Title

Impact of structural variation on gene regulation in humans and non-human primates

Author

Shew, Colin James

Publication Date

2022

Peer reviewed|Thesis/dissertation

Impact of Structural Variation on Gene Regulation in Humans and Non-human Primates

By

COLIN SHEW

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Integrative Genetics and Genomics

in the

OFFICE OF GRADUATE STUDIES

of the

UNIVERSITY OF CALIFORNIA

DAVIS

Approved:

Megan Dennis, Chair

Siobhan Brady

Gerald Quon

Committee in Charge

2022

Impact of structural variation on gene regulation in humans and non-human primates

Colin Shew
University of California, Davis
2022

Abstract

Genetic structural variation has strong potential to contribute to human evolution and disease, as large deletions, duplications, and inversions alter more base pairs of the human genome than single-nucleotide variants. However, structural variants (SVs) have remained systematically understudied because they are challenging to identify and resolve with short-read sequencing data, upon which modern genomics depends heavily. Nevertheless, targeted efforts and recent technological advances have unveiled a dynamic landscape of structural variation within the human population and between primate species, with examples tied to both genomic disease and adaptive human traits. Of particular interest is a class of SVs called segmental duplications, large blocks of sequence duplicated at low copy number that are enriched on the great ape lineage. Human-specific segmental duplications (HSDs) comprise millions of base pairs of DNA unique to our species and, intriguingly, contain genes that shape cortical development. Beyond the genes themselves, SVs can impact gene regulation by reorganizing genes and regulatory elements throughout the genome, rewiring their interactions or altering their copy number altogether. Gene regulation is itself considered a major driver of evolutionary change, and this work explores the interplay of structural variation and gene regulation to assess how SVs contribute to species-specific features in humans and other primates. We used long read sequencing, optical mapping, and single-cell template strand sequencing to discover novel SVs

in chimpanzees and rhesus macaques, demonstrating that deletions and inversions in these species preserve chromatin architecture but also are associated with divergent gene regulation. In addition, we compared mRNA levels of HSD genes and their single chimpanzee orthologs and found that derived, human-specific genes are more likely to exhibit novel expression patterns compared to ancestral paralogs. To understand the molecular mechanisms underlying this divergence, we identified novel candidate *cis*-regulatory elements in HSDs and demonstrate that promoters and enhancers in these regions have functionally diverged in recent evolutionary time. Finally, we performed a massively parallel reporter assay to quantify the regulatory activity of thousands paralogous HSD variants, finding mostly differences of small effect, but also uncovering variants that may generate human-specific expression patterns. Taken together, these studies highlight the regulatory consequences of SVs; duplicated and rearranged regulatory elements give rise to novel expression patterns that have the potential to underlie the emergence of new traits.

Acknowledgments

This work—and my entire experience of graduate school—would not have been possible without the contributions and support of numerous mentors, colleagues, friends, and family. I feel lucky and deeply grateful for all their professional and personal support during this chapter of my life.

First of all, I want to express my extraordinary gratitude for my advisor, Dr. Megan Dennis. As a scientist, she has been an incredible wealth of knowledge about genomics and segmental duplications, an endless source of new ideas, and a tenacious problem-solver. I have been inspired by her ambition and passion for understanding the human genome and hope some of it has rubbed off. More importantly, Megan has been an outstanding mentor and cheerleader; she was always available to guide me and celebrate every milestone of my PhD and has been a gateway to many opportunities.

I would also like to thank the other scientific role models who have propelled me along in my career, especially Dr. William Pastor, my undergraduate mentor. Will taught me all the basics of science, and then some; the molecular biology, epigenetics, and research skills I learned from him have been instrumental in completing my PhD. I also cannot thank him enough for encouraging me to pursue all kinds of scientific interests, even when it took me out of the lab. Thanks also to the professors and TAs of the Marine Biology Quarter at UCLA (Dawn Bailey, Dr. Malcolm Gordon, Dr. Laura Jordan-Smith, Dr. Dovi Kacev, and Dr. Julia Notar) for broadening my scientific horizons, not to mention introducing me to R. I also thank Dr. Robert Wayne and the members of his lab, especially Dr. Annabel Beichman, Dr. Devaughn Fraser, and Dr. Jacqueline Robinson for everything they taught me about population genomics and bioinformatics. Finally, many thanks to my dissertation committee members, Dr. Siobhan Brady

and Dr. Gerald Quon, for their valuable feedback, commentary, and scientific insights on my projects.

I also want to acknowledge all of my friends and colleagues in and around Davis. Working in the Dennis Lab, I have been constantly surrounded by bright scientists who made my PhD both fun and productive. Thanks especially to Dr. Paulina Carmona-Mora, Dr. Alexandra Colón-Rodríguez, Aedric Lim, Dr. Sierra Nishizaki, Sean McGinty, Brittany Radke, Joseph Rosas, Aarthi Sekar, Daniela Soto, and José Uribe-Salazar. A special thank you to Gulhan Kaya, who is the heart of the lab and an experimentalist extraordinaire, without whom much of my thesis would not have been possible. Thank you to my undergraduate researchers, Dhrithi Jagannathan and Elizabeth Roberts, for their hours of dedication and resolve in the struggles of cloning, and for being my first mentees. Thanks also to our neighbors in the Carvajal-Carmona and Segal Labs, as well as my Davis/IGG community: Drs. Nicole Halmai, Julian Halmai, Maika Malig, Giovanni Hanna, Krithi Bala, Shannon Kieran, Logan Blair, and so many others.

I am fortunate to have found friendships of a lifetime here in Davis. To Aarthi, Alex, Dani, José, and Phil, thanks for the adventures, deep and silly conversations, advice, beers, and all the rest. You all have truly kept me sane. To my housemate César, thanks for putting up with me for six years, the delicious homemade salsa, and all the times you didn’t ask how my dissertation was going Finally, to my parents Ben and Kirsten, my brother Cameron, and grandma Jacqueline, I can’t thank you enough for the unwavering love, support, food, and car/ electronic/household repairs. I would not have made it without you.

Table of Contents

List of Figures

List of Tables

Abbreviations

bp	base pair
BP	breakpoint
cCRE	candidate cis-regulatory element
ChIP	chromatin immunoprecipitation
CN	copy number
CNV	copy number variant
CRE	cis-regulatory element
DE	differential expression
eQTL	expression quantitative trait locus
FDR	false discovery rate
HSD	human-specific segmental duplication
kbp	kilobase pair
LCL	lymphoblastoid cell line
LoF	loss-of-function
Mbp	megabase pair
MPRA	massively parallel reporter assay
NMD	nonsense-mediated decay
ONT	Oxford Nanopore Technologies
PacBio	Pacific Biosciences
pLI	probability of loss-of-function intolerance score
PSV	paralogous sequence variant
RPKM	reads per kilobase per million mapped reads
SD	segmental duplication
SV	structural variant
TPM	transcripts per million
TSS	transcription start site
UTR	untranslated region

Chapter 1:
Introduction

1.1 BACKGROUND

1.1.1 Structural variation is a driver of human evolution and disease

A longstanding goal of human genomics is to understand the origin of our species. The characteristic features of humans—such as skeletal development enabling bipedalism, dietary adaptations to novel sources of nutrition, and the ability to use symbolic language—ultimately have an underlying genetic basis, which can be identified through comparison to other great apes. One of our two closest living relatives, the chimpanzee, differs from humans at about 1% of the bases within alignable portions of the two genomes (Chimpanzee Sequencing and Analysis Consortium 2005). While this comprises millions of base-pairs, a relatively small number of truly human-specific substitutions exist and fewer are known to be functional (Prüfer et al. 2014; Pääbo 2014). The total number of variable bases between species is higher when considering structural variation (Kehrer-Sawatzki and Cooper 2007; Varki and Altheide 2005), with structural variants (SVs) estimated to impact more than twice as many bases as single-nucleotide variants (SNVs) between humans and chimpanzees (Cheng et al. 2005). SVs are operationally defined as deletions, duplications, inversions, and translocations larger than 50 base pairs, often thousands or even millions of base pairs, giving them high potential to impact gene function (Figure 1.1).

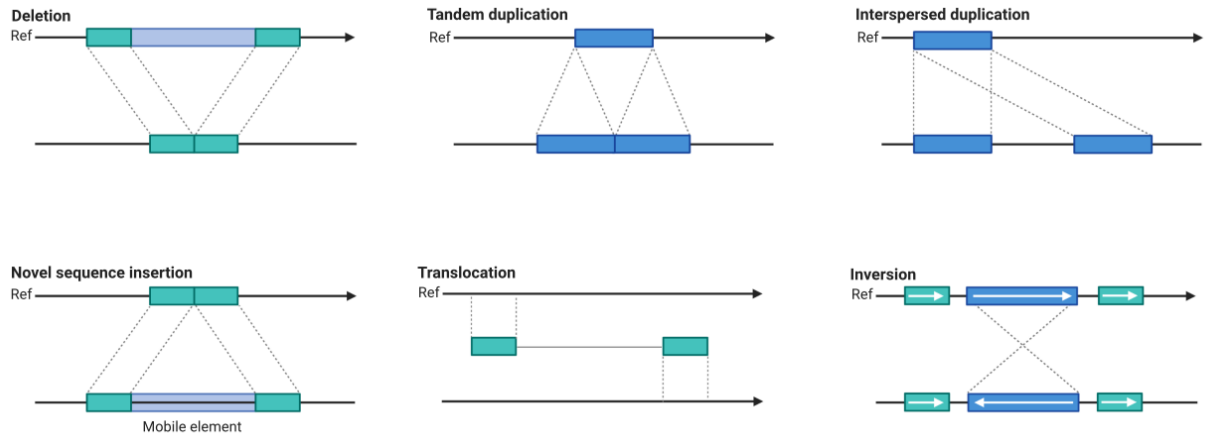


Figure 1.1: Examples of genomic structural variation. Categories of SVs are illustrated relative to a reference sequence (Ref). SVs that alter the copy number (CN) of a genomic segment (i.e., CNVs) include deletions and duplications; the largest, most similar duplications are termed segmental duplications (SDs) or low-copy repeats. CN-neutral SVs include insertions, translocations, and inversions. SVs frequently occur as complex rearrangements involving multiple types of variants. Figure is adapted from Alkan, Coe, and Eichler (2011) via "Genome Structural Variations" by BioRender.com (2022).

However, SVs are difficult to comprehensively identify and have consequently been understudied. Foundational genetic work in *Drosophila* linked developmental phenotypes to cytologically visible intrachromosomal duplications on the *Bar* locus (Bridges 1936), but SVs below the resolution of microscopy have proven elusive, even in the era of DNA sequencing and high-quality reference genomes. Genome assemblies themselves are typically fraught with collapsed and false duplications, as repeats longer than the sequencing read length cannot be traversed and unambiguously resolved (Vollger et al. 2019; Rhie et al. 2021). Over 90% of the human reference genome was painstakingly constructed by progressive assembly of bacterial artificial chromosome (BAC) libraries (Lander et al. 2001), but assembly gaps persisted until recent technological advances made it possible to span these regions (Nurk et al. 2022). The problematic loci consist of segmental duplications (SDs)—large blocks of sequence (typically 10s to 100s of kilobase pairs [kbp]) shared at high identity (>90%) at multiple locations throughout the genome—and the highly repetitive centromeres and ribosomal DNA arrays.

These regions are also prone to recurrent rearrangement (Stankiewicz and Lupski 2002; Shaw and Lupski 2004; Miga 2019) and continue to present a technical obstacle to the resolution of SVs in diploid genomes (Wang et al. 2022; Nurk et al. 2022). Finally, even with a near-perfect reference, the identification of SVs with prevalent and affordable short-read data remains a challenge. The short span of these reads means that few can be uniquely mapped to repetitive regions that often mediate rearrangement, and even outside these regions, accurate discovery requires high read depth (Alkan, Coe, and Eichler 2011).

Nevertheless, targeted effort has begun to characterize the SV landscape of primate genomes, human in particular, and pinpoint variants driving disease development and phenotypic divergence. Comparison of vertebrate reference genomes has identified lineage-specific deletions of highly conserved sequence, and functionally characterized human deletions include a likely forebrain enhancer of cell cycle-arresting *GADD45G*, an androgen receptor enhancer contributing to the loss of penile spines, and a hindlimb and digit enhancer of the bone morphogenetic protein *GDF6* (McLean et al. 2011; Reno et al. 2013; Indjeian et al. 2016). Array comparative genomic hybridization (aCGH) and depth-based analysis of short reads have also proven effective for identifying copy number variation, uncovering human-specific expansions of loci associated with brain development and genomic disease (Sudmant et al. 2013, 2010; Iskow et al. 2012). Many of these duplications, as well as deletions and inversions, are mediated by non-allelic homologous recombination of SD loci and are accordingly highly mutable within and between species (Marques-Bonet, Girirajan, and Eichler 2009). In humans, SDs colocalize with “hotspots” of genomic rearrangement and generate microdeletions and microduplications implicated in schizophrenia, epilepsy, intellectual disability, and autism (Itsara et al. 2009; Antonacci et al. 2014; Stefansson et al. 2008; Nuttle et al. 2016; Brunetti-Pierri et al. 2008; Sharp et al. 2008; Cuscó et al. 2008). These observations have led to the hypothesis that human-specific SDs (HSDs) have sensitized humans to unique neurodevelopmental pathologies while simultaneously enabling novel cognitive features.

Targeted studies of HSD loci and the genes within them have identified dozens of duplicate genes fixed in humans and absent from other ape species (Dennis et al. 2017) (Figure 1.2); strikingly, a handful of these genes (*SRGAP2C*, *ARHGAP11B*, and *NOTCH2NL*) contribute to corticogenesis and neuronal migration during early brain development (Dennis et al. 2012; Charrier et al. 2012; Florio et al. 2015; Heide et al. 2020; "Website," n.d.; Fiddes et al. 2018; Suzuki et al. 2018). Others appear to play a role in autoimmune response (*DUSP22* (J.-P. Li et al. 2014) and *NCF1* (Hultqvist et al. 2004; Zhao et al. 2017)), cause spinal muscular atrophy (*SMN2* (Kashima and Manley 2003)), and affect cold sensitivity (*TCAF1* and *TCAF2* (Kashima and Manley 2003; Gkika et al. 2015)). Many yet remain to be functionally investigated, but mounting evidence indicates that structural variation, and especially duplicated genes, are major drivers of human evolution and disease.

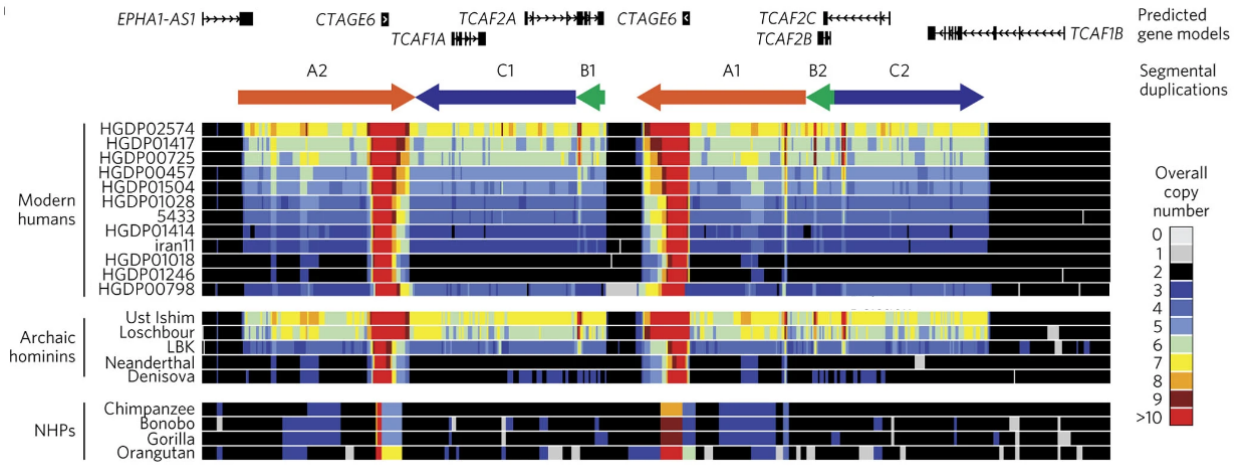


Figure 1.2: Example human-specific segmental duplication (HSD). The *TCAF* locus on chr7q35 contains multiple HSDs (colored and oriented with block arrows A, B, and C). CN is indicated for representative modern human, archaic human, and non-human primate genomes. Figure is adapted from Dennis *et al.* (2017).

1.1.2 Structural variation alters gene regulation

As suggested by the above examples, structural variation impacts not only genes, but also their regulatory elements, and likely to a greater extent. The vast majority (>98%) of the human

genome is noncoding, and changes to regulatory regions are thought to be better tolerated than changes to protein coding sequences; enhancers can contribute to cell-type specificity, so regulatory mutations tend to be modular, impacting the quantity, location, or developmental time of gene expression while leaving the genes themselves intact (Carroll 2000; Arnone and Davidson 1997). Accordingly, gene regulation is a major contributor to variation within and between species (Wray et al. 2003; Fay and Wittkopp 2008; Fraser 2013). This was suspected even before the genomic era, as comparison of human and chimpanzee sequences suggested that coding differences were insufficient to explain the phenotypic divergence between the species, and that most changes were likely regulatory (King and Wilson 1975). Given that SVs constitute a major component of intra- and inter-species variation, they may underlie much of this regulatory divergence, and indeed contribute to regulatory differences within humans and between primate species (Iskow et al. 2012; McLean et al. 2011; Stranger et al. 2007). At the same time, proper development relies on finely tuned spatiotemporal expression patterns, and many disease etiologies result from aberrant *cis*-regulatory activity of promoters and enhancers. Notably, many are also caused by structural rearrangements (Kleinjan and Coutinho 2009).

Compared to SNVs, SVs are more likely to result in regulatory changes, since their large size allows them to alter the copy number and genomic context of genes and regulatory elements. Intuitively, copy number variation (CNV) of coding sequence can impact gene dosage. Single- or multi-gene impacts of CNV on human health are numerous (Henrichsen, Chaignat, and Reymond 2009), including association of beta defensin copy number with higher mRNA levels and psoriasis susceptibility (Hollox et al. 2003; Hollox et al. 2008) and the >1 Mbp deletions at chromosome 7q11.23 underlying Williams-Beuren Syndrome (Bayés et al. 2003). However, like most of the genome, a majority of CNVs are noncoding and can cause indirect expression changes by deleting, duplicating, or rearranging regulatory elements. Genome-wide, analysis of 210 lymphoblastoid cell lines (LCLs) from the International HapMap Project found that CNVs explained 17.7% of the variation in mRNA levels (Stranger et al. 2007). More

recently, analysis of expression in diverse tissues in 613 individuals from the GTEx project predicts that common SVs are causal at 2.7% of eQTLs, which represents a 10.5-fold enrichment compared to SNVs, considering their relative abundance in the genome (Scott, Chiang, and Hall 2021). Beyond simply altering mRNA levels, individual regulatory SVs can have marked phenotypic effects. For example, duplication or deletion of enhancers upstream of *SOX9* causes XX and XY sex reversal, respectively,, and a human-specific loss of a conserved *GDF6* enhancer results in shortened hindlimb digits in mouse models (Croft et al., 2018; Indjeian et al., 2016).

The molecular mechanisms of SV-mediated non-coding changes have been best studied in the context of promoter-enhancer “rewiring”, in which a deletion, duplication, or inversion alters endogenous regulatory contacts, leading to aberrant gene expression as enhancers interact with non-target promoters. Functional dissection of the *WNT6*/*IHH*/*EPHA4*/*PAX2* locus in humans and mice demonstrated that rearrangements relative to insulating elements allowed *EPHA4* enhancers to interact with other promoters in the locus, driving ectopic expression in the limb buds and causing digit malformation phenotypes (Lupiáñez et al. 2015). This mechanism has been implicated in other disease contexts, including “enhancer hijacking” in cancer (Franke et al. 2022; M. Yang et al. 2020; Northcott et al. 2014). It is likely that altered chromatin conformation also contribute to non-pathogenic human variation in gene expression; for instance, different haplotypes of a common ~900-kb inversion at chromosome 17q21.31 exhibiting signatures of positive selection in Europeans are associated with up- and down-regulation of multiple genes (Stefansson et al. 2005; de Jong et al. 2012). Between species, comparison of long read-based great ape assemblies identified hundreds of human-specific SVs putatively altering gene expression, though these have not been functionally investigated (Kronenberg et al. 2018). In all, SVs are inextricably linked with the gene regulatory landscape.

Given the established roles of duplicated genes in human evolution and disease, it is

also critical to understand their regulation. Beyond simple copy number differences, HSD genes are often truncated, inverted, and translocated hundreds of kbp, and despite their high sequence identity (>99%), exhibit paralog-specific expression patterns across diverse post-mortem tissues (Dennis et al. 2017) (Figure 1.3A). Long read isoform sequencing of HSD genes in adult brain found that almost half of the 20 gene families examined contained novel features, including exapted or truncated exons, new transcription start or end sites, or altered splicing (Dougherty et al. 2018). Thus, the transcriptional structure of HSD genes has changed in short evolutionary time, though the contribution of duplicated and adjacent unique *cis*-regulatory elements (CREs) has not yet been examined. Spatiotemporal control of paralogous gene expression may be as critical as any novel biochemical functions to human-specific traits. Again in the context of the brain, 15 human-specific genes are known to be preferentially expressed in cortical progenitors (Florio et al. 2018). For instance, *ARHGAP11B* expression is largely restricted to the germinal zone, the location of cortical neuron progenitors, while *ARHGAP11A* is expressed at higher levels throughout the germinal zone and overlying cortical plate (Figure 1.3B–C). The molecular mechanisms underlying this and other cell type-specific regulatory activity are not known, but would illuminate the functions and evolutionary trajectories of these HSD genes. In addition, regulatory divergence may facilitate duplicate retention in the genome to begin with, allowing paralogs to become functionally distinct and experience negative selection rather than becoming pseudogenes (Kronenberg et al. 2018; Force et al. 1999). This can be achieved through the accumulation of independent loss-of-function mutations to the daughter paralogs, partitioning or subfunctionalizing ancestral expression, while additional mutations can introduce novel or neofunctionalized activity (Prince and Pickett 2002). HSD genes present a unique opportunity to learn about the regulatory divergence of evolutionarily recent, single-locus duplications, and understanding these loci in turn may nominate additional candidate genes contributing to human-specific traits in a tissue-specific context.

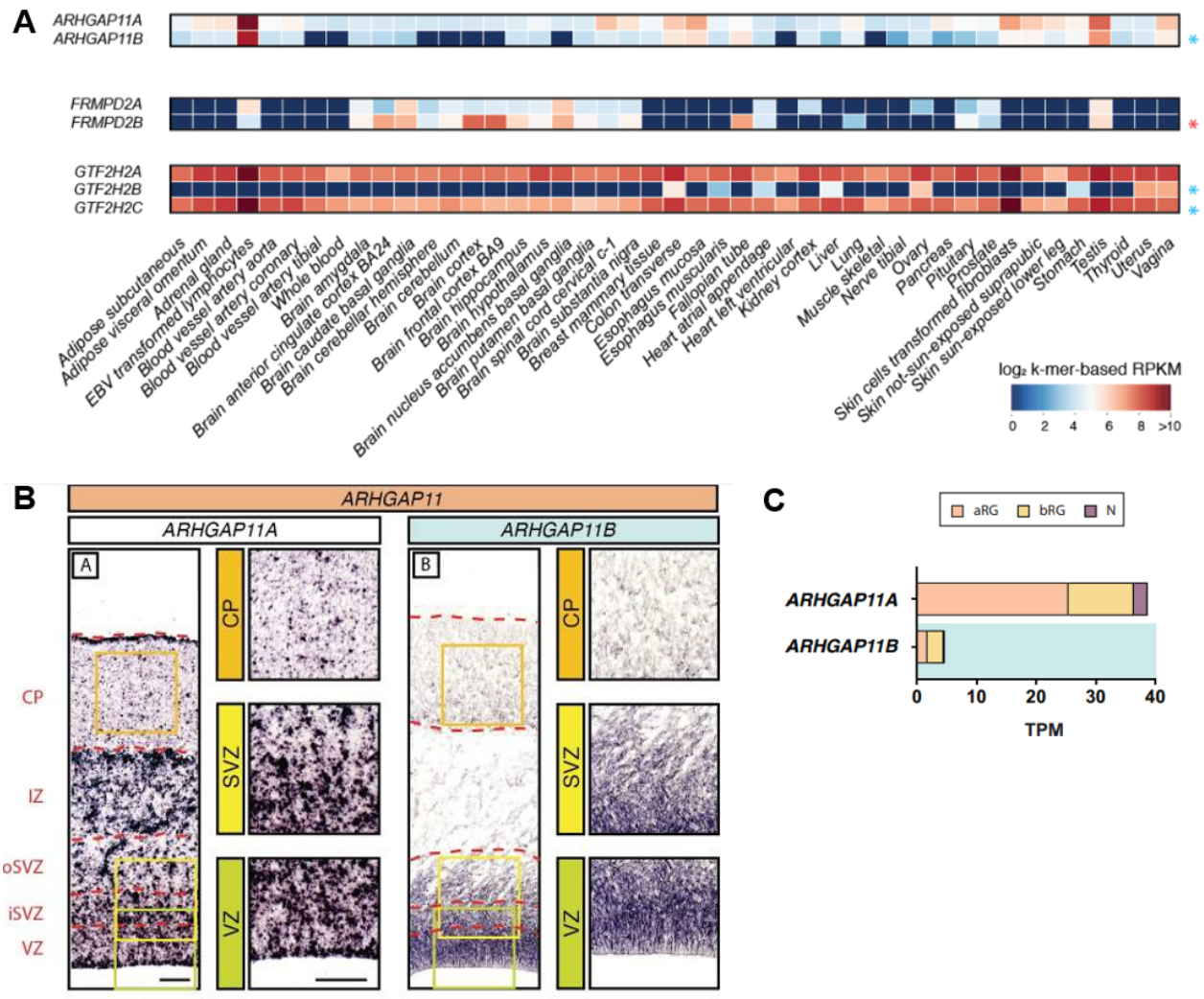


Figure 1.3: HSD genes exhibit paralog-specific expression patterns. (A) Quantification of mRNA levels of three selected HSD gene families (*ARHGAP11*, *FRMPD2*, *GTF2H2*). Asterisks indicate globally higher (red) or lower (blue) expression of derived paralogs relative to the ancestral paralog (top row). Panel is adapted from Dennis *et al.* (2017). (B) mRNA *in situ* hybridization with paralog-specific probes to distinguish *ARHGAP11A* and *ARHGAP11B* expression in a coronal section of human fetal cortex (13 weeks post-conception). Cortical layers are abbreviated as follows: ventricular zone (VZ), inner subventricular zone (iSVZ), outer subventricular zone (oSVZ), intermediate zone (IZ), and cortical plate (CP). (C) Quantification of mRNA levels for *ARHGAP11A* and *ARHGAP11B* in neurons (N) and the cortical progenitors apical radial glia (aRG) and basal radial glia (bRG). Panels B and C adapted from Florio *et al.* (2018).

1.2 GOALS AND CONTENTS

This work aims to characterize the regulatory landscape of structural variation in humans and nonhuman primates. While SVs have been difficult to comprehensively identify and resolve using genomic methods, particularly short-read sequencing, recent advances have made them more tractable to study. First, improvements to the quality, availability, and throughput of “third-generation” long-read technologies, namely Pacific Biosciences (PacBio) and Oxford Nanopore (ONT) sequencing, has allowed larger repeats and complex rearrangements to be resolved, in some cases for the first time (Nurk et al. 2022; Hsieh et al. 2021). Previous efforts relied on labor-intensive methods, such as targeted assembly of BAC libraries, which is not scalable to large numbers of genomes. Second, these technological advances have galvanized the development of novel assays and bioinformatic methods for data generation, genome assembly, and alignment. Finally, the increasing number of human and non-human primate resources, including population-scale genomic and epigenomic databases are enabling unprecedented new biological discoveries (Audano et al. 2019; Kronenberg et al. 2018) .

In this work, we leveraged this wealth of information to address open questions about structural variation and gene regulation: How has gene regulation diverged at SV loci between humans and non-human primates, and what impact does this have on function? In HSDs, we also considered how expression patterns have changed in recent evolutionary time and investigated potential underlying molecular mechanisms. Anticipating a major contribution of *cis*-regulatory changes to duplicated and structurally rearranged loci, we designed experiments to test two non-mutually exclusive hypotheses: (1) inversions and translocations of rearranged sequence can alter their regulatory environment, which is visible as changes to chromatin architecture of promoters and interacting CREs; (2) sequence divergence (such as SNVs) following duplications can cause gains and losses of activity of paralogous CREs (Figure 1.4)

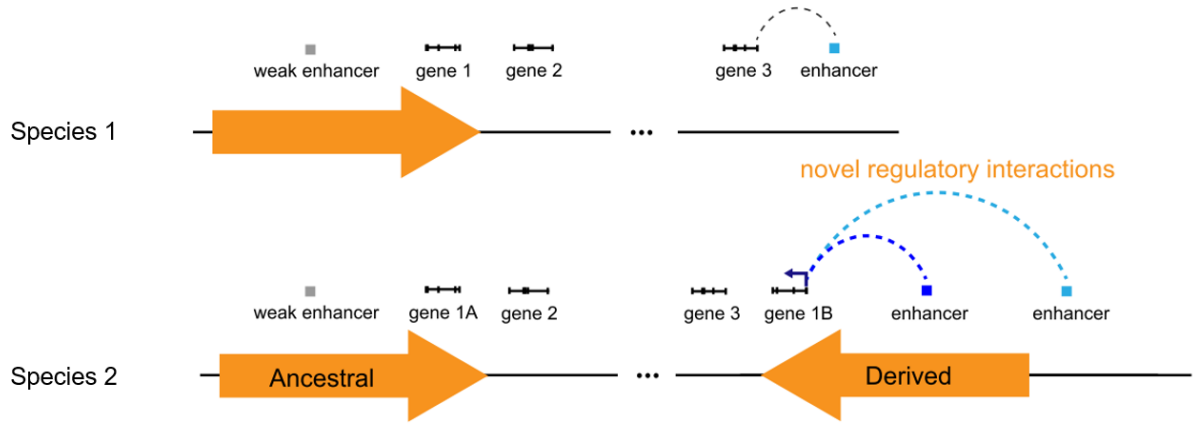


Figure 1.4: Potential impacts of SVs on gene regulation. Relative to the structural configuration of a reference, SVs in another genome or species may induce regulatory changes. Inversions and translocations can rewire promoter-enhancer contacts across SV breakpoints. In the case of duplications, sequence changes between paralogous loci can also result in changes to the activity of *cis*-regulatory elements (CREs) such as promoters and enhancers. This example illustrates a gain of expression of a derived gene in species 2, though gains or losses at the ancestral locus, or non-duplicated loci, are also possible.

In Chapters 2 and 3, we discovered novel SVs distinguishing non-human primate genomes from those of humans. Chapter 2 uses long-read sequencing and optical mapping in chimpanzee cell lines, while Chapter 3 focuses on inversions discovered in a rhesus macaque cell line using single-cell template strand sequencing (Strand-seq) (Sanders et al. 2017),. We identified differentially expressed genes between species in multiple cell types and compared gene expression and chromatin architecture at SV loci. In broad agreement with the existing literature, both studies found a significant tendency for SVs to preserve the structure of topologically associating domains (TADs) and avoid altering protein-coding genes, suggesting these types of SVs are strongly deleterious and experience negative selection. We showed this for the first time in inversions, likely due to the sensitivity to inversions of the technologies employed. We also found an enrichment for differentially expressed genes near SV breakpoints, indicating that species-specific SVs may disrupt gene regulation. Finally, we used epigenetic

data, signatures of selection, and human disease associations to speculate on SVs that may contribute to species-specific traits.

Chapters 4 and 5 focus on HSDs, quantifying differential gene expression at a subset of human paralogous and chimpanzee orthologous loci with high-quality assemblies (75 human genes in 30 gene families). In Chapter 4, we identified a tendency for derived genes to show reduced cross-tissue correlation and total expression compared to chimpanzee orthologs and considered molecular mechanisms that may drive these recent evolutionary changes. We implemented a multiple alignment and read allocation pipeline to generate a novel set of candidate CREs in HSDs, and showed with luciferase reporters that duplicated CREs can be differentially active. Additionally, we found evidence for paralog-specific contributions of adjacent non-duplicated loci. In Chapter 5, we explored HSD *cis*-regulatory divergence in high throughput, performing a massively parallel reporter assay (MPRA) to compare the activity of thousands of duplicated promoters and candidate enhancers. We were unable to successfully predict paralogous expression divergence using the measured differences in regulatory activity between duplicated regions alone but identified additional paralog-specific regulatory changes for further characterization. Future work is needed to characterize the chromatin structure of HSDs and integrate these data with the measured activity of CREs to create a more complete picture of regulatory divergence following genomic duplication. In all, we nominated regulatory changes that may contribute to human-specific traits, such as a putatively neofunctionalized gene (*ARHGEF35*), likely dosage increases in autoimmune-protective genes (*DUSP22* and *NCF1*), and paralog-specific activity loss of an intronic enhancer in *SRGAP2C*.

Chapter 2:
Identification of structural variation in chimpanzees using optical mapping and nanopore sequencing

2.1 ABSTRACT

Recent efforts to comprehensively characterize great ape genetic diversity using short-read sequencing and single-nucleotide variants have led to important discoveries related to selection within species, demographic history, and lineage-specific traits. Structural variants (SVs), including deletions and inversions, comprise a larger proportion of genetic differences between and within species, making them an important yet understudied source of trait divergence. Here, we used a combination of long-read and -range sequencing approaches to characterize the structural variant landscape of two additional *Pan troglodytes verus* individuals, one of whom carries 12% admixture from Pan troglodytes troglodytes. We performed optical mapping of both individuals followed by nanopore sequencing of one individual. Filtering for larger variants (>10 kbp) and combined with genotyping of SVs using short-read data from the Great Ape Genome Project, we identified 425 deletions and 59 inversions, of which 88 and 36, respectively, were novel. Compared with gene expression in humans, we found a significant enrichment of chimpanzee genes with differential expression in lymphoblastoid cell lines and induced pluripotent stem cells, both within deletions and near inversion breakpoints. We examined chromatin-conformation maps from human and chimpanzee using these same cell types and observed alterations in genomic interactions at SV breakpoints. Finally, we focused on 56 genes impacted by SVs in >90% of chimpanzees and absent in humans and gorillas, which may contribute to chimpanzee-specific features. Sequencing a greater set of individuals from diverse

subspecies will be critical to establish the complete landscape of genetic variation in chimpanzees.

2.2 CONTRIBUTIONS

This chapter is adapted with minimal modification from the following published work:

> Daniela C Soto, Colin Shew, Mira Mastoras, Joshua M Schmidt, Ruta Sahasrabudhe, Gulhan Kaya, Aida M Andrés, Megan Y Dennis. 2020. “Identification of Structural Variation in Chimpanzees Using Optical Mapping and Nanopore Sequencing.” *Genes* 11(3): 276. https://doi.org/10.3390/genes11030276.

D.C.S and C.S. co-authored the study. D.C.S., C.S., and M.Y.D. conceived the study. C.S., G.K., and R.S. prepared samples and generated sequencing data. D.C.S., C.S., M.M., J.M.S., and M.Y.D. analyzed data. D.C.S., C.S., M.M., J.M.S., R.S., G.K., A.M.A., and M.Y.D. wrote and edited the manuscript. All authors have read and agreed to the submitted version of the manuscript.

2.3 INTRODUCTION

Great apes have considerable phenotypic diversity despite being closely related species. For humans and chimpanzees, with only ~5 to 9 million years of independent evolution (Patterson et al. 2006; Langergraber et al. 2012), significant effort has gone into understanding the underlying genetic and molecular differences contributing to species differences, often with the primary focus on human-unique features (O’Bleness et al. 2012). Direct comparison of protein-coding genes has identified exciting candidates, but these only account for a minor proportion of species differences (Varki and Altheide 2005). Recent analysis of Illumina short-read sequencing has allowed identification and genotyping of single-nucleotide variants (SNVs) at the genome scale, which have been used to address questions related to the demographic history and genetic adaptations of each species, and lineage-specific traits (Prado-Martinez et

al. 2013). Further, transcriptome and epigenome comparisons of immortalized cell lines and tissues have revealed many thousands of individual genes and putative *cis*-acting regulatory elements that contribute to species differences in gene regulation (Gallego Romero et al. 2015; Khan et al. 2013; McLean et al. 2011; Prescott et al. 2015; Pollen et al. 2019; Brawand et al. 2011; Eres et al. 2019; Zhou et al. 2014), though often with varied results and reproducibiilty across studies.

Since the publication of the chimpanzee genome (Chimpanzee Sequencing and Analysis Consortium 2005), comparison with the human reference genome showed that structural variants (SVs), or genomic rearrangements such as inversions and copy-number variants (deletions and duplications), comprise a greater proportion of genetic differences than SNVs (Rogers and Gibbs 2014). Though important, SVs are difficult to discover and genotype using traditional short-read Sanger and Illumina data. As such, genome-wide analyses of SVs have leveraged alternative approaches, including fosmid-end mapping (Newman et al. 2005), array comparative genomic hybridization (CGH) (Gokcumen et al. 2013; Wilson et al. 2006; D. P. Locke et al. 2003), digital array CGH using whole-genome shotgun sequencing of Sanger (Marques-Bonet et al. 2009) and Illumina (Sudmant et al. 2013), and comparisons with improved genome assemblies (Catacchio et al. 2018; Feuk et al. 2005; Kuderna et al. 2017; Kronenberg et al. 2018). Most recently, the advent of long-read sequencing technologies, capable of completely traversing variant breakpoints, has significantly facilitated discovery of novel SVs (Mahmoud et al. 2019). To date, only one study has performed long-read sequencing of a chimpanzee; the most recent improvement to the chimpanzee reference genome (panTro6) used hybrid long-read (PacBio) and long-range sequencing approaches (Bionanogenomics (BNG) and HiC) of one individual, Clint, a male representing the subspecies *Pan troglodytes verus*, significantly increasing the number of known SVs (Kronenberg et al. 2018).

Recent comparisons of short- and long-read sequencing technologies using benchmark human genomic datasets revealed that multiple genomes (Audano et al. 2019) and

combinatorial platforms are required for comprehensive SV discovery (Chaisson et al. 2019); therefore, we performed long-range BNG optical mapping and Oxford Nanopore Technologies (ONT) long-read sequencing of additional chimpanzee individuals. These new datasets have allowed us to more comprehensively assess deletions and inversions in the chimpanzee genome. When compared with published whole-genome screens using orthogonal approaches, our approach validated existing variants and discovered many new variants. Knowing that SVs often alter gene functions and regulation (Spielmann, Lupiáñez, and Mundlos 2018), we characterized the association of our discovered SVs on differences in gene regulation and chromatin organization between human and chimpanzee, identifying a number of events that likely contribute to chimpanzee-specific differences.

2.4 RESULTS

2.4.1 Large-Scale SV Discovery and Genotyping in Chimpanzee

To date, one western chimpanzee individual (Clint) comprising the reference genome (PanTro6) has been subject to hybrid long-read sequencing for genome assembly and SV discovery (Kronenberg et al. 2018). We sought to expand SV discovery via long-read sequencing to two additional chimpanzee individuals (AG18359 and S003641) for which renewable LCLs and functional genomic information, including RNA-Seq and ChIP-Seq data (Khan et al. 2013; McVicker et al. 2013; Zhou et al. 2014), are available. To begin, we performed Illumina short-read sequencing (~30x coverage) of both individuals to confirm ancestry via SNV detection followed by comparisons of population-specific genetic markers and principal component analysis with chimpanzees from the GAGP (Prado-Martinez et al. 2013) (Figure S2.1). From this, we determined AG18359 to be a female western chimpanzee (*Pan troglodytes verus*) and S003641 to be a male western chimpanzee with some central chimpanzee ancestry

(*Pan troglodytes verus* × *Pan troglodytes troglodytes*). Notably, ~15% of the ancestry of this individual is assigned to the central-chimpanzee population, similar to one individual (Donald) that was sequenced as part of the GAGP.

To discover potentially novel chimpanzee SVs, we assayed AG18359 gDNA using ONT PromethION (29x) and BNG optical mapping (116x). To compare SV discovery of two individuals on the same platform, we also subjected S003641 to BNG optical mapping (70x). As it is the most accurate and well-annotated primate assembly, we mapped our sequence data to the human reference genome (GRCh38). We excluded SDs and insertions from our analysis of SVs due to challenges in their discovery and validation (Alkan, Coe, and Eichler 2011). Focusing exclusively on deletions and inversions, we discovered 49,579 deletions and 560 inversions using ONT and 4,790 deletions and 280 inversions using BNG from AG18359. Similarly, we identified 5,407 deletions and 207 inversions using BNG from S003641. For comparison purposes, we also mapped the AG18359 ONT sequence data to the most recent chimpanzee reference genome (PanTro6) and discovered fewer events (7,895 deletions and 142 inversions) suggesting that a significant proportion of SVs identified via mapping to the human reference represented species differences.

As the primary goal of our study was to identify species differences, we moved forward with SVs identified using the human reference genome. We next compared SV discovery across our two platforms. Although ONT had higher sensitivity to discover smaller variants, down to 50 bp, there was a higher chance of detecting false positives and errors at this resolution (Figure S2.2A). To properly compare across technologies, we filtered for large SVs (≥10 kbp) and compared similarities by consolidating variants with more than 50% reciprocal overlap. We found a comparable number of deletions in our three call set (586, 586, and 666 events in AG18359 ONT, AG18359 BNG, and S003641 BNG, respectively) with 138 deletions found by all three call sets (Figure S2.2B). Out of the 586 deletions found in the AG18359 ONT call set, 381 were uniquely discovered using this technology, while BNG contributed another 553 deletions,

out of which 307 (55.5%) had support from both individuals. As such, deletion call sets from the same technology exhibited a greater overlap than comparing calls from different technologies of the same individual. We also found a comparable number of inversions across all three call sets (243, 269, and 207 variants in AG18359 ONT, AG18359 BNG, and S003641 BNG, respectively) (Figure S2.2B), of which 34 variants were shared among them all. Again, the most overlap for inversions was identified between different individuals assayed using the same BNG technology, representing 80 shared out of the total 274 unique variants.

In order to narrow in on a higher-confidence set of SVs, we subsequently performed genotyping of this discovery set using short-read Illumina data from GAGP (>20-fold coverage) of all four chimpanzee subspecies (*n* = 25) using SVTyper (Chiang et al. 2015). We also compared our discovered SVs with previously-reported datasets from three recent whole-genome SV screens of chimpanzees (Kronenberg et al. 2018; Catacchio et al. 2018; Sudmant et al. 2013), each using diverse genomic methods for discovery. From this, we identified 425 deletions and 59 inversions that had support from short-read genotyping and/or intersecting with a previously-discovered SV (Supplementary Tables 7 and 8). In all, our discovery approach using ONT and BNG data achieved 88 novel deletions and 36 novel inversions when compared with the most recent genome-assembly alignment (Kronenberg et al. 2018; Catacchio et al. 2018) and read-depth (Sudmant et al. 2013) approaches (Figure 2.1A–B).

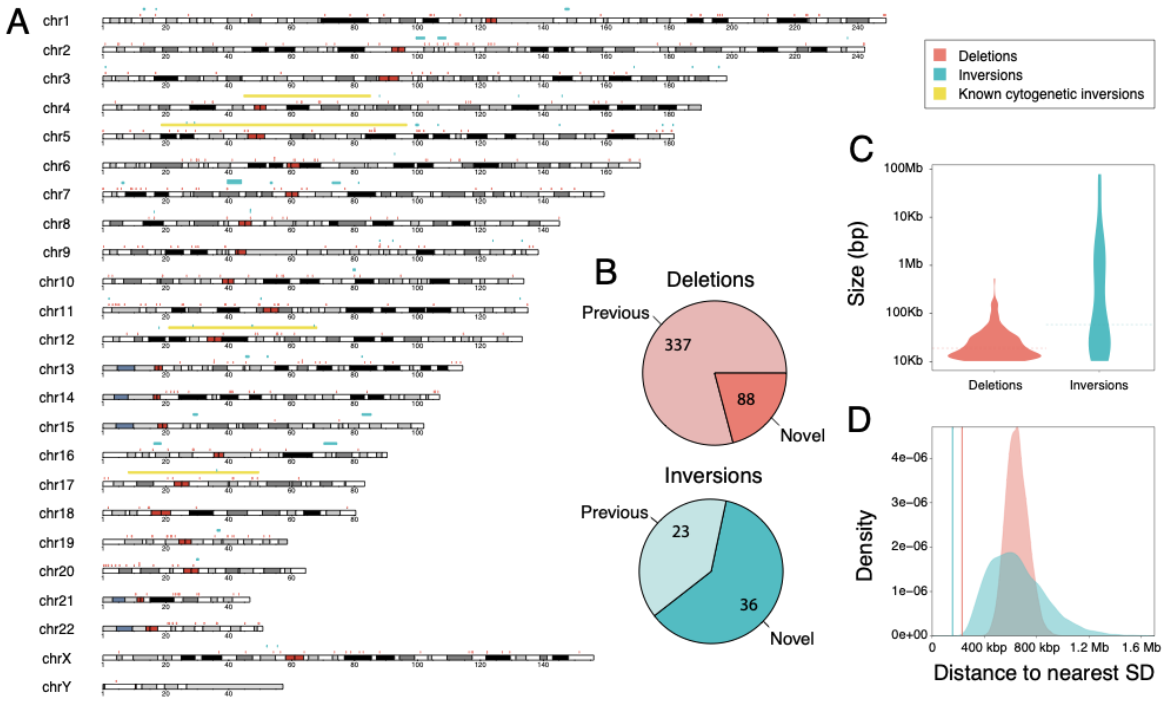


Figure 2.1: Genomic features of identified structural variants. (A) Deletions (red), inversions (cyan), and large-scale cytogenetic inversions (yellow) are interspersed across all 24 human orthologous chromosomes, depicted as ideograms. **(B)** Novel variants in our dataset defined as lacking 50% reciprocal overlap with previous reported variants in great apes. **(C)** Size distribution of deletions (red) and inversions (cyan). Median size is depicted as dashed lines. **(D)** Observed average distance of deletions (red line) and inversions (cyan line) to SDs, compared to randomly sampled regions across the genome of the same size of deletions (red distribution) and inversion (green distribution). We observed an enrichment of SV breakpoints residing near SDs (empirical p-value = 1 × 10−4).

2.4.2. Genomic features of identified SVs

Examining genomic features of our high-confidence set of chimpanzee SVs, we found that deletion sizes ranged between 10 kbp (our minimum threshold) up to ~526 kbp (31 kbp mean; 18.5 kbp median) (Figure 2.1C) and inversions ranged in size between 10 kbp and 78 Mbp (4.1 Mbp mean; 57.3 kbp median), including four of seven known chimpanzee pericentric inversions identified only with ONT (*n* = 2) or with both technologies (*n* = 2) (Nickerson and Nelson 1998; Kehrer-Sawatzki, Sandig, et al. 2005; Kehrer-Sawatzki et al. 2005; Goidts et al. 2005; Shimada et al. 2005; Szamalek et al. 2006; Kehrer-Sawatzki, Szamalek, et al. 2005). The majority of

novel inversions identified in our study tended to be smaller (57 kbp mean length), perhaps influenced by strict size cutoffs (>100 kbp) used in previous studies (Catacchio et al. 2018). The distribution of SVs across the human genome (Figure 2.1A, Figure S2.3) was relatively uniform for deletions, which were found on all 24 chromosomes. The greatest number of events were identified in chromosome 2 (n = 34); however, when normalizing by the total number of bases, chromosomes 19 (0.34 deletions per Mbp) and 21 (0.32 deletions per Mbp) exhibited the highest number of deletions (Figure S2.3). Inversions, on the other hand, were found on 19 chromosomes, with chromosomes 5 exhibiting the greatest number of variants (n = 8), and chromosomes 5, 7 and 12 displaying the greatest number of inversions per chromosome size (0.04 inversions per Mbp). Further, we found that breakpoints of SVs of both deletions and inversions were non-randomly distributed across the human genome near SDs (Figure 2.1D, empirical p-value = 1×10^{-4}), similar to previously reported results for distribution of SDs in primate genomes (Dennis et al. 2017; Sudmant et al. 2013; Cheng et al. 2005; Marques-Bonet et al. 2009). This observed clustering may be accounted for by SD-mediated deletions and inversions that can be created via non-allelic homologous recombination (Carvalho and Lupski 2016).

2.4.3 Genes impacted by SVs

To evaluate the functional impact of our high-confidence set of SVs, we retrieved all annotated transcribed features within deletions (±2.5 kbp) and at inversion breakpoints (±50 kbp). Deletions overlapped with 592 genes, out of which 162 were protein-coding genes (Figure 2.2A). To further refine the impact of the SVs and gene function, we focused on protein-coding genes and used Ensembl Variant Effect Predictor (VEP) to predict functional impact. VEP annotated 80 protein-coding genes as highly impacted by deletions (i.e., feature ablation or truncation), out of which 54 have been previously classified as loss of function (LoF) tolerant (probability of loss of function intolerance score [pLI] ≤ 0.1) by the Exome Aggregation

Consortium (Samocha et al. 2014; Lek et al. 2016) (Figure 2.2B). Also, three genes (*ATXN2L*, *SH2B1*, and *IL27*), which all reside within the same ~500 kbp "deletion" mapped to human chromosome 16p11.2, were classified as LoF intolerant (pLI ≥ 0.9). A search through the chimpanzee reference (panTro6) found *ATXN21* and *SH2B1* residing on an uncharacterized chimpanzee chromosome Un_NW_019937196v1, suggesting that these genes have been translocated to a new genomic locus. This is likely the case for other genes with predicted high-variant effect and LoF intolerance. Focusing on inversion, we found breakpoints overlapping with 342 transcribed elements of which 64 genes were within 2.5 kbp of breakpoints, including 95 and 21 protein-coding genes, respectively (Figure 2.2A). No highly impacted genes, as predicted by VEP, were found in this dataset. Using pLI scores, we identified 9 genes either modified or overlapped by inversions classified as loss-of-function intolerant in humans (Figure 2.2B).

In total, we found a significant depletion of protein-coding genes at deletion regions (162 genes within 2.5 kbp, empirical *p*-value = 0.001, Figure 2.3, Figure S2.4A) as well as at inversion breakpoints (21 protein-coding genes within 2.5 kbp, empirical *p*-value = 0.001, Figure 2.3, Figure S2.4B). Notably, this depletion did not persist when considering all transcribed elements intersecting SVs. Taking a closer look at genes with clear orthologs between chimpanzee and humans, we identified significantly fewer orthologs of deletion-impacted genes vs. inversion-impacted genes (67% vs. 89%, respectively; *p*-value = 1×10^{-5} Fisher's exact test). The majority of deletion-impacted genes with no orthologs were predicted to have high-VEP effect (179 out of 195 genes), suggesting that deletion of these genes completely ablated them from the chimpanzee genome.

Finally, we explored functional annotations of genes impacted by SVs. We found 208 transcribed elements impacted by deletions with known GO annotations as reported by DAVID (Huang, Sherman, and Lempicki 2009a; Huang, Sherman, and Lempicki 2009b) (Figure 2.2C). Compared to the complete set of human GO annotations, this gene list displays an

overrepresentation of genes associated with sensory perception of smell (GO:0050911, q-value = 8.7×10^{-11} and GO:0007608, q-value = 3.3×10^{-2}). We also found an overrepresentation of deletion-impacted genes involved in the G-protein coupled receptor signaling pathway (GO:0007186, q-value = 5×10^{-5}). Notably, both ontologies are primarily driven by known copy-number polymorphism that exists among olfactory-receptor genes (Nozawa, Kawahara, and Nei 2007). Inversions contained 140 genes with known GO functional annotation exhibiting an overrepresentation of regulation of cell differentiation (GO: 0045596, q-value = 1.2×10^{-4}).

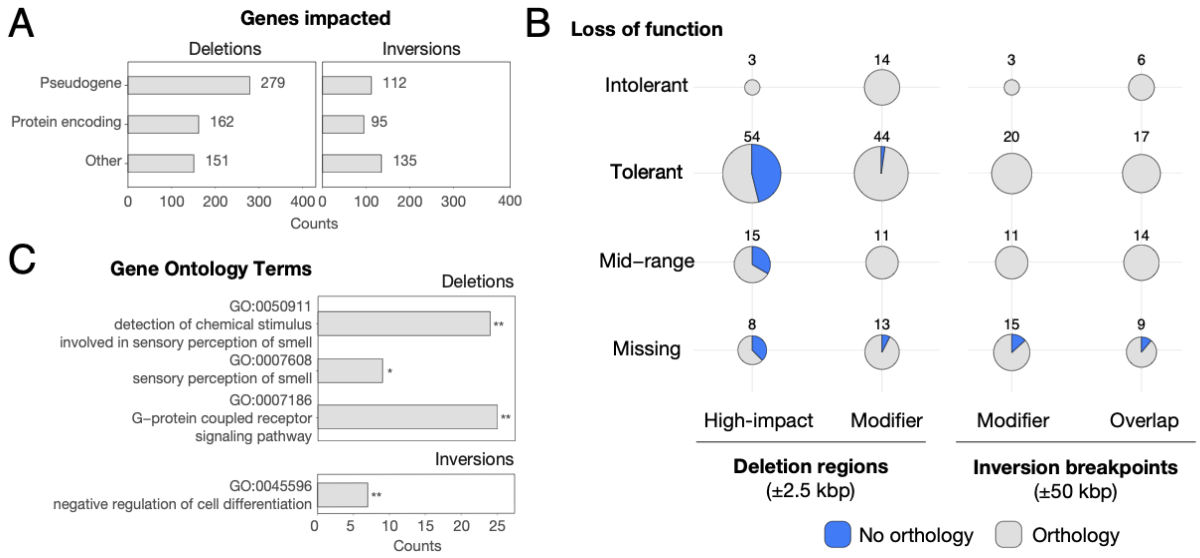


Figure 2.2: Description of genes overlapping identified structural variants. (A) Categories of genes overlapping deletion regions ±2.5 kbp (red) and inversion breakpoints ±50 kbp (cyan) as defined by ENSEMBL biotypes. **(B)** Number of protein-coding genes classified as LoF tolerant (pLI ≤ 0.1), intolerant (pLI ≥ 0.9) and middle range (pLI > 0.1 and pLI < 0.9) affected by deletions regions ±2.5 kbp (red) and inversion breakpoints ±50 kbp (cyan). Some affected genes lack LoF information (missing category). All genes impacted by deletions were classified by VEP as either highly impacted (feature ablation or truncation) or modified, while genes impacted by inversions were either modified or no effect was predicted (overlap only). Transcribed elements with no corresponding ENSEMBL transcript ID in humans were classified as no orthology (blue). **(C)** Overrepresented GO terms in genes impacted by deletions and inversions as reported by DAVID (* q-value < 0.05; ** q-value < 0.001). Counts represent the number of genes annotated with each GO term.

2.4.4 SVs and Gene Regulation

To understand if variants might affect gene regulation, we leveraged existing RNA-seq datasets generated from chimpanzee and human LCLs (Khan et al. 2013) and iPSCs (Pavlovic et al. 2018). From 55,461 human–chimpanzee orthologous transcribed features, we identified 6,565 and 8,946 genes in LCLs and iPSCs, respectively, as significantly DE between the two species. Among genes for which human-chimpanzee orthology was assigned that directly intersected SVs (N = 397 in deletions ±2.5 kbp; N = 61 for inversion breakpoints ±2.5 kbp), roughly half were significantly DE (57/135 LCL and 60/129 iPSC tested genes in deletions; 25/37 LCL and 22/36 iPSC tested genes in inversion breakpoints). We report a significant enrichment of DE genes from both cell types within (±2.5 kbp; permutation test empirical $p < 0.04$) and near (±50 kbp; $p < 0.01$) deletions and near (±50 kbp; $p < 0.002$) inversion breakpoints. DE gene enrichment was only significant within (±2.5 kbp) inversion breakpoints in LCLs (Figure 2.3, Figure S2.4).

Considering that gene regulation may be impacted by changes in genome organization, we next assayed the impact of SVs on chromatin structure by intersecting with previously identified TADs from a deeply-sequenced human LCL (GM12878) (Rao et al. 2014) and found 45 and 17 TAD boundaries likely disrupted by deletions and inversions, respectively, in chimpanzees. Similar to what others have reported previously (Fudenberg and Pollard, n.d.; Huynh and Hormozdiari 2019), deletions were less likely than expected by chance to straddle TAD boundaries, thereby generating putatively disrupted TADs (PDTs) (permutation test empirical $p < 0.01$ within 2.5 kbp and 50 kbp of deletions; Figure 2.3, Figure S2.4A). This is consistent with the hypothesis that regions maintaining chromatin structure are subject to negative selection. Not previously reported, we also found a significant depletion of PDTs intersecting inversions ($p = 0.001$ within 2.5 kbp and 50 kbp of inversions; Figure S2.4B). Within PDTs we identified 58 and 65 DE genes in LCLs and iPSCs, respectively. This suggests that disruption of genome

organization may have contributed to interspecific changes in gene expression for a subset of genes. Example loci are highlighted in Figure 2.4A and Figures S2.5, S2.7, and S2.8. Notably, chromatin structure was also apparently altered by variants near but not directly intersecting called TAD boundaries (Figure 2.4B, Figure S2.6)

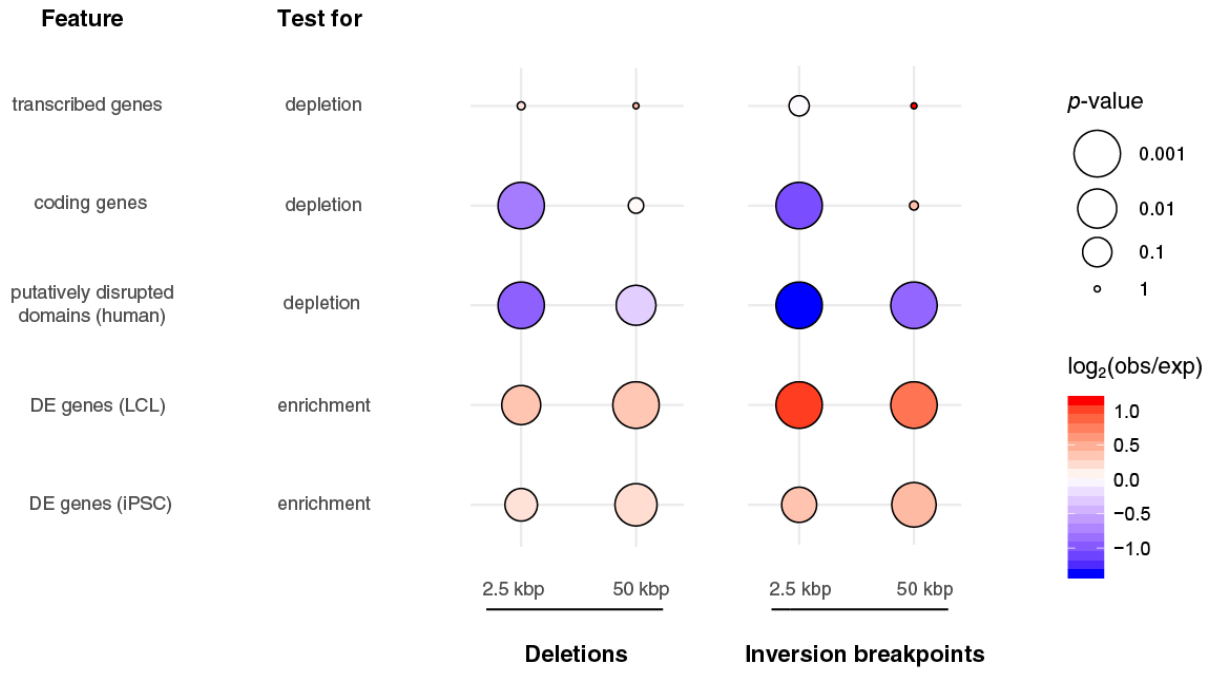


Figure 2.3: Enrichment and depletion tests of structural variants with genomic features. Both deletions and duplications were tested within 2.5 kbp (resolution of the SV calls) and 50 kbp. All annotated genes (GENCODE v27) and protein-coding genes were tested for depletion of SVs (top two rows) via permutation testing. Human TADs from the LCL GM12878 were tested for depletion of putatively disrupting SVs (i.e., SVs generating PDTs, third row). Human–chimpanzee DE genes from LCLs and iPSCs were also tested for enrichment in SVs via permutation testing (fourth and fifth rows). Circles are sized proportionally to the negative log of the empirical p-values and colored according to the strength of enrichment or depletion, represented by the log ratio of observed (number of features intersecting SVs) and expected (mean number of features intersecting 1000 permuted coordinate sets) counts.

To examine chromatin structure of PDTs, we generated orthologous Hi-C maps from human and chimpanzee LCLs and iPSCs (Eres et al. 2019) against the human reference (GRCh38) and directly compared differences in domain boundaries between species. Overall, domain calls

were similar between species (MoC 0.75 and 0.79 for LCLs and iPSCs, respectively (Zufferey et al. 2018)). We examined chimpanzee PDTs and identified more chimpanzee-unique boundaries than genome-wide boundaries (30.5% (18/59) versus 24.9% (1424/5714)). Similarly, for iPSCs we found 22.0% (13/59) of boundaries in PDTs are not shared with human, compared to 14.9% genome-wide boundaries (868/5834). These numbers suggest that TAD-altering SVs in may impact chromatin structure in chimpanzees.

Closer inspection of these regions revealed examples of altered gene expression coinciding with changes to three-dimensional chromatin structure. For example, the breakpoints of an inversion mapping to human chromosome 2q12.2-13 lie near altered domain boundaries and DE genes in iPSCs. Both *UXS1* and *SH3RF3* reside in altered domains and show increased contact frequency with chimpanzee-proximal inverted sequences that are over 1 Mbp away in the human genome (Figure 2.4A, Figure S2.5A). Similar gains of interactions are visible in the LCL Hi-C data and *UXS1* is also differentially expressed, though in the opposite direction (Figure S2.5B). A smaller inversion mapping to human chromosome 9q22.31 appears to mediate a domain fusion in both iPSCs and LCLs (Figure 2.4B, Figure S2.6). In both cell types, the nearby (<8 kbp) gene *SPTLC1* and truncated processed pseudogene *AL136097.2* are upregulated and downregulated, respectively, in chimpanzees compared with humans (Figure 2.4B, Figure S2.6). Other examples of domain-altering deletions and nearby DE genes are presented in Figures S2.7 and S2.8. Altogether, these data provide evidence that SVs may drive DE patterns, either through disruption of the transcribed sequence itself or through altered *cis* regulation, mediated by reorganization of physical interactions within chromatin.

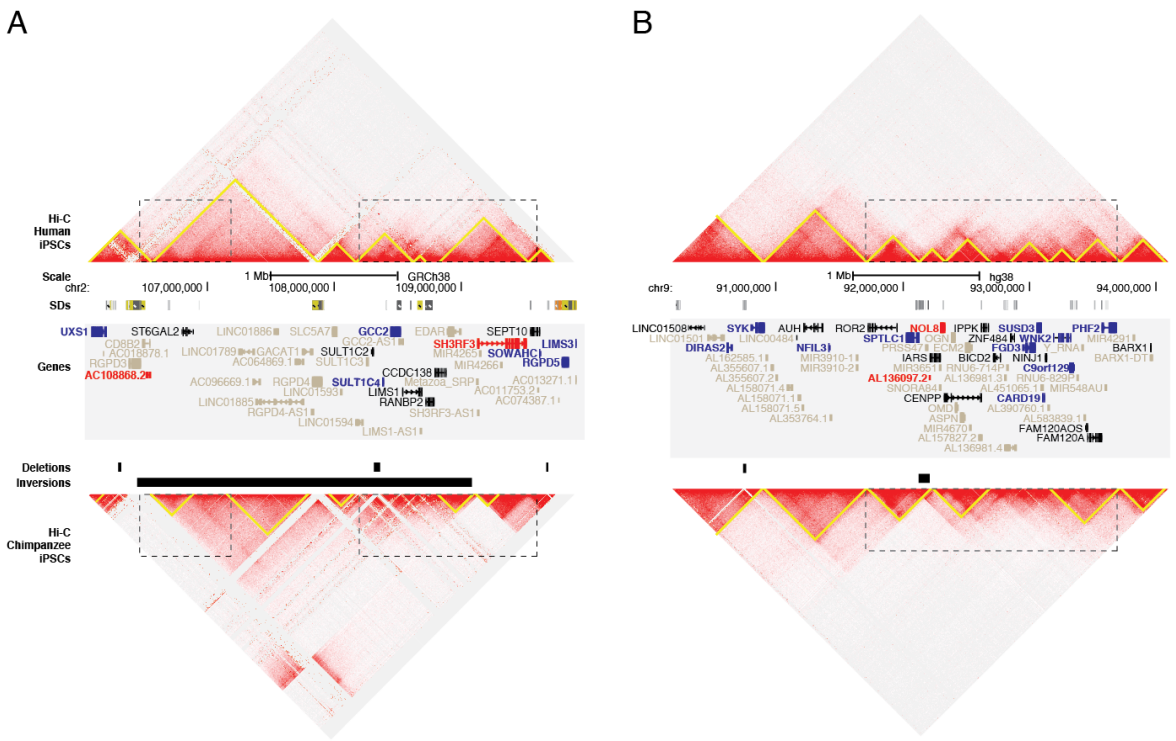


Figure 2.4: Genome organization of human and chimpanzee across regions with identified structural variants. The Hi-C genomic landscape of human (top) and chimpanzee (bottom) are depicted for iPSCs using Juicebox for (A) chromosome 2q12.2-q13 (chr2:106,095,001-109,905,000, GRCh38) and (B) chromosome 9q22.2-q22.32 (chr9:90,200,001-94,010,000, GRCh38). Predicted TADs (yellow triangles) were compared between species, noting differences at SVs (dotted boxes) including deletions and inversions. SDs are depicted as colored bars, taken from the UCSC Genome Browser track. Genes showing significant DE in chimpanzee versus humans are colored as red (up in chimpanzee) or blue (down in chimpanzee). Genes not included in the DE analysis are in gray.

2.4.5 Genes Showing Signatures of Natural Selection

Recent efforts to sequence diverse great ape genomes have led to identification of signatures of natural selection using SNV data that may help to explain features unique to chimpanzee species and subspecies (Prado-Martinez et al. 2013; Cagan et al. 2016; de Manuel et al. 2016; J. M. Schmidt et al. 2019). To understand if our identified SVs might impact the outcome of such studies or explain

signatures of selection previously identified, we compared our map of SVs with a recent study of natural selection in multiple genomes of the four chimpanzee subspecies (*Pan troglodytes verus*, *troglodytes*, *ellioti*, and *schweinfurthii*) mapped to the human reference genome (Cagan et al. 2016). In this study, among several other tests, the Hudson–Kreitman–Aguade (HKA) (Hudson, Kreitman, and Aguadé 1987) was used to identify the top 200 genes showing the strongest signatures of long-term balancing selection and positive selection in each subspecies. Intersecting this set of genes with our complete list of genes residing within or near deletions, we determined that of the 592 genes putatively disrupted by a deletion, 54 show strong signatures of natural selection using the HKA test (32 for positive and 22 for balancing selection). For inversions, of the 342 genes at or near inversion breakpoints, six show strong signatures of natural selection (five for positive, one for balancing). Of all genes affected by SVs and with strong signatures of natural selection, nine have evidence of DE in either LCLs or iPSCs, including two protein-coding genes showing signatures of balancing selection: *INPP4B*, which carries a deletion upstream of the transcription-start site and is upregulated in chimpanzee LCLs, and *HLA-F*, which is completely deleted and is upregulated in chimpanzee LCLs and downregulated in iPSCs. The possibility that these deletions generated beneficial expression changes that became strongly affected by natural selection makes these genes interesting candidates for follow up.

2.4.6 Genes Impacted by Chimpanzee-specific SVs

To hone in on SVs unique and universal to chimpanzees that may contribute to species-specific features, we consolidated the complete dataset of our newly discovered SVs and those previously published (Kronenberg et al. 2018; Catacchio et al. 2018; Sudmant et al. 2013). Filtering for only those with positive genotypes in >90% of chimpanzee individuals genotyped but found in neither humans (n = 8) nor gorillas (n = 8), we identified 209 deletions and 18 inversions. This set ranged in size from 10 kbp to 526 kbp for deletions and 12 kbp to 78 Mbp

for inversions (including the four large-scale cytogenetic events). Again due to the olfactory receptors at these loci, GO analysis shows that the genes contained within these SVs were overrepresented for the detection of chemical stimulus involved in sensory perception of smell (GO:0050911, *q*-value 4.1×10^{-2}). Focusing on genes with a higher likelihood of being functionally impacted by SVs, we identified 56 protein-coding genes with a high-impact VEP score (deletions) or within 2.5 kbp of a breakpoint (inversions) (Table 1).Of the 35 genes queried in our cross-species RNA-seq comparisons, 13 exhibited significant DE in chimpanzee versus human in LCLs and/or iPSCs, including *APOL4*, *CAST*, *CLN3*, *EFCAB13*, *EIF3C*, *IL18R1*, *NPIPB8*, *NPIPB9*, *NUPR1*, *RABEP2*, *SGF29*, *SLC01B3*, and *SULT1A1*. Additionally, six genes showed strong signatures of positive selection (*APOBR*, *IL27*, and *TUFM* at human chromosome 16p11.2 and *OR10H1* and *OR10H5* at human chromosome 19p13.12) or balancing selection (*CLC* at human chromosome 19q13.2). In all, this list of genes represents exciting candidates putatively implicated in chimpanzee-specific traits.

2.5 DISCUSSION

Most extensive SV analyses using comparative genomic approaches have used a single genome from one chimpanzee individual of the subspecies *Pan troglodytes verus* (i.e., Clint) (Kronenberg et al. 2018; Feuk et al. 2005; Newman et al. 2005; Catacchio et al. 2018; Chimpanzee Sequencing and Analysis Consortium 2005; Marques-Bonet et al. 2009). Here, we performed long-read sequencing of two additional individuals of the same subspecies, one of which carried admixture with *Pan troglodytes troglodytes*, using two orthogonal technologies: optical mapping and nanopore sequencing. To our knowledge, this represents the first nanopore sequence of a chimpanzee genome. From this, we discovered over 60,000 deletions and over 500 inversions (≥50 bp) when compared with the human reference (GRCh38), on the same scale as found in a recent comparison of the new chimpanzee assembly using a hybrid

assembly approach (panTro6) (Kronenberg et al. 2018). As expected, ONT sequencing was capable of detecting significantly more SVs, down to 50 bp with higher resolution at breakpoints (Figure S2.2A), compared to our BNG datasets. Notably many of the bioinformatically-identified SVs were redundant within and across technologies, which required additional filtering. To determine a higher-confidence set of SVs, we limited our analysis to variants ≥10 kbp in size with short-read Illumina sequencing evidence of the variant using SVtyper, a genotyping approach. Though the genotyping step significantly increased our confidence in variant calls, it also reduced the number of variants we identified (from 1,838 to 858 deletions and from 719 to 253 inversions), particularly for inversions, which are difficult to detect/genotype using short-read data. Additionally, our strict size cutoff limited our ability to discover transposable elements, which has been shown to represent a significant proportion of lineage divergence between chimpanzees and humans (Yohn et al. 2005). Furthermore, due to the uncertainty of the BNG breakpoints, most SVs discovered using only this approach were largely filtered from our subsequent analyses due to an inability to accurately genotype events. Nevertheless, our approach led to the discovery of 88 novel deletions and 36 novel inversions when compared to recent genome-wide scans. We note that we also excluded SDs and insertions from our analysis due to difficulties in discovery and subsequent validations using standard short-read genotyping approaches (Chander, Gibbs, and Sedlazeck 2019). As improved hybrid-based methods combining long- and short-read data are developed to more accurately identify SVs and their breakpoints, it will be a worthwhile endeavor to return to our dataset to discover additional SVs.

Our results implicated chimpanzee SVs in potentially impacting gene regulation and chromatin organization. It has been established that TAD structures are evolutionarily conserved (Rao et al. 2014; Dixon et al. 2012), and recent work finds that deletions altering TAD boundaries in humans are under purifying selection (Fudenberg and Pollard, n.d.; Huynh and Hormozdiari 2019). TAD structure is also conserved across apes, as evidenced by the incidence

of gibbon–human synteny breaks at domain boundaries (Lazar et al. 2018). Similarly, we find a depletion of PDTs generated by deletions in chimpanzees, as well as an expected but previously unreported reduction of inversions altering TADs. Taken together, the paucity of SVs altering domain boundaries suggests such variants experience strong negative selection in chimpanzees as in other species, perhaps due to conserved roles of TADs in modulating gene regulation. Despite the overall depletion of SVs at TAD boundaries, we did find an increased incidence of species-specific domain boundaries and significant enrichment of DE genes near SVs in the two cell types queried in this study, concordant with previous findings assessing the impact of deletions and duplications on differential gene expression in primate LCLs (Iskow et al. 2012). These analyses are subject to some limitations. Domain calling is highly sensitive to input parameters, but the pairs of Hi-C maps were subject to the same analysis and highly correlated at a variety of resolutions tested (MoC>0.7 at 100 kbp, 50 kbp, 25 kbp, and 10 kbp for iPSCs; 100 kbp and 50 kbp for LCLs) allowing for an assessment of genome-wide domain differences. Though the number of aligned reads were normalized to comparable levels, relative read depth is likely to vary across the genome due to differences in mappability. This is particularly likely at SV loci, where deletions and SDs generate discontinuities in the Hi-C matrix. As such, these domain calls should be interpreted primarily as a means of identifying regions of putatively disrupted chromatin structure.

Notably, many of the genes near SVs were not DE; however, it is plausible that these non-DE genes either remain connected to their regulatory elements or their associated elements are specific to cell types not assayed. Further, while it has been reported that topology-altering SVs can have little effect on gene expression (Ghavi-Helm et al. 2019), or that expression is not globally altered by loss of TADs (Rao et al. 2017), it could still be the case that expression-altering SVs are frequently subject to negative selection. For instance, TAD- and expression-altering SVs reported in humans are typically *de novo* and pathogenic (Franke et al. 2016; Lupiáñez et al. 2015). Regardless, our findings are concordant with those of (Kronenberg

et al. 2018), who reported an enrichment of human–chimpanzee cortical organoid DE genes near fixed human-specific SVs. While they find an enrichment for downregulated genes at insertions and deletions and upregulated genes at SDs, their analysis produced a much smaller set of DE genes (785 across both cell types, from single-cell RNA-seq) and a much larger set of variants (17,789). These findings are also in line with reports that SVs underlie many human expression quantitative trait loci (Chiang et al. 2017). However, considering the currently incomplete understanding of the relationship between gene regulation and three-dimensional chromatin structure, we emphasize that functional studies are necessary to causally implicate SVs in gene expression differences within or between species.

In addition to using Illumina genotyping of our identified SVs to filter out putatively false positive variants, we also used this information to query SV differences across subspecies. In our high-confidence set of SVs, we identified one novel deletion in chimpanzees (human chromosome 6q11.1; chr6:60639753-60662981, GRCh38) from our BNG data of the western individual carrying substantial central ancestry (S003641) that was also found uniquely in central chimpanzees (n = 4). Considering the relatively low ancestry contribution of this individual assigned to the central-chimpanzee population (~15%), this highlights the importance of sequencing more diverse individuals to identify additional subspecies-specific SVs to better survey the complete variant landscape. Using these same genotypes, we also focused on a set of genes universally impacted by SVs across all chimpanzees tested, but not detected in the other great apes studied (humans and gorillas), since these genes may putatively contribute to species-specific traits (Table 2.1). One example, *APOL4*, encoding Apolipoprotein L4, was completely deleted in all chimpanzees tested (n = 25) and also shown to be downregulated in both LCLs and iPSCs in chimpanzees when compared with humans. This gene is a member of a tandemly-duplicated family that has experienced a recent expansion in the primate lineage (Monajemi et al. 2002) and may play a role in lipid trafficking throughout the body. Human polymorphism at this locus has been shown to be associated with schizophrenia (Takahashi et

al. 2008). Several identified genes also exhibited signatures of natural selection. One example region putatively under balancing selection includes two deletions impacting the primate-expanded galectin gene cluster, a family of proteins that specifically bind β-galactoside sugars and are important in modulating immune response through interactions with T cells (Balogh et al. 2019). Both deletions (10 kbp and 35 kbp in size, respectively) are found homozygously in all chimpanzees tested (n = 25), and thus are likely not the target of balancing selection, but they completely ablated *CLC* (or *LGALS10*) and *LGALS14*, as well as the downstream region of *LGALS13* (Figure S2.9). Two of these genes (*LGALS13* and *14*), expressed exclusively in human placenta (Nandor Gabor Than et al. 2009), are important drivers of maternal adaptive immune response, with reductions in expression of either gene shown to be associated with an increased risk of preeclampsia (Nándor Gábor Than et al. 2014). Although the mechanisms are unclear, it is notable that other immune-related genes with connections to preeclampsia also exhibit signatures of balancing selection in humans (Andres et al. 2010; Wedenoja et al., n.d.; Tan, Shon, and Ober 2005). It is possible that deletions impacting this gene cluster may contribute to pregnancy-related outcomes in chimpanzees that could be subject to natural selective pressures.

Table 2.1: Protein-coding genes impacted by chimpanzee-specific deletions and inversions.

Gene	ENSEMBL ID	SV type	Description
APOBR	**ENSG00000184730**	**deletion**	**Apolipoprotein B receptor**
APOL1	ENSG00000100342	deletion	Apolipoprotein L1
*APOL4**	ENSG00000100336	deletion	Apolipoprotein L4
ATP2A1	ENSG00000196296	deletion	Sarcoplasmic/endoplasmic reticulum calcium ATPase 1
ATXN2L	ENSG00000168488	deletion	Ataxin 2 like
CARD18	ENSG00000255501	deletion	Caspase recruitment domain family member 18
*CAST**	ENSG00000153113	inversion	Calpastatin
CD19	ENSG00000177455	deletion	CD19 Molecule
CEACAM21	ENSG00000007129	deletion	CEA Cell Adhesion Molecule 21
CFHR2	ENSG00000080910	deletion	Complement Factor H Related 2

CFHR4	ENSG00000134365	deletion	Complement Factor H Related 4
CLC	**ENSG00000105205**	**deletion**	**Charcot-Leyden crystal Galectin**
*CLN3**	ENSG00000188603	deletion	CLN3 Lysosomal/Endosomal Transmembrane Protein, Battenin
CMPK1	ENSG00000162368	deletion	Cytidine/Uridine Monophosphate Kinase 1
CROCC	ENSG00000058453	inversion	Ciliary Rootlet Coiled-Coil, Rootletin
CYP2C18	ENSG00000108242	deletion	Cytochrome P450 Family 2 Subfamily C Member 18
DEFB128	ENSG00000185982	deletion	Defensin Beta 128
*EFCAB13**	ENSG00000178852	deletion	EF-Hand Calcium Binding Domain 13
*EIF3C**	ENSG00000184110	deletion	Eukaryotic Translation Initiation Factor 3 Subunit C
*IL18R1**	ENSG00000115604	inversion	Interleukin 18 Receptor 1
IL1RL1	ENSG00000115602	inversion	Interleukin 1 Receptor Like 1
IL27	**ENSG00000197272**	**deletion**	**Interleukin 27**
IL36B	ENSG00000136696	deletion	Interleukin 36B
IL37	ENSG00000125571	deletion	Interleukin 37
KRTAP19-6	ENSG00000186925	deletion	Keratin Associated Protein 19-6
KRTAP19-7	ENSG00000244362	deletion	Keratin Associated Protein 19-7
LCN10	ENSG00000187922	deletion	Lipocalin 10
LCN6	ENSG00000267206	deletion	Lipocalin 6
LGALS14	ENSG00000006659	deletion	Galectin 14
MERTK	ENSG00000153208	deletion	MER Proto-Oncogene, Tyrosine Kinase
*NPIPB8**	ENSG00000255524	deletion	Nuclear Pore Complex Interacting Protein Family Member B8
*NPIPB9**	ENSG00000196993	deletion	Nuclear Pore Complex Interacting Protein Family Member B9
*NUPR1**	ENSG00000176046	deletion	Nuclear Protein 1, Transcriptional Regulator
OBP2A	ENSG00000122136	deletion	Odorant Binding Protein 2A
OR10H1	**ENSG00000186723**	**deletion**	**Olfactory Receptor Family 10 Subfamily H Member 1**
OR10H5	**ENSG00000172519**	**deletion**	**Olfactory Receptor Family 10 Subfamily H Member 5**
OR2T33	ENSG00000177212	deletion	Olfactory Receptor Family 2 Subfamily T Member 33
OR6C2	ENSG00000179695	deletion	Olfactory Receptor Family 6 Subfamily C Member 2
OR6C3	ENSG00000205329	deletion	Olfactory Receptor Family 6 Subfamily C Member 3
OR6C65	ENSG00000205328	deletion	Olfactory Receptor Family 6 Subfamily C Member 65
OR6C70	ENSG00000184954	deletion	Olfactory Receptor Family 6 Subfamily C Member 70
OR6C75	ENSG00000187857	deletion	Olfactory Receptor Family 6 Subfamily C Member 75
OR6C76	ENSG00000185821	deletion	Olfactory Receptor Family 6 Subfamily C Member 76
POU6F2	ENSG00000106536	deletion	POU Class 6 Homeobox 2
*RABEP2**	ENSG00000177548	deletion	Rabaptin, RAB GTPase Binding Effector Protein 2
RACK1	ENSG00000204628	inversion	Receptor For Activated C Kinase 1

*SGF29**	ENSG00000176476	deletion	SAGA Complex Associated Factor 29
SH2B1	ENSG00000178188	deletion	SH2B Adaptor Protein 1
SLC35G4	ENSG00000236396	deletion	Solute Carrier Family 35 Member G4
*SLCO1B3**	ENSG00000111700	inversion	Solute Carrier Organic Anion Transporter Family Member 1B3
*SULT1A1**	ENSG00000196502	deletion	Sulfotransferase Family 1A Member 1
SULT1A2	ENSG00000197165	deletion	Sulfotransferase Family 1A Member 2
TUFM	**ENSG00000178952**	**deletion**	**Tumor Protein P53**
YAE1D1	ENSG00000241127	deletion	YAE1 Maturation Factor Of ABCE1
AC011604.2	ENSG00000257046	inversion	Uncharacterized
AL355987.1	ENSG00000204003	deletion	Uncharacterized

* Human and chimpanzee orthologs were tested and shown to be significant DE genes in either LCLs and/or iPSCs; Genes in bold were found to have strong signatures of positive or balancing selection using the HKA test (Cagan et al. 2016)

2.6 METHODS

Cell line Growth and DNA Extraction

Chimpanzee AG18359 and S003641 lymphoblastoid cell lines (LCLs) were generously shared with us by Dr. Yoav Gilad at the University of Chicago. LCLs were grown in T75 flasks with RPMI 1640 medium with L-Glutamine supplemented with 15% fetal bovine serum (Thermo Fisher Scientific, Waltham, MA, USA) and Penicillin-Streptomycin (100 U/ml, VWR, Radnor, PA, USA). For Illumina XTen sequencing, genomic DNA (gDNA) was isolated using DNeasy Blood and Tissue kit (Qiagen, Germantown, MD, USA) followed by RNase A treatment (Roche, Mannheim, Germany) and ethanol precipitation. For ONT PromethION sequencing, high molecular weight (HMW) gDNA was isolated from 5×10^7 cells following a modified Sambrook and Russell method as described previously (Kronenberg et al. 2018; Jain et al. 2018). The integrity of the HMW DNA was verified on a Pippin Pulse gel electrophoresis system (Sage Sciences, Beverly, MA). For the BNG assay, HMW gDNA was isolated from cells using the BNG Prep Blood and Cell Culture DNA Isolation Kit (BNG #80004). Briefly, 1.5×10^6 cells were resuspended in Cell Buffer and embedded in an agarose plug. The plug was treated with Proteinase K for 18 hours followed by RNase A digestion for one hour. After extensive washing,

the plug was melted, agarose was digested, and drop dialysis was performed to clean the DNA. A Qubit dsDNA BR Assay kit (Thermo Fisher Scientific) was used to quantify the DNA.

Determination of Chimpanzee Subspecies

gDNA isolated from AG18359 and S003641 LCLs was sequenced at ~30x coverage with Illumina HiSeq XTen (Novogene, Sacramento, CA and the UC Davis Genome Center DNA and Expression Analysis Core, Davis, CA, respectively) and SNVs were identified following a previously published approach (de Manuel et al. 2016). Briefly, reads were mapped using BWA (v0.7.17) against the chimpanzee reference genome (CHIMP2.1.4) using BWA-MEM with default parameters. Picard (v2.18.23) MarkDuplicates was used to remove duplicates with the flag "REMOVE_DUPLICATES = true." SNVs were called using FreeBayes (v1.2.0) with the following flags: "--standard-filters --no-population-priors -p 2 --report-genotype-likelihood-max --prob-contamination 0.05." We then filtered autosomal SNVs with QUAL ≥ 30 and intersected with data from de Manuel *et al.* (2016), callable genome regions, and finally merged with the 59 genomes from de Manuel *et al.* (2016), using bcftools merge with the following flags: "--missing-to-ref --force-samples." EIGENSOFT smartpca (Patterson, Price, and Reich 2006) was used to define principal components using the 59 Great Ape Genome Project (GAGP) chimpanzee genomes (de Manuel et al. 2016) and the genomes from AG18359 and S003641 were projected onto these components. We estimated the variance explained by each of the first 20 principal components as the eigenvalue / sum(top 20 eigenvalues). To expedite the analysis, it was run on 50% of the genome-wide SNVs. Admixture analysis was performed with the software ADMIXTURE (Alexander, Novembre, and Lange 2009) with a set the number of ancestral populations $K = 4$ corresponding to the four chimpanzee subspecies.

ONT Promethion Library Preparation and Sequencing

gDNA was sheared to an average size of 50 kbp using a Megaruptor instrument (Diagenode, Denville, NJ) and then verified on a Pippin Pulse gel. A sequencing library was prepared starting with 2 µg of sheared DNA using the ligation sequencing kit SQK-LSK109 (ONT, Oxford, UK) following the instructions of the manufacturer with the exception of extended incubation times for DNA damage repair, end repair, ligation, and bead elutions. Thirty femtomole of the final library was loaded on PromethION R9.4.1 flow cell (ONT, Oxford, UK) and the data were collected for 64 hours. Basecalling was performed live on the compute module using MinKNOW v2.1 (Oxford Nanopore Technologies, Oxford, UK).

BNG Saphyr Library Preparation and Sequencing

AG18359 and S003641 were sequenced at the McDonnell Genome Institute at Washington University and the UC Davis Genome Center DNA and Expression Analysis Core, respectively. A total of 750 ng of HMW gDNA was labeled with DLE-1 enzyme, followed by proteinase digestion and a membrane clean-up step using the BNG Prep DLS DNA Labeling Kit (#80005). After overnight staining with an intercalating dye, the labeled DNA was loaded onto a Saphyr Chip G2.3 (BNG #20366) and run on the Saphyr system (BNG #60325) using the Saphyr Instrument Control Software (ICS, version 3.1) to maximize throughput of molecules. Raw images of DNA were converted into digital molecules files using Saphyr ICS version 3.1.

Detection of SVs

To detect SVs, ONT long-reads were mapped to the human reference genome (GRCh38, no alternative haplotypes) using minimap2 (v2.17-r941) and SVs were identified using Sniffles (v1.0.11) with “--genotype” flag and default parameters. Large SVs were identified from BNG opticals maps using Bionano Solve (v3.5) (Hastie et al. 2017) *de novo* genome assembly and SV-discovery pipeline using human GRCh38 as the reference. The SV file in SMAP format was converted to VCF format using the smap_to_vcf_v2.py script contained in Solve software

(v3.4.1). Only the variants with “PASS” filter were considered in the analysis and homozygous reference calls were removed. SV size selection and filtering were performed with the bcftools (v1.9) view using the filter “INFO/SVLEN ≥ 10000 || INFO/SVLEN < −10,000” for both ONT and BNG datasets. To compare overlap between the SVs discovered by each method, we obtained 50% reciprocal overlap between features using bedtools intersect (v2.29.0) with flags “−f 0.5 −F 0.5.” Deletions and inversions were retrieved from the SVTYPE tag and processed separately in downstream analyses.

Genotyping and Filtering of SVs

Variants for each callset were genotyped independently using previously published Illumina data from 25 chimpanzees from all four subspecies, as well as eight gorillas and eight humans. SNV genotypes from non-human primates were retrieved from the GAGP (Prado-Martinez et al. 2013) and human SNV genotypes were obtained from the Simons Genome Diversity Project (Mallick et al. 2016). Reads were mapped to the human reference (GRCh38) using BWA MEM (0.7.17−r1188) (H. Li 2013) and subsequently merged and sorted with samtools (v1.9) for each individual. Large inversions and deletions (>10 kbp) were genotyped with SVtyper (v.0.7.1) (Chiang et al. 2015). Genotype information was retrieved using bedtools query (v2.29.0). To assess whether a variant was novel to this study, calls were compared to previously reported deletions and inversions larger than 10 kbp found in any great ape or any variant discovered in chimpanzee (Kronenberg et al. 2018; Catacchio et al. 2018; Sudmant et al. 2013) using bedtools intersect (v2.29.0) with 50% reciprocal overlap. SVs that were either (1) genotyped in one chimpanzee individual (1/1 or 0/1) or (2) reported as discovered in chimpanzee in previous studies, were selected to generate a higher confidence set (filter 1). This dataset was further refined by collapsing calls within the dataset with 50% reciprocal overlap. All novel calls were visually inspected in Integrative Genome Browser for ONT calls (Robinson et al. 2011) and Bionano Access for BNG calls. Also, SVs present in ≥90% of the chimpanzee individuals (22 or

more) as well as absent in the outgroups (human and gorilla) were included in the likely chimpanzee-specific dataset (filter 2). In Kronenberg *et al.* (2019), they genotyped eight chimpanzee individuals; as such, variants with evidence in seven or more individuals were also included in the chimpanzee-specific dataset. The distribution of high-confidence calls across the human reference (GRCh38) was plotted using the R package Karyoplotter (Gel and Serra 2017).

Annotation of Impacted Genes

Genes impacted by SVs were obtained by intersecting Gencode v27 genomics features annotation file to depletions coordinates ±2.5 kbp and inversions breakpoints (considered as estimated breakpoints ±2.5 kbp and ±50 kbp) using bedtools intersect (v2.29.0). The impact of the SVs on the function of the gene was predicted using Ensembl Variant Effect Predictor (VEP) (McLaren et al. 2016) with the Gencode v27 GTF file. The pLI score was obtained from the gene constraints scores table in the Exome Aggregation Consortium database (Karczewski et al. 2019). Gene ontology (GO) annotations and overrepresented terms were retrieved for each gene using DAVID (Huang, Sherman, and Lempicki 2009a, Huang, Sherman, and Lempicki 2009b) and by selecting terms at a 5% false-discovery rate (FDR). Genes previously identified as showing signatures of positive and balancing selection in chimpanzees were retrieved from previously published data (Cagan et al. 2016) and intersected with the set of genes impacted by SVs.

Differential Gene Expression

We obtained previously-published RNA-seq data from chimpanzee and human LCLs (Khan et al. 2013) and induced pluripotent stem cells (iPSCs) (Pavlovic et al. 2018). Raw data were trimmed using TrimGalore (v0.6.0) with the following parameters: “-q 20 --phred33 --length 20”. Transcripts per million (TPM) values were estimated using Salmon (v0.14.1) with the

“--validateMappings” flag (Patro et al. 2017) for all transcripts in GENCODE v27 and chimpanzee transcriptome published by (Kronenberg et al. 2018), which was based on a combination of orthologous genes identified via comparisons of human GENCODE v27 and novel transcripts identified through PacBio isoSeq of iPSCs. The R package tximport (Soneson, Love, and Robinson 2015a) was used to estimate gene-level counts from TPM values using the setting ‘countsFromAbundance = "lengthScaledTPM"’ for 55,461 annotated genes with equivalent identifiers in the two transcriptomes. Differential expression analysis was conducted with limma-voom (Smyth, n.d.; Law et al. 2014). Genes with fewer than 1 count per million across all samples were filtered from the analysis, and a two-factor model accounting for species and sex was implemented. Differentially expressed (DE) genes were called at a 5% FDR.

Topologically-Associated Domain (TAD) Analyses

We retrieved published TAD predictions from an LCL of a human female (GM12878) originally called with 4.9 billion Illumina reads (Rao et al. 2014). Domain coordinates were transformed from GRCh37 to GRCh38 using liftOver (UCSC Genome Browser; 9,262/9,274 domains successfully converted). Boundaries were defined as the start and end coordinates of each domain expanded to 5 kbp (resolution size of the TAD-calling analysis).

To directly compare domain boundaries between humans and chimpanzees, we generated DNase Hi-C libraries from three human (GM12878, GM20818, GM20543) and two chimpanzee (S007602, AG18359) LCLs as described by (Ramani et al. 2016). Raw data were processed using the Juicer pipeline (Durand et al. 2016) with the human reference GRCh38. Human alignments were downsampled to ~300 million reads to allow for equal comparison to chimpanzee, and Hi-C interaction matrices were generated with a (BWA) MAPQ filter of 30. Domains were called on Knight-Ruiz normalized contact matrices using TopDom (Shin et al. 2016) at 50 kbp resolution and the default window size (w = 5). Similarity between domain sets

was computed with the Measure of Concordance (MoC) as described previously (Zufferey et al. 2018) using chromosome 1. Domain calls were visualized with interaction maps (coverage normalized at 5 kbp resolution) using Juicebox (1.11.08). Across all chromosomes, boundaries unique to each species were considered to be the left and right coordinates of each domain, expanded to 50 kbp, when that region was not adjacent to (or overlapping) a boundary from the other species. This analysis was repeated using high-depth raw Hi-C data from four human and four chimpanzee iPSCs with approximately 1 billion reads per sample (combined across individuals; also normalized by downsampling) (Eres et al. 2019).

Permutation Analyses

For each variant, the distance to the nearest segmental duplication (SD; duplicated regions with >90% identity across >1 kbp, downloaded from UCSC Genome Browser GRCh38) was calculated using bedtools closest (v2.29.0). Regions of the same size (deletions ±2.5 kbp and inversions ±2.5 kbp) were randomly sampled from the human genome using bedtools shuffle (v2.29.0), and 5-kbp "breakpoints" were extracted from shuffled inversions. The distribution of the distance of these random regions to the nearest SD was plotted as density using the R package ggplot2. Permutation tests to assess the enrichment/depletion of genomic features (e.g., genes, boundaries) at SVs were similarly performed by shuffling the SV coordinates 1,000 times and counting the number of intersecting features with each set of coordinates. SVs were tested for enrichment of DE genes by generating 1,000 random samples of all genes tested in the expression analysis of equal size to the differential set. One-tailed empirical *p*-values were calculated as follows: $p = (M + 1) / (N + 1)$, where M is the number of iterations yielding a number of features less than (depletion) or greater than (enriched) observed and N is the number of iterations.

2.7 DATA AVAILABILITY

All sequence and optical mapping data generated as part of this project are available for download at the European Nucleotide Archive (accession number PRJEB36949).

2.8 ACKNOWLEDGMENTS

This work was funded in part by grants from the National Institutes of Health (NIH), including the National Institute of Neurological Disorders and Stroke (R00NS083627, M.Y.D.) and NIH Director's New Innovator award administered by the National Institute of Mental Health (DP2 OD025824 M.Y.D.). Additional support: M.Y.D. as a Sloan fellow (FG-2016-6814) and D.C.S. as a Fulbright fellow, A.M.A. and J.M.S. were supported by UCL's Wellcome Trust ISSF3 award (204841/Z/16/Z).

We would like to thank Y. Gilad and C. Chavarria for generously sharing chimpanzee LCLs with us, as well as the many labs participating in open-access research that made much of the genomic data used in this study available in the public domain. We thank F. Antonacci and T. Gill for thoughtful discussions and advice, and E. Georgian for critical review of the manuscript. Additionally, we are grateful to M. Kremitzki and T. Lindsay Graves at McDonnell Genome Institute and Washington University for supporting data analysis of our AG18359 BNG data.

2.9 SUPPLEMENTARY FIGURES

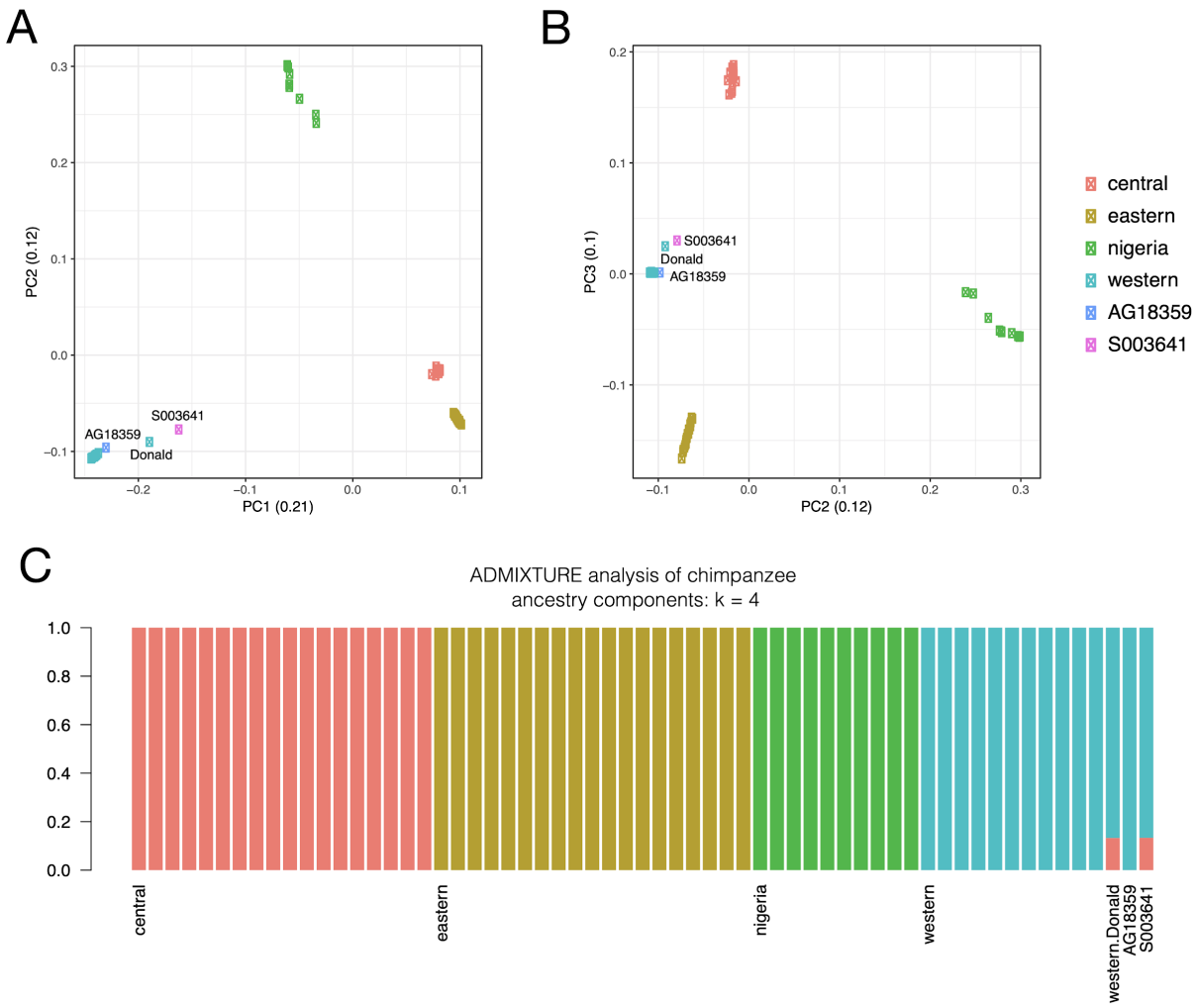


Figure S2.1: Chimpanzee subspecies identification. (A) and **(B)** PC analysis of chimpanzee genetic diversity. Both of the newly sequenced cell lines were projected onto PCs inferred from the 59 chimpanzees presented in de Manuel et al. (2016). Both cell lines show closest affinity to western chimpanzees (*Pan troglodytes verus*). While AG18359 clusters tightly with the western subspecies, S003461 also shows affinity to the central/eastern clade, with PC3 indicating that, like Donald, this cell line was derived from a hybrid individual with central ancestry. Values in parentheses are the proportion of variance explained by each PC. **(C)** ADMIXTURE analyses, assuming four ancestral components (K = 4) confirms the hybrid origin of S003641.

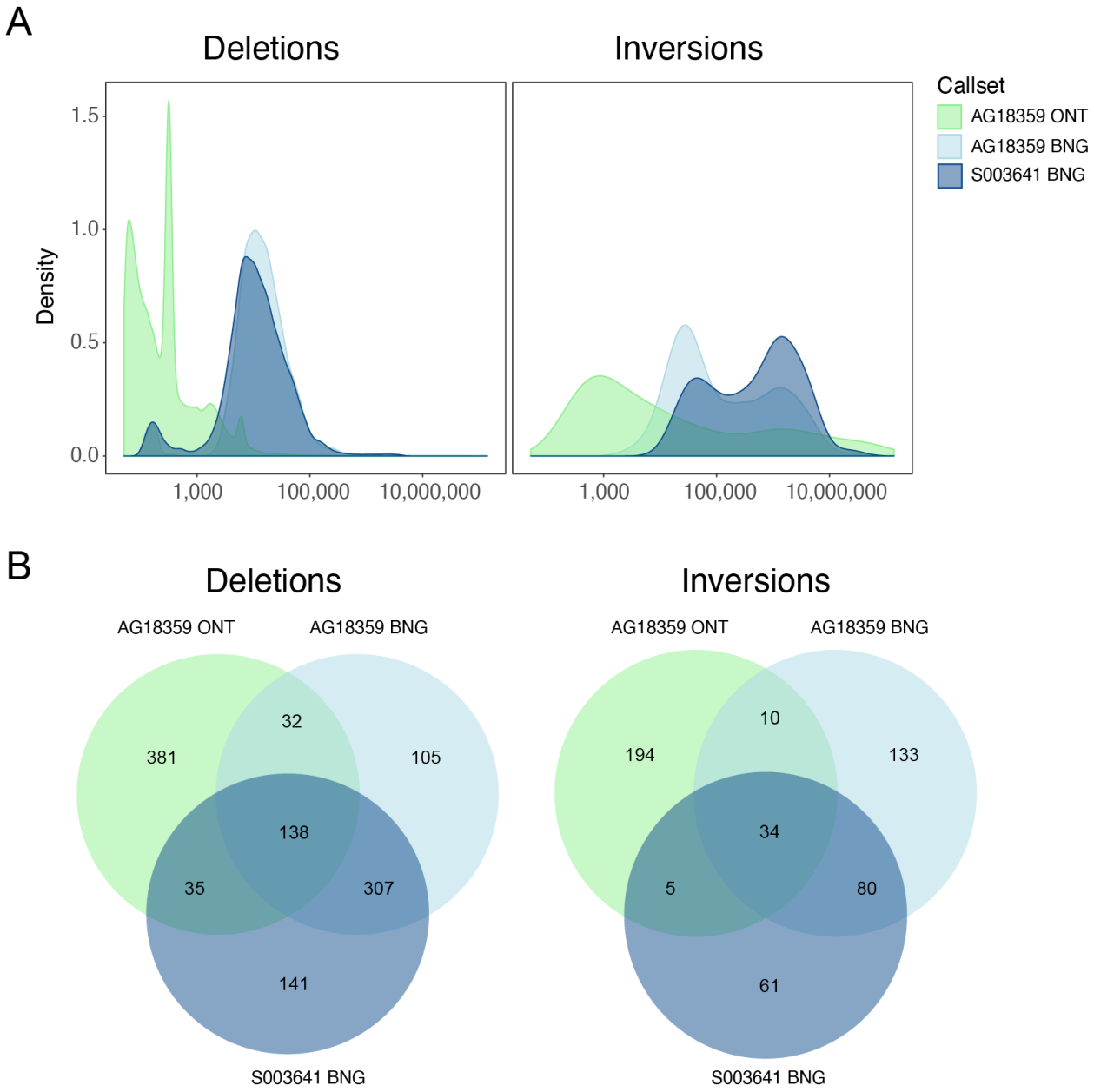


Figure S2.2: Description of SV discovery set. (A) Length distribution (x-axis in bp) of raw SV calls discovered by ONT (green) and BNG (light blue) from AG18359, and BNG from S003641 (dark blue). **(B)** Venn diagram comparing large (≥10 kbp) deletions (left) and inversions (right) discovered for each individual and technology (not to scale). Two variants were considered the same if they have a 50% reciprocal overlap.

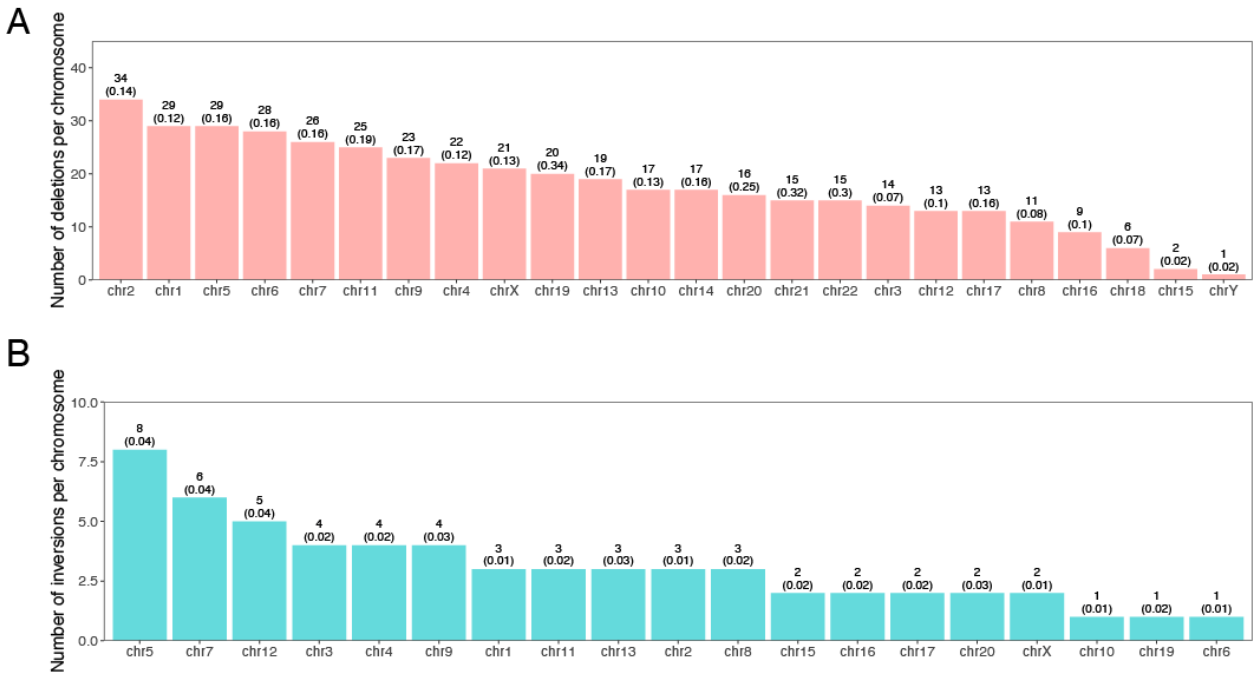


Figure S2.3: Histogram of identified SV events per chromosome. The number of high-confidence SV events discovered is depicted for **(A)** deletions and **(B)** inversions. The normalized number of events per Mbp for each chromosome is displayed in parentheses.

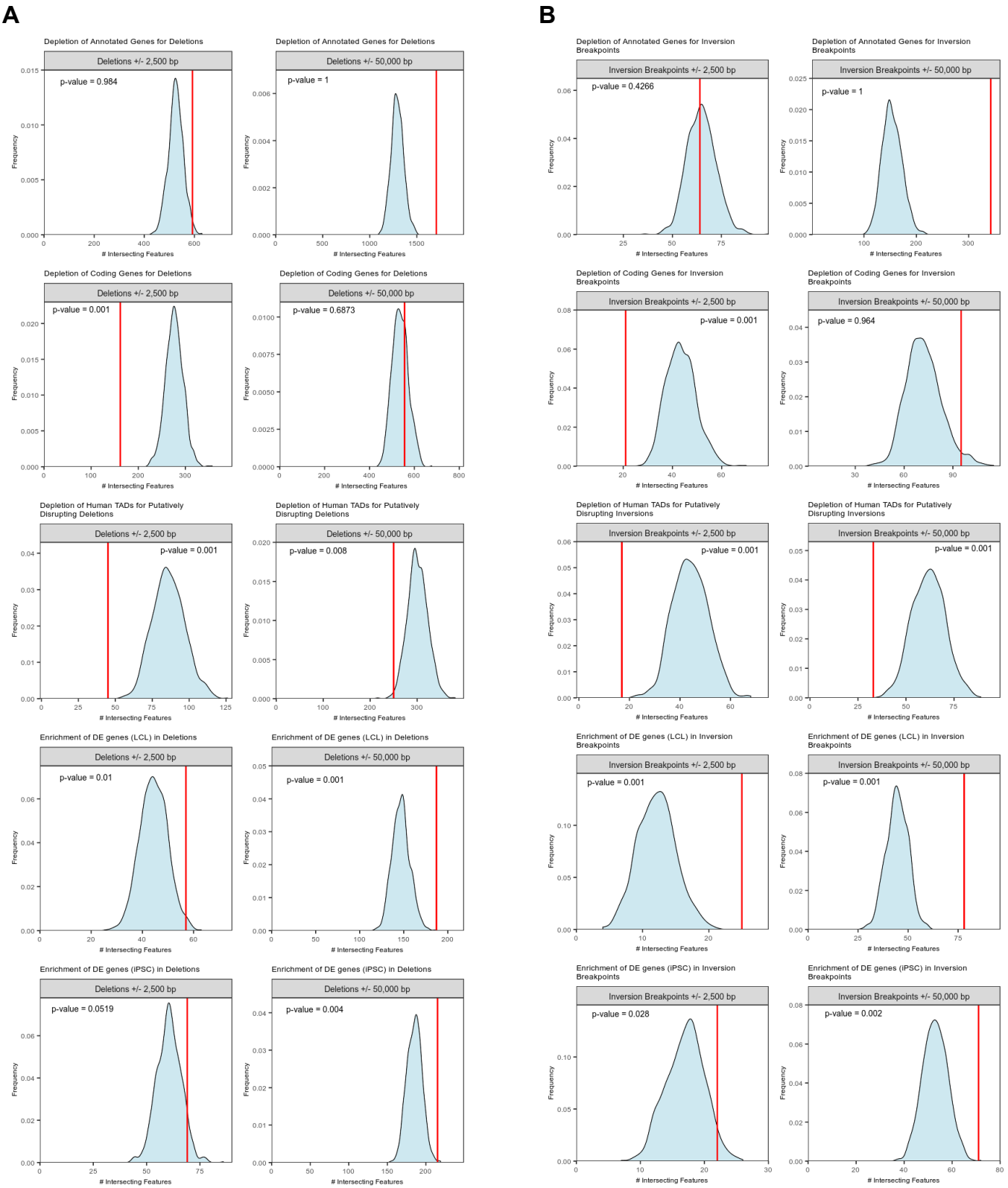

A
Depletion of Annotated Genes for Deletions
Deletions +/- 2,500 bp
p-value = 0.984
Deletions +/- 50,000 bp
p-value = 1
Depletion of Coding Genes for Deletions
p-value = 0.001
p-value = 0.6873
Depletion of Human TADs for Putatively Disrupting Deletions
p-value = 0.001
p-value = 0.008
Enrichment of DE genes (LCL) in Deletions
p-value = 0.01
p-value = 0.001
Enrichment of DE genes (iPSC) in Deletions
p-value = 0.0519
p-value = 0.004
B
Depletion of Annotated Genes for Inversion Breakpoints
Inversion Breakpoints +/- 2,500 bp
p-value = 0.4266
Inversion Breakpoints +/- 50,000 bp
p-value = 1
Depletion of Coding Genes for Inversion Breakpoints
p-value = 0.001
p-value = 0.964
Depletion of Human TADs for Putatively Disrupting Inversions
p-value = 0.001
p-value = 0.001
Enrichment of DE genes (LCL) in Inversion Breakpoints
p-value = 0.001
p-value = 0.001
Enrichment of DE genes (iPSC) in Inversion Breakpoints
p-value = 0.028
p-value = 0.002
Frequency
Intersecting Features

Figure S2.4: Enrichment/depletion of SV breakpoints for genomic features of interest as determined by permutation testing. Each plot compares the observed count of intersecting features (red vertical line) to a distribution of counts generated from 1000 permuted sets of coordinates (for testing depletion of SVs) or 1000 randomly selected genes from the background list of each DE analysis (for testing enrichment of DE genes in SVs) for **(A)** deletions and **(B)** inversions.

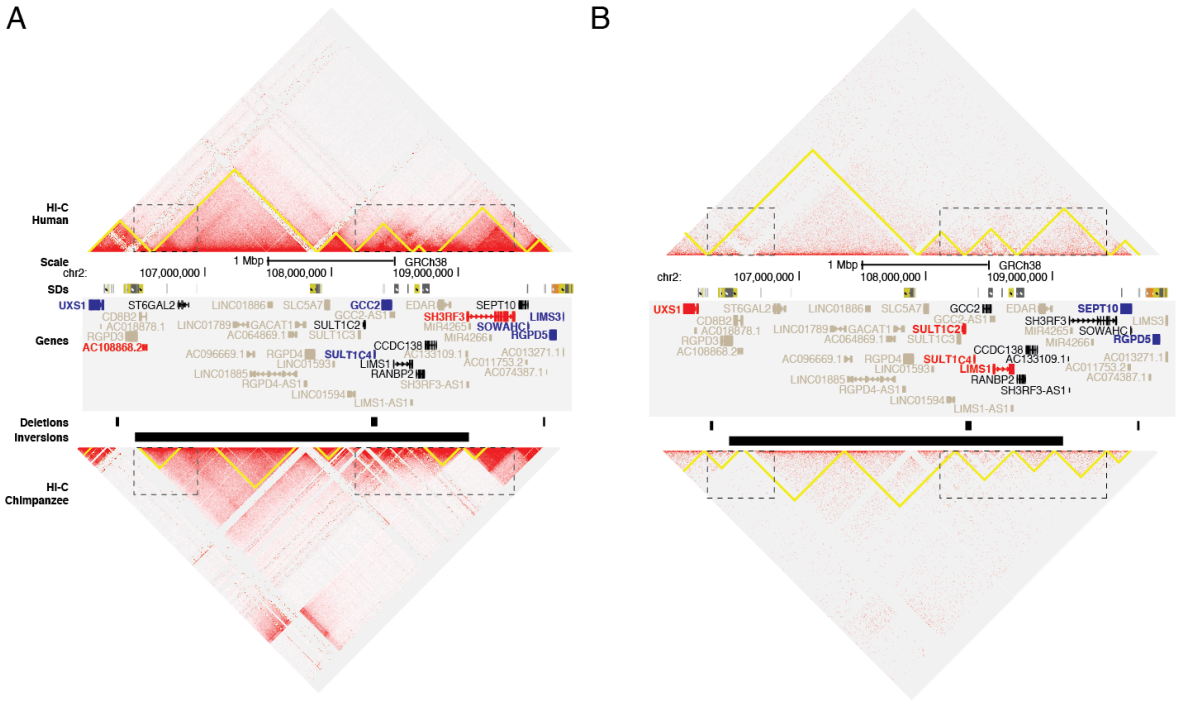


Figure S2.5: Genome organization of human chromosome 2q12.2-q13. The Hi-C genomic landscape of human (top) and chimpanzee (bottom) are depicted for iPSCs **(A)** and LCLs **(B)** using Juicebox at chr2:106,095,001-109,905,000 (GRCh38). Predicted TADs (yellow triangles) were compared between species, noting differences at SVs (dotted rectangles) including deletions and inversions. SDs are depicted as colored bars, taken from the UCSC Genome Browser track. Genes showing significant DE in chimpanzee versus humans are colored as blue (down in chimpanzee) or red (up in chimpanzee). Genes not included in the DE analysis are gray.

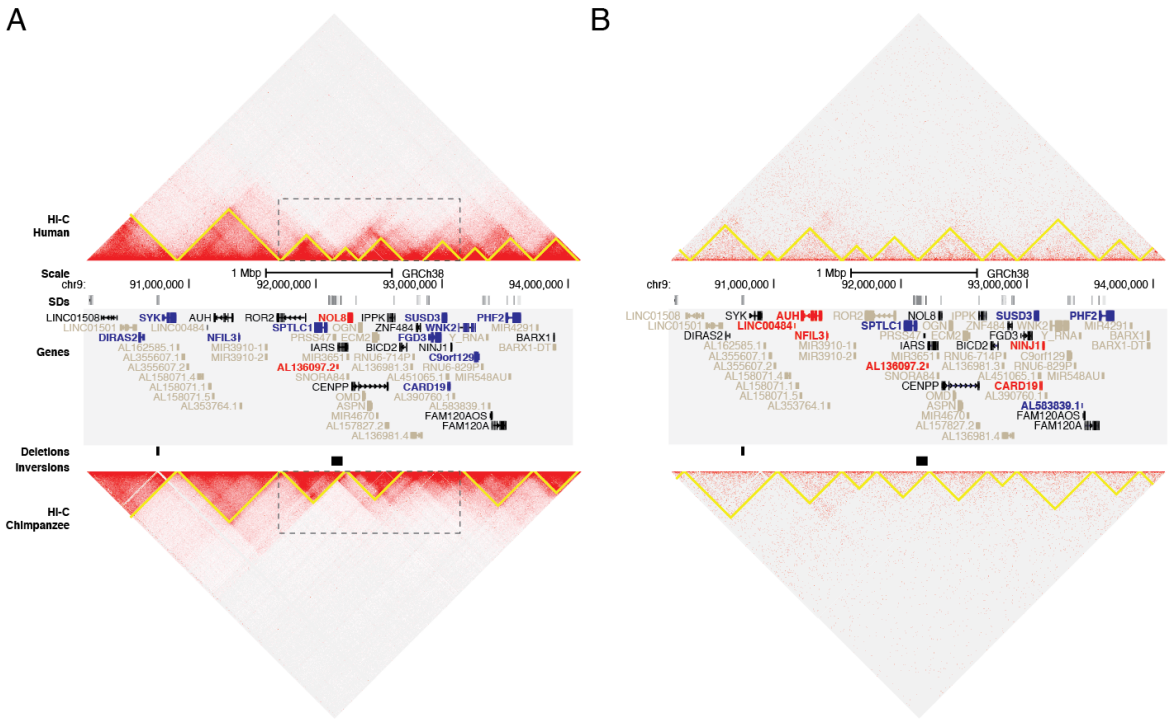


Figure S2.6: Genome organization of human chromosome 9q22.2-q22.32. The Hi-C genomic landscape of human (top) and chimpanzee (bottom) are depicted for iPSCs **(A)** and LCLs **(B)** using Juicebox at chr9:90,200,001-94,010,000 (GRCh38). Predicted TADs (yellow triangles) were compared between species, noting differences at SVs (dotted rectangles) including deletions and inversions. SDs are depicted as colored bars, taken from the UCSC Genome Browser track. Genes showing significant DE in chimpanzee versus humans are colored as blue (down in chimpanzee) or red (up in chimpanzee). Genes not included in the DE analysis are gray.

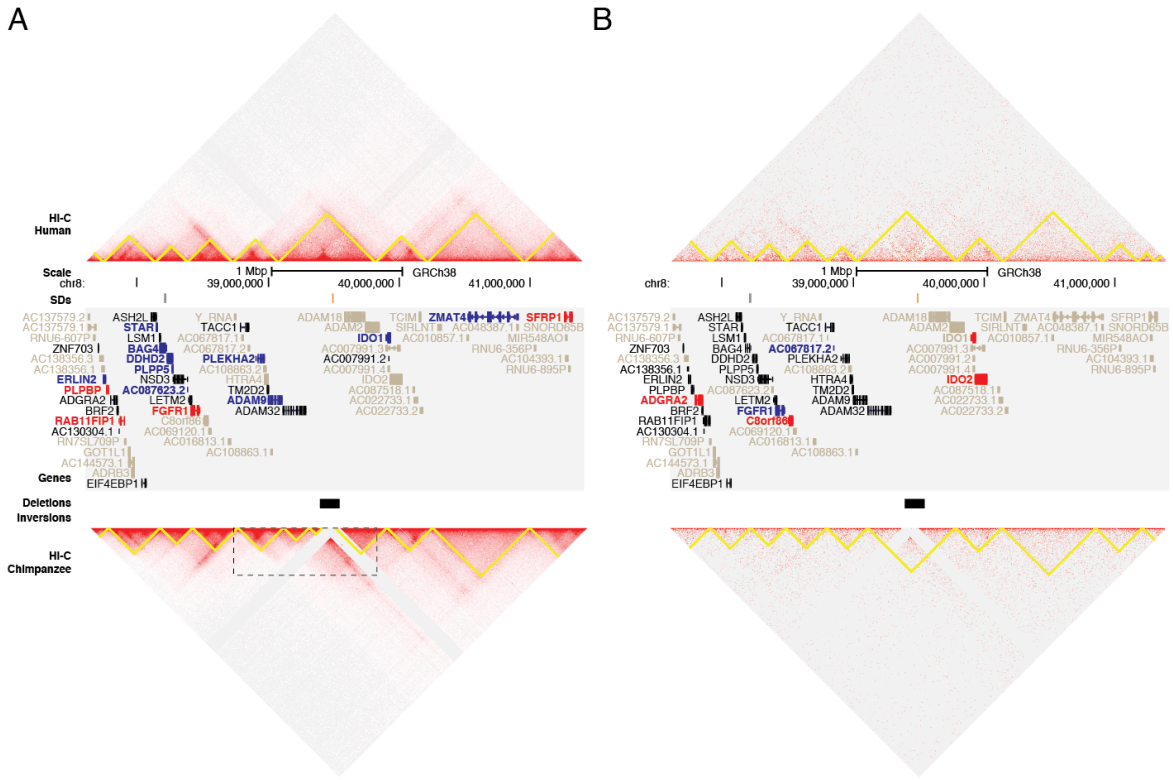


Figure S2.7. Genome organization of human chromosome 8p11.23-p11.21. The Hi-C genomic landscape of human (top) and chimpanzee (bottom) are depicted for iPSCs **(A)** and LCLs **(B)** using Juicebox at chr8:37,620,001-41,430,000 (GRCh38). Predicted TADs (yellow triangles) were compared between species, noting differences at SVs (dotted rectangle) including deletions and inversions. SDs are depicted as colored bars, taken from the UCSC Genome Browser track. Genes showing significant DE in chimpanzee versus humans are colored as blue (down in chimpanzee) or red (up in chimpanzee). Genes not included in the DE analysis are gray.

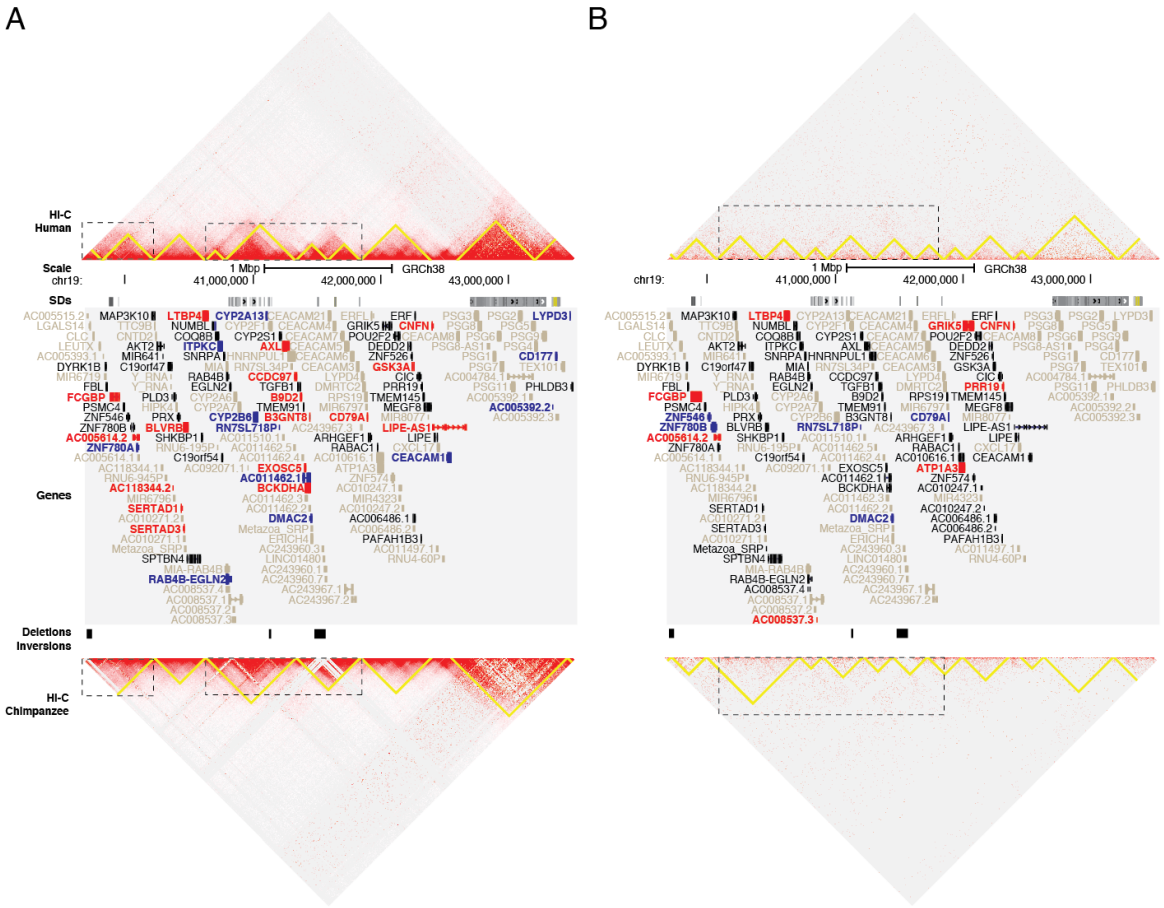


Figure S2.8: Genome organization of human chromosome 19q13.2-q13.31. The Hi-C genomic landscape of human (top) and chimpanzee (bottom) are depicted for iPSCs **(A)** and LCLs **(B)** using Juicebox at chr19:39,685,001-43,495,000 (GRCh38). Predicted TADs (yellow triangles) were compared between species, noting differences at SVs (dotted rectangles) including deletions (no inversions were identified as this locus). SDs are depicted as colored bars, taken from the UCSC Genome Browser track. Genes showing significant DE in chimpanzee versus humans are colored as blue (down in chimpanzee) or red (up in chimpanzee). Genes not included in the DE analysis are gray.

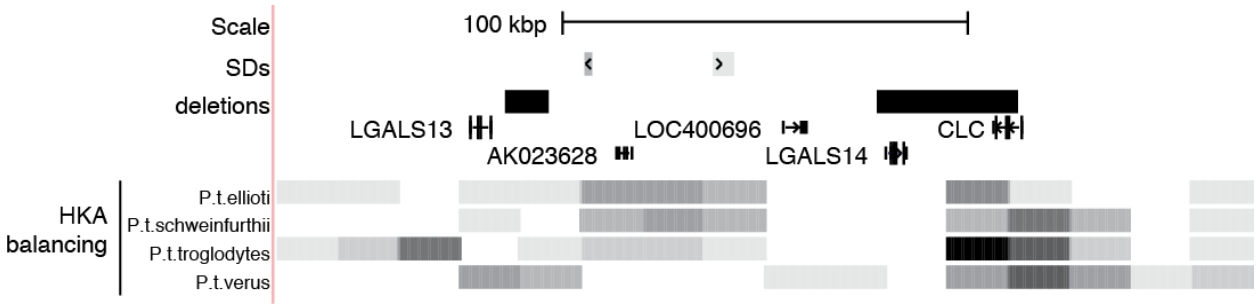


Figure S2.9. Chimpanzee-specific deletions of the galectin family of genes. Pictured is a UCSC Genome Browser snapshot of human chromosome 19p13.2. The locations of SDs (colored bars), deletions (black bars), and genes are indicated. For each subspecies of chimpanzee (*Pan troglodytes* (*P.t.*)), the -log-*p*-value for the HKA test of balancing selection is depicted as shades of gray in 15-kbp windows (darker shade indicates greater significance) as determined by Cagan et al., 2016 (human reference hg18).

Chapter 3:
Single-cell strand sequencing of a macaque genome reveals multiple nested inversions and breakpoint reuse during primate evolution

3.1 ABSTRACT

The rhesus macaque is an Old World monkey that shared a common ancestor with humans ~25 million years ago and is an important animal model for human disease studies. A deep understanding of its genetics is therefore required for both biomedical and evolutionary studies. Among structural variants, inversions represent a driving force in speciation and play an important role in disease predisposition. Here we generated a genome-wide map of inversions between human and macaque, combining single-cell strand sequencing with cytogenetics. We identified 375 total inversions between 859 bp and 92 Mbp, increasing by eightfold the number of previously reported inversions. Among these, 19 inversions flanked by segmental duplications overlap with recurrent copy number variants associated with neurocognitive disorders. Evolutionary analyses show that in 17 out of 19 cases, the Hominidae orientation of these disease-associated regions is always derived. This suggests that duplicated sequences likely played a fundamental role in generating inversions in humans and great apes, creating architectures that nowadays predispose these regions to disease-associated genetic instability. Finally, we identified 861 genes mapping at 156 inversion breakpoints, with some showing evidence of differential expression in human and macaque cell lines, thus highlighting candidates that might have contributed to the evolution of species-specific features. This study depicts the most accurate fine-scale map of inversions between human and macaque using a two-pronged integrative approach, such as single-cell strand sequencing and cytogenetics, and

represents a valuable resource toward understanding of the biology and evolution of primate species.

3.2 CONTRIBUTIONS

This chapter is adapted with minimal modification from the following published work:

> Flavia Angela Maria Maggiolini, Ashley D. Sanders, Colin James Shew, Arvis Sulovari, Yafei Mao, Marta Puig, Claudia Rita Catacchio, Maria Dellino, Donato Palmisano, Ludovica Mercuri, Miriana Bitonto, David Porubský, Mario Cáceres, Evan E. Eichler, Mario Ventura, Megan Y. Dennis, Jan O. Korbel, and Francesca Antonacci. 2020. "Single-cell strand sequencing of a macaque genome reveals multiple nested inversions and breakpoint reuse during primate evolution." *Genome Research* 30: 1680-1693. http://www.genome.org/cgi/doi/10.1101/gr.265322.120.

F.A. and F.A.M.M. designed the study. A.D.S., F.A.M.M., and C.R.C. performed Strand-seq single-cell libraries construction. A.D.S. performed Illumina sequencing and data analysis. F.A.M.M., M.D., D.Pa., and L.M. performed FISH experiments. F.A.M.M. and C.R.C. performed Illumina sequencing of BAC clones. M.P. and M.B. performed PCR experiments. C.J.S. and M.Y.D. performed differential gene expression and chromatin conformation analyses. D.P. and E.E.E. contributed to evolutionary analysis. A.S. and D.P. performed simulations of nested inversions. Y.M. performed heterozygosity analysis. F.A., F.A.M.M., A.D.S., M.V., C.J.S., M.Y.D., M.P., M.C., J.O.K., and E.E.E. contributed to data interpretation. F.A. and F.A.M.M. wrote the manuscript. All authors read and approved the final manuscript.

3.3 INTRODUCTION

Structural variants (SVs) are genomic alterations that involve segments of DNA that are >50 bp (Iafrate et al. 2004; Sebat et al. 2004; Tuzun et al. 2005; Kidd et al. 2008; Mills et al. 2011; Eichler 2019). SVs can include "balanced" rearrangements, such as inversions and translocations, or genomic imbalances (duplications and deletions), commonly referred to as

copy number variants (CNVs). Inversions represent an intriguing class of SVs, first identified by Sturtevant in 1917 (Sturtevant 1917), that play a key dual role in primate evolution and predisposition to disease. Chromosome inversions are the most common rearrangements differentiating humans and the great ape species at the karyotypic level (Yunis and Prakash 1982; Yunis, Sawyer, and Dunham 1980; Nickerson and Nelson 1998; Devin P. Locke et al. 2003). A key evolutionary effect of inversions is that they suppress recombination as heterozygotes. As a consequence, inversions can act as an initial step toward genomic divergence by protecting chromosomal regions from gene flow (Rieseberg 2001). Inversions are also the source of the majority of genetic structure within populations and affect polymorphisms chromosome-wide (Corbett-Detig and Hartl 2012). Despite the importance of inversions as a major mechanism of genome reorganization, we still struggle to understand how and why they evolve almost a century after Sturtevant's initial discovery owing to technical challenges in their discovery.

Recently, new advances in sequencing technologies, optical mapping, and novel assembly algorithms have deepened our understanding of SVs and their role in genome function, evolution, and disease. However, inversions still remain one of the most poorly studied types of genetic variation, mainly because of our insufficient ability to accurately detect them. Their balanced nature and the presence of segmental duplications (SDs) at inversion boundaries pose major challenges for inversion detection. A number of studies have identified and characterized large inversions (>2 Mbp) using laborious target-based cytogenetic studies (M. Ventura, Archidiacono, and Rocchi 2001; Mario Ventura et al. 2003, 2004, 2007, 2011; Carbone et al. 2002; Cardone et al. 2006, 2007, 2008; Kehrer-Sawatzki and Cooper 2008; Kehrer-Sawatzki, Sandig, et al. 2005; Kehrer-Sawatzki et al. 2005; Stanyon et al. 2008; Capozzi et al. 2012). With the advent of sequencing, inversions have been inferred from next-generation sequence data by abnormal paired-end mapping and split-read alignment signatures (Tuzun et al. 2005; Kidd et al. 2008). Because the human genome is highly enriched in SDs (J. A. Bailey

and Eichler 2006), these approaches often lead to false positives or fail to detect inversions flanked by highly identical sequences.

Recently, strand-specific sequencing technologies have been developed and successfully applied to detect inversions in human genomes (Porubsky, Sanders, Höps, et al. 2020). Single-cell template strand sequencing (Strand-seq) is an amplification-free sequencing technique that selectively sequences the template strands used for DNA replication during a single mitotic cell division. Although Strand-seq resolution is lower in repetitive regions, it is able to detect inversions with a size limit up to 1 kbp (Sanders et al. 2016; Chaisson et al. 2019). The power of Strand-seq lies in its ability to track the directionality of DNA template strands in every single cell. Inversions are detected solely by identifying DNA sequence strand switches internal to the inverted sequence, readily identifying inversions flanked by large SDs that can be neither assembled nor traversed using standard DNA sequencing technologies. These features make Strand-seq the leading genomic method for nontargeted inversion detection, especially suited for large repeat-embedded variants (Sanders et al. 2016; Chaisson et al. 2019).

In this study, we took advantage of this newly developed method to identify inversions in the rhesus macaque genome. Macaque is an Old World monkey sharing a common ancestor with humans ~25 million years ago. This primate showed some similarities with humans in physiology, neurobiology, and susceptibility to infectious diseases, making it one of the most important primate models for studies on human diseases (Rhesus Macaque Genome Sequencing and Analysis Consortium et al. 2007); a deeper understanding of its genetic features can help us better understand these processes. By applying Strand-seq in conjunction with molecular cytogenetics, we generated a complete map of inversions between human and macaque and identified variants affecting key genes that may be essential in understanding the evolution of specific human traits. Thus, this study represents a critical resource for genomic research that fills a major gap in the nonhuman primate research field.

3.4 RESULTS

3.4.1 Detection of inversions by Strand-seq

To detect inversions in the macaque genome, we generated 61 high-quality Strand-seq cell libraries for one macaque individual (MMU1). To overcome the low sequence coverage obtained for each single-cell Strand-seq library, selected libraries were merged into a high-coverage and directional composite file for each chromosome (Sanders et al. 2016). Inversions arising between macaque and human were discovered by aligning the macaque data to the human reference assembly genome (GRCh38/hg38) and performing breakpoint (BP) detection on the composite file using breakpointR (Porubsky, Sanders, Taudt, et al. 2020). This allowed us to predict the location and genotype of inversions based on segmental changes in read directionality arising within inverted loci. The BED-formatted composite file was additionally uploaded as custom track onto the UCSC Genome Browser (GRCh38/hg38 release) to facilitate manual curation and analysis of all predicted inversions. Because previous comparative studies were focused on autosomes, we excluded the X and Y Chromosomes from our analysis. By using this approach, we initially identified 373 inversions in the Strand-seq data that after validation and literature interrogation were extended to 375 (see “Validation of inversions in macaque” section) (Figure 3.1).

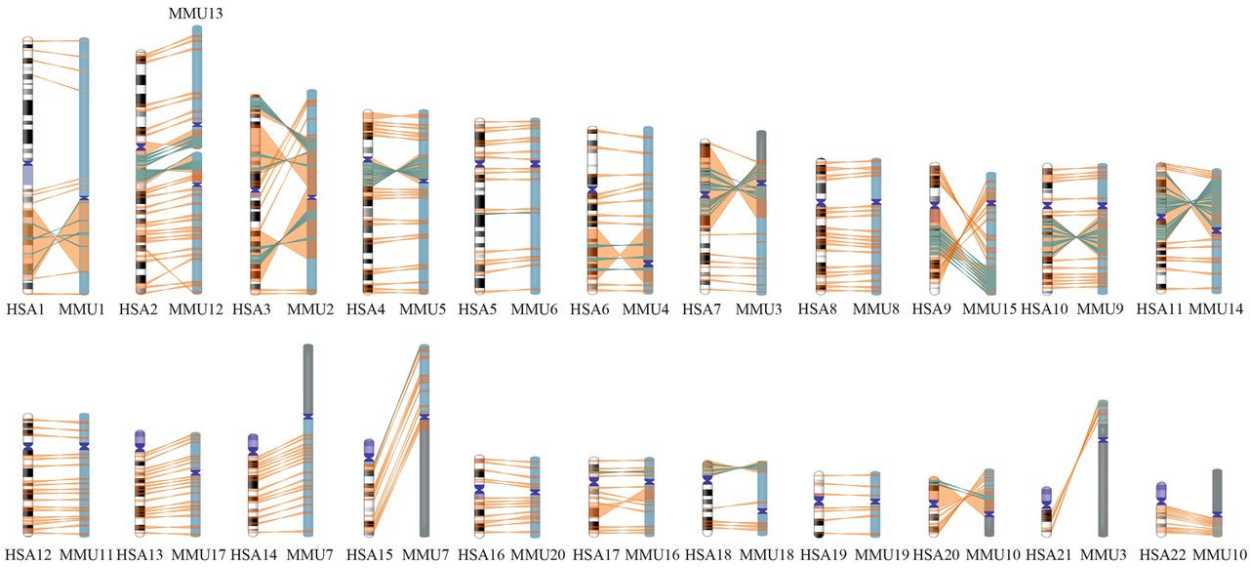


Figure 3.1: Genome-wide distribution of 375 inversions detected by Strand-seq between human and macaque genomes. Human chromosomes are shown on the left; orthologous macaque chromosomes, on the right. Orange lines between human and macaque ideograms show inversions detected by a simple strand switch. Green lines represent inversions within inversions, which are apparently direct by Strand-seq.

Inversions ranged in size from 859 bp to 92 Mbp and were distributed along all chromosomes with the highest density (number of inversions every 10 Mbp) on chromosome 22 and the lowest on chromosome 1. The vast majority of detected inversions (359 out of 375) appeared in homozygous state (i.e., both homologs being inverted and thus showing a "complete" switching of the read directionality within the locus). Conversely, the remaining 16 inversions were found in a heterozygous state (i.e., only one homolog was inverted and thus showed a "mixed" switch in read directionality); these likely represent polymorphic inversions among macaque individuals. Moreover, 87 out of 375 inversions were nested within larger inversions, in a "matryoshka" configuration, and are apparently direct by Strand-seq. One inversion (Chr7_inv12) flipped twice during evolution and appeared inverted by Strand-seq .

3.4.2 Comparison of human and macaque assemblies and published literature

We first compared the detected inversions with rearrangements previously reported for macaque (Mario Ventura et al. 2007; Antonacci et al. 2009; Catacchio et al. 2018; Maggiolini et al. 2019) and confirmed the orientation of 48 events, which correspond mainly to larger inversions (>130 kbp). Conversely, 327 regions (87%), ranging in size from 859 bp to 9 Mbp and amounting to 55.6 Mbp of inverted DNA, were novel and described here for the first time as human/macaque inversions. However, 58 of these 327 regions were previously found to be inverted in Hominidae in a study in which Strand-seq was applied to discover inversions between humans and great apes (Porubsky, Sanders, Höps, et al. 2020), and 11 out 327 have been described to be polymorphic inversions in human (Antonacci et al. 2009; Chaisson et al. 2019; Giner-Delgado et al. 2019; Puig et al. 2020).

All previously published macaque inversions (Mario Ventura et al. 2007; Catacchio et al. 2018) have been detected by our Strand-seq analysis, except for three regions (Chr1_inv5, Chr16_inv4, and Chr16_inv5) for which Strand-seq shows a direct orientation. However, this is not a Strand-seq error because these inversions have been reported to be potential misassemblies (Chr1_inv5) and minor alleles (Chr16_inv4 and Chr16_inv5) of the human reference genome (GRCh38/hg38) (Sanders et al. 2016; Catacchio et al. 2018), and therefore, the orientation of these three regions is the opposite of what appears by Strand-seq.

We also identified four regions (Chr11_inv4, Chr12_inv2, Chr12_inv10, and Chr21_inv6) that appeared inverted by Strand-seq but are reported as assembly errors in the human reference genome (GRCh38/hg38) (Sanders et al. 2016; Vicente-Salvador et al. 2017; Audano et al. 2019; Chaisson et al. 2019). Also, here the orientation of the region in macaque should be the opposite of what is shown by Strand-seq, and thus, these are not real inversions between human and macaque genomes but are an artifact because the reads are mapped against the human reference genome.

Next, we investigated 39 regions previously reported to be errors in the macaque BCM Mmul_8.0.1/rheMac8 release by Catacchio and colleagues (Catacchio et al. 2018) and

confirmed all previously reported errors and that all the regions were corrected in the latest Mmul_10/rheMac10 release.

3.4.3 Validation of inversions in macaque

To validate our novel 327 inversions, we tested 16 by fluorescence in situ hybridization (FISH) in the same *Macaca mulatta* individual (MMU1) for which Strand-seq data were generated. To also define if an inversion was polymorphic, 14 inversions were tested on two additional macaque individuals (*M. mulatta,* MMU2; *Macaca fascicularis*, MFA63). Owing to technical limitations, we were able to test only regions >500 kbp and for which the SD content was not an impediment for the FISH probe selection. In particular, 15 out of 16 inversions were tested by three-color FISH, whereas one, >2 Mbp, was tested by metaphase two-color FISH. Two of these were performed to refine the inversion BPs; for 10 regions, FISH experiments confirmed the inverted orientation in homozygous state in all macaque individuals, whereas for Chr10_inv9, one out of three individuals was found to carry the inversion in the heterozygous state. Moreover, for three inversions (Chr15_inv1, Chr15_inv3, and Chr16_inv3), the orientation could not be determined in all tested individuals. Testing multiple cell lines should allow investigation of the polymorphic nature of the inversions. However, we only tested three macaque individuals (six chromosomes), and therefore, we were unable to define if an inversion was polymorphic for allele frequencies <16.6%.

To further validate inversions not amenable to FISH, we performed BAC-end sequence (BES) paired mapping of a *M. mulatta* BAC library (CHORI-250) against the human reference genome. We expect BACs spanning inversion BPs discovered in macaque to be “discordant” when mapped to the human reference genome sequence with their ends mapping farther apart than expected in an incorrect orientation (Mario Ventura et al. 2007; Antonacci et al. 2009; Catacchio et al. 2018; Maggiolini et al. 2019). Seventy-six inversions detected as homozygous by Strand-seq had support from macaque discordant BAC clones spanning at least one BP. Ten

inversions identified as heterozygous by Strand-seq had just concordant (four inversions) or discordant (six inversions) clones spanning the inversion BPs, whereas one homozygous inversion had both concordant and discordant clones, suggesting that the macaque for which BAC ends are available might be heterozygous. Finally, just once BES paired mapping was inconsistent with Strand-seq data, because only one concordant clone was identified at the BPs of a homozygous inversion by Strand-seq. As a more direct means of validation, we selected 13 of these BAC clones for complete sequencing with Illumina as previously described (Tuzun et al. 2005). BAC-insert sequencing was 100% concordant with BES mapping and Strand-seq results.

Next, we selected 14 regions, without SDs at the BPs and ranging in size from 2 to 54 kbp, for polymerase chain reaction (PCR). In all tested regions, both orientations were detected in different species, and macaque was inverted as suggested by Strand-seq.

By combining different validation methods, we tested a total of 104 out of 327 novel inversions. Among these, 35 correspond to inversions within larger inversions, which appeared as direct by Strand-seq. After validation, the number of polymorphic inversions in macaque increased from 16 to 19. All validations, except for one, supported the inversion state identified by Strand-seq. Moreover, by intersecting data from previously published inversions (n = 5) and experimental analyses (n = 26), we confirmed the inverted status of 31 out of 87 nested inversions, which appeared to be in direct orientation by Strand-seq but were mapping within a larger region in an inverted orientation. Eighty-three of these represent cases of simple nested inversions without BP reuse. With the aforementioned analyses, the total number of inversions changed from 373 to 375. Through these efforts, we compiled a highly curated call-set of inversions that distinguishes macaque and humans, which we used for further analysis.

3.4.4 FISH analyses of complex inversions with BP reuse

Several FISH experiments were performed to better resolve the organization of complex regions. FISH analysis of a 2.7-Mbp inversion (BP2-BP3 inversion) on chromosome 10 allowed

us to refine the BPs of a previously known 36-Mbp inversion (BP4-BP5 inversion). BP2-BP3 and BP4-BP5 inversions were detected by Strand-seq as two inversions separated by a 940-kbp region in direct orientation between BP3 and BP4. Initial analyses of a BP2-BP3 inversion were performed using a reference probe mapping outside of the inversion and within the direct distal BP3-BP4 region. FISH experiments showed that this probe (blue) in macaque maps several megabases apart than expected, suggesting that the region detected in direct orientation by Strand-seq (BP3-BP4 region) is internal to the large 36-Mbp inversion (BP4-BP5 inversion). Consequently, the BP3-BP4 inversion appears to be direct by Strand-seq because it is nested within a larger inversion. Further experimental validations allowed us to define that the proximal BP of the 36-Mbp inversion is not BP4 but is BP3 (~900 kbp upstream than previously reported) (Catacchio et al. 2018).

Notably, the BP2-BP3 inversion is also flanked proximally by a ~900-kbp region between BP1 and BP2, which was previously reported to be inverted between human and macaque and is still polymorphic in human (Mario Ventura et al. 2007). The BP1-BP2 inversion, however, appeared to be in direct orientation by Strand-seq, leading us to hypothesize that this could be another example of an inversion within an inversion. Further investigation of this region confirmed that the BP1-BP2 inversion is nested within a larger one (BP1-BP3).

In total, we identified and validated four cases (Chr2_inv14, Chr9_inv14, Chr10_inv8, and Chr10_inv9) of nested inversions with BP reuse. In all four cases, SDs are flanking recurrently inverted regions. Moreover, we identified three inversions (Chr7_inv11, Chr7_inv13, and Chr10_inv7) for which the reused BPs are shared with adjacent inversions and not with the larger inversions that include them.

3.4.5 Nested inversions analyses

To assess the statistical significance of the observed nested inversions (n = 87), we conducted 100,000 simulations of the 375 observed inversions. First, we shuffled the observed inversion

coordinates (*n* = 375) across the entire GRCh38/hg38 at random, using BEDTools (v2.28.0) (Quinlan and Hall 2010), excluding assembly gaps and centromeres. Second, we limited our shuffling to occur only in the space between inter-chromosomal SD pairs of ≥98% sequence identity, accounting for the biological bias of inversion occurrence between high-identity SD pairs. We observed a trend toward significance for the enrichment in nested inversions. When the shuffling of the 375 inversion coordinates is restricted to the space between inter-chromosomal SD pairs, the enrichment is no longer significant, suggesting that nested inversions are likely driven by highly identical SD pairs. As expected, we observe that the number of nested inversions depends on the size of the inversion they are nested in. We further noticed that chromosomes 11 and 7 have a comparably high number of nested inversions, considering that there are two large inversions within these chromosomes.

3.4.6 Evolutionary analyses

To resolve the evolutionary history of the inversions detected by Strand-seq, we first took advantage of published data from previous studies (Catacchio et al. 2018) to establish the lineage specificity of 41 inversions. For 33 inversions, experimental analyses were performed; the remaining four regions were validated using a combination of both experimental and literature. Specifically, we tested nine regions >500 kbp by FISH in multiple primate cell lines, including two chimpanzees (*Pan troglodytes*), two gorillas (*Gorilla gorilla*), two orangutans (*Pongo pygmaeus*), and three macaques (two *M. mulatta* and one *M. fascicularis*); we used marmoset (*Callithrix jacchus*) as outgroup when necessary (Figure 3.2A).

We also tested 14 regions by PCR in the same species: Four are human specific; one occurred in the human and chimpanzee ancestor and another three in the human, chimpanzee, and gorilla ancestor; and one is macaque specific. Finally, for all the inversions detected by Strand-seq, we checked the BES paired mapping profiling from primate BAC and fosmid clones (CHORI-251, CHORI-277, CHORI-276, CHORI-250, CHORI-259, and CHORI-1277) as

previously described (Antonacci et al. 2009; Catacchio et al. 2018; Maggiolini et al. 2019; Sanders et al. 2016; Kronenberg et al. 2018; Giner-Delgado et al. 2019). We selected a total of 405 clones (257 concordant and 148 discordant) spanning the BPs of 176 putative inversions, and among these, 26 clones were fully sequenced with Illumina (Figure 3.2B).

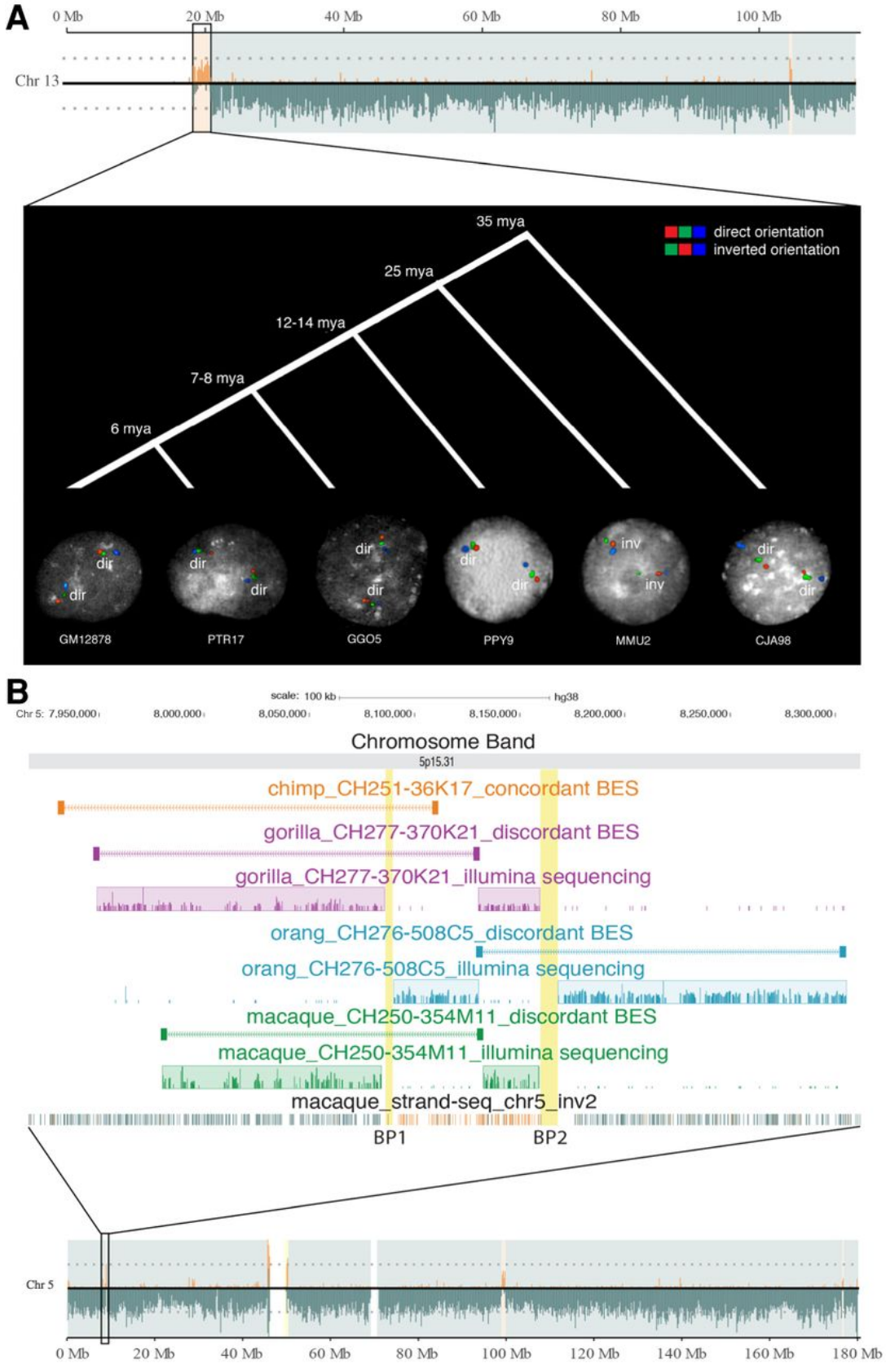

A
0 Mb
20 Mb
40 Mb
60 Mb
80 Mb
100 Mb
Chr 13
35 mya
25 mya
12-14 mya
7-8 mya
6 mya
direct orientation
inverted orientation
dir
inv
GM12878
PTR17
GGO5
PPY9
MMU2
CJA98
B
scale: 100 kb
hg38
Chr 5: 7,950,000
8,000,000
8,050,000
8,100,000
8,150,000
8,200,000
8,250,000
8,300,000
Chromosome Band
5p15.31
chimp_CH251-36K17_concordant BES
gorilla_CH277-370K21_discordant BES
gorilla_CH277-370K21_illumina sequencing
orang_CH276-508C5_discordant BES
orang_CH276-508C5_illumina sequencing
macaque_CH250-354M11_discordant BES
macaque_CH250-354M11_illumina sequencing
macaque_strand-seq_chr5_inv2
BP1
BP2
Chr 5
0 Mb
20 Mb
40 Mb
60 Mb
80 Mb
100 Mb
120 Mb
140 Mb
160 Mb
180 Mb

Figure 3.2: Evolutionary history of two inversions. (A) Strand-seq view of chromosome 13 shows the switch in orientation of a 2-Mbp region, suggesting the presence of an inversion (Chr13_inv1). The region was tested using FISH in interphase nuclei in multiple primate species and was inverted just in macaque, whereas all the other primates are in direct orientation similar to human. (HSA) Homo sapiens; (PTR) Pan troglodytes; (GGO) Gorilla gorilla; (PPY) Pongo pygmaeus; (MMU) Macaca mulatta; (CJA) Callithrix jacchus. **(B)** Strand-seq view of a 89-kbp inversion (Chr5_inv2) between BP1 and BP2 is shown. BES mapping and Illumina sequencing of primate clones show that the region is inverted in gorilla, orangutan, and macaque and is direct in chimpanzee.

In total, we reconstructed the evolutionary history of 78 out of 375 regions. Twelve (15.4%) of these are human specific; 10 (12.8%) occurred in the human–chimpanzee common ancestor; 23 (29.5%) occurred in the African great apes ancestor, although three of these show the direct orientation in chimpanzee (which means that either the region in chimpanzee flipped back to the direct orientation or it represents a case of incomplete lineage sorting); eight (10.3%) occurred in the great ape ancestor; and 25 (32.1%) in the macaque lineage (Figure 3.3A).

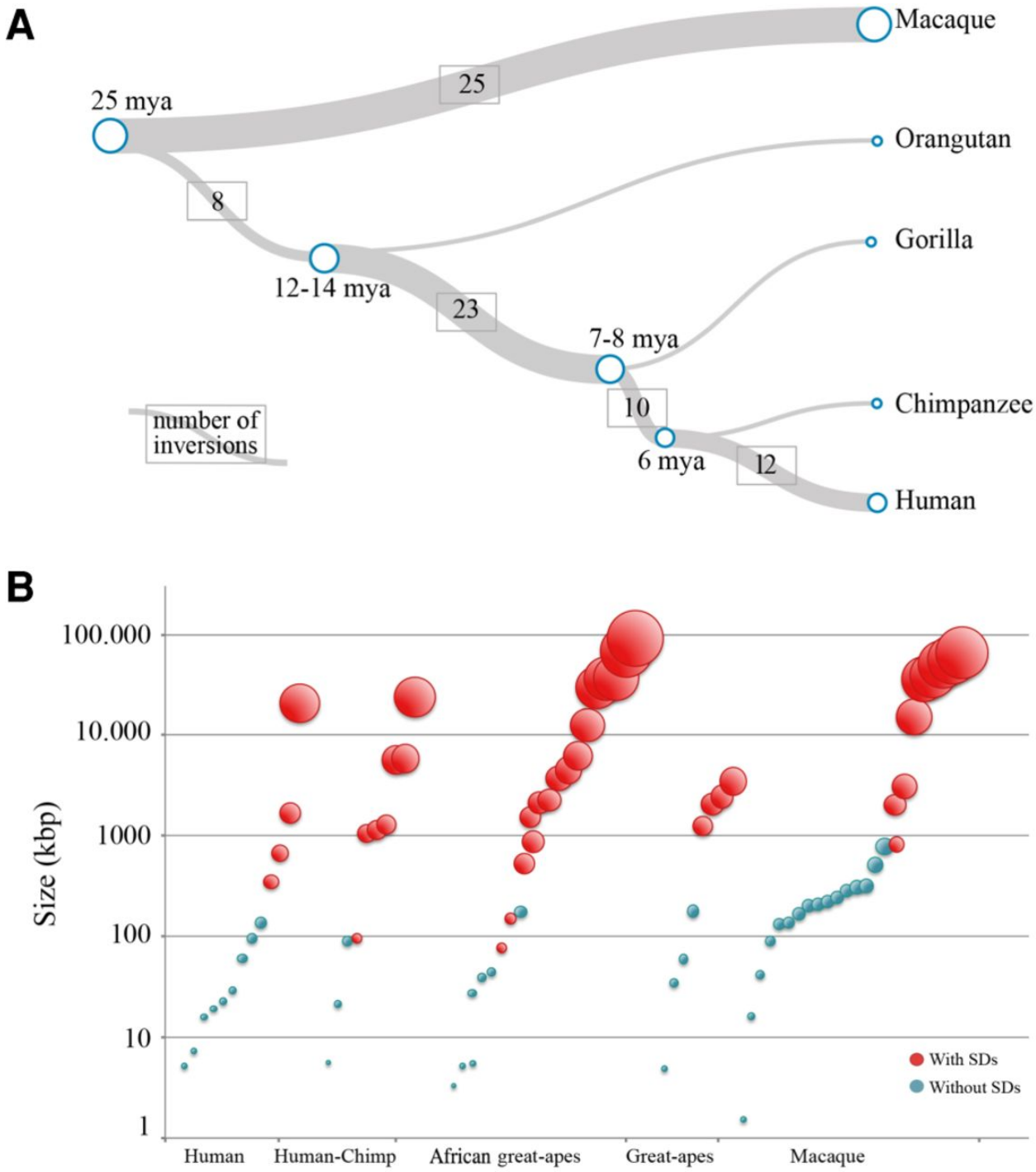


Figure 3.3: Evolutionary history and segmental duplication (SD) architecture of inverted region. **(A)** All inversions for which the evolutionary history has been determined are mapped on a phylogenetic tree in which the branch thickness is proportional to the number of inversions. **(B)** Inversions for which the lineage specificity has been determined are shown. The figure highlights the correlation between the presence of SDs at the inversion BPs and the size of the inversions.

To gather more information regarding the lineage specificity of the inversions, we used inversion calls generated for great apes, also from Strand-seq data (Porubsky, Sanders, Höps, et al. 2020), and net alignments for the most recent releases of genome assemblies of two New World monkey outgroup species (*C. jacchus*, calJac3; *Saimiri boliviensis*, saiBol1). This revealed that 49.1% (184/375) of the inversions occurred in the Old World monkeys, whereas 43.5% (163/375) are specific to Hominidae. We were not able to define the lineage specificity in only 28 cases (7.5%) as it was not possible to test the inversions in other species because the region structure makes validation difficult.

We compared the inversions identified between human and macaque genomes that are flanked by SDs, with the inversion list recently reported for other great ape genomes (Porubsky, Sanders, Höps, et al. 2020). We identified 51 (65%) inversions that are inverted in at least one of the great ape genomes, which we identify as candidate nonallelic homologous recombination (NAHR)–mediated inversions that may undergo recurrent rearrangements in primate genomes.

3.4.7 Analysis of genomic features

Because inversions can directly act on genes via direct breaking of structure and separation of promoters from *cis*-acting regulatory elements, we searched for human RefSeq genes mapping at the inversion BPs. We found that, of the 375 inversions, 156 have human genes spanning at least one BP. In particular, by removing duplicates, we detected 861 genes from the RefSeq curated subset overlapping with our inversion BPs. By considering these genes, we performed a Gene ontology analysis applying the ToppFun default parameters on the ToppGene portal and found matches for 855 out of 861 genes. Gene ontology analysis showed a high percentage of defensins, genes involved in the response to bacteria, and an enrichment of the golgin family members. We also asked whether inversions were less likely to fall on annotated genes and found that protein-coding, but not all, genes were significantly depleted at inversion BPs (p = 0.001, permutation analysis). Within 100 kbp of BPs, no such depletion was observed (Figure

S3.1A–D).

We tested for enrichment of protein-coding genes in all inversions identified between human and macaque and did not see any significant enrichment (*P*-value = 0.198, *Z*-score −0.82). However, we hypothesized that inversions predicted to be formed by NAHR and potentially undergoing recurrent rearrangement could show an effect on gene content owing to selective pressures acting at these inversions throughout evolution. To address this, we performed an enrichment analysis by focusing on the 51 inversions we identified as potential NAHR candidates, and tested for enrichment of protein-coding genes annotated in the human reference assembly (GRCh38/hg38). This revealed that NAHR-candidate inversions show an enrichment of protein-coding genes (*p*-value 0.031, *Z*-score 1.758) in support of the hypothesis.

At least five of the inversions validated by PCR overlap genes: Two inversions overlap protein-coding genes, and the other three overlap lncRNA genes. A detailed analysis of the human and macaque genome sequences (GRCh38/hg38 and rheMac10) shows that Chr17_inv12 and Chr21_inv5 inversions disrupt alternative transcripts of the protein-coding genes *CCDC40* and *ERG*, respectively, which cannot be generated from the macaque genome, although the remaining transcripts would be unaffected by the inversion. Also, four lncRNA genes are truncated by three inversions. Inversion Chr17_inv13 exchanges in macaque the last exons of two lncRNAs, leaving the first exons outside, so if these transcripts exist in macaque, they would be chimeric. In another macaque inversion (Chr4_inv14), all transcripts of lncRNA *LINC01094*, expressed in brain and placenta in humans, are disrupted by removing the first exon. Finally, lncRNA *LINC00605*, expressed in the human testis, is disrupted by inversion Chr14_inv16 in macaque and marmoset. In this last case, the generation of the inverted allele would have created the lncRNA, which would not exist in the ancestral inverted allele in its current human form.

We also performed a pairwise d_N/d_S analysis between macaque (Mmul_10/rheMac10) and human (GRCh38/hg38) in order to investigate what kind of evolutionary forces shaped the inversions. We created two orthologous gene sets, one including pairwise orthologs between macaque and human in inverted regions and the other including pairwise orthologs between macaque and human in noninverted regions. We used PAML (Z. Yang 1997) to calculate the d_N/d_S value of all orthologs and found that the d_N/d_S distribution of genes in the inverted regions is not significantly different from the d_N/d_S distribution of genes in noninverted regions (Wilcoxon rank-sum test, *P*-value = 0.5201).

Because duplications play a crucial role in the origin of inversions, we analyzed the SD content at inversion BPs. We found that 77 out of 375 inversions (20.5%) have SDs mapping at their BPs, whereas if we consider just inversions >300 kbp, 91.7% (55/60) have SDs at their BPs. We also investigated the link between SD regions and the lineage specificity of inversions and found that 29.4% of Hominidae-specific inversions have SDs at the BPs, whereas only 8.7% of the macaque-specific inversions have SDs. Of note, when filtering for inversions >300 kbp, the percentage of regions flanked by SDs increased to 100% for Hominidae-specific inversions and to 69% for Old World monkey inversions. When considering regions >1 Mbp, the percentage of Old World monkey–specific inversions flanked by SDs goes up to 89% (Figure 3.3B).

We compared our inversions with the UCSC ClinVar (Catacchio et al. 2018; Newman et al. 2005), development delay (Landrum et al. 2016), and ClinGen CNVs (Cooper et al. 2011; Coe et al. 2014) tracks and found 19 pathogenic CNVs overlapping inverted regions between human and macaque (Table 3.1).

Table 3.1: Inversions associated with human disease

Inversion	Coordinates (GRCh38/hg38)	Lineage specificity	Size	Disease association	References	Core duplicon/ Gene family
Chr1_inv5	Chr 1: 146,046,099–149,795,840	Hominidae	3.749.741	1q21.1-q21.2 deletion and duplication	Landrum et al. (2016); Coe et al. (2014); Cooper et al. (2011)	
Chr3_inv20*	Chr 3: 195,615,426–197,667,189	Hominidae	2.051.763	3q29 deletion and duplication	Kaminsky et al. (2011); Miller et al. (2010); Landrum et al. (2016); Coe et al. (2014); Cooper et al. (2011)	
Chr4_inv1*	Chr 4: 3,878,546–9,800,237	Hominidae	5.921.691	Recurrent t(4;8) (p16;p23) translocation	Giglio et al. (2002)	*ZNF705B/ZNF705G*
Chr7_inv13*	Chr 7: 72,519,724–74,982,331	Hominidae	2.462.607	7q11 Williams–Beuren syndrome	Osborne et al. (2001); Schubert (2009); Kaminsky et al. (2011); Miller et al. (2010); Landrum et al. (2016); Coe et al. (2014); Cooper et al. (2011)	*PMS2P7*
Chr8_inv2*	Chr 8: 7,058,306–12,722,555	Hominidae	5.664.249	8p23.1 deletion and duplication; inv dup(8p)	Coe et al. (2019); Kaminsky et al. (2011); Miller et al. (2010); Landrum et al. (2016) and Giglio et al. (2001)	
Chr10_inv7	Chr 10: 46,561,417–50,201,968	Hominidae	2.738.868	10q11.22-10q11.23 deletion and duplication	Kaminsky et al. (2011); Miller et al. (2010); Landrum et al. (2016); Coe et al. (2014); Cooper et al. (2011)	
Chr10_inv8*	Chr 10: 46,561,417–47,500,010	Hominidae	1.683.815	10q11 deletion and duplication	Landrum et al. (2016); Coe et al. (2014); Cooper et al. (2011)	
Chr15_inv1	Chr 15: 19,807,467–28,869,865	Hominidae	9.062.398	15q11.2-q13.1 deletion and duplication	Kaminsky et al. (2011); Miller et al. (2010); Landrum et al. (2016); Coe et al. (2014); Cooper et al. (2011)	
Chr15_inv3	Chr 15: 28,075,295–32,649,443	ND	4.574.148	15q13.1-q13.3 deletion and duplication	Kaminsky et al. (2011); Miller et al. (2010); Landrum et al. (2016); Coe et al. (2014); Cooper et al. (2011); Antonacci et al. (2014)	
Chr15_inv9*	Chr 15: 82,315,420–84,598,237	Hominidae	2.282.817	15q25.2 deletion	Cooper et al. (2011); Palumbo et al. (2012); Maggiolini et al. (2019)	*GOLGA2P7/ GOLGA6L4/ GOLGA6L5P*
Chr15_inv10	Chr 15: 84,147,736–85,279,737	Hominidae	1.132.001	15q25.2-q25.3 deletion	Kaminsky et al. (2011); Miller et al. (2010); Coe et al. (2014); Cooper et al. (2011)	*GOLGA2P10/ GOLGA6L10/ GOLGA6L17P/ GOLGA6L9*
Chr16_inv4*	Chr 16: 21,342,884–21,936,253	Hominidae	593.369	16p12.1 deletion and duplication	Coe et al. (2014; Cooper et al. (2011)	*NPIPB3/NPIPB4*
Chr16_inv5*	Chr 16: 21,728,768–22,611,067	Hominidae	882.299	16p12.1 deletion and duplication	Kaminsky et al. (2011); Miller et al. (2010); Landrum et al. (2016); Coe et al. (2014); Cooper et al. (2011)	*NPIPB4/NPIPB5*
Chr16_inv6	Chr 16: 28,590,202–29,638,840	Hominidae	1.048.638	16p11.2 deletion and duplication	Kaminsky et al. (2011); Miller et al. (2010); Landrum et al. (2016); Coe et al. (2014); Cooper et al. (2011)	*NPIPB8/NPIPB9*
Chr16_inv7	Chr 16: 29,035,196–30,339,222	Hominidae	1.304.026	16p11.2 deletion and duplication	Kaminsky et al. (2011); Miller et al. (2010); Landrum et al. (2016); Coe et al. (2014); Cooper et al. (2011)	*NPIPB11/NPIPB12*
Chr17_inv7*	Chr 17: 36,150,950–38,312,655	Hominidae	2.161.705	17q12 deletion and duplication	Kaminsky et al. (2011); Miller et al. (2010); Landrum et al. (2016); Coe et al. (2014); Cooper et al. (2011)	*TBC1D3B/TBC1D3F/ TBC1D3G/ TBC1D3H/ TBC1D3I/TBC1D3J*
Chr19_inv3	Chr 19: 23,444,060–23,990,525	ND	546.465	19p12 deletion and duplication	Landrum et al. (2016); Coe et al. (2014); Cooper et al. (2011)	
Chr22_inv3	Chr 22: 20,675,878–21,563,033	Hominidae	887.155	22q11.21 deletion and duplication	Kaminsky et al. (2011); Miller et al. (2010); Landrum et al. (2016); Coe et al. (2014); Cooper et al. (2011)	
Chr22_inv5	Chr 22: 23,303,592–24,300,127	Hominidae	996.535	22q11.23 deletion	Landrum et al. 2016; Coe et al. 2014; Cooper et al. 2011	

(*) Indicates inversions found to be polymorphic in human. ND, not determined.

Previous studies identified nine out of these 19 inversions as polymorphic in human, including the chromosome 4p16.2-4p16.1 and 8p23 inversions, both of which predispose to further rearrangements leading to complex neurological disorders (Miller et al. 2010; Kaminsky et al. 2011). In a 2.5-Mbp inversion involving the 7q11 locus predisposing to the deletion associated with Williams–Beuren syndrome (S. Giglio et al. 2001; Sabrina Giglio et al. 2002; Antonacci et al. 2009); a 2-Mbp inversion predisposing to RCAD syndrome (Osborne et al. 2001; Schubert 2009); a 1.5-Mbp inversion involving the 10q11 locus (Mefford et al. 2007); two inversions at the 16p12.1 locus associated with deletion and duplication of the same loci (Catacchio et al. 2018); and the inversion of the 15q25 locus predisposing to a deletion associated with developmental delay (Table 3.1) (Miller et al. 2010; Kaminsky et al. 2011; Cooper et al. 2011; Coe et al. 2014; Landrum et al. 2016). For 17 out of 19 regions previously shown to be associated to pathogenic CNVs, we were able to define the lineage specificity of the inversions and show that the Hominidae orientation is always derived.

Because ancestral duplications, termed core duplicons, have been shown to be hotspots of genomic rearrangements, including large-scale inversion polymorphisms and recurrent CNVs associated with disease (Zody et al. 2008; Giannuzzi et al. 2013; Antonacci et al. 2014; Dennis and Eichler 2016; Nuttle et al. 2016; Maggiolini et al. 2019), we compared genes present at the inversion BPs with gene families mapping at core duplicons reported by (Jiang et al. 2007; Antonacci et al. 2014). Almost half (nine out of 19) of these regions have one of these genes mapping at the inversion BPs (Table 3.1). This is also evident in our Gene Ontology analysis, which highlighted golgin genes, a core duplicon gene family previously implicated in other complex genomic rearrangements on human Chromosome 15 (Jiang et al. 2007; Antonacci et al. 2014; Maggiolini et al. 2019), as being enriched at BPs.

3.4.8 Recombination and heterozygosity

Moreover, because inversions can influence recombination, we analyzed the suppression of recombination over the inverted regions of the genome, relative to the background recombination rates (Wilcoxon rank sum test with continuity correction, *P*-value < 2.2^{-16}). We observed a significant (<10^{-15}) suppression of recombination in the inverted regions. Also, the recombination suppression effect was particularly pronounced in the case of polymorphic inversions (0.827 × background RC), followed by the fixed inverted regions (0.952 × background RC).

In addition, we investigated whether there is a difference of heterozygosity on inversions' flanking regions. We compared the heterozygosity distributions of four types of inversions and found that the polymorphic inversions' flank regions have higher heterozygosity than random 5-kbp regions' (*P*-value 0.02617, random vs. polymorphic inversions with SDs; *P*-value 1.68×10^{-8}, random vs. polymorphic) and fixed inversions (*P*-value 0.01655, fixed inversions with SDs vs. polymorphic with SDs; *P*-value 5.40×10^{-7}, fixed vs. polymorphic). However, we did not observe heterozygosity difference between fixed inversions and random regions (*P*-value 0.944, random vs. fixed with SDs; *P*-value 0.07333, random vs. fixed).

3.4.9 Effect of inversions on gene regulation

Because inversions have the potential to reorganize genes and regulatory elements, we sought to determine whether macaque inversions impact the expression of nearby genes. By using existing RNA-seq data from human and macaque lymphoblastoid cell lines (LCLs) and primary tissues (heart, kidney, liver, and lung) (Khan et al. 2013; Blake et al. 2020), we identified interspecific differentially expressed genes (DEGs) at a 5% FDR (on average approximately 4800 DEGs per tissue). We tested inversions, inversion BPs, and inversion BPs ±100 kbp for enrichment of DEGs in LCLs and the four tissues, either including or excluding genes overlapping SDs. However, after multiple testing (Benjamini–Hochberg) correction, none of the scenarios showed significant enrichment. A few tests displayed nominally significant enrichment

with SDs excluded (uncorrected $P \leq 0.05$): LCL DEGs (within inversions and at BPs) and kidney DEGs (BPs ± 100 kbp). When including SDs, LCL DEGs were also associated with inversions. Thus, we conclude that these results are compatible with SV alteration of gene expression reported in other species, but any true signal may be difficult to discern owing to the high overall proportion of DEGs between humans and macaques.

We next searched for changes to chromatin topology to which the observed differential expression may be attributable. By using a set of topologically associated domains (TADs) from the human LCL GM12878, we defined putatively disrupted TADs as those partially overlapping inversions (i.e., excluding those entirely within or containing inversions). We counted 48 inversions intersecting 69 putatively disrupted TADs, a number significantly lower than expected by chance (permutation test; empirical $p = 0.001$) (Figure S3.1). This depletion is also observed when SDs and inversions with BPs overlapping SDs are excluded from the analysis ($p = 0.001$) (Figure S3.1F). This is consistent with TAD-altering inversions being subject to negative selection. One such inversion (Chr18_inv4) is depicted in Figure 3.4A–B, along with chromatin domains predicted from a parallel analysis of paired human and rhesus Hi-C data generated for this study from LCLs (Methods), as well as previously published data from fibroblasts (Rao et al. 2014; Darrow et al. 2016).

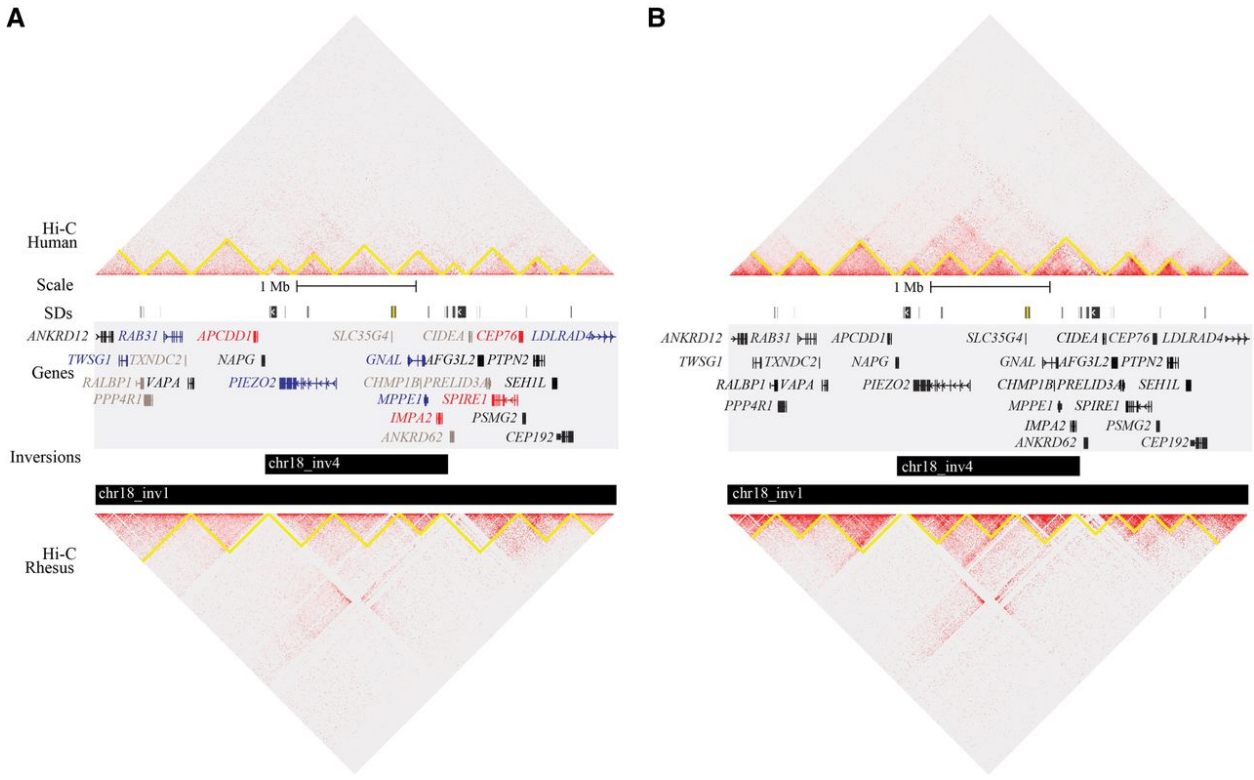


Figure 3.4: Comparison of chromatin structure and gene expression at a selected inversion (Chr18_inv4). Coordinates depicted are Chr 18: 9,140,001–13,490,000 (GRCh38). **(A)** Hi-C heatmap of human (top) and macaque (bottom) LCLs with predicted chromatin domains outlined in yellow, visualized in Juicebox. SDs are shown as colored blocks in the top track (taken from the UCSC Genome Browser). Genes are colored by differential expression: Red genes are up-regulated in macaque relative to human, blue genes are down-regulated, black genes are not differentially expressed, and gray genes were not tested. **(B)** The same locus is depicted with fibroblast Hi-C data. No differential expression analysis was conducted in fibroblasts.

Macaque-specific chromatin interactions are visible across the inversion from >1 Mbp away in the human reference, and the domain structure appears to be altered at the inversion BPs and associated SDs. In LCLs, in which the gene-expression analysis was performed, most of the genes tested within and adjacent to the inversion are DEGs (Figure 3.4B). Because of the lower sequencing depth of our Hi-Ci data, an alternative domain caller was used, which produced continuous annotations of domains, in contrast to the GM12878 TADs, which mostly contain gaps in between. We observed that BPs of large inversions (>100 kbp) often fell on domain boundaries (Figure S3.2) and confirmed significant enrichment for macaque domain

boundaries (permutation test; empirical $p < 0.03$ for LCL and fibroblast) and human fibroblast domain boundaries ($p = 0.002$ for fibroblast). As with the depletion of putatively disrupted TADs, this is suggestive of conservation of chromatin structure. Finally, as mentioned previously, many large inversions are flanked by SDs, which cannot be uniquely aligned to or may be missing from macaque. As such, identifying altered domain structure at BPs was not possible owing to missing Hi-C data (Figure S3.3).

3.5 DISCUSSION

Combining single-cell strand sequencing with cytogenetics, we created the most accurate fine-scale map of inversions between human and macaque to date. This approach was efficient in terms of time, cost, and resolution compared with high-throughput sequencing methods used to date. In total, we identified 375 inversions ranging in size from 859 bp to 92 Mbp, distributed along all the autosomes with the highest number on chromosome 2 and the lowest on chromosome 21. Despite this, considering the correlation between the size of each chromosome and the overall size of detected inversions for each of them, chromosome 7 and chromosome 3 show the highest percentage of inverted sequence (65% and 57%, respectively); indeed, these two chromosomes are two of the most rearranged among great ape and macaque genomes.

Of the 375 inversions, 48 were previously known, whereas the remaining 327 events (87.2%) are described here for the first time, increasing by eightfold the number of reported inversions between human and macaque. The vast majority (89.9%) of the 327 novel inversions are <100 kbp; this highlights the efficiency of Strand-seq in locating inversions, detecting even small events intractable by other methods.

To verify the reliability of Strand-seq, we validated a selection of 104 of the 327 novel inversions. All our results support Strand-seq data except for one case in which BES mapping

validation and Strand-seq seemed to be discordant. Yet, this does not exclude that the inversion might be polymorphic in the population, and therefore, the individual for whom the BES data are available may be in direct orientation. However, the size of the inversion and the SD content did not allow us to validate it in additional individuals with other methods.

Although our analysis showed the efficiency of Strand-seq in detecting inversions, it also highlights that Strand-seq analysis must be aware of cytogenetic rearrangements. Indeed, 23% of our inversions appeared direct by Strand-seq because they are nested inversions. Among these, the vast majority are “simple” cases of nested inversions without BP reuse and thus can be easily identified if large-scale cytogenetic inversions are already known. Although in four cases, the regions were much more complex, and several FISH experiments were necessary to resolve their genomic organization. These were cases of inversions in which BPs have been reused multiple times during evolution, contributing to their complexity. We observed that nested inversions are more likely to occur than expected by chance and that they are likely driven by highly identical SDs.

To reconstruct the lineage specificity of inversions, we tested several primate species by combining different methods and determined that 49% of the inversions occurred in Old World monkeys, whereas 44% are specific to *Hominidae*. Our analysis of the duplications at the inversion BPs suggests that NAHR mediated by SDs promoted most (89%) inversions >1 Mbp in Old World monkeys and all inversions >300 kbp in *Hominidae*. This observation is concordant with the expansion of SDs after the divergence of Hominidae from Old World monkeys and strongly suggests a link between SD expansion and the emergence of inversions. The use of Strand-seq is mandatory to resolve these structural variations as a genomic technique not hampered by SDs.

Although our knowledge on the impact of inversions on human health is limited, the strong correlation between some inversions and neurocognitive disorders is well documented. Thus, we searched for disease regions that are recurrently rearranged in humans and found 19

overlapping with our inversions, with nine being still polymorphic in humans (Table 3.1). For 17 of these regions, we were able to determine the lineage specificity. In 100% of the cases, the inversions are specific of the Hominidae, reinforcing the hypothesis that SDs played a fundamental role in generating inversions in humans and great apes that today, through their peculiar genomic structure, predispose to disease-causing rearrangements in humans.

Because SDs are frequently organized around core duplicons, we searched for their presence at the inversion BPs and found a total of 13 regions associated with cores. Among them, nine map to the BPs of inversions that overlap the aforementioned disease-associated regions (Table 3.1). Core duplicons have been previously described to be associated with the burst of SDs in the human–great ape ancestral lineage (Rao et al. 2014). This SD expansion likely set the stage for large-scale inversions to occur, ultimately leading to recurrent rearrangements associated with disease in humans.

Another interesting aspect about inversions is that they suppress recombination in heterokaryotype individuals. This results in independent genome evolution of direct and inverted arrangements and opportunities for divergence and speciation (Kirkpatrick 2010; Kirkpatrick and Barton 2006). Our results show a significant difference ($<10^{-15}$) in recombination suppression between inverted versus non-inverted regions in a size-independent way. Notably, by comparing homozygous and heterozygous inversions, we quantified how much the recombination was suppressed in fixed (~5% lower than background) versus polymorphic (~18% lower than background) inversions. This supports the role of inversions as a direct driving force in speciation because they suppress recombination when in heterozygous state. In addition, we observed a higher heterozygosity in polymorphic inversion flanking regions rather than in fixed inversions and random regions, supporting the idea that balancing selection has an important role in the maintenance of inversion polymorphisms, as previously reported (Wellenreuther and Bernatchez 2018; Mérot et al. 2020).

Because of the impact that inversions could have on the structure of genes, we searched for genes that can be altered by the presence of inversions. We identified 861 human genes overlapping with 156 inversion BPs; these include genes belonging to several groups, including members involved in the response to bacteria, genes with chemokine receptor binding activity, and golgin family members. At least five of the inversions that were validated by PCR overlap genes, with two located close to protein-coding genes and the other three in lncRNA genes. Future studies may evaluate the functional consequences of inversions on these genes in contributing to phenotypic differences among humans, great apes, and macaques.

Our assessment of the impact of inversions on gene regulation largely agrees with previous works that find structural variation alters gene expression in humans and nonhuman primates (Marques-Bonet, Girirajan, and Eichler 2009; Lazar et al. 2018). Although expression analysis was limited to a single cell type, we report an enrichment of DEGs within and nearby (<100 kbp) inversion BPs that suggests that inversions between human and macaque may have the same functional impact reported in other species. In parallel, we also report that macaque inversions tend to avoid disrupting chromatin domain structure, as is true for deletions and rearrangements in other primates (S. Giglio et al. 2001; Osborne et al. 2001; Zody et al. 2008; Stankiewicz and Lupski 2010; Lazar et al. 2018; Maggiolini et al. 2019). Chromatin domains are thought to play a role in orchestrating promoter–enhancer interactions, and their disruption is associated with pathological phenotypes in humans (Marques-Bonet and Eichler 2009). Together, these findings support a view in which inversions impacting critical genes or altering regulation are likely to be deleterious. At the same time, inversions between human and macaque are associated with differential expression of nearby genes. This study provides a list of 48 inversions that are candidates for driving rhesus-specific expression patterns, although this is by no means exhaustive given that TAD annotations vary by algorithm and that TAD alterations *per se* are not required to alter transcription.

In conclusion, our approach based on the combination of Strand-seq and cytogenetic data offered us the opportunity to create a complete and detailed map of genomic inversions between human and macaque. We identified many hotspots of genomic instability that pinpoint regions with complex rearrangement activity, likely implicated in evolutionary innovations, as well as medical conditions.

3.6 METHODS

Strand-seq detection of inversions

High-quality Strand-seq single-cell libraries (Iskow et al. 2012; Kronenberg et al. 2018) were obtained from an LCL derived from one macaque (M. mulatta, MMU1). The cells were maintained using standard culture conditions, and 40 μM of BrdU was added to the media for 23 h before sorting. Single cells were deposited in 96-well plate using the BD FACSMelody cell sorter, and Strand-seq library construction was pursued for single cells following the protocol previously described (Fudenberg and Pollard 2019; Huynh and Hormozdiari 2019). Libraries were sequenced on a NextSeq 500 (MID-mode, 75-bp paired-end protocol) and demultiplexed, and data were aligned to GRCh38/hg38 (BWA 0.7.15). Low-quality libraries, such as those with high background reads, were excluded from analysis, and 61 high-quality cells were obtained for inversion analysis. For each selected cell, only chromosomes inherited in the WW (plus-plus) or CC (minus-minus) state were considered and compiled into a directional composite file as previously described (Lupiáñez et al. 2015; Franke et al. 2016). The composite files were processed using breakpointR (v.1.2.0) (Falconer et al. 2012) to locate putative inversion BPs. To curate BPs and inversion calls, composite files were BED-formatted, uploaded to the UCSC Genome Browser, and manually inspected.

FISH analysis

Metaphases and interphase nuclei were obtained from two humans, two chimpanzees (*P. troglodytes*), two gorillas (*G. gorilla*), two orangutans (*P. pygmaeus*), three macaques (two *M. mulatta* and one *M. fascicularis*), and one marmoset (*C. jacchus*). Two-color and three-color FISH experiments were performed using human fosmid (*n* = 18) or BAC (*n* = 39) clones directly labeled by nick-translation with Cy3-dUTP (PerkinElmer), Cy5-dUTP (PerkinElmer), and fluorescein-dUTP (Enzo) as previously described (Sanders et al. 2017), with minor modifications. Briefly, 300 ng of labeled probe was used for the FISH experiments; hybridization was performed at 37°C in 2× SSC, 50% (v/v) formamide, 10% (w/v) dextran sulphate, and 3 mg sonicated salmon sperm DNA in a volume of 10 mL. Posthybridization washing was at 60°C in 0.1 × SSC (three times, high stringency, for hybridizations on human, chimpanzee, gorilla, and orangutan) or at 37°C in 2 × SSC and 42°C in 2 × SSC, 50% formamide (three times each, low stringency, for hybridizations on macaque and marmoset). Nuclei were simultaneously DAPI stained. Digital images were obtained using a Leica DMRXA2 epifluorescence microscope equipped with a cooled CCD camera (Princeton Instruments). DAPI, Cy3, Cy5, and fluorescein fluorescence signals, detected with specific filters, were recorded separately as grayscale images. Pseudocoloring and merging of images were performed using Adobe Photoshop software. For interphase three-color FISH, each region >500 kbp was interrogated using two probes within the predicted inversion and a reference probe outside. A change in the order of the probes mapping within the inversion was indicative of the presence of the inversion. For inversions >2 Mbp, two-color FISH on metaphase chromosomes was performed using two probes within the inverted region.

BES and fosmid-end sequence paired mapping

BESs of chimpanzee, gorilla, orangutan, and macaque BAC libraries (CHORI-251, CHORI-277, CHORI-276, CHORI-250, and CHORI-259) and fosmid-end sequences of gorilla fosmid library (CHORI-1277) were obtained from the NIH trace repository and mapped against the human

reference (GRCh38/hg38) following a protocol optimized by Sanders and colleagues (Sanders et al. 2016) for fosmids and adapted to BAC insert sizes as previously described (Porubsky, Sanders, Taudt, et al. 2020). BAC clones spanning regions in the same orientation as in human are concordant in size and orientation of the ends, whereas clones spanning inversion BPs are discordant because they have end pairs that are incorrectly oriented and map abnormally far apart when mapped to the human reference genome sequence. BES and fosmid-end sequence profiling of 372 BAC and 33 fosmid clones was used to study the orientation of 176 regions in different species.

Illumina sequencing of BAC clones

DNA from three CH251, five CH277, seven CH276, 13 CH250, and two CH259 BAC clones was isolated, prepped into sequencing libraries, and sequenced (PE250) on an Illumina MiSeq using a Nextera protocol (Lichter et al. 1990). DNA from one clone was barcoded before library preparation, whereas DNA from 25 clones mapping to different chromosomes and free of SDs was pooled two at a time before library preparation and then barcoded and sequenced. Sequencing data were mapped with mrsFAST (Hach et al. 2010) to the human reference genome, and singly unique nucleotide (SUN) identifiers were used to discriminate between highly identical SDs (Catacchio et al. 2018). Illumina sequencing data of the BAC clones were accessed from the NCBI BioProject database (https://www.ncbi.nlm.nih.gov/bioproject) under accession number PRJNA429373.

Polymerase chain reaction

PCR was used to test 14 inversions <54 kbp with simple BPs without large repeats. To validate inversions between different species, the first step was to identify the exact location of the inversion BPs through the NCBI “Blast2seq” tool to find the exact position range of the BPs for each species. Alignments were performed for human and chimpanzee, human and gorilla,

human and orangutan, and human and macaque. After that, we designed four different primers (A, B, C, and D) that amplify two regions for each haplotype and include the BPs so that in the direct haplotype the BP1 is inside the AB amplicon and the BP2 inside CD. The inverted haplotype instead is revealed by amplification of primers A and C and of primers B and D. In some cases, additional primers were required to detect one of the orientations owing to the presence of indels associated to the inversion. Primers were designed with “Primer 3 Plus” (http://www.bioinformatics.nl/cgi-bin/primer3plus/primer3plus.cgi) in order to amplify regions of 500–1000 bp. PCR amplification across inversion BPs was performed with genomic DNA from two humans (NA12878 and NA20528), two chimpanzees (PTR12 and N457/03), two gorillas (GGO2 and Z02/03), two orangutans (PPG9 and PPG10), two rhesus macaques (MMU1 and MMU2), one crab-eating macaque (MFA63), and one marmoset (CJA98), if needed. DNA N457/03 and Z02/03 were isolated from frontal cortex tissue samples of the Banc de Teixits Animals de Catalunya. PCR conditions were 30 sec at 94°C, 30 sec at 60°C–64°C, and 0.5–2 min at 72°C in 25-µL reactions with 100 ng of genomic DNA, 200 µM dNTPs, 10 pmol of each primer, and 1 U of Taq polymerase (Roche).

Simulations of nested inversions

In both simulation scenarios, we counted the number of times a nested inversion occurred, which was defined as an inversion that is 100% contained within another larger inversion. The null distributions from each scenario were constructed using the counts of the simulated nested inversions across 100,000 simulations. The empirical *p*-value was calculated after *Z*-score transformation using a one-tailed test, and the enrichment factor was estimated using 87/µ, where µ is the observed mean nested inversion count.

Gene ontology analysis

Genes at the inversion BPs were extracted from the curated subset of the RefSeq track from the UCSC Genome Browser. The obtained gene list has been analyzed using the ToppGene portal (Chen et al. 2009; https://toppgene.cchmc.org/), which is a one-stop portal for gene list enrichment analysis and candidate gene prioritization based on functional annotations and protein interaction networks. In particular, the ToppFun function has been used to detect functional enrichment of genes based on transcriptome, proteome, regulome (TFBS and miRNA), ontologies (GO, Pathway), phenotype (human disease and mouse phenotype), pharmacome (Drug-Gene associations), literature cocitation, and other features.

Recombination analysis

The 375 inversions were annotated as fixed (*n* = 356) and polymorphic (*n* = 19) following conversion of genomic coordinates from GRCh38/hg38 to MGSC Merged 1.0/rheMac2 using liftOver; because of large structural differences between these two genome assemblies, some of the coordinates failed to convert, resulting in 11 fixed and 214 polymorphic inversions successfully mapped onto rheMac2 space (60% liftOver success rate). All recombination data were obtained from the latest recombination estimates of the macaque genome (Xue et al. 2016).

Heterozygosity analysis

We downloaded the macaque whole-genome sequencing (WGS) population data and selected 94 Indian macaque individuals for which sequence coverage was greater than 10× (Xue et al. 2016). We used PLINK (v1.9) (Purcell et al. 2007) to calculate fixed/fixed + SD/polymorphic/polymorphic + SD 5-kbp flank regions heterozygosity. Moreover, we used BEDTools (v2.29.0) (Quinlan and Hall 2010) to randomly choose 200 5-kbp regions excluding inversions flanking regions, and we used PLINK to calculate their heterozygosity. We used a *t*-test to perform statistical analysis in R.

Differential gene expression

Gene expression was quantified in using RNA-seq data from LCLs (macaque *N* = 5 individuals; human *N* = 6) (Khan et al. 2013) and primary tissues (*N* = 4 each) (Blake et al. 2020). TrimGalore (v0.6.0; http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/) was used to trim FASTQ files using the flags "-q 20 --phred33 --length 20." Transcriptome indices were built (Salmon v1.1.0) (Sudmant et al. 2010) from species-specific sequences of all orthologous transcripts previously published (Chen et al. 2009; Zhu et al. 2014), and the most recent reference genomes (GRCh38/hg38 and Mmul_10/rheMac10) were included as decoy sequences. Transcripts per million (TPM) values were estimated using Salmon using "--validateMappings." To compute counts at the gene level for a total of 28,372 coding and noncoding genes, tximport (Khan et al. 2013) was used with the setting '"countsFromAbundance = 'lengthScaledTPM.'" A total of 15,920 genes were tested for differential expression after excluding those with less than one count per million in all samples. Length-normalized counts were passed to limma-voom (Patro et al. 2017), and each gene was fitted with a linear model accounting for species and sex. DEGs were called at a 5% FDR with no fold-change filter.

Chromatin conformation analyses

TADs were defined as a set of predictions generated from high-depth Hi-C of the human LCL GM12878 (approximately 4.9 billion Illumina reads) (Zhu et al. 2014). Coordinates of 9262/9274 TADs were converted to GRCh38 using the liftOver utility from the UCSC Genome Browser. The 5-kbp windows (resolution of the TAD-calling analysis) centered on the start and end coordinates of each TAD were considered to be TAD boundaries.

For an interspecies comparison of chromatin domain structure, we produced Hi-C libraries for LCLs of both species using a DNase-based method (Soneson, Love, and Robinson

2015b). Three human (GM12878, GM20818, GM20543, analyzed together) and one rhesus macaque (MM290-96) individual were included. Valid Hi-C contacts on the human reference (GRCh38/hg38) were produced with the Juicer pipeline (Law et al. 2014; Ritchie et al. 2015). Human alignments were down-sampled to about 200 million reads to match the number of macaque Hi-C contacts passing the MAPQ filter of 30 (BWA) (H. Li and Durbin 2009). Hi-C interaction matrices were generated using Juicer tools (MAPQ > 30, Knight–Ruiz normalization) (Durand et al. 2016) at a resolution of 50 kbp. TopDom (Shin et al. 2016) was used to identify chromatin domains with the default window size of five. The measure of concordance (MoC) as implemented by (Zufferey et al. 2018) was used to quantify similarity between domain sets on a scale of zero (no concordance) to one (identical) using Chromosome 1 matrices. Hi-C contact maps with coverage normalization and domain calls were visualized together in Juicebox (v1.11.08). Domain boundaries were defined as 50-kbp windows centered on the domain start and end coordinates and were considered to be shared between species if they intersected or were adjacent to a boundary in the other. This analysis was repeated in fibroblast cell lines using human IMR-90 (Rao et al. 2014) and macaque (Darrow et al. 2016) data (about 230 million reads). Human Hi-C data from LCLs are available under NCBI BioProject accession number PRJEB36949.

Enrichment and depletion analyses

Permutation tests were conducted to identify over- and underrepresentation of genomic features (genes and boundaries) at and within 100 kbp of inversion BPs. Inversions were shuffled (BEDTools v2.25.0) (Quinlan and Hall 2010) 1000 times in GRCh38/hg38, preserving the sizes and distances of BPs, and the number of features intersecting BPs was counted in each set. Empirical *P*-values were calculated as $p = (M + 1)/(N + 1)$, where M is the number of iterations yielding an equal or more extreme count than observed (greater for enrichment or fewer for depletion), and N is the number of permutations. To test BP regions for enrichment of DEGs, a

hypergeometric test was implemented to compare the ratio of DEGs at or near BPs to the overall ratio of DEGs.

3.7 DATA AVAILABILITY

The Strand-seq library sequence data generated from this study have been submitted to the NCBI BioProject (https://www.ncbi.nlm.nih.gov/bioproject) under accession number PRJNA625922. Illumina sequencing data of the BAC clones generated in this study have been submitted to BioProject database under accession numbers PRJNA627588. Rhesus Hi-C data from LCLs have been uploaded to the European Nucleotide Archive (ENA; https://www.ebi.ac.uk/ena/browser) under accession number PRJEB37908.

3.8 ACKNOWLEDGMENTS

We thank T. Brown for critical review of the manuscript, Alejandra Delprat for help with the PCR validations, and Mira Mastoras and Pietro D'Addabbo for bioinformatics support. This work was supported by “Fondi di Ateneo, University of Bari” grant (grant number CUP H92F17000190005) to F.A., by the Agencia Estatal de Investigación (AEI, Spain) and the European Regional Development Fund (FEDER, EU; grant number BFU2016-77244-R) grant to M.C., by European Research Council Consolidator Grant MOSAIC grant (grant number 773026) to J.O.K., by the National Institutes of Health, National Institute of Neurological Disorders and Stroke (R00NS083627) grant to M.Y.D., and by U.S. National Institutes of Health (NIH) Clinical Center grants (grant numbers HG002385 and HG010169) to E.E.E. E.E.E. is an investigator of the Howard Hughes Medical Institute, and M.Y.D. is a Sloan fellow (FG-2016-6814).

3.9 SUPPLEMENTARY FIGURES

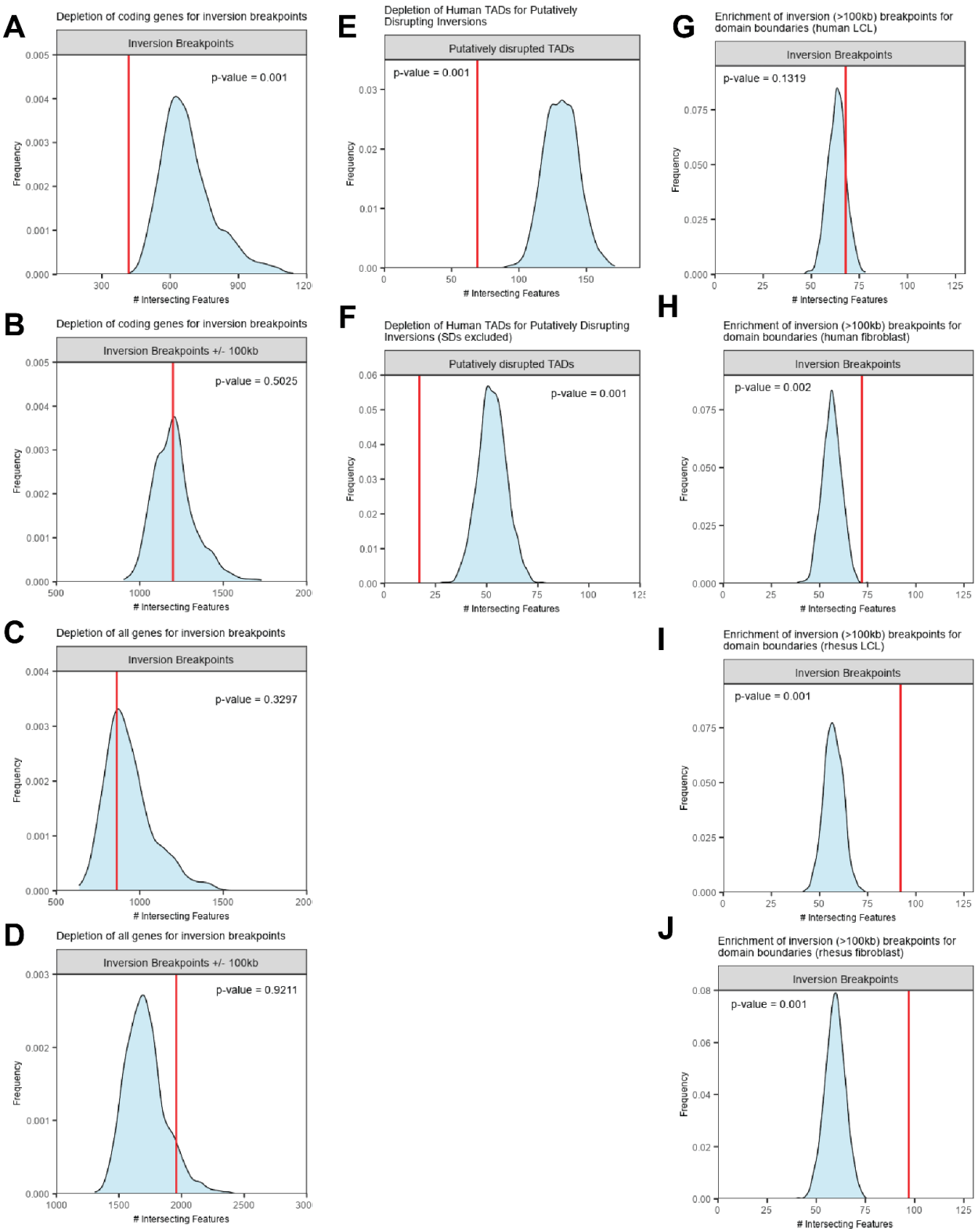


Figure S3.1: Permutation testing for enrichment/depletion of genomic features at inversion breakpoints (BPs). (A–D) The observed number of genes intersecting inversion BPs is depicted with a red line, and the distribution of genes intersecting 1,000 permuted sets of coordinates is shown in blue. The depletion test was conducted for coding genes and all genes, directly at BPs and within 100 kbp. **(E)** The observed number of putatively disrupted GM12878 TADs is marked with a red line, and the permuted distribution of putatively disrupted TADs is shown in blue. **(F)** The test for TAD disruption was repeated while excluding genes and inversions with BPs in SDs. **(G–J)** The observed number of inversions >100 kbp intersecting a domain boundary (from lower-depth Hi-C of human and rhesus) is marked with a red line, and the distribution of overlaps from 1,000 permutations is shown in blue.

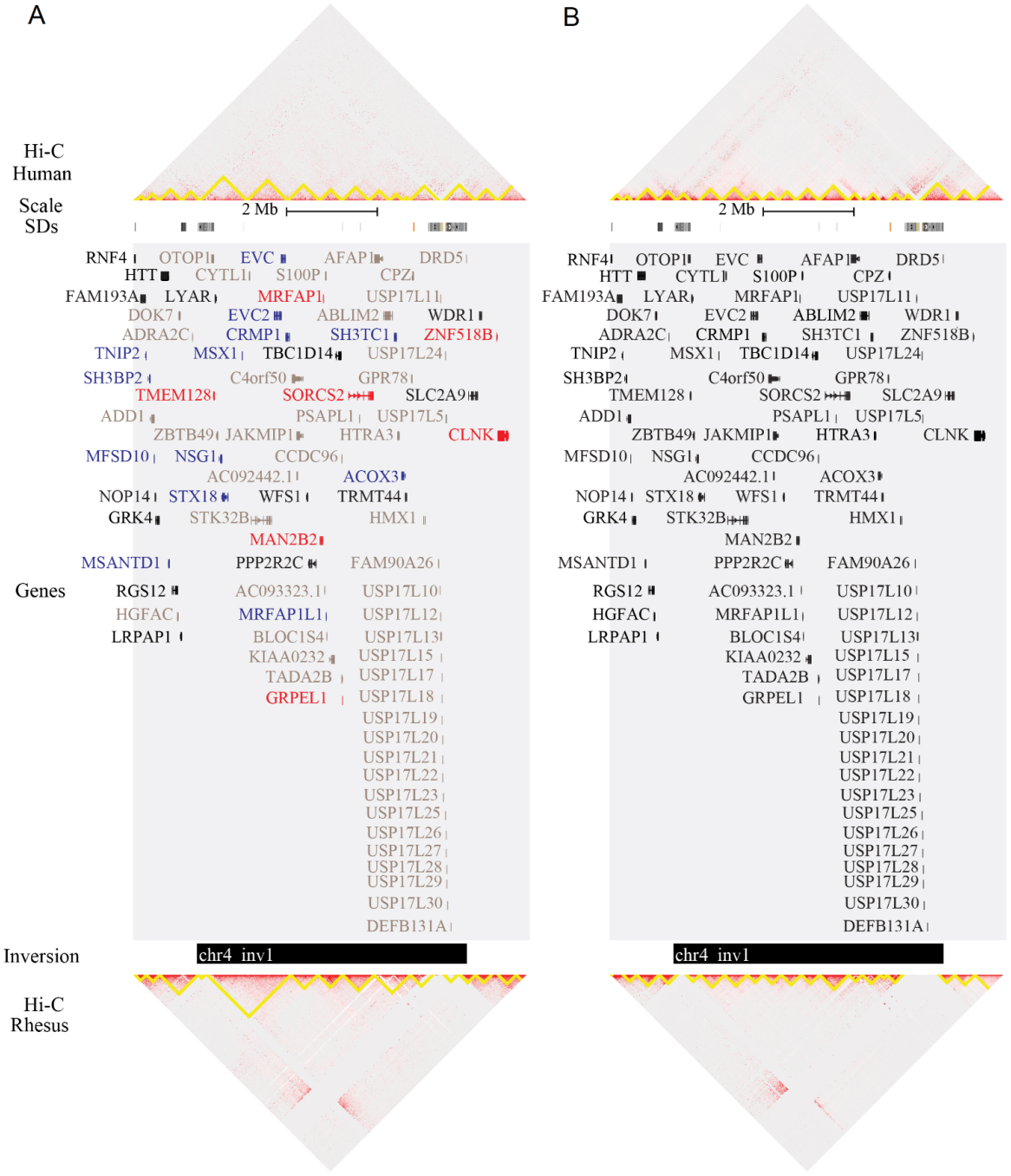
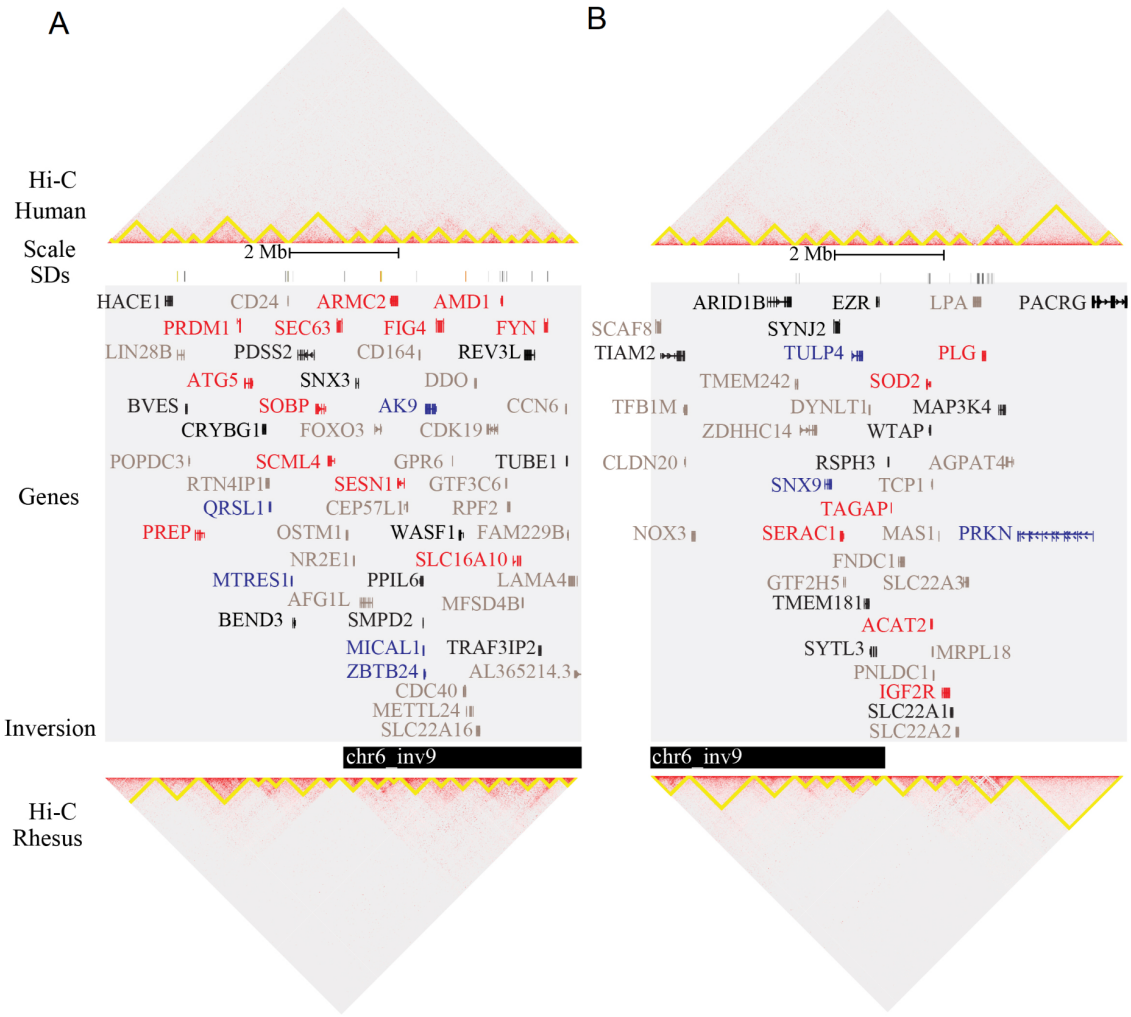


Figure S3.2: Comparison of chromatin structure and gene expression at a selected inversion (Chr4_inv1) with large segmental duplication (SD) blocks at the BPs. Coordinates depicted: chr4:2480001-11180000 (GRCh38). **(A)** Hi-C heatmap of human (top) and macaque (bottom) LCLs with predicted chromatin domains outlined in yellow, visualized in Juicebox. SDs are shown as colored blocks in the top track (taken from the UCSC Genome Browser). Genes are colored by differential expression: red genes were upregulated in macaque relative to human, blue genes were downregulated, black genes are not DE, and gray genes were not expressed. **(B)** The same locus depicted with fibroblast Hi-C data. Differential expression analysis was not conducted in fibroblasts.



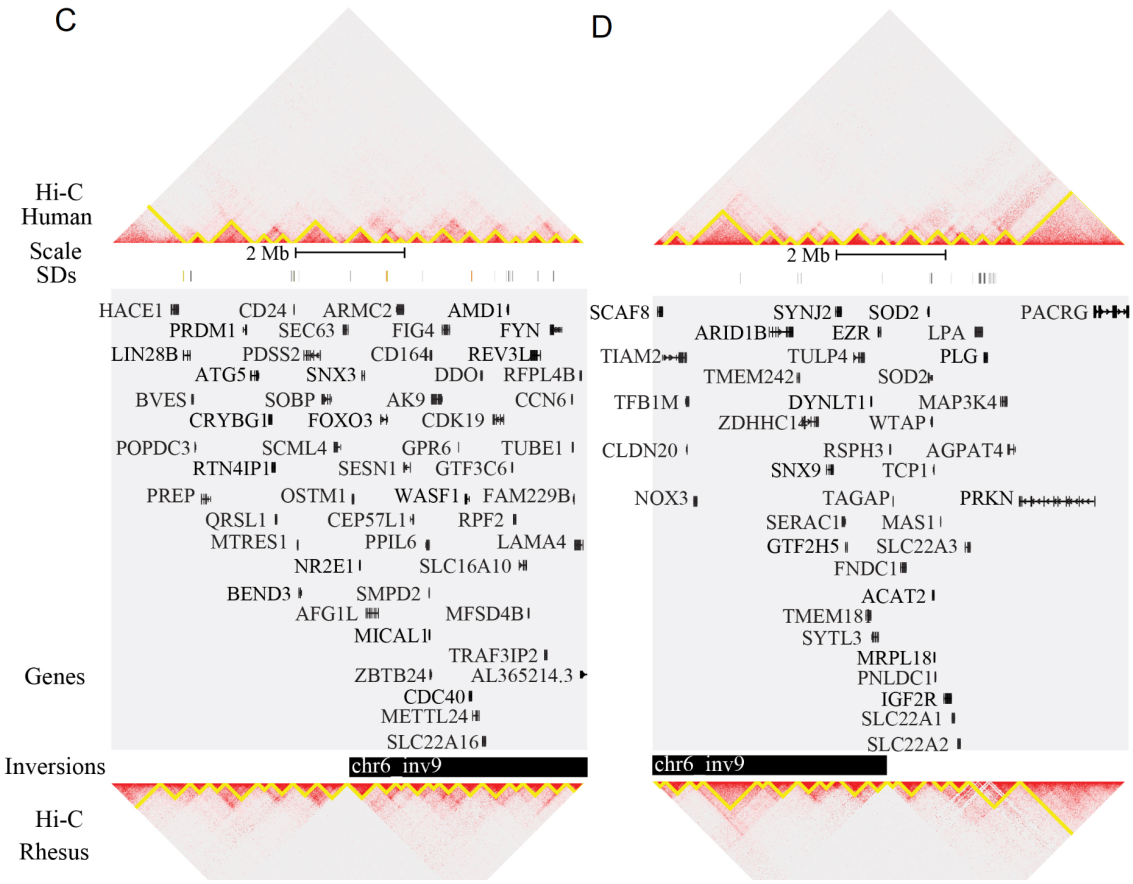


Figure S3.3: Comparison of chromatin structure and gene expression at a selected inversion (Chr6_inv9) with BPs falling on domain boundaries. Coordinates depicted: chr6:103650001-112350000 and ch6:154650000-163350000 (GRCh38). **(A–B)** Hi-C heatmap of human (top) and macaque (bottom) LCLs with predicted chromatin domains outlined in yellow, visualized in Juicebox. SDs are shown as colored blocks in the top track (taken from the UCSC Genome Browser). Genes are colored by differential expression: red genes were upregulated in macaque relative to human, blue genes were downregulated, black genes are not DE, and gray genes were not expressed. **(C–D)** The same locus depicted with fibroblast Hi-C data. Differential expression analysis was not conducted in fibroblasts.

Chapter 4:
Diverse molecular mechanisms contribute to differential expression of human duplicated genes

4.1 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. We compared cross-tissue mRNA levels of 75 HSD genes from 30 families between humans and chimpanzees and found expression patterns consistent with relaxed selection or neofunctionalization. In general, ancestral paralogs exhibited greatest expression conservation with chimpanzee orthologs, though exceptions suggest certain derived paralogs may retain or supplant ancestral functions. Concordantly, analysis of long-read isoform sequencing datasets from diverse human tissues and cell lines found that about half of derived paralogs exhibited globally lower expression. To understand mechanisms underlying these differences, we leveraged data from human lymphoblastoid cell lines (LCLs) and found no relationship between paralogous expression divergence and post-transcriptional regulation, sequence divergence, or copy number variation. Considering cis-regulation, we reanalyzed ENCODE data and recovered hundreds of previously unidentified candidate CREs in HSDs. We also generated large-insert ChIP-sequencing data for active chromatin features in an LCL to better distinguish paralogous regions. Some duplicated CREs were sufficient to drive differential reporter activity, suggesting they may contribute to

divergent cis-regulation of paralogous genes. This work provides evidence that cis-regulatory divergence contributes to novel expression patterns of recent gene duplicates in humans.

4.2 CONTRIBUTIONS

This chapter is adapted with minimal modification from the following published work:

> Colin J Shew, Paulina Carmona-Mora, Daniela C Soto, Mira Mastoras, Elizabeth Roberts, Joseph Rosas, Dhriti Jagannathan, Gulhan Kaya, Henriette O'Geen, Megan Y Dennis. 2021. Diverse Molecular Mechanisms Contribute to Differential Expression of Human Duplicated Genes." *Molecular Biology and Evolution* 38(8): 3060–3077. https://doi.org/10.1093/molbev/msab131.

C.S. and M.Y.D. conceived the study. C.S., P.C.M, E.R., J.R., D.J, G.K., and H.G.performed experiments. C.S., P.C.M., D.C.S., M.M., and M.Y.D. analyzed data. C.S. and M.Y.D. wrote and edited the manuscript.

4.3 INTRODUCTION

Gene duplication occurs universally and is considered a major source of evolutionary novelty; across eukaryotes, over 30% of genes are thought to have arisen from duplications (Zhang 2003). Although many duplicated genes rapidly become pseudogenes, some may share and maintain important ancestral functions via subfunctionalization, or gain novel functions entirely (neofunctionalization) (Lynch 2000). Expression divergence is likely integral to the survival of paralogous genes, as spatiotemporal partitioning of function places both daughter paralogs under purifying selection helping them escape pseudogenization (Rodin and Riggs 2003; Rodin et al. 2005). This may be the primary driver of duplicate gene retention, as gene regulation can be altered relatively easily while coding sequences remain intact (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 the control of regulatory elements from both paralogs (Tvrdik and Capecchi 2006). On a genome-wide scale, substantial expression

divergence has been 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 of these species, selection on gene-expression levels appears relaxed in one of the paralogs (Sandve et al. 2018). However, segmental duplications (SDs, regions defined as having >90% sequence similarity and being at least 1 kbp in size (Bailey 2002)) occur more commonly in vertebrates than whole-genome duplications and concomitantly generate structural rearrangements, potentially facilitating regulatory divergence and duplicate retention (Rodin et al. 2005). Although comparative 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), younger, human-specific duplications have yet to be analyzed in this light. Further, no molecular explanations have been provided for the observed expression changes between paralogs.

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 the split of the human and chimpanzee lineages, contain genes 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 assemblies 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 gene families containing >80 paralogs uniquely duplicated and existing in >90% of humans. Most derived HSD genes encode putatively functional proteins and exhibit

divergent expression patterns relative to ancestral paralogs across numerous primary tissues, despite HSDs being nearly identical by sequence (on average ~99.5%) (Dennis et al. 2017). Although there are examples of HSD genes exapting novel promoters and exons at the site of insertion (Dougherty et al. 2017), this cannot explain expression divergence that exists among whole-gene duplications. Differential regulation may be intertwined with associations of species-specific active chromatin modifications at SD loci (Giannuzzi et al. 2014) but historical reference errors and computational challenges in short-read mapping to highly-similar sequences has resulted in poorly annotated epigenetic information at duplicated loci (Chung et al. 2011; Ebbert et al. 2019).

In this study, we characterized patterns of regulatory divergence observed for HSD genes between humans and chimpanzees by quantifying cross-tissue conservation of orthologous gene expression. We found that even the youngest of duplicate genes have diverged in expression and, by comparing expression divergence between ancestral and derived paralogs, have begun to infer changes to HSD gene function. We leveraged genomic and epigenomic data from hundreds of human lymphoblastoid cell lines (LCLs) to identify differentially expressed (DE) ancestral-derived gene pairs and examined potential molecular contributors to paralogous expression divergence, including copy-number (CN) variation, post-transcriptional regulation, and cis-regulatory changes. Finally, we surveyed the active chromatin “landscape” for HSDs by reanalyzing ENCODE histone modification chromatin immunoprecipitation sequence (ChIP-seq) data, produced a novel “longer-read” ChIP-seq dataset to improve the unique alignment rate in SDs, and functionally validated candidate cis-regulatory elements (cCREs) 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.

4.4 RESULTS

4.4.1 Conservation of HSD gene expression following duplication

To assess the evolutionary trajectory of recent human duplicated genes, we quantified expression of 75 HSD genes from 30 gene families for which high-confidence sequences were available (Dennis et al. 2017) (Table S4.1). The SDs comprising these genes duplicated in an interspersed manner, often hundreds of kilobases away from the ancestral locus, with two of the 30 gene families residing on separate chromosomes. Each HSD gene family corresponded to a single-copy chimpanzee ortholog and multiple (2–4) human paralogs. If known, we classified the human paralog syntenic with the chimpanzee gene as ancestral and the human-specific paralog(s) as derived (Figure 4.1A, Table S4.1). To interpret the evolutionary fate of these genes, we compared expression of HSD paralogs (individual or summed) to chimpanzee orthologs using mRNA-sequencing (RNA-seq) data from three cell lines and four primary tissues (Khan et al. 2013; Pavlovic et al. 2018; Marchetto et al. 2019; Blake et al. 2020) using a lightweight mapping approach that shows high accuracy for paralogous genes (Soneson et al. 2015; Patro et al. 2017). Derived HSD paralogs tended to exhibit 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; $p = 0.028$, Fisher's Exact Test) (Figure 4.1B, Table S2). Altogether, these results suggest that ancestral genes tend to retain their expression patterns, while derived paralogs diverge and typically lose expression.

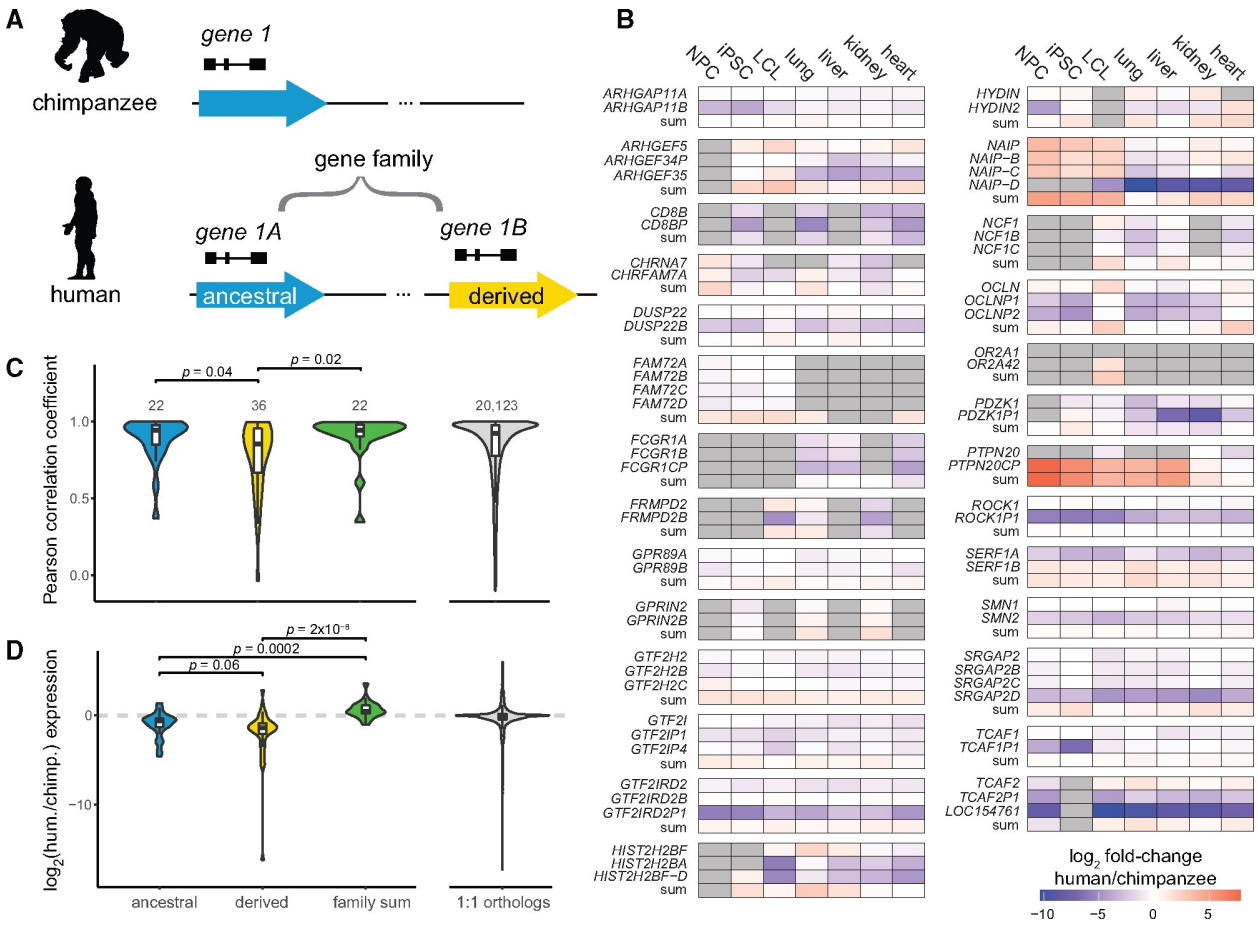


Figure 4.1: Expression patterns of HSD genes between species. (A) Illustration of genes residing within HSDs; the ancestral paralog (blue) corresponds to the chimpanzee ortholog, while derived paralogs (yellow) are human-specific. The ancestral and derived genes comprise a gene family. **(B)** HSD gene expression differences between humans and chimpanzees in three cell lines and four primary tissues. Cells are colored by the log2-fold change of human versus chimpanzee expression. Gray cells indicate nonexpressed genes. Note, PTPN20CP is expressed many fold higher than PTPN20 and the chimpanzee ortholog, but both paralogs are lowly expressed (<2 TPM) in most samples assayed. **(C, D)** Comparison of human gene expression with chimpanzee orthologs. Violin and box plots represent cross-tissue expression correlations (C) and relative expression levels (log2 ratio of human (hum.) versus chimpanzee (chimp.) expression, averaged across all cell and tissue types; Figure S4.1, Supplementary Material online) (D). HSD genes of known evolutionary status were classified as ancestral (blue) or derived (yellow) and compared with the aggregated gene family expression (green). P-values were calculated from Dunn's test following a Kruskal–Wallis test. Expression correlations of one-to-one orthologs are visualized for reference.

We next considered expression correlation across the four tissue types and three cell lines as a proxy for expression conservation between human genes and their chimpanzee orthologs. Our expectation was that in the case of subfunctionalization, the summed expression of all HSD paralogs would correlate best with chimpanzee expression, while all individual

paralogs would be less correlated; and in the cases of pseudogenization or neofunctionalization, a single paralog would exhibit high correlation with chimpanzee expression (Braasch et al. 2016; Sandve et al. 2018). We found that derived HSD paralogs exhibited significantly lower expression conservation than ancestral paralogs or summed expression, which were statistically equivalent (Kruskal-Wallis test followed by Dunn's test, Benjamini-Hochberg adjusted $p < 0.05$; Figure 4.1C). This pattern is broadly consistent with maintenance of the ancestral paralog and divergence of expression patterns of the others via relaxed selection or neofunctionalization. Further, the most conserved gene in each family was usually the ancestral paralog (14/22 of known status, $p < 0.001$, hypergeometric test). Nevertheless, eight derived paralogs showed strongest conservation of expression with chimpanzee orthologs and represent candidates for supplanting functions of their ancestral gene. For example, *SERF1B* exhibited higher expression correlation with chimpanzee than the ancestral *SERF1A* (Pearson's r of 0.81 and 0.74, respectively), while *SERF1A* expression was reduced relative to chimpanzee in lung, LCLs, and iPSCs. A few gene families (such as *CD8B*, *GTF2IRD2*, and *NAIP*) displayed expression patterns consistent with subfunctionalization, as their summed expression correlated better with that of chimpanzee than any individual paralog; however, in these cases the difference was small (difference in Pearson's r <0.05 between sum and most correlated paralog). We next considered relative expression levels between species and found that across tissues, ancestral paralogs trended toward higher expression than derived paralogs (Kruskal-Wallis test followed by Dunn's test, Benjamini-Hochberg adjusted $p = 0.058$; Figures 4.1D and S4.1A). As expected, summed HSD paralog expression was significantly higher than ancestral or derived paralogs alone. Finally, we calculated the tissue specificity index τ (Yanai et al. 2005) for HSD genes and one-to-one orthologs and found no significant differences between ancestral and derived genes (Figure S4.1B). Taken together, our analyses provide little evidence for subfunctionalization of HSD genes and are consistent with derived paralogs experiencing relaxed selection.

These results are concordant with our previous finding that derived paralogs globally show a reduction of expression relative to ancestral paralogs, with some exceptions, across diverse human tissues and cell lines from the Genotype-Tissue Expression project (Dennis et al. 2017). To validate this with a strict alignment-based approach, we used long-read PacBio isoform sequencing (Iso-Seq) data, which maps to paralogous loci with higher confidence, from a panel of 24 human biosamples and cell lines (Encyclopedia of DNA Elements (ENCODE) project). From this, we again found globally reduced expression of derived paralogs: 21/41 derived genes were expressed at a level below their ancestral paralog, while two derived genes were higher ($p < 0.05$, Wilcoxon Signed-Rank test with Benjamini-Hochberg correction; Figure S4.2). Though results should be interpreted cautiously given the low read depth and small number of replicates for each biosample, we also observed certain derived paralogs exhibit greater expression than the ancestral paralog in individual tissues or cell types; one compelling example was diverged expression of *ARHGAP11B* in excitatory neurons, which matches published findings related to the novel function of this gene in the human cortex (Florio et al. 2015; Kalebic et al. 2018; Heide et al. 2020).

4.4.2 Expression of HSD 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) and found high concordance with expression estimates from Iso-Seq data from the LCL GM12878 (Pearson's r=0.94 for 72 genes common to both analyses). We determined that over half (43/75) of HSD paralogs were expressed above 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 derived paralog *DUSP22B* is missing from the human reference

(GRCh38) (Dennis et al. 2017). Comparing expression profiles within gene families, 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, with a median TPM difference greater than two-fold in 20 of these. As was found across other cells/tissues, in most cases (25/31) the derived gene had lower expression, which we confirmed for three highly expressed gene families with RT-qPCR and Iso-Seq data (Figures 4.2A and S4.3). We noted that some paralogs exhibited high or low clustered values for derived to ancestral expression ratios, caused by lack of expression of one of the genes in a subset of individuals. This could not be reconciled as copy number (CN) or population of origin differences (Figure S4), sex, or technical effects due to sequencing depth or sequencing facility (data not shown). Altogether, these results indicate that paralogous HSD genes show divergent expression patterns in LCLs across hundreds of diverse samples.

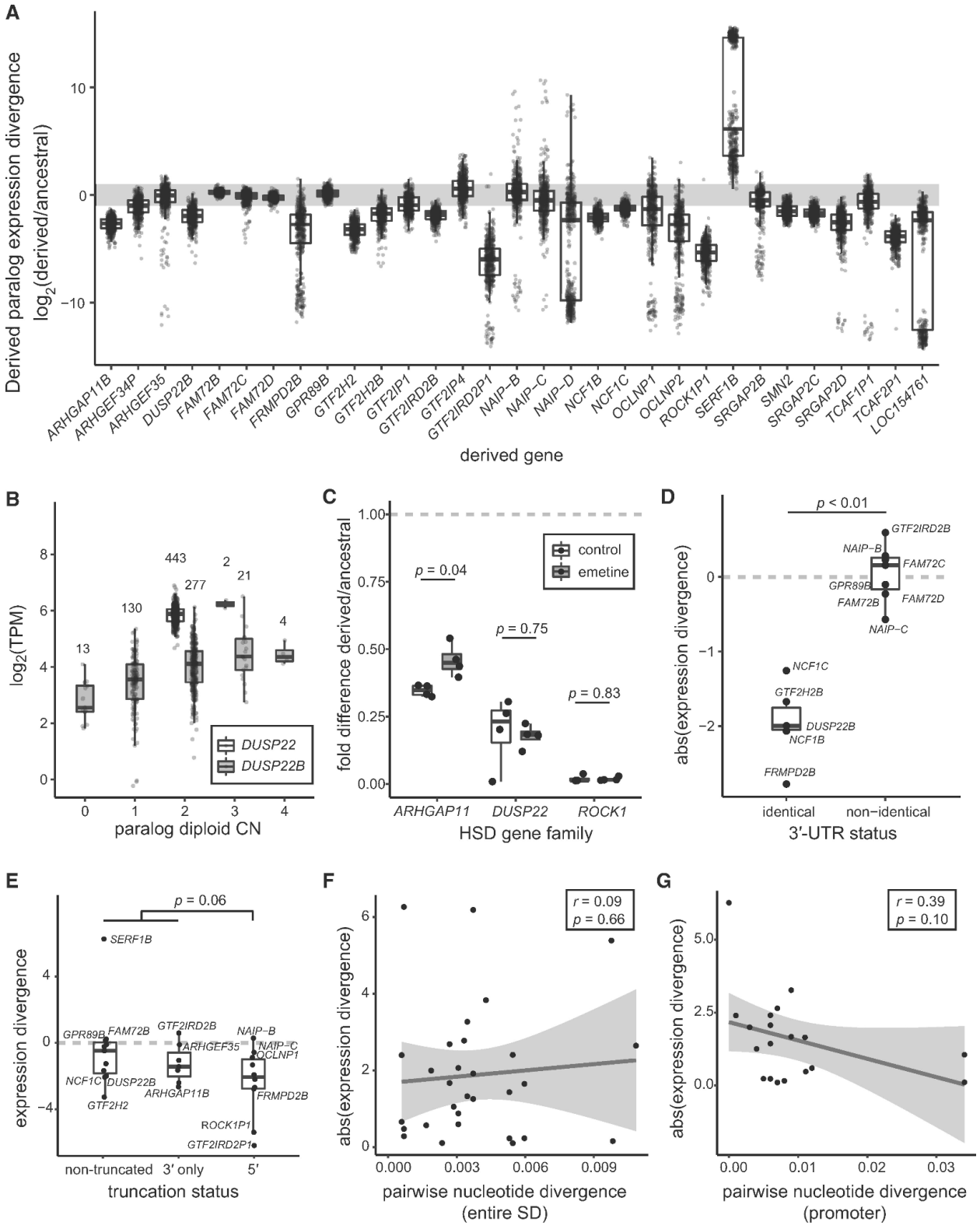


Figure 4.2: Differential expression of HSD genes in human LCLs. (A) Expression divergence of derived genes is plotted as the log2 ratio of median derived and ancestral expression for families with at least one LCL-expressed paralog. Each point represents an LCL from the Geuvadis consortium (total

N = 445) (Lappalainen et al. 2013). The gray bar indicates a 2-fold expression difference. **(B)** Expression values of ancestral DUSP22 (white) and derived DUSP22B (gray), stratified by CN. The number of individuals represented in each CN category is denoted over each boxplot. **(C)** Derived/ancestral fold-differences in expression determined from paralog-specific qPCR in control (white) and NMD-inhibited (gray) LCLs (N = 4). Statistical significance in panels C–E was assessed with a Wilcoxon signed-rank test. **(D)** Absolute value of expression divergence of ancestral-derived gene pairs, stratified by identical or nonidentical 3′ UTRs. **(E)** Comparison of expression divergence across truncation status for all expressed ancestral-derived gene pairs. **(F, G)** Scatterplot of the absolute value of expression divergence versus pairwise nucleotide identity for all expressed ancestral-derived gene pairs for whole duplicons (F) and promoters (G). Regression lines (black) and 95% confidence intervals are shown, along with the Pearson correlation coefficient (*r*) and significance of the regression slope (*p*).

4.4.3 CN variation and HSD expression

While the genes in this study were chosen for being nearly fixed in 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. Understanding that CN variation can alter gene expression levels (Stranger et al. 2007), we sought to characterize its impact 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, indicating that higher CN often but not always results in increased expression. Notably, derived genes tended to have higher CN (averaging 1.2-fold higher than ancestral over all individuals), but lower expression overall. We next used linear regression to remove the effect of CN from these comparisons 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). For example, while expression of *DUSP22B* was significantly associated with CN, these effects were insufficient to explain DE relative to *DUSP22* (Figure 4.2B). Thus, while CN differences alter the mRNA abundance of HSD paralogs, they do not provide an explanation for overall DE of these genes.

4.4.4 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 differentially subject 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 among identified DE genes. We also assessed directly if the ratio of derived to ancestral expression changed for each HSD gene family between NMD-deficient LCLs and controls and found no significant differences, though sample sizes were likely limiting (Figure S4.5). This result was largely recapitulated by paralog-specific RT-qPCR for three DE HSD genes families (*ARHGAP11*, *DUSP22*, and *ROCK1*) in four LCLs treated with the 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 4.2C). *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). Although miRNA binding sites were nearly identical between paralogs, we unexpectedly observed significantly greater expression divergence between paralogs with identical 3′ UTRs (N=5) from those that differed (N=7) (Wilcoxon signed-rank test $p < 0.01$, Figure 2D). While these data cannot rule out a role

for miRNAs in HSD transcriptional regulation, this mechanism does not explain observed differential expression of expressed gene families with identical 3′ UTRs, such as *DUSP22* and *NCF1*.

4.4.5 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 = 0.057, t-test; Figure 4.2E), in concordance with previous findings (Dougherty et al. 2018). While we may have limited power to detect differences with our small number of genes, these results hint that promoter activity is an important determinant of 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 4.2F–G).

Given that the vast majority of paralog-specific variants (PSVs) distinguishing HSDs are unlikely to be functional, we used publicly available chromatin immunoprecipitation sequencing (ChIP-seq) datasets from the 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 for which a wealth of functional genomic data exists (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 randomly sampled regions of equivalent

size (empirical p = 0.001, N=1000 replicates; Figure 4.3A, 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, with reads originating from these regions often discarded when mapping to multiple locations of the genome.

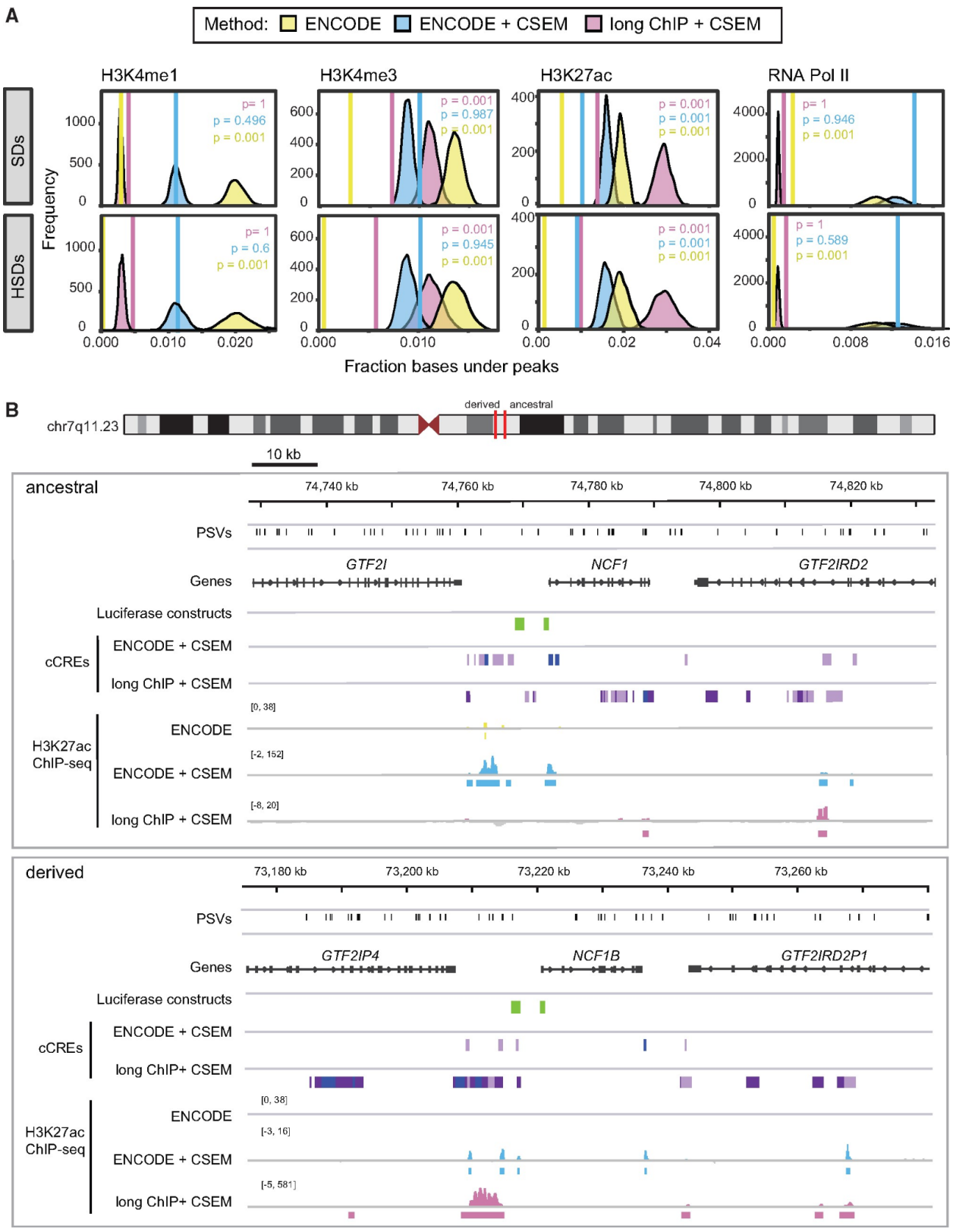


Figure 4.3: Depletion and recovery of ChIP peaks in SDs. (A) The fraction of bases covered by peaks (solid vertical line) was computed in SDs (top) and HSDs (SDs >98% sequence identity, bottom) for three

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 large-insert ChIP-sequencing ("long ChIP") data from this publication (magenta). 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 in each graph. **(B)** Chromatin landscape at the chromosome 7q11.23 HSD locus. The ancestral locus (top) and one of its derived loci (bottom) are shown with PSVs (black), genes (gray), and luciferase-tested regions (green). cCREs were identified with an 8-state ChromHMM model of GM12878 H3K4me3, H3K4me1, and H3K27ac data from multimapping reanalysis of ENCODE and long ChIP data after CSEM allocation (enhancer states in light and dark purple and promoter states in blue). H3K27ac ChIP-seq data (signal and peak calls) are also shown in yellow, blue, and magenta for published ENCODE, reanalyzed ENCODE + CSEM, and long ChIP + CSEM, respectively.

To recover this missing information, we implemented a pipeline that allowed reads to align to multiple locations in the genome and then, using CSEM (Chung et al. 2011), iteratively weighted alignments based on the nearby unique mapping rate. 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 4.3A, in blue). The peaks we discovered largely overlapped with the ENCODE peaks, though RNA PolII had a large proportion of peaks unique to our multi-mapping analysis (Figure S4.6). Using this new dataset, we observed greater enrichment of H3K27ac at the ancestral *DUSP22* versus *DUSP22B*, which we verified at three PSVs using ChIP-qPCR (1.1–2.9-fold difference of ChIP signal; 1.1–2.9-fold difference of dCt values) (Figure S4.7A). We also noted a correlation of *DUSP22*/*DUSP22B* expression divergence (Pickrell et al. 2010) (Table S4) and differential H2K27ac enrichment at two of these PSVs (Figure S4.7B). These findings suggest that reanalysis of ChIP-seq data can accurately identify enriched regions at HSD loci, uncovering potentially divergent regulatory environments.

4.4.6 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 S4.8A), allowing for higher-confidence discovery of novel peaks in duplicated regions using standard single-site mapping approaches. 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 assigned reads to one position (Bowtie and CSEM, Figures 3A and S8B). Long ChIP showed increased posterior assignment probabilities with respect to the short-read ENCODE data (Figure S4.8B), and the depletion of peaks in SDs was erased for H3K4me3, H3K4me1, and PolII (Figure 4.3A, in pink). Notably, for most libraries, fewer overall peaks were identified with long ChIP versus ENCODE data, though the peaks that did exist were largely replicated (on average, 73% of long ChIP peaks corresponded to ENCODE peaks (Chikina and Troyanskaya 2012); Figure S4.9). Long ChIP peaks tended to be larger with 2.4–3.7 times as many bases per peak except H3K4me1, which had slightly smaller peaks.

4.4.7 Identification of cCREs

To identify putatively functional cis-regulatory regions within HSDs, we integrated our reanalyzed ENCODE and long ChIP data into two 8-state chromHMM models (Ernst and Kellis 2012), from which we identified active promoter- and enhancer-like states that we considered to be cCREs (Figure S4.10). This generated a novel set of cCREs in SDs, as virtually no information is available in the current ENCODE release for these loci (Figure 4.3B). Because derived gene expression is broadly lower than ancestral, we quantified the proportion of cCREs covering HSDs in 100-kbp windows and observed no significant differences between ancestral and derived loci (defined in Dennis, 2017) (Wilcoxon rank-sum test; Figure S4.11A–B). We also observed no differences in the fraction of bases covered between ancestral and derived regions in individual ChIP-seq datasets: H3K27ac, H3K4me3, H3K27ac (data not shown), and heterochromatic H3K27me3 domains (Figure S4.11C; see Materials and Methods). 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. For instance, we found that genes whose transcription start site overlapped a cCRE, H3K4me3 peak, or H3K27ac peak had significantly higher expression than those that did not, while the presence of H3K27me3 domains showed the opposite effect ($p < 0.05$, Wilcoxon rank-sum test) (Figure S4.12). We also examined 5′-truncated paralogs, which have lost their ancestral promoters. The transcription start sites of the expressed genes *GTF2IP1* and *GTF2IP4* lie outside of the duplication block and overlapped active promoters (long ChIP cCREs). Other expressed 5′ truncations also show some evidence of active promoters; for example, the *NAIP-B* transcription start site is paralogous to an internal exon of the ancestral *NAIP* and overlaps an H3K4me3 peak not found on *NAIP* (ENCODE multimapping). Overall, we identified differences in the presence or absence of cCREs at paralogous sequences, suggesting a more nuanced approach is necessary in pinpointing mechanisms contributing to paralogous expression differences.

4.4.8 Impact of non-duplicated regions on HSD gene regulation

HSDs are often transposed many thousands of kilobases from their ancestral loci, and in some cases to different chromosomes. As such, we sought to understand if cCREs outside of our duplicated regions might contribute to paralog-specific regulatory patterns. To do this, we considered physical contacts generated by chromatin looping of HSD promoters with cCREs outside of HSD regions. 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 352 and 26 promoter-interacting regions, respectively (mean size ~5 kbp). We found 59 ENCODE multi-mapping and 106 long ChIP cCREs interacting with an HSD gene promoter. For instance, a chromatin loop connects the *ARHGAP11A* promoter with a cCRE overlapping its non-duplicated 3′-UTR (Figure 4.4A). The majority (>90%) of promoter-interacting regions reside

outside of HSDs, in part due to limitations of Hi-C analysis across duplicated loci (Zheng et al. 2019) (see 4.11 Supplementary Note). These findings indicate that proximal non-duplicated loci may play a role in regulating duplicated genes.

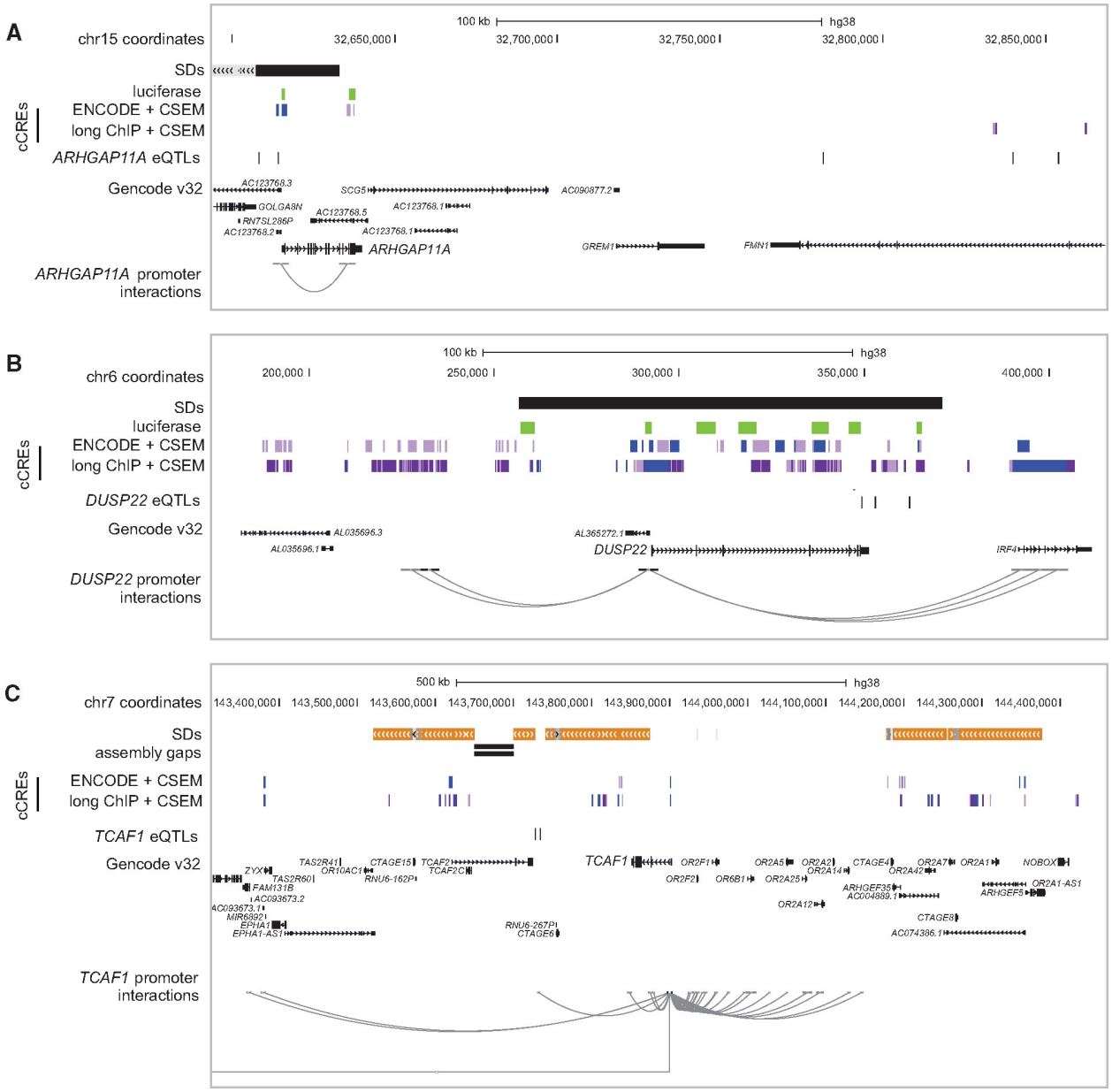


Figure 4.4: HSD gene regulation in adjacent, nonduplicated regions. Additional regulatory features were examined in the vicinity of the HSD loci, including **(A)** *ARHGAP11A* at chromosome 15q13.1, **(B)** *DUSP22* at chromosome 6p25.3, and **(C)** *TCAF1* at chromosome 7q35. In each panel, SDs are depicted as gray (>90% identical), orange (>98% identical), and black (unannotated) bars. cCREs as defined in this publication are shown in light and dark purple (active enhancer states 1 and 2) and blue (active transcription start site), with luciferase-tested regions in green. eQTLs defined in this publication and

regions previously found to interact with HSD promoters are shown for focal genes. Data were visualized in the UCSC Genome Browser (GRCh38).

We next performed expression quantitative trait locus (eQTL) mapping of HSD genes using our reanalyzed RNA-seq data and existing variant calls from the 1000 Genomes Project (N=460) (1000 Genomes Project Consortium et al. 2015). From this, we identified 40 HSD genes with significant eQTLs, an increase of 1.5- to 4-fold from published work (Lappalainen et al. 2013; Wen et al. 2015). These eQTLs consisted of 3,279 variants in 8,774 gene-variant pairs. A majority (68%) of eQTLs were located within annotated SDs, but variants identified within SDs are often unreliable (Hartasánchez et al. 2018; Ebbert et al. 2019). Accordingly, we focused on the 1,049 eQTLs in SD-proximal non-duplicated regions and found 439 of them had single-gene associations. For example, four variants were associated with *ARHGAP11A* expression (Figure 4.4A), while none were identified for *ARHGAP11B* located ~2 Mbp proximal to its ancestral locus. Similarly, four eQTLs were identified for DUSP22 on chromosome 6 (Figure 4.4B), though all were located in an SD, while 26 variants were linked with the derived paralog DUSP22B on chromosome 16. We intersected SD-proximal eQTLs with our cCREs, reasoning that functional elements would be sensitive to genetic variation and, thus, contain eQTLs. We found that five ENCODE multi-mapping and 15 long ChIP cCREs contained an HSD eQTL. Finally, 169 eQTLs fell within loci showing significant Hi-C interactions with HSD promoters (31 of these regions, total size ~160 kbp). For instance, the *TCAF1* promoter interacts with a region ~170 kbp downstream that is near two SNPs associated with *TCAF1* and *TCAF2* expression (Figure 4.4C). Altogether, these findings highlight the potential for adjacent, unique sequences to drive divergent regulation of HSDs genes.

4.4.9 Differential activity of cis-acting elements between paralogs

Using our combined datasets, we examined three HSD loci containing gene families expressed highly in LCLs (*ARHGAP11*, *NCF1*, and *DUSP22*) to identify functional changes in CREs that

may contribute to paralogous expression divergence (Figures 5, S4.13–S4.15). In all three cases, the ancestral paralog exhibited significantly greater expression compared to derived paralog(s) (Figure 4.5A). To determine if sequence differences within CREs identified from our chromHMM annotations were sufficient to drive differences in gene expression, we performed luciferase reporter assays on paralogous promoters and enhancer candidates in HeLa cells and LCLs.

ARHGAP11AARHGAP11B

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 5.5B, S4.16A, S4.17A). This was in contrast to mRNA levels in LCLs, where the ancestral *ARHGAP11A* was more highly expressed. With no CREs identified within the shared *ARHGAP11* HSD, we posited that distal elements may drive differential expression between these paralogs. We identified putative enhancers unique to each paralog outside of the shared HSD, which comprised one downstream of the *ARHGAP11A* duplicon (that was also found to interact with the promoter from our Hi-C analysis) and two downstream of *ARHGAP11B*. In HeLa cells, the *ARHGAP11A* element showed weak silencing 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 undetermined (Figures 5C, S4.16B). While these results were discordant with the mRNA expression of *ARHGAP11* paralogs in LCLs, they may help to explain the unique expression of *ARHGAP11B* in other cell types, such as cortical progenitor neurons (Florio et al. 2015).

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 differences in HeLa (0.8-fold

difference, $p < 0.001$; Figures 4A, S4.16A, S4.17A). 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 5, S4.17B). Thus, this enhancer, if targeted to *NCF1* and its paralogs, may contribute to differences in their mRNA levels.

DUSP22/DUSP22B

DUSP22 (ancestral) and *DUSP22B* (derived) promoters showed differential activity concordant with their gene expression in both HeLa and LCLs (i.e., the human ancestral paralog exhibited significantly greater activity than both the human derived and chimpanzee ortholog; ~1.5-fold difference; $p < 5\times10^{-13}$) (Figures 5.5B, S4.16A, S4.17A). We also tested six putative enhancers shared between the two paralogs in HeLa cells and found four active elements, of which two showed differential activity opposite to that of gene expression and one tracked with differential paralog expression (Figure 4.5C). We subsequently validated the latter enhancer element in LCLs (~1.4-fold difference; $p < 2\times10^{-16}$) (Figure S4.17C). From this, we concluded 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. Taken together, results from our reporter assays demonstrated that small sequence differences in HSDs can alter cis-regulatory activity.

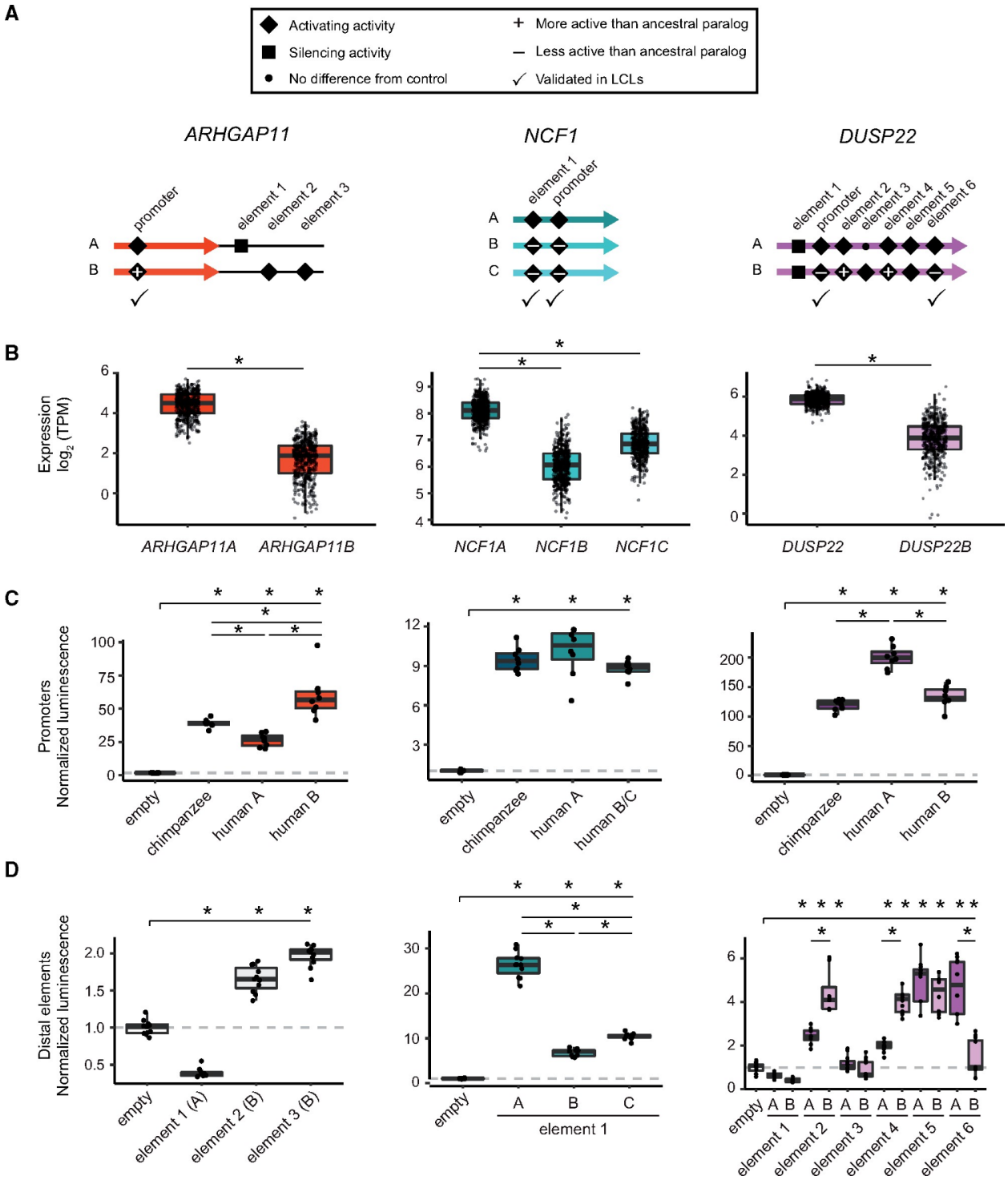


Figure 4.5: Functional characterization of cCREs in HSDs. cCREs (putative promoters and enhancers) from three HSD duplicate gene families (*ARHGAP11*, *NCF1*, and *DUSP22*) were tested in luciferase reporter assays for activity. **(A)** Cartoons indicating the relative locations of each candidate sequence within or adjacent to HSDs (thick, colored arrows). All experiments (Figures S4.16 and S4.17, Supplementary Material online) are summarized as follows: inactive sequences are shown with a small dot, activating sequences are shown with a diamond, and silencing sequences are shown with a square; differentially active derived sequences (relative to ancestral) are marked with a plus or minus sign;

elements tested (and validated) in LCLs are indicated with a check mark. **(B)** mRNA levels (TPM) for the three tested HSD gene families in human LCLs (N = 445). **(C)** 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 test following ANOVA) from the negative control is indicated along the top bar over each panel, and significant differences among homologous sequences are indicated between boxplots. **(D)** 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 test following ANOVA).

4.4.10 Putative mechanisms contributing to differential expression

In search of potential trans effectors driving differential expression of these HSD genes, we identified transcription factor binding sites within the assayed sequences. The derived *ARHGAP11B* promoter exhibited greater strength in a reporter assay versus the ancestral paralog and chimpanzee ortholog. We noted a single PSV that more than doubles the number of significant motif matches of the more active *ARHGAP11B* promoter for the transcriptional activators FLI1, GABPA, ETS1, and ELK1 (Figure S4.18). Based on chimpanzee homology, these are likely *ARHGAP11A*-specific losses, which matches its reduced activity relative to the derived paralog and chimpanzee ortholog (Figure 5). Examining predicted binding sites within *NCF1* promoters, which did not exhibit differential activity, we observed no gains or losses of any transcription factor recognition sites relative to chimpanzees. No predicted sites were unique to the most active *NCF1* enhancer, but the paralogous *NCF1B* and *NCF1C* possessed many binding sites that were missing from the ancestral, at least one of which belonged to the transcriptional repressor ZNF394. Finally, a deletion of four bases from a homopolymer repeat in the ancestral *DUSP22* promoter removes 13 similar binding sites found only in the less active *DUSP22B* and chimpanzee *DUSP22* ortholog. Some of these belonged to transcriptional repressors (ZNF394 and ZNF350), consistent with their differential transcription. Overall, these findings provide a plausible mechanism for the divergent regulatory activity of a targeted set of duplicated CREs within HSDs.

4.5 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. By comparing expression of human and chimpanzee homologs, we assayed potential mechanisms driving duplicate gene fates at relatively short evolutionary time scales (<6 million years). To simplify our comparisons of human and chimpanzee orthologs, we assayed gene families with unique duplications in the human lineage but found at single copies in other great apes. As a consequence, notable human-expanded genes such as *NOTCH2NL* (Fiddes et al. 2018; Suzuki et al. 2018), *AMY1* (Perry et al. 2007), and *TBC1D3* (Ju et al. 2016) were excluded from this study. Focusing on human LCLs, we characterized active chromatin features in HSDs and used these candidates to identify differentially active paralogous CREs. Our assessment failed to identify a universal factor responsible for the observed differential expression between paralogs, indicating the underlying molecular mechanisms vary across HSD genes families. Though this work represents an important step toward a more complete picture of HSD gene regulation, there are still some technical limitations to overcome primarily related to using short-read functional-genomic data to assess nearly-identical duplications (see Supplementary Note). Accurate long-read sequencing (e.g., PacBio HiFi) alleviates many of these issues and, as these technologies become more accessible, we will be able to more confidently identify paralog-specific regulatory features (expression, splicing, chromatin accessibility, histone modifications).

In agreement with previous analyses of whole-genome duplications in teleost fishes (Sandve, Rohlfs, and Hvidsten 2018) and focal duplications in yeast (Gu, Zhang, and Huang 2005), we found evidence for asymmetric conservation of duplicate gene expression. Specifically, human derived paralogs showed reduced and more divergent expression, recapitulating results in *Drosophila* (Assis and Bachtrog 2013). We suggest this is because

derived HSD genes may not be redundant if the full ancestral regulatory environment is missing, resulting in relaxed selection facilitating pseudogenization or neofunctionalization. This is likely for HSDs, which are interspersed throughout the human genome hundreds to thousands of kilobases from each other. As such, daughter paralogs may have been immediately removed from ancestral CREs or placed in novel regulatory environments, such as topological domains, heterochromatin, or transcriptional hubs, causing derived HSD genes to acquire new expression patterns at “birth”. This scenario is particularly likely for 5′-truncated genes. Accordingly, even very recent (<1 million years ago) duplications (Dennis et al. 2017), such as gene families *DUSP22*, *SERF1*, *SMN*, *TCAF1*, and *TCAF2*, exhibited differential expression between paralogs.

The young age of HSDs may also explain the lack of subfunctionalization observed in these data; while subfunctionalization is suggested to favor duplicate retention in the long term (Rastogi and Liberles 2005), compensatory mechanisms are not expected to arise instantaneously (Force et al. 1999). Indeed, (Lan and Pritchard 2016) concluded that in mammals neo- and subfunctionalization evolve slowly and are favored with greater genomic separation, especially for paralogs on different chromosomes. While their study discarded many of the HSD genes highlighted here, due to high sequence identity or classification as pseudogenes, our results are broadly in agreement. Meanwhile, our lack of evidence for dosage sharing stands in contrast to that of (Qian et al. 2010), who reported an inverse relationship between expression and number of paralogs in duplicates arising since the split of the human and mouse lineages, as well as the ancient split of the fission and budding yeasts (>300 million years). However, HSD genes are over an order of magnitude younger, providing a novel glimpse at gene regulation in very recent duplicates, many of which may not be retained. We again suggest that while expression changes reported here apparently arose rapidly, dosage compensation or subfunctionalization in general may take longer to evolve. Finally, we cannot discount increased dosage of functionally-redundant paralogs within a gene family as

contributing to unique human traits (Figure 4.1D), but we note that a little over half of our HSD genes represent partial duplications with likely altered protein functions, as observed for *SRGAP2C* (Charrier et al. 2012; Dennis et al. 2012) and *ARHGAP11B* (Florio et al. 2015). Thus, additive dosage effects must be considered for each gene on a case-by-case basis.

Our expression data offer some insights into the functions of previously uncharacterized HSD genes. Though our primary analysis used LCLs, a cell type not commonly associated with human-specific features such as altered brain and musculoskeletal morphology, there is evidence of immune-related differences across great apes (Barreiro et al. 2010). Further, it has been suggested that humans are more prone to autoimmune diseases than chimpanzees, particularly as a result of T- and B-cell response to viral infection (Jimenez and Piero-Velazquez 2013; Varki 2017). In our expression comparisons of chimpanzee and human orthologs, *ARHGEF35* stood out as a potentially neofunctionalized gene, as it exhibited lower cross-tissue correlation with chimpanzee, higher tissue specificity, yet globally higher expression in multiple human tissues versus its ancestral paralog *ARHGEF5* (Figure S4.2). Though little is known about its function, *ARHGEF35* encodes a truncated version of *ARHGEF5*, a Rho guanine nucleotide exchange factor (GEF) capable of activating Rho-family GTPases (Rossman, Der, and Sondek 2005) that plays a role in inflammatory response and dendritic cell migration (Z. Wang et al. 2009). We also speculate that two of our highlighted genes—*NCF1*, encoding Neutrophil Cytosolic Factor 1, and *DUSP22*, encoding a tyrosine phosphatase—may contribute to variation in protection against autoimmune response mediated by gene dosage. *NCF1* knockout causes increased T-cell activity in mice, resulting in arthritis and encephalomyelitis phenotypes (Hultqvist et al. 2004). While derived paralogs *NCF1B* and *NCF1C* are rendered nonfunctional in humans due to a frameshift mutation, in some individuals they encode the ancestral sequence as a result of interlocus gene conversion (Heyworth, Noack, and Cross 2002). Because increased *NCF1* CN is associated with reduced risk of systemic lupus erythematosus (Zhao et al. 2017), gene conversion of the derived paralogs could act to maintain

redundant, functional sequence variants (Kosuke M. Teshima 2008) with an advantageous additive effect. *DUSP22* also regulates immune response with knockout mice exhibiting enhanced T-cell proliferation, increased inflammation, and autoimmune encephalomyelitis (J.-P. Li et al. 2014). The full-length paralog *DUSP22B* is located on chromosome 16p12.1 at variable CN, but is functionally uncharacterized and missing from the human reference. No gene-disrupting mutations were identified for either paralog in hundreds of population controls (Dennis 2017) making it plausible that *DUSP22B*, which is expressed at variable dosage in humans (Figure 4.2B), is functionally redundant with *DUSP22* and could similarly play a protective role in autoimmunity. While only a proxy for function, our analysis of HSD gene expression is helpful in prioritizing genes for future assessments.

To better understand how altered CREs may contribute to paralogous expression divergence, we experimentally dissected three HSD gene families and found promoter activity was 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 pinpoint functional CREs impacted by PSVs or residing within non-duplicated regions that may differentially alter specific paralogs. We have produced and leveraged a variety of analyses to narrow down likely candidates by chromatin state, 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 reporter assays. This problem is exacerbated by the need to compare regulatory behavior across multiple cell types. To address this, massively parallel reporter assays should be employed 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 regulatory regions. We could also integrate additional types of data, such as targeted chromatin capture of CREs within SDs (such as 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. 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.

4.6 METHODS

Quantification of HSD gene expression

Iso-Seq filtered alignments were obtained from the ENCODE portal (Davis et al. 2018). Reads were counted per HSD gene with HTSeq (Anders et al. 2015) before calculating reads per kilobase of transcript, per million mapped reads values. For Figure S4.3B, *DUSP22* and *DUSP22B* reads were counted separately based on PSV-containing sequence using SAMtools mpileup. Human and chimpanzee RNA-seq data were quantified alignment-free with a custom reference transcriptomes. 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 a 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 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). See Supplementary Materials and Methods for more detailed descriptions.

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). For ancestral-derived comparisons in the human LCLs, TPM values were log-transformed using a pseudocount of 1×10-4 (an order of magnitude below the smallest nonzero value), compared with a Wilcoxon signed-rank test, and considered significant at a false discovery rate (Benjamini-Hochberg) of 5%.

CN-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, Sammeth, 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 log2-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 4.3B), *DUSP22* CN genotypes were adjusted to known values for GM12878 (as determined by fluorescence in situ hybridization in (Dennis et al. 2017)).

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 divergence and sequence divergence

Ancestral-derived paralog expression divergence was calculated as the absolute value of log2(derived/ancestral), using the median TPM values for each gene and a pseudocount 1×10-4. 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.

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% CO_2.

RNA extraction and cDNA generation

LCLs were harvested and added to an appropriate volume of TRIzol® solution (Invitrogen™) (1 ml per 107 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.

ChIP assays

ChIP assays were carried out as previously described with minor modifications (O'Geen et al. 2019) (see Supplementary Material and Methods). ChIP enrichments were confirmed by qPCR with ACTB (positive control) and HER2 (negative controls) (primers in Table S8). ChIP enrichment was calculated relative to input samples using the dCt method (dCt = Ct[HER2-ChIP]-Ct[input]). ChIP-seq libraries were prepared using the KAPA Hyper Prep Kit (Roche).

Analysis of ChIP-seq data

ChIP-seq raw data and peaks obtained with the ENCODE pipeline were directly downloaded from the online portal (Davis et al. 2018) (https://www.encodeproject.org/). All ChIP-seq analyses are available as a TrackHub for the UCSC Genome Browser (https://bioshare.bioinformatics.ucdavis.edu/bioshare/download/cpqqdfge5lfvovq/hsd_noncoding/hub.txt). Our ChIP-seq bioinformatic pipeline is freely available for use in Snakemake format (https://github.com/mydennislab/snake-chipseq), allowing the analysis to be replicated in any cell or tissue type of interest. Briefly, Illumina adapters and low quality bases (Phred score < 20) were trimmed using Trimmomatic (Bolger et al. 2014) 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. Paired-end long-ChIP reads were also mapped using paired-end BWA-MEM and filtered by MAPQ ≥ 20. After mapping, PCR duplicates and secondary alignments were removed using Picard MarkDuplicates and SAMtools v1.9, respectively. Bowtie multi-mapping reads were allocated to their most likely position using CSEM v2.4 (Chung et al. 2011) and a custom script was developed to select the alignment with the highest posterior probability. Peaks were called using MACS2 callpeak (v2.2.6) on default settings using the MACS2 shifting model (Zhang et al. 2008). Sets of peaks were compared between analysis methods using HOMER mergePeaks (parameters: “-d given”) (Heinz et al. 2010) and a unidirectional correlation metric derived from IntervalStats using peaks with an overlap p-value below 0.05 (Chikina and Troyanskaya 2012). See Supplementary Materials and Methods for more detailed descriptions.

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.

ChromHMM annotations

We generated ChromHMM (version 1.19) (Ernst and Kellis 2012) models separately for ENCODE short-read data and long ChIP after multi-mapping and CSEM allocation, using active chromatin histone modifications (H3K4me3, H3K4me1, and H3K27ac). States corresponding to active transcription start sites and active enhancers were identified manually (Ernst and Kellis 2017). In the ENCODE analysis, promoters were assigned to state 1, which corresponded to active transcription start sites, and enhancers were assigned to state 8, which corresponded to active enhancers (Figure S4.10A). Similarly, in the long ChIP analysis, promoters were assigned to state 3 and active enhancers were assigned to states 6; state 4 was considered to be an additional enhancer state lacking enrichment in H3K4me1 (Ernst and Kellis 2017) (Figure S4.10B). Together, these sets of elements were defined as cCREs.

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 S4.2). 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 CtFAM-CtVIC and ddCt as dCtcDNA-dCtgDNA from the same cell line. We calculated dCt of the input and ChIP-enriched library as CtFAM-CtVIC and ddCt as dCtChIP-dCtinput from the same cell line. For both expression and ChIP analyses, the ratio of abundance of the B to the A paralog is 2ddCt.

Luciferase reporter assays

Expression clones for luciferase assays were generated using reporter constructs pGL3-basic (Promega) for promoters and pE1B (Antonellis et al. 2008) for cCREs. Constructs were co-transfected (ThermoFisher Lipofectamine 3000) in equimolar amounts with 50 ng of the control plasmid pRL-TK (Renilla luciferase) into HeLa cells or electroporated using the Neon Transfection System for LCLs in accordance with previously published work (Tewhey et al. 2018). Luciferase assays were performed with the Dual-Luciferase Reporter Assay System (Promega E1910). Luminescence measurements were performed according to the manufacturer's instructions using a Tecan Infinite or Tecan Spark plate reader with injectors. See Supplementary Materials and Methods for more detailed descriptions.

Transcription factors binding motifs

Alignments of cloned sequences were scanned for HOmo sapiens COmprehensive MOdel COllection (HOCOMOCO) v11 (Kulakovskiy et al. 2018) transcription factor binding site 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. Transcription factor binding sites were compared across homologous sequences to identify putative paralog-specific gains and losses of binding sites.

4.7 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. In particular, we acknowledge the labs of Dr. Bradley Bernstein, Dr. Ali Mortazavi, Dr. Barbara Wold, Dr. Thomas Gingeras, and Dr. Brenton Graveley, which generated the ENCODE data used in this publication. We also thank Dr. Colin Kern for valuable advice concerning ChIP-seq analysis, Dr. Anthony Antonellis for sharing the pE1B enhancer reporter Gateway plasmid, as well as Drs. Gerald Quon, Siobhan Brady, and Torgeir Hvidsten for constructive feedback on the manuscript. This work was supported by the National Human Genome Research Institute (F31HG011205 to C.S.) and National Institute of Neurological Disorders and Stroke (R00NS083627 to M.Y.D.), and the Office of the Director and National Institute of Mental Health (DP2 OD025824 to M.Y.D.) at the National Institutes of Health (NIH). Statistical analysis advice was provided by Dr. Blythe Durbin-Johnson through the MIND Institute Intellectual and Developmental Disability Research Center, funded by the NIH 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).

4.8 DATA ACCESS

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

4.9 SUPPLEMENTARY FIGURES

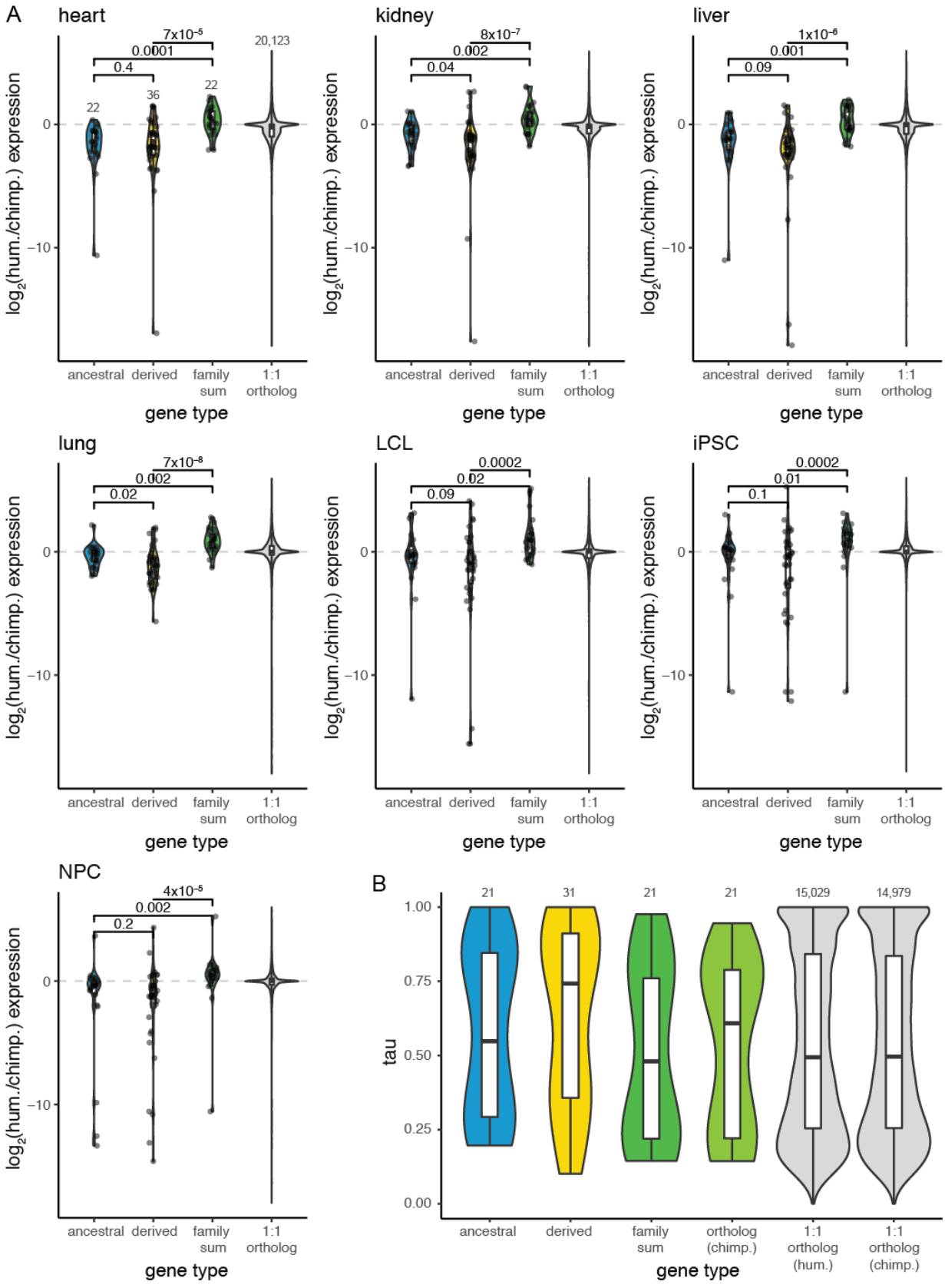

A
heart
kidney
liver
lung
LCL
iPSC
NPC
log2(hum./chimp.) expression
gene type
ancestral
derived
family sum
1:1 ortholog
7x10−5
0.0001
0.4
8x10−7
0.002
0.04
1x10−6
0.001
0.09
7x10−8
0.02
0.0002
0.01
0.1
4x10−5
0.2
22
36
20,123
B
tau
21
31
15,029
14,979
ortholog (chimp.)
1:1 ortholog (hum.)
1:1 ortholog (chimp.)

Figure S4.1. Comparison of human gene expression with chimpanzee orthologs across diverse tissues. Violin and box plots represent relative expression levels (log_2 ratio of human (hum.) versus chimpanzee (chimp.) expression **(A)** and cross-tissue expression correlations (tau) **(B)** across seven tissues/cell types. The pairwise *p*-values indicated above plots are from Dunn's test, following a Kruskal-Wallis test. Differences among tau values were not significant.

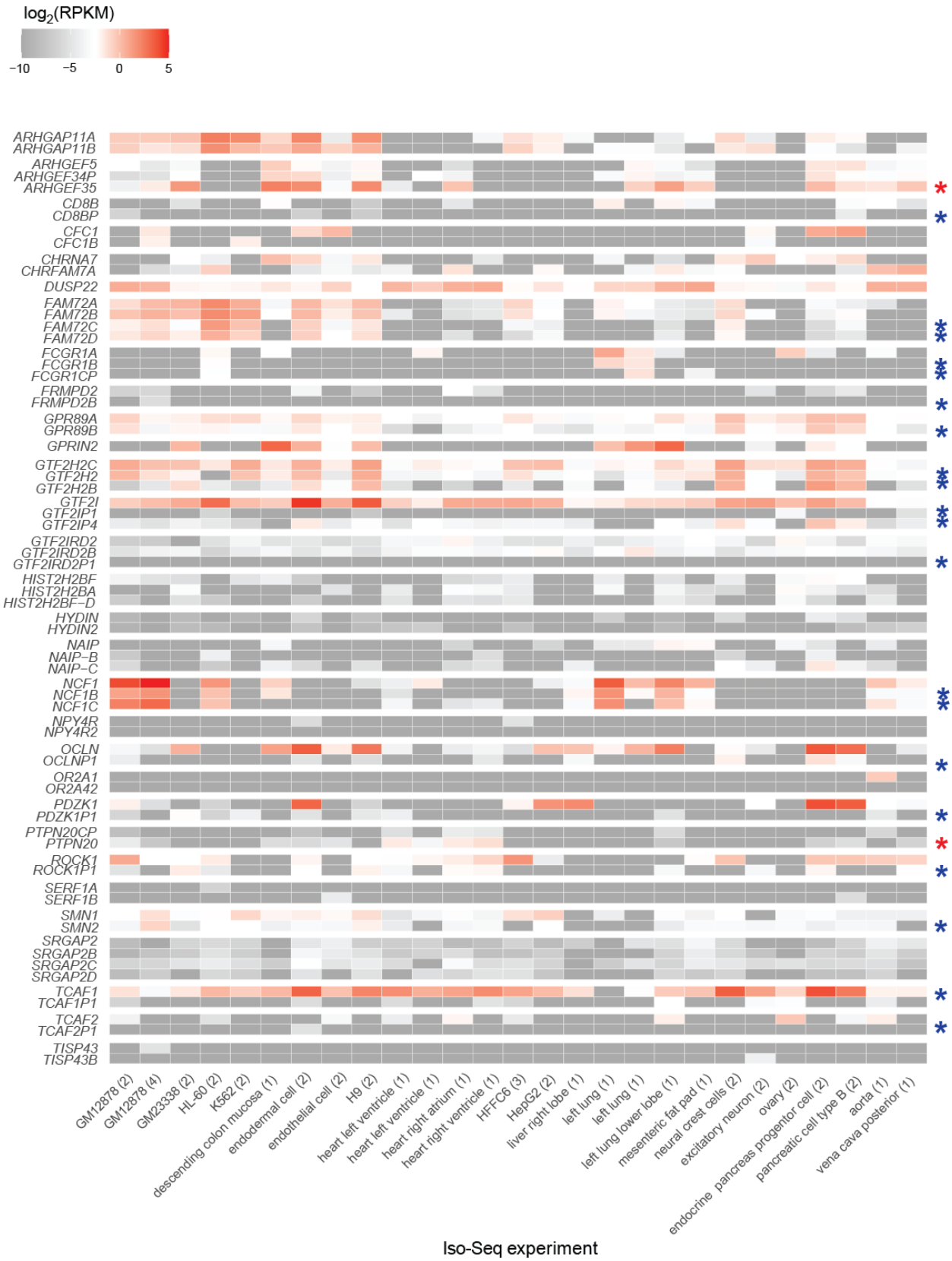


Figure S4.2. HSD gene expression in Iso-Seq datasets. For each available ENCODE Iso-Seq experiment (columns; number of technical replicates indicated in parentheses), HSD gene expression

was calculated in reads per kilobase per million mapped reads (RPKM) using only paralogous regions (i.e., excluding truncated portions and novel portions of genes). For derived genes, $\log_2$(RPKM) values were compared to the ancestral gene with a Wilcoxon signed-rank test. Significant differences (Benjamini-Hochberg adjusted $p < 0.05$) are indicated with an asterisk (blue for lower expression; red for higher). *DUSP22* and *GPRIN2* were not tested for differential expression because their derived genes are missing from GRCh38.

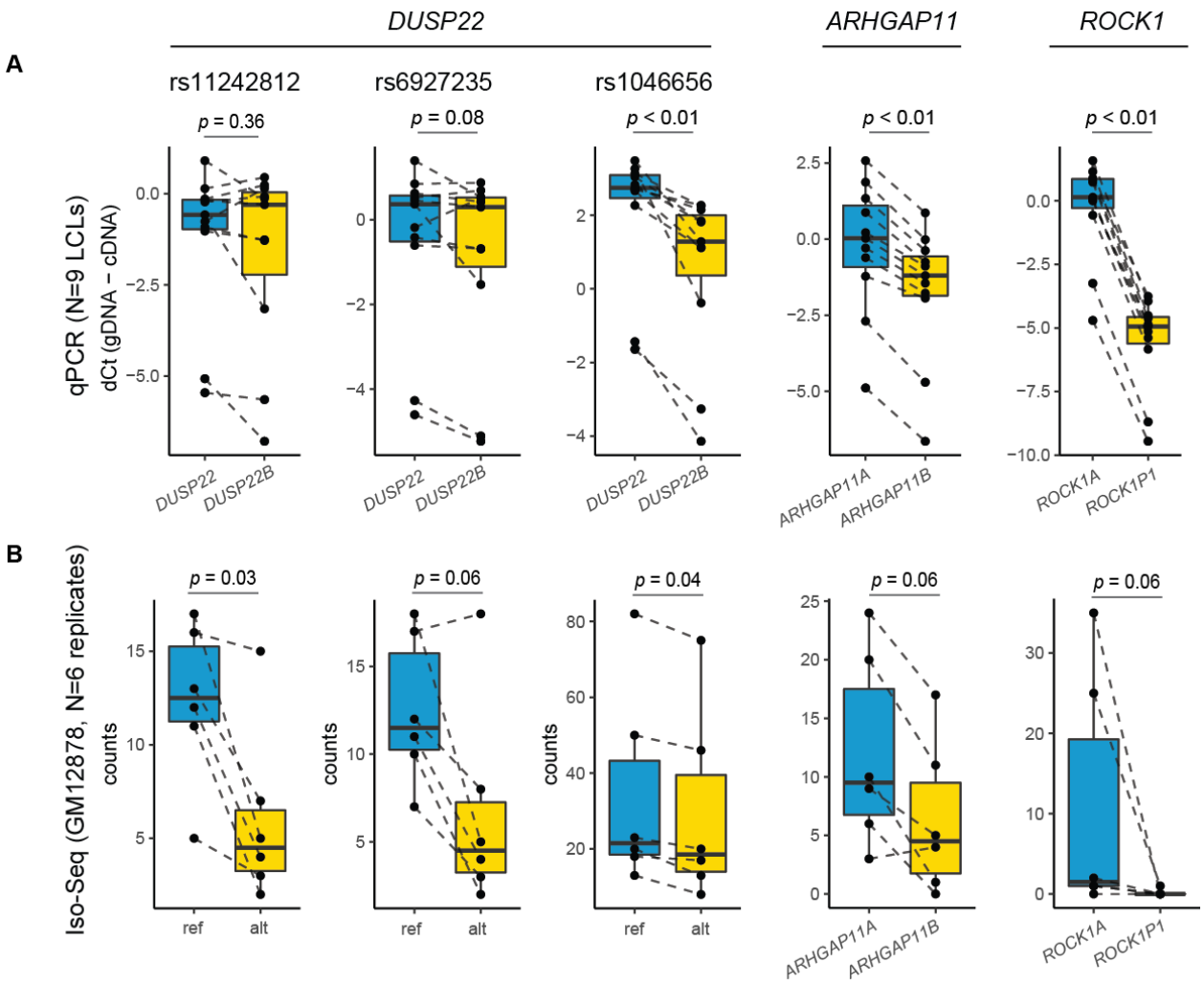


Figure S4.3. Validation of short-read RNA-seq quantification. (A) qPCR was conducted on genomic DNA (gDNA) and complementary DNA (cDNA) extracted from nine human LCLs to quantify expression differences at paralog-specific variants (PSVs). Because *DUSP22B* is missing from the reference, *DUSP22* PSVs are annotated as SNPs, where the alternate allele corresponds to a known *DUSP22B* substitution. The difference of cycle threshold (dCt) between gDNA and cDNA samples was calculated for each LCL and compared between paralogs with a paired Wilcoxon signed-rank test. Each point represents the mean of three technical replicates. **(B)** Read counts from six replicates of GM12878 Iso-Seq experiments (ENCODE) at *DUSP22* PSVs (raw alignments prior to sequence correction) or *ARHGAP11*/*ROCK1* paralogous regions (filtered alignments). Differences were quantified with a paired Wilcoxon signed-rank test.

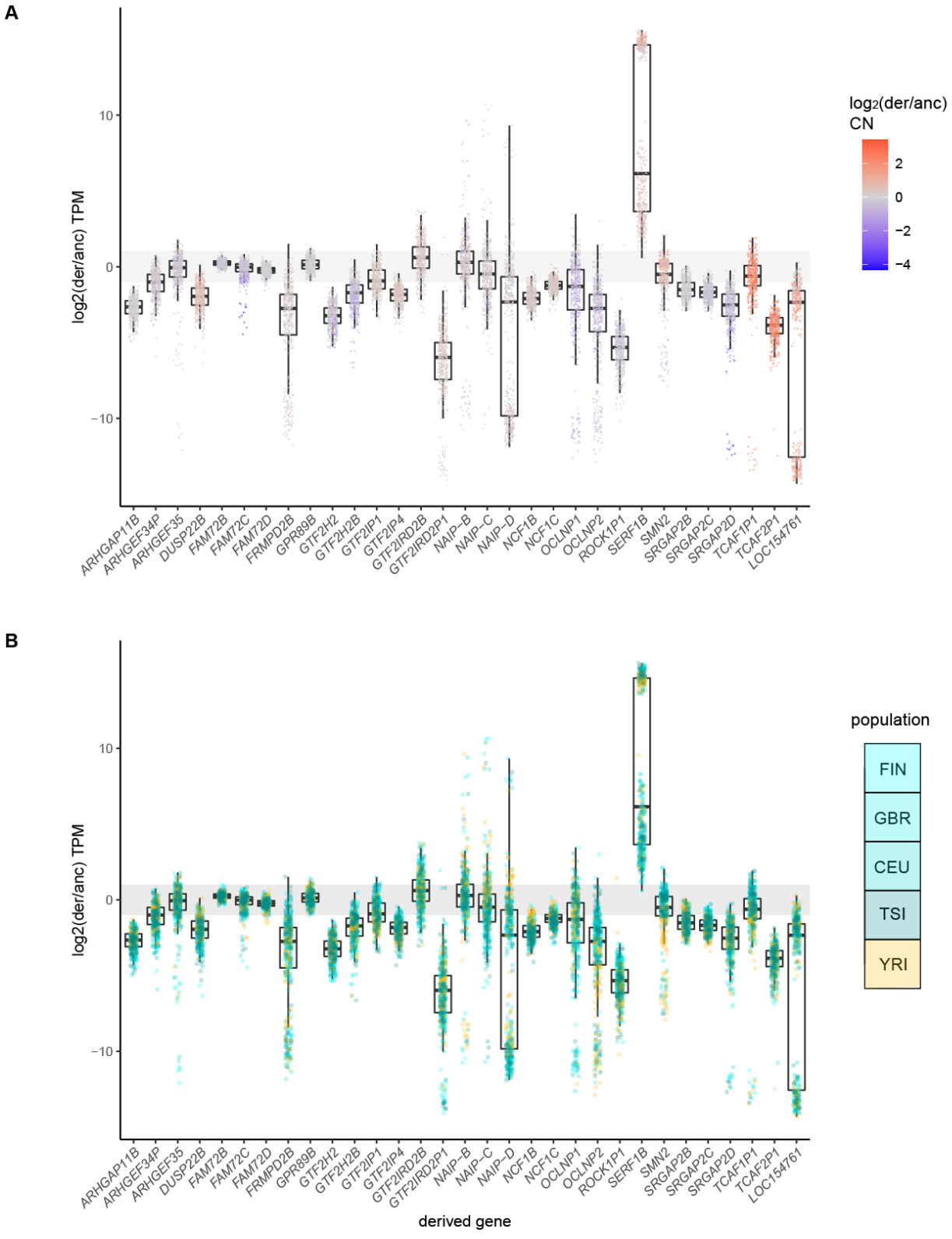


Figure S4.4. Expression divergence of derived HSD genes. Expression divergence of derived genes from families with at least one LCL-expressed paralog is plotted as the $\log_2$ ratio of median derived and ancestral TPM expression. Each point represents a different LCL from the Geuvadis consortium (total

N=445). The gray bar indicates a two-fold expression difference. Colors represent **(A)** relative CN and **(B)** LCL source population.

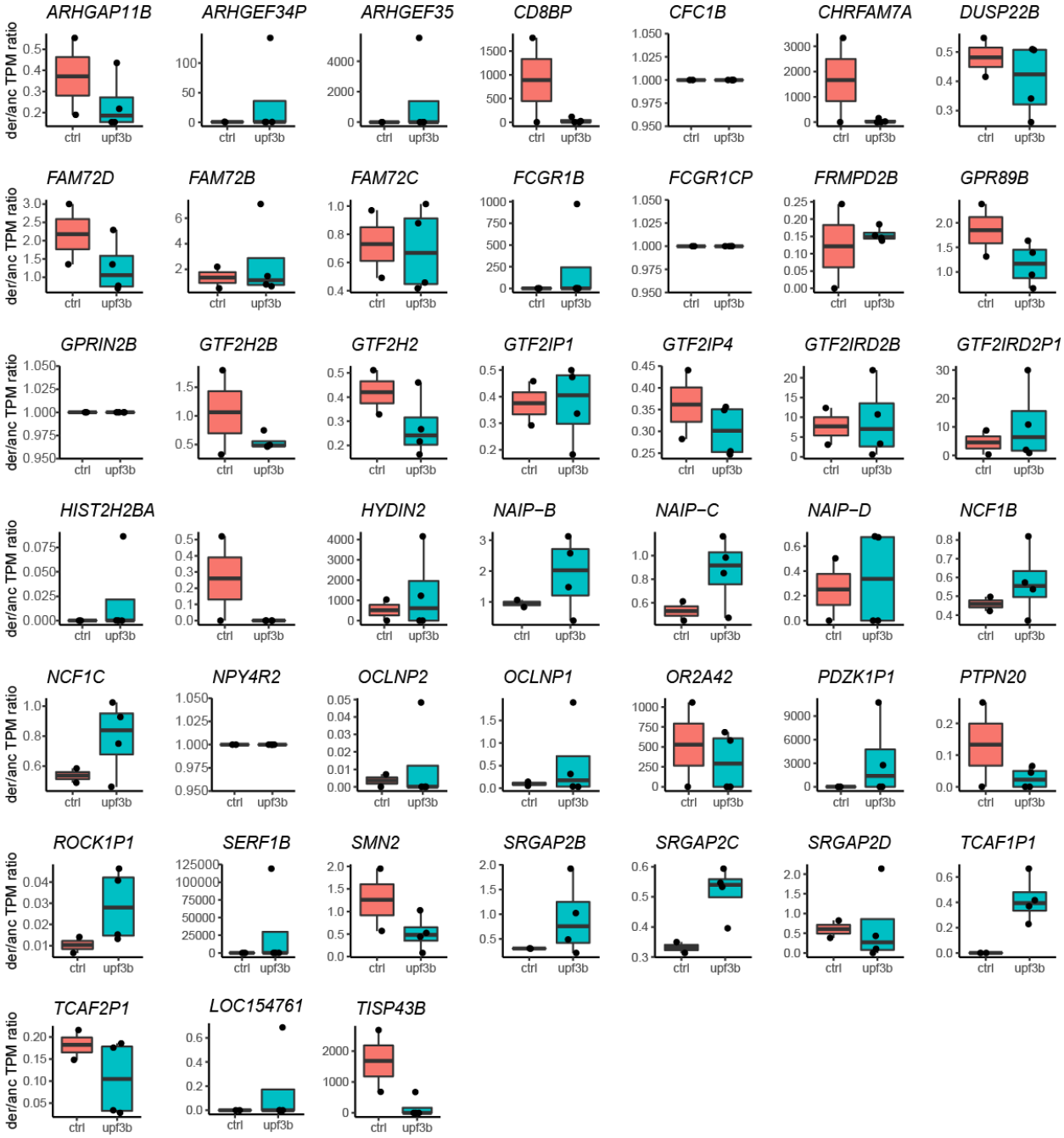


Figure S4.5. Nonsense-mediated decay of HSD transcripts. Derived/ancestral TPM ratios of HSD genes, calculated from RNA-seq of control (ctrl) and NMD-deficient cells (upf3b) (Nguyen et al. 2012). No differences were determined to be significant by differential expression analysis (limma-voom).

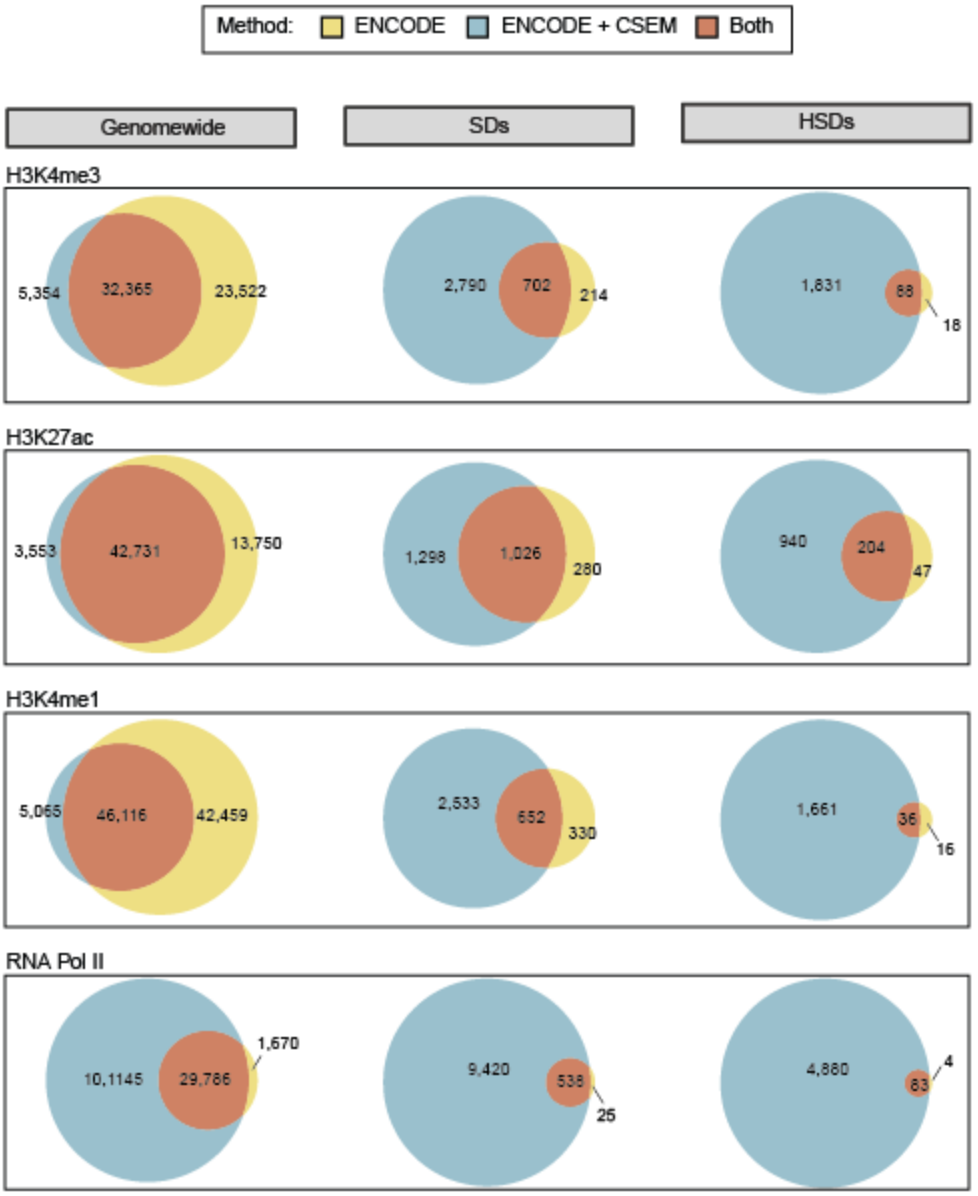


Figure S4.6. Comparison of ChIP-seq peaks from ENCODE and ENCODE data with multimapping and CSEM allocation. Overlap between data sets is shown for the whole genome, SDs, and HSDs (SDs with over 98% identity) for H3K4me1, H3K27ac, H4K4me3, and RNA PolII. Color indicates the method used for read mapping.

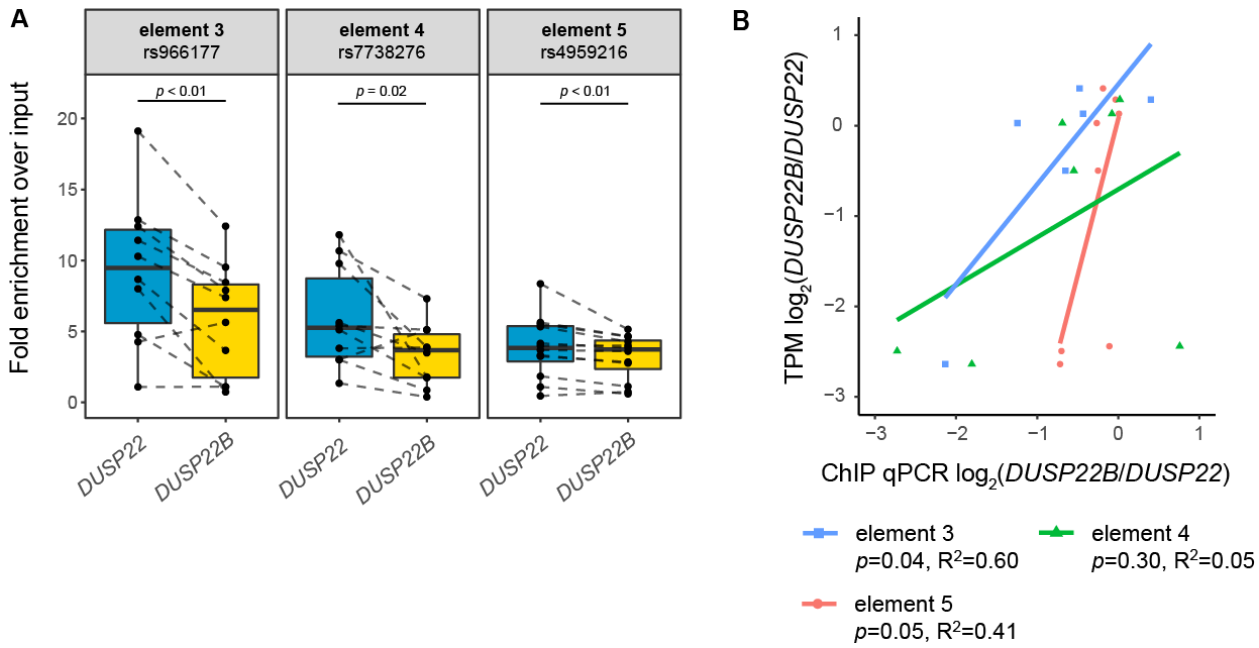


Figure S4.7. Paralogous differences of *DUSP22* H3K27ac peaks. To verify our H3K7ac ChIP-seq results in *DUSP22* and *DUSP22B*, as well as and assess biological reproducibility, we leveraged PSVs (annotated as SNPs in dbSNP) to perform paralog-specific ChIP-qPCR of three H3K27ac-enriched peaks in these genes. **(A)** *DUSP22* and *DUSP22B*-specific enrichment at three putative CREs, with paralogs distinguished using PSVs (N=12, 10, and 10 LCLs, respectively). Each variant lies within an element tested for enhancer activity with a luciferase reporter (Figures 5, S4.16, S4.17). Measurements were performed in triplicate and averaged. Differences in enrichment between paralogs were determined with a Wilcoxon signed-rank test, and *p*-values are denoted between the boxplots. **(B)** Correlation of expression divergence ($\log_2$ ratio of TPMs) with differences in enrichment (ChIP-qPCR signal) at the same three variants, for LCLs with RNA-seq data (Pickrell et al. 2010) (N=6, 7, and 8, respectively).

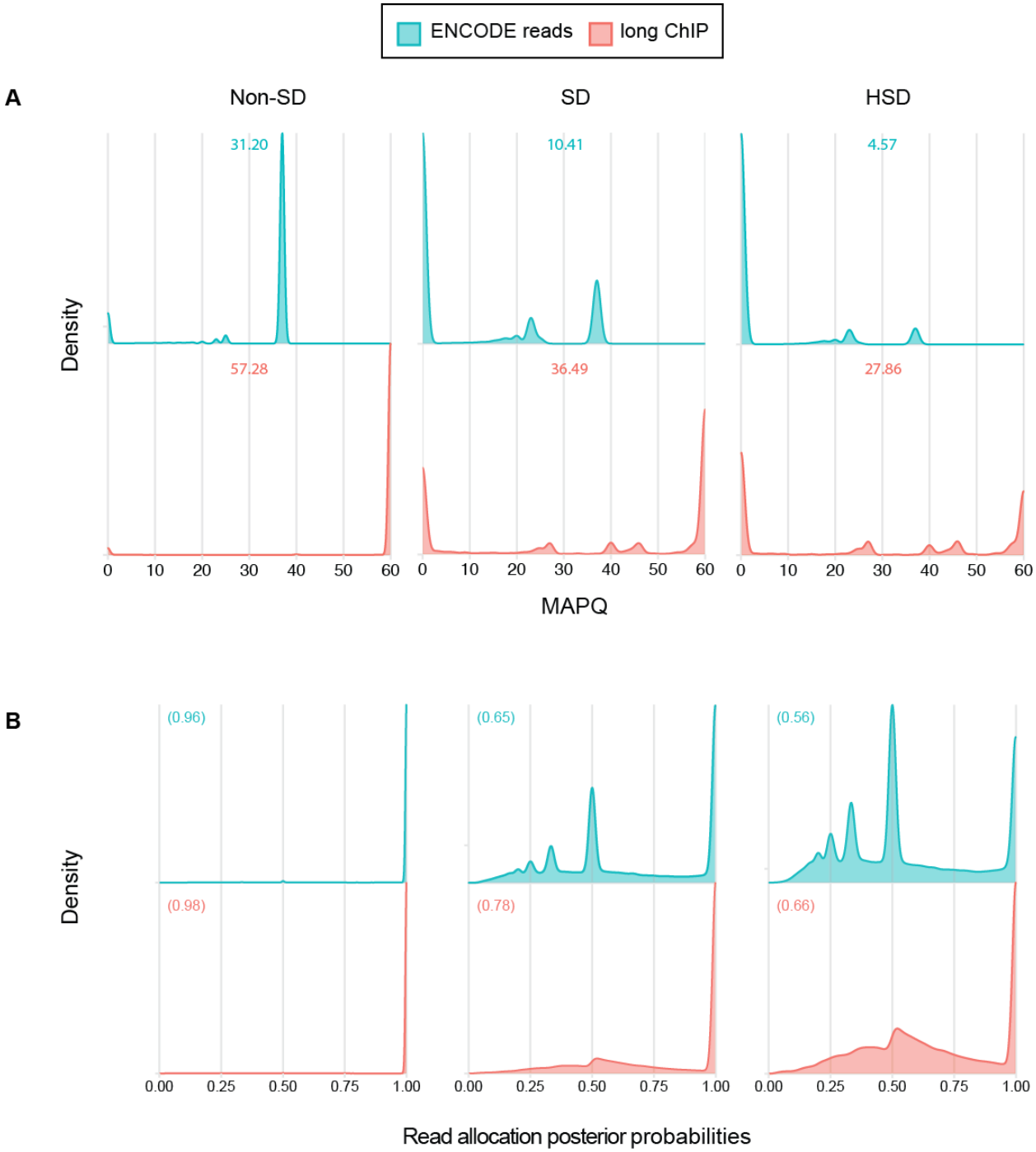


Figure S4.8. Summary of long ChIP data analysis (H3K27ac). Density plots are shown for the whole genome (left), SDs (middle), and SDs of over 98% sequence identity (HSD, left). **(A)** Distribution of BWA aln (ENCODE short reads, top) and BWA-MEM (long ChIP, bottom) alignment mapping quality (MAPQ) scores. The mean MAPQ score is shown on the top of each panel. **(B)** Distribution of bowtie (ENCODE short reads, top) and bowtie2 (long ChIP, bottom) CSEM posterior alignment probabilities. The mean posterior probability is shown on the top of each panel.

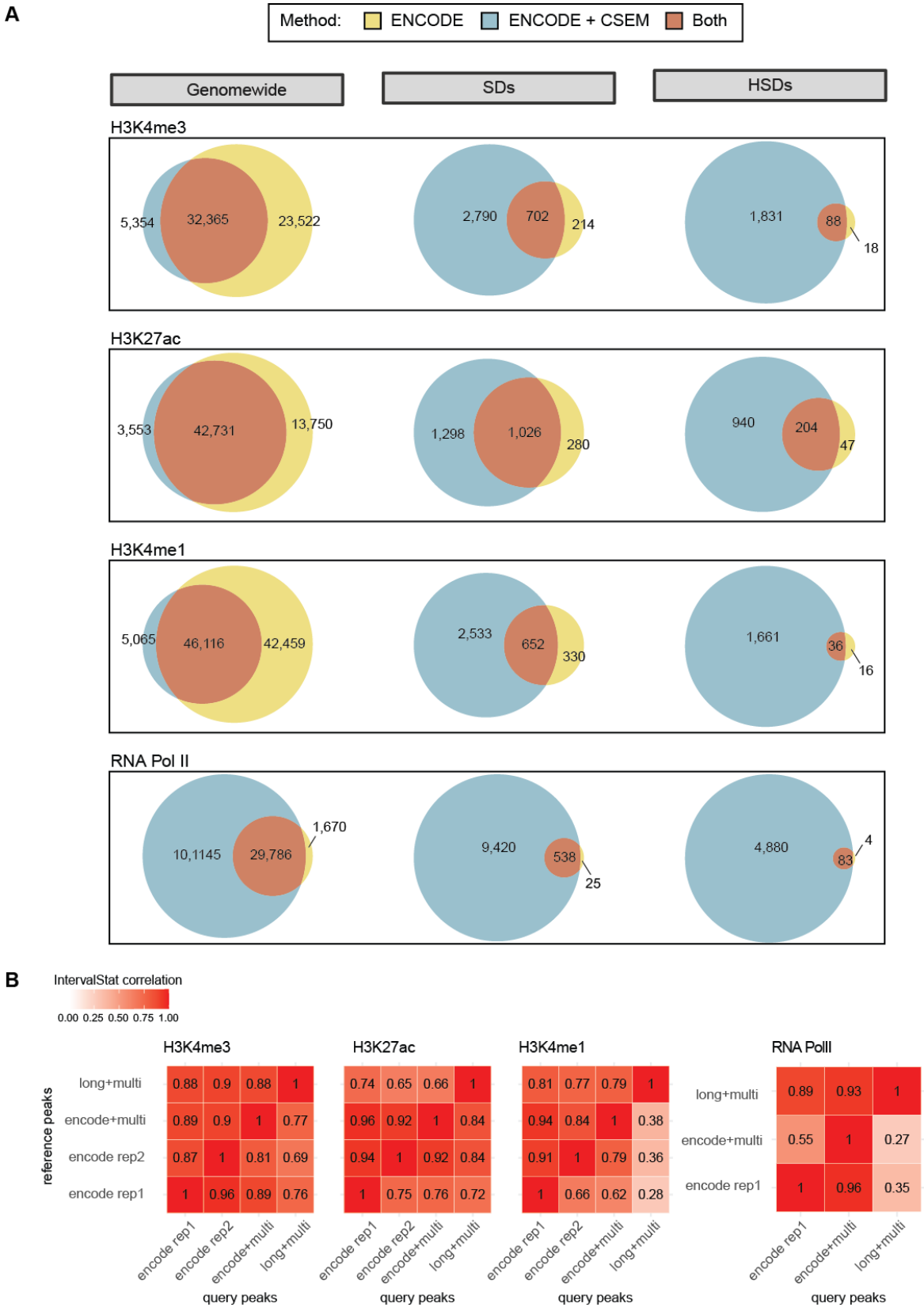

A
Method: ENCODE ENCODE + CSEM Both
Genomewide
SDs
HSDs
H3K4me3
5,354
32,365
23,522
2,790
702
214
1,831
88
18
H3K27ac
3,553
42,731
13,750
1,298
1,026
280
940
204
47
H3K4me1
5,065
46,116
42,459
2,533
652
330
1,661
36
16
RNA Pol II
10,1145
29,786
1,670
9,420
538
25
4,880
83
4
B
IntervalStat correlation
0.00 0.25 0.50 0.75 1.00
H3K4me3
H3K27ac
H3K4me1
RNA PolII
reference peaks
long+multi
encode+multi
encode rep2
encode rep1
0.88 0.9 0.88 1
0.89 0.9 1 0.77
0.87 1 0.81 0.69
1 0.96 0.89 0.76
0.74 0.65 0.66 1
0.96 0.92 1 0.84
0.94 1 0.92 0.84
1 0.75 0.76 0.72
0.81 0.77 0.79 1
0.94 0.84 1 0.38
0.91 1 0.79 0.36
1 0.66 0.62 0.28
0.89 0.93 1
0.55 1 0.27
1 0.96 0.35
encode rep1
encode rep2
encode+multi
long+multi
query peaks

Figure S4.9. Comparison of ChIP-seq peaks from long ChIP data with multimapping and CSEM allocation and published ENCODE. (A) Overlap between data sets is shown for the whole genome, SDs, and HSDs (SDs of over 98% identity) for H3K4me3, H3K27ac, H4K4me1, and RNA PolII. Color indicates the method used for read mapping. **(B)** Pairwise correlations of genome-wide peak sets from single ENCODE replicates, ENCODE multi-mapping with CSEM allocation, and long ChIP multimapping with CSEM allocation. Unidirectional correlations were determined using IntervalStats (Chikina and Troyanskaya 2012), with overlapping peaks defined at $p < 0.05$.

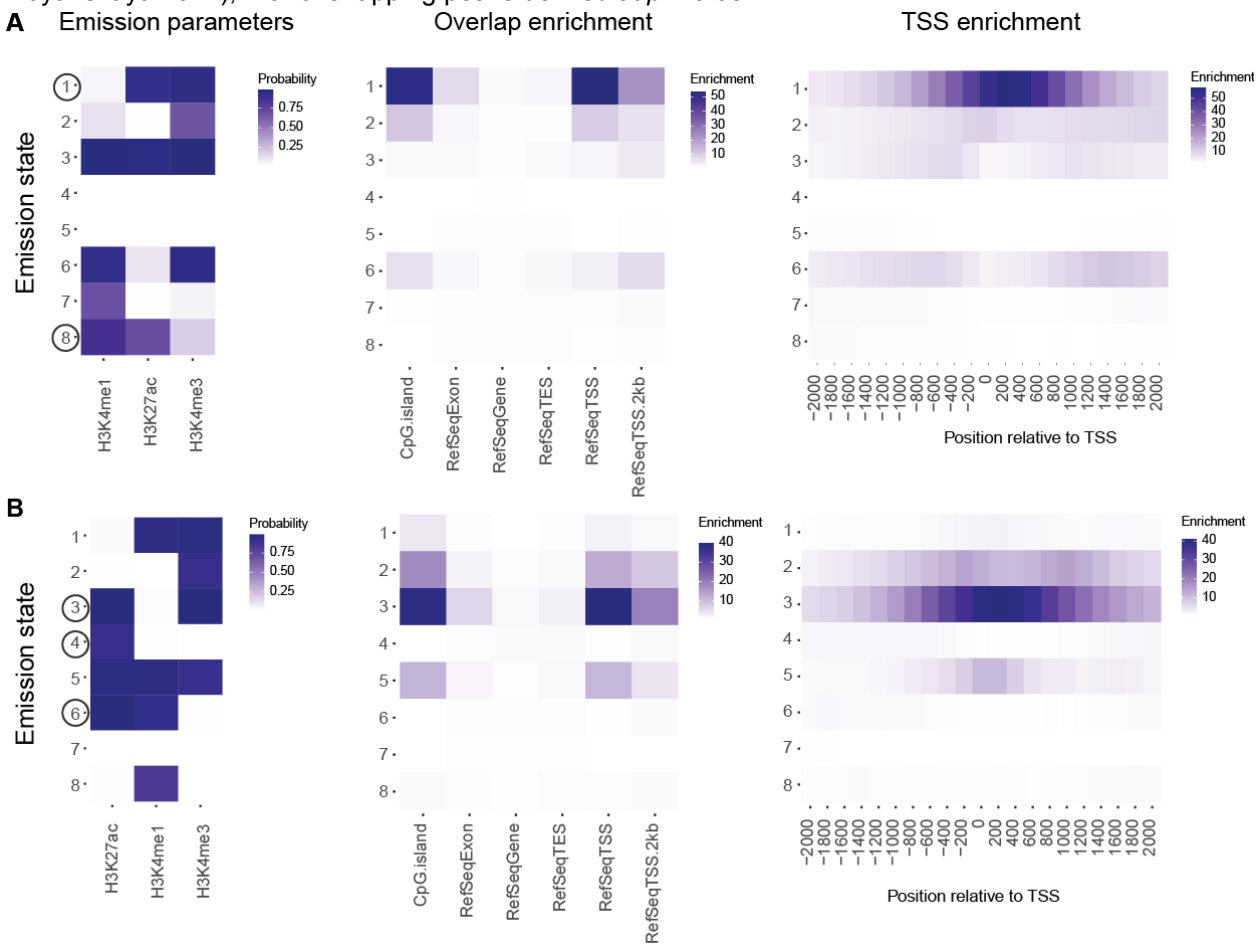


Figure S4.10. ChromHMM models used to cCREs. (A) Emission parameters, genomic features overlap, and enrichment relative to transcription start sites (TSSs) of an 8-state ChromHMM model built on ENCODE data (multiple-mapping and CSEM allocation). Darker blue indicates a higher probability of observing a given histone mark in each state, or a higher fold-enrichment in a given genomic feature or distance from TSS. State 1 was chosen to represent active promoters, and state 8 was chosen for active enhancers (circled). **(B)** Emission parameters, genomic features overlap, and enrichment relative to TSSs for an 8-state ChromHMM model built on long ChIP data (multiple-mapping and CSEM allocation). State 3 was chosen for active promoters, and states 4 and 6 were chosen for active enhancers (circled).

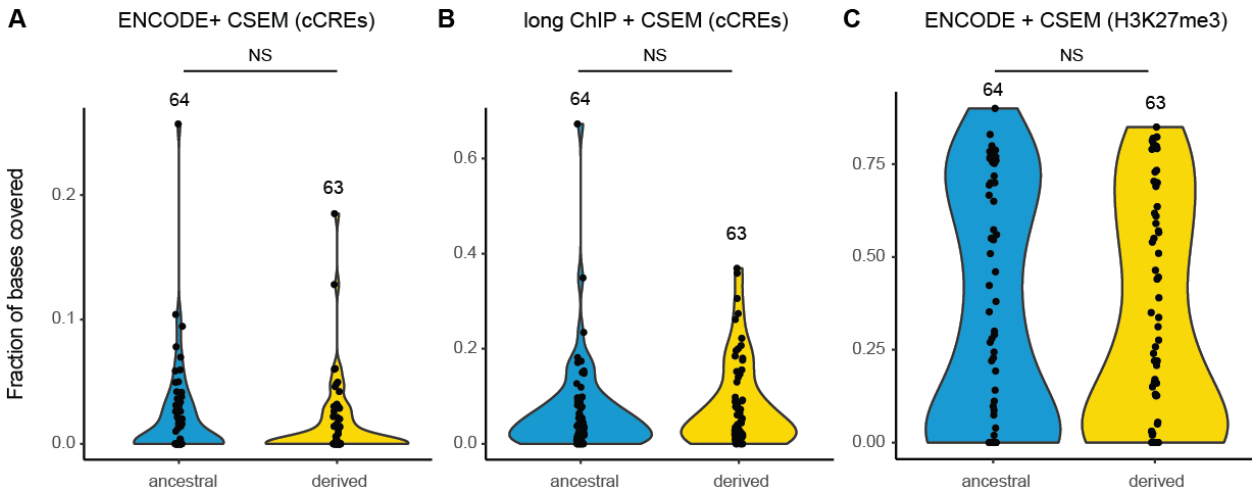


Figure S4.11. Global comparison of ancestral and derived HSDs. Violin plots represent the fraction of bases covered by (A) ENCODE multi-mapping cCREs, (B) long ChIP multi-mapping cCREs, and (C) ENCODE multi-mapping H3K27me3 domains. Fractional coverage was calculated in 100-kbp windows for ancestral and derived HSD regions. Values were compared with a Wilcoxon signed-rank test.

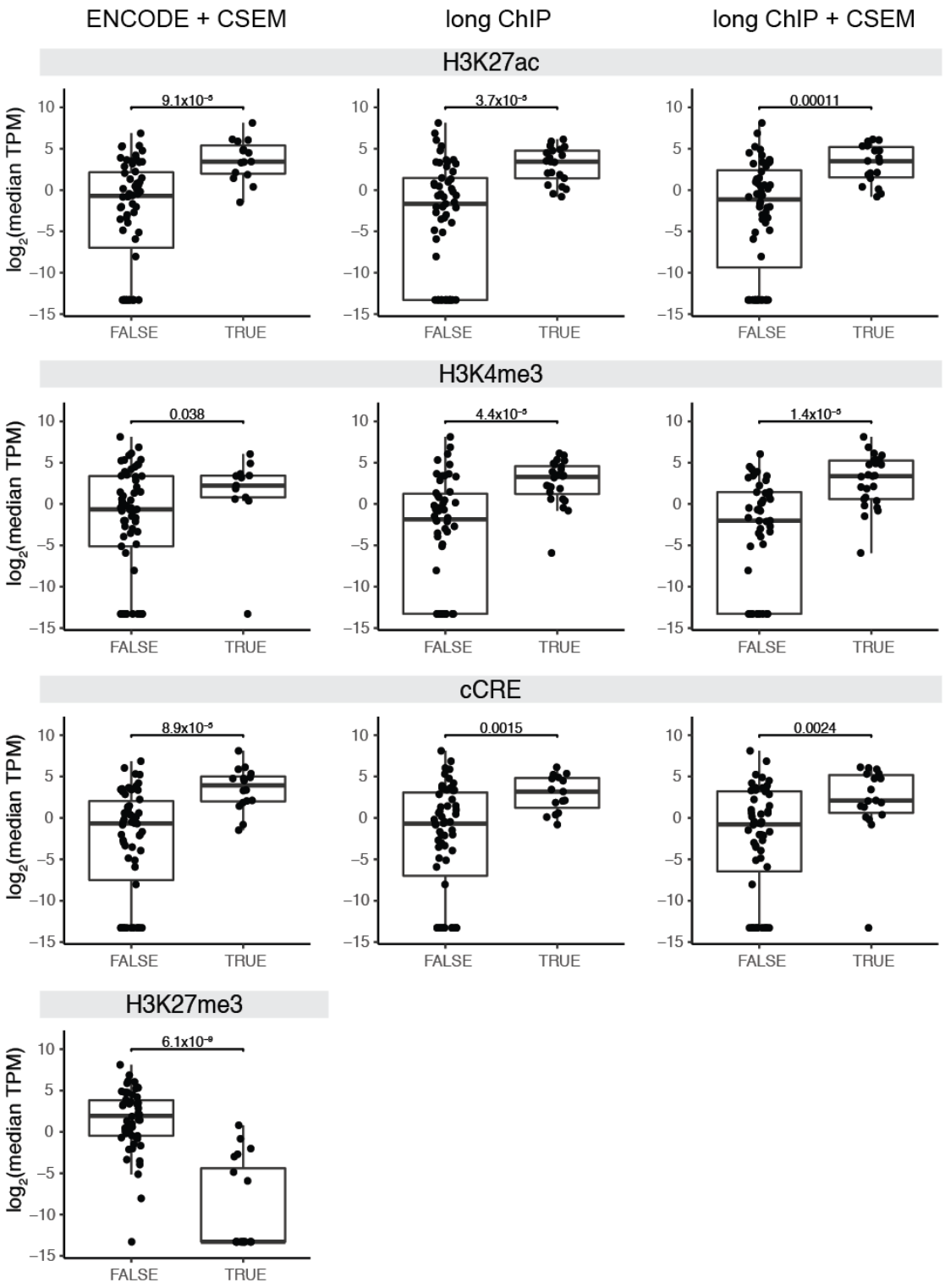


Figure S4.12. Associations of ChIP-seq marks and cCREs with HSD gene expression in LCLs. HSD genes were categorized by intersection of the transcription start site ("TRUE" or "FALSE") with a ChIP-seq peak or cCRE from the reanalyzed ENCODE data with multimapping and CSEM allocation (left), long ChIP data with single mapping (middle), and long ChIP data with multimapping and CSEM allocation

(right). Expression values (TPM) are from LCLs from the Geuvadis consortium (Lappalainen et al. 2013). *p*-values were generated from a Wilcoxon rank-sum test.

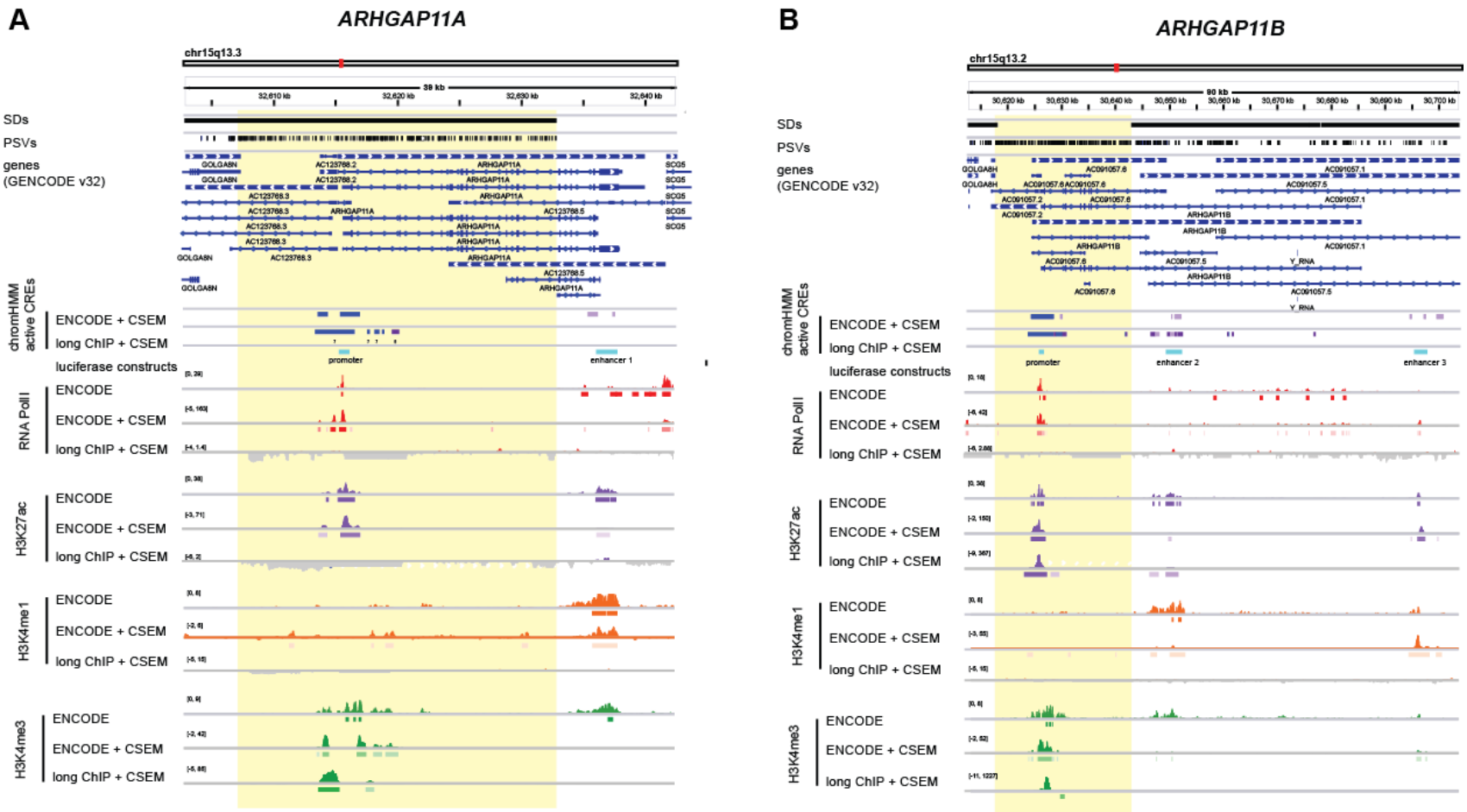


Figure S4.13. Epigenetic landscape of *ARHGAP11* genes. (A) *ARHGAP11A*. **(B)** *ARHGAP11B*. Coordinates indicate location on chromosome 15. The paralogous duplicated region is highlighted in yellow. Segmental duplications (SDs) and paralog-specific variants (PSVs) are indicated with black bars. ChromHMM segmentations are shown for active promoters (blue) and enhancers (lavender), as defined on ENCODE and long ChIP data (multimapping with CSEM allocation). Regions cloned and tested with luciferase reporters are shown in cyan. For each ChIP-seq target, a signal track is shown for published ENCODE; reanalyzed, multimapped ENCODE with CSEM allocation; and multimapped long ChIP with CSEM allocation. Visualized with the Integrative Genomics Viewer.

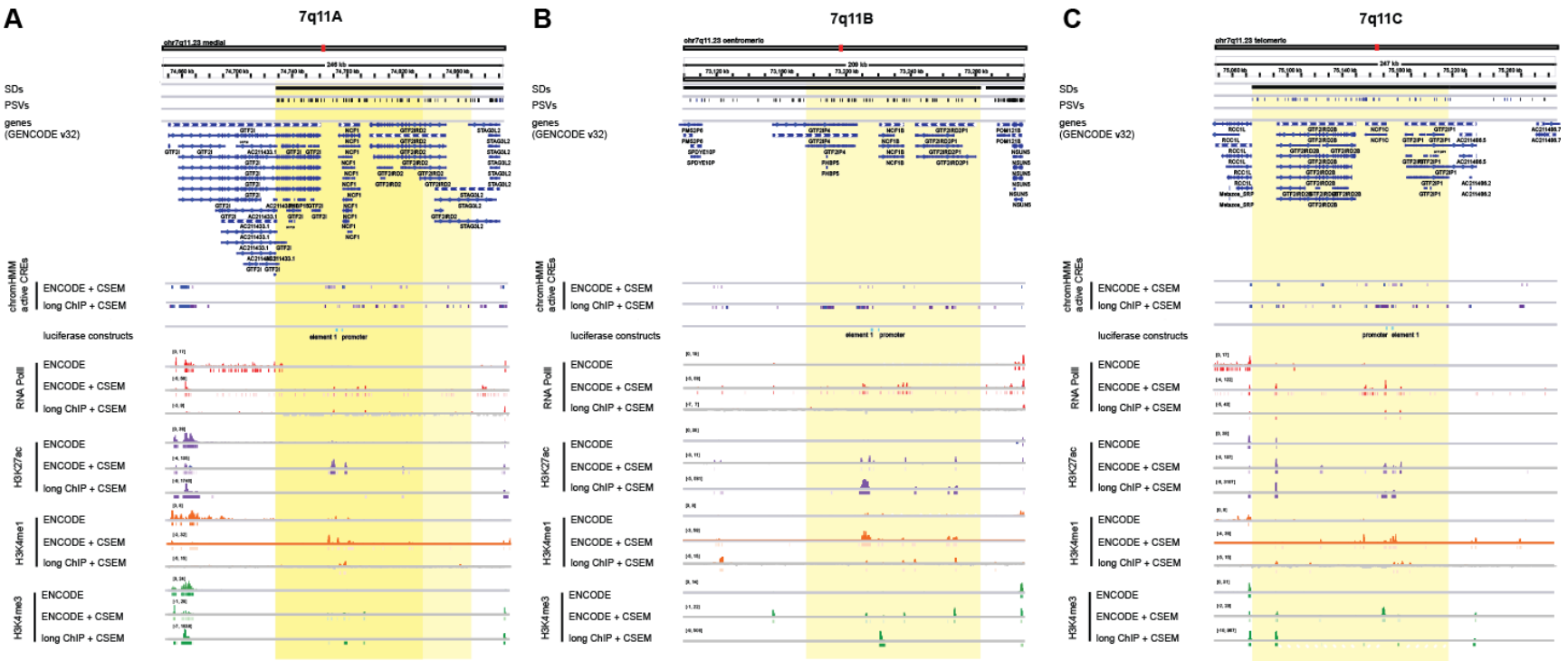


Figure S4.14. Epigenetic landscape of chromosome 7q11. (A) Ancestral locus. **(B)** Primary derived (centromeric) locus. **(C)** Secondary derived (telomeric) locus. Coordinates indicate location on chromosome 7. The paralogous duplicated regions are highlighted in yellow. Segmental duplications (SDs) and paralog-specific variants (PSVs) are indicated with black bars ChromHMM segmentations are shown for active promoters (blue) and enhancers (lavender), as defined on ENCODE and long ChIP data (multimapping with CSEM allocation). Regions cloned and tested with luciferase reporters are shown in cyan. For each ChIP-seq target, a signal track is shown for published ENCODE; reanalyzed, multimapped ENCODE with CSEM allocation; and multimapped long ChIP with CSEM allocation. Visualized with the Integrative Genomics Viewer.

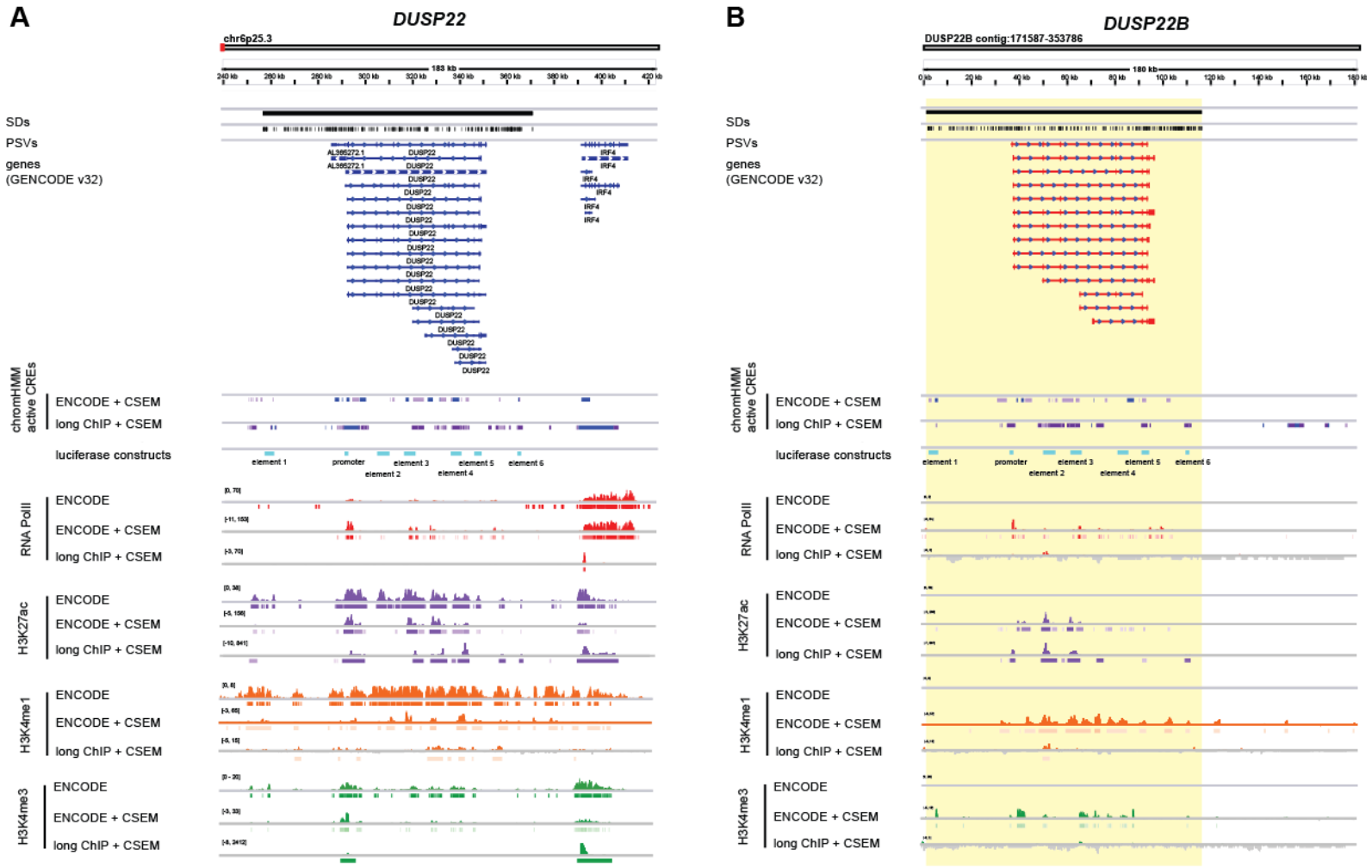


Figure S4.15. Epigenetic landscape of *DUSP22* genes. (A) *DUSP22*. **(B)** *DUSP22B*. Coordinates indicate location on chromosome 6 or the *DUSP22B* contig as published in (Dennis et al. 2017). *DUSP22B* gene models depict BLAT alignments of *DUSP22* transcripts. The paralogous duplicated region is highlighted in yellow. Segmental duplications (SDs) and paralog-specific variants (PSVs) are indicated with black bars. ChromHMM segmentations are shown for active promoters (blue) and enhancers (lavender), as defined on ENCODE and long ChIP data (multimapping with CSEM allocation). Regions cloned and tested with luciferase reporters are shown in cyan. For each ChIP-seq target, a signal track is shown for published ENCODE; reanalyzed, multimapped ENCODE with CSEM allocation; and multimapped long ChIP with CSEM allocation. Visualized with the Integrative Genomics Viewer.

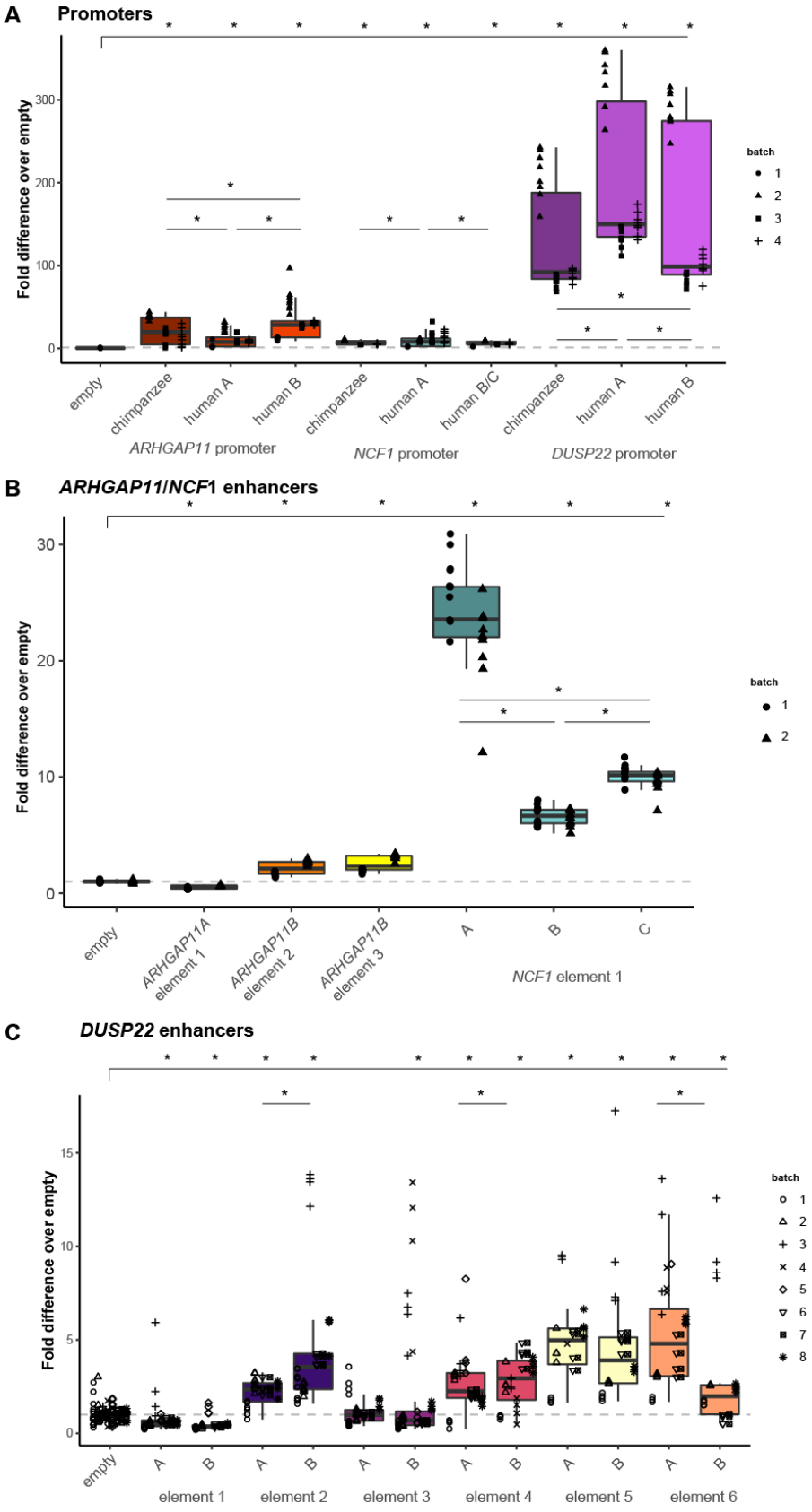

A
Promoters
Fold difference over empty
batch
empty
chimpanzee
human A
human B
chimpanzee
human A
human B/C
chimpanzee
human A
human B
ARHGAP11 promoter
NCF1 promoter
DUSP22 promoter
B
ARHGAP11/NCF1 enhancers
Fold difference over empty
batch
empty
ARHGAP11A element 1
ARHGAP11B element 2
ARHGAP11B element 3
A
B
C
NCF1 element 1
C
DUSP22 enhancers
Fold difference over empty
batch
element 1
element 2
element 3
element 4
element 5
element 6

Figure S4.16 Luciferase activity of candidate CREs from all HeLa experiments. Luciferase activity for each cloned construct is shown as the fold difference over the average negative control value for **(A)** promoters, **(B)** *ARHGAP11* and *NCF1* candidate enhancers, and **(C)** *DUSP22* candidate enhancers. Values are visually separated by experimental batch. Significant differences ($p < 0.05$, Tukey post-hoc test of two-way ANOVA for batch correction) from empty (top bar) and between homologous sequences are indicated with an asterisk.

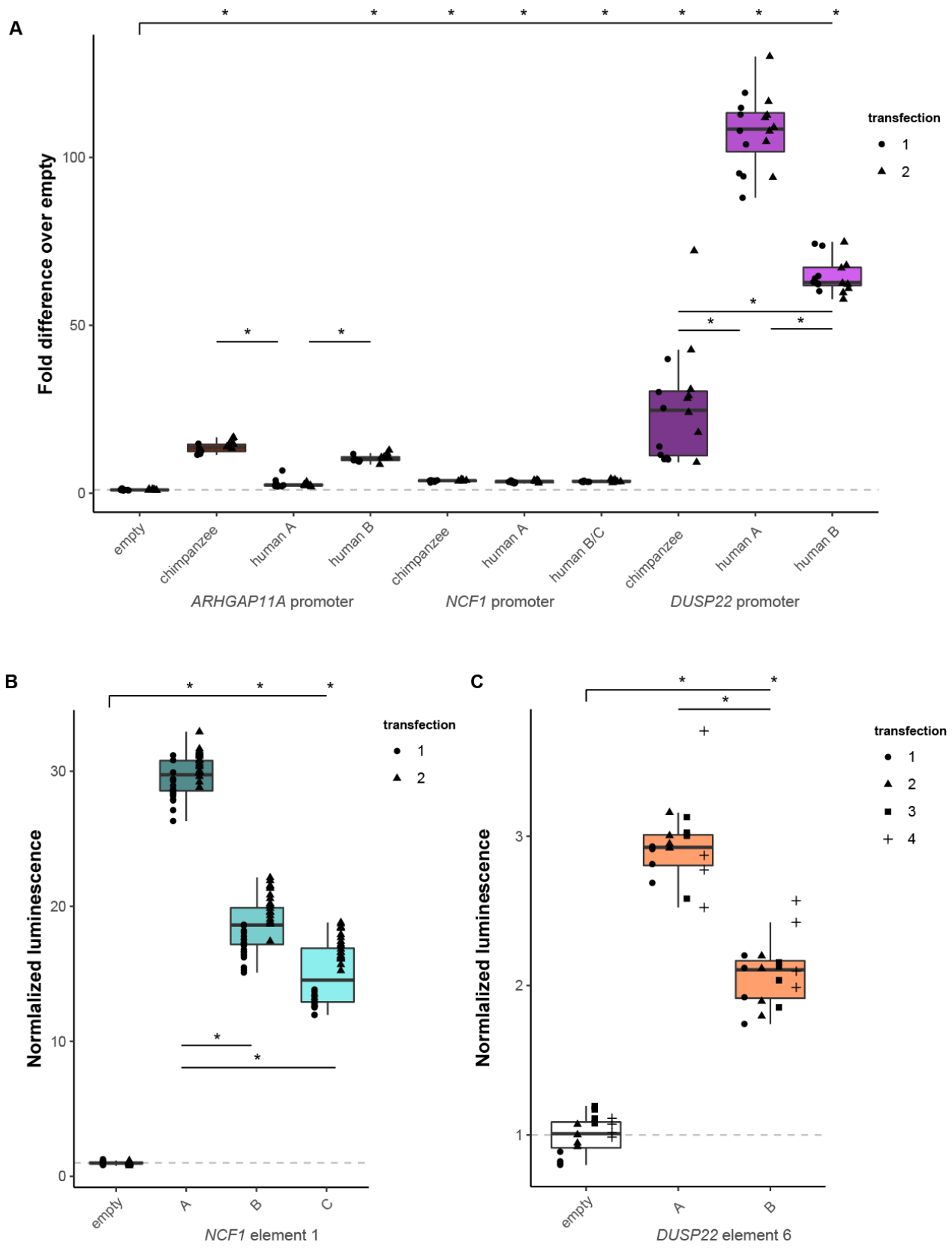


Figure S4.17. Luciferase activity of candidate CREs from all LCL experiments. Luciferase activity for each cloned construct is shown as the fold difference over the average negative control value for **(A)**

promoters, **(B)** *NCF1* candidate element 1, and **(C)** *DUSP22* candidate element 6. Values are visually separated by experimental batch. Significant differences ($p < 0.05$, Tukey post-hoc test of two-way ANOVA) from empty (top bar) and between homologous sequences are indicated with an asterisk.

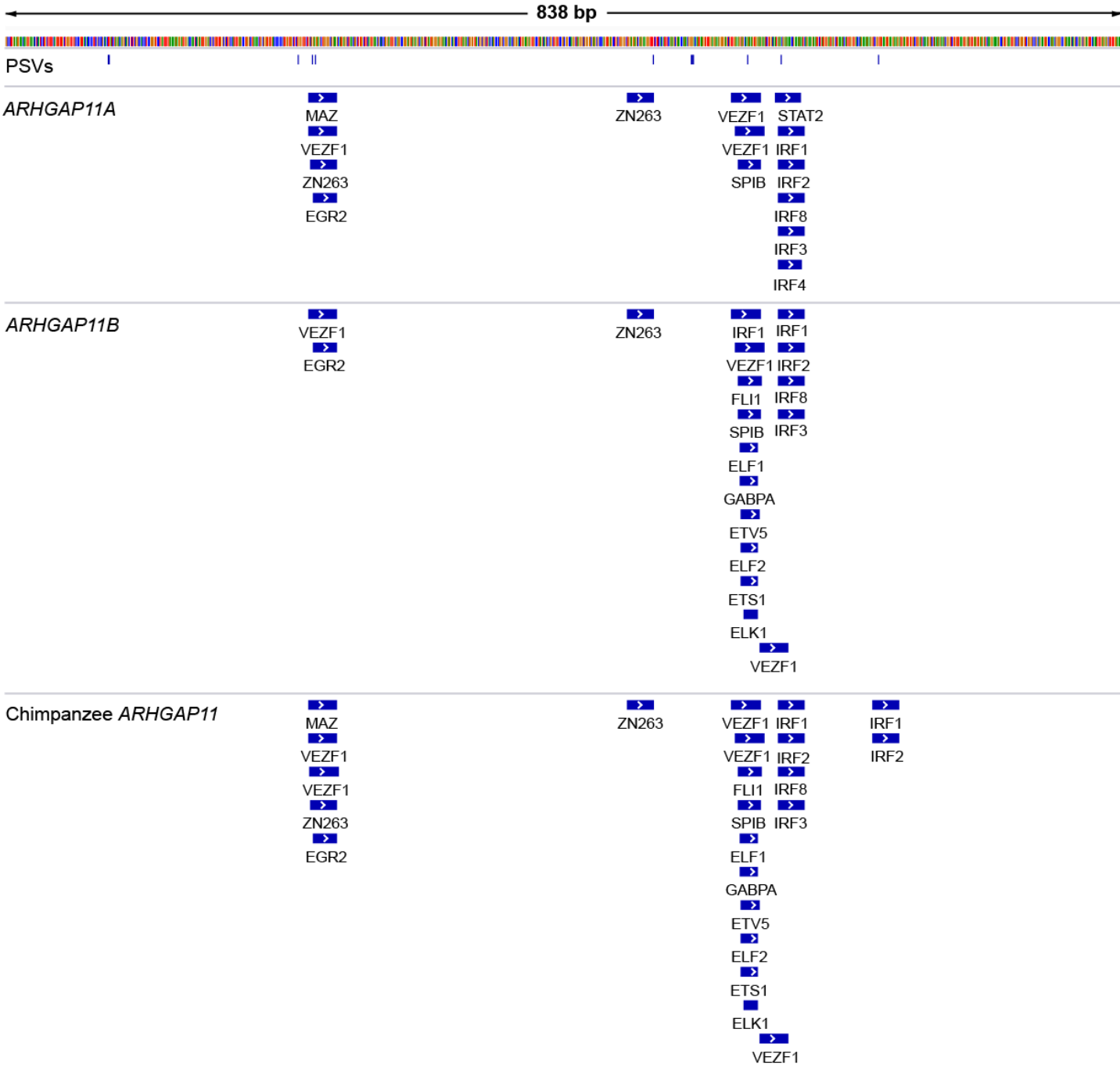


Figure S4.18. Transcription factor binding sites identified in *ARHGAP11* promoters. Significant matches (5% FDR) for HOCOMOCO v.12 Transcription factor binding site motifs intersecting PSVs are depicted under the cloned sequence tested with luciferase reporter. PSVs are depicted with blue vertical lines. Visualized with the Integrative Genomics Viewer.

4.10 SUPPLEMENTARY TABLES

Table S4.1: HSD genes assayed in this study.

Gene symbol	*Gencode v32*	*GRCh38 coordinates*	*Evolutionary status*	*Truncation status*
ARHGAP11A	ENSG00000198826.11	chr15:32615143-32639941	ancestral	ancestral
ARHGAP11B	ENSG00000284906.1	chr15:30624547-30685606	derived	3′ only
ARHGEF5	ENSG00000050327.15	chr7:144355287-144380632	ancestral	ancestral
ARHGEF34P	ENSG00000204959.4	chr7:144272444-144286966	derived	3′ only
ARHGEF35	ENSG00000213214.4	chr7:144186082-144195655	derived	3′ only
CD8B	ENSG00000172116.23	chr2:86815338-86861924	ancestral	ancestral
CD8BP/CD8B2	ENSG00000254126.8	chr2:106487363-106544297	derived	whole
CFC1	ENSG00000136698.8	chr2:130592167-130599575	unknown	whole
CFC1B	ENSG00000152093.8	chr2:130521196-130528604	unknown	whole
CHRFAM7A	ENSG00000166664.13	chr15:30357765-30393849	derived	5′
CHRNA7	ENSG00000175344.18	chr15:31923437-32173018	ancestral	ancestral
DUSP22	ENSG00000112679.14	chr6:291629-351355	ancestral	ancestral
-			derived	whole
FAM72A	ENSG00000196550.10	chr1:206186178-206204414	ancestral	ancestral
FAM72D	ENSG00000215784.6	chr1:145095973-145112696	derived	whole
FAM72B	ENSG00000188610.12	chr1:121167645-121185539	derived	whole
FAM72C	ENSG00000263513.5	chr1:143955363-143971965	derived	whole
FCGR1A	ENSG00000150337.13	chr1:149782670-149792518	ancestral	ancestral
FCGR1B	ENSG00000198019.13	chr1:121087344-121097161	derived	whole
FCGR1CP	ENSG00000265531.3	chr1:143874792-143883575	derived	whole
FRMPD2	ENSG00000170324.21	chr10:48153087-48274696	ancestral	ancestral
FRMPD2B	ENSG00000150175.13	chr10:46870857-46894562	derived	5′
GPR89A	ENSG00000117262.19	chr1:145607987-145670650	ancestral	ancestral
GPR89B	ENSG00000188092.15	chr1:147928392-147993592	derived	whole
GPRIN2	ENSG00000204175.5	chr10:46549043-46555530	unknown	NA
-			unknown	NA
GTF2H2C	ENSG00000183474.15	chr5:69560207-69594723	ancestral	ancestral
GTF2H2B	ENSG00000226259.10	chr5:70415351-70448015	derived	whole
GTF2H2	ENSG00000145736.14	chr5:71032669-71067689	derived	whole
GTF2I	ENSG00000263001.6	chr7:74650230-74760692	ancestral	ancestral
GTF2IP1	ENSG00000277053.4	chr7:75185384-75237696	derived	5′
GTF2IP4	ENSG00000233369.7	chr7:65084102-65100232	derived	5′
GTF2IRD2	ENSG00000196275.14	chr7:74796143-74851551	ancestral	ancestral

GTF2IRD2B	ENSG00000174428.18	chr7:75092572-75149817	derived	3′ only
GTF2IRD2P1	ENSG00000214544.7	chr7:73242750-73280119	derived	5′
HIST2H2BF	ENSG00000203814.6	chr1:149782688-149812373	ancestral	ancestral
HIST2H2BA	ENSG00000223345.3	chr1:121087566-121116676	derived	whole
-		chr1:143874964-143904037	derived	whole
HYDIN	ENSG00000157423.18	chr16:70802083-71230722	ancestral	ancestral
HYDIN2	ENSG00000276975.3	chr1:146472565-146914294	derived	5′
NAIP	ENSG00000249437.7	chr5:70968482-71025114	ancestral	ancestral
-		chr5:70093037-70128437	derived	5′
-		chr5:71101255-71128756	derived	5′
-		chr5:71123013-71128756; chr5:70473444-70479184; chr5:69618309-69624049; chr5:70122690-70128437	derived	5′
NCF1	ENSG00000158517.15	chr7:74773961-74789376	ancestral	ancestral
NCF1B	ENSG00000182487.12	chr7:73220623-73235945	derived	whole
NCF1C	ENSG00000165178.9	chr7:75156638-75172044	derived	whole
NPY4R	ENSG00000204174.8	chr10:46461098-46465958	unknown	whole
NPY4R2	ENSG00000264717.5	chr10:47918738-47923524	unknown	whole
OCLN	ENSG00000197822.11	chr5:69492291-69558104	ancestral	ancestral
-		chr5:70409765-70413251	derived	5′
LOC647859/OCLNP1	ENSG00000230847.4	chr5:71074224-71093193	derived	5′
OR2A1	ENSG00000221970.2	chr7:144312463-144322668	ancestral	ancestral
OR2A42	ENSG00000212807.2	chr7:144228243-144239605	derived	whole
PDZK1	ENSG00000174827.13	chr1:145670851-145708148	ancestral	ancestral
PDZK1P1	ENSG00000215859.9	chr1:147993861-148014956	derived	5′
PTPN20CP	ENSG00000278561.1	chr10:48064307-48119455	unknown	NA
PTPN20	ENSG00000204179.10	chr10:46911395-47002488	unknown	NA
ROCK1	ENSG00000067900.8	chr18:20946905-21111813	ancestral	ancestral
ROCK1P1	ENSG00000263006.6	chr18:109064-122219	derived	5′
SERF1A	ENSG00000172058.15	chr5:70900664-70918530	ancestral	ancestral
SERF1B	ENSG00000205572.9	chr5:70025246-70043113	derived	whole
SMN1	ENSG00000172062.16	chr5:70925029-70953942	ancestral	ancestral
SMN2	ENSG00000205571.13	chr5:70049611-70078522	derived	whole
SRGAP2	ENSG00000266028.7	chr1:206203344-206464443	ancestral	ancestral
SRGAP2B	ENSG00000196369.11	chr1:144887264-145095528	derived	3′ only
SRGAP2C	ENSG00000171943.12	chr1:121184810-121392874	derived	3′ only
SRGAP2D	ENSG00000270872.2	chr1:143975086-144068350	derived	3′ only

TCAF1P1	ENSG00000223459.6	chr7:143598038-143604839	unknown	NA
TCAF1	ENSG00000198420.10	chr7:143851374-143902198	unknown	NA
TCAF2	ENSG00000170379.20	chr7:143620951-143730409	unknown	NA
TCAF2P1	ENSG00000159860.7	chr7:143800731-143817973	unknown	NA
LOC154761		chr7:143811968-143836717	unknown	NA
TISP43/PRSS40A	ENSG00000183292.13	chr2:130570828-130584161	unknown	whole
TISP43B/PRSS40B/LOC646743	ENSG00000184761.8	chr2:130539094-130549883	unknown	whole

Table S4.2: Primers

Gene family	Application	Forward	Reverse	Locus Primer	Note
ARHGAP11	**Expression**	**/rhAmp-F/GCACAGAGGAAAAGAATAAAGCTATrACTGT/GT1/**	**/rhAmp-Y/GCACAGAGGAAAAGAATAAAGCTACrACTGT/GT1/**	**GCGTGGTCAGCCAGAAGACArGGAGA/GT2/**	
ROCK1	**Expression**	**/rhAmp-F/TTTTGTTCGTGCTTCCCCTTrGAACG/GT3/**	**/rhAmp-Y/TTTTGTTCGTGCTTCCCCTCrGAACG/GT3/**	**GCCACTTTCCGGAAAGACTGATTGrCAGTG/GT2/**	
DUSP22	**Expression**	**/rhAmp-F/CATGTGTGTATGTTGTGAAAGTrGTCTG/GT1/**	**/rhAmp-Y/CATGTGTGTATGTTGTGAAAGCrGTCTG/GT1/**	**GCCAGCAATATGAATTCTGTGACTTCrCAGCA/GT2/**	**DUSP22-1**
DUSP22	**Expression**	**/rhAmp-F/GTGGAGCAGTTTTCCGrGCACT/GT2/**	**/rhAMP-Y/GGTGGAGCAGTTTTCCArGCACT/GT2/**	**GCAAGACAAGCAGTGGGAAGrGAAGG/GT3/**	**DUSP22-3**
DUSP22	**ChIP**	**/rhAmp-F/GTCAATGGTGTATTTCTGTATTACATTTArUATGC/GT4/**	**/rhAMP-Y/GTCAATGGTGTATTTCTGTATTACATTTCrUATGC/GT4/**	**GCTCTCTAGGAAATCACCAGTTTGArGGAGT/GT3/**	**DUSP22 element 3**
DUSP22	**ChIP**	**/rhAmp-F/TGGACTTTTAGCGCATCCGrUCACC/GT3/**	**/rhAmp-Y/TGGACTTTTAGCGCATCCArUCACC/GT3/**	**GCATCCAGCCAAAGAAAGAGTTACAArAAGCA/GT4/**	**DUSP22 element 4**
DUSP22	**ChIP**	**/rhAmp-F/TCGAGTTTTCTCAGCTATATGAGGrGCTAG/GT4/**	**/rhAmp-Y/GTCGAGTTTTCTCAGCTATATGAGArGCTAG/GT4/**	**GCGATTGGGTAAGGGCTTCTCTArCATCA/GT1/**	**DUSP22 element 5**
ARHGAP11	**luciferase**	**GGCCTTGAAGGAACAAGTGA**	**TCCAATTTCCAAACGCTCTC**		

ARHGAP11	**luciferase**	**TTCTATCAACAATTGGGAAATGCTT**	**TTGTACCAGGCCCTTCCTCT**
ARHGAP11	**luciferase**	**TAAGTTTGTTTTCCCCTGTAACATA**	**CCACACCCATCCTGTAGGG**
ARHGAP11	**luciferase**	**AGGCTTCAGTGCCTTTGTGT**	**AGCTTCCCCATCCAGGAGTA**
NCF1	**luciferase**	**AAAAATTTAACCGGGCATGG**	**GATGTGACGGATGAAGGTGTC**
NCF1	**luciferase**	**GACGTGGGGGAATTCTTGGG**	**CCTGTCAAATGCCTCCTCGAA**
DUSP22	**luciferase**	**AACCTCACCCGTTTTCTCGG**	**GTGCTCGCAGTGTCAACAAG**
DUSP22	**luciferase**	**TCTGACTGCCTTTGGTTGACTA**	**CTCTTGCACAGCCTAGATGGTC**
DUSP22	**luciferase**	**GCTTTAGACTTCTCATGGGTGA**	**TTCATTGCTCCAACCTCTCAG**
DUSP22	**luciferase**	**CATAGCCCTTCAGGACTACAC**	**CTCCCCACTGAGTCAAACC**
DUSP22	**luciferase**	**CAGCTTTAGCAGTCCGTCTT**	**AAACCAGCCCACGTTTGTA**
DUSP22	**luciferase**	**CCTTAGCGTATTCTGGTCGG**	**ATTATCGTAGGTCAGCGGAG**
DUSP22	**luciferase**	**CTATTTCCGCTCTTCATTGTCG**	**CCTTGTACACTGTAGGCGAGT**
ACTB	**ChIP**	**AGGGTGAGGATGCCTCTCTT**	**GGGCTTCTTGTCCTTTCCTT**
HER2	**ChIP**	**TTGGAATGCAGTTGGAGGGG**	**GGTTTCTCCGGTCCCAATGG**

4.11 SUPPLEMENTARY NOTE

We encountered a number of technical limitations while studying highly similar duplicated genes. We implemented an alignment-free approach for transcript quantification, which accommodates ambiguous mappings resulting from methods and is demonstrated to accurately distinguish highly similar transcripts (Soneson et al. 2015; 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 non-zero

in individuals completely missing this paralog (Figure 3B) 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 (such *HIST2H2BF*) appeared entirely distinguishable between paralogs.

In addition, the available contigs of complete HSD loci have only been generated from a single haplotype (Dennis et al. 2017). As such, PSVs are not necessarily fixed, and may also be shared between genes as a result of interlocus gene conversion (Dumont 2015). Gene conversion can counteract the divergence of paralogous loci and has been predicted to prevent the fixation of neofunctionalizing variants (Kosuke M. Teshima 2008). While we recognize the impact gene conversion may have on our results, a previous study by Dumont *et al* (2015) found only ~3% of SDs show signatures of interlocus gene conversion in the human genome, a result we recapitulated in our initial sequencing and evolutionary comparisons of these HSD genes (Dennis et al. 2017). There, we identified a European-specific gene conversion event between *CFC1* paralogs but were unable to assess its impact here as neither gene is expressed in LCLs. Increasing availability of multi-modal data (genomic and RNA) from diverse human tissues will allow us to better explore the impact of interlocus gene conversion on expression. Nevertheless, assessment of alleles within and between HSD paralogous loci is still prohibitively difficult with short-read sequencing.

As expected, we found that longer read lengths allowed 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, though we used existing chromatin conformation information to connect non-duplicated adjacent regions with HSD gene promoters, improved experimental and bioinformatic methods are required to accurately link distal CREs within SDs to their target promoters. Like other genomic assays, available chromatin interactions from Hi-C data are sparse in SD regions due to poor mapping quality at similar paralogs. To our knowledge, one method exists (mHiC) that allocates multi-mapping reads to their possible alignments (Zheng et al. 2019), similar to CSEM. Future work might discover interactions within duplicated regions by reanalyzing chromatin conformation data in the most recent human reference (GRCh38) with this tool.

4.12 SUPPLEMENTARY MATERIALS AND METHODS

Quantification of HSD gene expression

The following aligned Iso-Seq filtered alignments were obtained from the ENCODE portal (Davis et al. 2018) (https://www.encodeproject.org/): ENCFF225CCJ, ENCFF648NAR, ENCFF192PJS, ENCFF538BNH, ENCFF596ODX, ENCFF694CBG, ENCFF049QGQ, ENCFF846YHI, ENCFF600MGT, ENCFF810FRP, ENCFF479SQR, ENCFF504GVG, ENCFF731THW, ENCFF936VUF, ENCFF939EUU, ENCFF100RGC, ENCFF927MKK, ENCFF292UIE, ENCFF738RAA, ENCFF437SYY, ENCFF989FKA, ENCFF911RNV, ENCFF305AFY, ENCFF016SHE, ENCFF479EHE, ENCFF914XOH, ENCFF054YYA, ENCFF470UHX, ENCFF158KCA, ENCFF757LOZ, ENCFF809QBD, ENCFF779VVX, ENCFF971JDY, ENCFF319JFG, ENCFF901XCR, ENCFF117DUA, ENCFF772MSZ, ENCFF644PGG, ENCFF955XSL, ENCFF509GHY, ENCFF803KIA, ENCFF058HQU, ENCFF973OML, ENCFF745HHL, ENCFF472TSL. Reads were counted per HSD gene with HTSeq (Anders, Pyl, and Huber 2015) before calculating RPKM values. For Figure S4.3B, *DUSP22* and *DUSP22B*

reads were counted separately based on PSV sequence using SAMtools mpileup and the raw alignments in the following accessions: ENCFF132HLS, ENCFF234YIJ, ENCFF407TMX, ENCFF592BQN, ENCFF615XZM, ENCFF810UWA.

Human and chimpanzee RNA-seq data were quantified alignment-free with a custom reference transcriptome. 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. (2018)) as a 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, Love, and Robinson 2015a).

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 S4.2). 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) (https://www.encodeproject.org/) 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, Lohse, and Usadel 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 multi-mapping 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 domains, enriched regions were identified with hiddenDomains, using the default settings (Starmer and Magnuson 2016). Paired-end long-ChIP reads were generated as described above. 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-MEM 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. Sets of peaks were compared between analysis methods using HOMER mergePeaks (parameters: "-d given") (Heinz et al. 2010) and a unidirectional correlation metric derived from IntervalStats using peaks with an overlap *p*-value below 0.05 (Chikina and Troyanskaya 2012).

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 kbp were amplified with KpnI and SacI restriction sites included in primers (Table S8) and cloned into the luciferase reporter vector pGL3-basic (Promega). Candidate enhancers within 50 kbp 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 kbp, 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, $\sim5\times10^5$ 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.

Chapter 5:
High-throughput characterization of *cis*-regulatory activity in human-specific duplications

5.1 ABSTRACT

Human-specific segmental duplications (HSDs) contain millions of base pairs of sequence unique to the human genome, including a number of genes recently implicated in driving neurodevelopment. Notably, 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. 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 (MPRA) of 8,145 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, mostly with small fold-differences. Combining measured activity levels across all assayed sequence, predicted differences in *cis*-regulatory activity did not correlate with mRNA levels. However, we identified four regions within derived *SRGAP2C* introns with greater than two-fold differences from the ancestral *SRGAP2* that may contribute to paralog-specific expression and thereby to human-specific traits. In all, this work suggests that functional divergence of duplicated CREs contributes to regulatory divergence of HSD genes and uncovers candidate drivers of human-specific regulatory patterns.

5.2 CONTRIBUTIONS

This study was conceived and designed by Colin Shew and Megan Y. Dennis. Experimental work was performed by Gulhan Kaya, Sean McGinty, and Colin Shew. Data analysis was conducted by Colin Shew.

5.3 INTRODUCTION

Gene duplication is a major contributor to 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 are thought to contribute to the evolution of new traits by facilitating relaxed selection via genetic redundancy (Ohno 1970; Lynch and Conery 2000; Kondrashov et al. 2002). While the vast majority of gene duplications 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 (Zhang 2003). Expression divergence is likely integral to this process, as the loss of *cis*-regulatory elements may drive expression partitioning of daughter paralogs, making them non-redundant and thus subject to purifying selection (Force et al. 1999). Indeed, paralogs exhibiting expression divergence are predicted to persist, and may subsequently accrue additional regulatory or functional changes (Rodin and Riggs 2003). Further, gene regulation is highly plastic, and a major driver of phenotypic evolution generally; alterations to spatiotemporal expression patterns are largely modular and leave the coding sequence of genes intact, making them less likely to be deleterious (Prud'homme, Gompel, and Carroll 2007). While duplications that occur in tandem tend to leave daughter paralogs in similar *cis*-regulatory environments, segmental duplications (SDs) in great apes are often interspersed hundreds of kilobases and typically involve concomitant rearrangements, like inversions (Marques-Bonet, Girirajan, and Eichler 2009; Lan and Pritchard 2016) . Since SDs are enriched along the great ape lineage

(Marques-Bonet et al. 2009), characterization of the regulatory landscape of these loci will improve understanding of the evolution of the human genome, and those of our closest relatives.

Human-specific segmental duplications (HSDs) are SDs unique to our species and consequently young (<6 million years) and highly similar (>99% nucleotide identity). Remarkably, genes within HSDs exhibit tissue-specific expression patterns across diverse primary tissues and cell lines (Dennis et al. 2017). Much attention has been focused on the role of HSD genes in neurodevelopment, and in addition to their novel biochemical functions, more than half of HSD gene families assayed display quantitative and spatial expression patterns specific to derived paralogs (Florio et al. 2018). For example, human-specific *ARHGAP11B* drives basal neuronal progenitor proliferation and increased cortical neuron numbers in mammalian models (Florio et al. 2015; Kalebic et al. 2018; Heide et al. 2020), and is preferentially expressed in the germinal zone of the developing brain, while the ancestral *ARHGAP11A* is expressed at higher levels and more broadly in the cortex (Florio et al. 2018). *ARHGAP11B* has attained novel biochemical functions (Namba et al. 2020), so its refined expression patterns are likely relevant to human brain development. The mechanisms underlying this high degree of cell type specificity are not well understood. Long read isoform sequencing of HSD genes identified extensive restructuring of gene models following duplication, and greater expression divergence was observed for 5'-truncated genes than 3' truncations or whole duplications, suggesting a role for *cis*-regulatory mechanisms (Dougherty et al. 2018). The authors also speculate that incomplete duplication of the suite of enhancers belonging to *CD8B* to *CD8B2* may contribute to its loss of expression in T cells, a mechanism which may be common in SDs, which typically reside hundreds or more kilobase pairs from their locus of origin. We also found evidence for paralog-specific regulatory contributions from adjacent non-duplicated sequence, as well as sequence-driven changes to the activity of duplicated *cis*-regulatory elements (CREs) (Shew et al. 2021). 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. We implemented an MPRA in the LCL GM12878 and SH-SY5Y neuroblastoma cells to measure regulatory activity of 8,145 paralogous human and orthologous chimpanzee sequences belonging to 2,675 homologous sites. Our findings suggest that individual paralogous changes mostly have a small effect size and that massively parallel screening effectively identifies functionally diverged CREs. We propose candidates for functional investigation from this differentially active set, including a duplicated intronic enhancer with *SRGAP2C*-specific activity loss in both cell types, which could contribute to human-specific expression patterns.

5.4 RESULTS

5.3.1 Most duplicated CREs exhibit similar activity levels

To characterize patterns of regulatory activity in HSDs, we designed an MPRA to directly compare paralogous and orthologous sequences comprising 75 genes from 30 families for which accurate paralog-resolved assemblies were available (Figure 5.1A) (Dennis et al. 2017). We chose to focus on primary fetal and adult prefrontal cortex due to the demonstrated role of HSD genes in cortical development, as well as LCLs due to their accessibility for humans and non-human primates. After re-mapping H3K27ac chromatin immunoprecipitation sequencing (ChIP-seq) and chromatin accessibility data (McVicker et al. 2013; Degner et al. 2012; Bryois et al. 2018; Vermunt et al. 2016; Reilly et al. 2015; de la Torre-Ubieta et al. 2018) to the human reference (GRCh38) with multiple alignments, we tiled these candidate CREs with 200mer

sequences at 2x density and identified homologous but non-identical tiles in the human and chimpanzee (panTro6) assemblies. In total, 8,145 test sequences and controls (N=500 LCL, 244 SH-SY5Y, 255 VISTA forebrain enhancer, and 500 scramble negatives) were synthesized, packaged into lentiviral vectors, and assayed in GM12878 and SH-SY5Y cells according to the lentiMPRA protocol (Gordon et al. 2020). 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 (integrated) and RNA (transcribed) libraries, respectively; Figure S5.1).

We next identified active and differentially active sequences in both cell lines using MPRAnalyze (Ashuach et al. 2019). In SH-SY5Y, 1,056/7,780 (14%) assayable sequences were active relative to the negative controls, while in GM12878 1,845/7,748 (24%) were active. Low p-values were abundant relative to a uniform distribution, indicating an enrichment for true active sequences (Figure S5.2). While a large proportion of human sequences scored as differentially active relative to the chimpanzee ortholog (of pairwise comparisons with at least one active sequence, 368/1,044 [35%] in SH-SY5Y and 175/5,037 in GM12878), less than 1% had a large fold-difference (eight loci greater than two-fold difference in SH-SY5Y and six in GM12878) (Figure 5.1B). We next considered whether relative gains or losses of activity were biased overall, in ancestral versus derived sequences, or in transcription start site (TSS)-proximal versus distal sequences, but found similar numbers for all categories (overall 171 lower and 190 higher activity in human than chimpanzee). Finally, we categorized all 426 families with at least one active homolog according to the number and types of differentially active sequences, relative to chimpanzee (Figure 5.1C). Significant differences were not more or less likely to affect ancestral sequences than expected by chance (permutation test of significant differences within families).

We next considered whether differentially active sequences might be driven by common transcription factors (TFs) in either cell type, thus pointing to biological pathways targeted by regulatory evolution in HSDs. We used SEA (T. L. Bailey and Grant 2021) to assess whether

any TF binding sites (TFBSs) were overrepresented in sequences with human-higher versus human-lower activity, or higher versus lower regardless of species, and found no significantly enriched motifs after multiple testing correction. However, coordinated evolution across HSDs is not necessarily expected, and an MPRA of rapidly evolved loci in the human genome produced similar results (Uebbing et al. 2021).

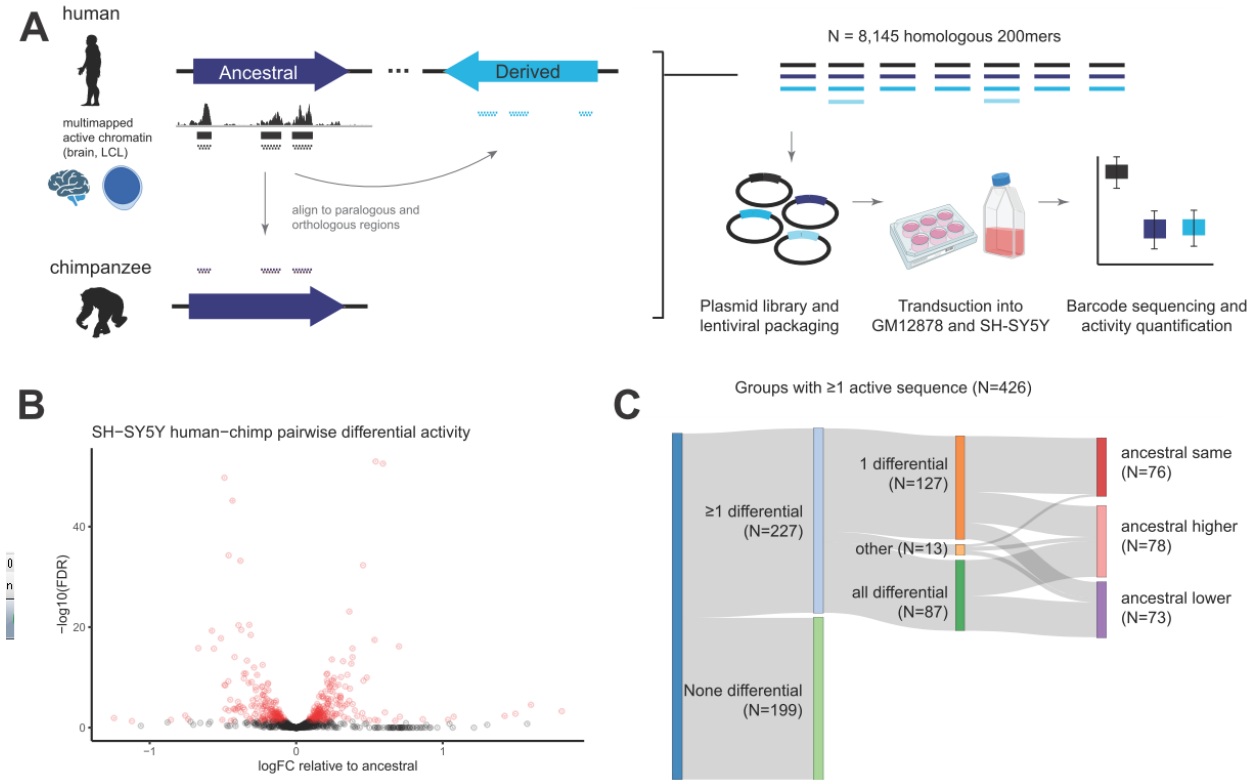


Figure 5.1. MPRA of HSD candidate enhancer sequences. (A) Illustration of experimental design; candidate CREs were identified by aligning H3K27ac ChIP-seq and chromatin accessibility (ATAC-seq or DNase-seq) to the human reference, allowing for multiple alignments. Peaks called in ancestral regions were used to design overlapping 200mers, and the corresponding paralogous human and orthologous chimpanzee sequences were also identified. The final pool of 8,145 test sequences (plus positive and negative controls) was cloned into a plasmid library, packaged in lentivirus, and transduced into GM12878 and SH-SY5Y. Sequence activity was quantified from RNA/DNA counts of associated barcodes. **(B)** Volcano plot of human-chimp pairwise comparisons in SH-SY5Y, for which at least one homolog was active. Significant differences (FDR<0.05) are colored red. **(B)** River plot classifying sequence groups with at least one active homolog by patterns of differential activity.

5.3.2 MPRA data agree with mRNA differential expression

Because our previous work identified examples of divergent *cis*-regulatory activity in HSDs, but not always in the expected direction, we sought to determine whether accounting for a more comprehensive suite of CREs would allow predictions of transcriptional divergence between paralogs. Inspired by the Activity-By-Contact (ABC) score and its implementation for predicting differential expression (Fulco et al. 2019); (Naqvi et al. 2022), we developed a similar metric using activity levels directly measured from the MPRA (Equation 5.1). Briefly, to predict differential expression of ancestral-derived paralog pairs, the baseline activity of active ancestral elements was weighted by the fold-difference of derived activity, if significantly different, as well as genomic distance (Equation 5.2). This ABC difference (ABCD) score was compared to the derived/ancestral mRNA expression ratio (as quantified by a Salmon, a probabilistic method robust to small sequence variation (Patro et al. 2017; Soneson, Love, and Robinson 2015b) for each gene pair in each cell type. For expressed derived-ancestral gene pairs (either TPM>1) in both cell types, expression divergence was not significantly correlated with the ABCD score ($p > 0.05$; Figure 5.2) While there was no relationship between expression estimates and ABCD scores, the sign of the two measures was concordant for a majority of derived-ancestral gene pairs (21/28 [75%] in SH-SY5Y and 16/23 [70%] in GM12878), a result that was maintained when only considering active and differentially active CREs. Still, the magnitude of divergence predicted was much smaller than that observed, with mRNA levels ranging over many more orders of magnitude. Taken together, these results indicate that comprehensive measurement of *cis*-regulatory activity may capture some information about paralogous regulatory divergence, but the duplicated 200mers assayed independently not able to predict these changes. This suggests that considering a larger sequence context, as well as the contributions of adjacent unique sequence, are likely critical to understanding the expression changes that have occurred in segmental duplications..

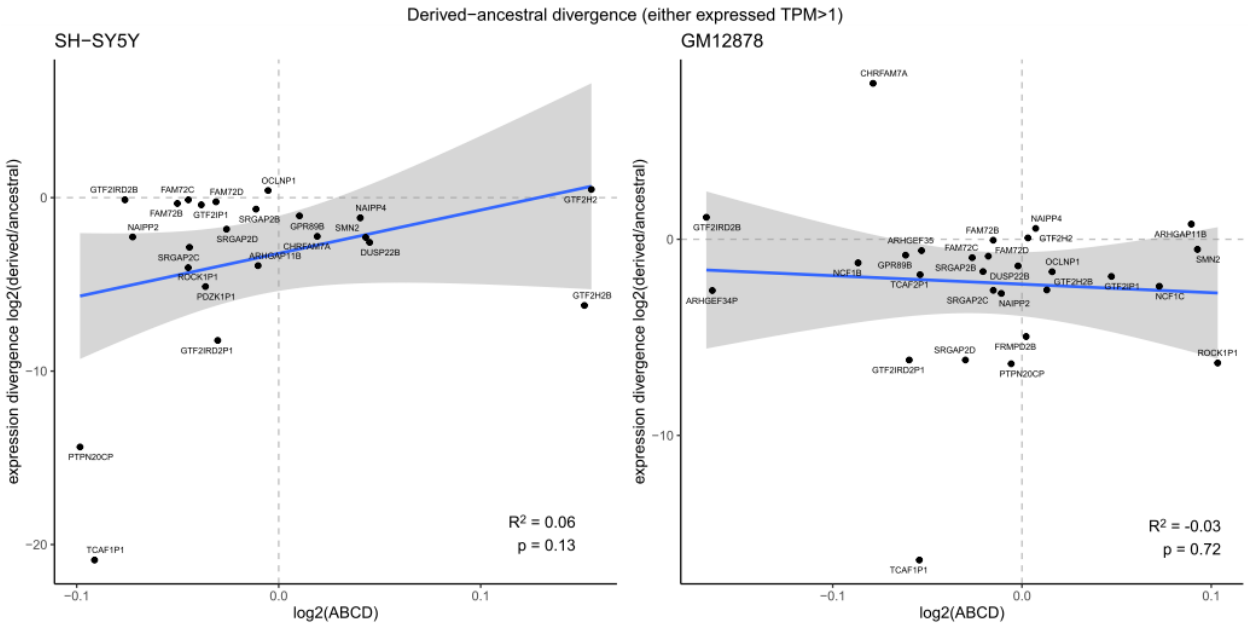


Figure 5.2. Prediction of differential regulation from MPRA data. Correlation of mRNA expression ratios to predicted divergence based on Activity-by-Contact difference (ABCD) score per derived-ancestral pair, calculated from CREs assayed in **(A)** SH-SY5Y and **(B)** GM12878.

5.3.3 Paralog-specific loss of enhancer activity in a SRGAP2C intron

While we are currently unable to generalize a relationship between the divergence of *cis*-regulation and gene expression, the MPRA highlighted paralog-specific changes to individual enhancers. Remarkably, two human-specific sequences with more than a two-fold activity difference relative to the ancestral were concordant in both GM12878 and SH-SY5Y: a loss of activity in a *SRGAP2C* intron (chr1:206333093-206333293) not observed in *SRGAP2* or *SRGAP2B* (Figure 5.3A–B), and a gain of activity of a sequence ~15 kbp upstream of *FRMPD2* (chr10:48140322-48140522), but not its human-specific paralog *FRMPD2B* (Tables S5.1 and S5.2). The *SRGAP2C* sequence contains three single-nucleotide variants relative to the ancestral *SRGAP2*, two of which were predicted to alter TFBS motifs (Figure 5.3C). Three additional *SRGAP2C* tiles showed strong paralog-specific activity, including one with more than

two-fold difference over the ancestral sequence in SH-SY5Y (Figure 5.3C). These enhancer regions could module expression or cell type specificity if targeted to *SRGAP2C*.

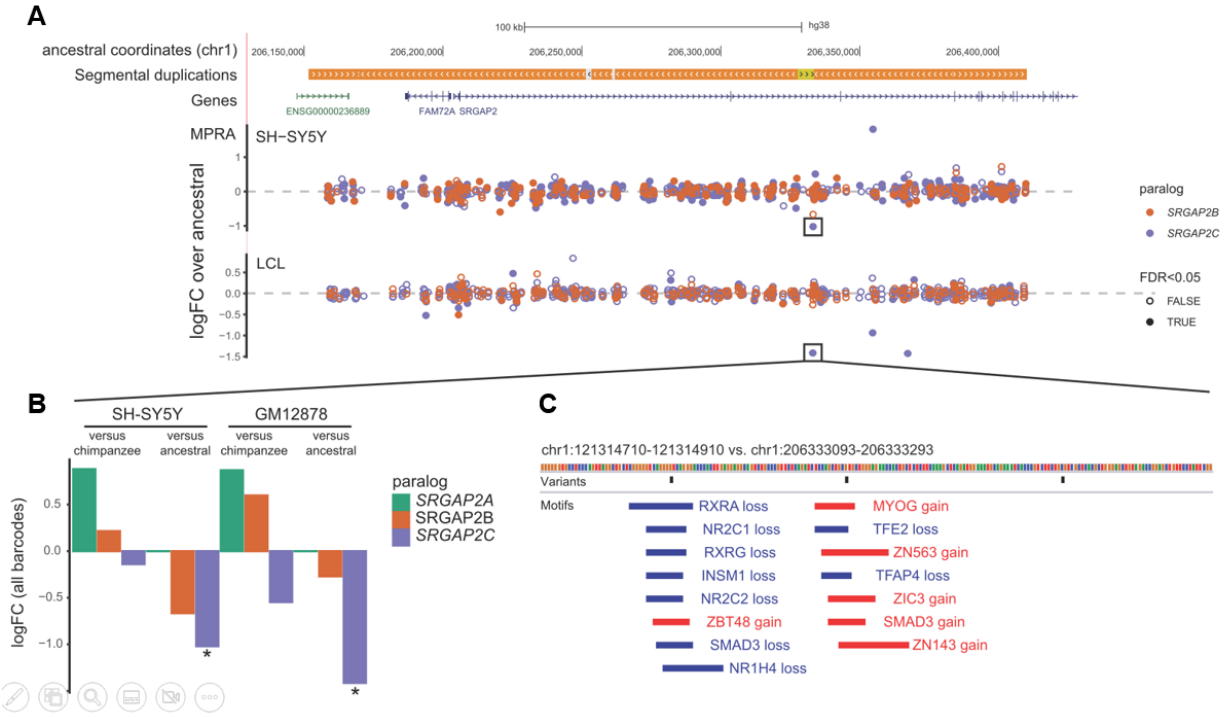


Figure 5.3. Paralog-specific activity loss of an enhancer in a *SRGAP2C* intron. (A) SH-SY5Y relative activity of derived paralogs *FAM72D*/*SRGAP2B* (orange) and *FAM72B*/*SRGAP2C* (purple), plotted on the ancestral *FAM72A*/*SRGAP2* locus. Comparisons are plotted as $\log_2$(fold-change) (logFC), with significant fold-differences (5% FDR) indicated as filled circles, and other comparisons plotted with open circles. Segmental duplications are indicated with orange bars. **(B)** LogFC for all comparisons of activity of chr1:206333093-206333293 and its paralogous sequences (black square in panel A). Significant differences (5% FDR) are shown with asterisks. **(C)** Annotated map of the *SRGAP2C* intronic 200mer sequence assayed, showing single-nucleotide differences relative to *SRGAP2* in black, and predicted *SRGAP2C*-specific changes to transcription factor binding site (TFBS) motifs in blue (loss) and red (gain).

5.5 DISCUSSION AND FUTURE DIRECTIONS

In this study, we presented the first high-throughput quantification of regulatory activity in duplicated regions. We noted a handful of example loci that exhibited strongly differential regulatory activity (near the genes *SRGAP2*, *PTPN20*, *GTF2H2*, and *CHRNA7;* Table S5.1, S5.2), though homologous sequences were broadly similar in activity. This is unsurprising given

the small number of differences 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 (Tewhey et al. 2016; Uebbing et al. 2021), However, this stands in contrast to examples of multi-fold differences of HSD promoter and enhancer activity demonstrated previously (Shew et al. 2021). One potential explanation may be the size of assayed sequences; MPRAs 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. Notably, this aspect of the technology is improving and can already be partially circumvented by assembling larger synthetic inserts from smaller oligos (Klein et al. 2020), which may increase the likelihood of detecting activity changes that depend on synergistic effects. Despite these limitations, MPRA is an ideal study design for directly comparing homologous sequences. We categorized the patterns of differential CRE activity of HSDs relative to chimpanzee, finding that about half of sequence families had no changes. Surprisingly, ancestral sequences were not less likely to change, even though these paralogs tend to maintain ancestral expression, and derived sequences were not biased toward reduced activity relative to chimpanzee. However, due to the small fold-difference of most sequence pairs, these figures may not necessarily be meaningful. We implemented an ABCD score to sum activity differences between ancestral and derived paralogs, accounting for the strength and distance of duplicated CREs, but were unable to predict expression differences with this information alone. Further refinements to the ABC score may help by factoring in the effect of adjacent non-duplicated sequences. Promoter truncations and exaptations from adjacent sequence are particularly likely to contribute a large effect to expression differences (Dougherty et al. 2018).

We succeeded in identifying individual differentially active elements that may contribute to paralog-specific gene regulation. One promising candidate located in a *SRGAP2C* intron exhibited a dramatic loss of activity unique to that paralog. Notably, *SRGAP2C* but not the other

human-derived genes *SRGAP2B* or *SRGAP2D* has a demonstrated role inhibiting the function of ancestral *SRGAP2*, resulting in prolonged radial migration of neurons, increased dendritic spine density, and increased synapse density (Charrier et al. 2012; Schmidt et al. 2019). *SRGAP2C* is expressed in the developing and adult brain, and at higher levels than the other human-specific paralogs (Dennis et al. 2012), it is less expressed than *SRGAP2* (Dennis et al. 2017; Shew et al. 2021). Given its mechanism of action by dimerization and dominant negative-like inhibition of the ancestral gene, expression modulation is likely critical to proper dosage. If this CRE indeed regulates *SRGAP2C*, it may play a role in reducing *SRGAP2C* expression. Understanding the cell type-specific activity of this and other CREs will also be critical to unraveling their function. We will next validate differential activity with luciferase reporters and examine single-cell ATAC-seq datasets from the developing brain to characterize enhancer accessibility patterns in different cell types.

To improve interpretation of variants contributing to differential activity of HSD genes, we will map expression quantitative loci (eQTLs) in HSDs. Previously, variants called in duplicated regions have been unreliable owing to false positives from paralogous variants or false negatives due to quality filtering (Aganezov et al. 2022), but a new set of variants identified from long-read capture of HSDs in diverse human populations will allow their analysis for the first time (Sekar and Soto et al., unpublished data). This will allow assessment of the overlap between variants in differentially active CREs and variants associated with expression levels, providing additional evidence for function and pointing to their target genes. In addition, these variants will be used to identify signatures of selection in noncoding regions, which will identify human-specific CREs that may contribute evolutionarily meaningful expression divergence.

Absent from this study is an analysis of chromatin conformation at HSD and orthologous chimpanzee loci. 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. We plan to reanalyze a high-depth human and chimpanzee iPSC Hi-C experiment (Eres et al. 2019) to address these questions. While published methods for allocation of multimapping Hi-C reads (Zheng, Ay, and Keles 2019) were not scalable to Hi-C maps of this size due to memory usage, we will analyze interactions overlapping singly unique kmers. In addition, we have generated long-read Hi-C libraries (Deshpande et al. 2022) with larger fragment sizes, allowing greater mappability in HSDs (Figure S5.3). While overall read depth will be lower, this will also serve to nominate additional candidate interactions not visible with short-read technologies.

Further, while identification of candidate CREs for the MPRA relied on multimapping short reads to generate a comprehensive list of elements to assay, new long read-based assays have recently been developed, allowing direct comparison of the epigenomic landscapes of paralogous regions. For example, DNA methylation was previously intractable using bisulfite sequencing with short reads, but cytosine methylation is directly available from Oxford Nanopore (ONT) basecalls from new “telomere-to-telomere” assemblies (Gershman et al. 2022). In addition, the ONT-based methods nanoNOMe and DiMeLo-seq allow for the characterization of open chromatin and target protein enrichment through the exogenous methylation of GpC dinucleotides and adenine residues, respectively (Gershman et al. 2022; Altemose et al. 2022; Lee et al. 2020). We will also mine these datasets for additional evidence of paralog-specific gene regulation.

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 differential activity, some may contribute to regulatory patterns of HSD genes such as *SRGAP2C*, driving unique features of our species. Integration of additional information, such as chromatin conformation, single nucleotide polymorphism, and long-read epigenomic data will be necessary to gain a more complete

picture of gene regulation within HSDs.

5.6 METHODS

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: (McVicker et al. 2013; Degner et al. 2012; Bryois et al. 2018; Vermunt et al. 2016; Reilly et al. 2015; de la Torre-Ubieta et al. 2018). In the case of Bryois *et al.*, only control samples were used. Reads were trimmed with Trimmomatic (Bolger, Lohse, and Usadel 2014) aligned to GRCh38 allowing for multiple mapping with bowtie v1.1.2 (Langmead et al. 2009) with the flags "-a -v2 -m 99". Peaks were called using MACS2 v2.1.2 (Liu 2014) 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 (Quinlan and Hall 2010): 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 roughly equalize sequence 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 (or a single HSD locus

if ancestral status unknown) and aligned with BLAT (Kent 2002) 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 (Dennis et al. 2017)). 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 (93% success rate). All tiles with this size 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 the 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.

Controls were generated from published MPRA data when available. For GM12878, active sequences from (Tewhey et al. 2018) were filtered for the reference haplotype, reference sequence exhibiting enhancer activity, and Bonferroni-corrected $p < 0.01$, and ranked by activity. Unique SNPs (only one strand kept per SNP) with an dbSNP identifier were used to extract a centered interval of 200 bp from GRCh37 (N=500 after restriction site filtering). SH-SY5Y positive controls were designed from (Myint 2019). Count data were obtained, and active sequences were defined with a 5% mean absolute deviation *p*-value from the MPRAnalyze (Ashuach et al. 2019) function analyzeQuantification(). The associated dbSNP identifiers were used to extract a 200-bp region from GRCh38 (final N=244). An additional 255 brain-positive sequences were sampled from random 200-bp windows within VISTA forebrain enhancers (Visel et al. 2007). Finally, 500 random ancestral tiles were selected and scrambled to generate

negative controls. Test and control 230mer sequences were synthesized by Agilent Technologies.

MPRA library preparation and sequencing

MPRA was carried out according to the lentiMPRA protocol as described (Gordon et al. 2020) with 3 technical replicates. For SH-SY5Y experiments, 2.5 million cells were transduced at a multiplicity of infection (MOI) of 40 in DMEM/F12 Glutamax medium with 10% FBS and 2.5 µg/m protamine sulfate. For GM12878, 55 million cells were transduced at an MOI of 13 in RPMI with 10% FBS and 8 µg/mL polybrene. Cells were incubated at 37 °C, 5% CO_2 and collected 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 (Gordon et al. 2020) 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 passed to MPRAnalyze (Ashuach et al. 2019). 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, with each homolog tracked in the annotation data, and barcodes were renumbered to reflect their unique identifiers. Models were fit in "scale" mode with the following model: "rnaDesign = ~homolog, reducedDesign = ~1". 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. Differential activity was also defined between cell types…

TFBS analysis

For each cell type, active and inactive sequences were defined by the median absolute deviation *p*-value from the quantification analysis, and filtered to keep only one homolog from each family of sequences (highest estimated transcription rate for active and lowest for inactive). Similarly, differentially active sequences relative to chimpanzee were separated into gains and losses (either human relative to chimp, or higher/lower regardless of species), and the homolog with the highest fold-difference in each family was selected. Sets of sequences were scanned for enrichment of HOCOMOCO v11 (Kulakovskiy et al. 2018) motifs using SEA (T. L. Bailey and Grant 2021). In the comparison of active vs. inactive sequences, enriched motifs were tested for enrichment for gene ontology terms with DAVID (Huang et al. 2007), using all HOCOMOCO TFs as a background list.

ABCD score

Predicted regulatory divergence between ancestral-derived pairs was calculated as in (Naqvi et al. 2022), using the ABC score framework defined in (Fulco et al. 2019). Briefly, ABC scores were calculated for all tested elements using the estimated ancestral transcription rate α from MPRAnalyze for activity and (genomic distance)$^{-0.7}$ as a proxy for contact frequency. ABC scores were used to weight the derived/ancestral fold-difference for all tested sequences within 5 Mbp of each TSS, and the resulting sum was considered to be the predicted fold-difference in expression.

RNA-seq analysis

RNA-seq data were obtained for GM12878 (ENCODE ENCSR000AEC, ENCSR000AEE, and ENCSR000CVT) and SH-SY5Y (Pezzini et al. 2017). Transcripts were quantified with Salmon v1.9.0 (Patro et al. 2017) 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 (Soneson, Love, and Robinson 2015b).

Luciferase validation

200mers 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 prewarmed 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.

5.7 ACKNOWLEDGMENTS

We thank Dr. Fumitaka Inoue and Dr. Tal Ashuach for technical advice related to MPRA library construction and data processing, as well as Dr. Sierra Nishizaki and Daniela C. Soto for valuable discussion of the bioinformatic analysis performed in this study. We also acknowledge Dr. Anthony Antonellis for sharing the pE1B enhancer reporter Gateway plasmid, as well as Dr. Gerald Quon and Dr. Siobhan Brady for constructive feedback on the manuscript. This work was supported by the National Human Genome Research Institute (F31HG011205 to C.S.) and National Institute of Neurological Disorders and Stroke (R00NS083627 to M.Y.D.), and the Office of the Director and National Institute of Mental Health (DP2 OD025824 to M.Y.D.) at the National Institutes of Health (NIH). Additionally, M.Y.D. is supported as a Sloan fellow (FG-2016-6814).

5.8 SUPPLEMENTARY FIGURES

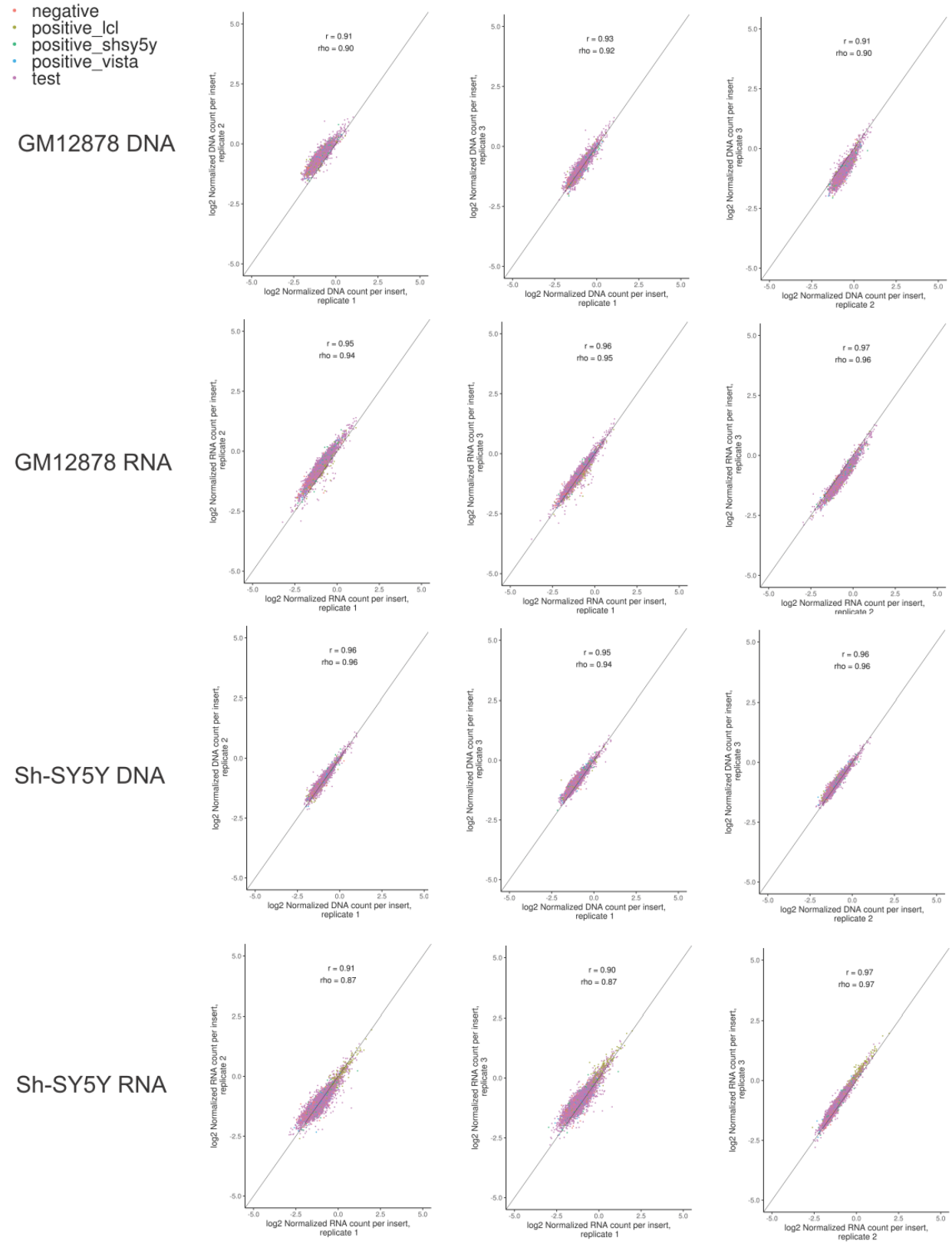

negative
positive_lcl
positive_shsy5y
positive_vista
test
GM12878 DNA
GM12878 RNA
Sh-SY5Y DNA
Sh-SY5Y RNA
r = 0.91
rho = 0.90
r = 0.93
rho = 0.92
r = 0.91
rho = 0.90
r = 0.95
rho = 0.94
r = 0.96
rho = 0.95
r = 0.97
rho = 0.96
r = 0.96
rho = 0.96
r = 0.95
rho = 0.94
r = 0.96
rho = 0.96
r = 0.91
rho = 0.87
r = 0.90
rho = 0.87
r = 0.97
rho = 0.97
log2 Normalized DNA count per insert, replicate 1
log2 Normalized DNA count per insert, replicate 2
log2 Normalized DNA count per insert, replicate 3
log2 Normalized RNA count per insert, replicate 1
log2 Normalized RNA count per insert, replicate 2
log2 Normalized RNA count per insert, replicate 3

Figure S5.1. Reproducibility of MPRA libraries. Plots of normalized barcode counts for all pairs of replicates (DNA, RNA, SH-SY5Y, GM12878). The black line tracks x=y, and the Pearson (*r*) and Spearman (rho) correlation coefficients are shown. Insert sequences are colored by category (test, positive, or negative control),

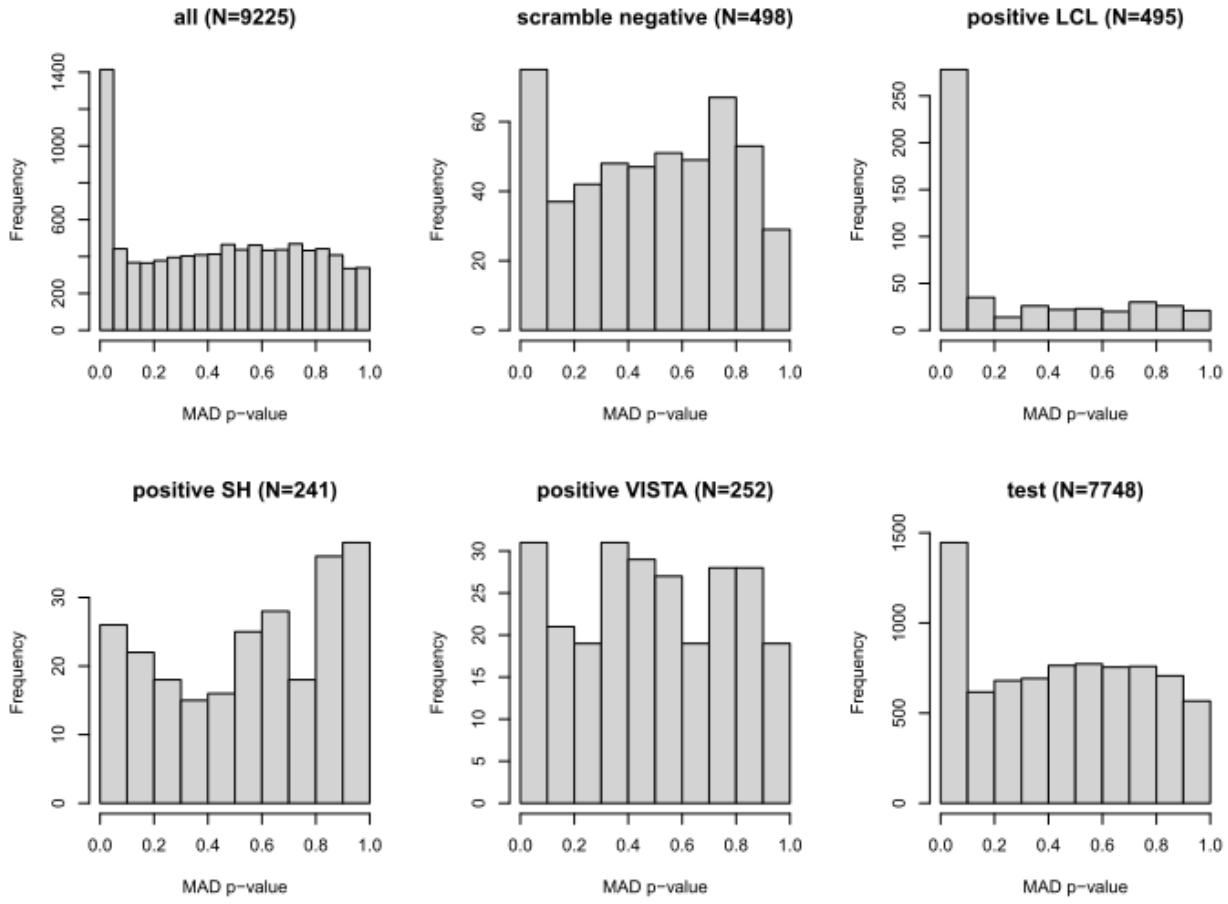


Figure S5.2. *P*-value distribution for identification of active sequences relative to scramble controls. Sequences are separated by control type (negative, LCL positive, SH-SY5Y positive, VISTA positive, and test) for experiments performed in **(A)** SH-SY5Y and **(B)** GM12878.

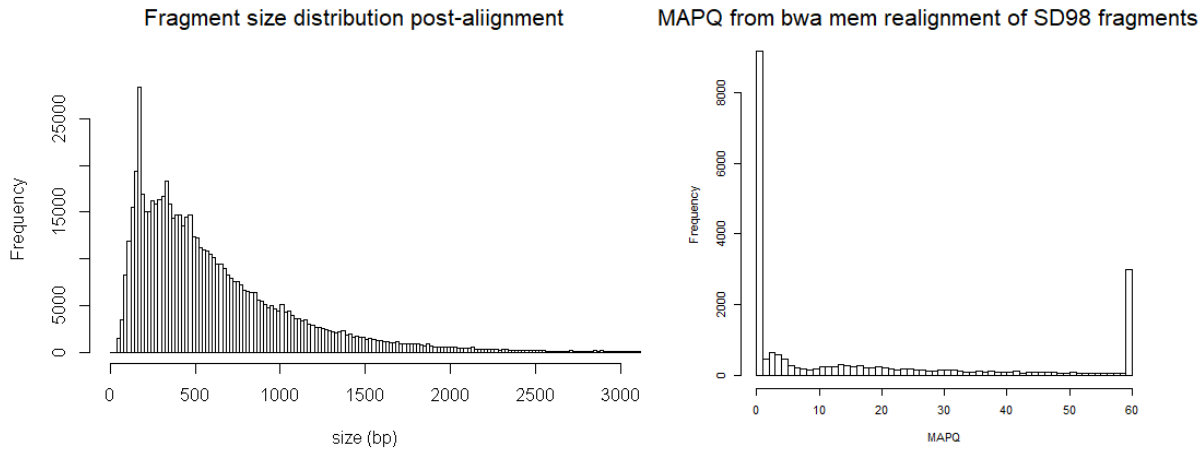


Figure S5.3: Mapping statistics from pilot long-read Hi-C. Histograms show the distribution of fragment sizes from sequenced concatemers (left) and BWA MEM MAPQ scores in SD with at least 98% identity (SD98) (right).

5.9 SUPPLEMENTARY TABLES

Table S5.1: Sequences with greater than two-fold change relative to ancestral in SH-SY5Y.

Test sequence	*Ancestral locus (hg38)*	*Nearest gene*	*logFC*	*Wald statistic*	*p-value*	*FDR*	*N deletions*
panTro6.chr15:10432929-10433129	chr15:32208666-32208866	CHRNA7	-1.12	18.56	1.65E-05	9.99 E-05	4
panTro6.chr15:10571507-10571707	chr15:32345908-32346108	ARHGAP11 /CHRNA7	-1.07	9.92	1.63E-03	4.41 E-03	0
panTro6.chr5_NW_019932884v1_random:170217-170417	chr5:69561512-69561712	GTF2H2	-1.04	6.90	8.63E-03	1.71 E-02	2
chr1:121314710-121314910	chr1:206333093-206333293	SRGAP2C	-1.02	10.54	1.17E-03	3.40 E-03	0
chr1:121336315-121336515|chr1:144943550-144943750	chr1:206354652-206354852	SRGAP2C, SRGAP2B	1.81	15.49	8.29E-05	3.88 E-04	0
GPRIN2_dist:151784-151984	chr10:46556008-46556208	GPRIN2B	1.23	9.85	1.70E-03	4.57 E-03	0

chr10:46911343-46911543	chr10:48140322-48140522	PTPN20	1.60	12.55	3.96E-04	1.40 E-03	-5
chr15:30321686-30321886	chr15:32208666-32208866	CHRFAM7A	-1.30	12.69	3.67E-04	1.33 E-03	8
chr5:69624353-69624553	chr5:70997615-70997815	GTF2H2C	-2.11	4.86	2.75E-02	4.19 E-02	0
chr7:75093531-75093731	chr7:74850409-74850609	GTF2IRD2B	1.37	12.47	4.13E-04	1.44 E-03	-1

Table S5.2: Sequences with greater than two-fold change relative to ancestral in GM12878.

Test sequence	*Ancestral locus (hg38)*	*Nearest gene*	*logFC*	*Wald statistic*	*p-value*	*FDR*	*N deletions*
panTro6.chr1:181584483-181584683	chr1:206367460-206367660	chimp SRGAP2	-1.30	13.40	2.52E-04	1.21 E-02	0
panTro6.chr7:72475913-72476113	chr7:74800500-74800700	chimp GTF2IRD2	1.04	14.34	1.53E-04	9.12 E-03	4
hg38.chr1:121314710-121314910	chr1:206333093-206333293	SRGAP2C	-1.41	26.25	2.99E-07	1.07 E-04	0
hg38.chr1:144931057-144931257	chr1:206367460-206367660	SRGAP2C	-1.43	16.10	6.02E-05	5.16 E-03	0
hg38.chr10:46911343-46911543	chr10:48140322-48140522	PTPN20	1.41	11.76	6.06E-04	2.15 E-02	-5
hg38.chr7:75157168-75157368	chr7:74788585-74788785	GTF2IRD2B	-2.04	21.23	4.07E-06	8.43 E-04	0

5.10 EQUATIONS

Equation 5.1: Activity-by-Contact (ABC) score, adapted for MPRA the power law model presented by Fulco et al. (2019). The contribution of a given CRE to a gene G is calculated as the product of its activity (the transcription rate α as modeled by MPRAnalyze (Ashuach et al. 2019)) and its genomic distance from the transcription start site (dist) to the power of −0.7, which serves as a proxy for contact frequency. ABC scores are normalized to represent relative contributions for all CREs assigned to a gene.

$$ABC_{CRE,G} = \frac{\alpha * dist^{-0.7}}{\sum(\alpha * dist^{-0.7})}$$

Equation 5.2: ABC difference (ABCD) score, adapted for MPRA from Naqvi et al. (2022). The ABCD score for two paralogous genes G1 and G2 is calculated as a weighted sum of the fold-changes (FC) in activity of paralogous CREs in G2 relative to G1. The ABC score for the CRE in G1 is used as a weight. A window of size *w* relative to the transcription start site is used to define the set of CREs considered.

$$ABCD_{G1,G2} = \sum_{CRE\ within\ \mathrm{w}} ABC_{CRE,G1} * FC_{G2/G1}$$

Chapter 6:
Conclusions and discussion

6.1 SUMMARY AND IMPACTS

After identifying novel SVs between humans and non-human primates, we found that chromatin architecture is generally preserved between these species, and SVs were depleted for positions predicted to alter TAD structure as defined in human LCLs. Though the popular notion that TADs are conserved across species has been called into question (Eres and Gilad 2021), our results are in line with previous work in humans and other primates that finds deletions impacting TAD structure likely experience negative selection (Lazar et al. 2018; Fudenberg and Pollard 2019; Huynh and Hormozdiari 2019). That we also implicate inversions in the same process suggests a more general rearrangement-driven phenomenon. Further, functional dissection of the three-dimensional architecture of individual loci indicates unequivocally that promoter-enhancer contacts and insulating elements can mediate SV-induced regulatory changes (Lupiáñez et al. 2015; Despang et al. 2019). Given that we observed an enrichment for differentially expressed genes near SVs, and that the definition of chromatin domains can be sensitive to the method and parameter tuning (Forcato et al. 2017), we suggest that it may not be as important to consider TADs *per se* but the observed promoter-enhancer contacts at a given locus, which requires higher resolution.

Another major finding of this work is that *cis*-regulatory activity has diverged, in some cases dramatically, between evolutionarily recent (<6 million years) segmental duplications.Our studies were the first to systematically test this, and while we not able to model paralogous regulatory divergence in a general sense, evidence suggests that contributions are likely to be locus specific; we highlighted individual examples of activity changes to promoters and

enhancers (*DUSP22* and *NCF1*), paralog-specific eQTLs and chromatin loops, and changes to gene structure across breakpoints. The MPRA also identified a few locations with strong paralog specific activity in *SRGAP2C* relative to ancestral *SRGAP2*, which are prime candidates to investigate for human-specific expression patterns in the developing brain. Finally, comparative expression analysis revealed that derived HSD genes are less likely to retain ancestral properties than the ancestral locus itself; it may be unlikely that the full suite of CREs is duplicated with a given gene, making true functional redundancy a rarity in the context of single-locus duplications. This work suggests differences in evolutionary pressures in SDs compared to whole-genome duplication (WGD); in WGD, subfunctionalization may be common (Braasch et al. 2016), while we found no compelling examples of this in HSDs.

6.2 FUTURE STUDIES

In addition to the previously proposed work to complement the experiments described in Chapter 5, functional validation will be necessary to confirm that CREs not only possess regulatory activity in an integrated reporter construct, but also contribute to transcriptional activation endogenously. After predicting the target genes of differentially active HSD enhancer sequences, follow-up work should establish a causal relationship between these CREs and genes. This is feasible at the family level, in which all paralogs can be targeted for deletion or epigenetic silencing by CRISPR editing or CRISPRi. In some cases, paralog-specific mRNA quantification is possible by qRT-PCR, as performed in Chapter 4.

A central challenge in studying gene regulation is that cell type-specificity necessitates trade-offs in performing experiments in a feasible model, which may only represent one or a few cell types *in vitro* or limited quantities of tissue *ex vivo*. In addition to performing experiments in more complex models (cortical organoids or mice, for example), we can generate additional hypotheses about regulatory divergence in a wide array of tissues and cell types by examining the predictions from machine learning methods. For example, Enformer is a deep learning

approach trained on over 5,000 epigenomic inputs to predict transcriptional activity to high accuracy (mean correlation to experimental data 0.85) (Avsec et al. 2021). Similarly, much effort has been focused on predicting three-dimensional chromatin structure from the genome itself, with models like Akita (Fudenberg, Kelley, and Pollard 2020). While caution should be exercised when interpreting machine learning predictions, these approaches entirely circumvent the multiple mapping problem in SDs and are becoming increasingly accurate and can inform experimental design to identify duplicated enhancers with paralog-specific activity and connectivity.

6.3 SHORTCOMINGS

While this work was motivated by the goal of identifying the genetic basis of species-specific traits, and succeeded in nominating candidate regulatory changes, demonstration of relevance to phenotypic differences between species ultimately requires functional evidence. Given the variety of divergent characteristics between humans and non-human primates, such as brain development, immune function, and musculoskeletal morphology, the experiments required will depend on the candidates being tested. For example, to implicate *SRGAP2C*-specific regulatory activity in humanizing neuronal migration patterns and dendritic spine morphology, these phenotypes could be quantified in a mouse model engineered with the *SRGAP2C* sequence of this locus, to determine if radial migration or dendritic spine density are increased (Charrier et al. 2012). Another avenue worth considering is to measure the effect of *NCF1* or *DUSP22* paralogs on T cell activation *in vitro*, to determine if their dosage might impact immune response. This could be achieved by transfecting human primary T cell cultures with expression constructs and assaying T cell proliferation, activation marker (IL2RA,CD25) upregulation, or effector cytokine production (IFN-γ, TNF-α) (Zappasodi et al. 2020).

Crucially, the interpretation of paralog-specific expression and regulatory activity depends on future characterization of single-nucleotide polymorphism across HSDs. Some

variants shared between paralogs are a result of interlocus gene conversion (Dumont 2015), which may mean that some of the regulatory divergence studied in this work is subject to allelic as well as paralogous variation. Ongoing work identifying SNVs in HSD with long-read capture sequencing (Sekar and Soto *et al.*, unpublished) and highly contiguous assemblies from the Human Pangenome Reference Consortium (HPRC) will allow accurate distinction of PSVs from SNVs. Nevertheless, the PSVs used in these studies were based on high quality assemblies from a true human haplotype (CHM1) (Dennis et al. 2017).

6.4 OUTLOOK

Advances in long-read sequencing and associated epigenomic assays will continue to propel research on duplications and other SVs in humans and other species. In the next few years, hundreds of telomere-to-telomere genomes from a diverse cohort are expected from the HPRC (Wang et al. 2022), and gapless genomes from non-human primates are also anticipated. These data will create a more complete picture of the evolutionary history, population variation, and selective pressures on SVs, and the availability of multiple high-quality assemblies will also open new HSD loci to study. Long-read methods for assaying gene regulation, such as Iso-seq, DiMeLo-seq, and nanoNOMe will all benefit from improved quality and throughput as cost and technical barriers to PacBio and ONT sequences continue to fall. These data types promise to allow distinction of transcribed, accessible, and active histone-associated DNA between paralogs and across SV haplotypes with relative ease. While it is already clear that structural variation and gene regulation are intimately linked, these fields and the overlap between them are poised for a productive future.

References

1000 Genomes Project Consortium, Adam Auton, Lisa D. Brooks, Richard M. Durbin, Erik P. Garrison, Hyun Min Kang, Jan O. Korbel, et al. 2015. “A Global Reference for Human Genetic Variation.” *Nature* 526 (7571): 68–74.

Aganezov, Sergey, Stephanie M. Yan, Daniela C. Soto, Melanie Kirsche, Samantha Zarate, Pavel Avdeyev, Dylan J. Taylor, et al. 2022. “A Complete Reference Genome Improves Analysis of Human Genetic Variation.” *Science* 376 (6588): eabl3533.

Alexander, David H., John Novembre, and Kenneth Lange. 2009. “Fast Model-Based Estimation of Ancestry in Unrelated Individuals.” *Genome Research* 19 (9): 1655–64.

Alkan, Can, Bradley P. Coe, and Evan E. Eichler. 2011. “Genome Structural Variation Discovery and Genotyping.” *Nature Reviews Genetics*. https://doi.org/10.1038/nrg2958.

Altemose, Nicolas, Annie Maslan, Owen K. Smith, Kousik Sundararajan, Rachel R. Brown, Reet Mishra, Angela M. Detweiler, et al. 2022. “DiMeLo-Seq: A Long-Read, Single-Molecule Method for Mapping Protein-DNA Interactions Genome Wide.” *Nature Methods* 19 (6): 711–23.

Anders, Simon, Paul Theodor Pyl, and Wolfgang Huber. 2015. “HTSeq--a Python Framework to Work with High-Throughput Sequencing Data.” *Bioinformatics* 31 (2): 166–69.

Andres, A. M., M. Y. Dennis, W. W. Kretzschmar, J. L. Cannons, S. Q. Lee-Lin, B. Hurle, Nisc Comparative Sequencing Program, et al. 2010. “Balancing Selection Maintains a Form of ERAP2 That Undergoes Nonsense-Mediated Decay and Affects Antigen Presentation.” *PLoS Genetics* 6 (10): e1001157.

Antonacci, Francesca, Megan Y. Dennis, John Huddleston, Peter H. Sudmant, Karyn Meltz Steinberg, Jill A. Rosenfeld, Mattia Miroballo, et al. 2014. “Palindromic GOLGA8 Core Duplicons Promote Chromosome 15q13.3 Microdeletion and Evolutionary Instability.” *Nature Genetics* 46 (12): 1293–1302.

Antonacci, Francesca, Jeffrey M. Kidd, Tomas Marques-Bonet, Mario Ventura, Priscillia Siswara, Zhaoshi Jiang, and Evan E. Eichler. 2009. “Characterization of Six Human Disease-Associated Inversion Polymorphisms.” *Human Molecular Genetics* 18 (14): 2555–66.

Antonellis, Anthony, Jimmy L. Huynh, Shih-Queen Lee-Lin, Ryan M. Vinton, Gabriel Renaud, Stacie K. Loftus, Gene Elliot, et al. 2008. “Identification of Neural Crest and Glial Enhancers at the Mouse Sox10 Locus through Transgenesis in Zebrafish.” *PLoS Genetics* 4 (9): e1000174..

Arnone, M. I., and E. H. Davidson. 1997. “The Hardwiring of Development: Organization and Function of Genomic Regulatory Systems.” *Development* 124 (10): 1851–64.

Ashuach, Tal, David S. Fischer, Anat Kreimer, Nadav Ahituv, Fabian J. Theis, and Nir Yosef. 2019. “MPRAnalyze: Statistical Framework for Massively Parallel Reporter Assays.” *Genome Biology* 20 (1): 183.

Assis, Raquel, and Doris Bachtrog. 2013. “Neofunctionalization of Young Duplicate Genes in Drosophila.” *Proceedings of the National Academy of Sciences of the United States of America* 110 (43): 17409–14.

Audano, Peter A., Arvis Sulovari, Tina A. Graves-Lindsay, Stuart Cantsilieris, Melanie Sorensen, Annemarie E. Welch, Max L. Dougherty, et al. 2019. “Characterizing the Major Structural Variant Alleles of the Human Genome.” *Cell* 176 (3): 663–75.e19.

Avsec, Žiga, Vikram Agarwal, Daniel Visentin, Joseph R. Ledsam, Agnieszka Grabska-Barwinska, Kyle R. Taylor, Yannis Assael, John Jumper, Pushmeet Kohli, and David R. Kelley. 2021. “Effective Gene Expression Prediction from Sequence by Integrating Long-Range Interactions.” *Nature Methods* 18 (10): 1196–1203.

Bailey, Jeffrey A., and Evan E. Eichler. 2006. “Primate Segmental Duplications: Crucibles of Evolution, Diversity and Disease.” *Nature Reviews. Genetics* 7 (7): 552–64.
Bailey, Timothy L., and Charles E. Grant. 2021. “SEA: Simple Enrichment Analysis of Motifs.” *bioRxiv*. https://doi.org/10.1101/2021.08.23.457422.
Balogh, Andrea, Eszter Toth, Roberto Romero, Katalin Parej, Diana Csala, Nikolett L. Szenasi, Istvan Hajdu, et al. 2019. “Placental Galectins Are Key Players in Regulating the Maternal Adaptive Immune Response.” *Frontiers in Immunology* 10 (June): 1240.
Barreiro, Luis B., John C. Marioni, Ran Blekhman, Matthew Stephens, and Yoav Gilad. 2010. “Functional Comparison of Innate Immune Signaling Pathways in Primates.” *PLoS Genetics* 6 (12): e1001249.
Bayés, Mònica, Luis F. Magano, Núria Rivera, Raquel Flores, and Luis A. Pérez Jurado. 2003. “Mutational Mechanisms of Williams-Beuren Syndrome Deletions.” *American Journal of Human Genetics* 73 (1): 131–51.
Blake, Lauren E., Julien Roux, Irene Hernando-Herraez, Nicholas E. Banovich, Raquel Garcia Perez, Chiaowen Joyce Hsiao, Ittai Eres, Claudia Cuevas, Tomas Marques-Bonet, and Yoav Gilad. 2020. “A Comparison of Gene Expression and DNA Methylation Patterns across Tissues and Species.” *Genome Research* 30 (2): 250–62.
Bolger, Anthony M., Marc Lohse, and Bjoern Usadel. 2014. “Trimmomatic: A Flexible Trimmer for Illumina Sequence Data.” *Bioinformatics* 30 (15): 2114–20.
Braasch, Ingo, Andrew R. Gehrke, Jeramiah J. Smith, Kazuhiko Kawasaki, Tereza Manousaki, Jeremy Pasquier, Angel Amores, et al. 2016. “The Spotted Gar Genome Illuminates Vertebrate Evolution and Facilitates Human-Teleost Comparisons.” *Nature Genetics* 48 (4): 427–37.
Brawand, D., M. Soumillon, A. Necsulea, P. Julien, G. Csardi, P. Harrigan, M. Weier, et al. 2011. “The Evolution of Gene Expression Levels in Mammalian Organs.” *Nature* 478 (7369): 343–48.
Bridges, Calvin B. 1936. “The Bar ‘Gene’ a Duplication.” *Science*. https://doi.org/10.1126/science.83.2148.210.
Brunetti-Pierri, Nicola, Jonathan S. Berg, Fernando Scaglia, John Belmont, Carlos A. Bacino, Trilochan Sahoo, Seema R. Lalani, et al. 2008. “Recurrent Reciprocal 1q21.1 Deletions and Duplications Associated with Microcephaly or Macrocephaly and Developmental and Behavioral Abnormalities.” *Nature Genetics*. https://doi.org/10.1038/ng.279.
Bryois, Julien, Melanie E. Garrett, Lingyun Song, Alexias Safi, Paola Giusti-Rodriguez, Graham D. Johnson, Annie W. Shieh, et al. 2018. “Evaluation of Chromatin Accessibility in Prefrontal Cortex of Individuals with Schizophrenia.” *Nature Communications* 9 (1): 3121.
Cagan, Alexander, Christoph Theunert, Hafid Laayouni, Gabriel Santpere, Marc Pybus, Ferran Casals, Kay Prüfer, et al. 2016. “Natural Selection in the Great Apes.” *Molecular Biology and Evolution* 33 (12): 3268–83.
Capozzi, Oronzo, Lucia Carbone, Roscoe R. Stanyon, Annamaria Marra, Fengtang Yang, Christopher W. Whelan, Pieter J. de Jong, Mariano Rocchi, and Nicoletta Archidiacono. 2012. “A Comprehensive Molecular Cytogenetic Analysis of Chromosome Rearrangements in Gibbons.” *Genome Research* 22 (12): 2520–28.
Carbone, Lucia, Mario Ventura, Sergio Tempesta, Mariano Rocchi, and Nicoletta Archidiacono. 2002. “Evolutionary History of Chromosome 10 in Primates.” *Chromosoma* 111 (4): 267–72.
Cardone, Maria Francesca, Alicia Alonso, Michele Pazienza, Mario Ventura, Gabriella Montemurro, Lucia Carbone, Pieter J. de Jong, et al. 2006. “Independent Centromere Formation in a Capricious, Gene-Free Domain of Chromosome 13q21 in Old World Monkeys and Pigs.” *Genome Biology* 7 (10): R91.
Cardone, Maria Francesca, Zhaoshi Jiang, Pietro D’Addabbo, Nicoletta Archidiacono, Mariano Rocchi, Evan E. Eichler, and Mario Ventura. 2008. “Hominoid Chromosomal Rearrangements on 17q Map to Complex Regions of Segmental Duplication.” *Genome*

Biology 9 (2): R28.
Cardone, Maria Francesca, Mariana Lomiento, Maria Grazia Teti, Doriana Misceo, Roberta Roberto, Oronzo Capozzi, Pietro D'Addabbo, Mario Ventura, Mariano Rocchi, and Nicoletta Archidiacono. 2007. "Evolutionary History of Chromosome 11 Featuring Four Distinct Centromere Repositioning Events in Catarrhini." *Genomics* 90 (1): 35–43.
Carroll, Sean B. 2000. "Endless Forms: The Evolution of Gene Regulation and Morphological Diversity." *Cell* 101 (6): 577–80.
Carvalho, C. M., and J. R. Lupski. 2016. "Mechanisms Underlying Structural Variant Formation in Genomic Disorders." *Nature Reviews. Genetics* 17 (4): 224–38.
Catacchio, Claudia Rita, Flavia Angela Maria Maggiolini, Pietro D'Addabbo, Miriana Bitonto, Oronzo Capozzi, Martina Lepore Signorile, Mattia Miroballo, et al. 2018. "Inversion Variants in Human and Primate Genomes." *Genome Research* 28 (6): 910–20.
Chaisson, Mark J. P., Ashley D. Sanders, Xuefang Zhao, Ankit Malhotra, David Porubsky, Tobias Rausch, Eugene J. Gardner, et al. 2019. "Multi-Platform Discovery of Haplotype-Resolved Structural Variation in Human Genomes." *Nature Communications* 10 (1): 1784.
Chander, Varuna, Richard A. Gibbs, and Fritz J. Sedlazeck. 2019. "Evaluation of Computational Genotyping of Structural Variation for Clinical Diagnoses." *GigaScience* 8 (9). https://doi.org/10.1093/gigascience/giz110.
Charrier, Cécile, Kaumudi Joshi, Jaeda Coutinho-Budd, Ji-Eun Kim, Nelle Lambert, Jacqueline de Marchena, Wei-Lin Jin, et al. 2012. "Inhibition of SRGAP2 Function by Its Human-Specific Paralogs Induces Neoteny during Spine Maturation." *Cell* 149 (4): 923–35.
Cheng, Ze, Mario Ventura, Xinwei She, Philipp Khaitovich, Tina Graves, Kazutoyo Osoegawa, Deanna Church, et al. 2005. "A Genome-Wide Comparison of Recent Chimpanzee and Human Segmental Duplications." *Nature* 437 (7055): 88–93.
Chiang, Colby, Ryan M. Layer, Gregory G. Faust, Michael R. Lindberg, David B. Rose, Erik P. Garrison, Gabor T. Marth, Aaron R. Quinlan, and Ira M. Hall. 2015. "SpeedSeq: Ultra-Fast Personal Genome Analysis and Interpretation." *Nature Methods* 12 (10): 966–68.
Chiang, Colby, Alexandra J. Scott, Joe R. Davis, Emily K. Tsang, Xin Li, Yungil Kim, Tarik Hadzic, et al. 2017. "The Impact of Structural Variation on Human Gene Expression." *Nature Genetics* 49 (5): 692–99.
Chikina, Maria D., and Olga G. Troyanskaya. 2012. "An Effective Statistical Evaluation of ChIPseq Dataset Similarity." *Bioinformatics* 28 (5): 607–13.
Chimpanzee Sequencing and Analysis Consortium. 2005. "Initial Sequence of the Chimpanzee Genome and Comparison with the Human Genome." *Nature* 437 (7055): 69–87.
Chung, Dongjun, Pei Fen Kuan, Bo Li, Rajendran Sanalkumar, Kun Liang, Emery H. Bresnick, Colin Dewey, and Sündüz Keleş. 2011. "Discovering Transcription Factor Binding Sites in Highly Repetitive Regions of Genomes with Multi-Read Analysis of ChIP-Seq Data." *PLoS Computational Biology* 7 (7): e1002111.
Coe, Bradley P., Kali Witherspoon, Jill A. Rosenfeld, Bregje W. M. van Bon, Anneke T. Vulto-van Silfhout, Paolo Bosco, Kathryn L. Friend, et al. 2014. "Refining Analyses of Copy Number Variation Identifies Specific Genes Associated with Developmental Delay." *Nature Genetics* 46 (10): 1063–71.
Cooper, Gregory M., Bradley P. Coe, Santhosh Girirajan, Jill A. Rosenfeld, Tiffany H. Vu, Carl Baker, Charles Williams, et al. 2011. "A Copy Number Variation Morbidity Map of Developmental Delay." *Nature Genetics* 43 (9): 838–46.
Corbett-Detig, Russell B., and Daniel L. Hartl. 2012. "Population Genomics of Inversion Polymorphisms in Drosophila Melanogaster." *PLoS Genetics* 8 (12): e1003056.
Cuscó, Ivon, Roser Corominas, Mònica Bayés, Raquel Flores, Núria Rivera-Brugués, Victoria Campuzano, and Luis A. Pérez-Jurado. 2008. "Copy Number Variation at the 7q11.23 Segmental Duplications Is a Susceptibility Factor for the Williams-Beuren Syndrome

Deletion." *Genome Research* 18 (5): 683–94.
Darrow, Emily M., Miriam H. Huntley, Olga Dudchenko, Elena K. Stamenova, Neva C. Durand, Zhuo Sun, Su-Chen Huang, et al. 2016. "Deletion of DXZ4 on the Human Inactive X Chromosome Alters Higher-Order Genome Architecture." *Proceedings of the National Academy of Sciences of the United States of America* 113 (31): E4504–12.
Davis, Carrie A., Benjamin C. Hitz, Cricket A. Sloan, Esther T. Chan, Jean M. Davidson, Idan Gabdank, Jason A. Hilton, et al. 2018. "The Encyclopedia of DNA Elements (ENCODE): Data Portal Update." *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkx1081.
Degner, Jacob F., Athma A. Pai, Roger Pique-Regi, Jean-Baptiste Veyrieras, Daniel J. Gaffney, Joseph K. Pickrell, Sherryl De Leon, et al. 2012. "DNase I Sensitivity QTLs Are a Major Determinant of Human Expression Variation." *Nature* 482 (7385): 390–94.
Dennis, Megan Y., and Evan E. Eichler. 2016. "Human Adaptation and Evolution by Segmental Duplication." *Current Opinion in Genetics & Development* 41 (December): 44–52.
Dennis, Megan Y., Lana Harshman, Bradley J. Nelson, Osnat Penn, Stuart Cantsilieris, John Huddleston, Francesca Antonacci, et al. 2017. "The Evolution and Population Diversity of Human-Specific Segmental Duplications." *Nature Ecology & Evolution* 1 (3): 69.
Dennis, Megan Y., Xander Nuttle, Peter H. Sudmant, Francesca Antonacci, Tina A. Graves, Mikhail Nefedov, Jill A. Rosenfeld, et al. 2012. "Evolution of Human-Specific Neural SRGAP2 Genes by Incomplete Segmental Duplication." *Cell* 149 (4): 912–22.
Deshpande, Aditya S., Netha Ulahannan, Matthew Pendleton, Xiaoguang Dai, Lynn Ly, Julie M. Behr, Stefan Schwenk, et al. 2022. "Identifying Synergistic High-Order 3D Chromatin Conformations from Genome-Scale Nanopore Concatemer Sequencing." *Nature Biotechnology* 40 (10): 1488–99.
Despang, Alexandra, Robert Schöpflin, Martin Franke, Salaheddine Ali, Ivana Jerković, Christina Paliou, Wing-Lee Chan, et al. 2019. "Functional Dissection of the Sox9-Kcnj2 Locus Identifies Nonessential and Instructive Roles of TAD Architecture." *Nature Genetics* 51 (8): 1263–71.
Dixon, Jesse R., Siddarth Selvaraj, Feng Yue, Audrey Kim, Yan Li, Yin Shen, Ming Hu, Jun S. Liu, and Bing Ren. 2012. "Topological Domains in Mammalian Genomes Identified by Analysis of Chromatin Interactions." *Nature* 485 (7398): 376–80.
Dougherty, Max L., Jason G. Underwood, Bradley J. Nelson, Elizabeth Tseng, Katherine M. Munson, Osnat Penn, Tomasz J. Nowakowski, Alex A. Pollen, and Evan E. Eichler. 2018. "Transcriptional Fates of Human-Specific Segmental Duplications in Brain." *Genome Research* 28 (10): 1566–76.
Dumont, Beth L. 2015. "Interlocus Gene Conversion Explains at Least 2.7% of Single Nucleotide Variants in Human Segmental Duplications." *BMC Genomics* 16 (June): 456.
Durand, Neva C., Muhammad S. Shamim, Ido Machol, Suhas S. P. Rao, Miriam H. Huntley, Eric S. Lander, and Erez Lieberman Aiden. 2016. "Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments." *Cell Systems* 3 (1): 95–98.
Eichler, Evan E. 2019. "Genetic Variation, Comparative Genomics, and the Diagnosis of Disease." *The New England Journal of Medicine* 381 (1): 64–74.
Eres, Ittai E., and Yoav Gilad. 2021. "A TAD Skeptic: Is 3D Genome Topology Conserved?" *Trends in Genetics: TIG* 37 (3): 216–23.
Eres, Ittai E., Kaixuan Luo, Chiaowen Joyce Hsiao, Lauren E. Blake, and Yoav Gilad. 2019. "Reorganization of 3D Genome Structure May Contribute to Gene Regulatory Evolution in Primates." *PLoS Genetics* 15 (7): e1008278.
Falconer, Ester, Mark Hills, Ulrike Naumann, Steven S. S. Poon, Elizabeth A. Chavez, Ashley D. Sanders, Yongjun Zhao, Martin Hirst, and Peter M. Lansdorp. 2012. "DNA Template Strand Sequencing of Single-Cells Maps Genomic Rearrangements at High Resolution." *Nature Methods* 9 (11): 1107–12.
Fay, J. C., and P. J. Wittkopp. 2008. "Evaluating the Role of Natural Selection in the Evolution of

Gene Regulation." *Heredity* 100 (2): 191–99.
Feuk, Lars, Jeffrey R. MacDonald, Terence Tang, Andrew R. Carson, Martin Li, Girish Rao, Razi Khaja, and Stephen W. Scherer. 2005. "Discovery of Human Inversion Polymorphisms by Comparative Analysis of Human and Chimpanzee DNA Sequence Assemblies." *PLoS Genetics* 1 (4): e56.
Fiddes, Ian T., Gerrald A. Lodewijk, Meghan Mooring, Colleen M. Bosworth, Adam D. Ewing, Gary L. Mantalas, Adam M. Novak, et al. 2018. "Human-Specific NOTCH2NL Genes Affect Notch Signaling and Cortical Neurogenesis." *Cell* 173 (6): 1356–69.e22.
Florio, Marta, Mareike Albert, Elena Taverna, Takashi Namba, Holger Brandl, Eric Lewitus, Christiane Haffner, et al. 2015. "Human-Specific Gene ARHGAP11B Promotes Basal Progenitor Amplification and Neocortex Expansion." *Science* 347 (6229): 1465–70.
Florio, Marta, Michael Heide, Anneline Pinson, Holger Brandl, Mareike Albert, Sylke Winkler, Pauline Wimberger, Wieland B. Huttner, and Michael Hiller. 2018. "Evolution and Cell-Type Specificity of Human-Specific Genes Preferentially Expressed in Progenitors of Fetal Neocortex." *eLife* 7 (March). https://doi.org/10.7554/eLife.32332.
Forcato, Mattia, Chiara Nicoletti, Koustav Pal, Carmen Maria Livi, Francesco Ferrari, and Silvio Bicciato. 2017. "Comparison of Computational Methods for Hi-C Data Analysis." *Nature Methods* 14 (7): 679–85.
Force, A., M. Lynch, F. B. Pickett, A. Amores, Y. L. Yan, and J. Postlethwait. 1999. "Preservation of Duplicate Genes by Complementary, Degenerative Mutations." *Genetics* 151 (4): 1531–45.
Franke, Martin, Adrian F. Daly, Leonor Palmeira, Amit Tirosh, Antonio Stigliano, Eszter Trifan, Fabio R. Faucz, et al. 2022. "Duplications Disrupt Chromatin Architecture and Rewire GPR101-Enhancer Communication in X-Linked Acrogigantism." *American Journal of Human Genetics* 109 (4): 553–70.
Franke, Martin, Daniel M. Ibrahim, Guillaume Andrey, Wibke Schwarzer, Verena Heinrich, Robert Schöpflin, Katerina Kraft, et al. 2016. "Formation of New Chromatin Domains Determines Pathogenicity of Genomic Duplications." *Nature* 538 (7624): 265–69.
Fraser, Hunter B. 2013. "Gene Expression Drives Local Adaptation in Humans." *Genome Research* 23 (7): 1089–96.
Fudenberg, Geoff, David R. Kelley, and Katherine S. Pollard. 2020. "Predicting 3D Genome Folding from DNA Sequence with Akita." *Nature Methods* 17 (11): 1111–17.
Fudenberg, Geoff, and Katherine S. Pollard. 2019. "Chromatin Features Constrain Structural Variation across Evolutionary Timescales." *Proceedings of the National Academy of Sciences of the United States of America* 116 (6): 2175–80.
———. n.d. "Chromatin Features Constrain Structural Variation across Evolutionary Timescales." https://doi.org/10.1101/285205.
Fulco, Charles P., Joseph Nasser, Thouis R. Jones, Glen Munson, Drew T. Bergman, Vidya Subramanian, Sharon R. Grossman, et al. 2019. "Activity-by-Contact Model of Enhancer-Promoter Regulation from Thousands of CRISPR Perturbations." *Nature Genetics* 51 (12): 1664–69.
Gallego Romero, I., B. J. Pavlovic, I. Hernando-Herraez, X. Zhou, M. C. Ward, N. E. Banovich, C. L. Kagan, et al. 2015. "A Panel of Induced Pluripotent Stem Cells from Chimpanzees: A Resource for Comparative Functional Genomics." *eLife* 4: e07103.
Gel, Bernat, and Eduard Serra. 2017. "karyoploteR: An R/Bioconductor Package to Plot Customizable Genomes Displaying Arbitrary Data." *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btx346.
Gershman, Ariel, Michael E. G. Sauria, Xavi Guitart, Mitchell R. Vollger, Paul W. Hook, Savannah J. Hoyt, Miten Jain, et al. 2022. "Epigenetic Patterns in a Complete Human Genome." *Science* 376 (6588): eabj5089.
Ghavi-Helm, Yad, Aleksander Jankowski, Sascha Meiers, Rebecca R. Viales, Jan O. Korbel,

and Eileen E. M. Furlong. 2019. “Highly Rearranged Chromosomes Reveal Uncoupling between Genome Topology and Gene Expression.” *Nature Genetics* 51 (8): 1272–82.
Giannuzzi, Giuliana, Priscillia Siswara, Maika Malig, Tomas Marques-Bonet, NISC Comparative Sequencing Program, James C. Mullikin, Mario Ventura, and Evan E. Eichler. 2013. “Evolutionary Dynamism of the Primate LRRC37 Gene Family.” *Genome Research* 23 (1): 46–59.
Giglio, Sabrina, Vladimiro Calvari, Giuliana Gregato, Giorgio Gimelli, Silvia Camanini, Roberto Giorda, Angela Ragusa, et al. 2002. “Heterozygous Submicroscopic Inversions Involving Olfactory Receptor-Gene Clusters Mediate the Recurrent t(4;8)(p16;p23) Translocation.” *American Journal of Human Genetics* 71 (2): 276–85.
Giglio, S., K. W. Broman, N. Matsumoto, V. Calvari, G. Gimelli, T. Neumann, H. Ohashi, et al. 2001. “Olfactory Receptor-Gene Clusters, Genomic-Inversion Polymorphisms, and Common Chromosome Rearrangements.” *American Journal of Human Genetics* 68 (4): 874–83.
Giner-Delgado, Carla, Sergi Villatoro, Jon Lerga-Jaso, Magdalena Gayà-Vidal, Meritxell Oliva, David Castellano, Lorena Pantano, et al. 2019. “Evolutionary and Functional Impact of Common Polymorphic Inversions in the Human Genome.” *Nature Communications* 10 (1): 4222.
Gkika, Dimitra, Loic Lemonnier, George Shapovalov, Dmitri Gordienko, Céline Poux, Michela Bernardini, Alexandre Bokhobza, et al. 2015. “TRP Channel–associated Factors Are a Novel Protein Family That Regulates TRPM8 Trafficking and Activity.” *The Journal of General Physiology*. https://doi.org/10.1085/jgp.1452oia1.
Goidts, Violaine, Justyna M. Szamalek, Pieter J. de Jong, David N. Cooper, Nadia Chuzhanova, Horst Hameister, and Hildegard Kehrer-Sawatzki. 2005. “Independent Intrachromosomal Recombination Events Underlie the Pericentric Inversions of Chimpanzee and Gorilla Chromosomes Homologous to Human Chromosome 16.” *Genome Research* 15 (9): 1232–42.
Gokcumen, Omer, Verena Tischler, Jelena Tica, Qihui Zhu, Rebecca C. Iskow, Eunjung Lee, Markus Hsi-Yang Fritz, et al. 2013. “Primate Genome Architecture Influences Structural Variation Mechanisms and Functional Consequences.” *Proceedings of the National Academy of Sciences of the United States of America* 110 (39): 15764–69.
Gordon, M. Grace, Fumitaka Inoue, Beth Martin, Max Schubach, Vikram Agarwal, Sean Whalen, Shiyun Feng, et al. 2020. “lentiMPRA and MPRAflow for High-Throughput Functional Characterization of Gene Regulatory Elements.” *Nature Protocols* 15 (8): 2387–2412.
Gu, Xun, Zhongqi Zhang, and Wei Huang. 2005. “Rapid Evolution of Expression and Regulatory Divergences after Yeast Gene Duplication.” *Proceedings of the National Academy of Sciences of the United States of America* 102 (3): 707–12.
Hach, Faraz, Fereydoun Hormozdiari, Can Alkan, Farhad Hormozdiari, Inanc Birol, Evan E. Eichler, and S. Cenk Sahinalp. 2010. “mrsFAST: A Cache-Oblivious Algorithm for Short-Read Mapping.” *Nature Methods* 7 (8): 576–77.
Hastie, Alex R., Ernest T. Lam, Andy Wing Chun Pang, Xinyue Zhang, Warren Andrews, Joyce Lee, Tiffany Y. Liang, et al. 2017. “Rapid Automated Large Structural Variation Detection in a Diploid Genome by NanoChannel Based Next-Generation Mapping.” *bioRxiv*. https://doi.org/10.1101/102764.
Heide, Michael, Christiane Haffner, Ayako Murayama, Yoko Kurotaki, Haruka Shinohara, Hideyuki Okano, Erika Sasaki, and Wieland B. Huttner. 2020. “Human-Specific Increases Size and Folding of Primate Neocortex in the Fetal Marmoset.” *Science* 369 (6503): 546–50.
Heinz, Sven, Christopher Benner, Nathanael Spann, Eric Bertolino, Yin C. Lin, Peter Laslo, Jason X. Cheng, Cornelis Murre, Harinder Singh, and Christopher K. Glass. 2010. “Simple

Combinations of Lineage-Determining Transcription Factors Prime Cis-Regulatory Elements Required for Macrophage and B Cell Identities." *Molecular Cell*. https://doi.org/10.1016/j.molcel.2010.05.004.
Henrichsen, Charlotte N., Evelyne Chaignat, and Alexandre Reymond. 2009. "Copy Number Variants, Diseases and Gene Expression." *Human Molecular Genetics* 18 (R1): R1–8.
Heyworth, Paul G., Deborah Noack, and Andrew R. Cross. 2002. "Identification of a Novel NCF-1 (p47-Phox) Pseudogene Not Containing the Signature GT Deletion: Significance for A47° Chronic Granulomatous Disease Carrier Detection." *Blood* 100 (5): 1845–51.
Hollox, E.J., J.A.L. Armour, and J.C.K. Barber. 2003. "Extensive Normal Copy Number Variation of a β-Defensin Antimicrobial-Gene Cluster." *American Journal of Human Genetics* 73 (3): 591–600.
Hollox, Edward J., Ulrike Huffmeier, Patrick L. J. M. Zeeuwen, Raquel Palla, Jesús Lascorz, Diana Rodijk-Olthuis, Peter C. M. van de Kerkhof, et al. 2008. "Psoriasis Is Associated with Increased Beta-Defensin Genomic Copy Number." *Nature Genetics* 40 (1): 23–25.
Hsieh, Pinghsun, Vy Dang, Mitchell R. Vollger, Yafei Mao, Tzu-Hsueh Huang, Philip C. Dishuck, Carl Baker, et al. 2021. "Evidence for Opposing Selective Forces Operating on Human-Specific Duplicated TCAF Genes in Neanderthals and Humans." *Nature Communications* 12 (1): 5118.
Huang, Da Wei, Brad T. Sherman, and Richard A. Lempicki. 2009a. "Systematic and Integrative Analysis of Large Gene Lists Using DAVID Bioinformatics Resources." *Nature Protocols* 4 (1): 44–57.
———. 2009b. "Bioinformatics Enrichment Tools: Paths toward the Comprehensive Functional Analysis of Large Gene Lists." *Nucleic Acids Research* 37 (1): 1–13.
Huang, Da Wei, Brad T. Sherman, Qina Tan, Jack R. Collins, W. Gregory Alvord, Jean Roayaei, Robert Stephens, Michael W. Baseler, H. Clifford Lane, and Richard A. Lempicki. 2007. "The DAVID Gene Functional Classification Tool: A Novel Biological Module-Centric Algorithm to Functionally Analyze Large Gene Lists." *Genome Biology* 8 (9): 1–16.
Hudson, R. R., M. Kreitman, and M. Aguadé. 1987. "A Test of Neutral Molecular Evolution Based on Nucleotide Data." *Genetics* 116 (1): 153–59.
Hultqvist, Malin, Peter Olofsson, Jens Holmberg, B. Thomas Bäckström, Jesper Tordsson, and Rikard Holmdahl. 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 of the United States of America* 101 (34): 12646–51.
Huynh, Linh, and Fereydoun Hormozdiari. 2019. "TAD Fusion Score: Discovery and Ranking the Contribution of Deletions to Genome Structure." *Genome Biology* 20 (1): 60.
Iafrate, A. John, Lars Feuk, Miguel N. Rivera, Marc L. Listewnik, Patricia K. Donahoe, Ying Qi, Stephen W. Scherer, and Charles Lee. 2004. "Detection of Large-Scale Variation in the Human Genome." *Nature Genetics* 36 (9): 949–51.
Indjeian, Vahan B., Garrett A. Kingman, Felicity C. Jones, Catherine A. Guenther, Jane Grimwood, Jeremy Schmutz, Richard M. Myers, and David M. Kingsley. 2016. "Evolving New Skeletal Traits by Cis-Regulatory Changes in Bone Morphogenetic Proteins." *Cell* 164 (1-2): 45–56.
Iskow, Rebecca C., Omer Gokcumen, Alexej Abyzov, Joanna Malukiewicz, Qihui Zhu, Ann T. Sukumar, Athma A. Pai, et al. 2012. "Regulatory Element Copy Number Differences Shape Primate Expression Profiles." *Proceedings of the National Academy of Sciences of the United States of America* 109 (31): 12656–61.
Itsara, Andy, Gregory M. Cooper, Carl Baker, Santhosh Girirajan, Jun Li, Devin Absher, Ronald M. Krauss, et al. 2009. "Population Analysis of Large Copy Number Variants and Hotspots of Human Genetic Disease." *American Journal of Human Genetics* 84 (2): 148–61.
Jain, Miten, Sergey Koren, Karen H. Miga, Josh Quick, Arthur C. Rand, Thomas A. Sasani, John R. Tyson, et al. 2018. "Nanopore Sequencing and Assembly of a Human Genome

with Ultra-Long Reads." *Nature Biotechnology* 36 (4): 338–45.
Jiang, Zhaoshi, Haixu Tang, Mario Ventura, Maria Francesca Cardone, Tomas Marques-Bonet, Xinwei She, Pavel A. Pevzner, and Evan E. Eichler. 2007. "Ancestral Reconstruction of Segmental Duplications Reveals Punctuated Cores of Human Genome Evolution." *Nature Genetics* 39 (11): 1361–68.
Jimenez, Sergio A., and Sonsoles Piera-Velazquez. 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." *Autoimmunity Reviews* 12 (11): 1046–51.
Jong, Simone de, Iouri Chepelev, Esther Janson, Eric Strengman, Leonard H. van den Berg, Jan H. Veldink, and Roel A. Ophoff. 2012. "Common Inversion Polymorphism at 17q21.31 Affects Expression of Multiple Genes in Tissue-Specific Manner." *BMC Genomics* 13 (September): 458.
Ju, Xiang-Chun, Qiong-Qiong Hou, Ai-Li Sheng, Kong-Yan Wu, Yang Zhou, Ying Jin, Tieqiao Wen, Zhengang Yang, Xiaoqun Wang, and Zhen-Ge Luo. 2016. "The Hominoid-Specific Gene TBC1D3 Promotes Generation of Basal Neural Progenitors and Induces Cortical Folding in Mice." *eLife* 5 (August). https://doi.org/10.7554/eLife.18197.
Kalebic, Nereo, Carlotta Gilardi, Mareike Albert, Takashi Namba, Katherine R. Long, Milos Kostic, Barbara Langen, and Wieland B. Huttner. 2018. "Human-Specific Induces Hallmarks of Neocortical Expansion in Developing Ferret Neocortex." *eLife* 7 (November). https://doi.org/10.7554/eLife.41241.
Kaminsky, Erin B., Vineith Kaul, Justin Paschall, Deanna M. Church, Brian Bunke, Dawn Kunig, Daniel Moreno-De-Luca, et al. 2011. "An Evidence-Based Approach to Establish the Functional and Clinical Significance of Copy Number Variants in Intellectual and Developmental Disabilities." *Genetics in Medicine: Official Journal of the American College of Medical Genetics* 13 (9): 777–84.
Karczewski, Konrad J., Laurent C. Francioli, Grace Tiao, Beryl B. Cummings, Jessica Alföldi, Qingbo Wang, Ryan L. Collins, et al. 2019. "Variation across 141,456 Human Exomes and Genomes Reveals the Spectrum of Loss-of-Function Intolerance across Human Protein-Coding Genes." *bioRxiv*. https://doi.org/10.1101/531210.
Kashima, Tsuyoshi, and James L. Manley. 2003. "A Negative Element in SMN2 Exon 7 Inhibits Splicing in Spinal Muscular Atrophy." *Nature Genetics* 34 (4): 460–63.
Kehrer-Sawatzki, Hildegard, and David N. Cooper. 2007. "Understanding the Recent Evolution of the Human Genome: Insights from Human-Chimpanzee Genome Comparisons." *Human Mutation*. https://doi.org/10.1002/humu.20420.
———. 2008. "Molecular Mechanisms of Chromosomal Rearrangement during Primate Evolution." *Chromosome Research: An International Journal on the Molecular, Supramolecular and Evolutionary Aspects of Chromosome Biology* 16 (1): 41–56.
Kehrer-Sawatzki, Hildegard, Catharina Sandig, Nadia Chuzhanova, Violaine Goidts, Justyna M. Szamalek, Simone Tänzer, Stefan Müller, Matthias Platzer, David N. Cooper, and Horst Hameister. 2005. "Breakpoint Analysis of the Pericentric Inversion Distinguishing Human Chromosome 4 from the Homologous Chromosome in the Chimpanzee (Pan Troglodytes)." *Human Mutation* 25 (1): 45–55.
Kehrer-Sawatzki, Hildegard, Justyna M. Szamalek, Simone Tänzer, Matthias Platzer, and Horst Hameister. 2005. "Molecular Characterization of the Pericentric Inversion of Chimpanzee Chromosome 11 Homologous to Human Chromosome 9." *Genomics* 85 (5): 542–50.
Kehrer-Sawatzki, H., C. A. Sandig, V. Goidts, and H. Hameister. 2005. "Breakpoint Analysis of the Pericentric Inversion between Chimpanzee Chromosome 10 and the Homologous Chromosome 12 in Humans." *Cytogenetic and Genome Research* 108 (1-3): 91–97.
Kent, W. James. 2002. "BLAT--the BLAST-like Alignment Tool." *Genome Research* 12 (4): 656–64.

Khan, Zia, Michael J. Ford, Darren A. Cusanovich, Amy Mitrano, Jonathan K. Pritchard, and Yoav Gilad. 2013. “Primate Transcript and Protein Expression Levels Evolve under Compensatory Selection Pressures.” *Science* 342 (6162): 1100–1104.
Kidd, Jeffrey M., Gregory M. Cooper, William F. Donahue, Hillary S. Hayden, Nick Sampas, Tina Graves, Nancy Hansen, et al. 2008. “Mapping and Sequencing of Structural Variation from Eight Human Genomes.” *Nature* 453 (7191): 56–64.
King, Mary-Claire, and A. C. Wilson. 1975. “Evolution at Two Levels in Humans and Chimpanzees.” *Science*. https://doi.org/10.1126/science.1090005.
Kirkpatrick, Mark. 2010. “How and Why Chromosome Inversions Evolve.” *PLoS Biology* 8 (9). https://doi.org/10.1371/journal.pbio.1000501.
Kirkpatrick, Mark, and Nick Barton. 2006. “Chromosome Inversions, Local Adaptation and Speciation.” *Genetics* 173 (1): 419–34.
Klein, Jason C., Vikram Agarwal, Fumitaka Inoue, Aidan Keith, Beth Martin, Martin Kircher, Nadav Ahituv, and Jay Shendure. 2020. “A Systematic Evaluation of the Design and Context Dependencies of Massively Parallel Reporter Assays.” *Nature Methods* 17 (11): 1083–91.
Kleinjan, Dirk-Jan, and Pedro Coutinho. 2009. “Cis-Ruption Mechanisms: Disruption of Cis-Regulatory Control as a Cause of Human Genetic Disease.” *Briefings in Functional Genomics & Proteomics* 8 (4): 317–32.
Kondrashov, Fyodor A., Igor B. Rogozin, Yuri I. Wolf, and Eugene V. Koonin. 2002. “Selection in the Evolution of Gene Duplications.” *Genome Biology* 3 (2): 1–9.
Kosuke M. Teshima, Hideki Innan. 2008. “Neofunctionalization of Duplicated Genes Under the Pressure of Gene Conversion.” *Genetics* 178 (3): 1385.
Kronenberg, Zev N., Ian T. Fiddes, David Gordon, Shwetha Murali, Stuart Cantsilieris, Olivia S. Meyerson, Jason G. Underwood, et al. 2018. “High-Resolution Comparative Analysis of Great Ape Genomes.” *Science* 360 (6393). https://doi.org/10.1126/science.aar6343.
Kuderna, Lukas F. K., Chad Tomlinson, Ladeana W. Hillier, Annabel Tran, Ian T. Fiddes, Joel Armstrong, Hafid Laayouni, et al. 2017. “A 3-Way Hybrid Approach to Generate a New High-Quality Chimpanzee Reference Genome (Pan_tro_3.0).” *GigaScience* 6 (11): 1–6.
Kulakovskiy, Ivan V., Ilya E. Vorontsov, Ivan S. Yevshin, Ruslan N. Sharipov, Alla D. Fedorova, Eugene I. Rumynskiy, Yulia A. Medvedeva, 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 Research* 46 (D1): D252–59.
Lander, E. S., L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, K. Devon, et al. 2001. “Initial Sequencing and Analysis of the Human Genome.” *Nature* 409 (6822): 860–921.
Landrum, Melissa J., Jennifer M. Lee, Mark Benson, Garth Brown, Chen Chao, Shanmuga Chitipiralla, Baoshan Gu, et al. 2016. “ClinVar: Public Archive of Interpretations of Clinically Relevant Variants.” *Nucleic Acids Research* 44 (D1): D862–68.
Landt, Stephen G., Georgi K. Marinov, Anshul Kundaje, Pouya Kheradpour, Florencia Pauli, Serafim Batzoglou, Bradley E. Bernstein, et al. 2012. “ChIP-Seq Guidelines and Practices of the ENCODE and modENCODE Consortia.” *Genome Research* 22 (9): 1813–31.
Langergraber, Kevin E., Kay Prüfer, Carolyn Rowney, Christophe Boesch, Catherine Crockford, Katie Fawcett, Eiji Inoue, et al. 2012. “Generation Times in Wild Chimpanzees and Gorillas Suggest Earlier Divergence Times in Great Ape and Human Evolution.” *Proceedings of the National Academy of Sciences of the United States of America* 109 (39): 15716–21.
Langmead, Ben, Cole Trapnell, Mihai Pop, and Steven L. Salzberg. 2009. “Ultrafast and Memory-Efficient Alignment of Short DNA Sequences to the Human Genome.” *Genome Biology* 10 (3): R25.
Lan, Xun, and Jonathan K. Pritchard. 2016. “Coregulation of Tandem Duplicate Genes Slows Evolution of Subfunctionalization in Mammals.” *Science* 352 (6288): 1009–13.

Lappalainen, Tuuli, Michael Sammeth, Marc R. Friedländer, Peter A. C. 't Hoen, Jean Monlong, Manuel A. Rivas, Mar Gonzàlez-Porta, et al. 2013. "Transcriptome and Genome Sequencing Uncovers Functional Variation in Humans." *Nature* 501 (7468): 506–11.
Law, Charity W., Yunshun Chen, Wei Shi, and Gordon K. Smyth. 2014. "Voom: Precision Weights Unlock Linear Model Analysis Tools for RNA-Seq Read Counts." *Genome Biology* 15 (2): 1–17.
Lazar, Nathan H., Kimberly A. Nevonen, Brendan O'Connell, Christine McCann, Rachel J. O'Neill, Richard E. Green, Thomas J. Meyer, Mariam Okhovat, and Lucia Carbone. 2018. "Epigenetic Maintenance of Topological Domains in the Highly Rearranged Gibbon Genome." *Genome Research* 28 (7): 983–97.
Lee, Isac, Roham Razaghi, Timothy Gilpatrick, Michael Molnar, Ariel Gershman, Norah Sadowski, Fritz J. Sedlazeck, Kasper D. Hansen, Jared T. Simpson, and Winston Timp. 2020. "Simultaneous Profiling of Chromatin Accessibility and Methylation on Human Cell Lines with Nanopore Sequencing." *Nature Methods* 17 (12): 1191–99.
Lek, Monkol, Konrad J. Karczewski, Eric V. Minikel, Kaitlin E. Samocha, Eric Banks, Timothy Fennell, Anne H. O'Donnell-Luria, et al. 2016. "Analysis of Protein-Coding Genetic Variation in 60,706 Humans." *Nature* 536 (7616): 285–91.
Lichter, P., C. J. Tang, K. Call, G. Hermanson, G. A. Evans, D. Housman, and D. C. Ward. 1990. "High-Resolution Mapping of Human Chromosome 11 by in Situ Hybridization with Cosmid Clones." *Science* 247 (4938): 64–69.
Li, Heng. 2013. "Aligning Sequence Reads, Clone Sequences and Assembly Contigs with BWA-MEM." *arXiv [q-bio.GN]*. arXiv. http://arxiv.org/abs/1303.3997.
Li, Heng, and Richard Durbin. 2009. "Fast and Accurate Short Read Alignment with Burrows-Wheeler Transform." *Bioinformatics* 25 (14): 1754–60.
Li, Ju-Pi, Chia-Yu Yang, Huai-Chia Chuang, Joung-Liang Lan, Der-Yuan Chen, Yi-Ming Chen, Xiaohong Wang, Alice J. Chen, John W. Belmont, and Tse-Hua Tan. 2014. "The Phosphatase JKAP/DUSP22 Inhibits T-Cell Receptor Signalling and Autoimmunity by Inactivating Lck." *Nature Communications* 5 (April): 3618.
Liu, Tao. 2014. "Use Model-Based Analysis of ChIP-Seq (MACS) to Analyze Short Reads Generated by Sequencing Protein-DNA Interactions in Embryonic Stem Cells." *Methods in Molecular Biology* 1150: 81–95.
Locke, Devin P., Nicoletta Archidiacono, Doriana Misceo, Maria Francesca Cardone, Stephane Deschamps, Bruce Roe, Mariano Rocchi, and Evan E. Eichler. 2003. "Refinement of a Chimpanzee Pericentric Inversion Breakpoint to a Segmental Duplication Cluster." *Genome Biology* 4 (8): R50.
Locke, D. P., R. Segraves, L. Carbone, N. Archidiacono, D. G. Albertson, D. Pinkel, and E. E. Eichler. 2003. "Large-Scale Variation among Human and Great Ape Genomes Determined by Array Comparative Genomic Hybridization." *Genome Research* 13 (3): 347–57.
Lupiáñez, Darío G., Katerina Kraft, Verena Heinrich, Peter Krawitz, Francesco Brancati, Eva Klopocki, Denise Horn, et al. 2015. "Disruptions of Topological Chromatin Domains Cause Pathogenic Rewiring of Gene-Enhancer Interactions." *Cell* 161 (5): 1012–25.
Lynch, M., and J. S. Conery. 2000. "The Evolutionary Fate and Consequences of Duplicate Genes." *Science*.
Maggiolini, Flavia A. M., Stuart Cantsilieris, Pietro D'Addabbo, Michele Manganelli, Bradley P. Coe, Beth L. Dumont, Ashley D. Sanders, et al. 2019. "Genomic Inversions and GOLGA Core Duplicons Underlie Disease Instability at the 15q25 Locus." *PLoS Genetics* 15 (3): e1008075.
Mahmoud, Medhat, Nastassia Gobet, Diana Ivette Cruz-Dávalos, Ninon Mounier, Christophe Dessimoz, and Fritz J. Sedlazeck. 2019. "Structural Variant Calling: The Long and the Short of It." *Genome Biology*. https://doi.org/10.1186/s13059-019-1828-7.
Mallick, Swapan, Heng Li, Mark Lipson, Iain Mathieson, Melissa Gymrek, Fernando Racimo,

Mengyao Zhao, et al. 2016. “The Simons Genome Diversity Project: 300 Genomes from 142 Diverse Populations.” *Nature* 538 (7624): 201–6.
Manuel, Marc de, Martin Kuhlwilm, Peter Frandsen, Vitor C. Sousa, Tariq Desai, Javier Prado-Martinez, Jessica Hernandez-Rodriguez, et al. 2016. “Chimpanzee Genomic Diversity Reveals Ancient Admixture with Bonobos.” *Science* 354 (6311): 477–81.
Marques-Bonet, T., and E. E. Eichler. 2009. “The Evolution of Human Segmental Duplications and the Core Duplicon Hypothesis.” *Cold Spring Harbor Symposia on Quantitative Biology* 74 (August): 355–62.
Marques-Bonet, Tomas, Santhosh Girirajan, and Evan E. Eichler. 2009. “The Origins and Impact of Primate Segmental Duplications.” *Trends in Genetics: TIG* 25 (10): 443–54.
Marques-Bonet, T., J. M. Kidd, M. Ventura, T. A. Graves, Z. Cheng, L. W. Hillier, Z. Jiang, et al. 2009. “A Burst of Segmental Duplications in the Genome of the African Great Ape Ancestor.” *Nature* 457 (7231): 877–81.
McLaren, William, Laurent Gil, Sarah E. Hunt, Harpreet Singh Riat, Graham R. S. Ritchie, Anja Thormann, Paul Flicek, and Fiona Cunningham. 2016. “The Ensembl Variant Effect Predictor.” *Genome Biology* 17 (1): 122.
McLean, Cory Y., Philip L. Reno, Alex A. Pollen, Abraham I. Bassan, Terence D. Capellini, Catherine Guenther, Vahan B. Indjeian, et al. 2011. “Human-Specific Loss of Regulatory DNA and the Evolution of Human-Specific Traits.” *Nature* 471 (7337): 216–19.
McVicker, Graham, Bryce van de Geijn, Jacob F. Degner, Carolyn E. Cain, Nicholas E. Banovich, Anil Raj, Noah Lewellen, Marsha Myrthil, Yoav Gilad, and Jonathan K. Pritchard. 2013. “Identification of Genetic Variants That Affect Histone Modifications in Human Cells.” *Science* 342 (6159): 747–49.
Mefford, Heather C., Severine Clauin, Andrew J. Sharp, Rikke S. Moller, Reinhard Ullmann, Raj Kapur, Dan Pinkel, et al. 2007. “Recurrent Reciprocal Genomic Rearrangements of 17q12 Are Associated with Renal Disease, Diabetes, and Epilepsy.” *American Journal of Human Genetics* 81 (5): 1057–69.
Mérot, Claire, Violaine Llaurens, Eric Normandeau, Louis Bernatchez, and Maren Wellenreuther. 2020. “Balancing Selection via Life-History Trade-Offs Maintains an Inversion Polymorphism in a Seaweed Fly.” *Nature Communications* 11 (1): 670.
Miga, Karen H. 2019. “Centromeric Satellite DNAs: Hidden Sequence Variation in the Human Population.” *Genes* 10 (5): 352.
Miller, David T., Margaret P. Adam, Swaroop Aradhya, Leslie G. Biesecker, Arthur R. Brothman, Nigel P. Carter, Deanna M. Church, et al. 2010. “Consensus Statement: Chromosomal Microarray Is a First-Tier Clinical Diagnostic Test for Individuals with Developmental Disabilities or Congenital Anomalies.” *American Journal of Human Genetics* 86 (5): 749–64.
Mills, Ryan E., Klaudia Walter, Chip Stewart, Robert E. Handsaker, Ken Chen, Can Alkan, Alexej Abyzov, et al. 2011. “Mapping Copy Number Variation by Population-Scale Genome Sequencing.” *Nature* 470 (7332): 59–65.
Monajemi, Houshang, Ruud D. Fontijn, Hans Pannekoek, and Anton J. G. Horrevoets. 2002. “The Apolipoprotein L Gene Cluster Has Emerged Recently in Evolution and Is Expressed in Human Vascular Tissue.” *Genomics* 79 (4): 539–46.
Namba, Takashi, Judit Dóczi, Anneline Pinson, Lei Xing, Nereo Kalebic, Michaela Wilsch-Bräuninger, Katherine R. Long, et al. 2020. “Human-Specific ARHGAP11B Acts in Mitochondria to Expand Neocortical Progenitors by Glutaminolysis.” *Neuron* 105 (5): 867–81.e9.
Naqvi, Sahin, Seungsoo Kim, Hanne Hoskens, Harry S. Matthews, Richard A. Spritz, Ophir D. Klein, Benedikt Hallgrímsson, et al. 2022. “Precise Modulation of Transcription Factor Levels Reveals Drivers of Dosage Sensitivity.” *bioRxiv*. https://doi.org/10.1101/2022.06.13.495964.
Newman, Tera L., Eray Tuzun, V. Anne Morrison, Karen E. Hayden, Mario Ventura, Sean D.

McGrath, Mariano Rocchi, and Evan E. Eichler. 2005. “A Genome-Wide Survey of Structural Variation between Human and Chimpanzee.” *Genome Research* 15 (10): 1344–56.
Nguyen, L. S., L. Jolly, C. Shoubridge, W. K. Chan, L. Huang, F. Laumonnier, M. Raynaud, et al. 2012. “Transcriptome Profiling of UPF3B/NMD-Deficient Lymphoblastoid Cells from Patients with Various Forms of Intellectual Disability.” *Molecular Psychiatry* 17 (11): 1103–15.
Nickerson, E., and D. L. Nelson. 1998. “Molecular Definition of Pericentric Inversion Breakpoints Occurring during the Evolution of Humans and Chimpanzees.” *Genomics* 50 (3): 368–72.
Northcott, Paul A., Catherine Lee, Thomas Zichner, Adrian M. Stütz, Serap Erkek, Daisuke Kawauchi, David J. H. Shih, et al. 2014. “Enhancer Hijacking Activates GFI1 Family Oncogenes in Medulloblastoma.” *Nature* 511 (7510): 428–34.
Nozawa, Masafumi, Yoshihiro Kawahara, and Masatoshi Nei. 2007. “Genomic Drift and Copy Number Variation of Sensory Receptor Genes in Humans.” *Proceedings of the National Academy of Sciences of the United States of America* 104 (51): 20421–26.
Nurk, Sergey, Sergey Koren, Arang Rhie, Mikko Rautiainen, Andrey V. Bzikadze, Alla Mikheenko, Mitchell R. Vollger, et al. 2022. “The Complete Sequence of a Human Genome.” *Science* 376 (6588): 44–53.
Nuttle, Xander, Giuliana Giannuzzi, Michael H. Duyzend, Joshua G. Schraiber, Iñigo Narvaiza, Peter H. Sudmant, Osnat Penn, et al. 2016. “Emergence of a Homo Sapiens-Specific Gene Family and Chromosome 16p11.2 CNV Susceptibility.” *Nature* 536 (7615): 205–9.
O’Bleness, M., V. B. Searles, A. Varki, P. Gagneux, and J. M. Sikela. 2012. “Evolution of Genetic and Genomic Features Unique to the Human Lineage.” *Nature Reviews. Genetics* 13 (12): 853–66.
O’Geen, Henriette, Sofie L. Bates, Sakereh S. Carter, Karly A. Nisson, Julian Halmai, Kyle D. Fink, Suhn K. Rhie, Peggy J. Farnham, and David J. Segal. 2019. “Ezh2-dCas9 and KRAB-dCas9 Enable Engineering of Epigenetic Memory in a Context-Dependent Manner.” *Epigenetics & Chromatin* 12 (1): 26.
Ohno, Susumu. 1970. *Evolution by Gene Duplication*. Springer Science & Business Media.
Osborne, L. R., M. Li, B. Pober, D. Chitayat, J. Bodurtha, A. Mandel, T. Costa, et al. 2001.“A 1.5 Million-Base Pair Inversion Polymorphism in Families with Williams-Beuren Syndrome.” *Nature Genetics* 29 (3): 321–25.
Pääbo, Svante. 2014. “The Human Condition-a Molecular Approach.” *Cell* 157 (1): 216–26.
Patro, Rob, Geet Duggal, Michael I. Love, Rafael A. Irizarry, and Carl Kingsford. 2017. “Salmon Provides Fast and Bias-Aware Quantification of Transcript Expression.” *Nature Methods* 14 (4): 417–19.
Patterson, Nick, Alkes L. Price, and David Reich. 2006. “Population Structure and Eigenanalysis.” *PLoS Genetics* 2 (12): e190.
Patterson, Nick, Daniel J. Richter, Sante Gnerre, Eric S. Lander, and David Reich. 2006. “Genetic Evidence for Complex Speciation of Humans and Chimpanzees.” *Nature* 441 (7097): 1103–8.
Pavlovic, Bryan J., Lauren E. Blake, Julien Roux, Claudia Chavarria, and Yoav Gilad. 2018. “A Comparative Assessment of Human and Chimpanzee iPSC-Derived Cardiomyocytes with Primary Heart Tissues.” *Scientific Reports* 8 (1): 15312.
Perry, G. H., N. J. Dominy, K. G. Claw, A. S. Lee, H. Fiegler, R. Redon, J. Werner, et al. 2007. “Diet and the Evolution of Human Amylase Gene Copy Number Variation.” *Nature Genetics* 39 (10): 1256–60.
Pezzini, Francesco, Laura Bettinetti, Francesca Di Leva, Marzia Bianchi, Elisa Zoratti, Rosalba Carrozzo, Filippo M. Santorelli, Massimo Delledonne, Maciej Lalowski, and Alessandro Simonati. 2017. “Transcriptomic Profiling Discloses Molecular and Cellular Events Related to Neuronal Differentiation in SH-SY5Y Neuroblastoma Cells.” *Cellular and Molecular*

Neurobiology 37 (4): 665–82.
Pickrell, Joseph K., John C. Marioni, Athma A. Pai, Jacob F. Degner, Barbara E. Engelhardt, Everlyne Nkadori, Jean-Baptiste Veyrieras, Matthew Stephens, Yoav Gilad, and Jonathan K. Pritchard. 2010. “Understanding Mechanisms Underlying Human Gene Expression Variation with RNA Sequencing.” *Nature*. https://doi.org/10.1038/nature08872.
Pollen, Alex A., Aparna Bhaduri, Madeline G. Andrews, Tomasz J. Nowakowski, Olivia S. Meyerson, Mohammed A. Mostajo-Radji, Elizabeth Di Lullo, et al. 2019. “Establishing Cerebral Organoids as Models of Human-Specific Brain Evolution.” *Cell* 176 (4): 743–56.e17.
Porubsky, David, Ashley D. Sanders, Wolfram Höps, Pinghsun Hsieh, Arvis Sulovari, Ruiyang Li, Ludovica Mercuri, et al. 2020. “Recurrent Inversion Toggling and Great Ape Genome Evolution.” *Nature Genetics* 52 (8): 849–58.
Porubsky, David, Ashley D. Sanders, Aaron Taudt, Maria Colomé-Tatché, Peter M. Lansdorp, and Victor Guryev. 2020. “breakpointR: An R/Bioconductor Package to Localize Strand State Changes in Strand-Seq Data.” *Bioinformatics* 36 (4): 1260–61.
Prado-Martinez, Javier, Peter H. Sudmant, Jeffrey M. Kidd, Heng Li, Joanna L. Kelley, Belen Lorente-Galdos, Krishna R. Veeramah, et al. 2013. “Great Ape Genetic Diversity and Population History.” *Nature* 499 (7459): 471–75.
Prescott, S. L., R. Srinivasan, M. C. Marchetto, I. Grishina, I. Narvaiza, L. Selleri, F. H. Gage, T. Swigut, and J. Wysocka. 2015. “Enhancer Divergence and Cis-Regulatory Evolution in the Human and Chimp Neural Crest.” *Cell* 163 (1): 68–83.
Prince, Victoria E., and F. Bryan Pickett. 2002. “Splitting Pairs: The Diverging Fates of Duplicated Genes.” *Nature Reviews. Genetics* 3 (11): 827–37.
Prud’homme, Benjamin, Nicolas Gompel, and Sean B. Carroll. 2007. “Emerging Principles of Regulatory Evolution.” *Proceedings of the National Academy of Sciences of the United States of America* 104 Suppl 1 (May): 8605–12.
Prüfer, Kay, Fernando Racimo, Nick Patterson, Flora Jay, Sriram Sankararaman, Susanna Sawyer, Anja Heinze, et al. 2014. “The Complete Genome Sequence of a Neanderthal from the Altai Mountains.” *Nature* 505 (7481): 43–49.
Puig, Marta, Jon Lerga-Jaso, Carla Giner-Delgado, Sarai Pacheco, David Izquierdo, Alejandra Delprat, Magdalena Gayà-Vidal, Jack F. Regan, George Karlin-Neumann, and Mario Cáceres. 2020. “Determining the Impact of Uncharacterized Inversions in the Human Genome by Droplet Digital PCR.” *Genome Research* 30 (5): 724–35.
Purcell, Shaun, Benjamin Neale, Kathe Todd-Brown, Lori Thomas, Manuel A. R. Ferreira, David Bender, Julian Maller, et al. 2007. “PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses.” *American Journal of Human Genetics* 81 (3): 559–75.
Qian, Wenfeng, Ben-Yang Liao, Andrew Ying-Fei Chang, and Jianzhi Zhang. 2010. “Maintenance of Duplicate Genes and Their Functional Redundancy by Reduced Expression.” *Trends in Genetics*. https://doi.org/10.1016/j.tig.2010.07.002.
Quinlan, Aaron R., and Ira M. Hall. 2010. “BEDTools: A Flexible Suite of Utilities for Comparing Genomic Features.” *Bioinformatics* 26 (6): 841–42.
Ramani, Vijay, Darren A. Cusanovich, Ronald J. Hause, Wenxiu Ma, Ruolan Qiu, Xinxian Deng, C. Anthony Blau, et al. 2016. “Mapping 3D Genome Architecture through in Situ DNase Hi-C.” *Nature Protocols* 11 (11): 2104–21.
Rao, Suhas S. P., Su-Chen Huang, Brian Glenn St Hilaire, Jesse M. Engreitz, Elizabeth M. Perez, Kyong-Rim Kieffer-Kwon, Adrian L. Sanborn, et al. 2017. “Cohesin Loss Eliminates All Loop Domains.” *Cell* 171 (2): 305–20.e24.
Rao, Suhas S. P., Miriam H. Huntley, Neva C. Durand, Elena K. Stamenova, Ivan D. Bochkov, James T. Robinson, Adrian L. Sanborn, et al. 2014. “A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping.” *Cell* 159 (7): 1665–80.
Rastogi, Shruti, and David A. Liberles. 2005. “Subfunctionalization of Duplicated Genes as a

Transition State to Neofunctionalization." *BMC Evolutionary Biology* 5 (April): 28.
Reilly, Steven K., Jun Yin, Albert E. Ayoub, Deena Emera, Jing Leng, Justin Cotney, Richard Sarro, Pasko Rakic, and James P. Noonan. 2015. "Evolutionary Genomics. Evolutionary Changes in Promoter and Enhancer Activity during Human Corticogenesis." *Science* 347 (6226): 1155–59.
Reno, Philip L., Cory Y. McLean, Jasmine E. Hines, Terence D. Capellini, Gill Bejerano, and David M. Kingsley. 2013. "A Penile Spine/Vibrissa Enhancer Sequence Is Missing in Modern and Extinct Humans but Is Retained in Multiple Primates with Penile Spines and Sensory Vibrissae." *PloS One* 8 (12): e84258.
Rhesus Macaque Genome Sequencing and Analysis Consortium, Richard A. Gibbs, Jeffrey Rogers, Michael G. Katze, Roger Bumgarner, George M. Weinstock, Elaine R. Mardis, et al. 2007. "Evolutionary and Biomedical Insights from the Rhesus Macaque Genome." *Science* 316 (5822): 222–34.
Rhie, Arang, Shane A. McCarthy, Olivier Fedrigo, Joana Damas, Giulio Formenti, Sergey Koren, Marcela Uliano-Silva, et al. 2021. "Towards Complete and Error-Free Genome Assemblies of All Vertebrate Species." *Nature* 592 (7856): 737–46.
Rieseberg, L. H. 2001. "Chromosomal Rearrangements and Speciation." *Trends in Ecology & Evolution* 16 (7): 351–58.
Ritchie, Matthew E., Belinda Phipson, Di Wu, Yifang Hu, Charity W. Law, Wei Shi, and Gordon K. Smyth. 2015. "Limma Powers Differential Expression Analyses for RNA-Sequencing and Microarray Studies." *Nucleic Acids Research* 43 (7): e47.
Robinson, J. T., H. Thorvaldsdottir, W. Winckler, M. Guttman, E. S. Lander, G. Getz, and J. P. Mesirov. 2011. "Integrative Genomics Viewer." *Nature Biotechnology* 29 (1): 24–26.
Rodin, Sergei N., and Arthur D. Riggs. 2003. "Epigenetic Silencing May Aid Evolution by Gene Duplication." *Journal of Molecular Evolution* 56 (6): 718–29.
Rogers, Jeffrey, and Richard A. Gibbs. 2014. "Comparative Primate Genomics: Emerging Patterns of Genome Content and Dynamics." *Nature Reviews. Genetics* 15 (5): 347–59.
Rossman, Kent L., Channing J. Der, and John Sondek. 2005. "GEF Means Go: Turning on RHO GTPases with Guanine Nucleotide-Exchange Factors." *Nature Reviews. Molecular Cell Biology* 6 (2): 167–80.
Samocha, Kaitlin E., Elise B. Robinson, Stephan J. Sanders, Christine Stevens, Aniko Sabo, Lauren M. McGrath, Jack A. Kosmicki, et al. 2014. "A Framework for the Interpretation of de Novo Mutation in Human Disease." *Nature Genetics* 46 (9): 944–50.
Sanders, Ashley D., Ester Falconer, Mark Hills, Diana C. J. Spierings, and Peter M. Lansdorp. 2017. "Single-Cell Template Strand Sequencing by Strand-Seq Enables the Characterization of Individual Homologs." *Nature Protocols* 12 (6): 1151–76.
Sanders, Ashley D., Mark Hills, David Porubský, Victor Guryev, Ester Falconer, and Peter M. Lansdorp. 2016. "Characterizing Polymorphic Inversions in Human Genomes by Single-Cell Sequencing." *Genome Research* 26 (11): 1575–87.
Sandve, Simen R., Rori V. Rohlfs, and Torgeir R. Hvidsten. 2018. "Subfunctionalization versus Neofunctionalization after Whole-Genome Duplication." *Nature Genetics*. https://doi.org/10.1038/s41588-018-0162-4.
Schmidt, Ewoud R. E., Justine V. Kupferman, Michelle Stackmann, and Franck Polleux. 2019. "The Human-Specific Paralogs SRGAP2B and SRGAP2C Differentially Modulate SRGAP2A-Dependent Synaptic Development." *Scientific Reports* 9 (1): 18692.
Schmidt, Joshua M., Marc de Manuel, Tomas Marques-Bonet, Sergi Castellano, and Aida M. Andrés. 2019. "The Impact of Genetic Adaptation on Chimpanzee Subspecies Differentiation." *PLoS Genetics* 15 (11): e1008485.
Schubert, C. 2009. "The Genomic Basis of the Williams-Beuren Syndrome." *Cellular and Molecular Life Sciences: CMLS* 66 (7): 1178–97.
Scott, Alexandra J., Colby Chiang, and Ira M. Hall. 2021. "Structural Variants Are a Major

Source of Gene Expression Differences in Humans and Often Affect Multiple Nearby Genes." *Genome Research*, September. https://doi.org/10.1101/gr.275488.121.
Sebat, Jonathan, B. Lakshmi, Jennifer Troge, Joan Alexander, Janet Young, Pär Lundin, Susanne Månér, et al. 2004. "Large-Scale Copy Number Polymorphism in the Human Genome." *Science* 305 (5683): 525–28.
Sharp, Andrew J., Heather C. Mefford, Kelly Li, Carl Baker, Cindy Skinner, Roger E. Stevenson, Richard J. Schroer, et al. 2008. "A Recurrent 15q13.3 Microdeletion Syndrome Associated with Mental Retardation and Seizures." *Nature Genetics* 40 (3): 322–28.
Shaw, Christine J., and James R. Lupski. 2004. "Implications of Human Genome Architecture for Rearrangement-Based Disorders: The Genomic Basis of Disease." *Human Molecular Genetics* 13 (suppl_1): R57–64.
Shew, Colin J., Paulina Carmona-Mora, Daniela C. Soto, Mira Mastoras, Elizabeth Roberts, Joseph Rosas, Dhriti Jagannathan, Gulhan Kaya, Henriette O'Geen, and Megan Y. Dennis. 2021. "Diverse Molecular Mechanisms Contribute to Differential Expression of Human Duplicated Genes." *Molecular Biology and Evolution* 38 (8): 3060–77.
Shimada, M. K., C-G Kim, T. Kitano, R. E. Ferrell, Y. Kohara, and N. Saitou. 2005. "Nucleotide Sequence Comparison of a Chromosome Rearrangement on Human Chromosome 12 and the Corresponding Ape Chromosomes." *Cytogenetic and Genome Research* 108 (1-3): 83–90.
Shin, Hanjun, Yi Shi, Chao Dai, Harianto Tjong, Ke Gong, Frank Alber, and Xianghong Jasmine Zhou. 2016. "TopDom: An Efficient and Deterministic Method for Identifying Topological Domains in Genomes." *Nucleic Acids Research* 44 (7): e70.
Smyth, G. K. n.d. "Limma: Linear Models for Microarray Data." *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*. https://doi.org/10.1007/0-387-29362-0_23.
Soneson, Charlotte, Michael I. Love, and Mark D. Robinson. 2015a. "Differential Analyses for RNA-Seq: Transcript-Level Estimates Improve Gene-Level Inferences." *F1000Research* 4 (December): 1521.
———. 2015b. "Differential Analyses for RNA-Seq: Transcript-Level Estimates Improve Gene-Level Inferences." *F1000Research* 4 (1521): 1521.
Spielmann, Malte, Darío G. Lupiáñez, and Stefan Mundlos. 2018. "Structural Variation in the 3D Genome." *Nature Reviews. Genetics* 19 (7): 453–67.
Stankiewicz, Pawel and James R. Lupski. 2002. "Genome Architecture, Rearrangements and Genomic Disorders." *Trends in Genetics: TIG* 18 (2): 74–82.
Stankiewicz, Paweł, and James R. Lupski. 2010. "Structural Variation in the Human Genome and Its Role in Disease." *Annual Review of Medicine* 61: 437–55.
Stanyon, R., M. Rocchi, O. Capozzi, R. Roberto, D. Misceo, M. Ventura, M. F. Cardone, F. Bigoni, and N. Archidiacono. 2008. "Primate Chromosome Evolution: Ancestral Karyotypes, Marker Order and Neocentromeres." *Chromosome Research: An International Journal on the Molecular, Supramolecular and Evolutionary Aspects of Chromosome Biology* 16 (1): 17–39.
Starmer, Joshua, and Terry Magnuson. 2016. "Detecting Broad Domains and Narrow Peaks in ChIP-Seq Data with hiddenDomains." *BMC Bioinformatics* 17 (March): 144.
Stefansson, Hreinn, Agnar Helgason, Gudmar Thorleifsson, Valgerdur Steinthorsdottir, Gisli Masson, John Barnard, Adam Baker, et al. 2005. "A Common Inversion under Selection in Europeans." *Nature Genetics* 37 (2): 129–37.
Stefansson, Hreinn, Dan Rujescu, Sven Cichon, Olli P. H. Pietiläinen, Andres Ingason, Stacy Steinberg, Ragnheidur Fossdal, et al. 2008. "Large Recurrent Microdeletions Associated with Schizophrenia." *Nature* 455 (7210): 232–36.
Stranger, Barbara E., Matthew S. Forrest, Mark Dunning, Catherine E. Ingle, Claude Beazley, Natalie Thorne, Richard Redon, et al. 2007. "Relative Impact of Nucleotide and Copy

Number Variation on Gene Expression Phenotypes." *Science* 315 (5813): 848–53.
Sturtevant, A. H. 1917. "Genetic Factors Affecting the Strength of Linkage in Drosophila." *Proceedings of the National Academy of Sciences of the United States of America* 3 (9): 555–58.
Sudmant, Peter H., John Huddleston, Claudia R. Catacchio, Maika Malig, Ladeana W. Hillier, Carl Baker, Kiana Mohajeri, et al. 2013. "Evolution and Diversity of Copy Number Variation in the Great Ape Lineage." *Genome Research* 23 (9): 1373–82.
Sudmant, Peter H., Jacob O. Kitzman, Francesca Antonacci, Can Alkan, Maika Malig, Anya Tsalenko, Nick Sampas, et al. 2010. "Diversity of Human Copy Number Variation and Multicopy Genes." *Science* 330 (6004): 641–46.
Suzuki, Ikuo K., David Gacquer, Roxane Van Heurck, Devesh Kumar, Marta Wojno, Angéline Bilheu, Adèle Herpoel, et al. 2018. "Human-Specific NOTCH2NL Genes Expand Cortical Neurogenesis through Delta/Notch Regulation." *Cell* 173 (6): 1370–84.e16.
Szamalek, Justyna M., Violaine Goidts, Jeremy B. Searle, David N. Cooper, Horst Hameister, and Hildegard Kehrer-Sawatzki. 2006. "The Chimpanzee-Specific Pericentric Inversions That Distinguish Humans and Chimpanzees Have Identical Breakpoints in Pan Troglodytes and Pan Paniscus." *Genomics* 87 (1): 39–45.
Takahashi, Sakae, Yu-Hu Cui, Yong-Hua Han, Jesen A. Fagerness, Brian Galloway, Yu-Cun Shen, Takuya Kojima, Makoto Uchiyama, Stephen V. Faraone, and Ming T. Tsuang. 2008. "Association of SNPs and Haplotypes in APOL1, 2 and 4 with Schizophrenia." *Schizophrenia Research*. https://doi.org/10.1016/j.schres.2008.05.028.
Tan, Zheng, Andrew Minsoo Shon, and Carole Ober. 2005. "Evidence of Balancing Selection at the HLA-G Promoter Region." *Human Molecular Genetics* 14 (23): 3619–28.
Tewhey, Ryan, Dylan Kotliar, Daniel S. Park, Brandon Liu, Sarah Winnicki, Steven K. Reilly, Kristian G. Andersen, et al. 2018. "Direct Identification of Hundreds of Expression-Modulating Variants Using a Multiplexed Reporter Assay." *Cell* 172 (5): 1132–34.
Than, Nándor Gábor, Andrea Balogh, Roberto Romero, Eva Kárpáti, Offer Erez, András Szilágyi, Ilona Kovalszky, et al. 2014. "Placental Protein 13 (PP13) - A Placental Immunoregulatory Galectin Protecting Pregnancy." *Frontiers in Immunology* 5 (August): 348.
Than, Nandor Gabor, Roberto Romero, Morris Goodman, Amy Weckle, Jun Xing, Zhong Dong, Yi Xu, et al. 2009. "A Primate Subfamily of Galectins Expressed at the Maternal-Fetal Interface That Promote Immune Cell Death." *Proceedings of the National Academy of Sciences of the United States of America* 106 (24): 9731–36.
Torre-Ubieta, Luis de la, Jason L. Stein, Hyejung Won, Carli K. Opland, Dan Liang, Daning Lu, and Daniel H. Geschwind. 2018. "The Dynamic Landscape of Open Chromatin during Human Cortical Neurogenesis." *Cell* 172 (1-2): 289–304.e18.
Tuzun, Eray, Andrew J. Sharp, Jeffrey A. Bailey, Rajinder Kaul, V. Anne Morrison, Lisa M. Pertz, Eric Haugen, et al. 2005. "Fine-Scale Structural Variation of the Human Genome." *Nature Genetics* 37 (7): 727–32.
Uebbing, Severin, Jake Gockley, Steven K. Reilly, Acadia A. Kocher, Evan Geller, Neeru Gandotra, Curt Scharfe, Justin Cotney, and James P. Noonan. 2021. "Massively Parallel Discovery of Human-Specific Substitutions That Alter Enhancer Activity." *Proceedings of the National Academy of Sciences of the United States of America* 118 (2). https://doi.org/10.1073/pnas.2007049118.
Varki, Ajit, and Tasha K. Altheide. 2005. "Comparing the Human and Chimpanzee Genomes: Searching for Needles in a Haystack." *Genome Research* 15 (12): 1746–58.
Varki, Ajit. 2017. "Are Humans Prone to Autoimmunity? Implications from Evolutionary Changes in Hominin Sialic Acid Biology." *Journal of Autoimmunity* 83 (September): 134–42.
Ventura, M., N. Archidiacono, and M. Rocchi. 2001. "Centromere Emergence in Evolution."

Genome Research 11 (4): 595–99.
Ventura, Mario, Francesca Antonacci, Maria Francesca Cardone, Roscoe Stanyon, Pietro D'Addabbo, Angelo Cellamare, L. James Sprague, Evan E. Eichler, Nicoletta Archidiacono, and Mariano Rocchi. 2007. "Evolutionary Formation of New Centromeres in Macaque." *Science* 316 (5822): 243–46.
Ventura, Mario, Claudia R. Catacchio, Can Alkan, Tomas Marques-Bonet, Saba Sajjadian, Tina A. Graves, Fereydoun Hormozdiari, et al. 2011. "Gorilla Genome Structural Variation Reveals Evolutionary Parallelisms with Chimpanzee." *Genome Research* 21 (10): 1640–49.
Ventura, Mario, Jonathan M. Mudge, Valeria Palumbo, Sally Burn, Elisabeth Blennow, Mauro Pierluigi, Roberto Giorda, et al. 2003. "Neocentromeres in 15q24-26 Map to Duplicons Which Flanked an Ancestral Centromere in 15q25." *Genome Research* 13 (9): 2059–68.
Ventura, Mario, Stefania Weigl, Lucia Carbone, Maria Francesca Cardone, Doriana Misceo, Mariagrazia Teti, Pietro D'Addabbo, et al. 2004. "Recurrent Sites for New Centromere Seeding." *Genome Research* 14 (9): 1696–1703.
Vermunt, Marit W., Sander C. Tan, Bas Castelijns, Geert Geeven, Peter Reinink, Ewart de Bruijn, Ivanela Kondova, et al. 2016. "Epigenomic Annotation of Gene Regulatory Alterations during Evolution of the Primate Brain." *Nature Neuroscience* 19 (3): 494–503.
Vicente-Salvador, David, Marta Puig, Magdalena Gayà-Vidal, Sarai Pacheco, Carla Giner-Delgado, Isaac Noguera, David Izquierdo, et al. 2017. "Detailed Analysis of Inversions Predicted between Two Human Genomes: Errors, Real Polymorphisms, and Their Origin and Population Distribution." *Human Molecular Genetics* 26 (3): 567–81.
Visel, Axel, Simon Minovitsky, Inna Dubchak, and Len A. Pennacchio. 2007. "VISTA Enhancer Browser--a Database of Tissue-Specific Human Enhancers." *Nucleic Acids Research* 35 (Database issue): D88–92.
Vollger, Mitchell R., Philip C. Dishuck, Melanie Sorensen, Annemarie E. Welch, Vy Dang, Max L. Dougherty, Tina A. Graves-Lindsay, Richard K. Wilson, Mark J. P. Chaisson, and Evan E. Eichler. 2019. "Long-Read Sequence and Assembly of Segmental Duplications." *Nature Methods* 16 (1): 88–94.
Wang, Ting, Lucinda Antonacci-Fulton, Kerstin Howe, Heather A. Lawson, Julian K. Lucas, Adam M. Phillippy, Alice B. Popejoy, Mobin Asri, Caryn Carson, Mark J. P. Chaisson, Xian Chang, Robert Cook-Deegan, Adam L. Felsenfeld, Robert S. Fulton, Erik P. Garrison, Nanibaa' A. Garrison, Tina A. Graves-Lindsay, Hanlee Ji, Eimear E. Kenny, Barbara A. Koenig, Daofeng Li, Tobias Marschall, Joshua F. McMichael, Adam M. Novak, Deepak Purushotham, Valerie A. Schneider, Baergen I. Schultz, Michael W. Smith, Heidi J. Sofia, Tsachy Weissman, Paul Flicek, Heng Li, Karen H. Miga, Benedict Paten, Erich D. Jarvis, Ira M. Hall, Evan E. Eichler, David Haussler, et al. 2022. "The Human Pangenome Project: A Global Resource to Map Genomic Diversity." *Nature*. https://doi.org/10.1038/s41586-022-04601-8.
Wang, Ting, Lucinda Antonacci-Fulton, Kerstin Howe, Heather A. Lawson, Julian K. Lucas, Adam M. Phillippy, Alice B. Popejoy, Mobin Asri, Caryn Carson, Mark J. P. Chaisson, Xian Chang, Robert Cook-Deegan, Adam L. Felsenfeld, Robert S. Fulton, Erik P. Garrison, Nanibaa' A. Garrison, Tina A. Graves-Lindsay, Hanlee Ji, Eimear E. Kenny, Barbara A. Koenig, Daofeng Li, Tobias Marschall, Joshua F. McMichael, Adam M. Novak, Deepak Purushotham, Valerie A. Schneider, Baergen I. Schultz, Michael W. Smith, Heidi J. Sofia, Tsachy Weissman, Paul Flicek, Heng Li, Karen H. Miga, Benedict Paten, Erich D. Jarvis, Ira M. Hall, Evan E. Eichler, and David Haussler. 2022. "The Human Pangenome Project: A Global Resource to Map Genomic Diversity." *Nature* 604 (7906): 437–46.
Wang, Zhenglong, Yosuke Kumamoto, Ping Wang, Xiaoqing Gan, David Lehmann, Alan V. Smrcka, Lauren Cohn, Akiko Iwasaki, Lin Li, and Dianqing Wu. 2009. "Regulation of Immature Dendritic Cell Migration by RhoA Guanine Nucleotide Exchange Factor Arhgef5." *The Journal of Biological Chemistry* 284 (42): 28599–606.

"Website." n.d. https://doi.org/10.7554/eLife.41241.
Wedenoja, S., M. Yoshihara, H. Teder, H. Sariola, M. Gissler, S. Katayama, J. Wedenoja, et al. n.d. "Balancing Selection at HLA-G Modulates Fetal Survival, Preeclampsia and Human Birth Sex Ratio." https://doi.org/10.1101/851089.
Wellenreuther, Maren, and Louis Bernatchez. 2018. "Eco-Evolutionary Genomics of Chromosomal Inversions." *Trends in Ecology & Evolution* 33 (6): 427–40.
Wilson, Gary M., Stephane Flibotte, Perseus I. Missirlis, Marco A. Marra, Steven Jones, Kevin Thornton, Andrew G. Clark, and Robert A. Holt. 2006. "Identification by Full-Coverage Array CGH of Human DNA Copy Number Increases Relative to Chimpanzee and Gorilla." *Genome Research* 16 (2): 173–81.
Wray, Gregory A., Matthew W. Hahn, Ehab Abouheif, James P. Balhoff, Margaret Pizer, Matthew V. Rockman, and Laura A. Romano. 2003. "The Evolution of Transcriptional Regulation in Eukaryotes." *Molecular Biology and Evolution* 20 (9): 1377–1419.
Xue, Cheng, Muthuswamy Raveendran, R. Alan Harris, Gloria L. Fawcett, Xiaoming Liu, Simon White, Mahmoud Dahdouli, et al. 2016. "The Population Genomics of Rhesus Macaques (Macaca Mulatta) Based on Whole-Genome Sequences." *Genome Research* 26 (12): 1651–62.
Yang, Minjun, Setareh Safavi, Eleanor L. Woodward, Nicolas Duployez, Linda Olsson-Arvidsson, Jonas Ungerbäck, Mikael Sigvardsson, et al. 2020. "13q12.2 Deletions in Acute Lymphoblastic Leukemia Lead to Upregulation of FLT3 through Enhancer Hijacking." *Blood* 136 (8): 946–56.
Yang, Z. 1997. "PAML: A Program Package for Phylogenetic Analysis by Maximum Likelihood." *Computer Applications in the Biosciences: CABIOS* 13 (5): 555–56.
Yohn, Chris T., Zhaoshi Jiang, Sean D. McGrath, Karen E. Hayden, Philipp Khaitovich, Matthew E. Johnson, Marla Y. Eichler, et al. 2005. "Lineage-Specific Expansions of Retroviral Insertions within the Genomes of African Great Apes but Not Humans and Orangutans." *PLoS Biology*. https://doi.org/10.1371/journal.pbio.0030110.
Yunis, J. J., and O. Prakash. 1982. "The Origin of Man: A Chromosomal Pictorial Legacy." *Science* 215 (4539): 1525–30.
Yunis, J. J., J. R. Sawyer, and K. Dunham. 1980. "The Striking Resemblance of High-Resolution G-Banded Chromosomes of Man and Chimpanzee." *Science* 208 (4448): 1145–48.
Zappasodi, Roberta, Sadna Budhu, Mohsen Abu-Akeel, and Taha Merghoub. 2020. "In Vitro Assays for Effector T Cell Functions and Activity of Immunomodulatory Antibodies." *Methods in Enzymology* 631: 43–59.
Zhang, Jianzhi. 2003. "Evolution by Gene Duplication: An Update." *Trends in Ecology & Evolution* 18 (6): 292–98.
Zhang, Yong, Tao Liu, Clifford A. Meyer, Jérôme Eeckhoute, David S. Johnson, Bradley E. Bernstein, Chad Nusbaum, et al. 2008. "Model-Based Analysis of ChIP-Seq (MACS)." *Genome Biology* 9 (9): R137.
Zhao, Jian, Jianyang Ma, Yun Deng, Jennifer A. Kelly, Kwangwoo Kim, So-Young Bang, Hye-Soon Lee, et al. 2017. "A Missense Variant in NCF1 Is Associated with Susceptibility to Multiple Autoimmune Diseases." *Nature Genetics* 49 (3): 433–37.
Zheng, Ye, Ferhat Ay, and Sunduz Keles. 2019. "Generative Modeling of Multi-Mapping Reads with mHi-C Advances Analysis of Hi-C Studies." *eLife* 8 (January). https://doi.org/10.7554/eLife.38070.
Zhou, Xiang, Carolyn E. Cain, Marsha Myrthil, Noah Lewellen, Katelyn Michelini, Emily R. Davenport, Matthew Stephens, Jonathan K. Pritchard, and Yoav Gilad. 2014. "Epigenetic Modifications Are Associated with Inter-Species Gene Expression Variation in Primates." *Genome Biology* 15 (12): 547.
Zhu, Ying, Mingfeng Li, André M. M. Sousa, and Nenad Sestan. 2014. "XSAnno: A Framework for Building Ortholog Models in Cross-Species Transcriptome Comparisons." *BMC*

Genomics 15 (May): 343.
Zody, Michael C., Zhaoshi Jiang, Hon-Chung Fung, Francesca Antonacci, Ladeana W. Hillier, Maria Francesca Cardone, Tina A. Graves, et al. 2008. “Evolutionary Toggling of the MAPT 17q21.31 Inversion Region.” *Nature Genetics* 40 (9): 1076–83.
Zufferey, Marie, Daniele Tavernari, Elisa Oricchio, and Giovanni Ciriello. 2018. “Comparison of Computational Methods for the Identification of Topologically Associating Domains.” *Genome Biology* 19 (1): 217.