4303 GBSF
Davis, CA 95616
Email: mydennis@ucdavis.edu

GLOSSARY OF ABBREVIATIONS

ABC	Activity-by-contact
CRE	*Cis*-regulatory element
HSD	Human-specific segmental duplication
kb	Kilobase pair
LCL	Lymphoblastoid cell line
MPRA	Massively parallel reporter assay

ABSTRACT

Human-specific segmental duplications (HSDs) contain millions of base pairs of sequence unique to the human genome, including genes that shape neurodevelopment. Despite their young age (<6 million years), HSD genes exhibit widespread regulatory divergence, with paralog-specific expression patterns documented across a variety of tissues and cell types. Using long-read expression and epigenomic data, we show that human-specific paralogs tend to have lower activity than the shared, ancestral ones. To systematically characterize the *cis*-regulatory elements (CREs) within HSDs and understand patterns of regulatory change in recently evolved gene families, we conducted a massively parallel reporter assay of 7,160 human duplicated and chimpanzee orthologous sequences in lymphoblastoid (GM12878) and neuroblastoma (SH-SY5Y) cell lines. A large proportion (14–24%) of sequences exhibited differential activity relative to the chimpanzee ortholog (or between human paralogs), mostly with small fold-differences. Combining measured activity levels across all assayed sequences, predicted differences in *cis*-regulatory activity correlated with mRNA levels in SH-SY5Y. Differentially active CREs were validated for *CHRFAM7A*, *HYDIN2*, and *SRGAP2C* that may contribute to paralog-specific expression patterns and thereby to human-specific traits. While we find some changes in CRE activity shared between duplicate paralogs likely driving regulatory divergence in gene expression, consideration of non-shared adjacent sequences to duplications suggests a larger role for altered genome positional effects. In all, this work suggests that functional divergence of duplicated CREs contributes moderately to regulatory divergence of HSD genes and uncovers enhancers that are candidate drivers of human-specific regulatory patterns.

INTRODUCTION

Gene duplication is a major driver of evolutionary innovation, generating novel genetic material on which mutation and selection can act. Duplications are widespread, comprising a substantial proportion of genes across all domains of life, and may enable evolution of new traits by facilitating relaxed selection via genetic redundancy [1–3]. While a majority of paralogs are predicted to become pseudogenes and lost from the genome, the universal presence of gene duplications across species indicates that this process ultimately yields advantageous variation [4]. Expression divergence is thought to be integral to gene retention post-duplication; the loss of *cis*-regulatory elements (CREs) can partition expression of daughter paralogs, eroding their redundancy and subjecting them to purifying selection [5]. Indeed, paralogs exhibiting expression divergence are predicted to persist, and may subsequently accrue additional regulatory or functional changes [6]. Further, gene regulation is highly plastic and itself a major driver of evolution. Alterations to spatiotemporal expression patterns are largely modular and preserve coding sequences and so are less likely to be deleterious [7].

In primates, segmental duplications (SDs) are large blocks of nearly identical sequence (>1 kilobase pair (kb), >90% similarity) that are particularly enriched in African great apes [8]. They are typically interspersed hundreds of kilobases apart and co-occur with additional structural variation [9,10]. Thus, SDs hold a strong potential to alter gene regulation by duplicating and translocating both genes and CREs. By contrast, tandem duplications are more common in other mammals but leave daughter genes in a similar regulatory environment [9,10]. Human-specific segmental duplications (HSDs) are SDs unique to our species, which are consequently young (<6 million years) and highly similar (>99% nucleotide identity). In spite of this, genes within HSDs exhibit tissue-specific expression patterns across diverse

primary tissues and cell lines, and a majority of HSD gene families display quantitative and spatial expression patterns specific to derived paralogs [11,12]. Some of these changes to gene regulation likely underlie human-specific traits, such as innovations in the development of the cerebral cortex. For example, human-specific *ARHGAP11B* drives basal neural progenitor proliferation and increased cortical neuron numbers in mammalian models [13–15], and is preferentially expressed in the germinal zone of the developing brain, while the ancestral *ARHGAP11A* is expressed at higher levels and more broadly throughout the neocortex [12]. While *ARHGAP11B* has also attained novel biochemical functions [16], these differentiated expression patterns are also critical for human brain development.

The mechanisms underlying such cell-type specificity have not been thoroughly characterized. *Cis*-regulatory changes were broadly implicated in a study of HSD gene families that found reduced expression conservation in 5′-truncated genes, which had either lost their original or exapted novel promoters [17]. Distal elements likely also contribute; we previously reported evidence for paralog-specific regulatory contributions from adjacent, non-duplicated genomic regions, as well as sequence-driven changes to the activity of duplicated enhancers [18]. However, only a few gene families were functionally tested, and one of them (*ARHGAP11*) showed discordant mRNA levels and cis-regulatory activity between paralogs. This highlights the need to more comprehensively dissect the regulatory landscape of many HSD gene families, in order to gain mechanistic insight into how gene regulation diverges on short evolutionary timescales and might contribute to human-specific traits.

In this work, we investigated *cis*-regulation in HSDs using a massively parallel reporter assay (MPRA), which measures CRE activity of integrated reporter constructs with a sequencing-based readout [19]. We tested 7,160 paralogous human and orthologous chimpanzee sequences selected from 2,675 candidate CREs in LCLs and primary brain in two cell lines: the lymphoblastoid cell line (LCL) GM12878 and neuroblastoma SH-SY5Y. Individual paralogous changes were less conserved in derived versus ancestral CREs, with mostly small effect sizes. While integrating CRE activities partially explains differential expression between HSD paralogs, consideration of long-read derived epigenomic data within and outside of HSDs suggests that adjacent non-paralogous sequences may be more impactful in influencing expression divergence.

RESULTS

Human-duplicated genes are differentially expressed

Our previous work characterizing a subset of HSD genes identified divergent expression patterns over a short evolutionary time span [18]. Leveraging recent work identifying additional duplicated gene families with at least one paralog unique to humans [20], we sought to expand our understanding of gene expression and regulation in HSDs. We also generated long-read RNA-seq data (PacBio Kinnex) in GM12878 and SH-SY5Y. Considering 605 duplicate gene families containing 1,394 genes, long- and short-read transcript abundance [21–24] (Table S1) are strongly correlated in both cell types (r=0.85 for GM12878 and 0.88 for SH-SY5Y; Figure S1). While the long-read quantification is expected to be more accurate for distinguishing paralogs from each other, it is less sensitive overall, with a larger proportion of genes not detected (25.3% with long reads versus 12.9% with short reads). Across duplicated gene families, we saw an enrichment of highest-expressed paralogs in regions with synteny to the chimpanzee ortholog (p=2.7□10^{-4} for GM12878 Kinnex; p=1.7□10^{-3} for GM12878 short-read; p=2.8□10^{-6} for SH-

SY5Y Kinnex; p=2.6□10^{-3} for SH-SY5Y short-read) (Figure 1). This is in agreement with our previous work showing ancestral paralogs exhibit highest expression, as opposed to human-specific, derived genes [18]. To the same end, we also found an enrichment of coding genes as opposed to pseudogenes among highest-expressed paralogs (p=2.7□10^{-8} for GM12878 Kinnex; p=1.5□10^{-3} for GM12878 short-read; p = 2.0□10^{-10} for SH-SY5Y Kinnex; p=7.7□10^{-3} for SH-SY5Y short-read; Figure 1). Taken together, these analyses confirm that human-specific genes exhibit novel expression patterns, and the ancestral paralogs tend to retain a conserved function.

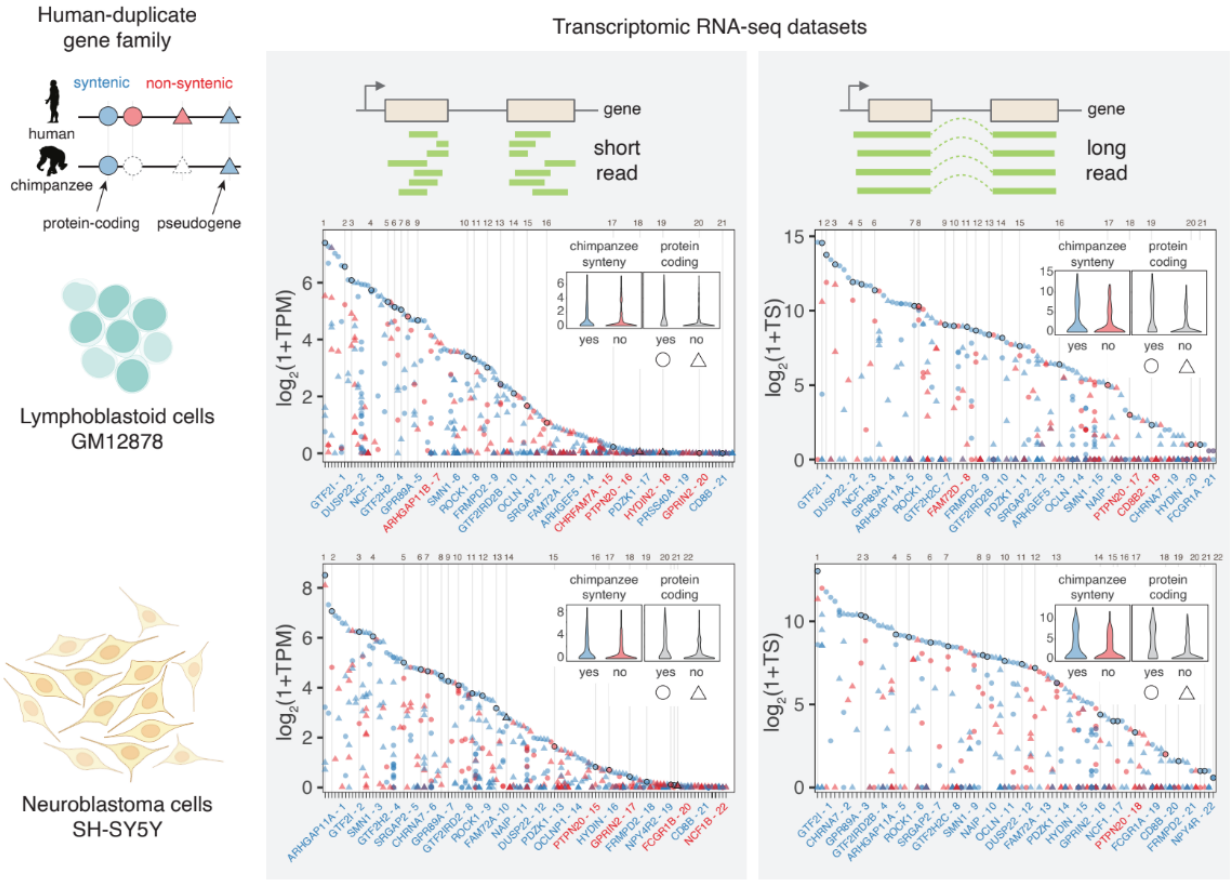

Figure 1. Human-duplicated genes are differentially expressed. The highest-expressed paralogs in each gene family (top point per column and inset violin plots) are more likely to be syntenic with chimpanzee (blue), and more likely to be protein-coding (circle). Comparisons were performed with transcriptomic data from GM12878 lymphoblastoid (top) and SH-SY5Y neuroblastoma cells (bottom) sequenced with short-read Illumina (left) and long-read PacBio Kinnex (right) reads. Gene families are represented on the x-axis, and transcript abundance on the y-axis (transcripts per million (TPM), TS (total transcript)). Numbered are high-confidence human-duplicated gene families with highest-expressed paralog labeled on the x-axis and colored to indicate syntenic, likely ancestral (blue) or non-syntenic, likely-derived (red).

MPRA identifies active regulatory elements in human duplications

To systematically quantify *cis*-regulatory activity in HSDs, we designed an MPRA to directly compare paralogous human and orthologous chimpanzee sequences from 67 genes in 26 gene families (Figure 2A; Table S2, S3). These gene families are well represented in high-quality assemblies [25,26], are found at elevated copy numbers in all humans surveyed [20], and have established ancestral and derived identities

[11]. In addition, in our expression analysis, 72% (18/25) of these gene families were detected with short reads (>1 TPM) in GM12878 or SH-SY5Y, while 68% (17/25) were with long reads (>1 CPM) (Table S4). Due to the known role of HSD genes in brain development, we chose to focus on *cis*-regulation in primary fetal and adult prefrontal cortex. We also included LCLs, whose accessibility has made them foundational for comparisons among humans and non-human primates. HSDs are typically excluded by standard bioinformatic approaches, as reads aligning to these regions have poor mapping qualities. To circumvent this, we permissively identified candidate CREs by re-mapping histone H3K27ac chromatin immunoprecipitation sequencing (ChIP-seq) and chromatin accessibility data [27–32] to the human reference (GRCh38) with multiple alignments. CREs in ancestral loci were divided into 200mer "tiles" with 100-bp overlap, and homologous tiles in the human and chimpanzee (panTro6) genomes were identified by sequence alignment. In total, we synthesized 8,145 test sequences and 500 scrambled negative controls (Table S5), which were packaged into lentiviral vectors and assayed in GM12878 and SH-SY5Y cells according to the lentiMPRA protocol [33]. Transduction efficiencies were comparable to previous lentiMPRA studies, with multiplicity of infection (MOI) of ~40 for SH-SY5Y and ~13 for GM12878. Technical replicates (N=3 per cell line) exhibited high reproducibility (mean pairwise r=0.93 and 0.93 for within-cell type comparisons of DNA and RNA libraries, respectively; Figure S2) [25,26].

We first quantified enhancer activity for all tested CREs in both cell lines using the estimated transcription rate (□), modeled from the integrated (DNA) and expressed (RNA) barcode counts (MPRAnalyze [34]). In SH-SY5Y, 1,056/7,760 (14%) tested sequences were active relative to the negative controls (median absolute deviation $p<0.05$), while in GM12878, 1,850/7,750 (24%) sequences were active (Figure 2C; Tables S6–S7). Of these, 472 (19%) of 2,434 active sequences were active in both cell types (Figure 2B). Notably, higher activity was measured overall in SH-SY5Y (median □=2.50, versus 0.73 in GM12878; Figure S3B), likely due to the three-fold greater infection rate achieved in this cell line. Activity measurements were correlated for overlapping tiles (r=0.21 for GM12878 and 0.43 for SH-SY5Y; Figure S3A), but we opted against combining measurements across tiles because any target nucleotide was covered by at most two constructs, and the 100-bp offset is much larger than transcription factor binding sites.

Sequences near transcription start sites (TSSs) were enriched for activity in both cell types (1.35-fold higher proportion of active sequences for promoters in GM12878; 3.46-fold in SH-SY5Y; Fisher's exact test $p<0.02$). While we observed a correlation between promoter activity (mean activity score for all active tiles) and mRNA expression driven by the contrast between active and inactive promoters (Figure S4A), MPRA promoter measurements did not predict differences in ancestral/derived gene pairs (Figure S4B). This suggests a role for distal elements in shaping expression patterns.

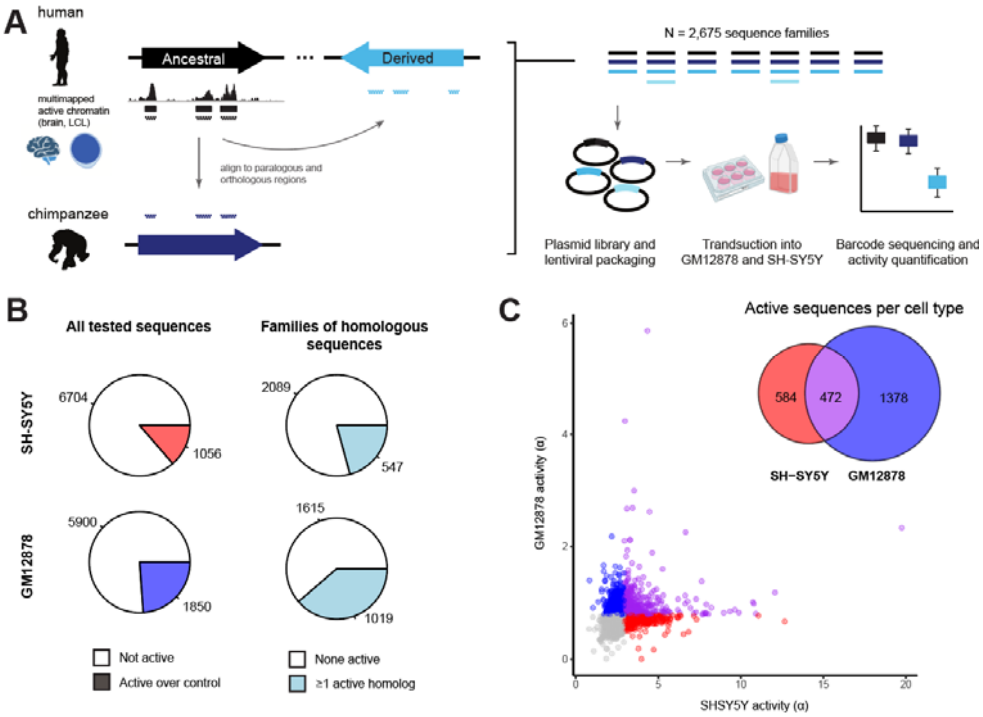

Figure 2. MPRA quantifies paralog-specific regulatory activity of duplicated CREs. (A) MPRA tiles were designed to multi-mapped chromatin accessibility and histone H3K27ac ChIP-seq data previously published from human primary brain (fetal and adult) and lymphoblastoid cell lines. Peak regions were tiled with 200mers at 2x density in ancestral loci, and homologous positions were identified by alignment to the human and chimpanzee genomes. All unique sequences were synthesized in an oligonucleotide pool and cloned into a lentiviral vector. MPRA was performed in the neuroblastoma cell line SH-SY5Y and lymphoblastoid cell line GM12878. **(B)** Pie charts depict the number and proportion of active sequences (left) and sequence families containing at least one active sequence (right), for each cell type. **(C)** Sequence activity (estimated transcription rate, □) is plotted type for 7,750 tiles measured in both assays. Sequences active over negative controls are colored red (SH-SY5Y), blue (GM12878), and purple (both). The Venn diagram indicates the number of sequences in each category.

Species- and paralog-specific CRE activity

We first examined sequence activity across species. Considering all human-chimpanzee sequence pairs with at least one active element, 35% (357/1,017) were differentially active in SH-SY5Y, and 11% (220/1,935) were differentially active in GM12878 (Figure S5; Tables S8–S9). We found an expected relationship between sequence divergence and the magnitude of differential activity between species in both cell types (r=-0.07, $p<1\times10^{-3}$; Figure S6). However, the strongest differentially active sequences tended to have the highest similarity with chimpanzee, demonstrating that a small number of substitutions can have large functional consequences. Based on our findings that ancestral genes generally exhibit higher expression and increased concordance with chimpanzee orthologs [18] (Figure 1), we next hypothesized that activity losses preferentially occur in derived CREs. We partitioned human sequences by ancestral or derived status but found no enrichment of activity gains or losses (considering chimpanzee a proxy for ancestral activity) in either group (hypergeometric test). However, the measured activities of ancestral sequences correlated significantly better than derived sequences with their chimpanzee orthologs (Fisher's z-test, p=0 in SH-SY5Y, p<0.01 in GM12878; Figure 3A). Thus, while *cis*-regulatory activity in HSDs is largely conserved between humans and chimpanzees, human-specific derived regions may experience relaxed constraint on CRE activity.

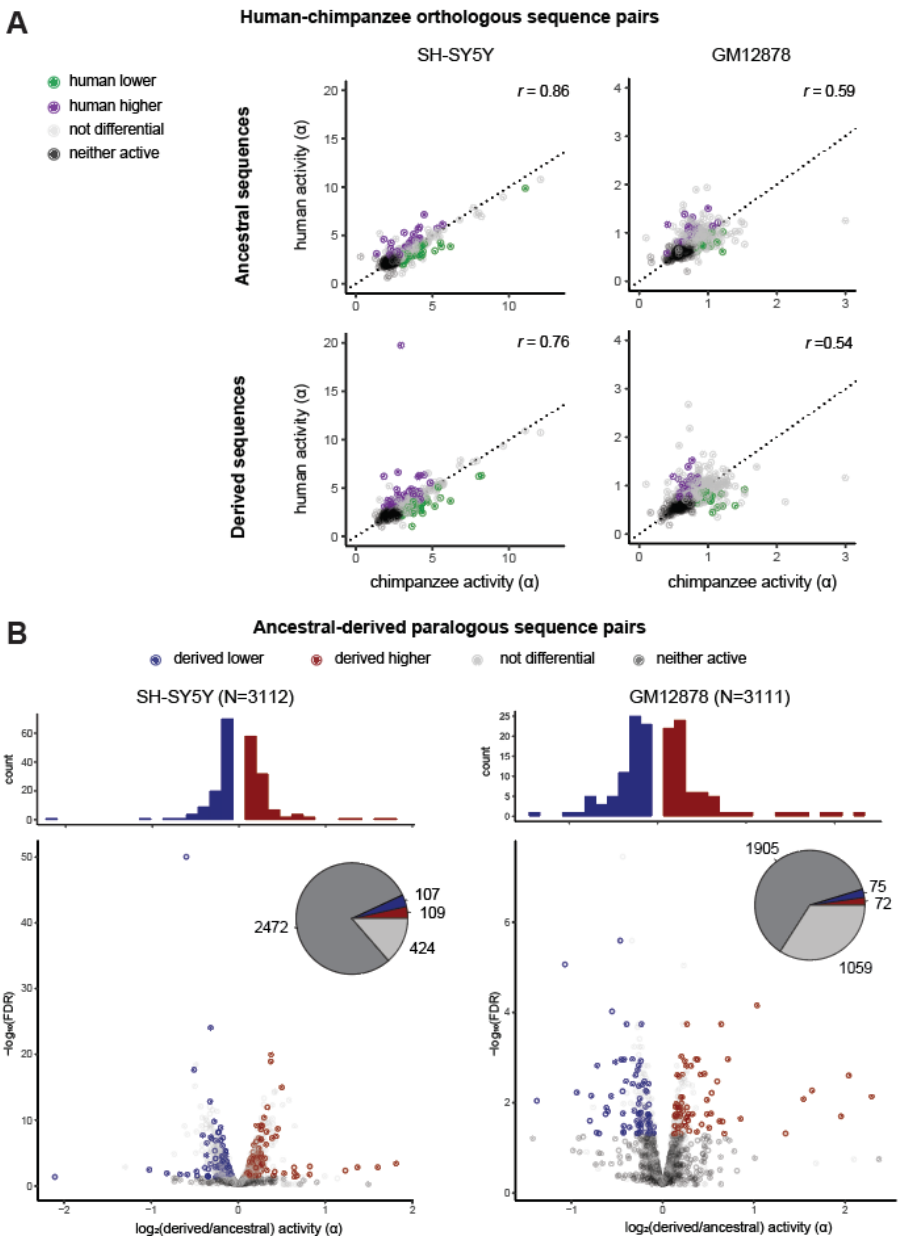

Figure 3. Differential activity between homologous sequences. (A) Scatterplots depict the activity of human and orthologous chimpanzee sequences, partitioned by human evolutionary status (ancestral above, derived below). Each sequence pair is colored by differential activity relative to chimpanzee: green (human lower), purple (human higher), gray (not differential), black (neither sequence active). The dotted line marks the identity y=x, and Pearson's correlation coefficient is printed on each plot. **(B)** Volcano plots depict differential activity between paralogous human sequences, with each unique derived-ancestral pair plotted separately. Points are colored by differential activity relative to the ancestral human paralog: blue (derived lower), red (derived higher), gray (not differential), black (neither sequence active). The marginal histograms depict the distribution of log-fold differences for differentially active sequences.

Comparing activities between human duplications, we found a similar proportion of derived-ancestral pairs were differentially active by the same definition: 216/640 (34%) of pairs in SH-SY5Y, with a median 1.13-fold difference; in GM12878, 147/1,206 (12%) were differentially active, with a median 1.20-fold difference (Figure 3B–C). While most paralogs exhibited modest divergence in activity (1.13-

fold and 1.20-fold differences in SH-SY5Y and GM12878, respectively), *SRGAP2* and *FRMPD2* contained strong differential human-specific sequences (>2-fold; Tables S10–S11) that were also concordant in both cell types, suggesting possible single CRE drivers of paralog differences. Targeting the putative CRE from *SRGAP2* (chr1:206333093-206333293), we used a luciferase reporter assay to validate that a larger homologous region containing the human-specific *SRGAP2B* (0.43-fold) and *SRGAPC* (0.36-fold) tile sequences were significantly less active than the ancestral *SRGAP2* in SH-SY5Y, with no differential activity observed in GM12878, matching MPRA results (Figure S7; Table S12).

Taking this same approach to test larger regions with luciferase reporters, we used 1-kb windows in HSDs to score total measured MPRA activity, secondarily ranking by the maximum fold difference of any contained tile, and manually curating select differential elements in both cell types (Figure S8). In the *CHRNA7* locus, we tested two candidate regions. The first (chr15:32183316-32184316) was predicted to have higher activity in the derived *CHRFAM7A*, which was true for GM12878 (>1.3-fold difference; Figures 4, S10; Table S12). The second was expected to have lower activity in SH-SY5Y, which was also verified by luciferase (<0.7-fold difference). In addition, an intronic region within ancestral *HYDIN* was predicted to have higher activity in the derived *HYDIN2* in both cell types. Reporter activity was significantly higher for both homologs in GM12878, and for *HYDIN2* in SH-SY5Y (>3.8-fold difference; Figures S9, S10; Table S12). Altogether, these experiments demonstrate paralog-specific activity for enhancers in a more biologically relevant sequence context. The reporter activities from these expanded constructs were generally in agreement with the MPRA, which measured activity of 200mer tiles in isolation.

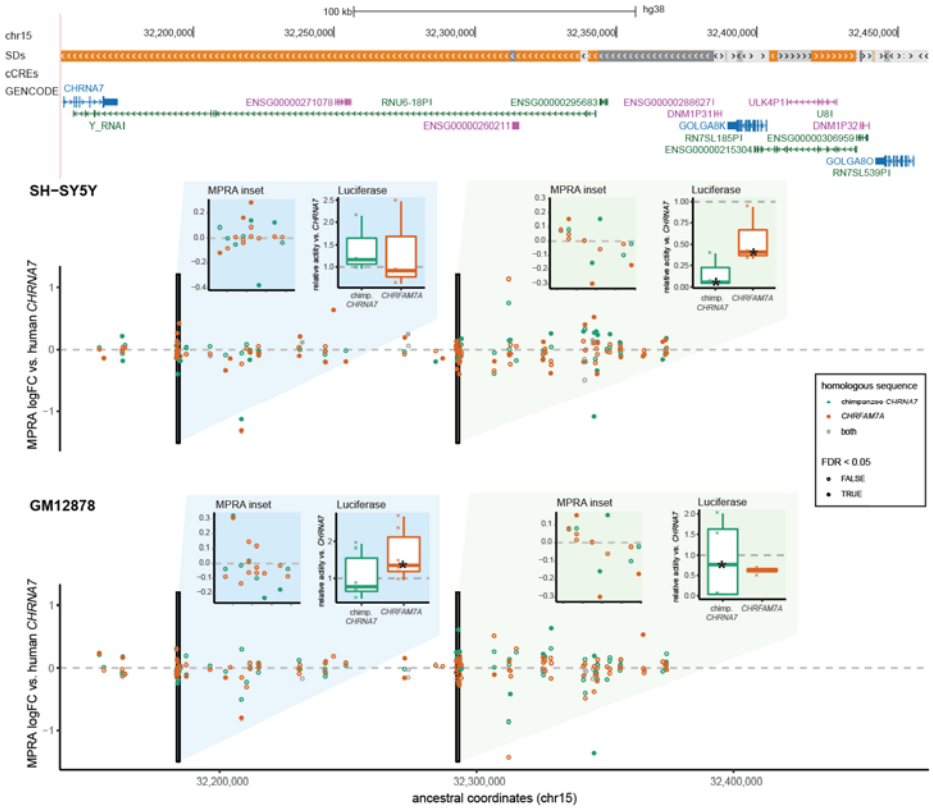

Figure 4: Paralog-specific regulation in the *CHRNA7* locus (chr15q13.3). A 143-kb region (chr15:32,153,206-32,460,660) is shown. Segmental duplications (SDs) are colored orange for >99% sequence identity, and gray for

<98% identity. No ENCODE candidate *cis*-regulatory elements (cCREs) are annotated in this locus. The MPRA data are visualized as activity relative to the human ancestral locus, and plotted on these coordinates. Each point represents a 200mer tile, and tiles are color by which homolog they correspond to (teal: chimpanzee *CHRNA7*; orange: human-specific *CHRFAM7A*; gray: identical in both) and filled in if they scored as differentially active. Two 1-kb regions are highlighted in the insets; the left panel shows a zoomed in plot of the MPRA measurements per tile, and the right shows a relative luciferase activity of the entire sequence, as assayed in the corresponding cell type. Asterisks on the box plots mark significantly different activity with respect to the human ancestral sequence. See also Table S12.

Collective analysis of CREs on paralog expression

Due to modest effect sizes of individually tested CREs in the MPRA, we next considered whether their collective activity might explain paralog-expression divergence. Building on the activity-by-contact (ABC) score [35], we developed a similar metric using activity levels directly measured from the MPRA (Methods; Figure S11A). For expressed derived-ancestral gene pairs (either TPM>1), expression divergence in the short-read data was significantly correlated with the summed ABC score ratio in SH-SY5Y ($p<0.05$, $r=0.70$; Figure S11B; Supplementary Note 1). This comparison was robust to various cutoffs for considering CREs active or differential and showed the best correlation for more stringent criteria. No relationship was observed using the long-read expression data, however, perhaps reflecting reduced sensitivity of Kinnex to detect transcripts (Figure S11C). The SH-SY5Y MPRA ABC score ratio also correlated with paralog expression divergence quantified in certain primary brain cell types (i.e., the cortical plate, cortical neurons without basal contact, and the outer subventricular zone; Figure S12), reflecting functional relevance beyond the cell lines assayed.

We searched for trends in CRE activity that could explain this, namely in strength, number, or location. In both cell types, neither measured CRE activity nor the number of CREs differed between ancestral and derived regions, but active CREs were slightly closer to ancestral TSSs than derived ones ($p=0.03$, Wilcoxon rank-sum test; Figure S13). This suggests a trend toward loss of proximal CRE activity in derived paralogs. Because promoter strength alone did not predict expression differences (Figure S4), the full complement of CREs can thus better explain differential mRNA levels across HSD genes in SH-SY5Y. Indeed, the median fraction of contribution of the top CRE to each HSD gene was 5%, and even genes with few (<5) CREs measured were not typically dominated by their promoters (Figure S14). Together, these results indicate that changes to promoter activity are not the primary driver of HSD expression divergence, and that paralogous CREs may influence this process.

For comparison, we also calculated the predicted difference using the ABC score as published, with chromatin accessibility and H3K27ac ChIP-seq. We aligned datasets from GM12878 [36] and SH-SY5Y [37,38] to the human reference (GRCh38), allocating multiple alignments probabilistically [36,39], enabling the calculation of the ABC difference metric across HSDs and adjacent, unique regions (Methods). For SH-SY5Y, we found a correlation of the ABC score ratio with short-read expression divergence ($p=0.05$, $r=0.42$), but this did not hold when subsetting to only the duplicated or unique space (Figure S15A–C). The long-read expression data for both cell types also trended in the expected direction for all and non-duplicated peaks. We also scored the exact regions tested by our MPRA, in place of peaks derived from the epigenomic data, and found a weaker correlation than the MPRA analysis on its own ($p=0.06$, $r=0.39$) (Figure S15D). Thus, the measured regulatory activity from the MPRA appears more informative than proxies based on chromatin state. The whole-genome chromatin activity data, however,

suggest that non-duplicated CREs have a greater impact on expression divergence.

Long-read epigenomics improves expression predictions

Recent advances in long-read epigenomic assays have better enabled differentiation of regulatory information between human duplications [40,41]. Using published DiMeLo-seq datasets from GM12878 [42], we identified expected correlations between DNA methylation (5mC), histone modifications (H3K4me3, H3K27ac) and HSD gene expression, including derived-ancestral differences (Figure S16A; Supplementary Note 2). Globally, H3K4me3 and 5mC signals were higher in chimpanzee-syntenic (likely ancestral) than non-syntenic (likely derived) regions, (respective 1.08- and 1.19-fold difference of medians), consistent with greater promoter activity and transcription-associated gene body methylation (Figure S16B). Interestingly, 5mC was not different when considering only promoters, counter to previous reports that CpG methylation tracked with expression in a few duplicated gene families [43]. Taken together, these results are in accord with observed differences in RNA levels.

We again calculated an ABC score per gene using DiMeLo signal at the same peaks defined with ChIP-seq. In agreement with the ChIP-based analysis in SH-SY5Y, ABC score ratios from DiMeLo in GM12878 correlated with expression divergence (short-read RNA, Figure 5). This result held when subsetting to only peaks in unique regions, but no correlation was found using HSD peaks only. This again suggests a larger contribution of adjacent, non-duplicated sequence to expression differences among HSD paralogs. We expect long-read epigenomic datasets will be critical for disentangling the mechanisms underlying paralogous expression divergence in other loci and cell types.

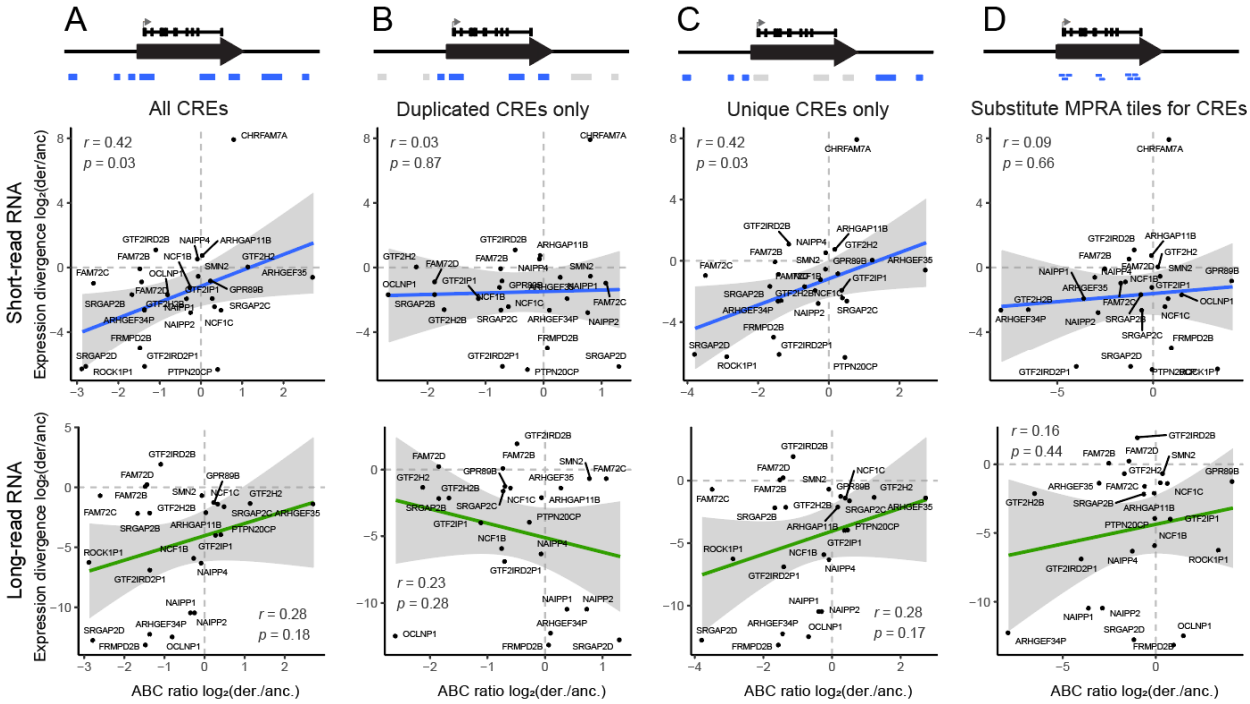

Figure 5. Differential expression predictions using DiMeLo-seq in GM12878 with unique and duplicated CREs. For each derived-ancestral (der/anc) gene pair, expression divergence was correlated to the ratio of summed ABC scores calculated from DiMeLo-seq. ABC scores were calculated from H3K27ac signal, using the following peak sets: **(A)** all CREs as defined by our ChIP-seq analysis within 5 Mb; **(B)** CREs within HSDs; **(C)** CREs outside

of HSDs; or **(D)** the exact positions tested by MPRA. The Pearson correlation and significance of the linear relationship is printed on each plot.

DISCUSSION

In this study, we present the first high-throughput quantification of regulatory activity in duplicated regions. Broadly, HSD sequences exhibit few strong differences in CRE activity. This is unsurprising given the small number of substitutions between human paralogous and chimpanzee orthologous sequences, and comparable to the fraction of differentially active sequences seen in MPRAs of polymorphic and species-specific variants [44,45]. However, this stands in contrast to examples of multi-fold differences of HSD promoter and enhancer activity demonstrated previously [18]. One explanation may be the size of assayed sequences; MPRA libraries constructed from synthetic fragments are currently limited to ~200 base pair insert sizes, while previously tested CREs in HSDs assayed fragments >1000 base pairs in size. Indeed, one of the regions we tested (*HYDIN*) in a luciferase assay displayed an almost four-fold difference in activity, while individual tiles measured below two-fold difference by MPRA (Figure S15). It is therefore likely that the synergistic effects of nearby enhancers are not completely captured in this design.

Despite these limitations, MPRA is well suited for directly comparing homologous sequences. Assessing CRE activity between humans and chimpanzees, we found that derived sequences were more weakly correlated than those from ancestral paralogs. This is in line with the predicted relaxation of selection on derived genes, though gains and losses of activity were seen in similar proportion. To account for both the strength and distance of duplicated elements across human paralogs, we used MRPA measurements to develop a metric based on the ABC score. This integrated score predicted expression differences, albeit only using the short-read RNA quantification. Meanwhile, consideration of a larger complement of CREs using chromatin-based proxies for activity suggests adjacent, non-paralogous regions are collectively more impactful. Overall, these results paint a picture in which sequence divergence of human-duplicated regulatory elements influences but does not primarily drive changes to expression of paralogous genes.

We also highlighted individual differentially active elements that may contribute to paralog-specific gene regulation. In the *CHRNA7* locus, we tested two regions containing multiple differentially active MPRA tiles: the first was significantly more active in *CHRFAM7A*, and the second was less active in SH-SY5Y, tracking with RNA expression and CRE activity in the MPRA. *CHRNA7* codes for the α7 neuronal nicotinic acetylcholine receptor and resides in a genomically unstable region in chromosome 15q11-q13 where copy-number variants are associated with neurodevelopmental features, including intellectual disability, autism, and epilepsy [46]. Human-specific *CHRFAM7A*, meanwhile, is a partial duplication and fusion gene linked to schizophrenia and shown to interact with *CHRNA7* as a dominant negative [47]. Accordingly, expression modulation is likely critical to proper dosage. If these CREs indeed regulate *CHRNA7*, the second may play a role in reducing *CHRFAM7A* expression in the brain, as indicated by its lower activity in SH-SY5Y. Understanding the cell type-specific activity of these and other CREs will be critical to unraveling their function.

This study focused on comparisons of CRE activities, leaving the contributions from changes in chromatin conformation at HSD and orthologous chimpanzee loci an open question. Identification of promoter-enhancer loops would enable more confident assignment of CREs to target genes, identification

of paralog-specific chromatin contacts across duplication breakpoints, and refinement of ABCD regulatory predictions by using true interaction frequency in place of a genomic distance proxy. While published methods for allocation of multimapping Hi-C reads exist, they still exhibit reduced performance in highly identical HSD regions [48]. Long-read based chromatin conformation capture methods such as Pore-C [49] and CiFi [50], with larger fragment sizes, will be necessary to confidently identify loops involving HSD regions. Finally, machine learning methods are rapidly improving prediction of epigenetic properties from DNA sequence alone. Chromatin conformation, accessibility, histone methylation, and RNA expression can all be compared *in silico*, though accurate variant effect prediction is currently an area of active development, and comparisons between nearly identical HSD paralogs would also require validation.

In all, this work measured the regulatory activity of thousands of candidate regulatory sequences distinguishing recent human duplications. We found that a large proportion of them are active, and while few individual 200mers exhibit strong differential activity, some contribute in combination to regulatory patterns of HSD genes such as *CHRFAM7A*, *HYDIN2*, and *SRGAP2C*, perhaps driving unique features of our species. Integration of additional information, such as new long-read epigenomic datasets, chromatin conformation, and single nucleotide and structural polymorphism will be necessary to gain a more complete picture of gene regulation within HSDs.

MATERIALS AND METHODS

RNA Library Preparation

Total RNA was extracted from GM12878 and SH-SY5Y cell lines using the Qiagen RNeasy Plus Kit following the manufacturer's protocol. RNA integrity was confirmed using the Agilent 2100 Bioanalyzer, ensuring RIN values ≥7.0. Full-length RNA libraries were prepared using the PacBio Kinnex Full-Length RNA Kit (103-238-700), including cDNA synthesis, Kinnex PCR amplification, and library cleanup. Libraries were quantified with the Qubit DNA HS Assay Kit and size-checked on the Agilent Femto Pulse system, which showed average insert sizes of ~15kb for both GM12878 and SH-SY5Y, before sequencing on the PacBio Revio platform.

RNA-seq analysis

RNA-seq data were obtained for GM12878 (ENCODE ENCSR000AEC, ENCSR000AEE, and ENCSR000CVT) and SH-SY5Y [21]. Transcripts were quantified with Salmon v1.9.0 [51] with the flags "--validateMappings --gcBias", the telomere-to-telomere CHM13 v2.0 CAT/Liftoff transcriptome, and the CHM13 v2.0 assembly as decoy sequence. All identical transcripts were removed from the transcriptome prior to index construction. Transcripts per million (TPM) values were summed to the gene level using tximport [52]. For relative expression analyses, derived-ancestral gene pairs were considered if at least one paralog was expressed >1 TPM. Expression divergence was calculated as the $\log_2$ ratio of derived/ancestral expression, with a pseudocount one order of magnitude smaller than the smallest nonzero value added to each.

Long-read PacBio Kinnex analysis

Initial processing of PacBio Kinnex data was done with the Isoseq workflow [53]. HiFi reads were first segmented using PacBio concatenated read splitter Skera [54] at adapter positions to generate segmented

reads. Lima is then used to identify and remove barcoded Iso-Seq primers. The output of lima are full length (FL) reads. These FL reads are then processed with `Isoseq refine` to identify and trim poly(A) tails, as well as identify and remove concatemers. The output of Isoseq refinement is full-length, non-concatemer (FLNC) reads. These FLNC reads are then quantified using Oarfish [55] using the same identical transcript-filtered CHM13 v2.0 CAT/Liftoff transcriptome described above. The estimated transcript counts are then summed to get gene-level counts, which were converted to counts per million (CPM) for expression analyses. No length normalization was performed because each transcript detected is expected to yield one FLNC read.

MPRA oligo design

Candidate CREs were identified from H3K27ac ChIP-seq, ATAC-seq, or DNase-seq data in LCLs, fetal cortex, and adult prefrontal cortex with data from the following publications: [27–32]. In the case of Bryois et al., only control samples were used. Reads were trimmed with Trimmomatic [56] aligned to GRCh38 allowing for multiple mapping with bowtie v1.1.2 [57] with the flags "-a -v2 -m 99". Peaks were called using MACS2 v2.1.2 [58] with a 5% FDR and the following additional settings for chromatin accessibility data: "--nomodel --shift -100 --extsize 200 --broad". Peaks were called on all samples in each study and combined by reporting genomic regions with nonzero coverage in a minimum number of samples using `bedtools genomecov` [59]: 3/10 LCL H3K27ac, 52/204 LCL DNase, 2/3 fetal cortex H3K27ac, 3/3 fetal ATAC-seq (cortical plate or germinal zone), 1/1 adult prefrontal cortex H3K27ac (peaks called jointly from samples HS1 and HS2), and 102/137 adult prefrontal cortex ATAC-seq. Cutoffs were chosen based on manual inspection of reproducibility and to equalize sequences represented from LCLs and brain.

To design the test oligos, all peaks within 400 bp were merged, and the resulting intervals were expanded by 100 bp to allow for full 2x tiling of the region. 200mer tiles were defined using `bedtools makewindows`. To match orientation of promoters, tiles were assigned to the strand of the nearest annotated feature (GENCODE v32), or randomly selected if multiple features were nearest. Tiles were intersected with ancestral HSD regions and aligned with BLAT [60] to a custom reference consisting of GRCh38 with non-HSD sequence masked, in addition to the missing portions of the DUSP22B and GPRIN2B contigs (ancestral loci and contigs from [11]). Alignments were then filtered in a two-step process to keep only paralogous positions: (1) keep if there were ≤4 hits >190 bp with >95% identity; (2) else keep if there were ≤4 hits >195 bp with >98% identity. To focus on paralogous differences, sequences for which all alignments were identical were discarded. To identify chimpanzee orthologs, human tiles were lifted over to the panTro6 assembly and filtered for being with 5% (10 bp) of the original 200 bp size (93% success rate). All tiles with this limit were trimmed or expanded to 200 bp; 91% of alignments were within 1 bp difference. FASTA sequences were extracted in a strand-aware manner from GRCh38 or panTro6 and deduplicated. Finally, a universal priming sequence (5'-AGGACCGGATCAACT-[200mer]-CATTGCGTGAACCGA-3') to the pLS-SceI vector (Addgene 137725) was added to all 200mers. Sequences were filtered to remove AgeI and SceI restriction sites, as well as those made singletons after filtering, leaving 8,145 test oligos. In addition, 500 ancestral tiles were sampled and scrambled to generate negative controls. All 230mer sequences were synthesized by Agilent Technologies.

MPRA library preparation and sequencing

MPRA was carried out according to the lentiMPRA protocol as described [33] with 3 technical replicates. We generated the lentivirus library by transfecting HEK293T cells with the plasmid library. To titrate the lentivirus, GM12878 and SH-SY5Y were plated into a 24-well plate and infected with varying volumes of the virus in each well. In these experiments, transduction efficiency was ~10% for GM12878 and ~30–40% for SH-SY5Y cells, as determined by GFP-positive cell counts on the Countess II FL. Following a 3-day incubation, genomic DNA was extracted from each well. Then, the MOI was calculated for each viral concentration via qPCR using plasmid backbone and genomic DNA control primers, as described in the protocol. From these titration results, the final infection conditions were set as follows: For SH-SY5Y experiments, 2.5 million cells were infected at a multiplicity of infection (MOI) of 40 in DMEM/F12 Glutamax medium (Thermo Fisher Scientific 10565018) with 10% FBS (Thermo Fisher Scientific, 10-438-026) and 2.5 µg/mL protamine sulfate (Spectrum chemical TCI-P0675-1G). For GM12878, 55 million cells were transduced at an MOI of 13 in RPMI (Fisher Scientific 11875135) with 10% FBS and 8 µg/mL polybrene (Millipore Sigma MILL-TR-1003-G). Cells were incubated at 37°C in 5% CO_2 and nucleic acid extractions were conducted 48 hours after infection. Barcode association was performed on a PE150 NextSeq mid-output run. Barcode counting from DNA and RNA libraries was performed on three PE15 NextSeq high-output runs.

MPRA data analysis

Barcode-insert association was performed with MPRAflow [33] using "--mapq 1" and all other parameters set to default. We manually confirmed that promiscuous barcodes had a majority assignment to one insert and were not mixed between paralogs. Barcodes were counted with the `count` utility and passed to MPRAnalyze [34]. Counts were depth-normalized with the total sum, and active sequences were defined per cell type against the negative controls with the following model: "dnaDesign = ~ barcode + replicate, rnaDesign = ~ 1". To identify differential activity of homologous sequences, counts matrices were reformatted to combine all homologs in the same row for model fitting, to use homolog identity used as a covariate. Models were fit in "scale" mode with the following design: "rnaDesign = ~homolog, reducedDesign = ~1". The alpha value used as a measure of sequence activity refers to the estimated transcription rate from the model. Differentially active homologs were defined in two ways: (1) against the chimpanzee ortholog to define relative gains/losses in human; and (2) against the human ancestral sequence to quantify divergence of ancestral-derived pairs. Differential activity was assessed in a pairwise manner with the `testCoefficient()` function (Wald test), and differential pairs were defined at a 5% FDR using the Benjamini-Hochberg procedure.

ABC score

Regulatory activity for each TSS was calculated using the ABC score framework defined in [35], substituting the estimated transcription rate from MPRAnalyze for activity and using (genomic distance)$^{-0.7}$ as a proxy for contact frequency. CREs within 5 Mb of each TSS were considered for the calculation. While ABC values per CRE are typically normalized and interpreted as a relative contribution to a given TSS, we compared summed ABC scores between paralogous genes. Regulatory divergence was scored in a similar manner to expression divergence: a pseudocount of one order of magnitude below the smallest nonzero value was added to all ABC values, and the $\log_2$ ratio of scores was calculated for each derived-ancestral gene pair.

ChIP-seq and ATAC-seq analyses

For calculation of ABC scores as published (using H3K27ac ChIP and ATAC data) and comparison with the MPRA, we aligned short reads (Table S1) to GRCh38 with bowtie v1.1.2 [57], allowing for multiple mapping with the flags "-a -v2 -m 99". CSEM was used to assign a probability to each alignment position on the basis of the local unique mapping rate [36,39]. Reads were then allocated to their most likely position by selecting the alignment with the highest posterior probability, or at random if all mappings were equally likely. Peaks were called on the ChIP data using input controls at a 5% FDR using MACS2 [58]. For ATAC and DHS data, the additional parameters "--nomodel --shift -100 --extsize 200 –broad" were added. Read counts per peak were determined with bedtools coverage [59].

Luciferase reporter assays

1-kb fragments were synthesized by Azenta Life Sciences and cloned into the pE1B vector using the NEBuilder HiFI Assembly Master Mix (NEB #E2621). GM12878 and SH-SY5Y cells were cultured as described above. SH-SY5Y cells at 70-90% confluence were cotransfected with 50 ng pRL-TK and equimolar test construct using Lipofectamine™ 3000 (ThermoFisher L3000001). Cells were lysed 48 hours post-transfection and assayed with the Dual-Luciferase Reporter Assay System (Promega E1910) and the Tecan Spark plate reader according to the manufacturer's instructions. GM12878 cells were split 48 and 24 hours prior to transfection and electroporated with 12.5 μg pRL-TK and equimolar test construct, brought to 100 μL with RPMI. For each transfection, 12.5 million cells were electroporated with the Neon system (ThermoFisher) using buffer E2 and the following program: 1200 V, 20 ms, 3 pulses. Cells were recovered in 4.3 mL of pre-warmed media (RPMI+15% FBS, no antibiotics). GFP controls were used to monitor transfection efficiency (~15%). 24 hours post-transfection, cells were washed in PBS, plated at 500,000 cells per well, and lysed in an optical plate. Luminescence measurements were performed as above. The activity of each construct was quantified relative to the empty vector, and differential activity was relative to the human ancestral sequence. Significant differences were determined by ANOVA, using batch as a covariate, followed by a Tukey post-hoc test.

DiMeLo-seq analyses

DiMeLo data for H3K4me3 and H3K27ac in GM12878 were obtained from GSE208125 [42] as BAM files aligned to T2T CHM13 v2.0. Base modifications (5mC and 6mA) were thresholded at 155 (corresponding to probability of ~60%) and the frequency of base modification per site was calculated 0using `modkit pileup` (Nanoporetech). For comparison with short-read ChIP data and MPRA tiles, coordinates were mapped to T2T CHM13 v2.0 using liftOver (UCSC Genome Browser Utilities). Then, the mean fraction of modified bases was calculated for each interval in all peak sets of interest using `bedtools map`. The summed ABC score was calculated per HSD gene as above, using only H3K27ac signal. Finally, differences in signal (mean fraction modified bases) for all assays (5mC, H3K27ac 6mA, and H3K4me3 6mA) were compared between chimpanzee-syntenic and non-syntenic regions, also converted to CHM13 v2.0 coordinates.

DATA AVAILABILITY

Raw sequence data generated from this study will be made publicly available through National Center for Biotechnology Information (NCBI) and European Nucleotide Archive (ENA) Accession PRJEB98375 at

the time of publication. Code used for this project are available is archived via Zenodo with the following DOI: 10.5281/zenodo.17240139

ACKNOWLEDGMENTS

We thank Drs. Fumitaka Inoue and Tal Ashuach for technical advice related to MPRA library construction and data processing, as well as Drs. Sierra Nishizaki and Daniela C. Soto for valuable discussion of the bioinformatic analyses performed in this study. We also acknowledge Dr. Anthony Antonellis for sharing the pE1B enhancer reporter Gateway plasmid, as well as Drs. Gerald Quon and Siobhan Brady for constructive feedback on the manuscript. Thank you to Dr. Nicolas Altemose for suggestions and code for analyzing the DiMeLo-seq data. Finally, we acknowledge the ENCODE Consortium, specifically the laboratories of Bradley Bernstein, J. Michael Cherry, Thomas Gingeras, Brenton Graveley, and John Stamatoyannopoulos for data used in this study. This work was supported by the UC Davis COVID-Impacted Research Funding Program, National Science Foundation (CAREER 2145885 to M.Y.D.), and National Human Genome Research Institute (F31HG011205 to C.S.) at the National Institutes of Health.

REFERENCES

1. Ohno S. Evolution by Gene Duplication. Springer Science & Business Media; 2013.

2. Lynch M, Conery JS. The evolutionary fate and consequences of duplicate genes. Science. 2000;290: 1151–1155.

3. Kondrashov FA, Rogozin IB, Wolf YI, Koonin EV. Selection in the evolution of gene duplications. Genome Biol. 2002;3: 1–9.

4. Evolution by gene duplication: an update. Trends Ecol Evol. 2003;18: 292–298.

5. Force A, Lynch M, Pickett FB, Amores A, Yan YL, Postlethwait J. Preservation of duplicate genes by complementary, degenerative mutations. Genetics. 1999;151: 1531–1545.

6. Rodin SN, Riggs AD. Epigenetic silencing may aid evolution by gene duplication. J Mol Evol. 2003;56: 718–729.

7. Prud'homme B, Gompel N, Carroll SB. Emerging principles of regulatory evolution. Proc Natl Acad Sci U S A. 2007;104 Suppl 1: 8605–8612.

8. Bailey JA, Yavor AM, Massa HF, Trask BJ, Eichler EE. Segmental duplications: organization and impact within the current human genome project assembly. Genome Res. 2001;11: 1005–1017.

9. Marques-Bonet T, Girirajan S, Eichler EE. The origins and impact of primate segmental duplications. Trends Genet. 2009;25: 443–454.

10. Lan X, Pritchard JK. Coregulation of tandem duplicate genes slows evolution of subfunctionalization in mammals. Science. 2016;352: 1009–1013.

11. Dennis MY, Harshman L, Nelson BJ, Penn O, Cantsilieris S, Huddleston J, et al. The evolution and population diversity of human-specific segmental duplications. Nat Ecol Evol. 2017;1: 69.

12. Florio M, Heide M, Pinson A, Brandl H, Albert M, Winkler S, et al. Evolution and cell-type specificity of human-specific genes preferentially expressed in progenitors of fetal neocortex. Elife. 2018;7. doi:10.7554/eLife.32332

13. Florio M, Albert M, Taverna E, Namba T, Brandl H, Lewitus E, et al. Human-specific gene ARHGAP11B promotes basal progenitor amplification and neocortex expansion. Science. 2015;347: 1465–1470.

14. Kalebic N, Gilardi C, Albert M, Namba T, Long KR, Kostic M, et al. Human-specific ARHGAP11B induces hallmarks of neocortical expansion in developing ferret neocortex. Elife. 2018;7. doi:10.7554/eLife.41241

15. Heide M, Haffner C, Murayama A, Kurotaki Y, Shinohara H, Okano H, et al. Human-specific ARHGAP11B increases size and folding of primate neocortex in the fetal marmoset. Science. 2020;369: 546–550.

16. Namba T, Dóczi J, Pinson A, Xing L, Kalebic N, Wilsch-Bräuninger M, et al. Human-Specific ARHGAP11B Acts in Mitochondria to Expand Neocortical Progenitors by Glutaminolysis. Neuron. 2020;105: 867–881.e9.

17. Dougherty ML, Underwood JG, Nelson BJ, Tseng E, Munson KM, Penn O, et al. Transcriptional fates of human-specific segmental duplications in brain. Genome Res. 2018;28: 1566–1576.

18. Shew CJ, Carmona-Mora P, Soto DC, Mastoras M, Roberts E, Rosas J, et al. Diverse Molecular Mechanisms Contribute to Differential Expression of Human Duplicated Genes. Mol Biol Evol. 2021;38: 3060–3077.

19. Gasperini M, Tome JM, Shendure J. Towards a comprehensive catalogue of validated and target-linked human enhancers. Nat Rev Genet. 2020;21: 292–310.

20. Soto DC, Uribe-Salazar JM, Kaya G, Valdarrago R, Sekar A, Haghani NK, et al. Human-specific gene expansions contribute to brain evolution. Cell. 2025. doi:10.1016/j.cell.2025.06.037

21. Pezzini F, Bettinetti L, Di Leva F, Bianchi M, Zoratti E, Carrozzo R, et al. Transcriptomic Profiling Discloses Molecular and Cellular Events Related to Neuronal Differentiation in SH-SY5Y Neuroblastoma Cells. Cell Mol Neurobiol. 2017;37: 665–682.

22. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012;489: 57–74.

23. Luo Y, Hitz BC, Gabdank I, Hilton JA, Kagda MS, Lam B, et al. New developments on the Encyclopedia of DNA Elements (ENCODE) data portal. Nucleic Acids Res. 2020;48: D882–D889.

24. Hitz BC, Lee J-W, Jolanki O, Kagda MS, Graham K, Sud P, et al. The ENCODE Uniform Analysis Pipelines. Res Sq. 2023. doi:10.21203/rs.3.rs-3111932/v1

25. Nurk S, Koren S, Rhie A, Rautiainen M, Bzikadze AV, Mikheenko A, et al. The complete sequence of a human genome. Science. 2022;376: 44–53.

26. Liao W-W, Asri M, Ebler J, Doerr D, Haukness M, Hickey G, et al. A draft human pangenome reference. Nature. 2023;617: 312–324.

27. McVicker G, van de Geijn B, Degner JF, Cain CE, Banovich NE, Raj A, et al. Identification of

genetic variants that affect histone modifications in human cells. Science. 2013;342: 747–749.

28. Degner JF, Pai AA, Pique-Regi R, Veyrieras J-B, Gaffney DJ, Pickrell JK, et al. DNase I sensitivity QTLs are a major determinant of human expression variation. Nature. 2012;482: 390–394.

29. Bryois J, Garrett ME, Song L, Safi A, Giusti-Rodriguez P, Johnson GD, et al. Evaluation of chromatin accessibility in prefrontal cortex of individuals with schizophrenia. Nat Commun. 2018;9: 3121.

30. Vermunt MW, Tan SC, Castelijns B, Geeven G, Reinink P, de Bruijn E, et al. Epigenomic annotation of gene regulatory alterations during evolution of the primate brain. Nat Neurosci. 2016;19: 494–503.

31. Reilly SK, Yin J, Ayoub AE, Emera D, Leng J, Cotney J, et al. Evolutionary genomics. Evolutionary changes in promoter and enhancer activity during human corticogenesis. Science. 2015;347: 1155–1159.

32. de la Torre-Ubieta L, Stein JL, Won H, Opland CK, Liang D, Lu D, et al. The Dynamic Landscape of Open Chromatin during Human Cortical Neurogenesis. Cell. 2018;172: 289–304.e18.

33. Gordon MG, Inoue F, Martin B, Schubach M, Agarwal V, Whalen S, et al. lentiMPRA and MPRAflow for high-throughput functional characterization of gene regulatory elements. Nat Protoc. 2020;15: 2387–2412.

34. Ashuach T, Fischer DS, Kreimer A, Ahituv N, Theis FJ, Yosef N. MPRAnalyze: statistical framework for massively parallel reporter assays. Genome Biol. 2019;20: 183.

35. Fulco CP, Nasser J, Jones TR, Munson G, Bergman DT, Subramanian V, et al. Activity-by-contact model of enhancer-promoter regulation from thousands of CRISPR perturbations. Nat Genet. 2019;51: 1664–1669.

36. Zhang J, Lee D, Dhiman V, Jiang P, Xu J, McGillivray P, et al. An integrative ENCODE resource for cancer genomics. Nature Communications. 2020;11: 1–11.

37. Gartlgruber M, Sharma AK, Quintero A, Dreidax D, Jansky S, Park Y-G, et al. Super enhancers define regulatory subtypes and cell identity in neuroblastoma. Nat Cancer. 2021;2: 114–128.

38. Zimmerman MW, Liu Y, He S, Durbin AD, Abraham BJ, Easton J, et al. Drives a Subset of High-Risk Pediatric Neuroblastomas and Is Activated through Mechanisms Including Enhancer Hijacking and Focal Enhancer Amplification. Cancer Discov. 2018;8: 320–335.

39. Chung D, Zhang Q, Keleş S. MOSAiCS-HMM: A Model-Based Approach for Detecting Regions of Histone Modifications from ChIP-Seq Data. Statistical Analysis of Next Generation Sequencing Data. 2014; 277–295.

40. Gershman A, Sauria MEG, Guitart X, Vollger MR, Hook PW, Hoyt SJ, et al. Epigenetic patterns in a complete human genome. Science. 2022;376: eabj5089.

41. Altemose N, Maslan A, Smith OK, Sundararajan K, Brown RR, Mishra R, et al. DiMeLo-seq: a long-read, single-molecule method for mapping protein-DNA interactions genome wide. Nat Methods. 2022;19: 711–723.

42. Maslan A, Altemose N, Marcus J, Mishra R, Brennan LD, Sundararajan K, et al. Mapping protein–

DNA interactions with DiMeLo-seq. Nature Protocols. 2024;19: 3697–3720.

43. Vollger MR, Guitart X, Dishuck PC, Mercuri L, Harvey WT, Gershman A, et al. Segmental duplications and their variation in a complete human genome. Science. 2022;376: eabj6965.

44. Uebbing S, Gockley J, Reilly SK, Kocher AA, Geller E, Gandotra N, et al. Massively parallel discovery of human-specific substitutions that alter enhancer activity. Proc Natl Acad Sci U S A. 2021;118. doi:10.1073/pnas.2007049118

45. Direct Identification of Hundreds of Expression-Modulating Variants using a Multiplexed Reporter Assay. Cell. 2016;165: 1519–1529.

46. Watson CT, Marques-Bonet T, Sharp AJ, Mefford HC. The genetics of microdeletion and microduplication syndromes: an update. Annu Rev Genomics Hum Genet. 2014;15: 215–244.

47. Gillentine MA, Berry LN, Goin-Kochel RP, Ali MA, Ge J, Guffey D, et al. The Cognitive and Behavioral Phenotypes of Individuals with CHRNA7 Duplications. J Autism Dev Disord. 2017;47: 549–562.

48. Zheng Y, Ay F, Keles S. Generative modeling of multi-mapping reads with mHi-C advances analysis of Hi-C studies. Elife. 2019;8. doi:10.7554/eLife.38070

49. Deshpande AS, Ulahannan N, Pendleton M, Dai X, Ly L, Behr JM, et al. Identifying synergistic high-order 3D chromatin conformations from genome-scale nanopore concatemer sequencing. Nat Biotechnol. 2022;40: 1488–1499.

50. McGinty SP, Kaya G, Sim SB, Corpuz RL, Quail MA, Lawniczak MKN, et al. CiFi: Accurate long-read chromatin conformation capture with low-input requirements. bioRxiv. 2025. doi:10.1101/2025.01.31.635566

51. Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. Salmon provides fast and bias-aware quantification of transcript expression. Nat Methods. 2017;14: 417–419.

52. Soneson C, Love MI, Robinson MD. Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. F1000Res. 2015;4: 1521.

53. Iso-Seq Home. In: Iso-Seq Docs [Internet]. [cited 24 Sep 2025]. Available: https://isoseq.how/

54. Skera Home. In: Skera Docs [Internet]. [cited 24 Sep 2025]. Available: https://skera.how/

55. Zare Jousheghani Z, Singh NP, Patro R. Oarfish: enhanced probabilistic modeling leads to improved accuracy in long read transcriptome quantification. Bioinformatics. 2025;41: i304–i313.

56. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30: 2114–2120.

57. Applied Research Press. Ultrafast and Memory-Efficient Alignment of Short DNA Sequences to the Human Genome. Createspace Independent Publishing Platform; 2015.

58. Liu T. Use model-based Analysis of ChIP-Seq (MACS) to analyze short reads generated by sequencing protein-DNA interactions in embryonic stem cells. Methods Mol Biol. 2014;1150: 81–95.

59. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26: 841–842.

60. Kent WJ. BLAT--the BLAST-like alignment tool. Genome Res. 2002;12: 656–664.