Diverse molecular mechanisms contribute to differential expression of human duplicated genes

Colin J. Shew[1,2], Paulina Carmona-Mora[1,3,4], Daniela C. Soto[1,2], Mira Mastoras[1], Elizabeth Roberts[1], Joseph Rosas[1,5], Dhriti Jagannathan[1], Gulhan Kaya[1], Henriette O'Geen[1], Megan Y. Dennis[1-6,†]

[1]Genome Center, [2]Integrative Genetics and Genomics Graduate Group, [3]MIND Institute, [4]Autism Research Training Program, [5]Postbaccalaureate Research Education Program, [6]Department of Biochemistry & Molecular Medicine, University of California, Davis, CA 95616

[†]Corresponding author:
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

ABSTRACT

Emerging evidence links genes within human-specific segmental duplications (HSDs) to traits and diseases unique to our species. Strikingly, despite being nearly identical by sequence (>98.5%), paralogous HSD genes are differentially expressed across human cell and tissue types, though the underlying mechanisms have not been examined. Comparing cross-tissue mRNA levels between humans and chimpanzees, we determined that broadly, HSD paralogs (75 genes from 30 families) display expression patterns consistent with pseudo- or neofunctionalization. In general, the ancestral paralog exhibited greatest expression conservation with the chimpanzee ortholog, though exceptions suggest duplicate paralogs that may retain or supplant ancestral functions. To understand mechanisms underlying this observed regulatory divergence, we reanalyzed data from human lymphoblastoid cell lines (LCLs) (n=445), showing that ~75% of derived HSD paralogs exhibit significant differential expression and a greater than two-fold difference from their ancestral counterpart. To identify active *cis*-regulatory elements (CREs) in HSDs, we reanalyzed ENCODE data to recover hundreds of candidate CREs in these regions. Further, we generated ChIP-seq data for active chromatin features in an LCL using longer Illumina reads to better distinguish peaks in paralogous regions. Some of these duplicated CREs are sufficient to drive differential reporter activity, suggesting they may contribute to divergent *cis*-regulation of paralogs. This work provides evidence that *cis*-regulatory divergence contributes to novel expression patterns of recent gene duplicates in humans.

INTRODUCTION

Gene duplication occurs universally and is considered a major source of evolutionary novelty. For example, from a survey of representative eukaryotic genomes, the origin of over 30% of genes was attributed to duplication(Zhang 2003). Although most duplicated genes are fated to be rapidly pseudogenized, some may retain important functions and experience strong purifying selection(Lynch 2000). These genes are thought to become variously subfunctionalized (sharing the ancestral function) or neofunctionalized (gaining a novel function). Expression divergence is likely integral to this process, as spatiotemporal partitioning of function places both daughter paralogs under negative selection, helping them escape pseudogenization(Rodin and Riggs 2003; Rodin et al. 2005). It has been suggested that this may be the primary driver behind duplicate gene retention, as gene regulation can be altered relatively easily, without the need to acquire rare beneficial coding mutations(Ohno 1970). For example, mouse *Hoxa1* and *Hoxb1* genes are functionally redundant but partitioned by expression, with normal development possible from a single gene under control of regulatory elements from both paralogs(Tvrdik and Capecchi 2006). On a genome-wide scale, substantial expression divergence is observed in vertebrates following whole-genome duplications specific to teleost and salmonid fishes(Kassahn et al. 2009; Braasch et al. 2016; Lien et al. 2016; Varadharajan et al. 2018). Meta-analysis suggests that across all these species, selection on expression levels appears relaxed in one of the paralogs(Sandve et al. 2018). However, duplication of individual loci rather than whole genome events is substantially more common in vertebrates, with segmental duplications (SDs, regions defined as having >90% sequence similarity and being at least 1 kb in size(Bailey 2002)) having the potential to structurally rearrange the genome, in turn facilitating regulatory divergence and duplicate retention(Rodin et al. 2005). Although studies characterizing expression divergence of duplicated

genes in humans, mice, and yeast have identified broad patterns of dosage sharing among daughter paralogs(Qian et al. 2010; Lan and Pritchard 2016), more recent duplications have not been analyzed. In addition, no molecular explanations have been provided for the observed expression changes.

Great apes have experienced a surge of SDs in the last ~10 million years, primarily interspersed throughout the genome and potentially contributing to phenotypic differences observed between these closely related species(Prado-Martinez et al. 2013). Human-specific SDs (HSDs), which arose in the last ~6 million years following divergence from a common ancestor with chimpanzee, contain genes unique to the human lineage that have compelling associations with neurodevelopmental features(Charrier et al. 2012; Dennis et al. 2012; Florio et al. 2015; Fiddes et al. 2018; Suzuki et al. 2018; Heide et al. 2020) and disorders(Dennis and Eichler 2016; Dennis et al. 2017; Ishiura et al. 2019). Historically such young duplications have been poorly resolved in genome assemblies due to their high sequence similarity. Recent sequencing efforts targeted to HSDs have generated high-quality references for many of these loci(Steinberg et al. 2012; Antonacci et al. 2014; O'Bleness et al. 2014; Dennis et al. 2017) resulting in the discovery of at least 30 duplicated gene families accounting for >80 gene paralogs. Most derived HSD paralogs encode putatively functional proteins and exhibit divergent expression patterns relative to ancestral paralogs across numerous primary tissues, despite being nearly identical (>98%) by sequence(Dennis et al. 2017). HSD genes also utilize novel promoters and exons exapted from the site of insertion(Dougherty et al. 2017). Differential regulation may also be explained by associations of species-specific active chromatin modifications at SD loci(Giannuzzi et al. 2014), though due to historical reference errors and computational challenges in short-read mapping to

highly-similar sequences(Chung et al. 2011; Ebbert et al. 2019), regulatory features of duplicated loci remain poorly characterized.

In this study, we characterized patterns of regulatory divergence observed for HSD genes between human and chimpanzee cell lines and primary tissues by quantifying the cross-tissue conservation of homologs, offering potential insight into the speed at which young duplicate genes diverge in expression and allowing for inference of likely functional HSD paralogs. We leveraged genomic and epigenomic data from hundreds of human lymphoblastoid cell lines (LCLs) to identify differentially expressed (DE) ancestral-derived gene pairs and tested potential molecular contributors to paralog-specific divergence in expression, including copy number (CN) variation, post-transcriptional regulation, and *cis*-regulatory changes. Finally, we surveyed the active chromatin "landscape" for HSDs by reanalyzing ENCODE chromatin immunoprecipitation sequence (ChIP-seq) data, producing novel data with longer reads to ameliorate data loss due to poor mapping in SDs, and functionally validating *cis*-regulatory elements (CREs) via a reporter assay. Overall, our work demonstrates that *cis*-regulatory divergence, among other mechanisms, drives differential expression following gene duplication and that useful regulatory information can be rescued from existing datasets for duplicated loci.

RESULTS

Conservation of HSD gene expression following duplication

To assess the evolutionary trajectory of recent human duplicated genes, we quantified (Soneson et al. 2015; Patro et al. 2017) expression of 75 HSD genes from 30 gene families for which high-

confidence sequences were available (Dennis et al. 2017) (Table S1). Each HSD gene family corresponds to a single-copy chimpanzee ortholog; the human paralog syntentic with the chimpanzee gene, if known, is considered to be the ancestral copy, while the human-specific duplicates are derived (Figure 1A) We compared expression HSD paralogs and chimpanzee orthologs using mRNA-sequencing (RNA-seq) data from three cell lines and four primary tissue types (Khan et al. 2013; Pavlovic et al. 2018; Marchetto et al. 2019; Blake et al. 2020). We generated a custom transcriptome with improved annotations of paralogous genes and quantified expression with a lightweight mapping approach that shows greater accuracy than alignment-based methods for paralogous genes (Soneson et al. 2015; Patro et al. 2017). To interpret the evolutionary fate of these paralogous genes, we compared the individual or summed expression of human paralogs within each gene family to the single chimpanzee ortholog. We found that derived HSD paralogs tended to show lower expression than the chimpanzee ortholog, summed family expression was mostly higher, and ancestral paralogs were less likely to be DE (9/21 expressed ancestral genes showed no differential expression across all cell/tissue types versus 6/37 of expressed derived genes) (Figure S1A, Table S2). This suggests that ancestral genes are most likely to retain their functions as duplicate paralogs diverge or are pseudogenized. We next considered expression correlation across the four tissue types and three cell lines as an indicator of expression conservation between human paralogs and the chimpanzee ortholog. Our expectation was that in the case of pseudogenization or neofunctionalization, a single paralog would correlate best with chimpanzee ortholog, while subfunctionalized genes would show a higher correlation of the total family expression with the chimpanzee ortholog. We identified the most conserved paralog in each duplicate gene family (versus chimpanzee) and found that these paralogs showed similar correlation to the overall sum expression of the gene family, while the

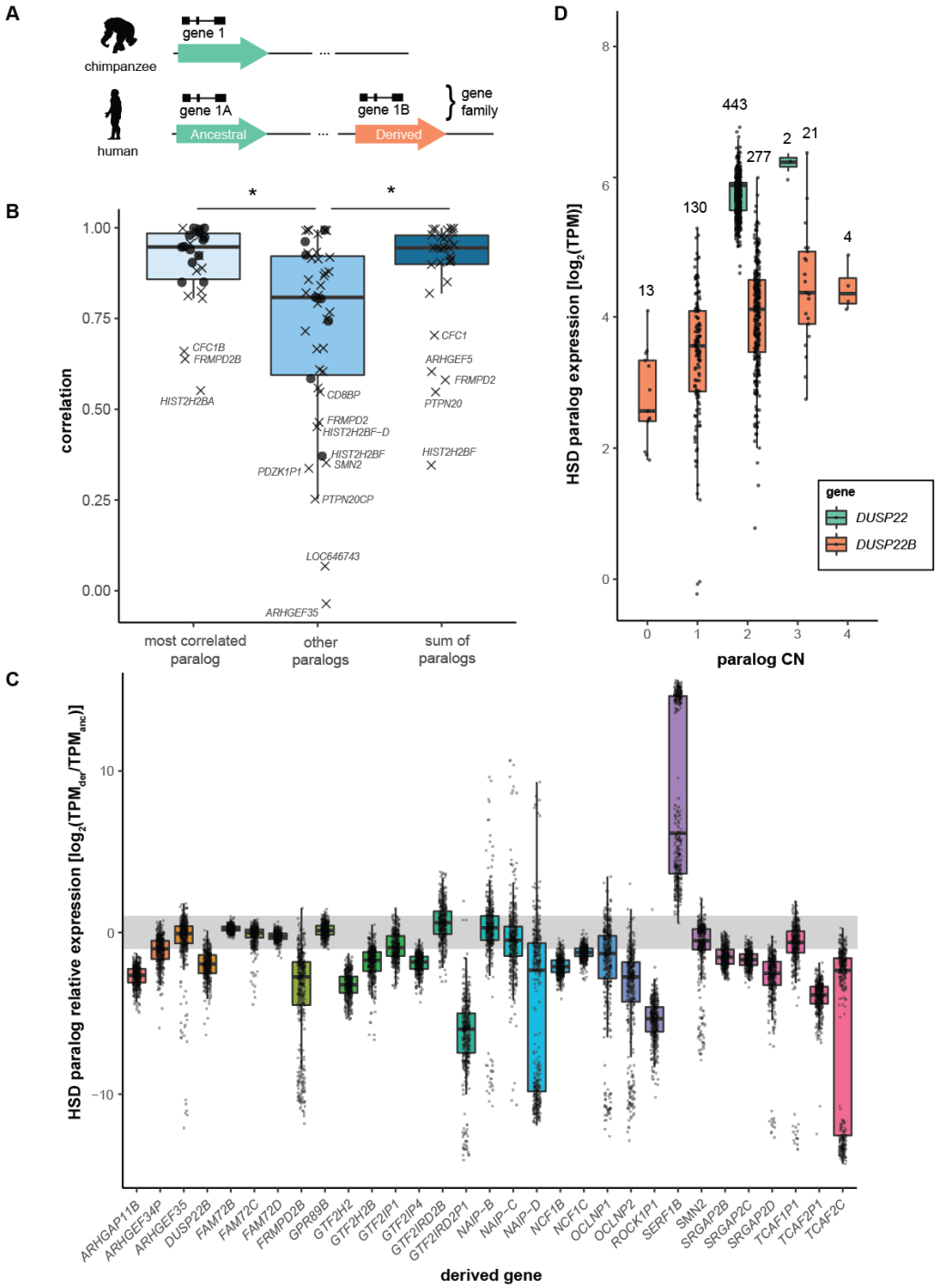

Figure 1. Expression patterns of HSD genes. (A) Illustration of genes residing within HSDs; the ancestral paralog (green) corresponds to the chimpanzee ortholog, while derived paralogs (orange) are human-specific. The ancestral and derived genes comprise a gene family. **(B)** Cross-tissue correlations of HSD genes or gene family summed expression versus expression of chimpanzee orthologs are shown in boxplots. For each gene family, the most conserved paralog ("highest"), all other paralogs ("lower"), and the sum of all paralogs ("sum") are plotted separately. Ancestral genes, if evolutionary status is known, are denoted as a filled circle. Asterisks denote significant differences (p<0.01, Tukey post-hoc test of ANOVA) **(C)** Expression divergence of derived genes from families with at least one LCL-expressed paralog is plotted as the $\log_2$ ratio of median derived and ancestral TPM expression. Each point represents a different LCL from the Geuvadis consortium (total N=445). The gray bar indicates a two-fold expression difference. **(D)** Expression values of ancestral *DUSP22* (green) and derived *DUSP22B* (orange), stratified by CN. Points represent individual LCLs, and the number of individuals represented in each CN category is indicated over each boxplot.

remaining paralogs had significantly lower correlation with the sum ($p<0.01$, Tukey post-hoc test of ANOVA) (Figure 1B). This pattern is consistent with maintenance of function of one paralog and neo- or pseudofunctionalization of the others. Notably, ancestral genes were overrepresented in the "highest" correlation category (14/22 of known status, $p<0.001$, hypergeometric test). Nevertheless, the remaining eight paralogs showing strongest conservation of expression with chimpanzee orthologs were derived, including *FRMPD2B*, *HIST2H2BA*, *OR2A42*, and *SERF1B*. These genes represent interesting candidates of a derived paralog usurping the function of the ancestral gene. Taken together, these findings suggest that HSD genes are variously pseudogenized or neofunctionalized, with the ancestral paralog often, but not always, maintaining its expression pattern across tissues.

Expression of human paralogs in lymphoblastoid cell lines (LCLs)

We next focused on LCLs to gain a more detailed understanding of HSD expression patterns across hundreds of individuals with matched genomic data. We estimated transcript abundance using RNA-seq data from 462 human LCLs (Lappalainen et al. 2013) (Table S3) and determined that over half (43/75) of HSD paralogs are expressed (at least one transcript per million (TPM)), with the most highly expressed genes including *ARHGAP11A*, *ROCK1*, the adjacent *GTF2I* and *NCF1* families, and the *DUSP22* family, whose human-specific paralog *DUSP22B* is missing from the reference (GRCh38) (Dennis et al. 2017). Comparing expression profiles within gene families, we found derived and ancestral paralogs globally showed divergent expression levels; in families with at least one expressed gene, all 31 derived genes showed significant differences from their ancestral counterpart. Twenty of these showed a difference in median TPM greater than two-fold, and in most cases (25/31) the derived gene had lower expression (Figures 1C, S2,

and Table S4). We noted that some paralogs exhibited clusters of outlier values for derived/ancestral expression ratios, which could not be reconciled as copy number (CN) or population of origin differences (Figure S3). Altogether, these results indicate that HSD genes show divergent expression patterns on LCLs.

Copy-number variation and HSD expression

While all the genes in this study were chosen for being fixed or nearly fixed in all modern human populations (Dennis et al. 2017), SD loci are known to be subject to recurrent rearrangement and consequently exhibit varying degrees of CN polymorphism. Anticipating that CN variation is likely to alter gene expression levels, we sought to characterize the impact of CN on differential expression of HSD genes. After performing paralog-specific CN genotyping (Shen and Kidd 2020) of a subset of individuals for which 1000 Genomes Illumina sequences were available (N=445), we found gene expression was positively associated with CN in about half (28/55) of genes in expressed families (Table S5), indicating that higher CN often but not always translates to increased expression. Notably, derived genes tended to have higher CN (1.3-fold on average, across all assayed genes and individuals) but lower expression overall. To account for the effect of CN on differential expression between paralogs, we also regressed CN out of comparisons of expressed ancestral-derived gene pair and found 23/25 derived paralogs were still DE with respect to the ancestral (six were not tested due to paralog-specific effects of CN; Table S4). For example, while expression of *DUSP22B* was significantly associated with CN, these effects were insufficient to explain DE relative to *DUSP22* (Figure 1D). Thus, while CN differences alter the mRNA levels of HSD paralogs, they do not provide an explanation for overall differential expression of these genes.

Post-transcriptional regulation of HSD genes

In order to determine if paralogous expression differences are driven by post-transcriptional regulation, we next considered whether HSD transcripts were being processed as nonfunctional pseudogenes. In this scenario, paralogs might be equally transcribed but then subjected to degradation via nonsense-mediated decay (NMD). To test this, we compared gene expression using available RNA-seq data from human NMD-deficient LCLs (N=4) against controls (N=2) (Nguyen et al. 2012) and found no HSD genes with significant differential expression. We also used these RNA-seq data to compare the ratio of derived to ancestral expression between NMD-deficient LCLs and controls and found no significant differences, though sample sizes were likely limiting (Figure S4A). This result is largely recapitulated by paralog-specific RT-qPCR for three DE HSD genes families (*ARHGAP11, DUSP22,* and *ROCK1*) in four LCLs treated with NMD-inhibiting drug emetine. Ratios of *ROCK1P1/ROCK1* and *DUSP22B/DUSP22* expression were unaltered by emetine treatment, while *ARHGAP11B*/*ARHGAP11A* expression ratio increased closer to one, consistent with NMD affecting *ARHGAP11B*, though not completely 'rescuing' derived expression levels to equal that of the ancestral (Figure S4B). *ARHGAP11B* is a 3′ truncation of *ARHGAP11A*, potentially explaining differences in transcript stability. Altogether, these results suggest that while NMD may alter steady-state expression levels of some HSD genes, it is not a primary driver of their differential expression.

We also examined HSD 3′ untranslated regions (UTRs) for recognition sites of miRNAs expressed in LCLs (Lappalainen et al. 2013) (N=13 3′ UTRs of expressed gene families; mean 94 binding sites per UTR) using TargetScan (Agarwal et al. 2015). miRNA binding sites were

nearly identical between paralogs. The small numbers of miRNA-binding differences between paralogs correlated with expression divergence (linear regression p=0.01; adjusted R^2=0.38; Figure S5A), but in the opposite direction as we might predict (i.e., more differences leading to more similar expression patterns between paralogs). When we compared paralogs with identical and non-identical 3′ UTRs, we observed significantly lower derived expression ratios from those that differ (Wilcoxon rank sum test p<0.01, Figure S5B). These data point to a potential role for miRNAs in HSD transcriptional regulation, but this mechanism is unlikely to cause differential expression within highly expressed gene families, such as the *DUSP22* and *NCF1*.

Role of *cis*-regulation in HSD differential expression

We next aimed to determine if *cis*-regulatory changes contribute to expression divergence of HSDs. Because SDs often generate gene truncations and fusions with adjacent transcribed sequences (Dougherty et al. 2017), we reasoned that gains or losses of promoters or UTRs would likely cause large changes in gene expression. We compared relative expression by truncation status (5′-, 3′-, or non-truncated) of all derived genes in expressed families to their ancestral paralogs. Ancestral and derived genes had more similar expression levels in non-truncating duplications, while truncated genes tended to be less expressed than their ancestral paralogs, particularly 5’ truncations compared to all other HSD genes (p=0.06, t-test; Figure S6A), in concordance with previous findings(Dougherty et al. 2018). While we may have limited power to detect differences with a small number of genes, these results hint that promoter activity may contribute to differential expression patterns. Considering sequence-level changes more broadly, however, we observed no relationship between expression divergence and pairwise nucleotide divergence across entire duplicons or within promoters (Figure S6B–C).

Given that the vast majority of paralog-specific variants (PSVs) are unlikely to be functionally important, we used publicly available chromatin immunoprecipitation sequencing (ChIP-seq) datasets from the Encyclopedia of DNA Elements (ENCODE) project (Consortium and The ENCODE Project Consortium 2012; Davis et al. 2018) to identify likely CREs (H3K4me3, H3K4me1, H3K27ac, and RNA PolII) in a single LCL (GM12878). In each data set, we observed a lower density of bases covered by peaks in SDs (>90% similarity) and HSDs (>98% similarity) compared to the rest of the genome (empirical p=0.001, N=1000 replicates; Figure 2, in yellow). We posit, as others have previously (Chung et al. 2011; McVicker et al. 2013; Giannuzzi et al. 2014), that this discrepancy is an artifact of the high sequence similarity of SDs, as reads originating from these regions map to multiple locations of the genome and are accordingly discarded at the alignment stage or by mapping-quality filtering.

To recover this missing information, we implemented a pipeline that allowed reads to align to multiple locations in the genome then, using CSEM (Chung et al. 2011), iteratively assigned posterior probabilities to each read alignment based on the unique mapping rate of nearby reads. Selecting the most likely alignment to allocate a read (i.e., mapping position with the highest posterior probability), we improved peak discovery in SDs and HSDs for the aforementioned chromatin features, erasing the depletion for all but H3K27ac, which was still substantially improved (Figure 2, in blue). The peaks we discovered largely overlapped with the ENCODE peaks, though RNA PolII had a large proportion of peaks unique to this analysis (Figure S7).

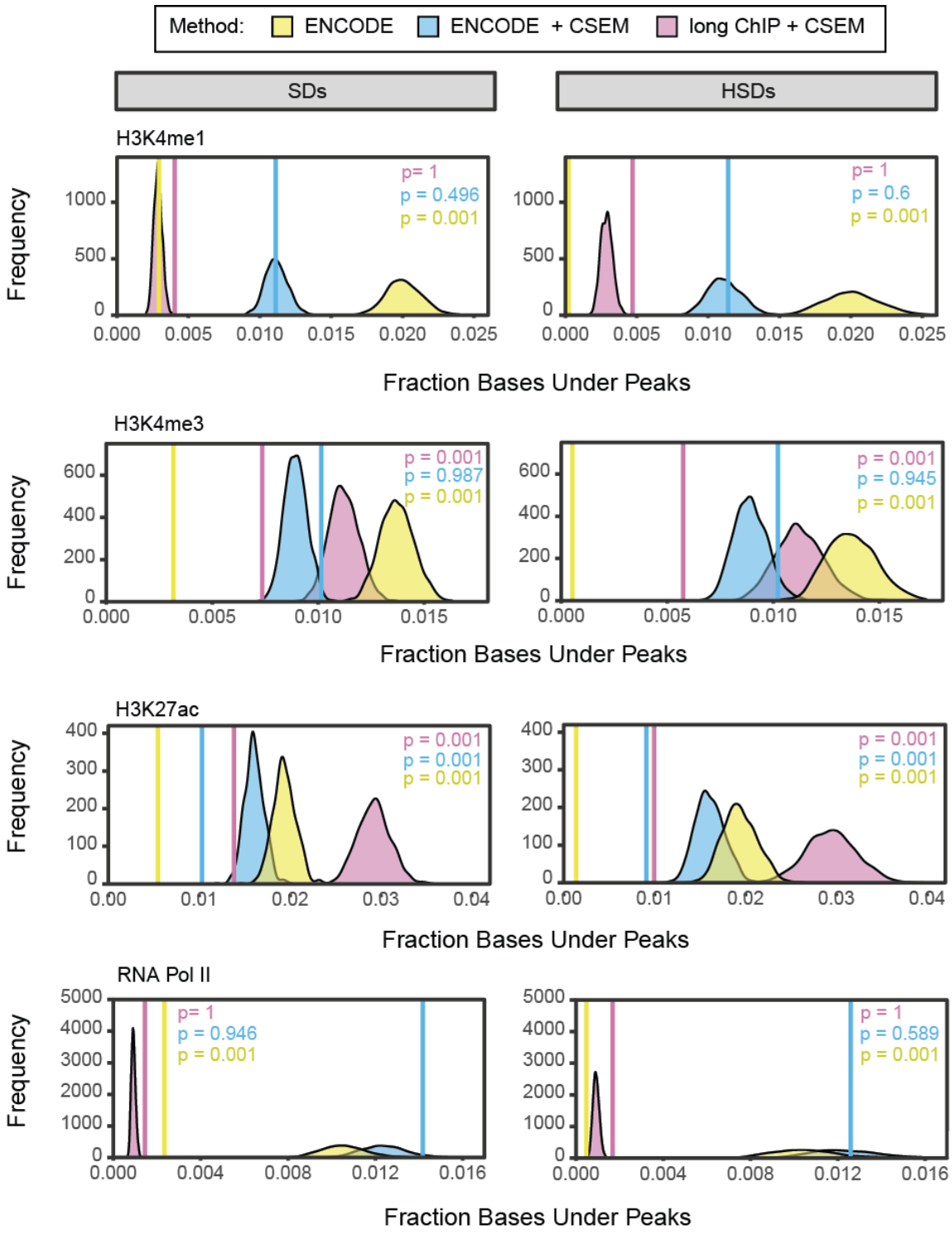

Figure 2. Depletion and recovery of ChIP peaks in SDs. Results are summarized for three ChIP-seq peak discovery approaches: publicly-available ENCODE peaks (yellow), peaks from multimapping and CSEM allocation of ENCODE raw data (blue), and peaks from multimapping and CSEM allocation of long ChIP data. For SDs (left) and SDs with greater than 98% sequence identity (or HSDs, right), the fraction of bases under peaks was calculated (solid vertical lines). The SD coordinates were permuted 1000 times within the human reference (GRCh38), and an expected distribution of the fraction of bases covered was generated. Empirical one-sided p-values for depletion are indicated above each plot.

229

Improved peak discovery using longer-read ChIP-seq

To improve our ability to align reads accurately to specific paralogs, we generated longer-read (~500 bp insert size, 2x250 bp PE Illumina) ChIP-seq ("long ChIP") libraries (H3K4me3, H3K27ac, H3K4me1, and RNA PolII) from the LCL GM12878. Longer reads mapped to SDs with greater accuracy (Figure S8A), allowing for higher-confidence discovery of novel peaks in duplicated regions using standard single-site mapping approaches (BWA mem, Figure S8B). However, all marks except H3K4me1 were still depleted for peaks in SDs relative to the rest of the genome. Subsequently, we analyzed the long ChIP data allowing for multiple alignments and probabilistically assigning reads to one position (Bowtie and CSEM, Figure S8B). Long ChIP showed increased posterior assignment probabilities with respect to the short-read ENCODE data, and the depletion of peaks in SDs was erased for H3K4me3, H3K4me1, and PolII (Figure 2, in pink). Notably, for most libraries fewer overall peaks were identified with long ChIP versus ENCODE data, though the peaks that do exist were largely replicated (on average, 73% of long ChIP peaks were corresponded to an ENCODE peak (Chikina and Troyanskaya 2012); Table S6; Figure S9). Long ChIP peaks tended to be larger, with 2.4–3.7 times as many bases per peak, excepting H3K4me1, which had slightly smaller peaks. We verified greater enrichment of H3K27ac at the ancestral *DUSP22* (1.1–2.9-fold difference) at three PSVs, also noting a correlation of expression divergence and differential enrichment (Figure S10). These findings suggest that reanalysis of ChIP-seq data can accurately identify enriched regions at HSD loci, uncovering potentially divergent regulatory environments.

To identify putatively functional *cis*-regulatory regions within HSDs, we integrated our ChIP-seq peaks in an 8-state chromHMM model (Ernst and Kellis 2012), providing a novel set of

candidate activating CREs, as virtually no information is available in the current ENCODE release for SD loci. Because derived gene expression is generally lower than ancestral, we compared putative CREs in 100-kb windows covering HSD loci (defined in Dennis, 2017). We observed no significant differences in the fraction of bases covered by putative CREs (Wilcoxon rank-sum test; Figure S11). We also observed no differences between ancestral and derived regions in the individual ChIP-seq datasets (active histone marks or RNA PolII), considering both the fraction of bases covered enrichment strength of peaks. We also failed to find a difference in the fraction of bases covered by repressed H3K27me3 domains identified from our multimapping analysis. Thus, explanations beyond the overall abundance of chromatin features are needed, as important functional changes in CRE activity may not be reflected in global differences. We did visually identify differences in presence or absence of putative *cis*-acting elements suggesting a more nuanced approach is necessary in pinpointing mechanisms contributing to paralog expression differences (Figure 3).

Impact of altered chromatin context due to structural variation on gene expression

HSDs are often transposed many thousands of kilobases from their ancestral paralogs, and in some cases on entirely different chromosomes, resulting in new adjacent genomic regions that could impact expression of genes contained within; as such, we sought to understand if *cis*-regulatory elements outside of our duplicated regions might contribute to gene expression divergence. To do this, we examined expression quantitative loci (eQTLs) associated with HSD gene expression. Using our reanalyzed expression data and variant calls from the 1000 Genomes Project (N=460) (Genomes Project, Consortium et al. 2015), we identified 42 HSD genes with significant eQTLs (N=3,603 variants in 9,524 gene-variant pairs), representing a 1.5- to 3-fold

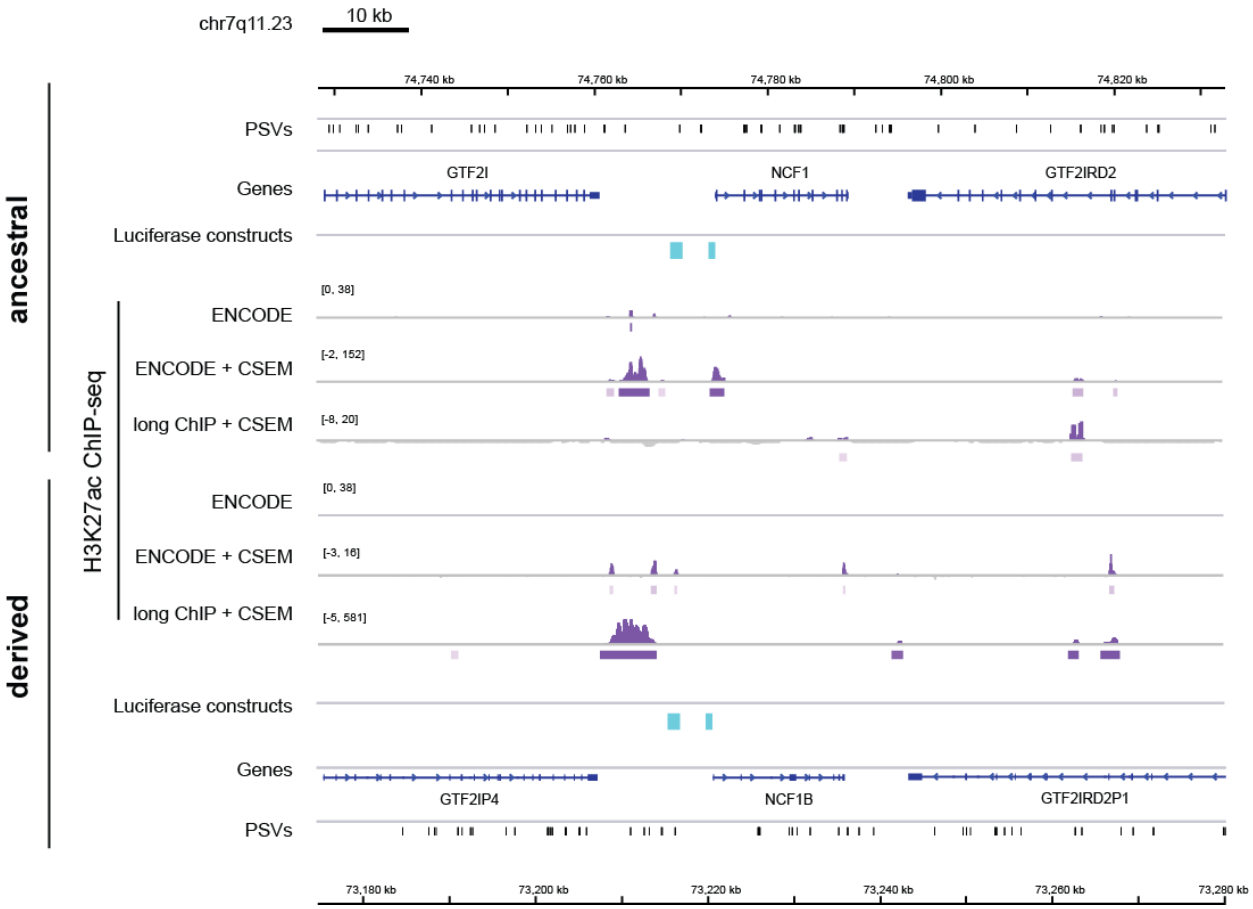

Figure 3. Histone H3K27ac landscape at the chromosome 7q11.23 HSD locus. The ancestral 7q11.23 HSD locus (top) and one of its derived loci (bottom) are aligned, and genes, PSVs, luciferase-tested regions (turquoise), and H3K27ac ChIP-seq signal (purple) are indicated. For each locus, publicly available ENCODE, reanalyzed ENCODE multimapping + CSEM allocation, and long ChIP multimapping + CSEM allocation data are shown.

increase from published work (Lappalainen et al. 2013; Wen et al. 2015). Sixty-four percent (N=2,300) of associated variants were located within SDs, of which 14% (N=498) fell within HSDs. Understanding that variants identified within SDs are likely to be unreliable (Hartasánchez et al. 2018; Ebbert et al. 2019) we focused on eQTLs in SD-proximal regions and found 577/1,303 eQTL variants with paralog-specific associations. For example, five SD-proximal variants were associated with *ARHGAP11A* expression, while none were identified for *ARHGAP11B*, which is located 2 Mb distal to its ancestral locus. Similarly, four eQTLs were identified for *DUSP22* on chromosome 6, while 26 eQTLs were linked with its derived paralog *DUSP22B* on chromosome 16. We also considered physical contacts generated by chromatin

looping of HSD promoters, as three-dimensional promoter-enhancer loops frequently indicate regulatory interactions. Using loops identified in GM12878 from promoter capture Hi-C (Mifsud et al. 2015) and H3K27ac HiChIP (Mumbach et al. 2017; Juric et al. 2019), we identified 354 and 26 regions, respectively, interacting with HSD promoters. Over 90% of these regions were located outside of HSDs, and 169 HSD eQTLs overlapped 31 of these regions (total size ~160 kb). These findings highlight the potential for adjacent, unique sequences to drive divergent regulation of HSDs genes.

Differential activity of *cis*-acting elements between duplicate paralogs

Using our combined datasets, we examined three HSD loci containing the gene families expressed most highly in LCLs (*ARHGAP11*, *NCF1*, and *DUSP22*) to identify functional changes in CREs that may contribute to paralogous expression divergence (Figures 3, S12–S14). In all three cases, the ancestral paralog exhibited significantly greater expression compared to derived paralog(s) (Figure 4A). To determine if sequence differences within CREs identified from our chromHMM annotations are sufficient to drive differences in gene expression, we performed luciferase-reporter assays in HeLa cells and LCLs on paralogous promoters and enhancer candidates.

ARHGAP11A/ARHGAP11B

The promoter of *ARHGAP11B* exhibited greater activity compared to the chimpanzee ortholog and ancestral paralog in both HeLa and LCLs (~4-fold difference in activity between HSD paralogs, $p<5\times10^{-10}$ in both cell lines; Figures 4B, S15A, S16A). This is in contrast to mRNA levels in LCLs, where the ancestral *ARHGAP11A* is more highly expressed. Positing that distal

elements may be driving differential expression between these paralogs, we identified putative enhancers unique to each paralog; these comprised one downstream of the *ARHGAP11A* duplicon and two downstream of *ARHGAP11B*. In HeLa cells, the *ARHGAP11A* element showed weak repressive activity (0.3-fold difference, $p<2\times10^{-16}$), while the *ARHGAP11B* elements showed modest activity over baseline (~2-fold difference, $p<2\times10^{-14}$ each), leaving the primary driver of differential expression for these genes unknown (Figures 4C, S15B).

NCF1/NCF1B/NCF1C

Promoters of the ancestral *NCF1* and its derived paralogs *NCF1B* and *NCF1C* genes did not exhibit significant differential activities in LCLs and modest difference in HeLa (0.8-fold difference, $p<0.001$; Figures 4A, S15A, S16A). However, an enhancer element common to all three paralogs showed the greatest activity for the ancestral *NCF1* paralog in both cell types. This was concordant with differential mRNA levels (~3-fold difference over either derived in LCL; $p<0.002$ for all comparisons) (Figures 4, S16B). Thus, this enhancer, if targeted to *NCF1* and its paralogs, may be responsible for differences in their mRNA levels.

DUSP22/DUSP22B

DUSP22 (ancestral) and *DUSP22B* (derived) promoters showed differential activity concordant with gene expression (i.e., the human ancestral paralog exhibited significantly greater activity than both the human derived and chimpanzee ortholog) (~1.5-fold difference; $p<5\times10^{-13}$) in both HeLa and LCLs (Figures 4B, S15A, S16A). We also tested six putative enhancers shared between the two paralogs in HeLa cells only. Four were active elements, two showed differential activity opposite to that of gene expression, and one element tracking with differential paralog

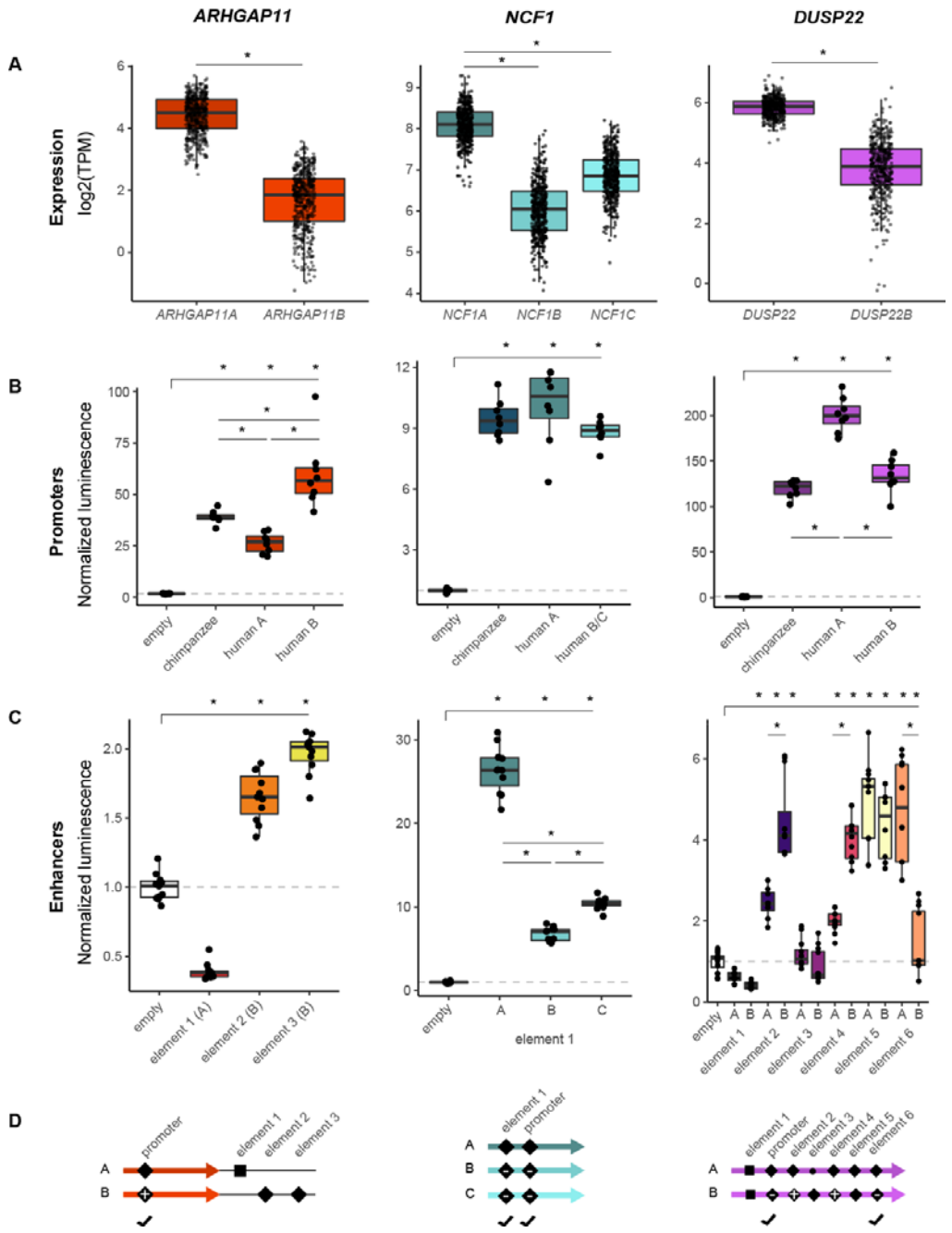

Figure 4. Functional characterization of putativef HSD CREs. Putative CREs (promoters and enhancers) from three HSD duplicate gene families (*ARHGAP11*, *NCF1*, and *DUSP22*) were tested in luciferase-reporter assays for function. **(A)** mRNA levels (TPM) for the three tested HSD gene families in human LCLs (N=446). **(B)** Representative luciferase reporter experiments for promoters of the paralogous HSD genes and orthologous chimpanzee sequences in HeLa cells. Significantly different activity (p<0.05, Tukey's post-hoc test of ANOVA) from the negative control is indicated along the top bar over each panel, and significant differences among homologous sequences are indicated between boxplots. The p-values for each comparison are available in Table S7. **(C)** Representative luciferase reporter experiments for candidate enhancers from the same gene families in HeLa cells, with significant activity over/under baseline indicated along the top bar, and significant differences between paralogous sequences between boxplots (p<0.05, Tukey's post-hoc test of ANOVA). **(D)** Cartoons indicating the relative locations of each candidate sequence within or adjacent to HSDs (thick, colored arrows), based on all experiments (Figure S15). Inactive sequences are shown with a small dot, enhancer sequences are shown with a diamond, and silencer sequences are shown with a square. Differentially active derived sequences (relative to ancestral) are marked with a plus or minus sign. Elements validated in LCLs are indicated with a check mark.

expression (Figure 4C). We subsequently validated the latter enhancer element in LCLs (~1.4-fold difference; $p<2\times10^{-16}$) (Figure S16C). We conclude that the difference in promoter activity is the primary driver of *DUSP22* and *DUSP22B* differential expression, though distal CREs also likely play a role in modulating transcription.

DISCUSSION

In this work, we provide evidence that recently duplicated, human-specific genes exhibit differential expression, at least in part due to divergent *cis*-acting regulation. Historically, these regions have been poorly characterized genetically and epigenetically. Here, we performed an improved analysis of gene expression and chromatin histone post-translational modification data to maximize information retained in HSDs and gain an understanding of the regulatory landscapes of these regions. Although previous work (Dennis et al. 2017; Dougherty et al. 2017) has quantified HSD gene expression across many tissues, these analyses were limited to a single canonical isoform from standard (RefSeq) annotations. This analysis is the first to consider all annotated isoforms for each HSD gene and to improve or create annotations for derived HSD genes. In addition, our implementation of CSEM (Chung et al. 2011) allowed ChIP-seq peaks to be identified in HSD loci, an improvement upon widely-used ENCODE data.

We highlight three loci whose regulation is of potential relevance to human-specific traits. Functional studies of *ARHGAP11B*, which encodes a Rho-GTPase activating protein, suggest this gene has been neofunctionalized in relation to its ancestral paralog by its unique expression in cortical progenitor neurons (basal radial glia) (Florio et al. 2015). Though the function(s) of this gene in LCLs is uncertain, it causes increased cortical neurogenesis in a number of model

systems (Kalebic et al. 2018; Namba et al. 2020). Our characterization of CRE in LCLs may be more directly related to immune phenotypes; it has been suggested that humans are more prone to autoimmune disease than chimpanzees, particularly as a result of T and B cell response to viral infection (Anon 2013; Anon 2017). The ancestral *NCF1*, which encodes Neutrophil Cytosolic Factor 1, has been shown to regulate autoimmune response via T cell activation in mice (Hultqvist et al. 2004)copy-number differences of the gene and its derived full-length paralogs (*NCF1B* and *NCF1C*) are associated with reduced risk of systemic lupus erythematosus, suggesting an additive functional effect (Zhao et al. 2017)(Hultqvist et al. 2004; Li et al. 2014) Finally, *DUSP22*, which encodes a tyrosine phosphatase, is also a regulator of autoimmune response, and *dusp22* knockout mice experience T cell proliferation, increased inflammation, and autoimmune encephalomyelitis (Li et al. 2014). The full-length paralog *DUSP22B* is uncharacterized, as it is missing from the human reference but located on chromosome 16p12.1 at variable CN (Figure 1D). Our characterization of promoters and distal regulatory elements of these three gene families uncovered potential drivers of differential expression; the the ancestral *DUSP22* promoter and an enhancer element upstream of *NCF1* both show activity patterns concordant with the mRNA levels of these gene families, for which the ancestral gene is most highly expressed. Searching for potential *trans* effectors, we identified an ancestral *DUSP22* deletion of four bases in a homopolymer repeat, with as many as 13 similar TFBSs found only in the less active *DUSP22B* and chimpanzee *DUSP22* ortholog. Some of these belonged to transcriptional repressors (ZNF394 and ZNF350). Examining predicted TFBSs within *NCF1* promoters, which did not exhibit differential activity, we observed no gains or losses of any TFBSs relative to chimpanzee. Puzzlingly, no predicted TFBSs were unique to the most active *NCF1* enhancer, but the paralogous *NCF1B* and *NCF1C* possessed many TFBSs that were

missing from the ancestral, at least one of which belonged to the transcriptional repressor ZNF394. Overall, these results point to diverse mechanisms contributing to differential expression between these three gene families and represent, to our knowledge, the first direct comparison and verification of active CREs between recently duplicated gene paralogs in humans.

In agreement with previous analysis of whole-genome duplications in teleost fishes (Sandve et al. 2018), we find that the dominant scenario for HSD genes involves a single paralog under stronger selection, resulting in maintenance of ancestral gene expression. This is largely consistent with nonfunctionalization or neofunctionalization of the other duplicate paralog; some gene families display patterns consistent with subfunctionalization (greater correlation of summed expression with chimpanzee than any individual paralog), but none of these exceed a difference in correlation of 0.05. Unlike previous work, these duplications arose much more recently (a few versus tens to hundreds millions of years) (Braasch et al. 2016; Lien et al. 2016), suggesting that these changes were rapidly acquired, though these analyses also rely on a much smaller set of genes and fewer tissues. Additionally, our finding that the majority of HSD paralogs show reduced expression is in line with a report that duplicate genes in mammals and yeast are less expressed, probably as result of neutral evolution (Qian et al. 2010). In contrast, a study by Lan and Pritchard (2016) found that mammalian tandem duplicated genes tend to be co-regulated and subfunctionalized by dosage. Notably, many of the HSD genes in our study were excluded from the previous analysis due to high sequence-identity. Further, most of our HSD genes are interspersed many hundreds of kilobases from each other throughout the human genome. They and others (Rodin and Riggs 2003) suggest that divergence of gene regulation is

more likely in the case of large genomic rearrangements, which can at birth place daughter paralogs in a novel regulatory environments, including topological domains, heterochromatin, or transcriptional hubs. We note that differential expression is still observed in even very recent (<1 million years) duplications (Dennis et al. 2017), such as the gene families *DUSP22*, *SERF1*, *SMN*, *TCAF1*, and *TCAF2*. Thus, these differences were either rapid to evolve or existed at the time of duplication.

Though this work represents an important step toward a more complete picture of the regulation of HSD genes, there are still some technical limitations to overcome. We implemented a mapping-free approach for transcript quantification, which circumvents multiple-mapping issues that affect "traditional" RNA-seq analyses and is demonstrated to accurately distinguish highly similar transcripts (Patro et al. 2017). However, reads originating from multiple transcripts are assigned probabilistically, and, as such, some genes may appear artificially similar in expression. For instance, *DUSP22B* expression was nonzero in individuals completely missing this paralog (Figure 1C) due to the recent nature of this duplication (<1 million years), which has resulted in very few PSVs differentiating the duplicates at the mRNA level. Conversely, RNA-seq quantification of older gene families appeared entirely distinguishable between paralogs (*HIST2H2BF*). In addition, the available contigs of complete HSD loci have only been generated from one effectively haploid cell line (Dennis et al. 2017). As such, PSVs are not necessarily fixed, and may also be shared between genes (Dumont 2015). Accordingly, a more advanced understanding of genetic variation at HSD loci is required to separate similarly expressed genes with maximum leverage. This would also allow for the identification of reliable eQTLs within duplications. We also find, as expected, that longer read lengths allow greater bioinformatic

distinction between duplicated loci, which remains an inherent limitation of many existing data sources, such as the relatively short reads (~30 nt) of ENCODE ChIP-seq. While longer reads are fundamentally more informative, some discrepancies between the ENCODE and long ChIP CSEM analyses could be a result of misallocation of reads. We were encouraged to find that a standard single-mapping approach (bwa mem) identified novel peaks in HSD, albeit at a lower rate than genome-wide. As such, longer ChIP read lengths can be used in future studies to further describe the epigenetic state of duplicated loci with a broader array of epitopes and cell/tissue types. Finally, chromatin conformation information will be necessary to physically link our tested distal elements to their target promoters, as linear proximity within the genome is an imperfect predictor. Like other genomic assays, these data are currently only sparsely available in SDs.

Despite limitations, we show that HSD regions contain differentially regulated genes and have built maps of potentially active chromatin in these regions. Using these data, we demonstrated that recently duplicated regulatory sequences are often functional, as expected, and that small sequence changes can significantly alter their activity. Our study did not pinpoint a universal factor that gives rise to the observed differential expression between paralogs, and the underlying molecular mechanisms are likely unique to each HSD gene. In the three duplicate gene families tested, promoter activity is only sometimes concordant with overall gene expression, suggesting that other types of regulatory elements, like enhancers and silencers, may cooperatively control overall expression. Currently, the challenge is to determine which PSVs and non-duplicated regions are functionally relevant. We have produced and leveraged a variety of analyses to narrow down likely candidates by chromatin state, changes to TFBSs, expression modulation,

and physical proximity to promoters. However, the number of candidate regions is too great to test via low-throughput methods such as luciferase reporters, and this problem is exacerbated by the need to compare regulatory behavior across multiple cell types. We propose that massively-parallel reporter assays could be implemented to validate and quantify CRE activity of thousands of candidate paralogous sequences. Such data could determine to what extent HSD gene expression is predicted by nearby CRE activity. Future studies could also integrate additional types of data, such as targeted chromatin capture data (such as 4C or capture Hi-C), or nascent transcription (GROseq, 5' CAGE). Finally, characterization of DNA methylation, which is especially challenging in duplicated loci, will be vital to build a more complete picture of the epigenetic landscape of HSD loci. This study represents a first step toward improving quantification of gene expression and active chromatin states in recent duplications and provides a foundation for future work characterizing regulatory and functional changes in recently duplicated loci.

MATERIALS AND METHODS

Quantification of HSD gene expression

Due to poor annotation of many HSD paralogs, custom transcriptomes were generated to ensure equivalent isoform models for paralogous genes, biasing against differential expression. First, transcript sequences for ancestral genes were extracted from GENCODE and mapped to derived human loci (contig from (Dennis et al. 2017) or GRCh38 for *SERF1B*) and a long-read chimpanzee assembly (Kronenberg et al. 2018) using BLAT (Kent 2002). GENCODE v27 transcripts were used for human-chimpanzee comparisons, since the chimpanzee transcriptome (Kronenberg et al. 2018) was built on this version; for human-only analyses, GENCODE v32

was used. Alignments were manually curated, and new derived transcripts were extracted from contigs. These transcripts, in addition to HSD transcripts generated from whole-isoform sequencing of brain tissue (Dougherty et al. 2018), were added to GENCODE (human) or the chimpanzee (after aligning to the chimpanzee assembly) transcriptome. Expression quantification was performed using Salmon v1.2.0 (Patro et al. 2017), the custom transcriptomes, and reference genomes (GRCh38 or Kronenberg et al.) as decoy sequence. For paired-end data, we used the flags "--validateMappings" and "--gcBias".. RNA-seq data were first lightly trimmed prior to quantification using trim_galore (https://github.com/FelixKrueger/TrimGalore) with the following flags: -q 20 --illumina --phred33 --length 20. Length-normalized TPM values or counts per gene were obtained using the tximport package in R (Soneson et al. 2015).

Differential expression analysis

Human and chimpanzee RNA-seq data from four primary tissues (Blake et al. 2020), LCLs (Khan et al. 2013; Blake et al. 2020), induced pluripotent stem cells (iPSCs) (Pavlovic et al. 2018), and iPSC-derived neural progenitor cells (Marchetto et al. 2019) were analyzed as described above. Count data from chimpanzee genes were duplicated to allow for pairwise comparison to each HSD duplicate, as well as the sum of all HSD genes in each family. Genes expressed below the 75% percentile (corresponding to 1-2 counts per million reads) were filtered from the analysis, leaving 16,752–18,225 genes. A linear model including species and sex was fitted to each shared gene (N=55,461) using limma-voom (Law et al. 2014; Ritchie et al. 2015), and differentially expressed genes were identified at a 5% false discovery rate (FDR) (Nguyen et al. 2012).

Copy number-controlled differential expression analysis

Paralog-specific CN estimates were generated using QuicK-mer2 (Shen and Kidd 2020),whole-genome sequence data from the 1000 Genomes Project (30X) (Fairley et al. 2020) and a custom reference consisting of GRCh38 plus an additional contig representing the *DUSP22B* duplicon (Dennis et al. 2017). Expression analysis was performed using RNA-seq data from LCLs included in the Geuvadis study (Lappalainen et al. 2013) for which CN genotypes were generated (N=445). Ancestral-derived gene pairs were compared with a linear model to identify significant differences in $\log_2$-transformed TPM values after controlling for continuous CN genotypes. Models were first fit with an interaction coefficient, and if no interaction was detected ($p > 0.05$), models were fit to expression and CN only. Resulting p-values were corrected via the Benjamini-Hochberg procedure using the R package qvalue (http://github.com/jdstorey/qvalue) and used to identify differential expression of ancestral-derived gene pairs at a 5% FDR. For visualization purposes (Figure 1D), *DUSP22* CN genotypes were adjusted to known values for GM12878 (as determined by fluorescence *in situ* hybridization in (Dennis et al. 2017)).

Cell Culture

Human LCLs were obtained from the Coriell Institute. The cells were grown in suspension in RPMI 1640 medium (Genesee Scientific) supplemented with 15% fetal bovine serum, 100 U/mL penicillin, and 100 µg/mL streptomycin and maintained at 37°C with 5% CO2. To test the impact of NMD inhibition, two million cells of each LCL (GM19204, GM18508, GM19193, GM19238, GM12878, and S003659_Chimp1) were grown overnight and subsequently treated with 100 µg/ml of emetine (Sigma) for seven hours (Noensie and Dietz 2001). Parallel cultures were left untreated and grown at standard conditions. HeLa cells were grown in Dulbecco's

Modified Eagle Medium (DMEM), High Glucose, with L-Glutamine (Genesee Scientific) supplemented with 10% fetal bovine serum (Gibco, Life Technologies), penicillin (100 U/mL) and streptomycin (100 µg/mL) (Gibco, Life Technologies) at 37°C with 5% CO2.

RNA extraction and cDNA generation

LCLs were harvested and added to an appropriate volume of TRIzol® solution (Invitrogen™) (1 ml per 10^7 cells) and stored at -80°C for ~24 hr before extraction to ensure complete lysis of cells. The next day, 200 µl of chloroform (Fisher Scientific) was added, and the homogenate was shaken vigorously for 20 seconds and incubated at room temperature for 2–3 min. Samples were spun at 10,000×*g* for 18 min at 4°C and the aqueous phase was transferred to a sterile RNase-free tube. An equal volume of 100% RNAse-free ethanol was added, samples were mixed by vortex, and then purified with an RNeasy Mini Kit (Qiagen). Samples were eluted in 30 µl RNase-free water and stored at -80°C. Transcriptor High Fidelity cDNA Synthesis Kit (Roche) was used for cDNA synthesis with OligodT primers. Following reverse transcription, samples were treated with RNase A (Qiagen) at 37°C, and cDNAs were stored at -20°C.

Identification of miRNA binding sites

For ancestral paralogs of each HSD gene family, the 3′-UTR was extracted from canonical transcript isoforms using the UCSC Genome Browser (GRCh38) and compared using blastn (Altschul et al. 1990) against existing alignments of homologs previously generated for human, chimpanzee, and rhesus (Dennis et al. 2017). Using TargetScan 7.0 and annotated miRNA sequences and families (release 7.1; Sept 2016) (Agarwal et al. 2015), we identified miRNA targets of individual human paralogs and non-human primate orthologs.

Correlation of expression skew and sequence divergence

Ancestral-derived paralog expression skew was calculated as the absolute value of $\log_2$(derived/ancestral), using the median TPM values for each gene and a pseudocount 1×10^{-4} (an order of magnitude below the smallest nonzero value). Sequence divergence as the pairwise identity with the ancestral sequence was taken from (Dennis et al. 2017). Gene families were included if at least one paralog was expressed at a level >1 TPM. For promoters, sequence divergence was tabulated as the sum of all mismatches and alignment gaps within ±500 base pairs of the transcription start site (Gencode v32). These quantities were correlated and the strength of the relationship was determined with a linear regression.

ChIP assays

ChIP assays were carried out as previously described with minor modifications (O'Geen et al. 2019). GM12878 cells were cross-linked in growth media containing 1% formaldehyde (Fisher Scientific BP531) for 10 min at room temperature and the reaction was stopped with 0.125 M glycine. Cross-linked cells were washed twice in PBS and stored at -80°C. 2×10^6 cells were used per ChIP assay. Cells from two biological replicates were lysed with ChIP lysis buffer (5 mM PIPES pH8, 85 mM KCl, 1% Igepal) with a protease inhibitor (PI) cocktail (Roche). Nuclei were collected by centrifugation at 2,000 rpm. for 5 min at 4°C and lysed in nuclei lysis buffer (50 mM Tris pH8, 10 mM EDTA, 1% SDS) supplemented with PI cocktail. Chromatin was fragmented in microTUBEs with the E220 (Covaris) using the low cell shearing protocol (Duty cycle 2%, PIP 105, CPB 200, 4 min) and diluted with 5 volumes of RIPA buffer (50 mM Tris pH 7.6, 150 mM NaCl, 1 mM EDTA pH8, 1% Igepal, 0.25% Deoxycholic acid). ChIP enrichment

was performed by incubation for 16 h at 4°C with the following antibodies: 2 µg H3K27ac antibody (Active Motif #39133), 4 µg H3K4me1 antibody (Millipore 07-436), 2 µg H3K4me3 antibody (Active Motif #39915), or 2 µg RNA Polymerase II (PolII) antibody clone 8WG16 (Covance MMS-126R). RNA PolII samples were incubated for an additional hour with 2 µg Rabbit Anti-Mouse IgG (MP Biomedical #55436). Immune complexes were bound to 20 µl magnetic protein A/G beads (ThermoFisher) for 2 hours at 4°C. Beads were washed 2x with RIPA, 3x with ChIP wash buffer (100 mM Tris pH8, 500 mM LiCl, 1% Deoxycholic acid) and once with ChIP wash buffer plus 150 mM NaCl. ChIP samples were eluted in 100 µl ChIP elution buffer (50 mM $NaHCO_3$, 1% SDS) and cross-linking reversed with addition of 0.5 M NaCl and heating at 65°C overnight. Samples were treated with 2 µg RNaseA (Qiagen) and DNA was purified using the QIAquick PCR Purification Kit (Qiagen). ChIP enrichments were confirmed by qPCR with 2× SYBR FAST mastermix (KAPA Biosystems) using the CFX384 Real-Time System C1000 Touch Thermo Cycler (BioRad). ACTB primers served as positive control and HER2 primers as negative controls (Table S8). ChIP enrichment was calculated relative to input samples using the dC_t method ($dC_t = C_t$[HER2-ChIP]-C_t[input]). Each entire ChIP sample was used to prepare Illumina sequencing libraries using the KAPA Hyper Prep Kit (Roche). Adapter-ligated DNA was separated on a 2% E-Gel EX (Invitrogen) and the 500-800 bp fraction was excised and purified using the QIAquick gel extraction Kit (Qiagen). Indexed primers were used to generate dual-indexed libraries and amplified libraries were size selected (500–700 bp) using the PippenHT (Sage Science). Equimolar library amounts were pooled and sequenced on the NovaSeq SP (Illumina).

Analysis of ChIP-seq data

ChIP-seq peaks obtained with the ENCODE pipeline were directly downloaded from the ENCODE portal (Davis et al. 2018) for H3K4me3 (ENCFF228GWY), H3K27ac (ENCFF367KIF), H3K4me1 (ENCFF453PEP), POLR2A (ENCFF455ZLJ), and H3K27me3 (ENCFF153VOQ). For "short" ChIP-seq peak calling using raw ENCODE data, GM12878 ChIP-seq reads were downloaded from the ENCODE portal for RNA Polymerase II (ENCSR000AKA), H3K4me3 (ENCSR000BGD), H3K4me1 (ENCSR000AKF), H3K27ac (ENCSR000AKC), and H3K27me3 (ENCFF000OBB). Illumina adapters and low quality bases (Phred score < 20) were trimmed using Trimmomatic (Bolger et al. 2014) (parameters SLIDINGWINDOW:4:20 MINLEN:20) and aligned to a custom reference genome (GRCh38 with an added *DUSP22B* contig) using single-end Bowtie (Langmead et al. 2009) configured to allow multiple mappings per read (parameters a -v2 -m99). After mapping, PCR duplicates were removed using Picard Markduplicates and secondary alignments were removed with samtools v1.9. Multi-mapping reads were allocated to their most likely position using CSEM v2.4 (Chung et al. 2011). CSEM was run using the --no-extending-reads option and the fragment size was calculated with phantompeakqualtools run_SPP.R script (Landt et al. 2012). A custom script was developed to select the alignment with the highest posterior probability as assigned by CSEM for each multimapping read, choosing one alignment randomly in case of a tie. Peaks were called using MACS2 callpeak (v2.2.6) on default settings using MACS2's shifting model (Zhang et al. 2008) (https://github.com/macs3-project/MACS). Broad peaks were called at a FDR of 5%, while narrow peaks were called at a FDR of 1%. BigWig files for peak's visualization were obtained with MACS2 bdgcmp tool and UCSC bedGraphToBigWig. For H3K27me3, which occurs in large repressive domains, enriched regions were identified with hiddenDomains (Starmer and Magnuson 2016). Paired-end long-ChIP reads were generated as described above

and Illumina adapters were removed using Trimmomatic (parameters SLIDINGWINDOW:4:30 MINLEN:50). Reads were mapped using both paired-end BWA mem and single-end Bowtie allowing for multiple mappings (parameters -a -n -S -e 200 -m 99). For single-end alignments, forward and reverse reads were concatenated into a single file and properly renamed to secure unique reads IDs. Reads aligned with BWA were filtered by MAPQ ≥ 20 while reads with multiple mappings aligned with Bowtie were allocated with CSEM and most likely alignments were selected with the custom script. Duplicates and secondary alignments were removed as explained above. Peaks were called using MACS2 with identical parameters used for short-reads, adding the BAMPE option in the case of paired-end reads aligned with BWA mem.

For depletion analyses, SD coordinates were directly downloaded from UCSC Table Browser and HSD coordinates were obtained by filtering alignments with sequence identity over 98% in the fracMatch column, converting them to BED format and merging overlapping entries using bedtools merge. The number of peaks and bases under peaks on each region of interest were obtained with bedtools intersect. To obtain depletion statistics, 1000 regions of the same size as SD and HSD were randomly sampled from the human genome GRCh38. Empirical *p*-values of depletion tests were calculated as *p*-value = (M+1)/(N+1), where M is the number of iterations less than the observed value and N is the number of iterations.

Additionally, mapping quality scores (MAPQ) distributions for H3K27ac following a similar approach as explained before, but using BWA aln and BWA mem for short and long ChIP-seq reads respectively, after PCR duplicates and secondary alignments removal. Posterior probabilities distributions for H3K27ac were examined using the output of CSEM after selecting

the most likely alignment with the custom script. Entries in unique space were subsampled to 10 million and plots were obtained with the geom_density function in ggplot R package.

All ChIP-seq analyses are available as a TrackHub for the UCSC Genome Browser (TBD). Further, the bioinformatic pipeline is freely available for use in Snakemake format (Supplementary File 1, github TBD), allowing the analysis to be replicated in any cell or tissue type of interest.

Paralog-specific validation of RNA expression and ChIP data

Following published protocols (Integrated DNA Technologies), we used the rhAMP assay in 10 µl total reaction volumes to quantify abundance of PSVs (for all assays except *ARHGAP11* expression, the fluorophores FAM=A paralog and VIC=B paralog) as a proxy for paralog-specific expression (RNA) and enrichment (ChIP) (Table S7). We used 10 ng total of RNA converted to cDNA to validate gene expression for duplicated gene families *ARHGAP11*, *ROCK1*, and *DUSP22*. We calculated dCt of cDNA and gDNA as Ct_{FAM}-Ct_{VIC} and ddCt as dCt_{cDNA}-dCt_{gDNA} from the same cell line. We calculated dCt of the input and ChIP-enriched library as Ct_{FAM}-Ct_{VIC} and ddCt as dCt_{ChIP}-dCt_{input} from the same cell line. For both expression and ChIP analyses, the ratio of abundance of the B to the A paralog is 2^{ddCt}.

chromHMM annotations

We generated models separately for ENCODE short-read data and long ChIP, after multimapping and allocation, using active chromatin histone modifications (H3K4me3, H3K4me1, and H3K27ac). States corresponding to active transcription start sites and active

enhancers were identified manually. These annotations are publicly available at UCSC Genome Browser TrackHub (TBD)).

Luciferase reporter assays

Promoters of highly and differentially expressed HSD gene families (*ARHGAP11*, *NCF1*, and *DUSP22*) were chosen for screening in a reporter assay. Fragments containing the TSS and spanning ~1 kb were amplified with KpnI and SacI restriction sites and cloned into the luciferase reporter vector pGL3-basic (Promega). Candidate enhancers within 50 kb of genes bodies were selected based on the presence of ChromHMM CREs in the re-analyzed data from human LCLs. Target regions were a maximum size of 5 kb, and peaks larger than this were tiled with multiple targets. Gateway homology arms were added to primers in accordance with the manual (ThermoFisher), and PCR products were cloned into the entry vector pDONR221 (ThermoFisher 12536017). Expression clones for luciferase assays were generated by cloning pDONR221 inserts into the luciferase reporter pE1B (Antonellis et al. 2008) with the Gateway system.

Constructs were co-transfected (ThermoFisher Lipofectamine 3000) in equimolar amounts with 50 ng of the control plasmid pRL-TK (Renilla luciferase) into HeLa cells in 96-well plates. Cells were at 70-90% confluence at the time of transfection. Luciferase assays were performed with the Dual-Luciferase Reporter Assay System (Promega E1910). 48 hours post-transfection, cells were washed with PBS, and lysed with Passive Lysis Buffer for at least 15 min sharking at 500 rpm. Lysates were stored at -80C. For LCLs, cells were split 48 and 24 hours pre-transfection to ensure active division. Cells were counted, washed in PBS, and resuspended such that each transfection contained 12.5×10^6 cells, 6.25 ug of test construct, and equimolar pRL-TK in RPMI.

Cells were electroporated using the Neon Transfection System in accordance with previously published work (Tewhey et al. 2018) and recovered at a density of 3×10^6 cells/mL in pre-warmed RPMI including 15% FBS without antibiotics. Transfection efficiencies of ~15% were achieved. To perform luciferase assays, ~$5\times x10^5$ cells were pipetted into each well of a 96-well plate, washed with PBS, and lysed with Passive Lysis Buffer as described for HeLa. Luminescence measurements were performed according to the manufacturer's instructions using a Tecan Infinite or Tecan Spark plate reader with injectors.

Transcription factors binding motifs

Alignments of cloned sequences were scanned for HOmo sapiens COmprehensive MOdel COllection (HOCOMOCO) v11 (Kulakovskiy et al. 2018) TFBS motifs using FIMO (Grant et al. 2011). HOCOMOCO motifs were limited to transcription factors expressed above 1 TPM in >75% of ENCODE mRNA-seq libraries generated for GM12878 (ENCSR077AZT, ENCLB555AQG, ENCLB555AQH, ENCLB555ANP, ENCLB555ALI, ENCLB555ANM, ENCLB555ANN, ENCLB037ZZZ, ENCLB038ZZZ, ENCLB043ZZZ, ENCLB044ZZZ, ENCLB041ZZZ, ENCLB042ZZZ, ENCLB045ZZZ, ENCLB046ZZZ, ENCLB700LMU, ENCLB150CGC). Significant matches above a 5% FDR were retained for the analysis. TFBSs were compared across homologous sequences to identify putative paralog-specific gains and losses of binding sites.

FIGURE LEGENDS

Figure 1. Expression patterns of HSD genes. (A) Illustration of genes residing within HSDs; the ancestral paralog (green) corresponds to the chimpanzee ortholog, while derived paralogs

(orange) are human-specific. The ancestral and derived genes comprise a gene family. **(B)** Cross-tissue correlations of HSD genes or gene family summed expression versus expression of chimpanzee orthologs are shown in boxplots. For each gene family, the most conserved paralog ("highest"), all other paralogs ("lower"), and the sum of all paralogs ("sum") are plotted separately. Ancestral genes, if evolutionary status is known, are denoted as a filled circle. Asterisks denote significant differences ($p<0.01$, Tukey post-hoc test of ANOVA) **(C)** Expression divergence of derived genes from families with at least one LCL-expressed paralog is plotted as the $\log_2$ ratio of median derived and ancestral TPM expression. Each point represents a different LCL from the Geuvadis consortium (total N=445). The gray bar indicates a two-fold expression difference. **(D)** Expression values of ancestral *DUSP22* (green) and derived *DUSP22B* (orange), stratified by CN. Points represent individual LCLs, and the number of individuals represented in each CN category is indicated over each boxplot.

Figure 2. Depletion and recovery of ChIP peaks in SDs. Results are summarized for three ChIP-seq peak discovery approaches: publicly-available ENCODE peaks (yellow), peaks from multimapping and CSEM allocation of ENCODE raw data (blue), and peaks from multimapping and CSEM allocation of long ChIP data. For SDs (left) and SDs with greater than 98% sequence identity (or HSDs, right), the fraction of bases under peaks was calculated (solid vertical lines). The SD coordinates were permuted 1000 times within the human reference (GRCh38), and an expected distribution of the fraction of bases covered was generated. Empirical one-sided p-values for depletion are indicated above each plot.

Figure 3. Histone H3K27ac landscape at the chromosome 7q11.23 HSD locus. The ancestral 7q11.23 HSD locus (top) and one of its derived loci (bottom) are aligned, and genes, PSVs, luciferase-tested regions (turquoise), and H3K27ac ChIP-seq signal (purple) are indicated. For each locus, publicly available ENCODE, reanalyzed ENCODE multimapping + CSEM allocation, and long ChIP multimapping + CSEM allocation data are shown.

Figure 4. Functional characterization of putativef HSD CREs. Putative CREs (promoters and enhancers) from three HSD duplicate gene families (*ARHGAP11*, *NCF1*, and *DUSP22*) were tested in luciferase-reporter assays for function. **(A)** mRNA levels (TPM) for the three tested HSD gene families in human LCLs (N=446). **(B)** Representative luciferase reporter experiments for promoters of the paralogous HSD genes and orthologous chimpanzee sequences in HeLa cells. Significantly different activity ($p<0.05$, Tukey's post-hoc test of ANOVA) from the negative control is indicated along the top bar over each panel, and significant differences among homologous sequences are indicated between boxplots. The *p*-values for each comparison are available in Table S7. **(C)** Representative luciferase reporter experiments for candidate enhancers from the same gene families in HeLa cells, with significant activity over/under baseline indicated along the top bar, and significant differences between paralogous sequences between boxplots ($p<0.05$, Tukey's post-hoc test of ANOVA). **(D)** Cartoons indicating the relative locations of each candidate sequence within or adjacent to HSDs (thick, colored arrows), based on all experiments (Figure S15). Inactive sequences are shown with a small dot, enhancer sequences are shown with a diamond, and silencer sequences are shown with a square. Differentially active derived sequences (relative to ancestral) are marked with a plus or minus sign. Elements validated in LCLs are indicated with a check mark.

ACKNOWLEDGMENTS

We thank the many groups/consortia that have made their data publicly available, including Dr. Yoav Gilad, the 1000 Genomes Project, GTEX, and ENCODE for which this research would not be possible without its use. We also thank Dr. Colin Kern for valuable advice concerning ChIP-seq analysis, Dr. Anthony Antonellis for sharing the pE1B enhancer reporter Gateway plasmid, and Drs. Gerald Quon and Siobhan Brady for constructive feedback on the manuscript. This work was supported by the National Institute of Neurological Disorders and Stroke at the National Institutes of Health (NIH) (R00NS083627 to M.Y.D.). Statistical analysis advice was provided by Dr. Blythe Durbin-Johnson through the MIND Institute Intellectual and Developmental Disability Research Center, funded by the National Institute of Child Health and Human Development (U54 HD079125). Additionally, M.Y.D. is supported as a Sloan fellow (FG-2016-6814), P.C.M as an NIH National Institute of Mental Health T32 UC Davis Autism Research Training Program fellow (5T32MH073124-17), D.C.S. as a Fulbright fellow, and J.R. as an NIH National Institute of General Medical Sciences UC Davis Postbaccalaureate Research Education Program fellow (R25GM116690).

DATA ACCESS

Long-insert ChIP-sequencing data generated for this study are available from the European Nucleotide Archive under the accession PRJEB40356.

DISCLOSURE DECLARATION

Authors have no disclosures to declare.

REFERENCES

Agarwal V, Bell GW, Nam J-W, Bartel DP. 2015. Predicting effective microRNA target sites in mammalian mRNAs. *eLife* [Internet] 4. Available from: http://dx.doi.org/10.7554/elife.05005

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J. Mol. Biol.* 215:403–410.

Anon. 2013. Potential role of human-specific genes, human-specific microRNAs and human-specific non-coding regulatory RNAs in the pathogenesis of Systemic Sclerosis and Sjögren's Syndrome. *Autoimmun. Rev.* 12:1046–1051.

Anon. 2017. Are humans prone to autoimmunity? Implications from evolutionary changes in hominin sialic acid biology. *J. Autoimmun.* 83:134–142.

Antonacci F, Dennis MY, Huddleston J, Sudmant PH, Steinberg KM, Rosenfeld JA, Miroballo M, Graves TA, Vives L, Malig M, et al. 2014. Palindromic GOLGA8 core duplicons promote chromosome 15q13.3 microdeletion and evolutionary instability. *Nature Genetics* [Internet] 46:1293–1302. Available from: http://dx.doi.org/10.1038/ng.3120

Antonellis A, Huynh JL, Lee-Lin S-Q, Vinton RM, Renaud G, Loftus SK, Elliot G, Wolfsberg TG, Green ED, McCallion AS, et al. 2008. Identification of neural crest and glial enhancers at the mouse Sox10 locus through transgenesis in zebrafish. *PLoS Genet.* 4:e1000174.

Bailey JA. 2002. Recent Segmental Duplications in the Human Genome. *Science* [Internet] 297:1003–1007. Available from: http://dx.doi.org/10.1126/science.1072047

Blake LE, Roux J, Hernando-Herraez I, Banovich NE, Perez RG, Hsiao CJ, Eres I, Cuevas C, Marques-Bonet T, Gilad Y. 2020. A comparison of gene expression and DNA methylation patterns across tissues and species. *Genome Res.* 30:250–262.

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

Braasch I, Gehrke AR, Smith JJ, Kawasaki K, Manousaki T, Pasquier J, Amores A, Desvignes T, Batzel P, Catchen J, et al. 2016. The spotted gar genome illuminates vertebrate evolution and facilitates human-teleost comparisons. *Nature Genetics* [Internet] 48:427–437. Available from: http://dx.doi.org/10.1038/ng.3526

Charrier C, Joshi K, Coutinho-Budd J, Kim J-E, Lambert N, de Marchena J, Jin W-L, Vanderhaeghen P, Ghosh A, Sassa T, et al. 2012. Inhibition of SRGAP2 Function by Its Human-Specific Paralogs Induces Neoteny during Spine Maturation. *Cell* [Internet] 149:923–935. Available from: http://dx.doi.org/10.1016/j.cell.2012.03.034

Chikina MD, Troyanskaya OG. 2012. An effective statistical evaluation of ChIPseq dataset similarity. *Bioinformatics* 28:607–613.

Chung D, Kuan PF, Li B, Sanalkumar R, Liang K, Bresnick EH, Dewey C, Keleş S. 2011. Discovering transcription factor binding sites in highly repetitive regions of genomes with multi-read analysis of ChIP-Seq data. *PLoS Comput. Biol.* 7:e1002111.

Consortium TEP, The ENCODE Project Consortium. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* [Internet] 489:57–74. Available from: http://dx.doi.org/10.1038/nature11247

Davis CA, Hitz BC, Sloan CA, Chan ET, Davidson JM, Gabdank I, Hilton JA, Jain K, Baymuradov UK, Narayanan AK, et al. 2018. The Encyclopedia of DNA elements (ENCODE): data portal update. *Nucleic Acids Research* [Internet] 46:D794–D801. Available from: http://dx.doi.org/10.1093/nar/gkx1081

Dennis MY, Eichler EE. 2016. Human adaptation and evolution by segmental duplication. *Current Opinion in Genetics & Development* [Internet] 41:44–52. Available from: http://dx.doi.org/10.1016/j.gde.2016.08.001

Dennis MY, Harshman L, Nelson BJ, Penn O, Cantsilieris S, Huddleston J, Antonacci F, Penewit K, Denman L, Raja A, et al. 2017. The evolution and population diversity of human-specific segmental duplications. *Nat Ecol Evol* 1:69.

Dennis MY, Nuttle X, Sudmant PH, Antonacci F, Graves TA, Nefedov M, Rosenfeld JA, Sajjadian S, Malig M, Kotkiewicz H, et al. 2012. Evolution of Human-Specific Neural SRGAP2 Genes by Incomplete Segmental Duplication. *Cell* [Internet] 149:912–922. Available from: http://dx.doi.org/10.1016/j.cell.2012.03.033

Dougherty ML, Nuttle X, Penn O, Nelson BJ, Huddleston J, Baker C, Harshman L, Duyzend MH, Ventura M, Antonacci F, et al. 2017. The birth of a human-specific neural gene by incomplete duplication and gene fusion. *Genome Biol.* 18:49.

Dougherty ML, Underwood JG, Nelson BJ, Tseng E, Munson KM, Penn O, Nowakowski TJ, Pollen AA, Eichler EE. 2018. Transcriptional fates of human-specific segmental duplications in brain. *Genome Research* [Internet] 28:1566–1576. Available from: http://dx.doi.org/10.1101/gr.237610.118

Dumont BL. 2015. Interlocus gene conversion explains at least 2.7 % of single nucleotide variants in human segmental duplications. *BMC Genomics* [Internet] 16. Available from: http://dx.doi.org/10.1186/s12864-015-1681-3

Ebbert MTW, Jensen TD, Jansen-West K, Sens JP, Reddy JS, Ridge PG, Kauwe JSK, Belzil V, Pregent L, Carrasquillo MM, et al. 2019. Systematic analysis of dark and camouflaged genes reveals disease-relevant genes hiding in plain sight. *Genome Biol*. 20:97.

Ernst J, Kellis M. 2012. ChromHMM: automating chromatin-state discovery and characterization. *Nature Methods* [Internet] 9:215–216. Available from: http://dx.doi.org/10.1038/nmeth.1906

Fairley S, Lowy-Gallego E, Perry E, Flicek P. 2020. The International Genome Sample Resource

(IGSR) collection of open human genomic variation resources. *Nucleic Acids Res.* 48:D941–D947.

Fiddes IT, Lodewijk GA, Mooring M, Bosworth CM, Ewing AD, Mantalas GL, Novak AM, van den Bout A, Bishara A, Rosenkrantz JL, et al. 2018. Human-Specific NOTCH2NL Genes Affect Notch Signaling and Cortical Neurogenesis. *Cell* [Internet] 173:1356–1369.e22. Available from: http://dx.doi.org/10.1016/j.cell.2018.03.051

Florio M, Albert M, Taverna E, Namba T, Brandl H, Lewitus E, Haffner C, Sykes A, Wong FK, Peters J, et al. 2015. Human-specific gene ARHGAP11B promotes basal progenitor amplification and neocortex expansion. *Science* [Internet] 347:1465–1470. Available from: http://dx.doi.org/10.1126/science.aaa1975

Genomes Project, Consortium, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, Marchini JL, McCarthy S, McVean GA, et al. 2015. A global reference for human genetic variation. *Nature* 526:68–74.

Giannuzzi G, Migliavacca E, Reymond A. 2014. Novel H3K4me3 marks are enriched at human- and chimpanzee-specific cytogenetic structures. *Genome Research* [Internet] 24:1455–1468. Available from: http://dx.doi.org/10.1101/gr.167742.113

Grant CE, Bailey TL, Noble WS. 2011. FIMO: scanning for occurrences of a given motif. *Bioinformatics* [Internet] 27:1017–1018. Available from: http://dx.doi.org/10.1093/bioinformatics/btr064

Hartasánchez DA, Brasó-Vives M, Heredia-Genestar JM, Pybus M, Navarro A. 2018. Effect of Collapsed Duplications on Diversity Estimates: What to Expect. *Genome Biol. Evol.* 10:2899–2905.

Heide M, Haffner C, Murayama A, Kurotaki Y, Shinohara H, Okano H, Sasaki E, Huttner WB. 2020. Human-specific ARHGAP11B increases size and folding of primate neocortex in the fetal marmoset. *Science* [Internet]:eabb2401. Available from: http://dx.doi.org/10.1126/science.abb2401

Hultqvist M, Olofsson P, Holmberg J, Backstrom BT, Tordsson J, Holmdahl R. 2004. Enhanced autoimmunity, arthritis, and encephalomyelitis in mice with a reduced oxidative burst due to a mutation in the Ncf1 gene. *Proceedings of the National Academy of Sciences* [Internet] 101:12646–12651. Available from: http://dx.doi.org/10.1073/pnas.0403831101

Ishiura H, Shibata S, Yoshimura J, Suzuki Y, Qu W, Doi K, Asem Almansour M, Kikuchi JK, Taira M, Mitsui J, et al. 2019. Noncoding CGG repeat expansions in neuronal intranuclear inclusion disease, oculopharyngodistal myopathy and an overlapping disease. *Nature Genetics* [Internet] 51:1222–1232. Available from: http://dx.doi.org/10.1038/s41588-019-0458-z

Juric I, Yu M, Abnousi A, Raviram R, Fang R, Zhao Y, Zhang Y, Qiu Y, Yang Y, Li Y, et al. 2019. MAPS: Model-based analysis of long-range chromatin interactions from PLAC-seq and HiChIP experiments. *PLoS Comput. Biol.* 15:e1006982.

Kalebic N, Gilardi C, Albert M, Namba T, Long KR, Kostic M, Langen B, Huttner WB. 2018. Human-specific ARHGAP11B induces hallmarks of neocortical expansion in developing ferret neocortex. *eLife* [Internet] 7. Available from: http://dx.doi.org/10.7554/elife.41241

Kassahn KS, Dang VT, Wilkins SJ, Perkins AC, Ragan MA. 2009. Evolution of gene function and regulatory control after whole-genome duplication: comparative analyses in vertebrates. *Genome Res.* 19:1404–1418.

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

Khan Z, Ford MJ, Cusanovich DA, Mitrano A, Pritchard JK, Gilad Y. 2013. Primate transcript and protein expression levels evolve under compensatory selection pressures. *Science* 342:1100–1104.

Kronenberg ZN, Fiddes IT, Gordon D, Murali S, Cantsilieris S, Meyerson OS, Underwood JG, Nelson BJ, Chaisson MJP, Dougherty ML, et al. 2018. High-resolution comparative analysis of great ape genomes. *Science* [Internet] 360. Available from: http://dx.doi.org/10.1126/science.aar6343

Kulakovskiy IV, Vorontsov IE, Yevshin IS, Sharipov RN, Fedorova AD, Rumynskiy EI, Medvedeva YA, Magana-Mora A, Bajic VB, Papatsenko DA, et al. 2018. HOCOMOCO: towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-Seq analysis. *Nucleic Acids Res.* 46:D252–D259.

Landt SG, Marinov GK, Kundaje A, Kheradpour P, Pauli F, Batzoglou S, Bernstein BE, Bickel P, Brown JB, Cayting P, et al. 2012. ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Res.* 22:1813–1831.

Langmead B, Trapnell C, Pop M, Salzberg SL. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* 10:R25.

Lan X, Pritchard JK. 2016. Coregulation of tandem duplicate genes slows evolution of subfunctionalization in mammals. *Science* [Internet] 352:1009–1013. Available from: http://dx.doi.org/10.1126/science.aad8411

Lappalainen T, The Geuvadis Consortium, Sammeth M, Friedländer MR, ‘t PA, Monlong J, Rivas MA, Gonzàlez-Porta M, Kurbatova N, Griebel T, et al. 2013. Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* [Internet] 501:506–511. Available from: http://dx.doi.org/10.1038/nature12531

Law CW, Chen Y, Shi W, Smyth GK. 2014. voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biology* [Internet] 15:R29. Available from: http://dx.doi.org/10.1186/gb-2014-15-2-r29

Lien S, Koop BF, Sandve SR, Miller JR, Kent MP, Nome T, Hvidsten TR, Leong JS, Minkley DR, Zimin A, et al. 2016. The Atlantic salmon genome provides insights into rediploidization. *Nature* [Internet] 533:200–205. Available from: http://dx.doi.org/10.1038/nature17164

Li J-P, Yang C-Y, Chuang H-C, Lan J-L, Chen D-Y, Chen Y-M, Wang X, Chen AJ, Belmont JW, Tan T-H. 2014. The phosphatase JKAP/DUSP22 inhibits T-cell receptor signalling and autoimmunity by inactivating Lck. *Nature Communications* [Internet] 5. Available from: http://dx.doi.org/10.1038/ncomms4618

Lynch M. 2000. The Evolutionary Fate and Consequences of Duplicate Genes. *Science* [Internet] 290:1151–1155. Available from: http://dx.doi.org/10.1126/science.290.5494.1151

Marchetto MC, Hrvoj-Mihic B, Kerman BE, Yu DX, Vadodaria KC, Linker SB, Narvaiza I, Santos R, Denli AM, Mendes APD, et al. 2019. Species-specific maturation profiles of human, chimpanzee and bonobo neural cells. *eLife* [Internet] 8. Available from: http://dx.doi.org/10.7554/elife.37527

McVicker G, van de Geijn B, Degner JF, Cain CE, Banovich NE, Raj A, Lewellen N, Myrthil M, Gilad Y, Pritchard JK. 2013. Identification of Genetic Variants That Affect Histone Modifications in Human Cells. *Science* [Internet] 342:747–749. Available from: http://dx.doi.org/10.1126/science.1242429

Mifsud B, Tavares-Cadete F, Young AN, Sugar R, Schoenfelder S, Ferreira L, Wingett SW, Andrews S, Grey W, Ewels PA, et al. 2015. Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. *Nat. Genet.* 47:598–606.

Mumbach MR, Satpathy AT, Boyle EA, Dai C, Gowen BG, Cho SW, Nguyen ML, Rubin AJ, Granja JM, Kazane KR, et al. 2017. Enhancer connectome in primary human cells identifies target genes of disease-associated DNA elements. *Nat. Genet.* 49:1602–1612.

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

Nguyen LS, Jolly L, Shoubridge C, Chan WK, Huang L, Laumonnier F, Raynaud M, Hackett A, Field M, Rodriguez J, et al. 2012. Transcriptome profiling of UPF3B/NMD-deficient lymphoblastoid cells from patients with various forms of intellectual disability. *Molecular Psychiatry* [Internet] 17:1103–1115. Available from: http://dx.doi.org/10.1038/mp.2011.163

Noensie EN, Dietz HC. 2001. A strategy for disease gene identification through nonsense-mediated mRNA decay inhibition. *Nat. Biotechnol.* 19:434–439.

O'Bleness M, Searles VB, Dickens CM, Astling D, Albracht D, Mak AC, Lai YY, Lin C, Chu C, Graves T, et al. 2014. Finished sequence and assembly of the DUF1220-rich 1q21 region using a haploid human genome. *BMC Genomics* 15:387.

O'Geen H, Bates SL, Carter SS, Nisson KA, Halmai J, Fink KD, Rhie SK, Farnham PJ, Segal DJ. 2019. Ezh2-dCas9 and KRAB-dCas9 enable engineering of epigenetic memory in a context-dependent manner. *Epigenetics Chromatin* 12:26.

Ohno S. 1970. Evolution by Gene Duplication. Available from: http://dx.doi.org/10.1007/978-3-

642-86659-3

Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. 2017. Salmon provides fast and bias-aware quantification of transcript expression. *Nature Methods* [Internet] 14:417–419. Available from: http://dx.doi.org/10.1038/nmeth.4197

Pavlovic BJ, Blake LE, Roux J, Chavarria C, Gilad Y. 2018. A Comparative Assessment of Human and Chimpanzee iPSC-derived Cardiomyocytes with Primary Heart Tissues. *Sci. Rep.* 8:15312.

Prado-Martinez J, Sudmant PH, Kidd JM, Li H, Kelley JL, Lorente-Galdos B, Veeramah KR, Woerner AE, O'Connor TD, Santpere G, et al. 2013. Great ape genetic diversity and population history. *Nature* 499:471–475.

Qian W, Liao B-Y, Chang AY-F, Zhang J. 2010. Maintenance of duplicate genes and their functional redundancy by reduced expression. *Trends in Genetics* [Internet] 26:425–430. Available from: http://dx.doi.org/10.1016/j.tig.2010.07.002

Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK. 2015. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* 43:e47.

Rodin SN, Parkhomchuk DV, Riggs AD. 2005. Epigenetic changes and repositioning determine the evolutionary fate of duplicated genes. *Biochemistry* 70:559–567.

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

Sandve SR, Rohlfs RV, Hvidsten TR. 2018. Subfunctionalization versus neofunctionalization after whole-genome duplication. *Nature Genetics* [Internet] 50:908–909. Available from: http://dx.doi.org/10.1038/s41588-018-0162-4

Shen F, Kidd JM. 2020. Rapid, Paralog-Sensitive CNV Analysis of 2457 Human Genomes Using QuicK-mer2. *Genes* [Internet] 11. Available from: http://dx.doi.org/10.3390/genes11020141

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

Starmer J, Magnuson T. 2016. Detecting broad domains and narrow peaks in ChIP-seq data with hiddenDomains. *BMC Bioinformatics* 17:144.

Steinberg KM, Antonacci F, Sudmant PH, Kidd JM, Campbell CD, Vives L, Malig M, Scheinfeldt L, Beggs W, Ibrahim M, et al. 2012. Structural diversity and African origin of the 17q21.31 inversion polymorphism. *Nature Genetics* [Internet] 44:872–880. Available from: http://dx.doi.org/10.1038/ng.2335

Suzuki IK, Gacquer D, Van Heurck R, Kumar D, Wojno M, Bilheu A, Herpoel A, Lambert N,

Cheron J, Polleux F, et al. 2018. Human-Specific NOTCH2NL Genes Expand Cortical Neurogenesis through Delta/Notch Regulation. *Cell* [Internet] 173:1370–1384.e16. Available from: http://dx.doi.org/10.1016/j.cell.2018.03.067

Tewhey R, Kotliar D, Park DS, Liu B, Winnicki S, Reilly SK, Andersen KG, Mikkelsen TS, Lander ES, Schaffner SF, et al. 2018. Direct Identification of Hundreds of Expression-Modulating Variants using a Multiplexed Reporter Assay. *Cell* 172:1132–1134.

Tvrdik P, Capecchi MR. 2006. Reversal of Hox1 gene subfunctionalization in the mouse. *Dev. Cell* 11:239–250.

Varadharajan S, Sandve SR, Gillard GB, Tørresen OK, Mulugeta TD, Hvidsten TR, Lien S, Vøllestad LA, Jentoft S, Nederbragt AJ, et al. 2018. The Grayling Genome Reveals Selection on Gene Expression Regulation after Whole-Genome Duplication. *Genome Biology and Evolution* [Internet] 10:2785–2800. Available from: http://dx.doi.org/10.1093/gbe/evy201

Wen X, Luca F, Pique-Regi R. 2015. Cross-Population Joint Analysis of eQTLs: Fine Mapping and Functional Annotation. *PLOS Genetics* [Internet] 11:e1005176. Available from: http://dx.doi.org/10.1371/journal.pgen.1005176

Zhang J. 2003. Evolution by gene duplication: an update. *Trends in Ecology & Evolution* [Internet] 18:292–298. Available from: http://dx.doi.org/10.1016/s0169-5347(03)00033-8

Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, Li W, et al. 2008. Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* 9:R137.