Structural Variation and Expression of Segmentally Duplicated Human Genes

Philip Dishuck

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2024

Reading Committee:

Evan Eichler, Chair

Willie Swanson

Phil Green

Program Authorized to Offer Degree:

Genome Sciences

University of Washington

Abstract

Structural Variation and Expression of Segmentally Duplicated Human Genes

Philip Dishuck

Chair of the Supervisory Committee:
Evan Eichler
Genome Sciences

Gene duplication is a major driver of evolution, and the African ape lineage, including humans, experienced a burst of segmental duplications (SDs). Recent gene duplications help explain the rapid phenotypic changes in humans despite a slowdown in point mutations in primates. However, these genes are particularly difficult to study due to limitations in sequencing and assembly—the first complete human assembly, including all segmentally duplicated genes, was not finished until 2022. Long-read DNA sequencing (PacBio HiFi [high-fidelity] and ONT [Oxford Nanopore Technologies]) now enables the routine assembly of highly contiguous human genomes, and long-read cDNA sequencing (Iso-Seq) allows paralog-specific assessment of gene models and identification of isoforms. In this thesis, I analyze the gene duplications of some of the first human HiFi and ONT assemblies and use Iso-Seq to functionally annotate recent duplications. I characterized 170 highly contiguous human haplotypes containing 47 Mbp of additional SD content absent from the first complete reference assembly. Using Iso-Seq, I

annotated the segmentally duplicated genes in these assemblies, discovering 201 new genes in copy number polymorphic gene families. These include a coding gene fusion *NSFP1-LRRC37A2* in an inverted form of the *MAPT* (tau) locus, and a KRAB-zinc finger gene present in 36% haplotypes that has only 69% amino acid identity to the best-matching annotated human gene. To validate long-read assemblies, I created a method, called GAVISUNK, that uses the distance between singly unique nucleotide k-mers (SUNKs) in ultra-long ONT reads to validate the structure of HiFi assemblies. This method identifies structural errors in assemblies and allows confident downstream of analysis of structural variation, unbiased by assembly artefacts. I performed a detailed analysis of a high copy number gene family, *NPIP*, which displays signatures of positive selection on the human and African ape lineage. Of 28 named human paralogs, I found that just three are fixed at a single copy (*NPIPB2*, *B11*, and *B14*). I found evidence of ongoing gene duplications, deletions, interlocus gene conversion, and large inversions mediated by *NPIP* duplication blocks. Two paralogs (*B9* and *B15*) were within the most extreme percentile of tests for positive selection and selective sweeps. Full-length cDNA from 101 tissue/cell types revealed distinct gene models for subgroups of *NPIP*s, including a variable number tandem repeat (VNTR) that encodes a variably sized beta helix. Paralogs in that subgroup show enriched expression in brain tissue, while others retain the ancestral testis-enriched expression. These analyses reveal mechanisms for rapid evolution of duplicated genes and demonstrate their polymorphism among humans.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

Scientific discovery is the greatest thrill I know, and for allowing me to contribute to it, I am deeply indebted to all those who have led me here. To my family who gave me the freedom, encouragement, and means to pursue my passions no matter how obscure, thank you.
I am grateful to my high school physics teacher, William Hooper, who showed me the simple beauty of a good thought experiment and introduced me to Jane Rasco at the University of Alabama, who integrated me into her lab's experimental work as a teenager, allowing me to believe I could make science part of my life.

Thanks to my undergraduate PI, Natalia Toporikova, and our collaborators Cleyde Helena and Joel Tabak for showing me what a fun place collaborative science can be, and convincing me that if you can't simulate it, you don't fully understand it.

I am grateful for my thesis advisor, Evan Eichler, for taking a chance on someone without much experience in genetics and providing constant mentorship and encouragement throughout. He is one of the clearest thinkers I have ever worked with and has done his best to teach me the value functions of modern science. Thanks to all the Eichler lab members, current and former, who make it a great place to learn and explore genetics, especially to Max Dougherty, Mitchell Vollger, Glennis Logsdon, Katy Munson, Alex Lewis, Jason Underwood, William Harvey, PingHsun Hsieh, DongAhn Yoo, Hyeonsoo Jeong, Michelle Noyes, Xavi Guitart, Taylor Real, and Lizzie Plender. Tonia Brown has been a steadfast supporter and provided valuable advice and edits for all of my writing.

I would not have made it here without the support of my cohort – thanks in particular to Will DeWitt for letting me borrow a quiet cubicle for writing and encouraging me to think more broadly.

My committee members Willie Swanson, Kelley Harris, Phil Green, Adam Leache, and Armita Nourmohammad have given me invaluable support, guidance, and perspective throughout my PhD, and I am grateful that they have been so generous with their expertise.

Most of all, my sincerest thanks to the tissue donors who make human genetics possible, and to the funders who see our work as a worthwhile endeavor.

CHAPTER 1. INTRODUCTION

1.1 EVOLUTION THROUGH GENE DUPLICATION

Gene duplication presents a tantalizingly tractable mechanism for functional innovation. In 1970, Ohno argued for gene duplication as "the major force of evolution," because the functional redundancy between paralogs relaxes selective pressure, allowing exploration of a larger mutational space (Ohno 1970). Pseudogenization is presumed to be the most common fate of a new paralog, as most mutations are deleterious, yet analysis of human and mouse gene families finds that about half of duplicated genes are retained as functional genes, more than expected by neutral evolution (Nadeau and Sankoff 1997), and experience a period of accelerated evolution (Pegueroles et al. 2013). If they survive, paralogs may neofunctionalize, evolving a novel function; or subfunctionalize, distributing their functions between the two (Force et al. 1999).

We estimate that 7% of the human genome consists of segmental duplications (SDs), operationally defined as regions of >1 kbp sharing at least 90% sequence identity, though initial estimates were ~5% (Vollger, Guitart, Dishuck et al. 2022; Bailey et al. 2002). SDs are enriched in exons compared to the remainder of euchromatic regions (She et al. 2006) and are particularly dynamic, with a tenfold excess of copy number polymorphism (Sudmant et al. 2015a), largely because they are subject to non-allelic homologous recombination (NAHR) (Lupski and Stankiewicz 2005).

The interspersed nature of primate SDs (Bailey and Eichler 2006) creates perhaps even simpler modes of neofunctionalizion than Ohno imagined. A gene in a new genomic context may co-opt regulatory elements and alter the context of its expression, fuse with an adjacent gene, or a

truncated gene may instantaneously gain a dominant negative interaction with the ancestral copy as in the case *SRGAP2C* (Dennis et al. 2012). In addition to *SRGAP2C*, several other recently duplicated human genes have been implicated in brain development or neuronal function, including *ARHGAP11B* (Fischer et al. 2022), *CROCCP2* (Van Heurck et al. 2023), *NOTCH2NL* (Fiddes et al. 2018), and *TBC1D3* (Ju et al. 2016). In the endeavor to understand our uniquely human traits, these regions should not be ignored.

King and Wilson recognized an apparent mismatch between human phenotypic evolution and the observed rate of molecular evolution relative to chimpanzee (King and Wilson 1975). In contrast to the slowdown in the point mutation rate in hominoids (Steiper et al. 2004), primates have had an accelerated rate of gene gain and loss (Hahn et al. 2007), and the ancestral great ape in particular had an accelerated rate of duplications, perhaps providing a template for genetic innovation (Marques-Bonet et al. 2009a; Yoo et al. 2024).

1.2 THE COST OF GENE DUPLICATION: GENOMIC INSTABILITY AND DISEASE

The high-identity interspersed duplications that have created this genetic innovation present an evolutionary trade-off, as they predispose the sequence intervening the duplications to microduplications, microdeletions, and inversions through NAHR (Stankiewicz and Lupski 2002). An alluring hypothesis is that the evolution of genes that make us uniquely human has left portions of our chromosomes in a fragile state, causing uniquely human disease. For example, the recurrent microdeletion at 16p11.2 mediated by *NPIP* duplications, the topic of Chapter 4, is the second-most common cause of autism, and its reciprocal duplication is associated with a 14.5-fold increased risk of schizophrenia (McCarthy et al. 2009). Structural variants (SVs), commonly mediated by SDs, create more base-pair variation between individuals than single-nucleotide

variants (SNVs) or indels. SVs are estimated to be 28-54 times more likely to affect gene expression than a SNV or indel (Sudmant et al. 2015b; Chiang et al. 2017), making them an important source of phenotypic variation. Though short-read and array-based disease association studies may detect a tagging single-nucleotide polymorphism in linkage disequilibrium with an SV, this is dependent on proper assembly of the locus and the proximity of mappable variants to the SV. Understanding the role of structural variation in diversity and disease will require assembly and functional annotation of many individual genomes.

1.3 Recent developments in genome assembly and annotation

Though the Human Genome Project declared victory in 2003, it was not until 2022 that we fully completed the assembly of a single human haplotype (Nurk … Dishuck et al. 2022). This delay was not due to lack of effort, but because the longest and most recent genomic duplications are recalcitrant to standard assembly and analysis methods. Shotgun Sanger sequencing failed to resolve most SDs, though read-depth pileups can be used to infer the copy number, but not structure, of duplications (She et al. 2004). Bacterial artificial chromosome (BAC) libraries improve SD assembly via physical separation, but the approach is laborious and still leaves assembly gaps (Bailey et al. 2001, 2002). Two innovations in long-read sequencing, namely PacBio HiFi (>10 kbp, >99.9% accuracy) and ultra-long ONT (>100 kbp, >95% accuracy) allow the assembly of even the longest and most repetitive genome regions (Chaisson et al. 2015; Jain et al. 2018; Wenger et al. 2019). The first complete assembly of a human haplotype and the advancement in methods spurred on by that effort created the potential to study SDs more comprehensively than ever before.

Two consortia are using LRS to create highly contiguous genome assemblies of diverse panels of humans: the Human Genome Structural Variation Consortium (HGSVC) and Human Pangenome Reference Consortium (HPRC) (Ebert et al. 2021; Wang et al. 2022). These assemblies have been a critical resource for my research, serving as the beginning of a catalog of normal human variation within segmentally duplicated genes.

These same innovations in long-read sequencing also enable functional annotation of duplicated sequence. Because short-read RNA-sequencing (RNA-seq) does not align uniquely to high sequence identity paralogs, their annotated gene models are often incorrect, missing paralog-specific isoforms or mistaking genes for pseudogenes and vice versa (Chaisson et al. 2015; Vollger et al. 2019). Similarly, tissue-specific expression differences are obscured by the ambiguous read alignments. Full-length cDNA sequencing with PacBio circular consensus sequencing (Iso-Seq) enables transcripts to be uniquely assigned to even recent gene duplications (Dougherty et al. 2018). This development allows the most recent human genes to be studied on equal footing with the rest of the genome to further our understanding of human evolution and disease.

1.4 Research Goals

The overarching goal of this thesis is to further the understanding of human variation in segmentally duplicated genes through the application of long-read sequencing. We have just finished the assembly of the first complete human haplotype (Nurk et al. 2022), yet we know a single reference cannot represent all individuals As long-read technologies have become more affordable, we are now able to create contiguous assemblies of these recent duplications for a sample of the human population. First, which genes exist across humans – are they fixed in copy number or polymorphic? Using highly contiguous assemblies of PacBio HiFi and ONT sequence,

I characterized the structural variation that creates new genes in different individuals. The absence of deletions of a paralog in an otherwise copy number polymorphic gene family hints at essentiality, though stochasticity or differences in susceptibility to various mutational mechanisms may also explain the absence of deletions. Second, are these duplicate genes even expressed, and if so, do they retain the ancestral gene model and expression pattern? Because short-read RNA-seq does not align uniquely to recently duplicated genes, their gene models are often mis-annotated, hampering attempts to understand their function. I performed full-length cDNA sequencing to correctly annotate these genes and used hybridization capture to target the cDNA sequencing to gene families of interest, allowing more comprehensive characterization of their gene models. Finally, how much can we trust these LRS assemblies? At the cutting edge of sequencing technology, ground truth is hard to come by. I developed a method, GAVISUNK, to use orthogonal ONT sequencing of HiFi assembles to validate their structure using the spacing of unique k-mers.

1.5 TOPICS IN THIS DISSERTATION

In Chapter 2, I characterize segmentally duplicated genes in 170 of the first human haplotypes assembled using HiFi sequencing, which fully resolves the majority of autosomal SDs. We observe 47 Mbp of SDs absent from the first complete reference assembly and found that African genomes are more likely to have higher copy number of recently duplicated gene families compared to non-African genomes. I used full-length cDNA sequencing of 563 million reads from 67 tissues to discover 201 new genes in copy number polymorphic gene families, including a coding gene fusion in an inverted form of the *MAPT* (tau) locus and a KRAB-zinc finger protein present in 36% of haplotypes that has only 69% amino acid identity to its closest-matching annotated human gene.

In Chapter 3, I develop a method for orthogonal validation of diploid HiFi assemblies with ultra-long ONT sequence. The contiguity of a genome assembly does not guarantee its validity—the same high identity between SDs (often >99%) that allows for high rates of structural variation also create ambiguous assembly graphs leading to frequent misassembly. Ultra-long ONT reads (100 kbp – ~4 Mbp) can span most SD blocks, but their relatively low sequence accuracy (~95%) relative to SDs means that simple read alignments are not reliable enough to confirm assembly accuracy. Instead, my method compares the spacing of SUNKs between the HiFi assembly and ONT reads to detect misassemblies that prior approaches missed.

In Chapter 4, I narrow my focus to a single high-copy number gene family, *NPIP*. Using 169 highly contiguous assembled haplotypes from HiFi and ONT sequencing, I catalog the structural diversity of this gene family. Of 28 paralogs, only three (*NPIPB2*, *B11*, and *B14*) are fixed at copy number one, while four may be duplicated but are never deleted (*A2*, *A4*, *B12/B13*, and *B15*), a possible indication of loss intolerance. Short-read RNA-seq does not map uniquely to individual paralogs, so I use full-length cDNA sequencing data from 101 tissues to create 55 paralog-specific gene models, 50 of which were not represented in RefSeq, and observed rapid diversification of gene models and predicted protein features. Five paralogs showed enriched fetal or adult brain expression, while eight are testis-enriched, the presumed ancestral state. I search for unique k-mers in an atlas of developmental short-read RNA-seq, showing brain-enriched paralogs increasing in abundance postnatally and testis-enriched paralogs at puberty. Extreme positive selection (dN/dS) had previously been reported for the *NPIP* family, yet the function remains unknown. I applied haplotype-based selection tests to these assemblies and detected two paralogs (*NPIPB9* and

NPIPB15) with positive selection signatures in the one percent of most extreme values chromosome-wide.

CHAPTER 2. STRUCTURAL POLYMORPHISM AND DIVERSITY OF HUMAN SEGMENTAL DUPLICATIONS

Chapter 2 is adapted with minimal modification from:

Hyeonsoo Jeong[*], Philip C. Dishuck[*], DongAhn Yoo[*], William T. Harvey, Katherine M. Munson, Alexandra P. Lewis, Jennifer Kordosky, Gage H. Garcia, Human Genome Structural Variation Consortium (HGSVC), Feyza Yilmaz, Pille Hallast, Charles Lee, Tomi Pastinen, Evan E. Eichler

This work has been accepted for publication in *Nature Genetics*.

[*]These authors contributed equally to this work.

Author contributions:

H.J., D.Y., P.C.D., and E.E.E. conceived the project. K.M.M., A.P.L., J.K., G.H.G., and T.P. generated sequencing data. F.Y., P.H., and C.L. generated genome assemblies. W.T.H. performed quality-control analyses. H.J. and D.Y. analyzed sequencing data and segmental duplications (Figures 1-4). P.C.D. analyzed gene duplications, Iso-Seq expression, and short-read copy number estimates (Figures 5-7). H.J., D.Y., P.C.D., and E.E.E. drafted the manuscript.

2.1 ABSTRACT

Segmental duplications (SDs) contribute significantly to human disease, evolution, and diversity yet have been difficult to resolve at the sequence level. We present a population genetics survey of SDs by analyzing 170 human genome assemblies (from 85 samples representing 38 Africans and 47 non-Africans) where the majority of autosomal SDs are fully resolved using long-read sequence assembly. Excluding the acrocentric short arms and sex chromosomes, we identify 173.2 Mb of duplicated sequence (47.4 Mb not present in the telomere-to-telomere reference) distinguishing fixed from structurally polymorphic events. We find that intrachromosomal SDs are among the most variable with rare events mapping near their progenitor sequences. African genomes harbor significantly more intrachromosomal SDs and are more likely to have recently duplicated gene families with higher copy number when compared to non-African samples. A comparison to a resource of 563 million full-length Iso-Seq reads identifies 201 novel, potentially protein-coding genes corresponding to these copy number polymorphic SDs.

2.2 INTRODUCTION

The first draft sequences of the human genome (International Human Genome Sequencing Consortium 2001) revealed a surprising degree of high-identity duplications dispersed both interchromosomally and intrachromosomally. Segmental duplications (SDs) have been operationally defined as blocks of homologous DNA greater than 1 kb in length with >90% sequence identity (Bailey et al. 2001). In humans, ~60% of the pairwise alignments are interspersed, i.e., separated by more than 1 Mb within a given chromosome or mapping to nonhomologous chromosomes (Eichler 1997; Trask et al. 1998). Because of their size and high

degree of sequence identity, SDs have been some of the last regions of the human genome to be fully resolved (Church 2022). Originally estimated at 5% of the genome, the relative proportion within the telomere-to-telomere (T2T) genome has increased to ~7%, especially as the acrocentric regions of the short arms of human chromosomes have become fully characterized at the sequence level (Nurk et al. 2022; Vollger et al. 2022).

SDs show a wide range of copy number variation in the human species and contribute to structural variation as a result of unequal crossing over (aka non-allelic homologous recombination or NAHR). These structural variants (SVs) contribute to more base-pair differences between humans than those contributed by single-nucleotide variants (SNVs) or indel polymorphisms. Based on sequence read-depth analysis of the short-read sequencing data from the 1000 Genomes Project (1KG) (1000 Genomes Project Consortium et al. 2012), for example, we estimated that 50% of all copy number polymorphisms in the human species >1 kb in length map to SDs—an ~10-fold enrichment (Sudmant et al. 2013). Importantly, almost all copy number polymorphic genes in the human species map to these particular regions of the genome (Liao et al. 2023). Such copy number polymorphic genes have been strongly implicated in a variety of human diseases ranging from immune/autoimmune (*FCGR*) (Hargreaves et al. 2015; Rahbari et al. 2017; Hujoel et al. 2024), neurological (*C3/C4*) (Yang et al. 2007; Sekar et al. 2016), to coronary heart disease (*LPA*) (Trégouët et al. 2009; Clarke et al. 2009; Fitzgerald and Birney 2022). More recently, it has become apparent that genes embedded within SDs play an important role in the evolution of our species, including the expansion of the human frontal cortex (*SRGAP2C* (Dennis et al. 2012), *ARHGAP11B* (Florio et al. 2015), *TBC1D3* (Ju et al.

2016)), adaptation to starch-rich diets (amylase (Groot et al. 1989; Perry et al. 2007)), or even the development of color vision within the primate lineage (green and red opsins (Dulai et al. 1999)).

Notwithstanding their importance, understanding the genetic diversity of these more complex regions of the genome has been challenging. Most efforts have focused on estimating copy number by mapping short-read data back to a singular reference to discover copy number variant (CNV) regions (Tuzun et al. 2005; Mills et al. 2011; Sudmant et al. 2015b). Such short-read investigations are useful but incomplete with respect to genetic characterization of these loci. For example, read-depth analyses can be used to accurately estimate copy number differences in a diploid genome; however, they provide limited information about the location or the structure of the duplicated genes or the structure of the associated CNVs. Similarly, while actual protein-coding differences can be inferred from short-read alignments, these differences are not readily phased, especially in high-identity SDs and, thus, genes cannot be fully reconstructed limiting the potential to distinguish pseudogenes from genes. Finally, mapping short reads to a reference genome introduces reference bias since, until recently, the human reference genome was incomplete—with gaps enriched precisely over the most duplicated regions. Advances in long-read sequencing technology over the past four years have addressed these limitations by allowing high-identity regions to be fully phased and assembled allowing the haplotype, structure, and gene annotation to be investigated in many cases for the first time in the human population (Chaisson et al. 2015; Vollger et al. 2019, 2022). In particular, the development of PacBio HiFi (high-fidelity) sequencing technology and associated assembly algorithms (Cheng et al. 2021; Rautiainen et al. 2023) has meant that most SD regions can be fully sequence resolved at the haplotype level. In this study, we sought to investigate the population genetic diversity of SDs by

focusing on 170 human genome assemblies for which HiFi sequence data had been collected as part of the Human Pangenome Reference Consortium (HPRC) and Human Genome Structural Variation Consortium (HGSVC) (Ebert et al. 2021; Liao et al. 2023).

2.3 RESULTS

2.3.1 *Distribution of shared versus polymorphic SDs*

In this study, we analyzed 170 independent genome assemblies and identified SDs (>1 kb and >90%) from 85 human specimens representing 38 African and 47 non-African samples (Supplementary Tables 1 and 2). To investigate how autosomal SD patterns vary among human genomes, we mapped SDs back to the T2T human reference genome (T2T-CHM13) classifying events as either known or new with respect to that reference and then assessed whether they were shared or variable among the 170 human haplotypes. We used these data to estimate the allele frequency of inter and intrachromosomal duplications creating a pangenome representation of human SDs (Fig. 1). Because of the difficulties in both assembly and mapping of acrocentric SDs, we excluded all short arms or acrocentric chromosomes from this analysis. Acrocentric portions of the human genome are almost entirely composed of repetitive sequences—in fact, the largest and most identical duplications map to this portion of the genome. Moreover, ectopic recombination is rampant among these five chromosomes making reference mapping almost impossible and delineation of inter and intrachromosomal SDs extremely challenging. Consequently, these are frequently the last portions of the genome to be accurately assembled and sequenced and require the generation of T2T genomes (Nurk et al. 2022; Vollger et al. 2022). In total, we identified 2,742 intrachromosomal and 4,772 interchromosomal nonoverlapping SD regions, constituting 6.1% of the genome or 173.21 Mb (150.12 Mb and

73.95 Mb for intra and interchromosomal SDs, respectively; 50.86 Mb overlapped between intra and interchromosomal SDs) based on the genomic coordinates of the T2T-CHM13 genome.

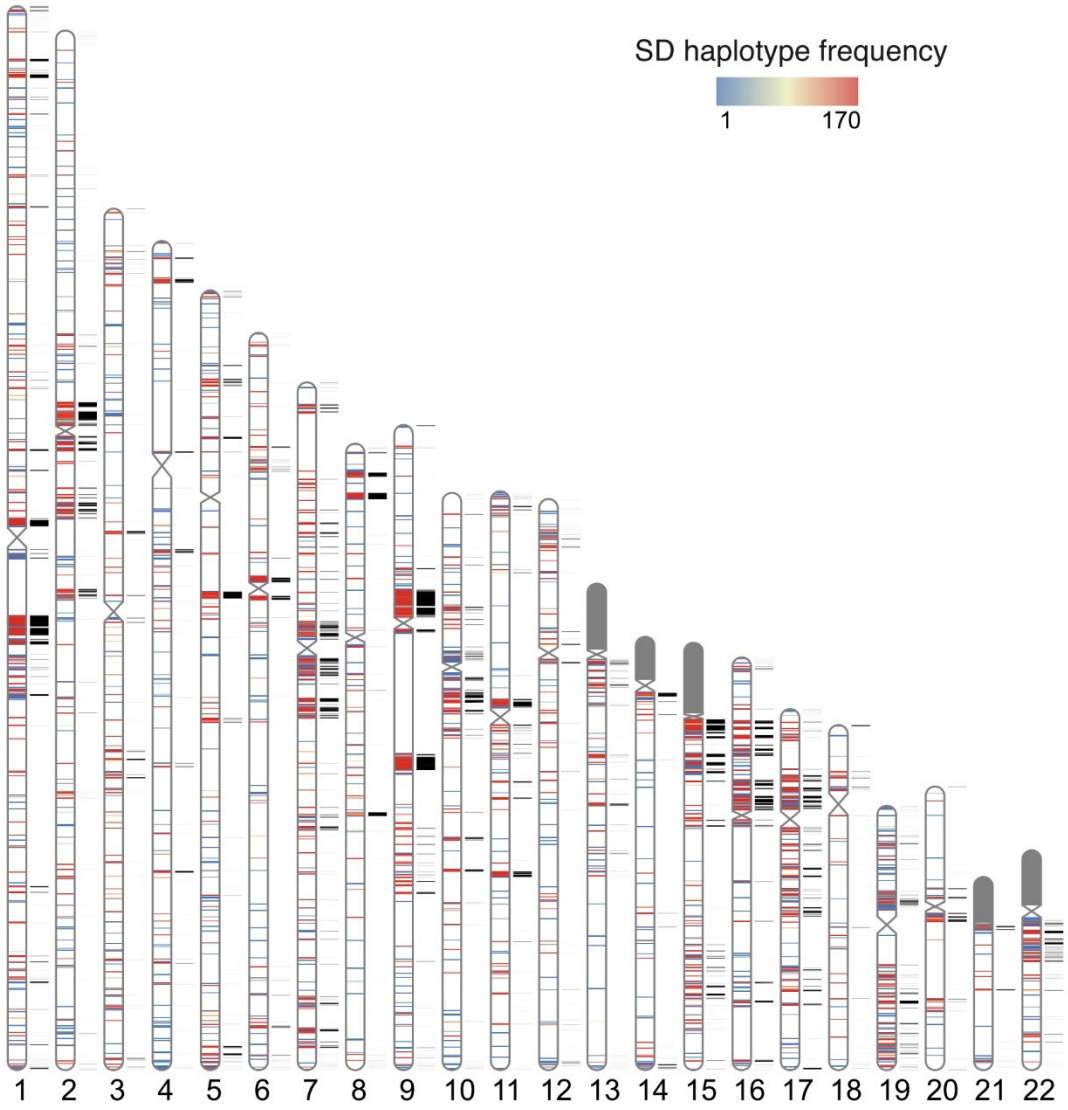


Figure 2.1. Pangenome representation of human segmental duplications (SDs)

Haplotype frequency distribution of intrachromosomal SD content from HPRC and HGSVC haplotype genome assemblies (n = 170). SDs are colored by the haplotype frequency. SD content on the p-arms of acrocentric chromosomes (chr13, chr14, chr15, chr21, and chr22) was excluded due to assembly errors and potential chromosomal misassignment compared to other autosomal

chromosomes. The known SDs of T2T-CHM13 are shown in black next to the ideograms on each chromosome.

Compared to the T2T-CHM13 human genome, we classify 47.4 Mb overall as newly discovered with an estimated rate of accumulation of 408.3 kb SDs being added for each additional sequenced and assembled human genome (Fig. 2). The majority of these novel SDs map intrachromosomally (41.7 Mb), although we classify 7.4 Mb as interchromosomal, and a significant fraction (24.6%) of interchromosomal SDs map to subtelomeric and pericentromeric regions of the human genome (p-value < 0.01, odds ratio = 2.39). Overall, a greater fraction of interchromosomal SDs (78.4%; 16.1 Mb of variable vs. 57.8 Mb of invariant SD regions) is fixed when compared to intrachromosomal events (59.7%; 60.3 Mb of variable vs. 89.7 Mb of invariant SD regions). With respect to intrachromosomal duplications, we find the majority of novel SDs tend to occur in close proximity to previously known SDs (empirical p-value < 0.01, permutation test) although we do note that certain chromosome arms and regions (e.g., p-arms of chromosomes 8, 10, 16, 17, 19 and q-arms of chromosomes 1, 15, 22) show an excess of these rarer SDs (Fig. 1 and Supplementary Table 3). As expected from other structural variation studies, the accumulation of novel SDs shows an asymptotic relationship with increasing sample size and the accumulation is greater for the additional African samples owing to their overall increased genetic diversity (Fig. 2).

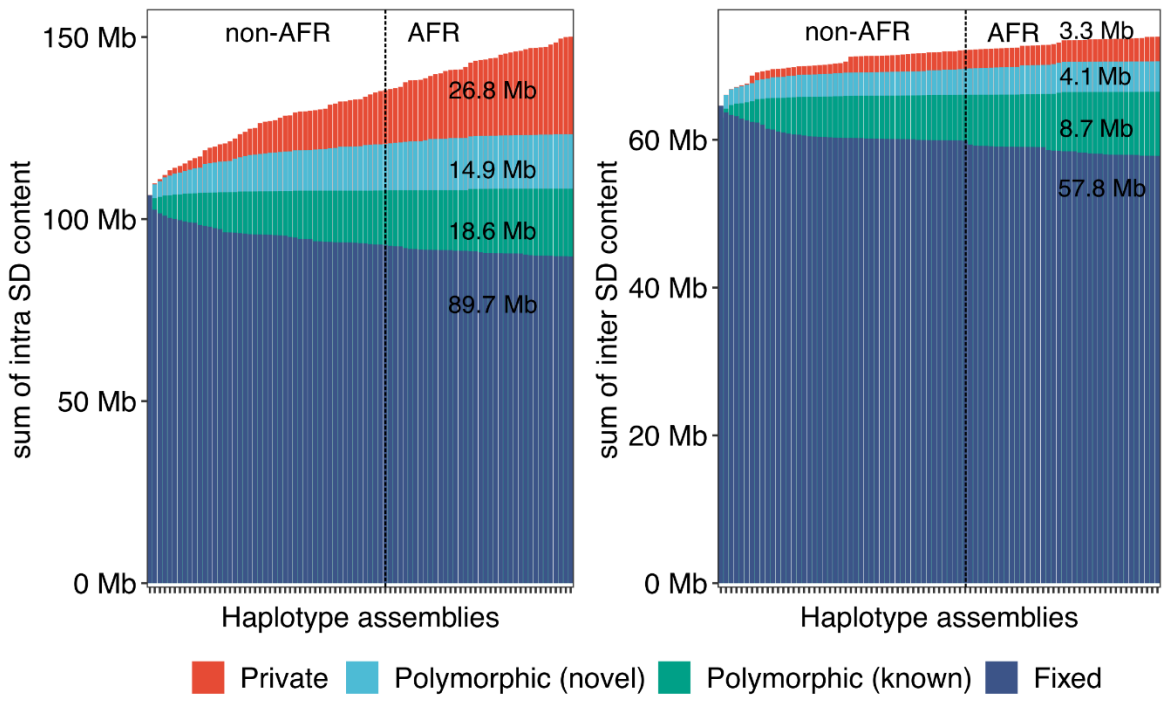


Figure 2.2. Cumulative sum of SDs by frequency.

Bar plot displays the cumulative sum of SD content by adding genomes (from left to right) for intrachromosomal and interchromosomal SDs. Four SD frequency categories are considered: "Fixed" are SDs present in all 170 human genome assemblies (i.e., conserved in all samples); "Polymorphic (known)" are SDs in the reference genome (T2T-CHM13) that are not fixed; "Polymorphic (novel)" refers to SDs observed in two or more HPRC/HGSVC assemblies yet not present in T2T-CHM13; "Private" is an SD found in one sample. Samples are grouped by non-African (non-AFR) and then African (AFR) genetic ancestry due to the expected increased diversity among the latter.

2.3.2 *Sequence properties of polymorphic and rare SDs*

Among the polymorphic SDs, we further distinguished two groups: rare SDs observed in up to five human genomes (<3% allele frequency) and common SDs observed between 6-20 times (~3-10% allele frequency). We find that rare SDs tend to be longer and have higher sequence identity between SD pairs when compared to common SDs (empirical p-value < 0.01, permutation test) (Fig. 3A and Supplementary Fig. 1). These features are consistent with a more recent origin for the majority of rare SDs; however, there are still SDs that show high sequence divergence that occur at low frequency in the human population and these may represent ancient SDs that are being lost (Fig. 3A). Notably, we find that low-frequency SDs also tend to be more distant from known SDs than those with a higher allele frequency in the population suggesting that the interspersion process characteristic of ape genomes is still ongoing in the human population (She et al. 2006; Marques-Bonet et al. 2009b).

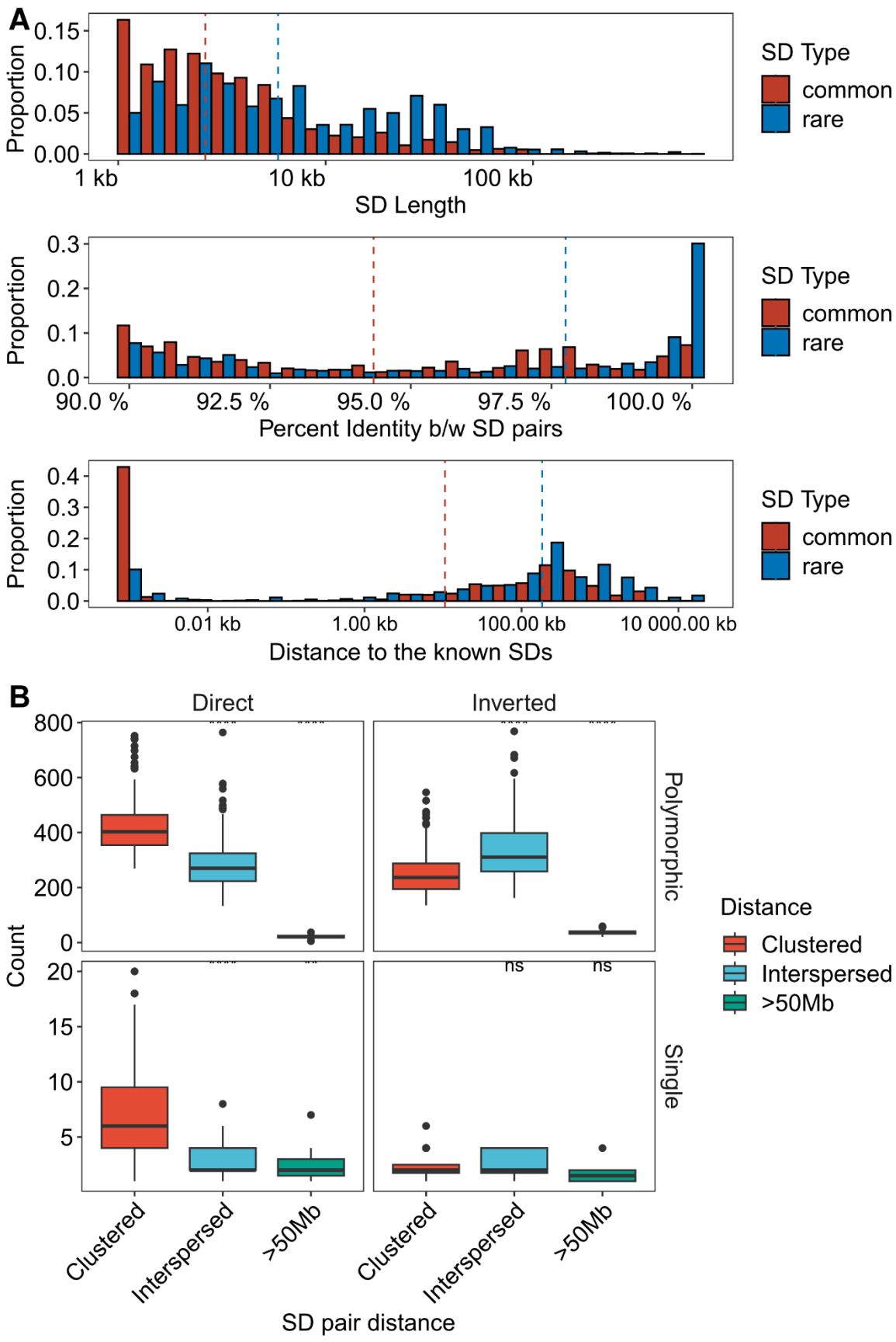

A
Proportion
0.15
0.10
0.05
0.00
1 kb
10 kb
100 kb
SD Length
SD Type
common
rare
Proportion
0.3
0.2
0.1
0.0
90.0 %
92.5 %
95.0 %
97.5 %
100.0 %
Percent Identity b/w SD pairs
SD Type
common
rare
Proportion
0.4
0.3
0.2
0.1
0.0
0.01 kb
1.00 kb
100.00 kb
10 000.00 kb
Distance to the known SDs
SD Type
common
rare
B
Direct
Inverted
Polymorphic
Single
Count
800
600
400
200
0
20
15
10
5
****
**
ns
Clustered
Interspersed
>50Mb
SD pair distance
Distance
Clustered
Interspersed
>50Mb

Figure 2.3. Sequence properties of polymorphic versus rare SDs.

(A) Histogram comparing the sequence identity and length of rare and common SDs (see Supplementary Fig. 1 for polymorphic SDs with more subclassified haplotype frequencies). **(B)** Orientation and pairwise dispersion of polymorphic and singleton SDs. Each data point represents haplotype assembly, and their counts of clustered, interspersed (>1 Mb apart), and distant (>50 Mb apart) SDs. Left and right panels summarize the SDs in direct or inverted orientation while the top and bottom panels contrast polymorphic vs. singleton SDs.

We also considered the orientation and configuration of singleton (single occurrence in a genome validated by read depth) versus polymorphic (not fixed in all human genome assemblies with allele frequency below 90%) SDs. We find that the vast majority (89.1%) of all SD singletons are clustered irrespective of whether they exist in a direct or an inverted orientation (Fig. 3B; Supplementary Table 4). We find that the proportion of polymorphic SDs classified as interspersed (SD pairs separated by more than 1 Mb) increases in approximately equal proportions between inverted and directly orientated SDs (Fig. 3B). However, interspersed polymorphic SDs favor an inverted orientation ($p = 9.6 \times 10^{-10}$, odds ratio = 1.99, Fisher's exact test). Figure 4 depicts examples of the structure of rare SDs in an inverted orientation.

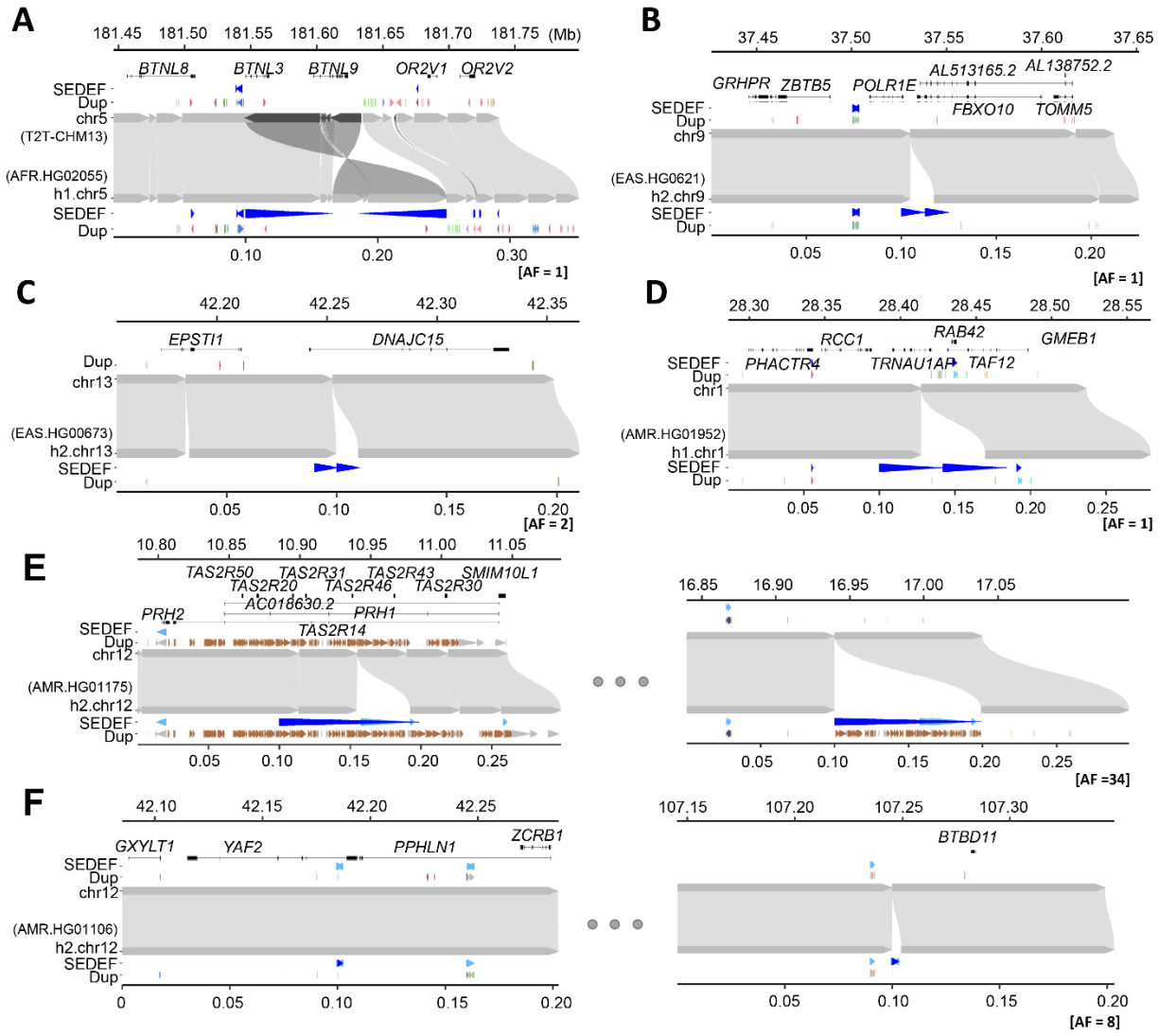


Figure 2.4. Examples of clustered (A-D) and interspersed (E-F; >1 Mb apart) SDs associated with genes.

In each plot, the top represents the T2T-CHM13 genome aligned to bottom, new genome assemblies. **(A)** Clustered duplication with inverted orientation (65.8 kb; with allele frequency [AF] = 1) found in chr5 and **(B-D)** clustered and tandem duplications (12.6, 10.3 and 42.3 kb; with AF of 1, 2 and 1, respectively) in chr9, chr13 and chr1. **(E-F)** Interspersed duplications of chr 12 (98.9 and 2.5 kb; with AF = 34 and 8) showing duplicated regions in left and right panels. The gene track of the T2T-CHM13 genome assembly is shown at the top, followed by SDs predicted by SEDEF and the respective direction indicated by blue arrowheads. The DupMasker track shows the duplicon structure.

2.3.3 *Gene content and population differences in copy number*

Based on the current gene annotation of the T2T-CHM13 genome assembly, we estimate that there are 1,156 duplicated protein-coding genes (standard deviation = 49) per diploid. Considering all SDs identified in the 170 human genomes, we estimate that 1,340 protein-coding genes are duplicated to copy number four in at least one sample. Of note, 173 of these correspond to single-copy genes in the T2T-CHM13 reference (Methods; Supplementary Table 5). The majority of the low-frequency SDs are incomplete with respect to their ancestral gene model, often involving a subset of the original exons. We caution, however, that incomplete SDs do not guarantee that the duplicates are pseudogenes (Dennis et al. 2012; Florio et al. 2015; Suzuki et al. 2018).

Next, we considered multicopy SDs based on gene content, grouping 1,095 multicopy genes in the T2T-CHM13 reference into 314 gene families. As a control for potential assembly artifacts, we orthogonally evaluated gene copy by correlating (R-squared = 0.94) assembly gene copy by predicted copy number from Illumina short-read sequencing read depth (Supplementary Fig. 2). We applied the index of dispersion as a metric of the level of copy number variation for each gene family, which is computed simply as the variance divided by the mean copy number. We identified the 25 most variable gene families in the human genome and contrasted them with the 25 least variable (Fig. 5A-B and Supplementary Fig. 3). As expected, higher copy number gene families (10-50 members) were among the most variable in the human population while the most invariant typically were fixed at four or six copies (diploid copy number). The least-variable gene family, *HYDIN*, includes the human-specific duplication *HYDIN2*, which gained neural expression by adopting a new promoter (Dougherty et al. 2017). Similarly, the *RGPD3* family is

the eighth least variable based on its copy number and includes two human-specific copies that we find to be under selection (Mao et al. 2024). A gene ontology analysis showed that highly variable gene copy associated with female pregnancy (*PSG*), amylase activity (*AMY2*, *AMY1*) and immune response (defensin, *KIR2DL*) and unknown biological function while fixed copy number SD genes were particularly common among Kruppel-associated box (KRAB) zinc-finger proteins (KRAB-ZFPs) and genes associated with metabolic process (*CYP1*, *CYP2* and *CYP4*) (Supplementary Fig. 4). However, even among high copy number genes, both variable and invariant members are observed.

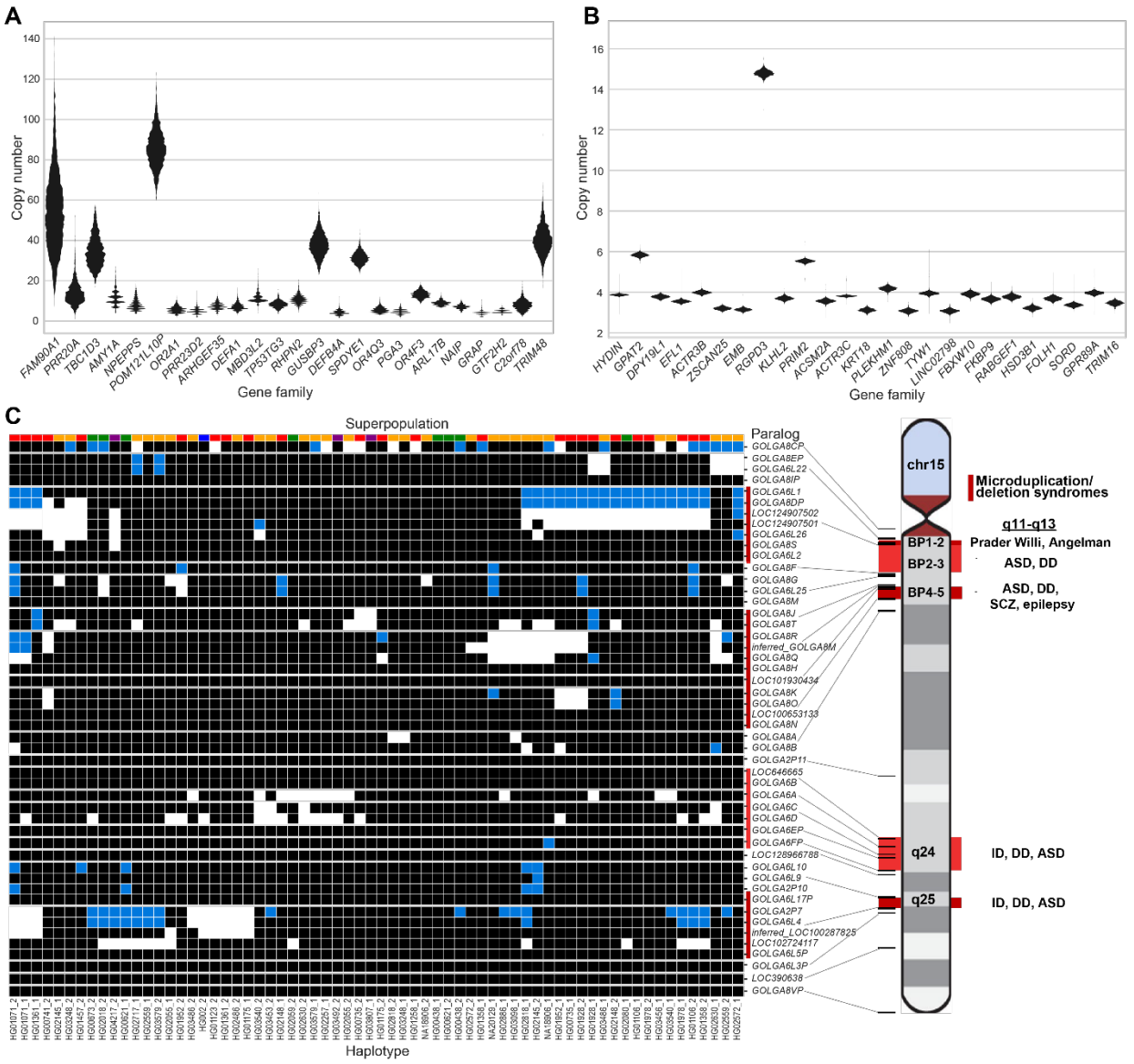


Figure 2.5. Variable copy number of duplicated genes.

(A-B) Gene families with highly variable (A) and nearly fixed (B) copy number are displayed. Gene families are selected and ordered by dispersion index, requiring an average diploid copy number greater than three. Read-depth copy number was estimated with fastCN, using Illumina reads for each sample and comparing to the T2T-CHM13 genome. **(C)** Estimated copy number of *GOLGA6/8* paralogs in each assembled haplotype, based on assembly alignments (white:0, black:1, blue:2). The continental population groups for each haplotype are indicated by color above each column (Africa: gold, East Asia: green, South Asia: purple,

Europe: blue, the Americas: red). ASD: autism spectrum disorder, DD: developmental delay, ID: intellectual disability, SCZ: schizophrenia.

Using these highly contiguous assemblies, we are now able to assign copy number polymorphism to specific paralogs in addition to assaying copy number on the level of gene families. For example, for one of the most variable gene families in the human genome (*GOLGA6/8*), we analyzed the copy number variation of each *GOLGA* paralog across our assembled haplotypes (Fig. 5C). To avoid false duplicates, we only included haplotypes with no assembly breaks within 30 kb of the *GOLGA* paralog. Of the named protein-coding paralogs, only *GOLGA6L2*, *GOLGA8M*, *GOLGA8H*, *GOLGA8N*, and *GOLGA6B* are fixed at a single copy across all haplotypes, while four are variable but never deleted, and the remaining 18 are deleted or absent in some haplotypes. This paralog specificity identifies those five single-copy genes as higher-priority candidates for functional analysis, given they are fixed in the human population. *GOLGA* paralogs mediate pathogenic microduplications and deletions at 15q11-q13, 15q24, and 15q25 causing forms of intellectual delay, including Prader-Willi syndrome (Amos-Landgraf et al. 1999; El-Hattab et al. 2009; Mefford et al. 2012; Antonacci et al. 2014; Paparella et al. 2023). While we observe multi-gene deletions in these regions (Fig. 5C), including genes such as the human-specific fusion gene *CHRFAM7A* whose deletion has been implicated in Alzheimer's disease pathology (Ihnatovych et al. 2024), none of these deletions extend beyond the SD into the unique critical regions for named syndromes in our samples.

During this gene analysis, we noticed that samples of African ancestry tended to show overall higher copy number for multicopy SDs. We tested this more formally in three ways. First, we compared the intra and interchromosomal content irrespective of gene content between genomes

of African or non-African origin. Genomes of African origin harbor significantly more intrachromosomal SDs ($p = 1.6 \times 10^{-6}$, Mann-Whitney U test) (Fig. 6A and Supplementary Fig. 5). Next, we examined the copy number of gene families by two methods: by counting assembled paralogs in our long-read assemblies and by using read depth to estimate copy number in a larger cohort with short reads (n = 2,196). In the long-read assemblies, we tested the 90 protein-coding gene families with variable copy number (dispersion index ≥ 0.1) and mean copy number greater than two for African or non-African samples. Seventeen gene families showed shifted copy number distribution, and 16/17 showed the same effect by read-depth analysis (Mann-Whitney U test, Benjamini-Hochberg corrected $p \leq 0.05$). Consistent with the increase in intrachromosomal SDs, for 13/16 gene families (81%), the copy number distribution is higher in African than non-African samples, as shown in Figure 6B (binomial test, $p = 0.01$). Finally, with the larger sample of high-coverage Illumina data from unrelated individuals in the 1KG, excluding highly admixed populations (n = 2,196), we considered the 1,171 gene families with dispersion index ≥0.1 and mean copy number greater than two in African or non-African samples (Supplementary Fig. 6). Population-differentiated copy numbers are observed in 263 gene families (Mann-Whitney U test, Benjamini-Hochberg-corrected $p < 0.05$), with 164/263 (62%) shifted towards higher copy number in African samples ($p = 0.00004$, binomial test). The gene families with largest shifts (greater than 15%) are shown in Figure 6C, with 17/22 (77%) shifted towards higher copy number in African samples. All statistically significant gene families are shown in Supplementary Fig. 7. From the assembly test of copy number differentiation, only *GUSBP3* did not replicate in the larger read-depth cohort.

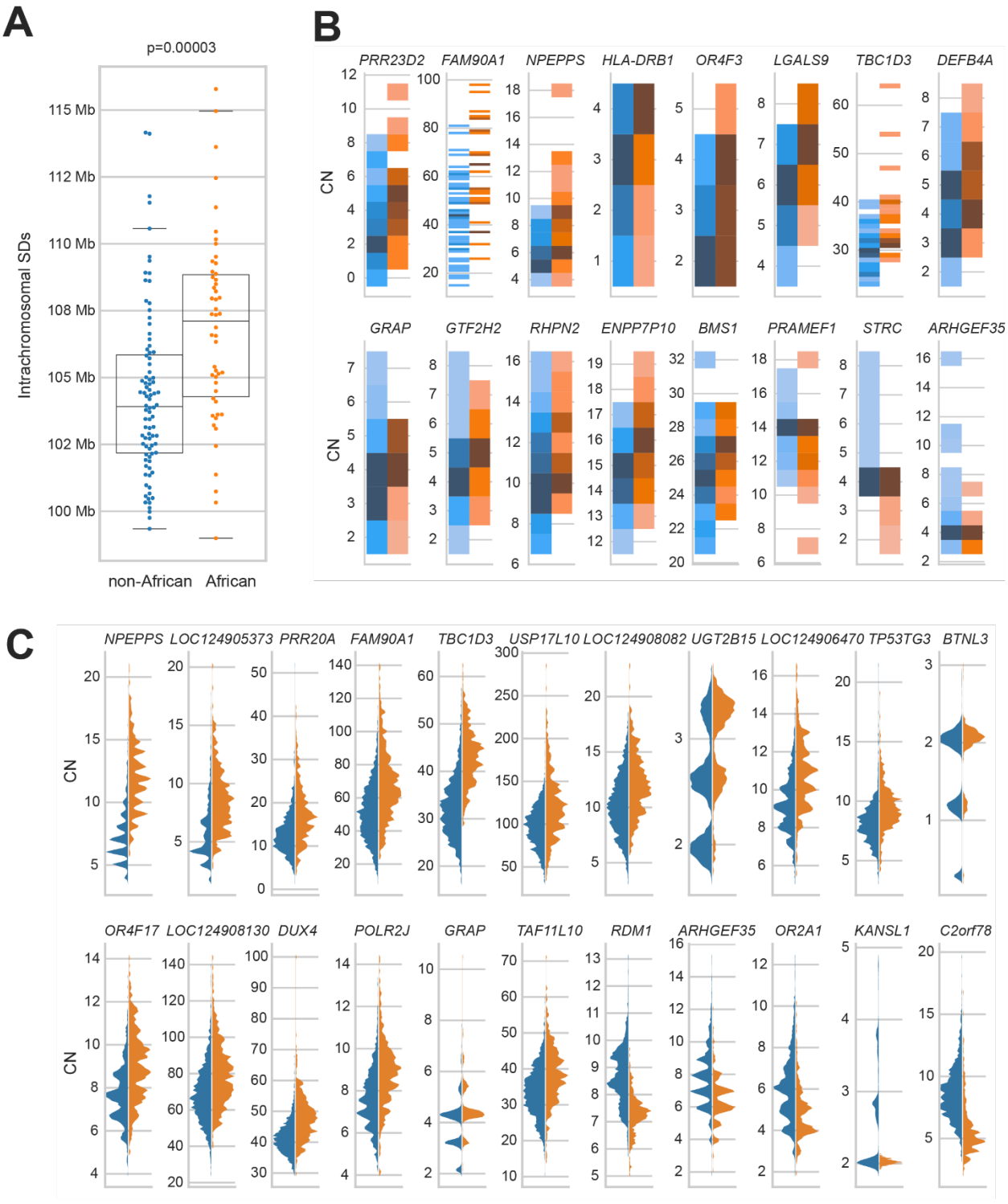


Figure 2.6. African vs. non-African SD copy number variation.

(A) Proportion of intrachromosomal SD content between African and non-African populations. African genomes have a higher SD content compared to non-African genomes, and

the difference is significant for intrachromosomal SDs. **(B)** Gene family copy number variation between populations. Gene families with significant copy number differences between African and non-African populations are shown (Mann-Whitney U test, Benjamini-Hochberg adjusted p-value <0.05), excluding *GUSPB3*, which did not replicate in the larger cohort. Gene copy number (CN) was estimated from the assemblies by whole-genome alignment; 13/16 gene families average higher copy number in individuals of African ancestry (binomial, p = 0.01). **(C)** Gene copy number evaluated by Illumina read depth. The 22 gene families with the largest distribution shift are shown.

2.3.4 *Genic potential of polymorphic SDs*

We sought to assess the transcriptional potential of the structurally polymorphic SDs identified in this study. Because of the high degree of sequence identity among the SDs, gene annotation has been difficult with standard RNA-seq datasets because short reads map equally well to distinct loci. This is especially true for copy number polymorphic genes where individual copies are >99% identical and can range in copy from 5-40 among different individuals in the population (Guitart et al. 2024). To address this limitation, we assembled a long-read Iso-Seq resource of 563 million full-length non-chimeric (FLNC) cDNA sequences generated from 241 libraries and 67 distinct tissues (Supplementary Table 6). We mapped each FLNC read both to the T2T-CHM13 genome and a pangenome of 170 human genomes searching specifically for FLNC reads that mapped better to the pangenome. Specifically, we required at least 99.9% sequence identity to an assembled haplotype and less than 99.7% gap-compressed identity to T2T-CHM13—below the expected allelic divergence for most protein-coding regions of the genome. We focused on putative protein-coding genes and constructed 7,081 gene models for cDNA alignments spanning 476 Mb of T2T-CHM13.

We used these reference-divergent cDNA reads that matched better to other assembled haplotypes (n = 1,279,037) to predict protein-coding genes (Fig. 7A). Each additional human haplotype contained an average of 46 protein-coding gene predictions (range 13–77) that showed more than 1% divergence from T2T-CHM13 reference annotations, highlighting the importance of additional human genome references to fully assess human genic variation. To count novel gene annotations across haplotypes, we grouped genes/transcripts into gene families by counting only predictions from the haplotype with greatest number of novel paralogs. This resulted in a total count of 260 putative novel protein-coding genes from 206 gene families. Of these 260 genes, 183 mapped to SD regions, 18 genes mapped to SD regions for at least one sample but not the T2T-CHM13 reference, and the remaining 59 genes mapped to unique sequence (not SDs). Gene ontology biological process enrichment analysis of these genes compared to the background of protein-coding genes within the 476 Mb of sequence examined yielded 13 significantly enriched driver terms, largely related to immunity: positive regulation of leukocyte mediated immunity, antigen processing and presentation of endogenous peptide antigen, rRNA metabolic process, symbiont entry into host, T cell extravasation, regulation of deoxyribonuclease activity, leukocyte cell-cell adhesion, regulation of type II interferon production, regulation of lymphocyte activation, dendritic cell differentiation, detection of bacterium, peptide antigen assembly with MHC class II protein complex, and positive regulation of cell-cell adhesion (Benjamini-Hochberg adjusted $p < 0.05$). Twenty of the novel genes belong to the immunoglobulin superfamily (Smith and Xue 1997) while ten are within core duplicons, a group of loci hypothesized to drive the evolution of interspersed larger SD blocks (Jiang et al. 2007) during ape evolution. Only 10.8% (28/260) of these predicted protein sequences had previously been submitted to GenBank.

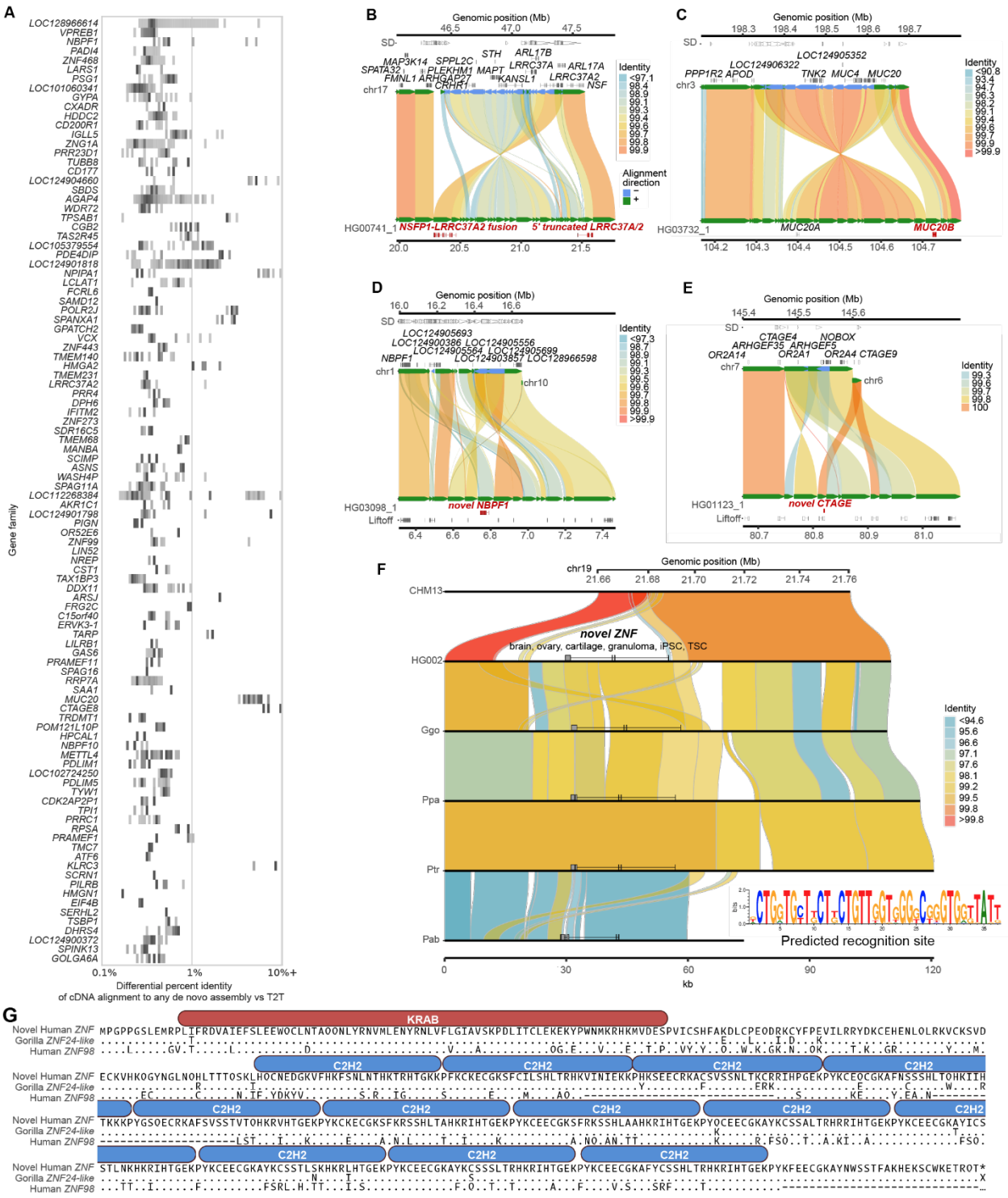


Figure 2.7. Discovery of novel gene/transcripts in rare and polymorphic SD regions.

(A) 2D histogram display of copy number polymorphic gene families where FLNC generated from Iso-Seq map better to the pangenome than to the T2T-CHM13 human genome reference.

(B-E) Selected haplotypes containing novel gene predictions for *MUC20*, *NBPF1*, *CTAGE*, and *LRRC37A* compared to T2T-CHM13 reference where there is FLNC transcript support. Alignment color indicates percent identity**. (F)** Comparison of T2T-CHM13 (top) and HG002 maternal haplotype (bottom) depicts 48 kb polymorphic SD region present in 66/170 haplotypes. Nonhuman apes all carry a copy of the duplicated sequence. ZNF predicted recognition site shown (inset). **(G)** Comparison of the novel ZNF to its best human match (ZNF98, 68% identity), and the most similar existing primate annotation (low-quality protein ZNF724-like in gorilla, 95% identity). ProSite-predicted KRAB-ZFP is shown above the sequence.

Notable examples of novel gene annotations include additional copies of *MUC20*, *GSTM*, *TUBB8*, *SIRPB1*, *GOLGA8*, *LRRC37A*, *NBPF1*, *CTAGE*, and *UPK3BL1* (Fig. 7B-E) (Seidegård et al. 1988; Miller et al. 1989; Pallaoro et al. 1999). The paralogs with the lowest identity cDNA sequence compared to T2T-CHM13 often have modified isoform structures, predominately modified N- or C-termini due to the structural rearrangements that led to their formation. For example, the 17q21.31 *KANSL1* inversion haplotype H2 includes a partial *KANSL1* duplication, which acts as an alternate 5' promoter for *LRRC37A/2*, producing a putatively protein-coding transcript with 39 novel amino acids at its N-terminus followed by sites 870 to 1700 of the canonical *LRRC37A/2* isoform (Fig. 7B). The same haplotype also encodes a *NSFP1-LRRC37A2* fusion transcript, which maintains an open reading frame, predicted to produce a protein with the first 492 amino acids of *NSF* followed by amino acids 41-903 belonging to the core duplicon gene *LRRC37A2* (Supplementary Fig. 8).

We discovered a novel expressed copy of *MUC20* in the paternal haplotype of HG03732 within a 214 kb inversion at chr3q29 that duplicates 73 kb and 37 kb on its edges (Fig. 7C); its expression is supported by full-length cDNA from 14 Iso-Seq libraries from chondrocytes, soft

tissue, left colon, induced pluripotent stem cells (iPSCs), human embryonic stem cells, and fibroblast cell lines. Structural diversity at 3q29 has been documented in prior work (Smith and Xue 1997) but the expression and gene model of the additional *MUC20* paralog has not yet been reported to our knowledge. Two assembled haplotypes have a complex rearrangement at chr1p36.13 that creates 86 kb of additional sequence compared to T2T-CHM13, including an additional copy of *NBPF1* supported by two Iso-Seq reads from brain tissue and a mammary epithelial cell line (Fig. 7D). In the paternal haplotype of HG01123, an additional copy of *CTAGE* is created by a 16 kb insertion from chr6q23.2 into chr7p35 in the context of a 59 kb duplicated inversion, with expression detected with a single read from each of four Iso-Seq libraries from promyeloblast cells and iPSCs (Fig. 7E). Even among *HLA* genes, whose polymorphisms have been extensively documented due to their clinical significance, we identified 62 novel alleles across seven distinct genes not currently represented in GenBank or the IPD-IMGT/HLA database (*HLA-A, -B, -C, -G, -DQB1, -DRB1, -DRB5*).

During this analysis, we identified low-identity alignments to *ZNF724* corresponding to a novel KRAB-ZFP absent from the T2T-CHM13 reference. This duplicate gene, provisionally named *ZNF972*, has only 69% identity to its best-matching annotated human gene, *ZNF98* (Fig. 7). *ZNF972* cDNA reads were found in 18 (6%) Iso-Seq libraries and correspond to a 48 kb region within the chr19p12 ZNF cluster, not present in previous human reference genome assemblies (GRCh38 and T2T-CHM13), though it exists in 35.9% of the assembled human haplotypes we analyzed. This region is also present in *Pan*, gorilla, and *Pongo* genomes, but an orthologous gene has only been annotated in gorilla as *ZNF972* coding sequence. Its open reading frame is disrupted in *Pan* and *Pongo* relative to gorilla and human (Fig. 7F-G). Thus, ZNF972 is an

example of an ancestral ape duplicated gene that is still present in gorilla and is present in a subset of humans but likely pseudogenized in other ape lineages.

2.4 DISCUSSION

The last two decades of human genomics research have shown that SDs play an important role in human health and evolution, contributing to genetic diversity, adaptation, genomic instability, and susceptibility to disease (Sharp et al. 2006; Zody et al. 2008; Lupski 2010; Dennis et al. 2012; Boettger et al. 2012; Steinberg et al. 2012; Florio et al. 2015; Suzuki et al. 2018; Porubsky et al. 2022). Despite their importance, understanding how humans vary with respect to this structural feature of our genomes and its potential functional consequence has always been challenging in large part because the size and high sequence identity of SD repeats have made interrogation of these regions and ~1000 protein-coding genes mapping within almost impossible with traditional sequencing and genotyping approaches. As a result, most genome-wide association studies as well as genome-wide surveys of selection, gene regulation (ENCODE), and transcription (GTEx) have explicitly excluded the most identical SDs from study (ENCODE Project Consortium 2012; GTEx Consortium 2013). Even early long-read sequencing-based approaches failed to adequately resolve these particular regions (Ebert et al. 2021; Porubsky et al. 2022). The advent of PacBio HiFi sequencing data (Wenger et al. 2019) along with improved assembly algorithms has fundamentally changed the calculus (Vollger et al. 2020; Cheng et al. 2021; Rautiainen et al. 2023). The sequence accuracy of HiFi data (>99.9%) meant that paralogs and alleles could be fully resolved in a phased genome assembly making these regions systematically accessible for the first time (Vollger et al. 2020; Jarvis et al. 2022; Liao et al. 2023). In this study, we took advantage of HiFi data generated as part of the HGSVC and HPRC

to analyze SDs in a total of 170 genome assemblies and compare the results to a complete human reference genome (T2T-CHM13). We harmonized the data using the same assembly algorithm and validated copy number in individual genomes using Illumina whole-genome sequence data to reveal the location and structure of copy number of SD variation at a population level for the first time. Thus, this pangenome representation provides one of the first glimpses of human structural diversity of SDs genome-wide.

While SDs have been known to be enriched in copy number polymorphisms (Sharp et al. 2006; Sudmant et al. 2010, 2015b), the phased genome assemblies allow us to quantify, map, and compare this variation revealing some unexpected findings. Our analysis of 170 genomes identifies 76.4 Mb of variable (60.3 Mb intra and 16.1 Mb interchromosomal) versus 147.5 Mb of invariant (89.7 Mb intra and 57.8 Mb interchromosomal) SD DNA—the latter may be more likely to harbor genes that will be functionally constrained (Dennis et al. 2012). Although fundamentally different in nature, the number of variable nucleotides in this 6% of the genome in these 85 individuals is comparable to the estimated 84.7 million single-nucleotide polymorphisms discovered genome-wide from sequencing the 2,500 individuals from the 1KG (1000 Genomes Project Consortium et al. 2015). We find that intrachromosomal SDs are twice as likely to be polymorphic when compared to interchromosomal SDs, although we should caution that we excluded the acrocentric regions of human short arms from this analysis where we anticipate rampant ectopic recombination and interchromosomal copy number variation to occur (Guarracino et al. 2023). While most of the 41.4 Mb of novel SDs (with respect to the finished T2T-CHM13 genome) occur in close proximity to existing regions of SD, we discovered novel sites of interspersed duplications. Such interspersed rare SDs are more likely to

be configured in an inverted orientation minimizing predisposition to large-scale microdeletions although potentially promoting rare inversion polymorphisms in the population (Porubsky et al. 2023). Also, we note that certain chromosome arms, including chromosomes 1q, 8p, 10p, 15q, 16p, 17p, 19p, and 22q, appear enriched for novel SDs—the basis for this chromosomal bias is unknown.

From a population genetics perspective, it is noteworthy that samples of African ancestry show significantly greater intrachromosomal SD content when compared to 1KG populations belonging to other continental groups. This translates into an overall higher gene copy number for duplicated genes (Fig. 6)—an observation we confirmed both by genome assembly as well as Illumina read-depth analyses (Methods). While increased variance in copy number would be consistent with the overall 15-20% increase in genetic diversity and greater population substructure that has been reported for populations of African ancestry (Campbell and Tishkoff 2008), there are other explanations. Overall higher copy number for duplicated gene families, especially those related to environmental interaction (e.g., drug detoxification, immunity), may have provided ancestral human populations with increased genetic diversity in terms of duplicated genes allowing for selection to operate on different copies to evolve new or modified functions and, therefore, increased fitness. Higher copy number, however, would also lead to greater susceptibility to NAHR-mediated rearrangements with potential negative consequences. Alternatively, genetic drift in ancestral populations may have introduced copy number differences and, if the ancestral African populations had higher copy number, mutational biases such as NAHR may have promoted subsequent increases in copy number. It is interesting that read-depth sequencing analysis of some archaic hominins such as Denisova have suggested

overall higher copy number for many gene families when compared to modern humans (Sudmant et al. 2015a).

There are several limitations of the current study. First, we sampled only 85 individuals (170 human genomes) and this represents only a small proportion of potential human genetic diversity. As more human genomes are sequenced and pushed toward T2T status (Miga and Eichler 2023), a more complete picture of human genetic diversity will begin to emerge. This will include population-specific paralogs and insights into the mechanisms underlying the formation of interspersed SDs as well as the role of SDs in driving ectopic recombination of acrocentric short arms in the human population (Hsieh et al. 2019). Similarly, our attempt to identify novel genes using a deep resource of human Iso-Seq data should be regarded only as a starting point. The challenge especially for assessing rarer SDs that harbor duplicated genes is that full-length cDNA was derived from different individuals from those whose genomes were sequenced and assembled (Dougherty et al. 2018; Liao et al. 2023). Genomic resources, such as those being generated from the SMaHT (Somatic Mosaicism across Human Tissues) initiative where donor-specific assemblies and Iso-Seq data from different human tissues from the same source are gathered, will be required (Johnson and Voight 2018; Souilmi et al. 2022). Such matched transcription and assembly data from the same donor will provide a clearer picture of the transcription as well as the tissue specificity of the thousand genes mapping to human duplicated sequence. Ultimately, functional characterization will be required to confirm the missing protein-coding copy number polymorphic genes in our genome.

2.5 METHODS

2.5.1 *PacBio HiFi sequence production*

University of Washington: Isolated DNA was sheared using the Megaruptor 3 instrument (Diagenode) twice using settings 31 and 32 to achieve a peak size of ~15–20 kb. The sheared material was processed for SMRTbell library preparation using the Express Template Prep Kit v2 and SMRTbell Cleanup Kit v2 (PacBio). After checking for size and quantity, the libraries were size-selected on the Pippin HT instrument (Sage Science) using the protocol '0.75% agarose, 15–20 kb high pass' and a cut-off of 14–15 kb. Size-selected libraries were checked by fluorometric quantitation (Qubit) and pulse-field sizing (FEMTO Pulse). All cells were sequenced on the Sequel II instrument (PacBio) with 30 h video times using version 2.0 sequencing chemistry and 2 h pre-extension. HiFi/CCS analysis was performed using SMRT Link (v.10.1) using an estimated read-quality value of 0.99.

The Jackson Laboratory: High-molecular-mass DNA was extracted from 30 million frozen pelleted cells using the Gentra Puregene extraction kit (Qiagen). Purified gDNA was assessed using fluorometric (Qubit, Thermo Fisher Scientific) assays for quantity and FEMTO Pulse (Agilent) for quality. For HiFi sequencing, samples exhibiting a mode size above 50 kb were considered to be good candidates. Libraries were prepared using the SMRTbell Express Template Prep Kit 2.0 (PacBio). In brief, 12 µl of DNA was first sheared using gTUBEs (Covaris) to target 15–18 kb fragments. Two 5 µg of sheared DNA were used for each prep. DNA was treated to remove single-stranded overhangs, followed by DNA damage repair and end repair/A-tailing. The DNA was then ligated with a V3 adapter and purified using Ampure beads. The adapter ligated library was treated with Enzyme mix 2.0 for nuclease treatment to remove

damaged or non-intact SMRTbell templates, followed by size selection using Pippin HT (Sage Science) generating a library with a size >10 kb. The size-selected and purified >10 kb fraction of libraries was used for sequencing on the Sequel II (PacBio) system.

We would like to note that ONT data from matched samples generated as part of the HGSVC are available but were generated using ONT R9 flow cells while more recent data from the HPRC and HGSVC are being generated from R10 flow cells. The ONT R9 flow cell generates sequencing reads with an error rate of 2-3% even with the most accurate base-calling model. The high error rate of ONT reads was a major concern for this particular analysis because we wanted to fully characterize highly identical duplicated regions. A hybrid approach using both HiFi and ONT sequencing could increase the continuity of the assembly; however, for the purposes of this study, the HiFi-only-based assembly approach provides sufficient assembly continuity (average contig N50 of 49.59 Mb) and accuracy (QV >50) allowing data to be harmonized between HPRC and HGSVC samples.

2.5.2 *Genome assembly and SD annotation.*

We initially considered a diverse set of 106 human samples (212 haplotype assemblies), all of which originated from the 1KG and for which sufficient HiFi sequence data had been generated as part of previous efforts (Ebert et al. 2021; Liao et al. 2023). This included 47 HPRC (all trio binning assemblies using parental short reads) and 53 HGSVC (14 trio and 39 non-trio) samples. We sequenced and assembled all genomes using the same assembly algorithm, hifiasm (v0.14/v0.16) (Cheng et al. 2021), which had been shown previously to accurately resolve most (although not all) SD regions (Porubsky et al. 2023; Liao et al. 2023). We predicted collapses

and misassemblies by assessing the read-depth of HiFi reads realigned back onto the assemblies using NucFreq implemented in the assembly_eval pipeline (https://github.com/EichlerLab/assembly_eval; Supplementary Table 7). Because of the potential for assembly collapse, we further restricted our analysis to 1KG samples where matched Illumina short-read sequence data were available, and the genomes passed QC and were all assembled with the same algorithm. SDs are particularly prone to assembly errors or collapses and this procedure both harmonized the results and allowed for all duplicated sequence to be validated by Illumina read-depth analysis. We limited SD analyses to the autosomes because of ploidy differences between males and females and the challenges associated with Y chromosome and pseudoautosomal region (PAR) in phased assembly. Because of the difficulties in mapping acrocentric SDs to specific chromosomes, we also excluded sequence mapping to the short arms of chromosomes 13, 14, 15, 21 and 22. Analysis of these will require T2T genome assemblies. The autosomal contigs are scaffolded using RagTag (v2.1) (Alonge et al. 2022). We masked repeat content using RepeatMasker (v4.1) and called SDs using SEDEF (Numanagic et al. 2018). To call SDs, we followed the operational definition of SDs (>90% and >1 kb) from Bailey et al., 2001 (Bailey et al. 2001). Under neutral evolution, 90% sequence identity allows us to identify SDs that occurred ~35-40 million years ago and a length threshold >1 kb excludes the effective insertion length of most retrotransposons other than some full-length elements. We matched Illumina short-read sequencing data for all 170 haplotypes, which were used for additional read-depth support of the putative duplicated regions (fastCN) (Pendleton et al. 2018).

2.5.3 *Variant calling.*

We used PAV, an assembly-based phased assembly variant caller, to call variants for 164 genome assemblies, 58 of which were phased into paternal and maternal haplotypes using parental Illumina short-read data (https://github.com/EichlerLab/pav; v.2.1.0). The regions that align 1-to-1 via minimap2 (Li 2018) "-x asm20 --secondary=no -s 25000 -K 8G", showing no variants, were assigned as the reference, 0|0 genotype, while the regions outside of the alignment blocks were considered as missing genotypes when merging the variant calls of individual samples. This was done by via BCFtools merge --missing-to-ref followed by BCFtools view with the aligned regions (v.1.9) (Li 2011). In addition, in order to focus on confident variant callset, we additionally defined a 1-to-1 alignment block, which is syntenic with length >1 Mb, in at least 80% of the samples. The population statistics were calculated across this 1-to-1 syntenic, shared alignment blocks of length 2.605 Gb (90%), across the autosomes.

2.5.4 *Iso-Seq and transcript analyses.*

We used long-read RNA-seq (PacBio Iso-Seq) data to look for evidence of expression of newly discovered low-frequency gene duplications. Examining regions where 10 or fewer haplotypes have a duplication relative to the remainder of the samples, we align 563 million FLNC reads from 241 libraries to *de novo* assemblies and T2T-CHM13 v2.0 as a reference. Only alignments with >99.9% identity to the novel duplication and <99.7% identity to T2T-CHM13 were considered. To generate gene models for each *de novo* assembly, we transferred GENCODE v44 gene models with Liftoff (v1.6.3) (Shumate and Salzberg 2021), classified Iso-Seq reads compared to the Liftoff gene models with PacBio Pigeon and SQANTI3 (v5.2) (Pardo-Palacios et al. 2024), and predicted open reading frame sequences with GeneMark (Besemer and

Borodovsky 2005). Each coding gene prediction was compared to the NCBI nonredundant protein database (nr) with BLAST (v.2.15) (Altschul et al. 1990). We limited our Iso-Seq analysis to transcripts aligning to reference SDs (227.4 Mb), their boundaries (defined as 10% of the SD block size on each edge, 28.1 Mb), SDs seen in at least one assembled haplotype but not the reference (17.9 Mb), and regions corresponding to highly divergent loci in the *de novo* assemblies (unaligned to T2T-CHM13 with the asm20 preset of minimap2 but forced to align with -r2k,200k -N50 parameters, 202.7 Mb), totaling 476.1 Mb of theT2T-CHM13 genome. Gene ontology was performed with g:Profiler (database e111_eg58_p18_30541362) (Kolberg et al. 2023).

2.5.5 *Copy number estimation.*

Assembly-based methods. To estimate copy number, we mapped protein-coding genes (genic sequences from T2T-CHM13) overlapping with SDs to each haplotype assembly using minimap2 (>60% coverage and >90% identity) (Li 2018). Single exons or short genes (coding sequence < 200 bp) were excluded. If genes are composed with high repeat content, copy number can be overestimated due to partial mapping of repeat content. To remove this incorrect mapping, we removed alignments that completely matched the repeat sequence. To increase contiguity of the alignment and to estimate counts of high copy number genes, we customized the minimap2 options as follows: `minimap2 -cx map-ont -f 5000 -k15 -w10 -p 0.05 -N 200 -m200 -s200 -z10000 --secondary=yes --eqx `. To avoid any bias due to switch errors, we estimated gene copy number in diploid genomes. We also excluded duplicate genes found only in one haplotype.

Illumina fastCN. Read-depth copy number was estimated with fastCN (Pendleton et al. 2018), using Illumina reads for each sample and comparing to the T2T-CHM13 genome. To remove signal from variable number tandem repeats (VNTRs), we excluded fastCN windows that overlapped TRF (Benson 1999) or WindowMasker (Morgulis et al. 2006) calls by more than 10%; to estimate gene copy number, we only considered windows contained within the bounds of each annotated gene, taking the median value. To check for biased copy number estimates, we decomposed the T2T-CHM13 genome into 36-mers and estimated copy number with the fastCN pipeline, simulating fastCN results for a perfectly matched sample and reference. By default, fastCN overcorrected read depth based on GC content for these unbiased artificial reads, due to overrepresentation of extreme GC values in the human genome itself. We recalibrated the GC correction, window-by-window, based on these results. Read-depth-based methods also underestimate copy number to a variable extent based on sequence divergence between paralogs, due to unaligned sequence. To correct for this, we calculated an adjustment factor for each gene to match fastCN results to alignment-based assembly copy number with T2T-CHM13 as ground truth (Supplementary Fig. 9). Genes that required more than a 50% adjustment to match copy number between methods were excluded from further analysis (n = 82).

Please note that the correlation of determination between Illumina fastCN and assembly copy number differs slightly from that previously reported (Vollger, Dishuck, et al. 2023) (R^2 = 0.94 vs. 0.994). This is because Vollger et al. restricted the analysis to 19 large genes and compared Illumina fastCN estimates to k-merized assembly fastCN estimates, not quantifying gene copy number in the assemblies directly. In this analysis, we directly quantified the copy number for all SD gene families (n = 314) in the assembled autosomes, excluding p-arms of acrocentric

chromosomes and sex chromosomes. Short-read copy number estimates are noisier for the shorter and higher copy number loci, and Illumina fastCN estimates have some residual error from repetitive elements within the genes despite our best efforts to exclude such regions. Vollger et al. were able to mitigate this issue because they compared k-merized assemblies instead of direct gene copy number estimates.

2.6 DATA AVAILABILITY

The raw sequencing data generated in this study are available under project ID PRJEB58376 (https://www.ebi.ac.uk/ena/browser/view/PRJEB58376) and the HPRC year 1 PacBio HiFi data is available under PRJNA730823 (https://ncbi.nlm.nih.gov/bioproject/PRJNA730823) or https://github.com/human-pangenomics/HPP_Year1_Assemblies. The raw genome sequencing data generated from this study are available online (https://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/HGSVC3/).

2.7 COMPETING INTERESTS

E.E.E. is a scientific advisory board (SAB) member of Variant Bio, Inc. C.L. is an SAB member of Nabsys and Genome Insight. The other authors declare no competing interests.

2.8 ACKNOWLEDGMENTS

We thank T. Brown for assistance with manuscript editing and preparation. This research was supported, in part, by funding from the National Institutes of Health (NIH) grants

R01 HG002385 and R01 HG010169 (to E.E.E.) and U24 HG007497 (to E.E.E. and C.L.). E.E.E. is an investigator of the Howard Hughes Medical Institute.

CHAPTER 3. GAVISUNK: GENOME ASSEMBLY VALIDATION VIA INTER-SUNK DISTANCES IN OXFORD NANOPORE READS

Chapter 3 has been published:

Philip C. Dishuck, Allison N. Rozanski, Glennis A. Logsdon, David Porubsky, and Evan E. Eichler. *Bioinformatics* 39 (1) https://doi.org/10.1093/bioinformatics/btac714

3.1 ABSTRACT

Motivation: Highly contiguous *de novo* phased diploid genome assemblies are now feasible for large numbers of species and individuals. Methods are needed to validate assembly accuracy and detect misassemblies with orthologous sequencing data to allow for confident downstream analyses.

Results: We developed GAVISUNK, an open-source pipeline that detects misassemblies and produces a set of reliable regions genome-wide by assessing concordance of distances between unique *k*-mers in Pacific Biosciences high-fidelity (HiFi) assemblies and raw Oxford Nanopore Technologies reads.

Availability: GAVISUNK is available at https://github.com/pdishuck/GAVISUNK.

3.2 INTRODUCTION

Highly accurate and contiguous phased diploid *de novo* assemblies of long-read sequencing data have made reference-grade assemblies feasible for many species and individuals (Rhie et al. 2021; Ebert et al. 2021). Pacific Biosciences high-fidelity (HiFi) sequencing in particular has facilitated some of the first complete assembly of centromeres, acrocentric regions, as well as other complex SDs (Miga et al. 2020; Logsdon et al. 2021; Nurk et al. 2022). These assemblies make possible comprehensive, whole-genome evaluations of sequence variation, including some of the most difficult-to-assemble regions, unbiased by reference alignments for the first time. Phased genome assemblies, however, are still subject to the collapse of repetitive sequences, incorrect orientations, and misassemblies. Thus, any discoveries based on these automated shotgun sequence assemblies raise the question of whether the assembled sequence is, in fact, valid.

Here, we present genome assembly validation via inter-singly unique nucleotide *k*-mer (SUNK) distances in Oxford Nanopore Technologies (ONT) reads, known hereafter as GAVISUNK. GAVISUNK is a method of validating phased diploid HiFi-driven assemblies with orthogonal ONT sequence. It specifically assesses the contiguity of regions, flagging potential haplotype switches or misassemblies. Although the ONT platform has a significantly higher error rate than that of HiFi (Logsdon et al. 2020), such reads are typically much longer, making it a powerful orthogonal approach for assessing both contiguity and read depth across regions of interest. Whereas previous genome blacklists or masks of inaccessible regions, such as those used by the ENCODE Consortium, are determined based on annotation of a reference genome (Amemiya et al. 2019), GAVISUNK may be applied to any region or genome assembly to identify misassemblies and potential collapses and is, thus, particularly valuable for validating the integrity of regions with large and highly identical repeats that are more prone to assembly error. This method can be applied genome-wide or at fine scale to closely examine regions of interest across multiple haplotype assemblies.

3.3 METHODS

This assembly validation method relies on identifying SUNKs, *k*-mers that occur just a single time within the HiFi-based assembly (Sudmant et al. 2010) and confirming these SUNKs within long ONT sequencing reads. Because of the relatively lower accuracy of ONT data, false SUNK overlaps may occur at an appreciable frequency between paralogous regions of the genome or haplotypes. Therefore, we leverage not only the presence or absence of SUNKs, but also the intervening distance between pairs of SUNKs, referred to hereafter as inter-SUNK distances. The approach then compares the expected inter-SUNK distance in the assembly and observed distance within ONT reads to recruit reads to their corresponding genomic location. The failure

of ONT reads to span between SUNKs in the assembly defines a misjoin, while an excess of reads flags a potential collapse.

We apply Jellyfish (Marçais and Kingsford 2011) to identify unique *k*-mers based on all HiFi contigs within an assembly, generating a set of SUNKs for validation. ONT reads are haplotype phased using parental Illumina whole-genome sequencing data via Canu (Koren et al. 2017), or, in the absence of parental data, with Hi-C and HapCUT2 (Edge *et al.*, 2016). The position of all SUNKs within each ONT read are identified and used for downstream analysis. Each ONT read is assigned to its best-matching HiFi-assembled contig and orientation by comparing the locations of read SUNKs to assembly SUNKs within a diagonal band centered on the median SUNK location for that read. SUNKs observed in ONT reads at an implausibly high or low frequency (i.e., greater than four standard deviations above the mean or fewer than twice) for either haplotype are excluded from consideration.

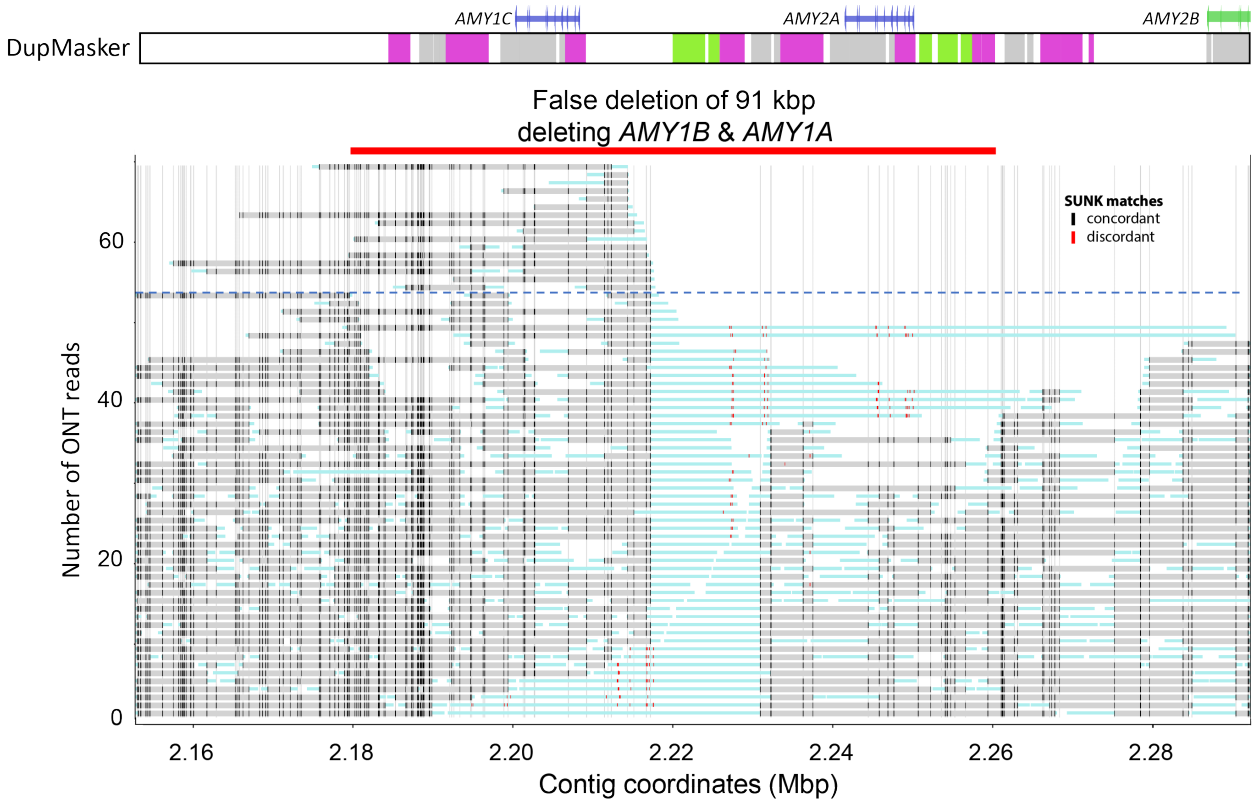


Figure 3.1. Example detected misassembly (HG02723 paternal haplotype) within the amylase duplication locus.

This assembly contains a false deletion of *AMY1B* and *AMY1A*, which is confirmed by the failure of ONT to anchor across the misjoin. Each ONT read is represented by a horizontal bar: the gray region contains valid inter-SUNK distances, while the remainder of the read is cyan. Concordant SUNK matches are black tick marks, while discordant matches are red. All possible assembly SUNKs are shown as dotted lines in the background, and a DupMasker track marks SDs above. The dotted blue line indicates the mean genome-wide coverage of ONT sequencing data.

To validate the assembly of each HiFi contig, all reads identified in the previous step are considered. For each read, a matrix of all pairwise inter-SUNK distances within the read is generated using NumPy and compared to expected distances from the assembly, allowing ±2% variation in length for a given distance by default (Harris et al. 2020). Only the largest set of SUNKs with fully inter-consistent inter-SUNK distances are retained from each read for subsequent validation. In this way, chimeric reads and other spuriously connected SUNKs are

separated. A graph is generated for each contig, with SUNKs as nodes and reads as edges, connecting pairs of SUNKs with consistent inter-SUNK distances, using the graph-tool library. This graph is decomposed into its connected components, each of which now corresponds to a validated region of the contig, identifying SUNKs spanned by ONT reads. These validated regions are sent as output to a BED file, and assembly SUNKs with no read support are listed as potentially artifactual.

3.4 Usage and Examples

GAVISUNK can run on consumer hardware, research clusters, or in cloud computing environments, as its steps are automated as part of a configurable Snakemake workflow (Mölder et al. 2021). To install GAVISUNK, clone https://github.com/pdishuck/GAVISUNK. Upon first run, dependencies will install automatically as a conda environment.

As described in the README, two configuration files must be edited to specify the assembly and *k*-mer size (config.yaml) and input nanopore read files (ont.tsv). To execute the pipeline, submit `snakemake –use-conda –cores {thread.count}`

By default, the pipeline produces a BED file of validated regions (hap#.validated.bed) and the complementary unconfirmed "gaps" between validated regions (hap#.gaps.bed) for each haplotype of the assembly. For the region of each validation gap, the pipeline produces a visualization to show SUNK-tagged read support, in PDF, SVG, and PNG formats (contig_start_end.format), as shown in Figure 1 and Supplementary Figure 1. Optionally, a BED file with regions of interest for visualization can be specified for each haplotype in ont.tsv.

To annotate the visualizations, BED files encoding regions of interest and color may be supplied. In the example shown in Figure 1, DupMasker annotations demarcate SDs across the locus (Jiang et al. 2008). Additionally, the sizes of unspanned inter-SUNK gaps are output for comparison with the distribution of inter-SUNK distances and expectation of spanning that distance given empirical ONT coverage at that read length (Supplementary Figure 2). Highly identical loci may have insufficient SUNK density for validation with this method, given current ONT read lengths (Supplementary Figure 3).

We constructed a pseudo-diploid genome assembly as a benchmark of false discovery rate for a well-curated reference assembly with matched ultra-long ONT data. We used the long-read assemblies of the CHM13 and CHM1 human cell lines, both of which are derived from complete hydatidiform moles exhibiting genome-wide uniparental disomy and are therefore guaranteed to represent a single haplotype. CHM13 was assembled by the T2T Consortium with a variety of methods and manual validation (Nurk et al. 2022), and CHM1 was assembled from HiFi data (Vollger et al., 2022, using hifiasm v0.12 (Cheng et al. 2021)), each with >30× coverage of ONT reads longer than 100 kbp. Results for this pseudo-diploid, along with true diploid HG02723 (hifiasm v0.14), are summarized in Supplementary Table 1**,** with detailed results for T2T-CHM13 validation gaps in Supplementary Table 2.

Other assembly validation methods TandemTools and VerityMap (Mikheenko et al. 2020) also used rare k-mers for validation but have distinct goals to GAVISUNK. Both are intended for use on extra-long tandem repeats and developed to use HiFi to validate haploid assemblies. GAVISUNK uses ultra-long ONT data genome-wide to validate diploid assemblies, focused on

long interspersed repeats such as SDs.

For CHM13, 1.0% of the assembled genome (103 gaps, 31.7 Mbp) is unsupported by ONT inter-SUNK distances, and 0.7% (274 gaps, 23.9 Mbp) for CHM1. For comparison, assuming random distribution of the empirical ONT read lengths compared to the distances between SUNKs for the CHM13 assembly, 24.9 Mbp is expected to be unsupported (92 gaps, Supplementary Figure 6). Extra-long tandem repeats are difficult to validate with this method, particularly the qh regions of chromosomes 1, 9, and 16, with 73 of 103 CHM13 validation gaps falling in the region. An optional "2pass" mode performs a second pass of analysis on putative validation gaps with more-permissive read recruitment and validates the higher-order repeats of 21/23 CHM13 centromeres (Table S3).

3.5 Conclusion

We developed GAVISUNK, a method for assembly validation using inter-SUNK distances in ONT reads. Applied to HiFi genome assemblies, this tool provides orthogonal validation of regions for downstream analysis, allowing for subsequent genome analyses and annotation. GAVISUNK provides easily interoperable BED outputs and interpretable visualizations of supported and unsupported regions of interest.

3.6 Data availability

ONT sequencing reads and their corresponding assemblies are available from the Telomere-to-Telomere and Human Pangenome Reference consortia at https://github.com/marbl/CHM13 and https://github.com/human-pangenomics/HPP_Year1_Data_Freeze_v1.0. ONT and HiFi sequencing of CHM1 are available on SRA as PRJNA869061 and PRJNA726974, and its

hifiasm assembly is available at https://zenodo.org/record/5502036.

3.7 ACKNOWLEDGEMENTS

We thank Mitchell Vollger and William Harvey for their helpful algorithmic, programming, and source control advice and Tonia Brown for proofreading.

3.8 FUNDING

This work was supported, in part, by US National Institutes of Health (NIH) grants HG002385 and HG010169 to E.E.E and 1F32GM134558 to G.A.L. E.E.E. is an investigator of the Howard Hughes Medical Institute. *Conflict of Interest:* E.E.E. is a scientific advisory board (SAB) member of Variant Bio, Inc.

CHAPTER 4. STRUCTURAL GENETIC DIVERSITY OF THE *NPIP* GENE FAMILY AND EVIDENCE OF SELECTIVE SWEEPS AND BRAIN-SPECIFIC EXPRESSION IN HUMANS

Philip C. Dishuck, Max L. Dougherty, Katherine M. Munson, Alexandra P. Lewis, Jason G. Underwood, William T. Harvey, PingHsun Hsieh, Tomi Pastinen, Evan E. Eichler

In preparation

4.1 ABSTRACT

The *NPIP* gene family has expanded to high copy number in humans and African apes where it has been subject to an excess of amino acid replacement consistent with positive selection (Johnson et al. 2001). Due to limitations of short-read sequencing, *NPIP* human genetic diversity has been poorly understood. Using highly accurate assemblies generated from long-read sequencing, we completely characterize 169 human haplotypes (4,752 *NPIP* paralogs and alleles). Of the 28 *NPIP* paralogs, just three (*NPIPB2*, *B11*, and *B14*) are fixed at a single copy, and only a single locus, *B2*, shows no structural variation. Four *NPIP* paralogs map to large segmental duplication blocks that mediate polymorphic inversions (355 kbp – 1.6 Mbp) as well as microdeletions associated with developmental delay and autism. Comparing haplotypes, we find evidence of ongoing gene duplications, deletions, and interlocus gene conversion. Haplotype-based tests of positive selection and selective sweeps identify two paralogs, *B9* and *B15*, within the top percentile for both tests. Using full-length cDNA data from 101 tissue/cell types, we explore differences in the structure and expression of the gene family. We define six distinct translation start sites and other protein structural features that distinguish paralogs, including a variable number tandem repeat (VNTR) that encodes a beta helix of variable size that emerged ~3.8 million years ago in human evolution. Among the 28 *NPIP* paralogs, we identify distinct tissue and developmental patterns of expression with only a few maintaining the ancestral testis-enriched expression. A subset of paralogs (*NPIPA1*, *A5*, *A6-9*, *B3-5* and *B12/B13*) show increased brain expression. Our results suggest ongoing positive selection of this gene family in the human population with potential neofunctionalization of *NPIP* subfamilies.

4.2 INTRODUCTION

NPIP (also known as *Morpheus*) is a gene family of unknown function that has undergone independent duplication in several primate lineages (Johnson et al. 2001; Cantsilieris et al. 2020). The gene family was first described based on the observation of a rapid expansion in African apes where the underlying genes show a significant excess of amino acid replacements consistent with the action of positive selection (Johnson et al. 2001). The ~20 kbp duplicon that contains *NPIP*, LCR16a, is interspersed across human chromosome 16 (Fig. 1A) (Loftus et al. 1999) with a solitary copy on human chromosome 18 and mediates recurrent duplications and deletions frequently associated with neurodevelopmental delay (Sharp et al. 2006; Ballif et al. 2007; Girirajan et al. 2010), including one of the most common genetic causes of autism spectrum disorder. Altogether, the segmental duplications (SDs) associated with LCR16a (Johnson et al. 2006) span more the 10% of the euchromatic portion of human chromosome 16p, having emerged and expanded since great ape divergence from the Old World monkeys (25 million years ago [mya]). The LCR16a encoding *NPIP* has been described as a "core duplicon" for its characteristic overabundance within these intrachromosomal duplications (Loftus et al. 1999; Jiang et al. 2007; Stallings et al. 2008). It has independently duplicated at least five times over the course of primate evolution leading each time to the formation of interspersed SDs where lineage-specific duplications accrue flanking the core LCR16a duplicon (Cantsilieris et al. 2020).

Because *NPIP* is frequently embedded in large blocks of SDs that share >97% sequence identity (Fig. 1), standard sequencing and assembly methods have limited our understanding of its genetic diversity and, consequently, our ability to make genetic associations or perform standard population genetic analyses. FISH analysis with LCR16a probes along with read-depth analyses

using short reads, however, have been used to estimate a range of 20-30 copies per human haplotype (Johnson et al. 2006; Cantsilieris et al. 2020). Targeted bacterial artificial chromosome (BAC) assemblies have partially resolved the *NPIP* loci in the most commonly used reference genomes. Because of the high sequence identity of the underlying duplicated segments, misassembly of the loci is a common problem. Even in one of the most recent references, GRCh38, there is evidence of at least two chimeric misassemblies created as a result of inadvertently assembling paralogous loci whose sequence identity approximates allelic variation. Not surprisingly, GRCh38 contains just 24 copies of LCR16a, compared to the median 25 copies estimated by read depth to be present in most human haplotypes (Sudmant et al. 2010).

Over the last few years, a series of resources and methods have been developed making it possible to systematically characterize human genetic variation and expression across these regions of chromosome 16 arguably for the first time. First, the T2T (Telomere-to-Telomere) Consortium recently completed the assembly of a single human haplotype, CHM13, by combining highly accurate long HiFi reads with ultra-long ONT reads (Nurk et al. 2022). As a result, all *NPIP* gene copies are fully resolved providing a reference for comparisons. Second, both the HPRC (Human Pangenome Reference Consortium) and HGSVC (Human Genome Structural Variation Consortium) using similar approaches have published and released contiguous phased assemblies of 87 unrelated individuals. The availability of both short-read sequencing (SRS) and long-read sequencing (LRS), including phasing information, enables the characterization and validation of entire chromosomal haplotypes for even the most identical gene families (Liao et al. 2023; Hallast et al. 2023; Guitart et al. 2024). Third, the recent advancements of multiple sequence alignment (MSA) and phylogenetic methods optimized for

comparing thousands of viral genomes (Katoh and Standley 2013; Nguyen et al. 2015) are useful for the evolutionary reconstruction of rapidly evolving 20 kbp segment of human DNA like LCR16a. We directly apply these methods to characterize thousands of *NPIP* paralogs and alleles to reconstruct the complex population genetic history underlying these regions of human chromosome 16, including the mutational forces that have shaped them. Finally, the recent release of 1.4 billion full-length cDNA from 384 Iso-Seq libraries from the Genomic Answer for Kids Study and ENCODE, among others (Zhang et al. 2014; Zook et al. 2016; Sun et al. 2021; Kim et al. 2022; Caballero et al. 2022; Miller et al. 2022; Abood et al. 2023; Reese et al. 2023; Cheung et al. 2023; Rybak-Wolf et al. 2023; Schertzer et al. 2023; D et al. 2023; Garza et al. 2023; Maeng et al. 2023; Shimada et al. 2024), makes it possible to assign transcript data to specific paralogs and alleles—a near impossibility with traditional short-read RNA-seq data. We use this data to accurately construct gene models, define transcription start sites, distinguish potential protein-coding genes from pseudogenes, and interrogate expression for specific *NPIP* copies.

4.3 RESULTS

4.3.1 *Human genetic diversity*

Using the complete sequence of the T2T human genome assembly, we first annotated LCR16a and its associated SDs using DupMasker for the T2T-CHM13 reference genome (Fig. 1a). The analysis reveals 27 *NPIP* genes—26 of which map to 12 duplication blocks on chromosome 16 with a solitary copy mapping to chromosome 18, as expected (Johnson et al, 2001). This is in stark contrast to the macaque genome (Fig. 1A) where only a single copy of LCR16a was identified. We assigned gene names based on best matches back to the annotation on GRCh38.

In order to estimate the copy number distribution in the human population, we mapped whole-genome shotgun sequence data (WGS) (Byrska-Bishop et al. 2022) from the 2,609 unrelated individuals from the 1000 Genomes Project (1KG) using read depth to estimate the diploid and haploid copy number across each superpopulation. We estimate the haploid copy number ranges from 21 to 33 copies with the highest copy number observed among individuals of African descent (Wilcoxon rank-sum test, p=0.000001, Fig. 1b).

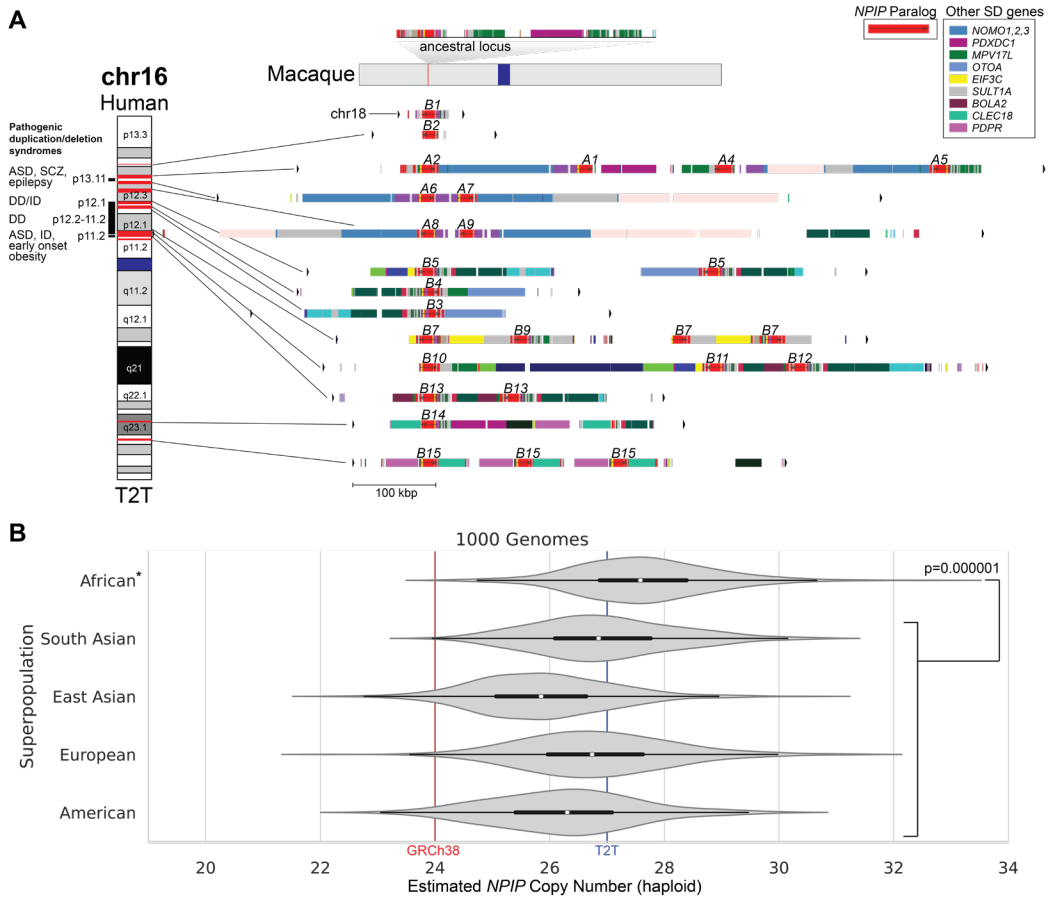


Figure 4.1. *NPIP* locus organization and copy number variation.

A) *NPIP* regional organization in the T2T-CHM13 genome. The single-copy sequence in the macaque genome (Mmul10) is compared to the duplicated sequence on human chromosomes 16 and 18. Red highlights on chromosome 16 (left) correspond to *NPIP* loci, with segmental duplication (SD) content annotated by DupMasker colors (bars). *NPIP* gene names (*A1-9, B1-15*) are labeled above DupMasker tracks, with selected *NPIP* SD-associated genes shown in the legend. To the left of the ideogram, pathogenic duplication/deletion syndromes associated with

NPIP SDs are shown as red horizontal bars on the ideogram. **B)** Read-depth estimates (fastCN) of modern human per-haplotype *NPIP* copy number from the 1000 Genomes Project (n=2,609), grouped by superpopulation. Short-read shotgun sequence from each individual is split into 36 bp segments and aligned to a reference genome, allowing up to two single-nucleotide mismatches. For this estimate, regions of *NPIP*-containing VNTRs are excluded.

To understand the variation in structure of *NPIP* loci across the human population, we collected previously assembled and released genomes from the HPRC (n=44) and HGSVC (n=38), along with a draft T2T assembly of HG002 (Rautiainen et al. 2023), two individuals from Papua New Guinea (Hsieh et al. 2019), the reference genome T2T-CHM13v2, and an additional haploid cell line CHM1 (Vollger et al. 2022; Dishuck et al. 2022), for a total of 169 haplotypes (Supplementary Table S2). We identified and extracted *NPIP* loci from each assembly by aligning the *NPIP* locus from GRCh38 to each haplotype with minimap2 and wfmash (Methods), for a total of 4,961 copies of *NPIP.* As *NPIP* loci are known to be structurally variable and subject to gene conversion, we did not rely on synteny alone to determine the paralog identity. Instead, we created an MSA and maximum likelihood phylogeny of the 4,961 *NPIP* loci from the 169 assembled haplotypes, using *Pongo pygmaeus* and *Siamang syndactylus* as outgroups (Fig. 2A). Clades with >75% branch support (SH-aLRT) were used to assign copies to one of 28 defined paralogs, named based on phylogenetic identity to T2T-CHM13 and GRCh38. In cases where a clade did not have an anchor in T2T-CHM13 or GRCh38, we defined it based on its nearest neighbor (i.e., *NPIPA1L*). In cases where there was insufficient genetic distance to distinguish paralogs, they were grouped into a single clade comprising the two copies (i.e., *NPIPB12/B13).* For expediency, we subsequently shorten gene names by dropping the *NPIP* prefix in this report. Additionally, to estimate the age of each branch, we created a timetree with Reltime (Koichiro Tamura, Tao, and Kumar 2018) (Methods), incorporating paralogs from

human assemblies CHM13, CHM1, GRCh38, HG002, and PNG15, along with nonhuman primate sequences from the primary *Pan troglodytes*, *Pan paniscus*, *Gorilla gorilla*, *Pongo pygmaeus*, *Pongo abelii*, and *Siamang syndactylus* haplotypes from the T2T Primate Project (Yoo et al. 2024) (Methods).

Even among LRS-assembled genomes, high sequence identity duplications that are hundreds of kbp in length remain a common source of misassembly and collapse (Porubsky et al, 2023). We, therefore, validated the integrity of assembled haplotypes using tools designed to detect misassemblies (i.e., NucFreq, Flagger, and GAVISUNK (Vollger et al. 2019; Dishuck et al. 2022; Liao et al. 2023)). For a haplotype structure to be classified as correctly assembled, we required contiguous assembly without collapse across all duplicated segments (not just *NPIP*) (Fig. 1a), including at least 30 kbp of unique sequence. The assembly validation rate varied from 52-85% (Fig. 2A, right). The copy number of *NPIP* paralogs varies widely across assembled haplotypes (Fig. 2B). Copy number heterozygosity for the paralogs, defined as the frequency of discordant copy numbers between the two haplotypes of a sample, ranges from 0 for *B2*, *B11*, and *B14*, with all individuals having just one copy, to 0.74 for *A6/A9*. While only *B2*, *B11*, and *B14* are fixed in copy number, *A2*, *A4*, *B12/B13*, and *B15* always have at least one copy in all assemblies that pass QC. Individual members of *NPIP* subfamilies *B3-B5* and *B6-9* are not always present when considered individually, yet at least one paralog from each of these larger subfamilies is always present in a given haplotype.

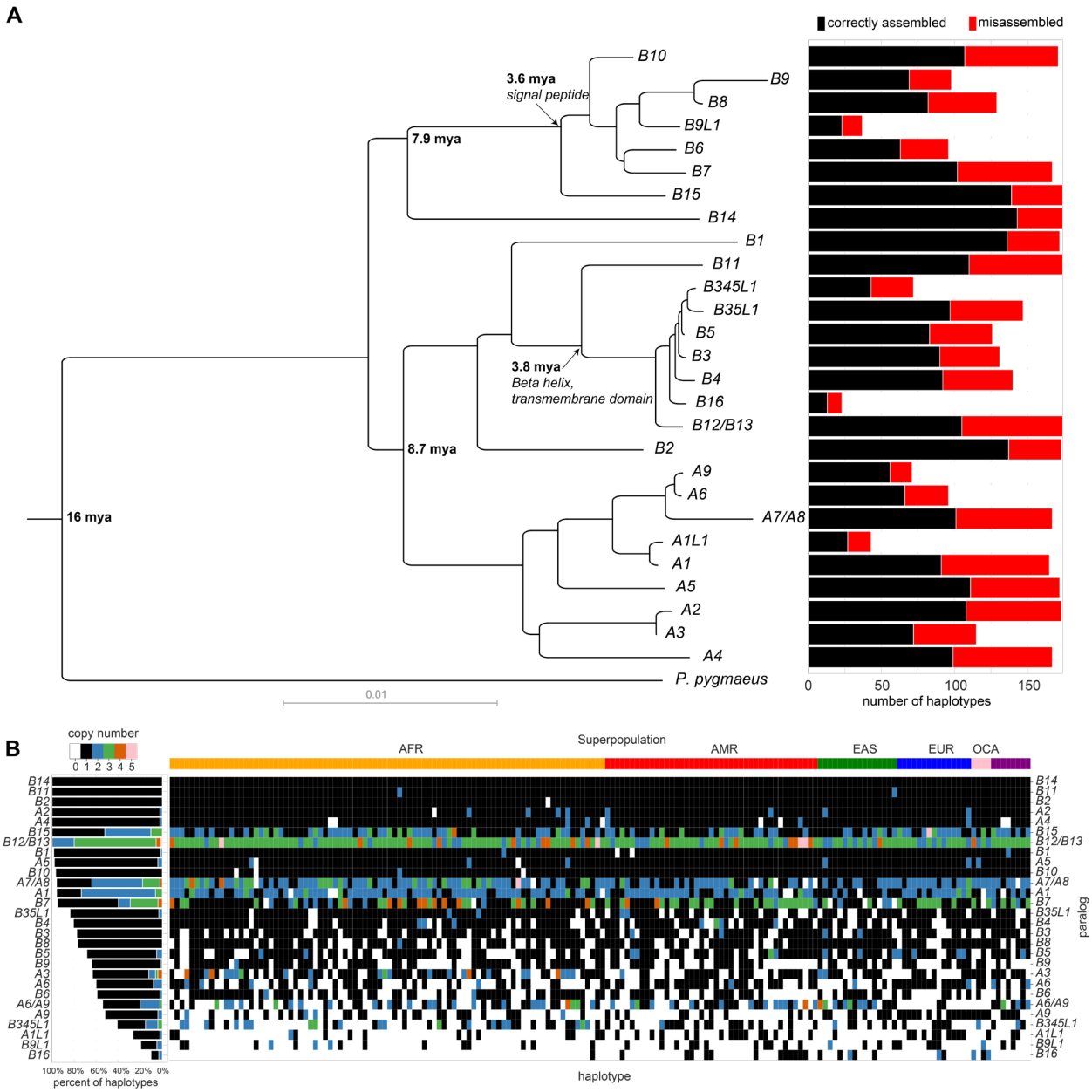


Figure 4.2. Classification of human *NPIP* haplotypes and locus-specific copy number.

A) Left: Phylogeny of human *NPIP* loci outgrouped to *Pongo pygmaeus* based on 15 kbp of intronic sequence Right: Frequency of each paralog among the 169 haplotypes passing assembly validation (black) and number of misassembled loci typically where a potential collapse was identified (red). **B)** Copy number summary of 169 assembled haplotypes. Color indicates copy number of each gene, as defined by the phylogenetic grouping (panel A). Paralogs are sorted by fraction of validated haplotypes containing at least one copy, and haplotypes are grouped by continental superpopulation. Left: Percent of haplotypes with each copy number state for each

paralog, restricted to assembled regions passing QC. Right: copy number states for all assembled haplotypes, with each column representing a separate haplotype. Haplotypes are grouped by superpopulation (above).

In addition to copy number variation, interlocus gene conversion (IGC) is another common source of *NPIP* variation, as the high sequence identity among paralogs enables the replacement of *NPIP* sequence from one paralog to another. As a result, the sequence content of a paralog does not always correspond to the syntenic location in human haplotypes. We reanalyzed a recent genome-wide IGC callset for a subset of these (n=94 haplotypes) to classify IGC patterns among *NPIP* duplication blocks (Vollger, Dishuck, et al. 2023). As expected, IGC between paralogs is frequent (exceeding >50% of haplotypic configurations) and is driven primarily by proximity (1-2 Mbp) with six distinct IGC "hotspots" identified (Fig. S1, Fig. 3A) on the short arm of human chromosome 16. IGC is restricted by the two major subfamilies *(NPIPB* copies undergo IGC only with *NPIPB* but not *NPIPA* loci). We also observe particular biases in donor/acceptor directionality. For example, the putative ancestral paralog *NPIPA1* acts only as a donor to *A5*, *A6*, *A8*, and *A9* locations, but never as an acceptor, reflecting either functional constraint or bias in the mutation process itself.

During our comparative analysis of validated chromosome 16 structural haplotypes, we frequently noted that the gene order of unique (nonduplicated) genes flanking *NPIP* copies are inverted. In total, we identify four inversion polymorphisms ranging in size from 350 kbp to 1.6 Mbp in size (Fig. 3B-E) across chromosome 16p. The breakpoints of these inversions map either at *NPIP* copies or associated SDs flanking *NPIP.* All of these large inversions are common polymorphisms (>5% allele frequency) and in some cases represent the major allelic

configuration in the human population. In several cases, the inverted unique sequence shows considerable allelic divergence (<99.8%) suggesting a deep coalescence as has been observed for other human inversion polymorphisms (Zody et al. 2008; Porubsky et al. 2022). Indeed, the coalescence of the D1 inversion polymorphism at 16p11.2 has been estimated as 1.35 mya and associated with susceptibility to asthma and obesity (González et al. 2014). González et al. estimated that at least six distinct haplotypes exist at this locus based on multidimensional scaling of single-nucleotide polymorphisms (SNPs); we are able to resolve 13 structural configurations at this locus distinguished by orientation, *NPIP* paralog identity, and *SULT1A* copy number that were previously indistinguishable. This complete sequence resolution may help explain their observed association of the inversion with increased *SULT1A4* expression and decreased *SULT1A1* expression (González et al. 2014). Importantly, we note that many of the human haplotypic configurations occurred in conjunction with copy number variation and IGC events associated with specific *NPIP* loci.

In order to classify different structural configurations, we encode haplotypes by the identity, order, and orientation of *NPIP* paralogs and marker genes. We apply a double-cut-and-join rearrangement distance metric (Bohnenkämper 2024) (Methods) between each configuration to create corresponding neighbor joining trees for each of the major *NPIP* clusters (Fig. 3E-F). For example, at the ancestral chromosome 16p13.11 locus, we observe a 545 kbp inversion and the variable presence or absence of *A3* and a newly discovered paralog, *A1L1.* By contrast, *A5* maps invariably at the proximal end of this cluster (Fig. 3A,E). The *A1L1* paralog only associates with 16p13.11 haplotypes that are inverted relative to T2T-CHM13; this 545 kbp inversion is the major allele (AF=0.69). The chromosome 16p11.2 locus contains *NPIPB6, B7, B8*, and *B9*

spanning a 650 kbp inversion polymorphism (Fig. 3D,G). Through IGC and inversions, *B7* sequence can occupy any of the four canonical *NPIP* locations in this cluster. Additionally, 8/13 configurations also have a 355 kbp inversion with respect to T2T-CHM13, and only the inverted orientation configurations carry the *B6* or *B9* genes. At 99.6% sequence identity to the reference genome, this inverted region is in the lowest 9.5% of sequence identity genome-wide. We also observed 1.6 and 1.3 Mbp inversions at chromosome 16p12.3 and p12.2, respectively (Fig. 3D,E). Altogether, of the nine loci containing *NPIP* paralogs, only the locus at 16p13.3 containing *B2* is structurally invariant.

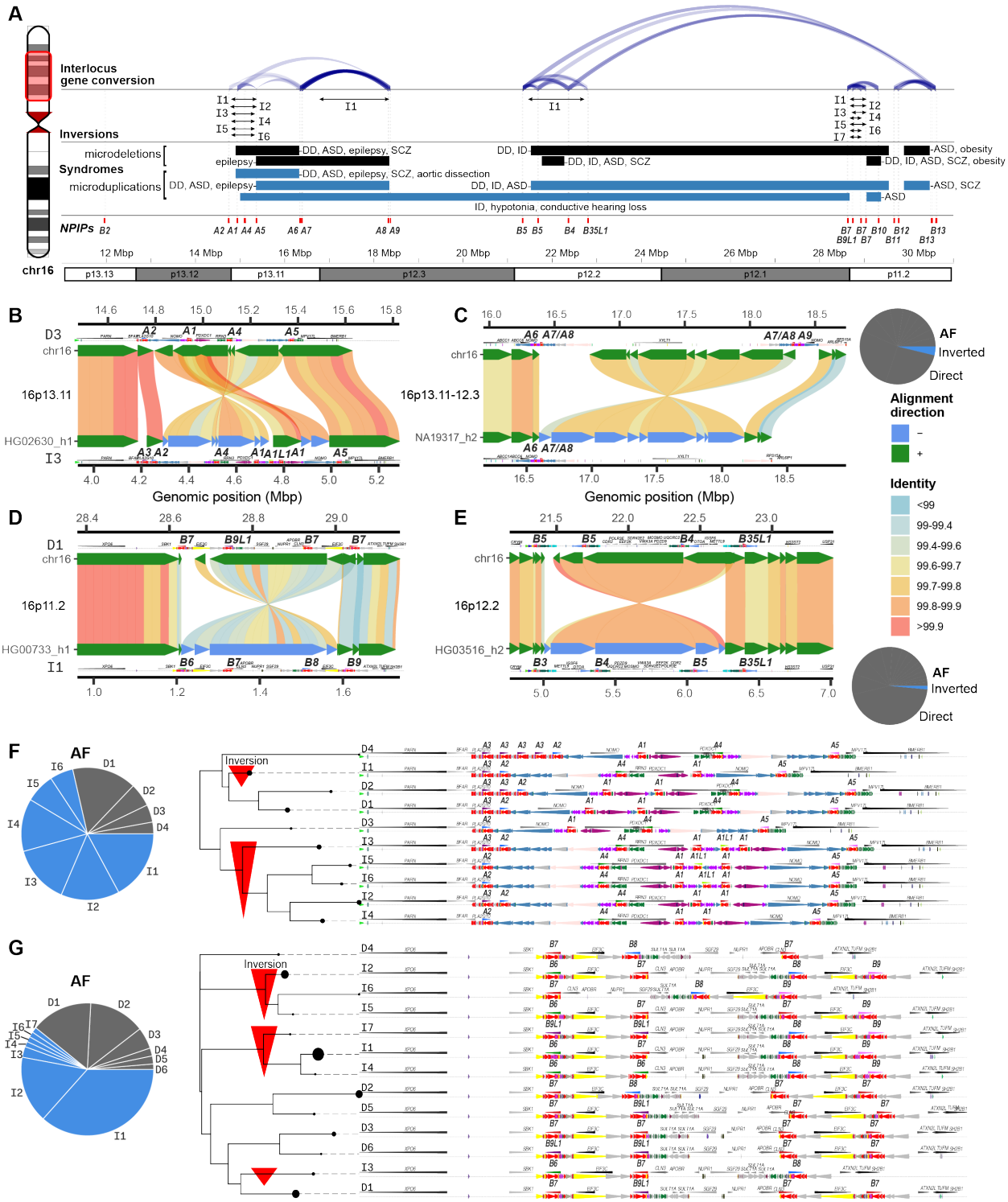


Figure 4.3. *NPIP* interlocus gene conversion (IGC) and complex structural changes.

A) Overview of *NPIP* loci on chromosome 16p (highlighted ideogram region). The location of each T2T-CHM13 *NPIP* paralog is shown as a red bar. The count and location of IGC between *NPIP* pairs is shown as blue arcs at the top, with opacity corresponding to number of observed

haplotypes. Inversions mediated by *NPIP* are shown as black arrows, named corresponding to structures in panels B-G. Known pathogenic microdeletions and microdeletions with breakpoints at *NPIP* are shown as black and blue bars, respectively (Sharp et al. 2006; Ballif et al. 2007; Kumar et al. 2008; Weiss et al. 2008; Hannes et al. 2009; de Kovel et al. 2010; Bochukova et al. 2010; Girirajan et al. 2010; Heinzen et al. 2010; Antonacci et al. 2010; Ingason et al. 2011; Kuang et al. 2011; Ramalingam et al. 2011; Cooper et al. 2011; Barber et al. 2012; Quintela et al. 2015; Loureiro et al. 2017; Loviglio et al. 2017; Coe et al. 2019; Pop-Jordanova et al. 2021; Nicolle et al. 2022). **B-E)** Large-scale inversion polymorphisms associated with *NPIP* loci (below) as compared to T2T-CHM13v2 (top). Inversions are shown with SVbyEye, with DupMasker annotations for each haplotype. Allele frequency (AF) for inverted (I) and direct (D) orientation haplotypes are shown with the pie charts at right or in panels F-G. **F-G)** The duplication architecture (Dupmasker) of the *A1-5* and *B6-9* loci for the most common haplotype configurations, grouped by a neighbor-joining tree of double-cut-and-join edit distance (pairwise number of rearrangements between configurations). *NPIP* sequence denoted in red. The size of the circle for each clade corresponds to frequency of each configuration, and direct and inverted orientation configurations named by frequency. Red arrows under the cladogram indicate configurations inverted with respect to T2T-CHM13.

4.3.2 *Diversity-based tests of selection*

The *NPIP* gene family members were previously shown to harbor a significant excess of amino-acid replacements, exhibiting one of the most extreme signals of positive selection in the human and African ape lineage (i.e., dN/dS>1.0) (Johnson et al. 2001). To assess whether positive selection is still ongoing in the human population and narrow down signatures to individual loci, we performed complementary tests of Tajima's D and nS_L for extended haplotype homozygosity (Tajima 1989; Ferrer-Admetlla et al. 2014). Tajima's D compares the number of segregating sites to pairwise differences to find deviations from the neutral expectation; negative values correspond to an abundance of rare alleles and are consistent with positive selection, while positive values correspond to a scarcity of rare alleles and are consistent with balancing

selection. nS_L (number of segregating sites by length) is a test of extended haplotype homozygosity designed to detect recent hard and soft selective sweeps (Ferrer-Admetlla et al. 2014) and is more robust than Tajima's D to artifactual signals arising from bottlenecks and population growth. Unlike other haplotype-based tests of selective sweeps like iHS, it is robust to phasing errors and does not rely on detailed recombination maps, as such maps are either nonexistent or unreliable in SD regions (Methods; Szpiech 2024).

Previous attempts have been confounded by the inability of short reads to align to these duplicated regions, but our contiguous haplotype-resolved assemblies now allow us to investigate whether there is evidence of selective sweeps across these regions in the human population. We calculated nS_L and Tajima's D with HiFi assemblies, restricting to individuals of African ancestry. To evaluate consistency of Tajima's D within unique sequence flanking SDs, we compared Tajima's D using short reads from the 1KG (Gambian individuals). Signals are comparable in unique regions (Fig. 4) but dropout over SDs for short-read sequence data. Using Tajima's D, we find positive selection signatures in the 1% most extreme chromosome-wide for *NPIPB9*, *B12*, and *B15*, while *A1*, *A2*, *A5*, *B3*, *B4*, *B11*, *B13*, and *B14* are within the 5th percentile. *B7* and *A7* are, however, within the 5% most extreme windows for balancing selection (Fig. 4A). Similarly, with nS_L we find signatures of selective sweeps for *B7*, *B9*, and *B15* within the 1st percentile of most extreme values, and for *A8* within the top 5th percentile (Fig. 4C,D). The only two regions of consecutive nS_L values in the first percentile correspond to *B7/9* and *B15 NPIP* loci. However, *B3*, *B4*, *B5*, *B7*, and *B9* are located within or near the boundaries of inversions, raising the possibility that suppressed recombination may be contributing to this signal (Fig. 3E,F).

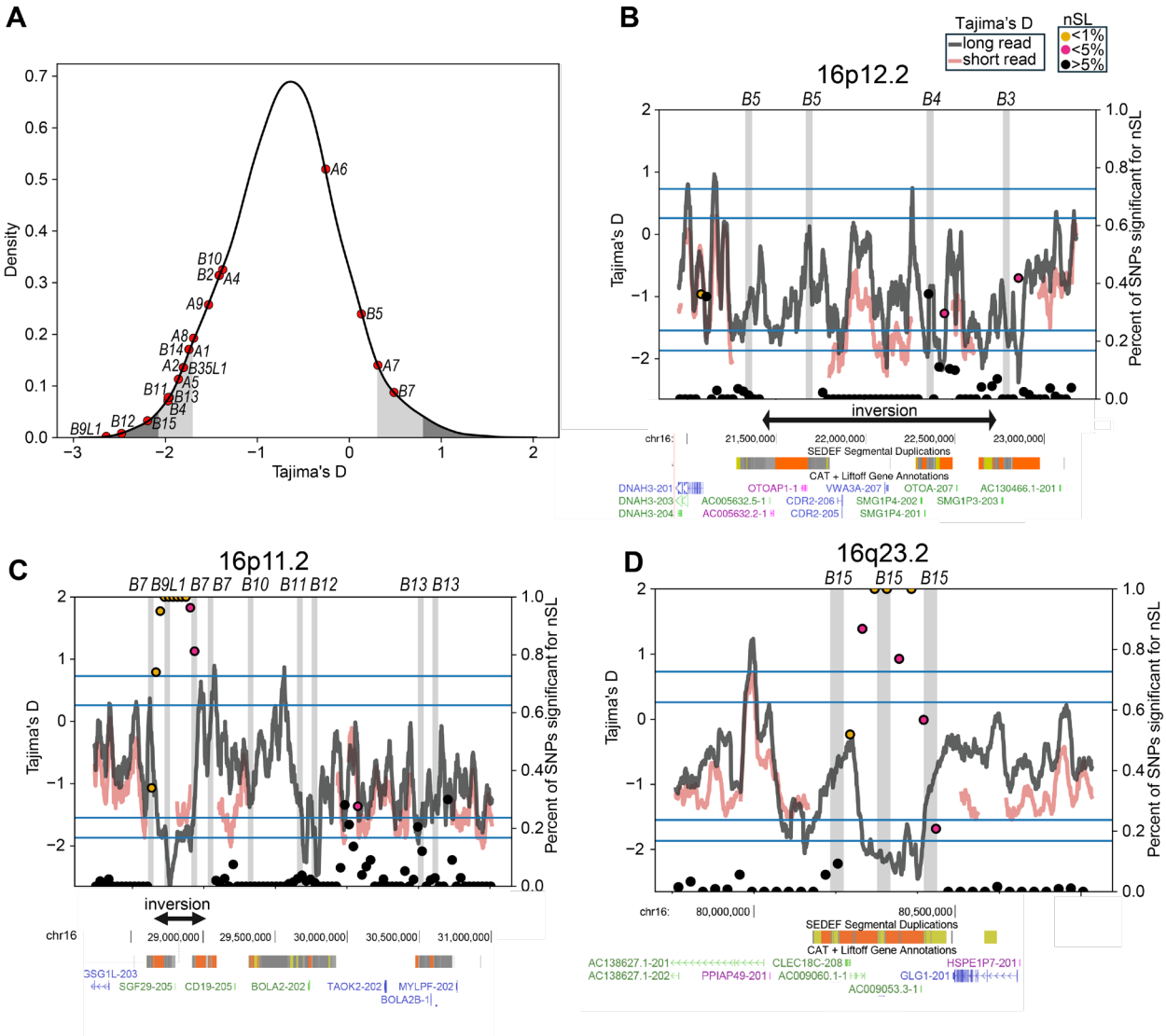


Figure 4.4. Selection signatures at *NPIP* loci in the human population.

A) Tajima's D distribution on chromosome 16, from long-read sequencing and assembly. The most extreme 1% and 5%, both positive (balancing selection) and negative (positive selection) are colored in gray and dark gray. The values for various *NPIP* paralogs are represented with red dots along the distribution. **B-D**) Results of Tajima's D and nSL selection tests for three loci showing signatures of positive selection. Short-read and long-read Tajima's D results are shown across each locus as red and blue lines, respectively. nSL values are plotted as filled circles with color indicating significance. Known inversion polymorphisms are indicated (black arrows below) along with SDs and T2T gene annotations. Horizontal lines indicate 1% and 5% cutoffs for Tajima's D, both positive and negative. Vertical gray highlights indicate locations of *NPIP* paralogs, with gene names and SDs (SEDEF) shown below.

4.3.3 *NPIP gene models and differential expression*

Previous research demonstrated ubiquitous expression of *NPIP* paralogs in apes, as compared to the largely testis-specific expression in Old and New World monkeys, along with slightly different gene models for human *NPIPA* and *NPIPB* subfamilies (Cantsilieris et al. 2020). With our more complete catalog of human *NPIP* paralogs, we sought to determine whether we could identify additional paralog-specific changes to gene models, and if there is evidence tissue-specific expression when considering particular *NPIP* paralogs instead of the family as a whole. Short-read RNA-seq does not align uniquely to *NPIP* paralogs due to their high sequence identity, preventing the construction of complete gene models and paralog-specific expression estimates. Instead, we used PacBio HiFi sequencing of full-length cDNA (Iso-Seq), facilitating the unambiguous assignment of the majority of Iso-Seq reads to specific *NPIP* paralogs.

To this end, we assembled a database of full-length non-chimeric (FLNC) cDNA generated from 1.4 billion Iso-Seq reads from 384 libraries, representing 101 human tissue and cell types (Supplementary Table S1). To complement this effort, we also performed hybridization capture experiments against select tissues using *NPIP*-targeting capture probes in order to enrich in *NPIP* FLNC molecules (Methods; Dougherty et al. 2018). We extracted Iso-Seq reads aligning to any *NPIP* paralog, totaling 1.07 million reads with an average length of 1960 nt. To create paralog-specific gene models, we considered open reading frames (ORFs) seen in at least five Iso-Seq molecules as valid and only display the most abundant and longest isoforms for each paralog (Fig. 5A).

We observe Iso-Seq molecules encoding full-length ORFs for most *NPIP* paralogs (Fig. 5). This includes four paralogs that had previously been annotated as noncoding pseudogenes, *NPIPB1P*, *NPIPP1*, *NPIPB10P*, and *NPIPB14P*, which we refer to as *B1*, *A4*, *B10*, and *B14*, respectively. The African-ape specific *B1* paralog, the only human paralog on chromosome 18 and therefore not predisposed to the same level of structural variation, was previously reported to neither be transcribed nor maintain an ORF (Cantsilieris et al. 2020). By contrast, we find that it maintains an ORF and is expressed, albeit at low levels, in testis and brain organoids.

Closer inspection of these gene models reveals a considerable amount of variation in predicted amino acid composition across *NPIP* paralogs and their isoforms due to alternative promoters, differences in translation initiation, and expansion of protein-encoding VNTRs. Consequently, ORFs range from 155 to 1217 amino acids. Of the 55 most common *NPIP* isoforms, only seven begin with the canonical first coding exon "MFCC...," which is shared with African apes (Cantsilieris et al. 2020), and only five were represented in RefSeq. Eleven begin with an alternate translation initiation "MVKL" sequence previously identified as the start sequence for the *NPIPB* subfamily (Bekpen et al. 2017). In addition to *NPIPB* paralogs, we also observe this start sequence for *NPIPA2* and determine that this 40 amino acid exon arose from an independent duplication of the twelfth exon of *ACSM1*, an acyl-CoA synthetase gene (Fig. S2), including half of its AMP-binding enzyme C-terminal domain (InterPro domain IPR025110). Cantsilieris et al. reported an "MRVR" start sequence in non-African ape primates, perhaps the ancestral sequence; we observe this start site used in 11 human *NPIPA* isoforms. A subset of *NPIPB* members (*B6*, *B7*, *B8*, *B9*, *B10*, *B14*, *B15*) use a previously undocumented "MRLR" start site, encoding a 19-26 amino acid signal peptide, as predicted by SignalP-6.0 (Teufel et al.

2022). Though the sequence that encodes the signal peptide is present in all human *NPIP* paralogs and shared with nonhuman primates, we estimate the clade that uses this sequence as its transcription start site to be human specific and only 3.6 mya.

Finally, for the six *NPIP* paralogs adjacent to *PKD1* pseudogenes (*A1*, *A4*, *A6*, *A7*, *A8*, *A9*), we observe 10 distinct *PKD1-NPIP* fusions, four of which are multi-exonic fusing up to nine *PKD1* exons (530 aa) with eight *NPIP* exons (343 aa). Remarkably, these fusions maintain long ORFs up to 843 aa in length. *PKD1* variants are implicated in polycystic kidney disease, as well as estimated glomerular filtration rate (Hellwege et al. 2019). Though the *NPIP*-adjacent *PKD1* copies have been considered pseudogenes because their predicted ORFs are truncated (Bogdanova et al. 2001) several partial gene duplications like *SRGAP2C*, *NOTCH2NL*, and *ARHGAP11B* have been shown to be functional through dominant negative interaction (Dennis et al. 2012; Florio et al. 2015; Fiddes et al. 2018).

The final coding exon of the human-specific *NPIPB* subfamily (*B3*, *B4*, *B5*, *B11*, *B12*, *B13*) contains an expanded in-frame VNTR. Our analysis of 169 haplotypes and hundreds of cDNA libraries demonstrates that even within single paralogs, the copy number of this VNTR is variable among individuals. The VNTR encodes a repetitive amino acid motif of 19 (SADDNLKTPSERQLTPLPP) or 23 (SADDNIKTPAERLRGPLPPSAPP) residues, with the two lengths alternating. Within each paralog, the sequence frameshifts from the SADDN… form (7-15 repeats) to a MIISRHLPSVSSLPFHPQLHPQQMI form (6-14 repeats), and back to SADDN… (5-11 repeats) in the genomic annotations, resulting in a repeat domain ranging in size from 297 to 1,298 amino acids (Fig. 5B). Analyzing Iso-Seq cDNA directly, we observe

8,986 molecules sharing this VNTR switching pattern with up to 25, 20, and 17 repeat units. Computational protein structure prediction suggests that both frames of the VNTR repeat may form a left-handed beta helix, with each VNTR repeat unit corresponding to an additional turn of the helix, kinked as the frame shifts (Fig. 5C). The gene models for this subfamily also encode a transmembrane domain as predicted by DeepTMHMM (Hallgren et al. 2022). We estimate this *NPIP* branch to have arisen 3.8 mya (Fig. 2A).

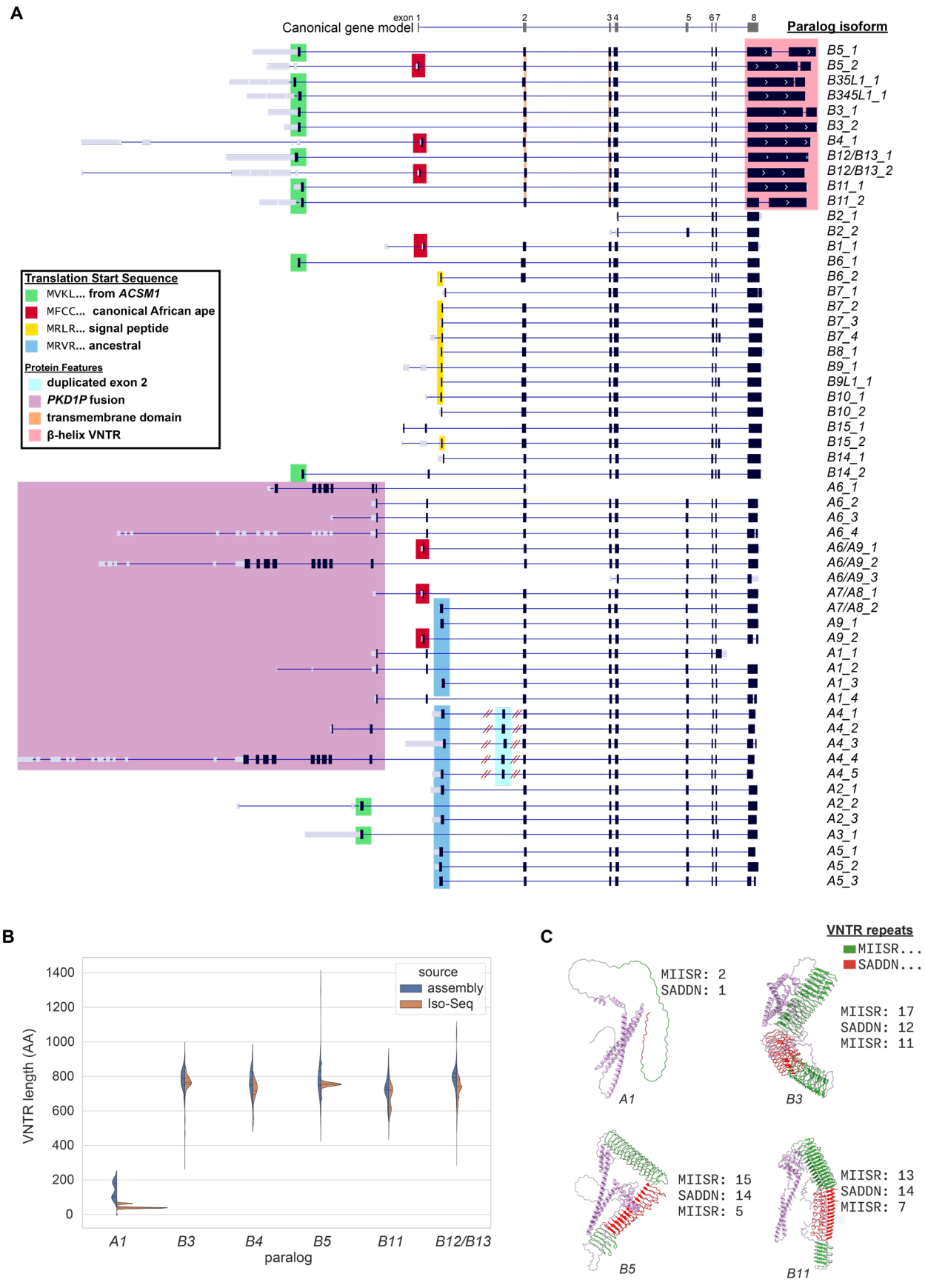

A
Canonical gene model
exon 1
2
3 4
5
6 7
8
Paralog isoform
B5_1
B5_2
B35L1_1
B345L1_1
B3_1
B3_2
B4_1
B12/B13_1
B12/B13_2
B11_1
B11_2
B2_1
B2_2
B1_1
B6_1
B6_2
B7_1
B7_2
B7_3
B7_4
B8_1
B9_1
B9L1_1
B10_1
B10_2
B15_1
B15_2
B14_1
B14_2
A6_1
A6_2
A6_3
A6_4
A6/A9_1
A6/A9_2
A6/A9_3
A7/A8_1
A7/A8_2
A9_1
A9_2
A1_1
A1_2
A1_3
A1_4
A4_1
A4_2
A4_3
A4_4
A4_5
A2_1
A2_2
A2_3
A3_1
A5_1
A5_2
A5_3
Translation Start Sequence
MVKL... from ACSM1
MFCC... canonical African ape
MRLR... signal peptide
MRVR... ancestral
Protein Features
duplicated exon 2
PKD1P fusion
transmembrane domain
β-helix VNTR
B
VNTR length (AA)
1400
1200
1000
800
600
400
200
0
source
assembly
Iso-Seq
A1
B3
B4
B5
B11
B12/B13
paralog
C
VNTR repeats
MIISR...
SADDN...
MIISR: 2
SADDN: 1
A1
MIISR: 17
SADDN: 12
MIISR: 11
B3
MIISR: 15
SADDN: 14
MIISR: 5
B5
MIISR: 13
SADDN: 14
MIISR: 7
B11

Figure 4.5. Paralog-specific gene models.

A) Most common isoforms for each *NPIP* paralog based on full-length cDNA Iso-Seq mapping. The _x suffix indicates relative abundance (i.e., B5_1 is the most abundant B5 gene model). Predicted protein-coding regions (black), untranslated regions (gray) with different protein start sequences and structural features are highlighted (color) over the gene models, including transcripts with the expanded protein-encoding beta-helix (pink) and the signal peptide (yellow). The canonical *NPIP* gene model is depicted (top). **B)** A comparison of VNTR length encoding the beta-helix of exon 8 in the genome assemblies versus Iso-Seq data. C) Predicted protein structures for four paralogs with exon 8 VNTR sizes. The copy number of repeating amino acid motifs by type are indicated and projected onto Chai-1 structure predictions. (MIISR… repeat protein domain shown in green, while frameshifted VNTR SADDN… repeat protein domain in red.)

We attempted to assess paralog-specific expression levels of *NPIP* paralogs using both Iso-Seq reads and short-read RNA-seq using a unique k-mer approach to specifically tag the short read data. To determine paralog identity, the 1.07 million *NPIP* Iso-Seq reads were aligned to each of the 169 assembled haplotypes, recording location of best mapping, and requiring a difference of at least one additional mismatch to the next-best mapping paralog to consider a read uniquely identified. The approach allowed us to assign ~55.6% of Iso-Seq reads to create paralog-specific gene models. Grouping highly similar paralogs (*A2/3*, *A6-9*, *B3-5*, and *B12/13*) allowed us to assign ~93.2% of Iso-Seq reads for expression analysis. While these estimates are not quantitative due to errors and biases inherent in library preparation and sequencing, we observe relative and reproducible differences in paralog expression across tissue types. Comparing expression between tissue types, specific clusters of paralogs have increased relative expression in distinct tissues. In particular, *NPIPA1*, *A5*, *A6-9*, *B3-5*, and *B12/13* show increased expression in fetal or adult brain relative to other tissues, while *A2-3*, *A4*, *B1*, *B2*, *B6-9*, *B10*, *B14*, and *B15*

retain the presumed ancestral testis-enriched expression pattern (Bekpen et al. 2017). Immune function-related tissues like tonsil, B cells, granuloma, and blood also significantly overexpress paralogs seen in brain or testis.

We also examined developmental time-point specificity by classifying short-read RNA-seq reads from an atlas of organ development (Cardoso-Moreira et al. 2019) based on the presence of uniquely identifying k-mers from the 169 haplotypes. Paralogs that contained few uniquely-identifying k-mers were combined into larger paralog groups for this analysis (Methods). Altogether, 25.2% of reads containing any *NPIP* k-mer (n=1.06 million) are uniquely assigned to a paralog or paralog group. *NPIPA1*, *A4*, and *B3-5* tend to increase in expression in the cerebellum after birth (Fig. 6B). In contrast, *B1*, *B2*, *B6-9*, *B10*, and *B15* expression is almost entirely testis-specific with levels increasing after puberty (Figures 6C, S3).

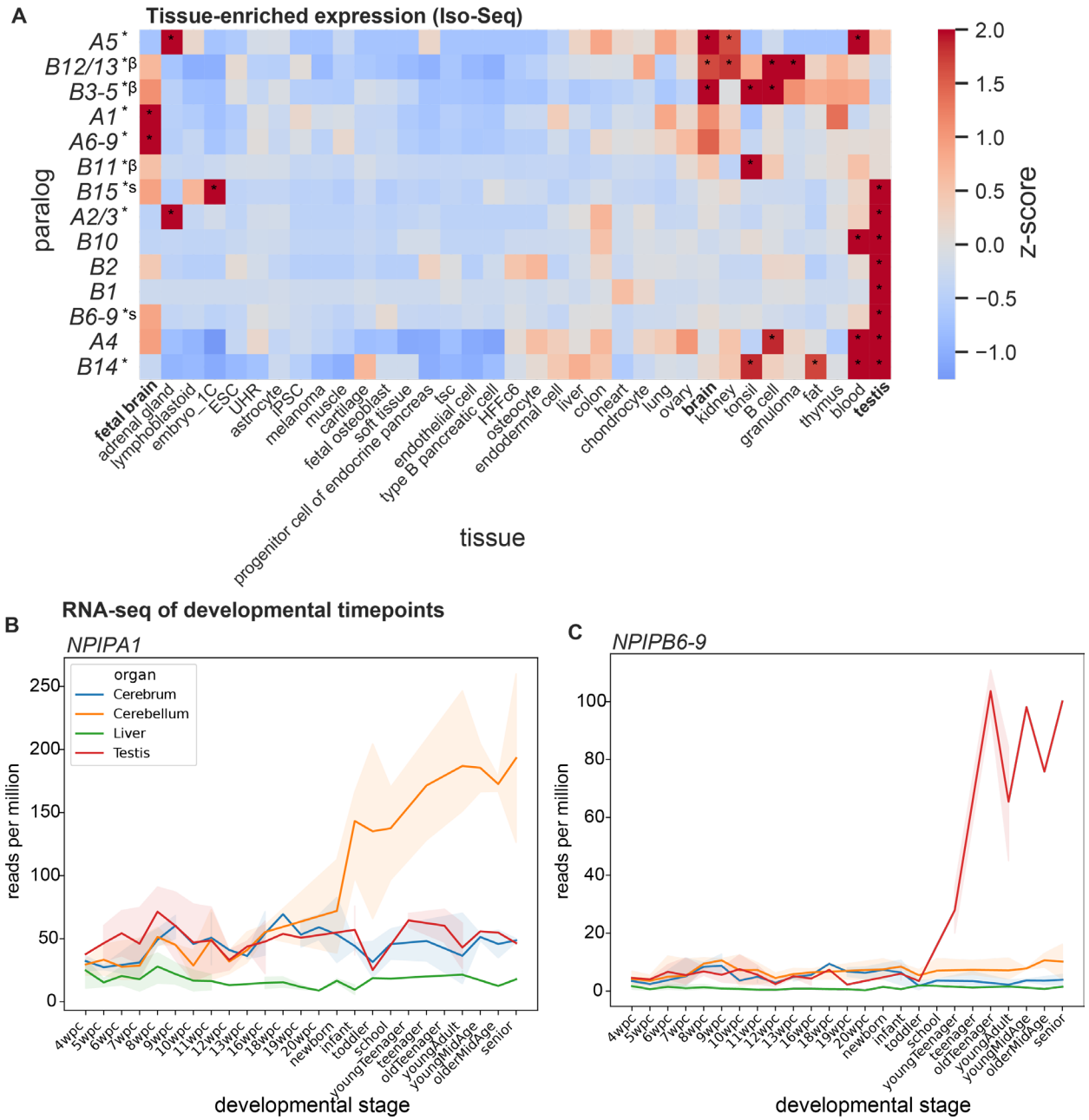


Figure 4.6. Variable expression of *NPIP* paralogs across tissues, cell types, and developmental time points.

A) Relative enrichment of Iso-Seq expression estimates for 35 tissues, clustered with UPGMA. Significantly positive z-scores are indicated with * ($p<0.05$). Paralogs with selection signatures are indicated with * at left; beta helix: β; signal peptide: s. **B-C)** Short-read RNA-seq expression estimates for human developmental timepoints in four tissues for *NPIPA1* and *NPIPB6-9* paralogs (aggregate), using unique k-mers for paralog identity. Transparent error bands represent 95% confidence interval of replicates.

4.4 DISCUSSION

The expansion of *NPIP* and its associated SDs across the short arm of chromosome 16 predispose humans to frequent recurrent pathogenic duplications and deletions associated with autism, developmental delay, and obesity (Sharp et al. 2006; Ballif et al. 2007; Kumar et al. 2008; Weiss et al. 2008; Hannes et al. 2009; de Kovel et al. 2010; Bochukova et al. 2010; Girirajan et al. 2010; Heinzen et al. 2010; Antonacci et al. 2010; Ingason et al. 2011; Kuang et al. 2011; Ramalingam et al. 2011; Cooper et al. 2011; Barber et al. 2012; Quintela et al. 2015; Loureiro et al. 2017; Loviglio et al. 2017; Coe et al. 2019; Pop-Jordanova et al. 2021; Nicolle et al. 2022). Despite this negative effect on fitness, the duplications not only persist but have expanded in the African great apes. Moreover, these same sites have homogenized via IGC, ensuring high sequence identity and driving NAHR. In light of the strong signals of positive selection for this hominid gene family (Johnson et al. 2001; Cantsilieris et al. 2020), we hypothesized that an evolutionary tradeoff exists between disease susceptibility and, as of yet, unknown adaptive function. In this work, we catalog, for the first time, normal human variation at each *NPIP* locus and identify paralog-specific features potentially relevant to understanding the function of this enigmatic gene family.

Because functional human-specific genes are frequently invariant (Dennis et al. 2012; Florio et al. 2015; Fiddes et al. 2019), we systematically assessed copy number for each paralog. Based on our 28 distinct human phylogenetic groups, we find only three loci that are copy number invariant (*NPIPB2*, *B11*, and *B14)*. We also distinguish loci that always have at least one copy in humans although often more (*A2*, *A4*, *B12/B13*, and *B15*). It should be noted that several of these copies are associated with larger scale structural changes such as inversion polymorphism or

gene conversion events. Based on the human genomes we surveyed here, only the *NPIPB2* locus is invariant by all analyses. While copy number polymorphic, the ancestral locus, *NPIPA1* (Johnson et al. 2001; Cantsilieris et al. 2020), shows a striking asymmetry for IGC, serving only as a donor and never an acceptor of a gene conversion event. Such asymmetry may reflect either mutational bias or functional restraint.

We previously identified extreme signatures of positive selection for *NPIP* in African apes, and particularly the *NPIPB* subfamily based solely on tests for an excess of amino-acid replacement among paralogs (Johnson et al. 2001; Cantsilieris et al. 2020). With highly contiguous haplotype-resolved assemblies, we were able to apply population-level selection tests for the first time. Using Tajima's D and nSL, we find that *NPIPB9* and *B15* are within the top percentile of most extreme values for both tests, while nine additional paralogs occur within the top 5th percentile for at least one test. We also find some evidence of balancing selection for a few loci (*A7* and *B7*). While these findings strongly suggest ongoing positive selection in humans, caution must be exercised given the large-scale structural changes associated with these regions. For example, *NPIPB3-B9* are located within or near the boundaries of inversions, raising the possibility that suppressed recombination may be contributing to this signal, including extended haplotypes (Fig. 3D-E). These signals may not, however, be mutually exclusive with inversions enriched for adaptively evolving genes (Kirkpatrick and Barton 2006; Charlesworth and Barton 2018). Such is the case of the 17q21.31 inversion polymorphism—a locus associated within increased fecundity (Stefansson et al. 2005; Zody et al. 2008), positive selection in humans (Boettger et al. 2012; Steinberg et al. 2012), and the dynamic evolution of newly minted gene family

LRRC37A1/2 (Zody et al. 2006; Bekpen et al. 2012; Giannuzzi et al. 2013) expressed highly in human astrocytes (Bowles et al. 2022).

With a database of 1.4 billion FLNC reads (Iso-Seq), we were able to comprehensively construct paralog-specific gene models. 91% (50/55 most abundant isoforms) have not been previously described in RefSeq. All but one paralog maintains a full-length ORF, while *B2*, the most copy number invariant, is predicted to encode a truncated protein. The full-length gene models reveal new features—such as *NPIP* subfamilies gaining a start sequence co-opted from *ACSM1*, a novel signal peptide, transmembrane domain, or a variably sized coding VNTR that is predicted to form a beta helix. We estimate that the signal peptide and beta helix evolved independently 3.6-3.8 mya and are innovations specific to the human lineage of evolution. The specificity afforded by LRS or paralog-specific k-mer analysis also reveals tissue-specific differences. For example, the subfamily encoding the novel signal peptide includes the two paralogs with the strongest signal of positive selection (*B6-9* and *B15*). This set shows testis-enriched expression, and analysis of a short-read development dataset additionally indicates that these paralogs increase in abundance at the onset of puberty. The paralogs with the novel beta helix and transmembrane domain, by contrast, tend to be enriched in brain samples (*B3-5*, *B12/B13*), with *B12* also among the strongest signals of positive selection.

In summary, the dynamic changes in copy number, gene model, and expression specificity across *NPIP* paralogs, along with strong signals of positive selection, suggest neofunctionalization of specific copies during human evolution. Both the paralogs that have maintained the presumed ancestral testis expression pattern (*B15*) and those that have gained enriched brain expression

(*B12/B13*) exhibit clear signatures of positive selection. Concurrently, these two subfamilies evolved radically distinct gene models and associated protein structural changes in the human lineage. All of these changes have occurred and potentially been accelerated in a milieu of recurrent structural variation and IGC. Notwithstanding, it is noteworthy that *B15* and *B12/B13* are among just four paralogs that are sometimes duplicated but never deleted among humans, a potential indication of their intolerance to loss. Now that the organization and variation of these oft-overlooked loci has been resolved and their variation and gene structures understood, the next step will be associating this variation with human phenotypes including disease.

4.5 METHODS

4.5.1 *Short-read copy number estimation*

We applied fastCN to high-coverage Illumina data for 2,609 unrelated individuals from the 1KG to estimate *NPIP* copy number (Pendleton et al. 2018; Byrska-Bishop et al. 2022). Windows overlapping the exon 8 VNTR were excluded from copy number estimation to avoid biasing the estimate.

4.5.2 *NPIP gene identification*

To identify *NPIP* gene locations within assemblies, we aligned the ancestral *NPIP* locus from GRCh38 (chr16:14,935,711-14,954,790) to each haplotype separately with wfmash (v0.7; parameters: -p 80 --num-mappings-for-segment=10000) and minimap2 (v2.22; parameters -x map-ont -f 5000 -N 300 -p 0.5) (Li 2018; Marco-Sola et al. 2023), restricting to aligned regions of at least 15 kbp. We also applied Dupmasker (v1.11) (Jiang et al. 2008) to identify the LCR16a

duplicon where *NPIP* is located (SD9443). Dupmasker identified additional copies of *NPIP* only in nonhuman primates but was not necessary for detecting *NPIP* copies in human haplotypes.

4.5.3 *Genome-wide gene annotation*

We annotated genes on each haplotype with Liftoff (v1.6.3; parameters: -flank 0.1 -polish -sc 0.85 -copies -mm2_options="-a --end-bonus 5 --eqx -N 10000 -p 0.3 -f 1000") (Shumate and Salzberg 2021), using protein-coding genes in GENCODE v44 (Frankish et al. 2023) on GRCh38 as the reference annotation set.

4.5.4 *Phylogenetic paralog identity*

We created an MSA of *NPIP* genes from each human and Pongo pygmaeus haplotype using MAFFT (v7.487; FFT-NS-2) (Katoh and Standley 2013). To create a phylogenetic tree of *NPIP* paralogs, we trimmed VNTRs, exons, and poorly aligned regions from the MSA visually. We estimated a maximum-likelihood phylogeny from this MSA using IQ-TREE (v2.2.3 COVID-edition; parameters -B 1000 -alrt 1000) with the GTR+F+R6 substitution model selected with ModelFinder (Kalyaanamoorthy et al. 2017; Minh et al. 2020). Ultrafast bootstrap and SH-aLRT were used as measures of clade confidence (Anisimova et al. 2011; Hoang et al. 2018). Clades were named based on annotations of GRCh38, T2T-CHM13v2, and T2T-HG002 and defined with SH-aLRT branch support values >75.

4.5.5 *Genome assembly validation*

Assembled regions were validated with NucFreq, Flagger, and GAVISUNK depending on availability of orthogonal sequencing data (Vollger et al. 2019; Cartney et al. 2022; Dishuck et al. 2022; Liao et al. 2023). Regions with no read support, only ONT support, or only HiFi support were removed. Flagger and NucFreq were applied to hifiasm and Verkko assemblies and

excluded erroneous, falsely duplicated, collapsed, low confidence, or unreliable blocks. GAVISUNK was applied to hifiasm assemblies as described in Vollger, 2023, and supported regions were kept for downstream analyses (Vollger, Dishuck, et al. 2023). Only assemblies that were contiguous between proximal and distal non-segmentally duplicated marker genes were considered.

4.5.6 *Locus configuration comparisons*

To compare structural configurations for each *NPIP* locus across samples, 10 loci were defined based on adjacent non-duplicated genes from Liftoff annotations. Configurations were definited based on order and orientation of *NPIP* paralogs and protein-coding genes from the Liftoff annotations relative to adjacent marker genes. Only configurations that passed assembly validation in at least one haplotype and in were detected in at least two haplotypes were considered for further analysis. To calculate rearrangement distance between each configuration at each locus, the order and orientation of DupMasker annotations of at least 1 kbp and protein-coding marker genes was used as input to the capping-free double-cut-and-join indel model (Bohnenkämper 2024), and the matrix of pairwise rearrangement distances was transformed into a midpoint-rooted neighbor-joining tree with Bio.Phylo (Talevich et al. 2012). The resulting trees, gene annotations, and DupMasker content were visualized with custom scripts and Baltic (Dudas 2024).

4.5.7 *VNTR analysis*

To measure the length of *NPIP* exon 8 VNTRs, Tandem Repeats Finder (TRF v4.10; parameters 2 5 7 80 10 10 2000 -d -ngs) was applied to each *NPIP* copy from each haplotype (Benson 1999). The longest contiguous region of tandem repeats with period of at least 40 was considered

for each *NPIP* copy. Exon 8 VNTR size was also called directly from Iso-Seq predicted ORFs by counting substrings containing “SADD” and “ISR” for the two frames of the repeat.

4.5.8 *Gene model and ORF prediction*

Iso-Seq reads were used to generate gene models on each human haplotype with PacBio Pigeon and SQANTI3 (v5.2), and ORF sequences with GeneMark (Besemer and Borodovsky 2005; Pardo-Palacios et al. 2024). Only uniquely-mapping Iso-Seq reads were used for gene model prediction, defined as a delta of at least one additional mismatch between the best-mapping paralog and second-best mapping. Mono-exonic reads were excluded. For comparison to gene models, ORFs were called directly from each *NPIP* Iso-Seq read with ANGEL (2023), keeping the longest ORF per molecule.

4.5.9 *Selection analysis*

PAV v2.4.0.1 was used to call variants for each assembled HGSVC3 (Freeze 4) haplotype relative to T2T-CHM13 v2.0 (Ebert et al. 2021). Analysis was restricted to chromosome 16, containing all but one *NPIP* paralog, and African samples (n=20) to reduce the impact of population bottlenecks in the human demography. Variants were restricted to biallelic SNPs with BCFtools (Danecek et al. 2021). Tajima’s D was estimated for sliding 30 kbp windows with VCF-kit (Cook and Andersen 2017). For comparison, Tajima’s D was called in the same way using high-coverage Illumina data for Gambian samples in the 1KG samples (n=119), restricting to 95% mappable regions as defined by the Genome in a Bottle Consortium (Dwarshuis et al. 2024). nSL was called for 30 kbp windows with selscan v2.0.2 (Szpiech 2024), for PAV (long-read) and 1KG (short-read) samples. PAV African nSL results were jointly normalized for variant frequency with 95% mappable short-read calls for Gambian (GWD) samples (parameters

--nsl --bins 100 --qbins 10 --min-snps 10 --bp-win --winsize 30000). Windows overlapping a T2T-CHM13 v2.0 *NPIP* copy by at least 5 kbp were considered valid.

4.5.10 *Protein structure prediction*

For the long exon 8 VNTR isoforms predicted with SQANTI3, protein structures were predicted with Chai-1, using MSA-free mode as *NPIP* does not have the deep homology exploited by MSA-based methods for structure prediction (Chai Discovery et al. 2024). Protein structure predictions were visualized with ChimeraX (Pettersen et al. 2021).

4.5.11 *Short-read RNA-seq expression analysis*

To quantify *NPIP* paralog-specific expression from short-read RNA-seq, reads were first aligned to T2T-CHM13 v2.0 with hisat2 (Kim et al. 2019). Jellyfish was used to find all possible 31-mers from *NPIP* gene models that were not found in the rest of the T2T-CHM13v2.0 genome (Marçais and Kingsford 2011). The uniqueness of each k-mer was classified by the number of T2T-CHM13 *NPIP* paralogs in which it was found, and paralogs were iteratively merged to form detectable paralog groups until each group contained at least five uniquely identifying k-mer positions. A custom script was then used to count each identifying k-mer with each RNA-seq read and classify reads by paralog group.

4.5.12 *Visualization*

SV and phylogenetic visualizations were created with SVbyEye, archaeopteryx, augur, MEGA, and augur/auspice (Huddleston et al. 2021; Tamura et al. 2021; Porubsky et al. 2024).

4.5.13 *Timetree analysis*

A timetree was inferred by applying the RelTime method in MEGA11 (Tamura et al. 2012, 2018, 2021) to the neutral *NPIP* phylogenetic tree, including human sequences from CHM13, CHM1, GRCh38, HG002, and PNG15, along with nonhuman primate sequences from the primary *Pan troglodytes*, *Pan paniscus*, *Gorilla gorilla*, *Pongo pygmaeus*, *Pongo abelii*, and *Siamang syndactylus* haplotypes from the T2T Primate Project (v1.1). Branch lengths were calculated using the Maximum Likelihood (ML) method and the Tamura-Nei substitution model (Tamura and Nei 1993). The timetree was calibrated with *Homo-Pongo* divergence set as 16 mya. The estimated log likelihood value of the tree is -106,468.55. There were a total of 15,985 positions in the final dataset.

4.5.14 *Probe design, cDNA generation, enrichment, and sequencing*

Biotinylated oligonucleotide probes targeting *NPIP* (Supplementary Table S3) were designed as described in Dougherty 2018. Briefly, probes were designed to target constitutive exons for subfamilies A and B (exons 2, 3, 5, 6, and 7), avoiding repeat-masked sequence. 5' biotinylated sense strand oligonucleotides were synthesized by IDT for *NPIP* enrichment.

cDNA were generated using the Clontech SMARTer PCR cDNA Synthesis Kit for CHM1 (BioSample SAMN02205338), adult brain (Clontech catalog no. 636102), fetal brain (Clontech catalog no. 636106), heart (Takara catalog no. 636532, lot 1902102A), lung (Origene sample ID FR5B3386C1), ovary (Origene sample ID FR00027E9B), thymus (Origene sample ID FR5B338054), and testis (Takara catalog no. 636533 lot 1402004; BioSample SAMN15935045). For fetal brain and testis samples, cDNA were also generated using the TeloPrime Full-Length

cDNA Amplification Kit V2 (Lexogen), which aims to avoid generating truncating cDNA by requiring the 5' mRNA cap.

Unenriched polyA cDNA were sequenced from heart, lung, ovary, and thymus samples, which were barcoded and pooled on a PacBio Sequel II SMRT cell with 30 hour movie time and two-hour pre-extension.

Hybridization capture was performed on cDNA from the remaining tissues using the biotinylated *NPIP* probes, as described in Dougherty 2018. A single Sequel SMRT cell was used for each of CHM1 and adult brain, with the remaining samples barcoded and pooled for sequencing.

Iso-Seq data from previous publications and public data depositions was obtained from ANVIL, ENCODE, and SRA, as referenced in Supplementary Table S1, and analyzed together with our generated FLNC data (Zhang et al. 2014; Zook et al. 2016; Sun et al. 2021; Kim et al. 2022; Caballero et al. 2022; Miller et al. 2022; Abood et al. 2023; Reese et al. 2023; Cheung et al. 2023; Rybak-Wolf et al. 2023; Schertzer et al. 2023; D et al. 2023; Garza et al. 2023; Maeng et al. 2023; Shimada et al. 2024).

CHAPTER 5. SUMMARY AND FUTURE DIRECTIONS

5.1 IMPLICATIONS OF RESULTS

5.1.1 *Population-scale variation in segmental duplications in highly contiguous assemblies*

In Chapter 2, by creating a pangenome representation of the SDs in 170 human haplotypes, we observed several important features of variation across the population. At this point, the more genomes we sequence, the less amount of segmental duplication appears "fixed", and the more rare and common polymorphic SVs we observe, i.e., the discovery curve is not saturated (Fig. 2.2). With the idea that functionally important SD genes are more likely to be fixed, additional samples will provide additional power to identify the most important loci. The least variable gene families include *HYDIN2* with its newly gained neural expression (Dougherty et al. 2017) and the *RGPD* family, which we found to be under selection (Mao … Dishuck et al. 2024). We observed fixed copy number for five of the 27 *GOLGA* paralogs, which mediate pathogenic microduplications and microdeletions at 15q11-q13, 15q24, and 15q25 causing forms of intellectual delay, including Prader-Willi syndrome (Amos-Landgraf et al. 1999; El-Hattab et al. 2009; Mefford et al. 2012; Antonacci et al. 2014; Paparella et al. 2023). African genomes showed higher intrachromosomal SD content and higher average copy number for 13/16 gene families with shifted distributions compared to out-of-African samples, further highlighting the necessity of sequencing African individuals to fully represent the breadth of genetic diversity. By aligning full-length cDNA reads to each assembled haplotype, we identified and created gene models for 260 putative novel protein-coding genes, including truncated and fusion genes at an SV of the *MAPT* (tau) locus, and a KRAB zinc-finger gene present in 36% of assembled haplotypes and only 69% identity to its best-matching annotated human gene. Future functional

studies of human disease should consider the diversity of copy number variable genes as a possible source of disease susceptibility, requiring a pangenome approach – as in this study, many high quality assemblies can be used as multiple references to represent human diversity and ameliorate reference bias, or alternatively, new graph-based methods may emerge to enable efficient interrogation of the genome without reference bias.

5.1.2 *SUNK-based validation of HiFi assemblies with orthogonal ultra-long ONT-sequencing*

GAVISUNK, the method I created to validate HiFi assemblies with orthogonal ultra-long ONT using unique k-mers, detects misassemblies invisible to other approaches. The false deletion of two amylase genes in HG02723 was invisible by read depth, as the assembler produced orphan contigs in addition to creating a misjoin on the main contig. Studies that depend on HiFi assembly accuracy, particularly in high copy number duplications, should be aware of this potential for misassembly and apply validation methods. We applied GAVISUNK to additional genomes and regions to be more confident of the validity of our results (Vollger, Dishuck et al. 2023; Fornezza … Dishuck et al. 2024; Mao … Dishuck et al. 2024). Since the release of GAVISUNK, assembly methods that incorporate both HiFi and ultra-long ONT have been developed (Rautiainen et al. 2023; Cheng et al. 2024). These methods correctly resolve more of the difficult-to-assemble SDs, but by using both HiFi and ONT in the assembly itself, the orthogonality is lost and validation becomes circular. I encourage researchers studying difficult regions to remember the importance of assembly validation, and perhaps hold out some of their sequence data from the assembler to confirm their assemblies' validity.

5.1.3 *Structural variation, selection, and brain-enriched expression of the NPIP gene family*

The *NPIP* gene family is an interesting subject for more detailed analyses, as it has a strong positive selection signature in the African ape lineage yet an unknown function (Johnson et al. 2001; Cantsilieris et al. 2020). Its copy number varies from 20-30 copies per human haplotype, and by examining 169 highly contiguous assemblies, I was able to distinguish paralog-specific copy number variation and the SVs associated with particular *NPIP* copies. Only three paralogs, *NPIPB2*, *B11*, and *B14*, had fixed copy numbers, and all but one locus, *B2*, showed structural variation in the form of inversions, duplications, and deletions (Fig. 4.2A). Rampant IGC means that the same paralog can be present in non-syntenic locations across samples, and we observed a strong directionality bias with the ancestral paralog *NPIPA1* often serving as an IGC donor, but never as an acceptor, a possible indication of functional importance. We were able to apply site-frequency spectrum and haplotype-based tests of selection within SDs, as contiguous assemblies allow confident alignment of orthologous regions between haplotypes. We find evidence of positive selection and recent selective sweeps for at least two *NPIP* loci, *B9* and *B15*. Similar methods can be applied genome-wide and for additional populations as the number of LRS genomes increases. Analyses of full-length cDNA reads reveal rapid changes to the gene model between paralogs, including a transmembrane domain and an expanded VNTR encoding a variably sized beta helix that has enriched brain expression, while testis-enriched paralogs use an alternative start site that encodes a signal peptide. I also developed a unique k-mer approach to assign short-read RNA-seq reads to subgroups of paralogs instead of individual paralogs, confirming testis and brain enrichment of particular paralogs. These methods could be extended in a straightforward manner to single-cell RNA-seq data to look for enrichment in particular cell types and for other genes.

The variable rate of IGC that we observe between paralogs hints at a partial resolution to the so-called "Ohno's dilemma": that a new gene must avoid the more common loss-of-function mutations on the way to acquiring a novel function that can be selected, yet continuous selection for the original function would constrain the new copy's evolutionary space (Bergthorsson et al. 2007). IGC means that a duplicate gene can accumulate mutations freely, and crucially, reversibly. Even loss-of-function mutations will be reverted periodically, allowing for selection on the resurrected paralog in a sort of punctuated, patchwork evolution. The preference for interspersed rather than tandem duplications in the African great ape lineage may modulate the IGC rate and provide a mechanism for paralogs to escape concerted evolution by relocating, and facilitate the development of diverse cis-regulatory elements at the flanks of insertions as their duplication architecture differs.

5.2 FUTURE DIRECTIONS FOR THE FIELD

5.2.1 *Experimental paralogy: functional interrogation and annotation*

A significant caveat of my expression results is that I did not know the genotypes of the tissue samples I used for expression analysis but could merely infer the expression of specific paralogs by sequence identity. With matched samples, we could address a few additional possibilities of paralog expression. For copy number variable paralogs, does gene copy number increase expression linearly, or are there regulatory mechanisms for some paralogs that buffer expression levels? Put another way, if a particular paralog is lost, is another copy upregulated to compensate for it? Experimental evidence in zebrafish has demonstrated the existence of a nonsense-mediated decay pathway-dependent mechanism for truncating mutations to activate the

expression of paralogs, compensating for the loss-of-function mutation (El-Brolosy et al. 2019; Ma et al. 2019). How do SVs alter expression of nearby genes and chromatin architecture? An NIH Common Fund initiative, Somatic Mosaicism across Human Tissues (SMaHT), name notwithstanding, will begin to address this shortcoming by creating long-read donor-specific assemblies paired with long- and short-read RNA-seq for ~12 tissues from 50 donors. These data will allow us to assess whether transcriptionally buffering compensates for copy number variation between individual haplotypes for recently duplicated genes, as well as exploring whether somatic SVs affect expression, as has been suggested for L1 retrotransposons in hippocampal neurons (Upton et al. 2015).

Advances in genome editing and *in vitro* differentiation will also enable forward genetic screens of paralogs. A recent study used CRISPR-Cas9 knockout to identify synthetic lethal paralog pairs in cancer cell lines (Parrish et al. 2021). Higher copy number gene families have more opportunity for neofunctionalization, but because CRISPR-Cas9 induces double-strand breaks, this method would induce genomic instability instead of measuring the actual effect of multiple knockouts. Methods like bridge editing (Durrant et al. 2024) will enable the insertion of specific SVs onto a fixed background, while CRISPRi allows inactivating paralogs without problematic double-stranded breaks, as has been used to investigate the expression of the L1-repressor *ZNF558* in cerebral organoids (Johansson et al. 2022). Cerebral organoids and other *in vitro* tissues will allow more relevant functional readouts than have previously been possible with cell death assays or model organisms. For example, expressing an *NPIP* transgene in mouse brain had no obvious phenotype, but *NPIP*'s function may depend on human- or primate-specific brain features.

Most genes have a wealth of functional annotations (histone methylation, chromatin accessibility, etc.) available from projects like ENCODE (ENCODE Project Consortium 2012), but because of short-read mapping ambiguity, these annotations are not accurate for recent duplications. New long-read-based methods promise to fill this gap. Fiber-seq captures chromatin accessibility equivalent to DNase-seq (Stergachis et al. 2020); DiMeLo-seq maps protein-DNA interaction like ChIP-seq to reveal features like histone modifications or CENP-A binding (Altemose et al. 2022); Pore-C is the long-read equivalent of Hi-C and can detect multiway chromatin contacts even within SDs (Deshpande et al. 2022; Chen et al. 2023). Similar to the high-identity nucleolus organizer regions on the short arms of human acrocentric chromosomes that allow for ectopic recombination (Guarracino et al. 2023), it is possible that SDs play a role in nuclear organization. Could nuclear organization explain our observed variation in IGC rates and other mutational biases (Vollger, Dishuck et al. 2023)? The tools now exist to find out.

5.2.2 *Disease associations in recent duplications*

During my PhD, watching GWAS (genome-wide association study) presentation after GWAS presentation, I kept noticing variants clustering on Manhattan plots near regions that I knew to be segmentally duplicated. This happened enough times that I suspected there must be a real enrichment, and now is the time to find out. I took the NHGRI-EBI GWAS Catalog (2022-03-08 freeze), lifted to T2T-CHM13v2.0 by Rajiv McCoy. Restricting to short-read mappable regions (Dwarshuis et al. 2024), I simulated random GWAS hits and compared their distances to SDs relative to the empirical GWAS hits, with 1000 permutations. I observed a 12% enrichment for GWAS variants within 50 kbp of an SD relative to expectation ($p<0.001$, permutation test) (Fig.

5.1). Excluding HLA regions, an 8% enrichment within 50 kbp of SDs remains ($p<0.001$, permutation test), with the ten regions with highest enrichment shown in Table 5.1.

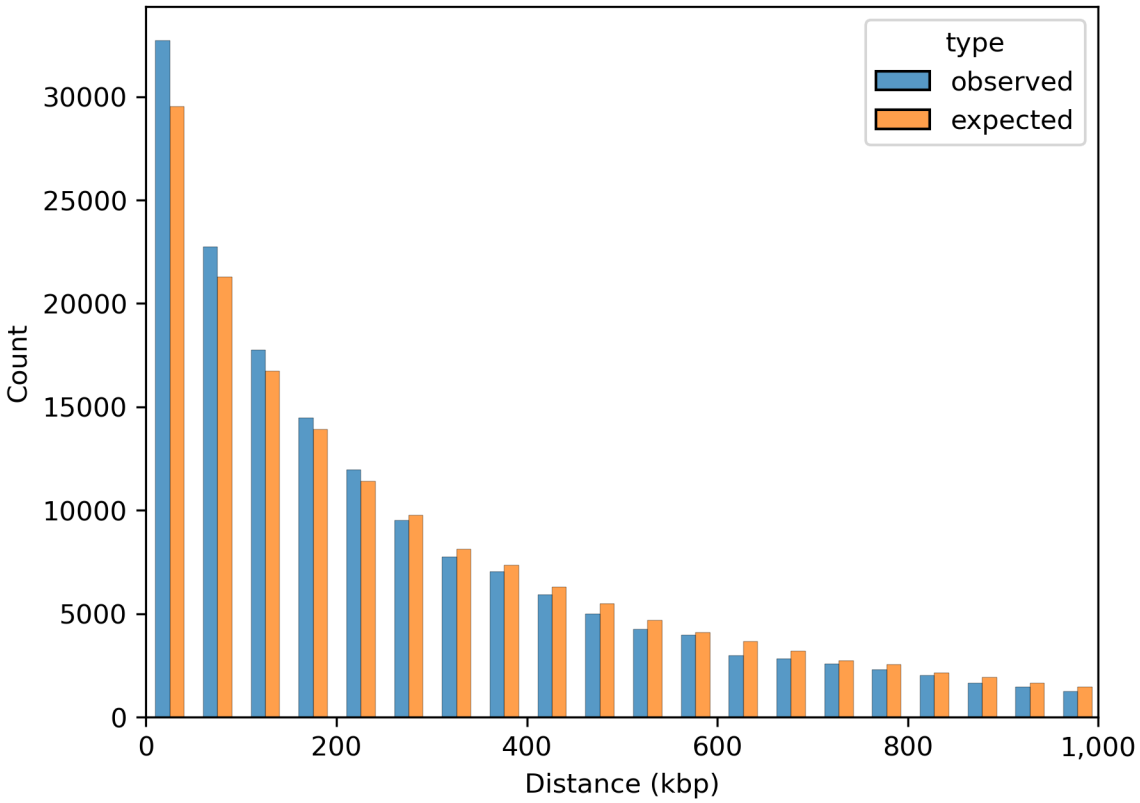


Figure 5.1. Distance of GWAS hits relative to SDs.

Empirical distances of GWAS variants to the nearest SD in T2T-CHM13v2 (SEDEF) compared to random expectation within mappable regions.

Table 5.1. GWAS-enriched SD regions

Chromosome	Start	End	Length	*Genes*	Accessible bp	GWAS hits	Accessible bp/hit
chr11	61,797,533	61,898,875	101,342	*FADS1, FADS2, FADS3, RAB3IL1*	51,913	56	927.0
chr4	147,839,600	147,940,359	100,759	*AC098588.3*	99,029	101	980.5
chr14	57,892,072	57,993,625	101,553	*SGPP1, SYNE2*	81,202	74	1097.3
chr11	0	287,999	287,999	*WASHC1, OR4F4, BET1L, SCGB1C1, ODF3, RIC8A, SIRT3, PSMD13*	35,498	29	1224.1
chr11	102,808,072	102,909,234	101,162	*WTAPP1, MMP12, BOLA3P1*	95,606	78	1225.7
chr16	14,916,215	15,081,962	165,747	*NPIPA1, PDXDC1, NTAN1, RRN3*	24,815	20	1240.8
chr16	94,748,786	94,850,519	101,733	*SNAI3, RNF166, CTU2, PIEZO1*	9,922	7	1417.4
chr19	43,579,652	43,697,941	118,289	*SNRP1, MIA, RAB4B, EGLN2, CYP2T1P, CYP2F2P, CYP2A6, CYP2A7*	37,010	26	1423.5
chr4	102,369,634	102,474,856	105,222	*METAP1, ADH5, ADH4*	90,701	63	1439.7
chr2	113,356,858	113,458,088	101,230	*IL36G, IL36A, IL36B*	97,340	66	1474.8

I suspect this enrichment is due to a combination of factors: the increased mutation rate we observed in SDs (Vollger, Dishuck et al. 2023), increased exon density in duplicated regions (She et al. 2006), suppressed recombination at inversions that are enriched near SDs allowing the accumulation of mutations, variants in linkage disequilibrium with unobserved SVs (serving as tagging SNPs), and perhaps uncontrolled confounders. Whatever the source, this enrichment gives me hope that our expectation that SDs hold many cryptic disease-associated variants will hold true.

For now, though, large cohorts have only short-read or array data, so we must infer SVs instead of directly observing them. As we assemble large cohorts of long-read assemblies that capture most common SVs, we can associate short-read genotypes with SVs using methods like Pangenie and Locityper (Ebler et al. 2022; Prodanov et al. 2024). I am particularly interested in replicating the obesity and asthma associations previously reported for the 16p11.2 inversion mediated by *NPIP* (González et al. 2014), as well as investigating whether the novel KRAB-ZNF gene that I described in Chapter 2 has any disease associations.

5.3 CLOSING THOUGHTS

The dramatic changes in *NPIP* gene models and extreme variation across humans should serve as a reminder of the evolutionary dynamism possible within SDs. Similarly, the 201 putatively novel expressed genes I discovered with an analysis of just 170 haplotypes is indicative of underappreciated levels of functional diversity in the human population. Emerging methods will allow detailed annotation and *in vitro* validation of the functions of recently duplicated genes, shedding light on cryptic sources of disease and phenotypic variation. The maturation of long-read sequencing will imminently enable routine telomere-to-telomere genome sequencing and

assembly, but understanding all the variation that we are bound to discover will undoubtedly reveal a new set of challenges for the future of genomics.

BIBLIOGRAPHY

1000 Genomes Project Consortium, Abecasis GR, Auton A, et al (2012) An integrated map of genetic variation from 1,092 human genomes. Nature 491:56–65. https://doi.org/10.1038/nature11632

1000 Genomes Project Consortium, Auton A, Brooks LD, et al (2015) A global reference for human genetic variation. Nature 526:68–74. https://doi.org/10.1038/nature15393

Abood A, Mesner LD, Jeffery ED, et al (2023) Long-read proteogenomics to connect disease-associated sQTLs to the protein isoform effectors of disease. bioRxiv 2023.03.17.531557. https://doi.org/10.1101/2023.03.17.531557

Alonge M, Lebeigle L, Kirsche M, et al (2022) Automated assembly scaffolding using RagTag elevates a new tomato system for high-throughput genome editing. Genome Biol 23:258. https://doi.org/10.1186/s13059-022-02823-7

Altemose N, Maslan A, Smith OK, et al (2022) DiMeLo-seq: a long-read, single-molecule method for mapping protein–DNA interactions genome wide. Nat Methods 19:711–723. https://doi.org/10.1038/s41592-022-01475-6

Altschul SF, Gish W, Miller W, et al (1990) Basic local alignment search tool. J Mol Biol 215:403–410. https://doi.org/10.1016/S0022-2836(05)80360-2

Amemiya HM, Kundaje A, Boyle AP (2019) The ENCODE Blacklist: Identification of Problematic Regions of the Genome. Sci Rep 9:9354. https://doi.org/10.1038/s41598-019-45839-z

Amos-Landgraf JM, Ji Y, Gottlieb W, et al (1999) Chromosome breakage in the Prader-Willi and Angelman syndromes involves recombination between large, transcribed repeats at proximal and distal breakpoints. Am J Hum Genet 65:370–386. https://doi.org/10.1086/302510

Anisimova M, Gil M, Dufayard J-F, et al (2011) Survey of Branch Support Methods Demonstrates Accuracy, Power, and Robustness of Fast Likelihood-based Approximation Schemes. Systematic Biology 60:685–699. https://doi.org/10.1093/sysbio/syr041

Antonacci F, Dennis MY, Huddleston J, et al (2014) Palindromic GOLGA8 core duplicons promote chromosome 15q13.3 microdeletion and evolutionary instability. Nat Genet 46:1293–1302. https://doi.org/10.1038/ng.3120

Antonacci F, Kidd JM, Marques-Bonet T, et al (2010) A large, complex structural polymorphism at 16p12.1 underlies microdeletion disease risk. Nat Genet 42:745–750. https://doi.org/10.1038/ng.643

Bailey JA, Eichler EE (2006) Primate segmental duplications: crucibles of evolution, diversity and disease. Nat Rev Genet 7:552–564. https://doi.org/10.1038/nrg1895

Bailey JA, Gu Z, Clark RA, et al (2002) Recent Segmental Duplications in the Human Genome. Science 297:1003–1007. https://doi.org/10.1126/science.1072047

Bailey JA, Yavor AM, Massa HF, et al (2001) Segmental duplications: organization and impact within the current human genome project assembly. Genome Res 11:1005–1017. https://doi.org/10.1101/gr.gr-1871r

Ballif BC, Hornor SA, Jenkins E, et al (2007) Discovery of a previously unrecognized microdeletion syndrome of 16p11.2–p12.2. Nat Genet 39:1071–1073. https://doi.org/10.1038/ng2107

Barber JCK, Hall V, Maloney VK, et al (2012) 16p11.2–p12.2 duplication syndrome; a genomic condition differentiated from euchromatic variation of 16p11.2. European Journal of Human Genetics 21:182. https://doi.org/10.1038/ejhg.2012.144

Bekpen C, Baker C, Hebert MD, et al (2017) Functional Characterization of the Morpheus Gene Family. https://doi.org/10.1101/116087

Bekpen C, Tastekin I, Siswara P, et al (2012) Primate segmental duplication creates novel promoters for the LRRC37 gene family within the 17q21.31 inversion polymorphism region. Genome Res 22:1050–1058. https://doi.org/10.1101/gr.134098.111

Benson G (1999) Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Research 27:573–580. https://doi.org/10.1093/nar/27.2.573

Bergthorsson U, Andersson DI, Roth JR (2007) Ohno's dilemma: Evolution of new genes under continuous selection. Proceedings of the National Academy of Sciences 104:17004–17009. https://doi.org/10.1073/pnas.0707158104

Besemer J, Borodovsky M (2005) GeneMark: web software for gene finding in prokaryotes, eukaryotes and viruses. Nucleic Acids Res 33:W451-454. https://doi.org/10.1093/nar/gki487

Bochukova EG, Huang N, Keogh J, et al (2010) Large, rare chromosomal deletions associated with severe early-onset obesity. Nature 463:666–670. https://doi.org/10.1038/nature08689

Boettger LM, Handsaker RE, Zody MC, McCarroll SA (2012) Structural haplotypes and recent evolution of the human 17q21.31 region. Nat Genet 44:881–885. https://doi.org/10.1038/ng.2334

Bogdanova N, Markoff A, Gerke V, et al (2001) Homologues to the first gene for autosomal dominant polycystic kidney disease are pseudogenes. Genomics 74:333–341. https://doi.org/10.1006/geno.2001.6568

Bohnenkämper L (2024) Recombinations, chains and caps: resolving problems with the DCJ-indel model. Algorithms for Molecular Biology 19:8. https://doi.org/10.1186/s13015-024-00253-7

Bowles KR, Pugh DA, Liu Y, et al (2022) 17q21.31 sub-haplotypes underlying H1-associated risk for Parkinson's disease are associated with LRRC37A/2 expression in astrocytes. Molecular Neurodegeneration 17:48. https://doi.org/10.1186/s13024-022-00551-x

Byrska-Bishop M, Evani US, Zhao X, et al (2022) High-coverage whole-genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. Cell 185:3426-3440.e19. https://doi.org/10.1016/j.cell.2022.08.004

Caballero M, Ge T, Rebelo AR, et al (2022) Comprehensive analysis of DNA replication timing across 184 cell lines suggests a role for MCM10 in replication timing regulation. Hum Mol Genet 31:2899–2917. https://doi.org/10.1093/hmg/ddac082

Campbell MC, Tishkoff SA (2008) African genetic diversity: implications for human demographic history, modern human origins, and complex disease mapping. Annu Rev Genomics Hum Genet 9:403–433. https://doi.org/10.1146/annurev.genom.9.081307.164258

Cantsilieris S, Sunkin SM, Johnson ME, et al (2020) An evolutionary driver of interspersed segmental duplications in primates. Genome Biol 21:202. https://doi.org/10.1186/s13059-020-02074-4

Cardoso-Moreira M, Halbert J, Valloton D, et al (2019) Gene expression across mammalian organ development. Nature 571:505–509. https://doi.org/10.1038/s41586-019-1338-5

Cartney AMM, Shafin K, Alonge M, et al (2022) Chasing perfection: validation and polishing strategies for telomere-to-telomere genome assemblies. Nature methods 19:687. https://doi.org/10.1038/s41592-022-01440-3

Chai Discovery, Boitreaud J, Dent J, et al (2024) Chai-1: Decoding the molecular interactions of life

Chaisson MJP, Huddleston J, Dennis MY, et al (2015) Resolving the complexity of the human genome using single-molecule sequencing. Nature 517:608–611. https://doi.org/10.1038/nature13907

Charlesworth B, Barton NH (2018) The Spread of an Inversion with Migration and Selection. Genetics 208:377–382. https://doi.org/10.1534/genetics.117.300426

Chen Y, Lin Z-B, Wang S-K, et al (2023) High-resolution diploid 3D genome reconstruction using Pore-C data. 2023.08.29.555243

Cheng H, Asri M, Lucas J, et al (2024) Scalable telomere-to-telomere assembly for diploid and polyploid genomes with double graph. Nat Methods 21:967–970. https://doi.org/10.1038/s41592-024-02269-8

Cheng H, Concepcion GT, Feng X, et al (2021) Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. Nat Methods 18:170–175. https://doi.org/10.1038/s41592-020-01056-5

Cheung WA, Johnson AF, Rowell WJ, et al (2023) Direct haplotype-resolved 5-base HiFi sequencing for genome-wide profiling of hypermethylation outliers in a rare disease cohort. Nat Commun 14:3090. https://doi.org/10.1038/s41467-023-38782-1

Chiang C, Scott AJ, Davis JR, et al (2017) The impact of structural variation on human gene expression. Nature genetics 49:692. https://doi.org/10.1038/ng.3834

Church DM (2022) A next-generation human genome sequence. Science 376:34–35. https://doi.org/10.1126/science.abo5367

Clarke R, Peden JF, Hopewell JC, et al (2009) Genetic variants associated with Lp(a) lipoprotein level and coronary disease. N Engl J Med 361:2518–2528. https://doi.org/10.1056/NEJMoa0902604

Coe BP, Stessman HAF, Sulovari A, et al (2019) Neurodevelopmental disease genes implicated by de novo mutation and copy number variation morbidity. Nat Genet 51:106–116. https://doi.org/10.1038/s41588-018-0288-4

Cook DE, Andersen EC (2017) VCF-kit: assorted utilities for the variant call format. Bioinformatics 33:1581–1582. https://doi.org/10.1093/bioinformatics/btx011

Cooper GM, Coe BP, Girirajan S, et al (2011) A copy number variation morbidity map of developmental delay. Nat Genet 43:838–846. https://doi.org/10.1038/ng.909

D T, Nj F, Y K, et al (2023) Isoform-resolved transcriptome of the human preimplantation embryo. Nature communications 14:. https://doi.org/10.1038/s41467-023-42558-y

Danecek P, Bonfield JK, Liddle J, et al (2021) Twelve years of SAMtools and BCFtools. Gigascience 10:giab008. https://doi.org/10.1093/gigascience/giab008

de Kovel CGF, Trucks H, Helbig I, et al (2010) Recurrent microdeletions at 15q11.2 and 16p13.11 predispose to idiopathic generalized epilepsies. Brain 133:23–32. https://doi.org/10.1093/brain/awp262

Dennis MY, Nuttle X, Sudmant PH, et al (2012) Evolution of Human-Specific Neural *SRGAP2* Genes by Incomplete Segmental Duplication. Cell 149:912–922. https://doi.org/10.1016/j.cell.2012.03.033

Deshpande AS, Ulahannan N, Pendleton M, et al (2022) Identifying synergistic high-order 3D chromatin conformations from genome-scale nanopore concatemer sequencing. Nat Biotechnol 40:1488–1499. https://doi.org/10.1038/s41587-022-01289-z

Dishuck PC, Rozanski AN, Logsdon GA, et al (2022) GAVISUNK: genome assembly validation via inter-SUNK distances in Oxford Nanopore reads. Bioinformatics 39:btac714. https://doi.org/10.1093/bioinformatics/btac714

Dougherty ML, Nuttle X, Penn O, et al (2017) The birth of a human-specific neural gene by incomplete duplication and gene fusion. Genome Biol 18:49. https://doi.org/10.1186/s13059-017-1163-9

Dougherty ML, Underwood JG, Nelson BJ, et al (2018) Transcriptional fates of human-specific segmental duplications in brain. Genome Res 28:1566–1576. https://doi.org/10.1101/gr.237610.118

Dudas G (2024) evogytis/baltic

Dulai KS, von Dornum M, Mollon JD, Hunt DM (1999) The evolution of trichromatic color vision by opsin gene duplication in New World and Old World primates. Genome Res 9:629–638

Durrant MG, Perry NT, Pai JJ, et al (2024) Bridge RNAs direct programmable recombination of target and donor DNA. Nature 630:984–993. https://doi.org/10.1038/s41586-024-07552-4

Dwarshuis N, Kalra D, McDaniel J, et al (2024) The GIAB genomic stratifications resource for human reference genomes. Nat Commun 15:9029. https://doi.org/10.1038/s41467-024-53260-y

Ebert P, Audano PA, Zhu Q, et al (2021) Haplotype-resolved diverse human genomes and integrated analysis of structural variation. Science 372:eabf7117. https://doi.org/10.1126/science.abf7117

Ebler J, Ebert P, Clarke WE, et al (2022) Pangenome-based genome inference allows efficient and accurate genotyping across a wide spectrum of variant classes. Nat Genet 54:518–525. https://doi.org/10.1038/s41588-022-01043-w

Edge P, Bafna V, Bansal V (2016) HapCUT2: robust and accurate haplotype assembly for diverse sequencing technologies. Genome Res gr.213462.116. https://doi.org/10.1101/gr.213462.116

Eichler E (1997) Interchromosomal duplications of the adrenoleukodystrophy locus: a phenomenon of pericentromeric plasticity. Human Molecular Genetics 6:991–1002. https://doi.org/10.1093/hmg/6.7.991

El-Brolosy MA, Kontarakis Z, Rossi A, et al (2019) Genetic compensation triggered by mutant mRNA degradation. Nature 568:193–197. https://doi.org/10.1038/s41586-019-1064-z

El-Hattab AW, Smolarek TA, Walker ME, et al (2009) Redefined genomic architecture in 15q24 directed by patient deletion/duplication breakpoint mapping. Hum Genet 126:589–602. https://doi.org/10.1007/s00439-009-0706-x

ENCODE Project Consortium (2012) An integrated encyclopedia of DNA elements in the human genome. Nature 489:57–74. https://doi.org/10.1038/nature11247

Ferrer-Admetlla A, Liang M, Korneliussen T, Nielsen R (2014) On Detecting Incomplete Soft or Hard Selective Sweeps Using Haplotype Structure. Molecular Biology and Evolution 31:1275–1291. https://doi.org/10.1093/molbev/msu077

Fiddes IT, Lodewijk GA, Mooring M, et al (2018) Human-Specific NOTCH2NL Genes Affect Notch Signaling and Cortical Neurogenesis. Cell 173:1356-1369.e22. https://doi.org/10.1016/j.cell.2018.03.051

Fiddes IT, Pollen AA, Davis JM, Sikela JM (2019) Paired involvement of human-specific Olduvai domains and NOTCH2NL genes in human brain evolution. Human Genetics 138:715–721. https://doi.org/10.1007/s00439-019-02018-4

Fischer J, Ortuño EF, Marsoner F, et al (2022) Human-specific ARHGAP11B ensures human-like basal progenitor levels in hominid cerebral organoids. EMBO Reports 23:e54728. https://doi.org/10.15252/embr.202254728

Fitzgerald T, Birney E (2022) CNest: A novel copy number association discovery method uncovers 862 new associations from 200,629 whole-exome sequence datasets in the UK Biobank. Cell Genom 2:100167. https://doi.org/10.1016/j.xgen.2022.100167

Florio M, Albert M, Taverna E, et al (2015) Human-specific gene ARHGAP11B promotes basal progenitor amplification and neocortex expansion. Science 347:1465–1470. https://doi.org/10.1126/science.aaa1975

Force A, Lynch M, Pickett FB, et al (1999) Preservation of duplicate genes by complementary, degenerative mutations. Genetics 151:1531–1545

Fornezza S, Delvecchio VS, Harvey WT, et al (2024) AGAP duplicons associate with structural diversity at Chromosome 10q11.22. Genome Res 34:1487–1499. https://doi.org/10.1101/gr.279454.124

Frankish A, Carbonell-Sala S, Diekhans M, et al (2023) GENCODE: reference annotation for the human and mouse genomes in 2023. Nucleic Acids Res 51:D942–D949. https://doi.org/10.1093/nar/gkac1071

Garza R, Atacho DAM, Adami A, et al (2023) LINE-1 retrotransposons drive human neuronal transcriptome complexity and functional diversification. Sci Adv 9:eadh9543. https://doi.org/10.1126/sciadv.adh9543

Giannuzzi G, Siswara P, Malig M, et al (2013) Evolutionary dynamism of the primate LRRC37 gene family. Genome Res 23:46–59. https://doi.org/10.1101/gr.138842.112

Girirajan S, Rosenfeld JA, Cooper GM, et al (2010) A recurrent 16p12.1 microdeletion supports a two-hit model for severe developmental delay. Nat Genet 42:203–209. https://doi.org/10.1038/ng.534

González JR, Cáceres A, Esko T, et al (2014) A Common 16p11.2 Inversion Underlies the Joint Susceptibility to Asthma and Obesity. American Journal of Human Genetics 94:361. https://doi.org/10.1016/j.ajhg.2014.01.015

Groot PC, Bleeker MJ, Pronk JC, et al (1989) The human alpha-amylase multigene family consists of haplotypes with variable numbers of genes. Genomics 5:29–42. https://doi.org/10.1016/0888-7543(89)90083-9

GTEx Consortium (2013) The Genotype-Tissue Expression (GTEx) project. Nat Genet 45:580–585. https://doi.org/10.1038/ng.2653

Guarracino A, Buonaiuto S, de Lima LG, et al (2023) Recombination between heterologous human acrocentric chromosomes. Nature 617:335–343. https://doi.org/10.1038/s41586-023-05976-y

Guitart X, Porubsky D, Yoo D, et al (2024) Independent expansion, selection and hypervariability of the TBC1D3 gene family in humans. Genome Res gr.279299.124. https://doi.org/10.1101/gr.279299.124

Hahn MW, Demuth JP, Han S-G (2007) Accelerated Rate of Gene Gain and Loss in Primates. Genetics 177:1941–1949. https://doi.org/10.1534/genetics.107.080077

Hallast P, Ebert P, Loftus M, et al (2023) Assembly of 43 human Y chromosomes reveals extensive complexity and variation. Nature 621:355–364. https://doi.org/10.1038/s41586-023-06425-6

Hallgren J, Tsirigos KD, Pedersen MD, et al (2022) DeepTMHMM predicts alpha and beta transmembrane proteins using deep neural networks. 2022.04.08.487609

Hannes FD, Sharp AJ, Mefford HC, et al (2009) Recurrent reciprocal deletions and duplications of 16p13.11: the deletion is a risk factor for MR/MCA while the duplication may be a rare benign variant. J Med Genet 46:223–232. https://doi.org/10.1136/jmg.2007.055202

Hargreaves CE, Rose-Zerilli MJJ, Machado LR, et al (2015) Fcγ receptors: genetic variation, function, and disease. Immunol Rev 268:6–24. https://doi.org/10.1111/imr.12341

Harris CR, Millman KJ, van der Walt SJ, et al (2020) Array programming with NumPy. Nature 585:357–362. https://doi.org/10.1038/s41586-020-2649-2

Heinzen EL, Radtke RA, Urban TJ, et al (2010) Rare Deletions at 16p13.11 Predispose to a Diverse Spectrum of Sporadic Epilepsy Syndromes. Am J Hum Genet 86:707–718. https://doi.org/10.1016/j.ajhg.2010.03.018

Hellwege JN, Velez Edwards DR, Giri A, et al (2019) Mapping eGFR loci to the renal transcriptome and phenome in the VA Million Veteran Program. Nat Commun 10:3842. https://doi.org/10.1038/s41467-019-11704-w

Hoang DT, Chernomor O, von Haeseler A, et al (2018) UFBoot2: Improving the Ultrafast Bootstrap Approximation. Molecular Biology and Evolution 35:518–522. https://doi.org/10.1093/molbev/msx281

Hsieh P, Vollger MR, Dang V, et al (2019) Adaptive archaic introgression of copy number variants and the discovery of previously unknown human genes. Science 366:eaax2083. https://doi.org/10.1126/science.aax2083

Huddleston J, Hadfield J, Sibley TR, et al (2021) Augur: a bioinformatics toolkit for phylogenetic analyses of human pathogens. Journal of open source software 6:2906. https://doi.org/10.21105/joss.02906

Hujoel MLA, Handsaker RE, Sherman MA, et al (2024) Protein-altering variants at copy number-variable regions influence diverse human phenotypes. Nat Genet 56:569–578. https://doi.org/10.1038/s41588-024-01684-z

Ihnatovych I, Saddler R-A, Sule N, Szigeti K (2024) Translational implications of CHRFAM7A, an elusive human-restricted fusion gene. Mol Psychiatry. https://doi.org/10.1038/s41380-023-02389-1

Ingason A, Rujescu D, Cichon S, et al (2011) Copy number variations of chromosome 16p13.1 region associated with schizophrenia. Mol Psychiatry 16:17–25. https://doi.org/10.1038/mp.2009.101

International Human Genome Sequencing Consortium (2001) Initial sequencing and analysis of the human genome. Nature 409:860–921. https://doi.org/10.1038/35057062

Jain M, Koren S, Miga KH, et al (2018) Nanopore sequencing and assembly of a human genome with ultra-long reads. Nat Biotechnol 36:338–345. https://doi.org/10.1038/nbt.4060

Jarvis ED, Formenti G, Rhie A, et al (2022) Semi-automated assembly of high-quality diploid human reference genomes. Nature 611:519–531. https://doi.org/10.1038/s41586-022-05325-5

Jiang Z, Hubley R, Smit A, Eichler EE (2008) DupMasker: A tool for annotating primate segmental duplications. Genome Res 18:1362–1368. https://doi.org/10.1101/gr.078477.108

Jiang Z, Tang H, Ventura M, et al (2007) Ancestral reconstruction of segmental duplications reveals punctuated cores of human genome evolution. Nat Genet 39:1361–1368. https://doi.org/10.1038/ng.2007.9

Johansson PA, Brattås PL, Douse CH, et al (2022) A cis-acting structural variation at the ZNF558 locus controls a gene regulatory network in human brain development. Cell Stem Cell 29:52-69.e8. https://doi.org/10.1016/j.stem.2021.09.008

Johnson KE, Voight BF (2018) Patterns of shared signatures of recent positive selection across human populations. Nat Ecol Evol 2:713–720. https://doi.org/10.1038/s41559-018-0478-6

Johnson ME, National Institute of Health Intramural Sequencing Center Comparative Sequencing Program, Cheng Z, et al (2006) Recurrent duplication-driven transposition of DNA during hominoid evolution. Proc Natl Acad Sci USA 103:17626–17631. https://doi.org/10.1073/pnas.0605426103

Johnson ME, Viggiano L, Bailey JA, et al (2001) Positive selection of a gene family during the emergence of humans and African apes. Nature 413:514–519. https://doi.org/10.1038/35097067

Ju X-C, Hou Q-Q, Sheng A-L, et al (2016) The hominoid-specific gene TBC1D3 promotes generation of basal neural progenitors and induces cortical folding in mice. Elife 5:e18197. https://doi.org/10.7554/eLife.18197

Kalyaanamoorthy S, Minh BQ, Wong TKF, et al (2017) ModelFinder: fast model selection for accurate phylogenetic estimates. Nat Methods 14:587–589. https://doi.org/10.1038/nmeth.4285

Katoh K, Standley DM (2013) MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. Mol Biol Evol 30:772–780. https://doi.org/10.1093/molbev/mst010

Kim D, Paggi JM, Park C, et al (2019) Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nat Biotechnol 37:907–915. https://doi.org/10.1038/s41587-019-0201-4

Kim H, Jeon S, Kim Y, et al (2022) KOREF_S1: phased, parental trio-binned Korean reference genome using long reads and Hi-C sequencing methods. GigaScience 11:giac022. https://doi.org/10.1093/gigascience/giac022

King M-C, Wilson AC (1975) Evolution at Two Levels in Humans and Chimpanzees. Science 188:107–116. https://doi.org/10.1126/science.1090005

Kirkpatrick M, Barton N (2006) Chromosome inversions, local adaptation and speciation. Genetics 173:419–434. https://doi.org/10.1534/genetics.105.047985

Kolberg L, Raudvere U, Kuzmin I, et al (2023) g:Profiler-interoperable web service for functional enrichment analysis and gene identifier mapping (2023 update). Nucleic Acids Res 51:W207–W212. https://doi.org/10.1093/nar/gkad347

Koren S, Walenz BP, Berlin K, et al (2017) Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Res 27:722–736. https://doi.org/10.1101/gr.215087.116

Kuang S-Q, Guo D-C, Prakash SK, et al (2011) Recurrent Chromosome 16p13.1 Duplications Are a Risk Factor for Aortic Dissections. PLoS Genetics 7:e1002118. https://doi.org/10.1371/journal.pgen.1002118

Kumar RA, KaraMohamed S, Sudi J, et al (2008) Recurrent 16p11.2 microdeletions in autism. Human Molecular Genetics 17:628–638. https://doi.org/10.1093/hmg/ddm376

Li H (2018) Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics 34:3094–3100. https://doi.org/10.1093/bioinformatics/bty191

Li H (2011) A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics 27:2987–2993. https://doi.org/10.1093/bioinformatics/btr509

Liao W-W, Asri M, Ebler J, et al (2023) A draft human pangenome reference. Nature 617:312–324. https://doi.org/10.1038/s41586-023-05896-x

Loftus BJ, Kim U-J, Sneddon VP, et al (1999) Genome Duplications and Other Features in 12 Mb of DNA Sequence from Human Chromosome 16p and 16q. Genomics 60:295–308. https://doi.org/10.1006/geno.1999.5927

Logsdon GA, Vollger MR, Eichler EE (2020) Long-read human genome sequencing and its applications. Nat Rev Genet 21:597–614. https://doi.org/10.1038/s41576-020-0236-x

Logsdon GA, Vollger MR, Hsieh P, et al (2021) The structure, function and evolution of a complete human chromosome 8. Nature 593:101–107. https://doi.org/10.1038/s41586-021-03420-7

Loureiro S, Almeida J, Café C, et al (2017) Copy number variations in chromosome 16p13.11-The neurodevelopmental clinical spectrum. Current Pediatric Research

Loviglio MN, Leleu M, Männik K, et al (2017) Chromosomal contacts connect loci associated with autism, BMI and head circumference phenotypes. Mol Psychiatry 22:836–849. https://doi.org/10.1038/mp.2016.84

Lupski JR (2010) Retrotransposition and structural variation in the human genome. Cell 141:1110–1112. https://doi.org/10.1016/j.cell.2010.06.014

Lupski JR, Stankiewicz P (2005) Genomic Disorders: Molecular Mechanisms for Rearrangements and Conveyed Phenotypes. PLOS Genetics 1:e49. https://doi.org/10.1371/journal.pgen.0010049

Ma Z, Zhu P, Shi H, et al (2019) PTC-bearing mRNA elicits a genetic compensation response via Upf3a and COMPASS components. Nature 568:259–263. https://doi.org/10.1038/s41586-019-1057-y

Maeng JH, Jang HJ, Du AY, et al (2023) Using long-read CAGE sequencing to profile cryptic-promoter-derived transcripts and their contribution to the immunopeptidome. Genome Res 33:2143–2155. https://doi.org/10.1101/gr.277061.122

Mao Y, Harvey WT, Porubsky D, et al (2024) Structurally divergent and recurrently mutated regions of primate genomes. Cell S0092-8674(24)00121–1. https://doi.org/10.1016/j.cell.2024.01.052

Marçais G, Kingsford C (2011) A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics 27:764–770. https://doi.org/10.1093/bioinformatics/btr011

Marco-Sola S, Eizenga JM, Guarracino A, et al (2023) Optimal gap-affine alignment in O(s) space. Bioinformatics 39:btad074. https://doi.org/10.1093/bioinformatics/btad074

Marques-Bonet T, Kidd JM, Ventura M, et al (2009a) A burst of segmental duplications in the genome of the African great ape ancestor. Nature 457:877–881. https://doi.org/10.1038/nature07744

Marques-Bonet T, Ryder OA, Eichler EE (2009b) Sequencing primate genomes: what have we learned? Annu Rev Genomics Hum Genet 10:355–386. https://doi.org/10.1146/annurev.genom.9.081307.164420

McCarthy SE, Makarov V, Kirov G, et al (2009) Microduplications of 16p11.2 are associated with schizophrenia. Nat Genet 41:1223–1227. https://doi.org/10.1038/ng.474

Mefford HC, Rosenfeld JA, Shur N, et al (2012) Further clinical and molecular delineation of the 15q24 microdeletion syndrome. J Med Genet 49:110–118. https://doi.org/10.1136/jmedgenet-2011-100499

Miga KH, Eichler EE (2023) Envisioning a new era: Complete genetic information from routine, telomere-to-telomere genomes. Am J Hum Genet 110:1832–1840. https://doi.org/10.1016/j.ajhg.2023.09.011

Miga KH, Koren S, Rhie A, et al (2020) Telomere-to-telomere assembly of a complete human X chromosome. Nature 585:79–84. https://doi.org/10.1038/s41586-020-2547-7

Mikheenko A, Bzikadze AV, Gurevich A, et al (2020) TandemTools: mapping long reads and assessing/improving assembly quality in extra-long tandem repeats. Bioinformatics 36:i75–i83. https://doi.org/10.1093/bioinformatics/btaa440

Miller AR, Wijeratne S, McGrath SD, et al (2022) Pacific Biosciences Fusion and Long Isoform Pipeline for Cancer Transcriptome–Based Resolution of Isoform Complexity. The Journal of Molecular Diagnostics 24:1292–1306. https://doi.org/10.1016/j.jmoldx.2022.09.003

Miller JS, Westin EH, Schwartz LB (1989) Cloning and characterization of complementary DNA for human tryptase. J Clin Invest 84:1188–1195. https://doi.org/10.1172/JCI114284

Mills RE, Walter K, Stewart C, et al (2011) Mapping copy number variation by population-scale genome sequencing. Nature 470:59–65. https://doi.org/10.1038/nature09708

Minh BQ, Schmidt HA, Chernomor O, et al (2020) IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. Molecular Biology and Evolution 37:1530–1534. https://doi.org/10.1093/molbev/msaa015

Mölder F, Jablonski KP, Letcher B, et al (2021) Sustainable data analysis with Snakemake. F1000Res 10:33. https://doi.org/10.12688/f1000research.29032.2

Morgulis A, Gertz EM, Schäffer AA, Agarwala R (2006) WindowMasker: window-based masker for sequenced genomes. Bioinformatics 22:134–141. https://doi.org/10.1093/bioinformatics/bti774

Nadeau JH, Sankoff D (1997) Comparable Rates of Gene Loss and Functional Divergence after Genome Duplications Early in Vertebrate Evolution. Genetics 147:1259–1266

Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ (2015) IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. Molecular Biology and Evolution 32:268–274. https://doi.org/10.1093/molbev/msu300

Nicolle R, Siquier-Pernet K, Rio M, et al (2022) 16p13.11p11.2 triplication syndrome: a new recognizable genomic disorder characterized by optical genome mapping and whole genome sequencing. Eur J Hum Genet 30:712–720. https://doi.org/10.1038/s41431-022-01094-x

Numanagic I, Gökkaya AS, Zhang L, et al (2018) Fast characterization of segmental duplications in genome assemblies. Bioinformatics 34:i706–i714. https://doi.org/10.1093/bioinformatics/bty586

Nurk S, Koren S, Rhie A, et al (2022) The complete sequence of a human genome. Science 376:44–53. https://doi.org/10.1126/science.abj6987

Ohno S (1970) Evolution by Gene Duplication. Springer-Verlag

Pallaoro M, Fejzo MS, Shayesteh L, et al (1999) Characterization of genes encoding known and novel human mast cell tryptases on chromosome 16p13.3. J Biol Chem 274:3355–3362. https://doi.org/10.1074/jbc.274.6.3355

Paparella A, L'Abbate A, Palmisano D, et al (2023) Structural Variation Evolution at the 15q11-q13 Disease-Associated Locus. Int J Mol Sci 24:15818. https://doi.org/10.3390/ijms242115818

Pardo-Palacios FJ, Arzalluz-Luque A, Kondratova L, et al (2024) SQANTI3: curation of long-read transcriptomes for accurate identification of known and novel isoforms. Nat Methods 21:793–797. https://doi.org/10.1038/s41592-024-02229-2

Parrish PCR, Thomas JD, Gabel AM, et al (2021) Discovery of synthetic lethal and tumor suppressor paralog pairs in the human genome. Cell Reports 36:. https://doi.org/10.1016/j.celrep.2021.109597

Pegueroles C, Laurie S, Albà MM (2013) Accelerated Evolution after Gene Duplication: A Time-Dependent Process Affecting Just One Copy. Mol Biol Evol 30:1830–1842. https://doi.org/10.1093/molbev/mst083

Pendleton AL, Shen F, Taravella AM, et al (2018) Comparison of village dog and wolf genomes highlights the role of the neural crest in dog domestication. BMC Biol 16:64. https://doi.org/10.1186/s12915-018-0535-2

Perry GH, Dominy NJ, Claw KG, et al (2007) Diet and the evolution of human amylase gene copy number variation. Nat Genet 39:1256–1260. https://doi.org/10.1038/ng2123

Pettersen EF, Goddard TD, Huang CC, et al (2021) UCSF ChimeraX: Structure visualization for researchers, educators, and developers. Protein Sci 30:70–82. https://doi.org/10.1002/pro.3943

Pop-Jordanova PN, Zorcec T, Sukarova-Angelovska E (2021) Duplication of Chromosome 16p13.11-p12.3 with Different Expressions in the Same Family. Balkan Journal of Medical Genetics : BJMG 24:89. https://doi.org/10.2478/bjmg-2021-0010

Porubsky D, Guitart X, Yoo D, et al (2024) SVbyEye: A visual tool to characterize structural variation among whole genome assemblies. 2024.09.11.612418

Porubsky D, Höps W, Ashraf H, et al (2022) Recurrent inversion polymorphisms in humans associate with genetic instability and genomic disorders. Cell 185:1986-2005.e26. https://doi.org/10.1016/j.cell.2022.04.017

Porubsky D, Vollger MR, Harvey WT, et al (2023) Gaps and complex structurally variant loci in phased genome assemblies. Genome Res 33:496–510. https://doi.org/10.1101/gr.277334.122

Prodanov T, Plender EG, Seebohm G, et al (2024) Locityper: targeted genotyping of complex polymorphic genes. 2024.05.03.592358

Quintela I, Barros F, Lago-Leston R, et al (2015) A maternally inherited 16p13.11-p12.3 duplication concomitant with a de novo SOX5 deletion in a male patient with global developmental delay, disruptive and obsessive behaviors and minor dysmorphic features. American Journal of Medical Genetics Part A 167:1315–1322. https://doi.org/10.1002/ajmg.a.36909

Rahbari R, Zuccherato LW, Tischler G, et al (2017) Understanding the Genomic Structure of Copy-Number Variation of the Low-Affinity Fcγ Receptor Region Allows Confirmation of the Association of FCGR3B Deletion with Rheumatoid Arthritis. Hum Mutat 38:390–399. https://doi.org/10.1002/humu.23159

Ramalingam A, Zhou X-G, Fiedler SD, et al (2011) 16p13.11 duplication is a risk factor for a wide spectrum of neuropsychiatric disorders. J Hum Genet 56:541–544. https://doi.org/10.1038/jhg.2011.42

Rautiainen M, Nurk S, Walenz BP, et al (2023) Telomere-to-telomere assembly of diploid chromosomes with Verkko. Nat Biotechnol 41:1474–1482. https://doi.org/10.1038/s41587-023-01662-6

Reese F, Williams B, Balderrama-Gutierrez G, et al (2023) The ENCODE4 long-read RNA-seq collection reveals distinct classes of transcript structure diversity. 2023.05.15.540865

Rhie A, McCarthy SA, Fedrigo O, et al (2021) Towards complete and error-free genome assemblies of all vertebrate species. Nature 592:737–746. https://doi.org/10.1038/s41586-021-03451-0

Rybak-Wolf A, Wyler E, Pentimalli TM, et al (2023) Modelling viral encephalitis caused by herpes simplex virus 1 infection in cerebral organoids. Nat Microbiol 8:1252–1266. https://doi.org/10.1038/s41564-023-01405-y

Schertzer MD, Stirn A, Isaev K, et al (2023) Cas13d-mediated isoform-specific RNA knockdown with a unified computational and experimental toolbox. 2023.09.12.557474

Seidegård J, Vorachek WR, Pero RW, Pearson WR (1988) Hereditary differences in the expression of the human glutathione transferase active on trans-stilbene oxide are due to a gene deletion. Proc Natl Acad Sci U S A 85:7293–7297. https://doi.org/10.1073/pnas.85.19.7293

Sekar A, Bialas AR, de Rivera H, et al (2016) Schizophrenia risk from complex variation of complement component 4. Nature 530:177–183. https://doi.org/10.1038/nature16549

Sharp AJ, Hansen S, Selzer RR, et al (2006) Discovery of previously unidentified genomic disorders from the duplication architecture of the human genome. Nat Genet 38:1038–1042. https://doi.org/10.1038/ng1862

She X, Jiang Z, Clark RA, et al (2004) Shotgun sequence assembly and recent segmental duplications within the human genome. Nature 431:927–930. https://doi.org/10.1038/nature03062

She X, Liu G, Ventura M, et al (2006) A preliminary comparative analysis of primate segmental duplications shows elevated substitution rates and a great-ape expansion of intrachromosomal duplications. Genome Res 16:576–583. https://doi.org/10.1101/gr.4949406

Shimada M, Omae Y, Kakita A, et al (2024) Identification of region-specific gene isoforms in the human brain using long-read transcriptome sequencing. Science Advances 10:eadj5279. https://doi.org/10.1126/sciadv.adj5279

Shumate A, Salzberg SL (2021) Liftoff: accurate mapping of gene annotations. Bioinformatics 37:1639–1643. https://doi.org/10.1093/bioinformatics/btaa1016

Smith DK, Xue H (1997) Sequence profiles of immunoglobulin and immunoglobulin-like domains. J Mol Biol 274:530–545. https://doi.org/10.1006/jmbi.1997.1432

Souilmi Y, Tobler R, Johar A, et al (2022) Admixture has obscured signals of historical hard sweeps in humans. Nat Ecol Evol 6:2003–2015. https://doi.org/10.1038/s41559-022-01914-9

Stallings RL, Whitmore SA, Doggett NA, Callen DF (2008) Refined physical mapping of chromosome 16-specific low-abundance repetitive DNA sequences. Cytogenetics and Cell Genetics 63:97–101. https://doi.org/10.1159/000133509

Stankiewicz P, Lupski JR (2002) Genome architecture, rearrangements and genomic disorders. Trends in Genetics 18:74–82. https://doi.org/10.1016/S0168-9525(02)02592-1

Stefansson H, Helgason A, Thorleifsson G, et al (2005) A common inversion under selection in Europeans. Nat Genet 37:129–137. https://doi.org/10.1038/ng1508

Steinberg KM, Antonacci F, Sudmant PH, et al (2012) Structural diversity and African origin of the 17q21.31 inversion polymorphism. Nat Genet 44:872–880. https://doi.org/10.1038/ng.2335

Steiper ME, Young NM, Sukarna TY (2004) Genomic data support the hominoid slowdown and an Early Oligocene estimate for the hominoid–cercopithecoid divergence. Proceedings of the National Academy of Sciences 101:17021–17026. https://doi.org/10.1073/pnas.0407270101

Stergachis AB, Debo BM, Haugen E, et al (2020) Single-molecule regulatory architectures captured by chromatin fiber sequencing. Science 368:1449–1454. https://doi.org/10.1126/science.aaz1646

Sudmant PH, Huddleston J, Catacchio CR, et al (2013) Evolution and diversity of copy number variation in the great ape lineage. Genome Res 23:1373–1382. https://doi.org/10.1101/gr.158543.113

Sudmant PH, Kitzman JO, Antonacci F, et al (2010) Diversity of human copy number variation and multicopy genes. Science 330:641–646. https://doi.org/10.1126/science.1197005

Sudmant PH, Mallick S, Nelson BJ, et al (2015a) Global diversity, population stratification, and selection of human copy-number variation. Science 349:aab3761. https://doi.org/10.1126/science.aab3761

Sudmant PH, Rausch T, Gardner EJ, et al (2015b) An integrated map of structural variation in 2,504 human genomes. Nature 526:75–81. https://doi.org/10.1038/nature15394

Sun YH, Wang A, Song C, et al (2021) Single-molecule long-read sequencing reveals a conserved intact long RNA profile in sperm. Nat Commun 12:1361. https://doi.org/10.1038/s41467-021-21524-6

Suzuki IK, Gacquer D, Van\ Heurck R, et al (2018) Human-Specific NOTCH2NL Genes Expand Cortical Neurogenesis through Delta/Notch Regulation. Cell 173:1370-1384.e16. https://doi.org/10.1016/j.cell.2018.03.067

Szpiech ZA (2024) selscan 2.0: scanning for sweeps in unphased data. Bioinformatics 40:btae006. https://doi.org/10.1093/bioinformatics/btae006

Tajima F (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123:585–595. https://doi.org/10.1093/genetics/123.3.585

Talevich E, Invergo BM, Cock PJ, Chapman BA (2012) Bio.Phylo: A unified toolkit for processing, analyzing and visualizing phylogenetic trees in Biopython. BMC Bioinformatics 13:209. https://doi.org/10.1186/1471-2105-13-209

Tamura K, Battistuzzi FU, Billing-Ross P, et al (2012) Estimating divergence times in large molecular phylogenies. Proceedings of the National Academy of Sciences 109:19333–19338. https://doi.org/10.1073/pnas.1213199109

Tamura K, Nei M (1993) Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. Mol Biol Evol 10:512–526. https://doi.org/10.1093/oxfordjournals.molbev.a040023

Tamura K, Stecher G, Kumar S (2021) MEGA11: Molecular Evolutionary Genetics Analysis Version 11. Molecular Biology and Evolution 38:3022–3027. https://doi.org/10.1093/molbev/msab120

Tamura K, Tao Q, Kumar S (2018) Theoretical Foundation of the RelTime Method for Estimating Divergence Times from Variable Evolutionary Rates. Molecular Biology and Evolution 35:1770–1782. https://doi.org/10.1093/molbev/msy044

Teufel F, Almagro Armenteros JJ, Johansen AR, et al (2022) SignalP 6.0 predicts all five types of signal peptides using protein language models. Nat Biotechnol 40:1023–1025. https://doi.org/10.1038/s41587-021-01156-3

Trask BJ, Friedman C, Martin-Gallardo A, et al (1998) Members of the olfactory receptor gene family are contained in large blocks of DNA duplicated polymorphically near the ends of human chromosomes. Human Molecular Genetics 7:13–26. https://doi.org/10.1093/hmg/7.1.13

Trégouët D-A, König IR, Erdmann J, et al (2009) Genome-wide haplotype association study identifies the SLC22A3-LPAL2-LPA gene cluster as a risk locus for coronary artery disease. Nat Genet 41:283–285. https://doi.org/10.1038/ng.314

Tuzun E, Sharp AJ, Bailey JA, et al (2005) Fine-scale structural variation of the human genome. Nat Genet 37:727–732. https://doi.org/10.1038/ng1562

Upton KR, Gerhardt DJ, Jesuadian JS, et al (2015) Ubiquitous L1 Mosaicism in Hippocampal Neurons. Cell 161:228–239. https://doi.org/10.1016/j.cell.2015.03.026

Van Heurck R, Bonnefont J, Wojno M, et al (2023) CROCCP2 acts as a human-specific modifier of cilia dynamics and mTOR signaling to promote expansion of cortical progenitors. Neuron 111:65-80.e6. https://doi.org/10.1016/j.neuron.2022.10.018

Vollger MR, Dishuck PC, Harvey WT, et al (2023) Increased mutation and gene conversion within human segmental duplications. Nature 617:325–334. https://doi.org/10.1038/s41586-023-05895-y

Vollger MR, Dishuck PC, Sorensen M, et al (2019) Long-read sequence and assembly of segmental duplications. Nat Methods 16:88–94. https://doi.org/10.1038/s41592-018-0236-3

Vollger MR, Guitart X, Dishuck PC, et al (2022) Segmental duplications and their variation in a complete human genome. Science 376:eabj6965. https://doi.org/10.1126/science.abj6965

Vollger MR, Logsdon GA, Audano PA, et al (2020) Improved assembly and variant detection of a haploid human genome using single-molecule, high-fidelity long reads. Ann Hum Genet 84:125–140. https://doi.org/10.1111/ahg.12364

Wang T, Antonacci-Fulton L, Howe K, et al (2022) The Human Pangenome Project: a global resource to map genomic diversity. Nature 604:437–446. https://doi.org/10.1038/s41586-022-04601-8

Weiss LA, Shen Y, Korn JM, et al (2008) Association between Microdeletion and Microduplication at 16p11.2 and Autism. New England Journal of Medicine 358:667–675. https://doi.org/10.1056/NEJMoa075974

Wenger AM, Peluso P, Rowell WJ, et al (2019) Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. Nat Biotechnol 1–8. https://doi.org/10.1038/s41587-019-0217-9

Yang Y, Chung EK, Wu YL, et al (2007) Gene copy-number variation and associated polymorphisms of complement component C4 in human systemic lupus erythematosus (SLE): low copy number is a risk factor for and high copy number is a protective factor against SLE susceptibility in European Americans. Am J Hum Genet 80:1037–1054. https://doi.org/10.1086/518257

Yoo D, Rhie A, Hebbar P, et al (2024) Complete sequencing of ape genomes. 2024.07.31.605654

Zhang S-J, Liu C-J, Yu P, et al (2014) Evolutionary interrogation of human biology in well-annotated genomic framework of rhesus macaque. Mol Biol Evol 31:1309–1324. https://doi.org/10.1093/molbev/msu084

Zody MC, Garber M, Adams DJ, et al (2006) DNA sequence of human chromosome 17 and analysis of rearrangement in the human lineage. Nature 440:1045–1049. https://doi.org/10.1038/nature04689

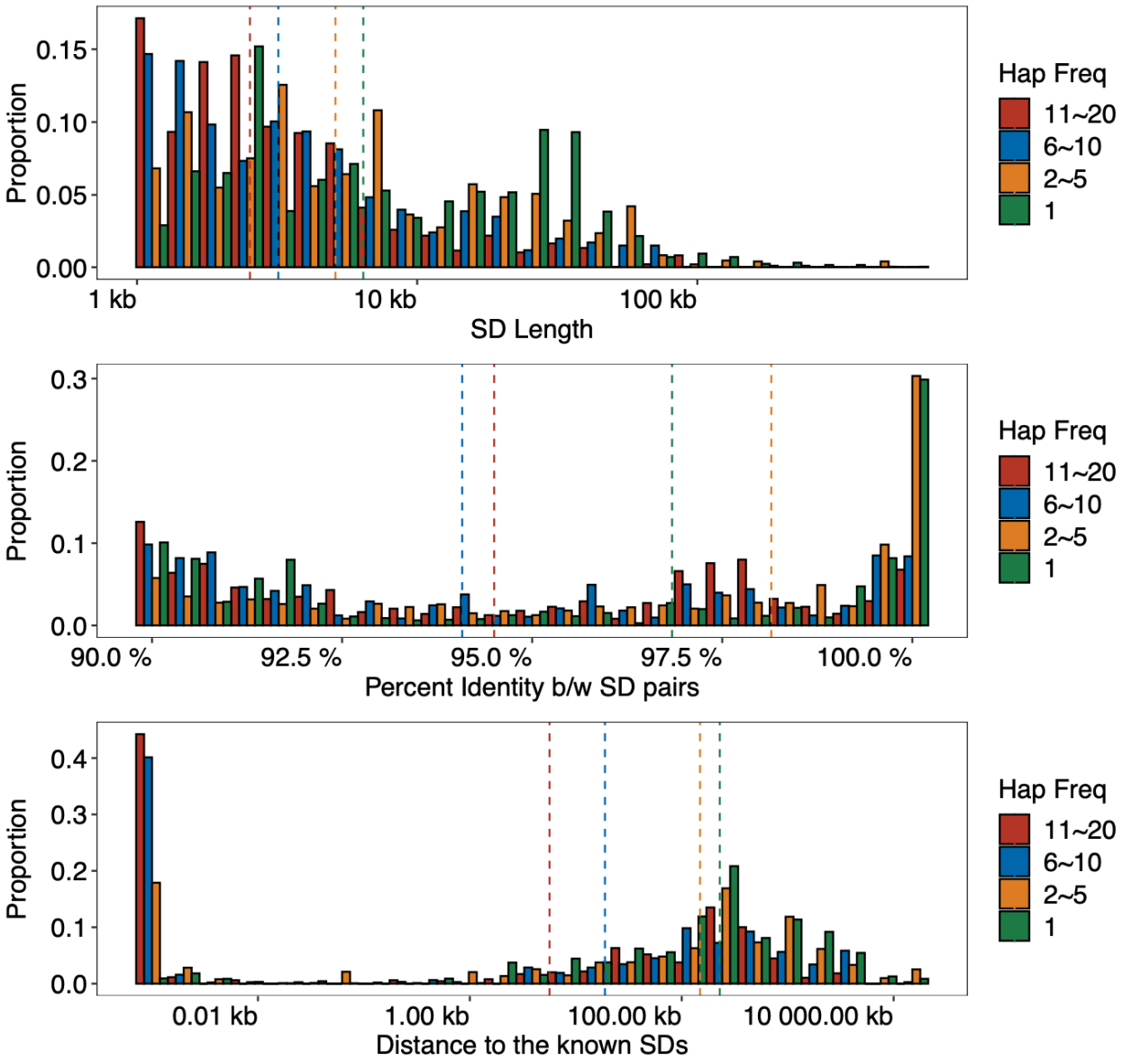
Zody MC, Jiang Z, Fung H-C, et al (2008) Evolutionary toggling of the MAPT 17q21.31 inversion region. Nat Genet 40:1076–1083. https://doi.org/10.1038/ng.193

Zook JM, Catoe D, McDaniel J, et al (2016) Extensive sequencing of seven human genomes to characterize benchmark reference materials. Sci Data 3:160025. https://doi.org/10.1038/sdata.2016.25

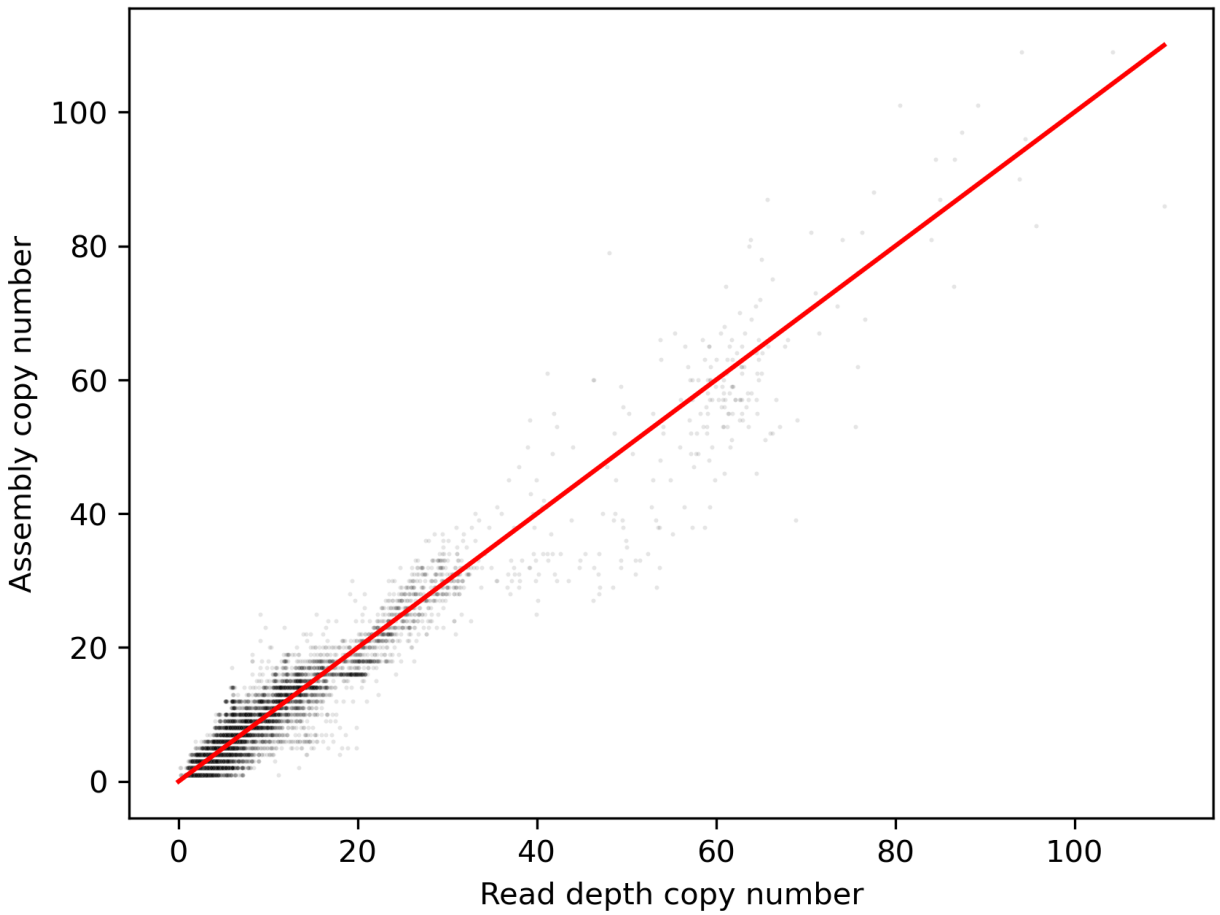
(2023) PacificBiosciences/ANGEL

Archaeopteryx. http://www.phylosoft.org/archaeopteryx/. Accessed 31 Oct 2024

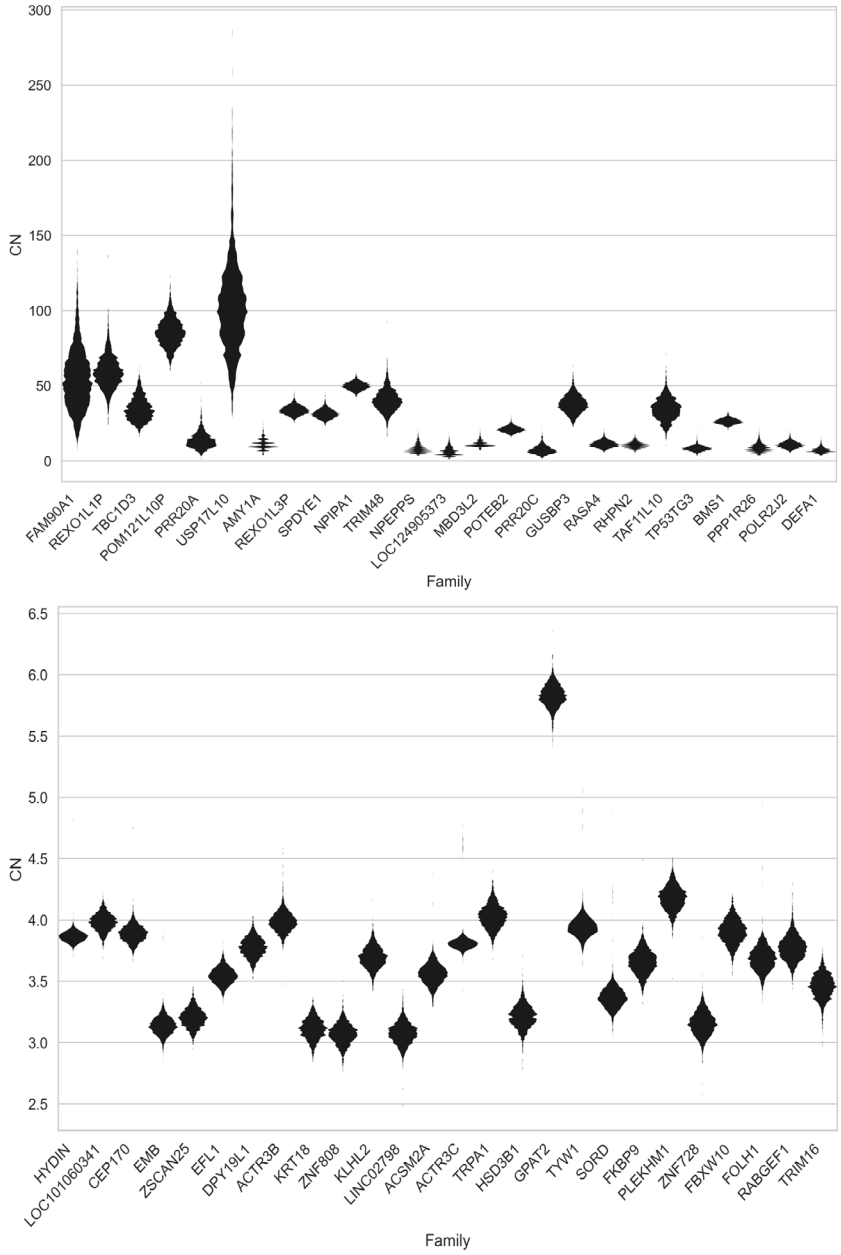
APPENDIX A. SUPPLEMENT FOR CHAPTER 2



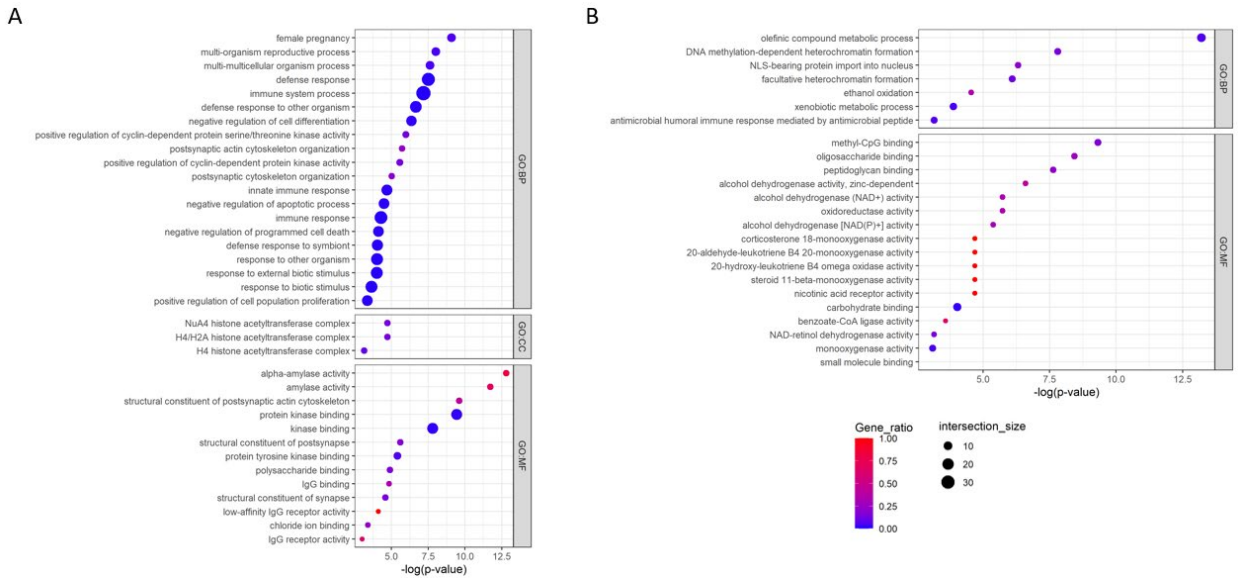
Supplementary Figure 1. Histogram comparing the sequence identity and length of polymorphic segmental duplications (SDs) at different haplotype frequencies.



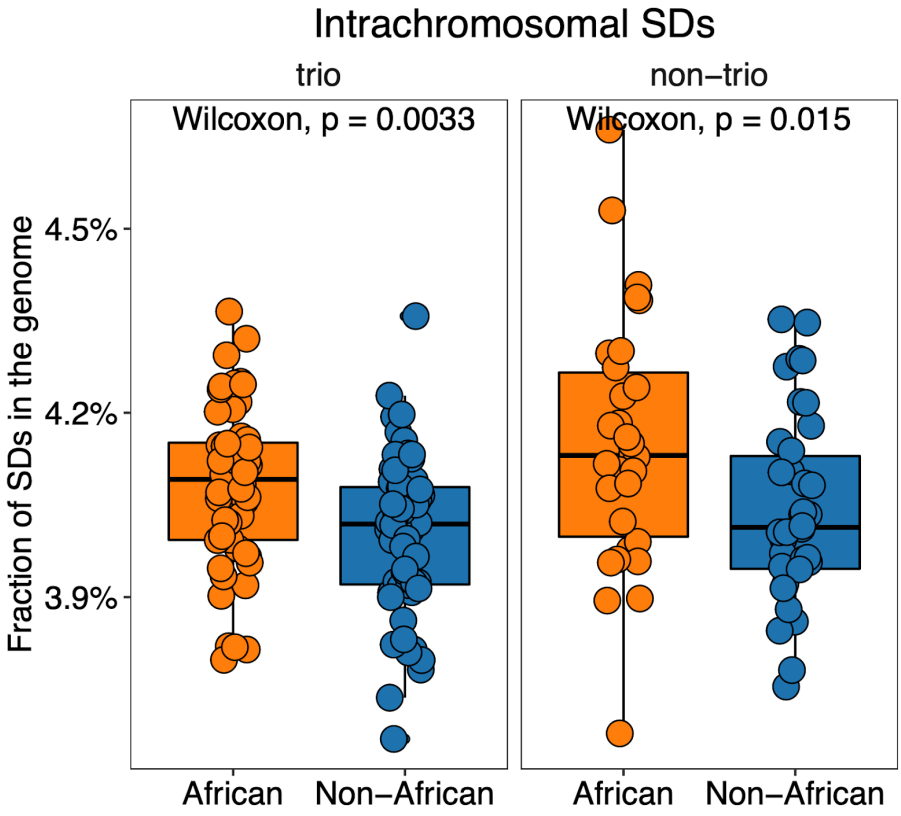
Supplementary Figure 2. Read-depth-based copy numbers estimated with fastCN compared to assembled copy number for each sample, summed between the two haplotypes. Each point represents the copy number estimates for a gene family in a sample (R^2 = 0.94).



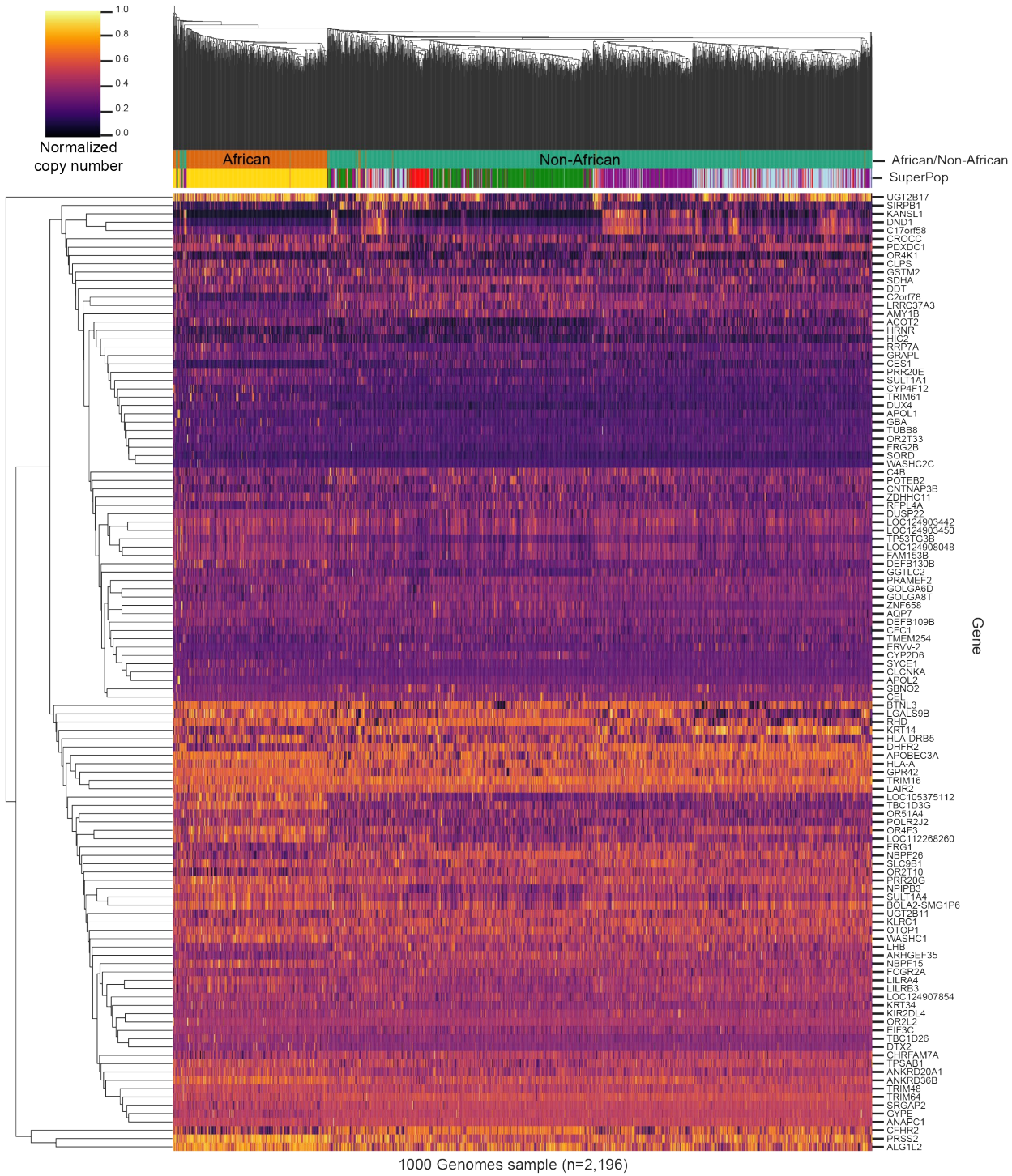
Supplementary Figure 3. Copy number distribution of high- and low-variance gene families. The read-depth copy number of gene families with highly variable (above) and nearly fixed copy number (below) are displayed. Gene families are selected and ordered by variance, requiring an average diploid copy number greater than three.



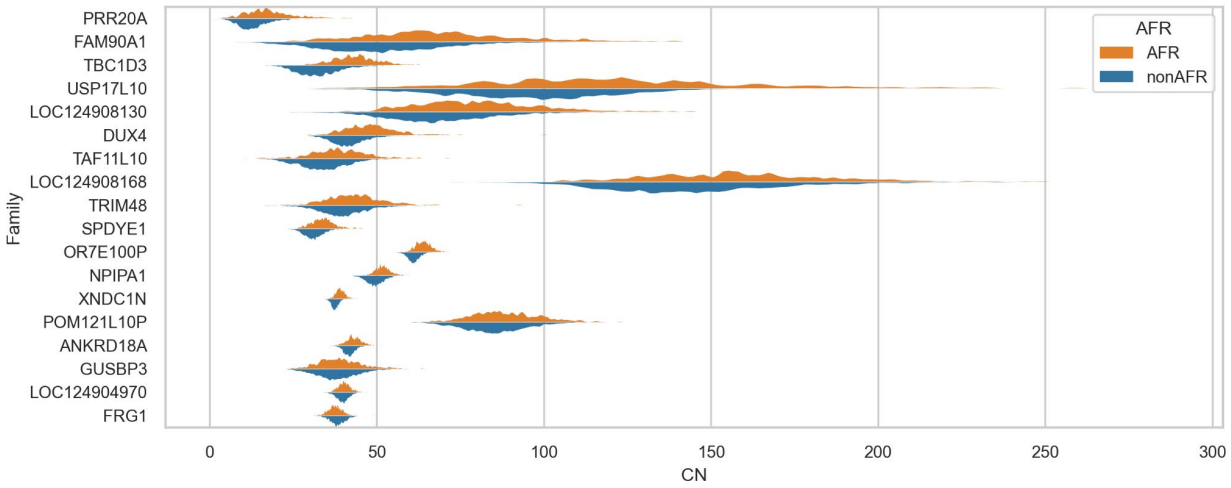
Supplementary Figure 4. Gene ontology enrichment of the (A) top 100 variable gene families (n = 358) and (B) invariable genes (n = 115). X-axis indicates negative log transformed adjusted p-value. The number of intersecting genes is indicated by the size of the circle and the gene ratio represents the number of intersect/term size.

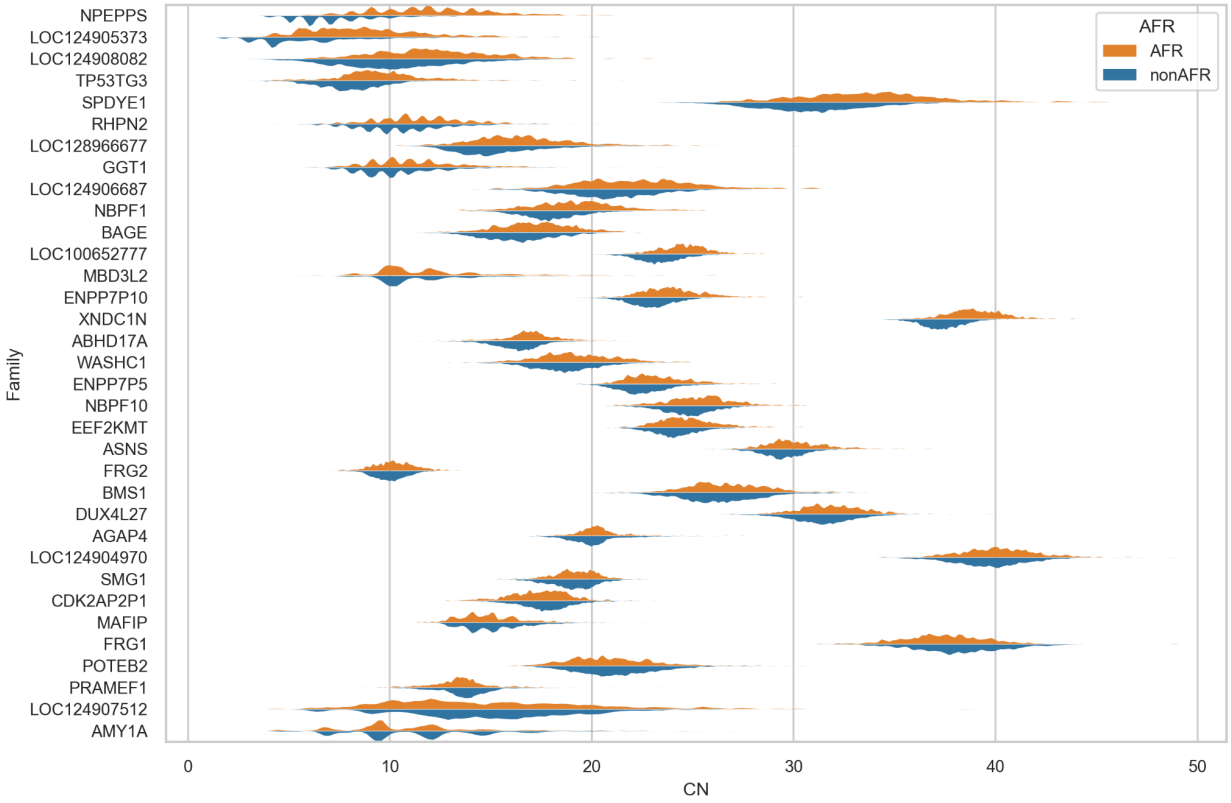


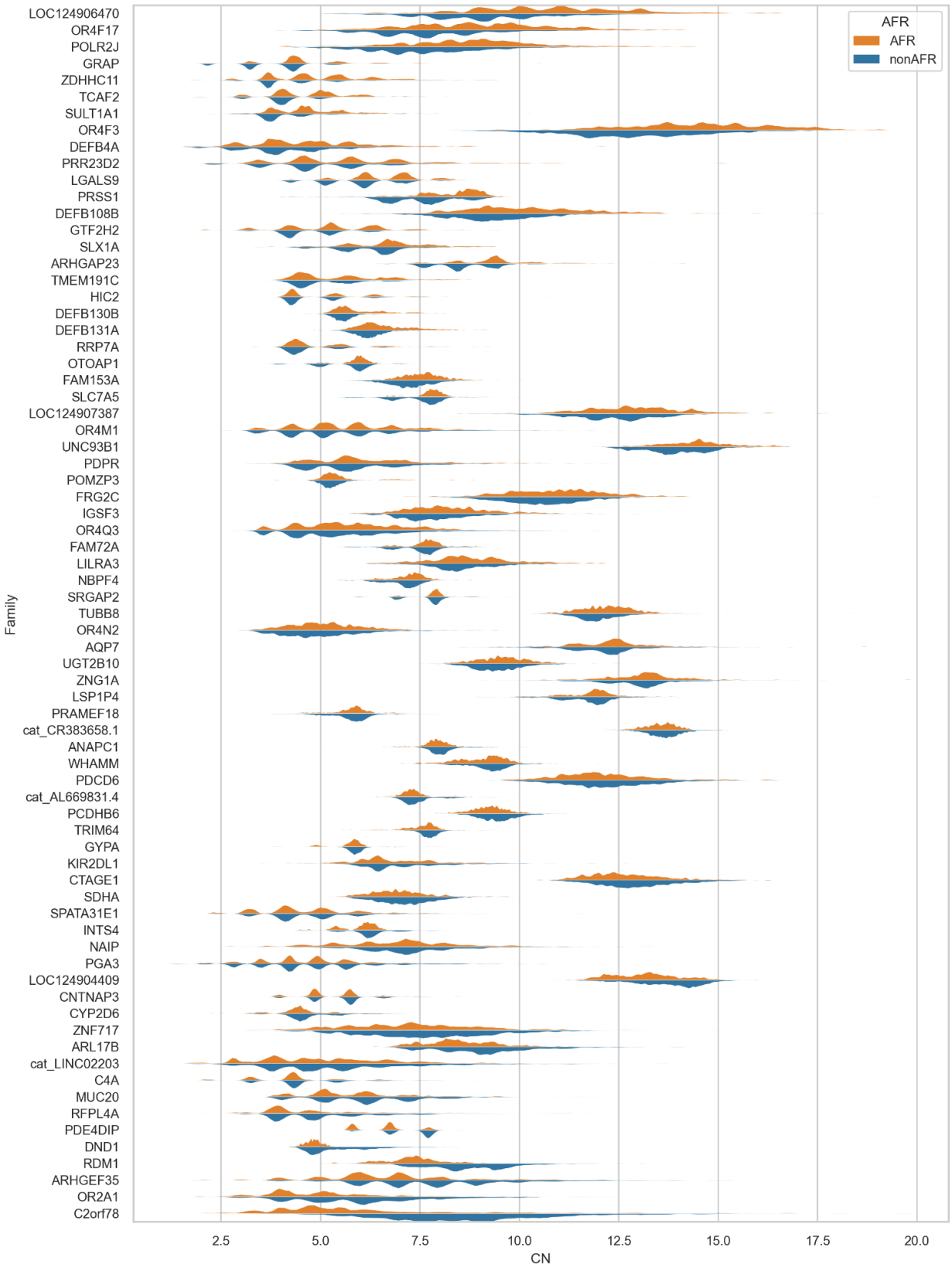
Supplementary Figure 5. Comparison of SD content in trio vs. non-trio genomes. Both datasets show a significant excess of SD content in African.

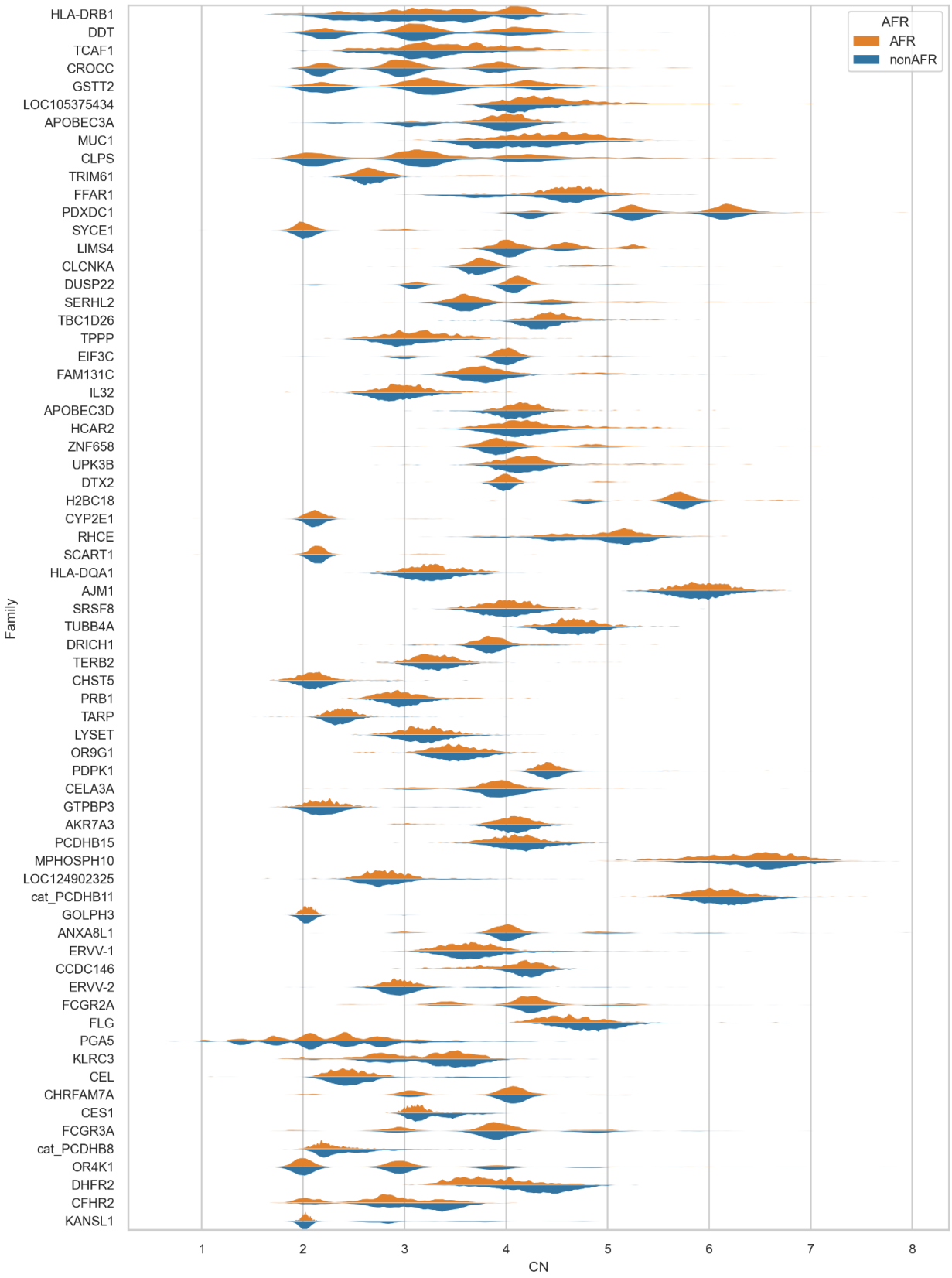


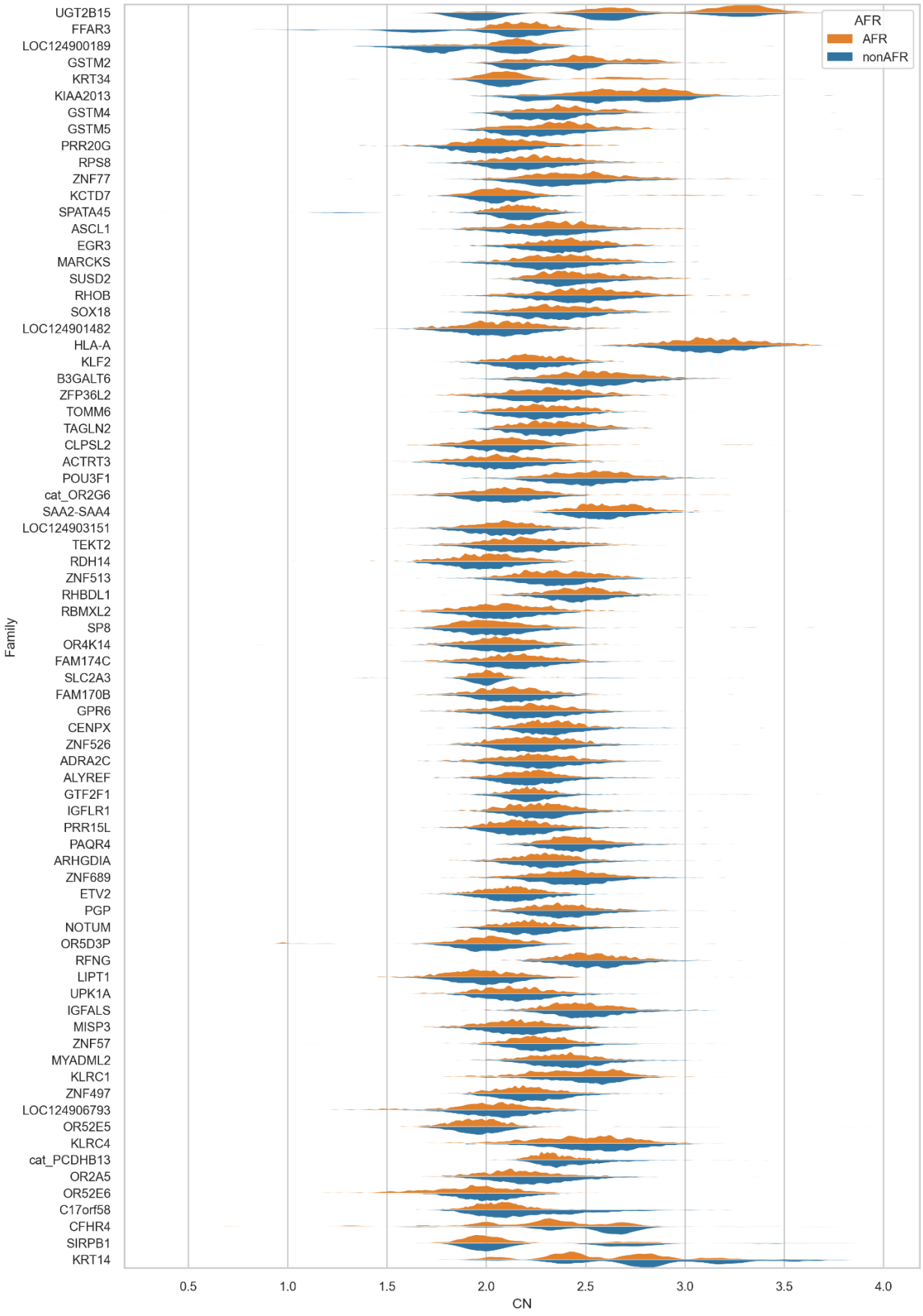
Supplementary Figure 6. Population-stratified genic copy number in 2,196 unrelated individuals from the 1000 Genomes Project. Gene copy number values are centered on the mean for each gene and scaled by unit variance to range from 0-1. One paralog per gene family and duplication block is shown. 73/115 deduplicated population-stratified genes have higher mean copy number in the African group as compared to the non-African group (p=0.002). Superpopulations as described in the 1000 Genomes Project are shown above (Africa: gold, East Asia: green, South Asia: purple, Europe: blue, the Americas: red). Copy number estimates by population group (below).


PRR20A
FAM90A1
TBC1D3
USP17L10
LOC124908130
DUX4
TAF11L10
LOC124908168
TRIM48
SPDYE1
OR7E100P
NPIPA1
XNDC1N
POM121L10P
ANKRD18A
GUSBP3
LOC124904970
FRG1
Family
AFR
AFR
nonAFR
0
50
100
150
200
250
300
CN

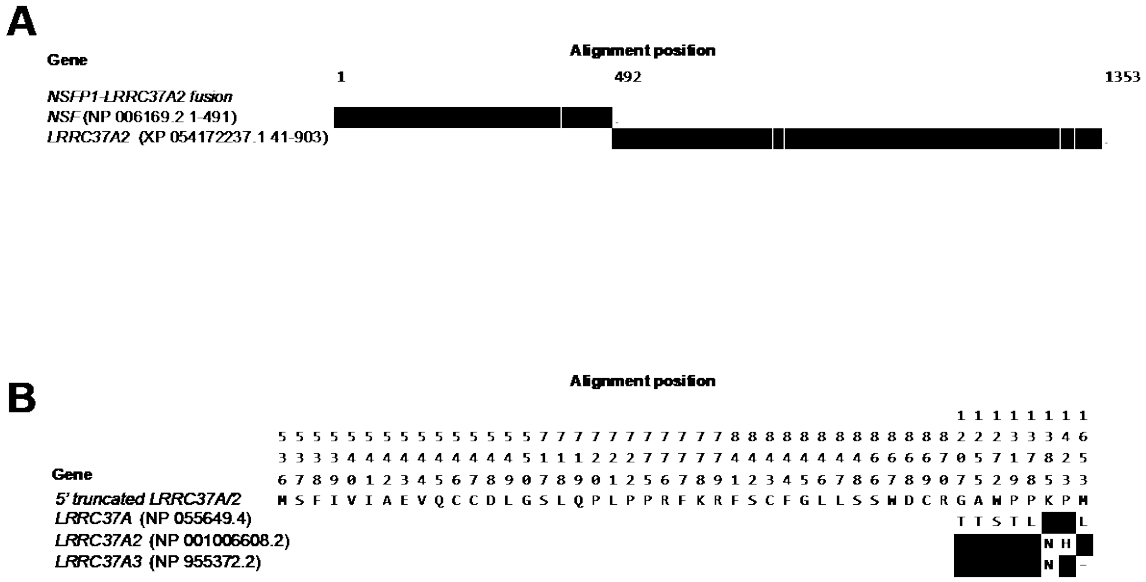

NPEPPS
LOC124905373
LOC124908082
TP53TG3
SPDYE1
RHPN2
LOC128966677
GGT1
LOC124906687
NBPF1
BAGE
LOC100652777
MBD3L2
ENPP7P10
XNDC1N
ABHD17A
WASHC1
ENPP7P5
NBPF10
EEF2KMT
ASNS
FRG2
BMS1
DUX4L27
AGAP4
LOC124904970
SMG1
CDK2AP2P1
MAFIP
FRG1
POTEB2
PRAMEF1
LOC124907512
AMY1A
Family
AFR
AFR
nonAFR
0
10
20
30
40
50
CN


AFR
AFR
nonAFR
Family
LOC124906470
OR4F17
POLR2J
GRAP
ZDHHC11
TCAF2
SULT1A1
OR4F3
DEFB4A
PRR23D2
LGALS9
PRSS1
DEFB108B
GTF2H2
SLX1A
ARHGAP23
TMEM191C
HIC2
DEFB130B
DEFB131A
RRP7A
OTOAP1
FAM153A
SLC7A5
LOC124907387
OR4M1
UNC93B1
PDPR
POMZP3
FRG2C
IGSF3
OR4Q3
FAM72A
LILRA3
NBPF4
SRGAP2
TUBB8
OR4N2
AQP7
UGT2B10
ZNG1A
LSP1P4
PRAMEF18
cat_CR383658.1
ANAPC1
WHAMM
PDCD6
cat_AL669831.4
PCDHB6
TRIM64
GYPA
KIR2DL1
CTAGE1
SDHA
SPATA31E1
INTS4
NAIP
PGA3
LOC124904409
CNTNAP3
CYP2D6
ZNF717
ARL17B
cat_LINC02203
C4A
MUC20
RFPL4A
PDE4DIP
DND1
RDM1
ARHGEF35
OR2A1
C2orf78
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
CN

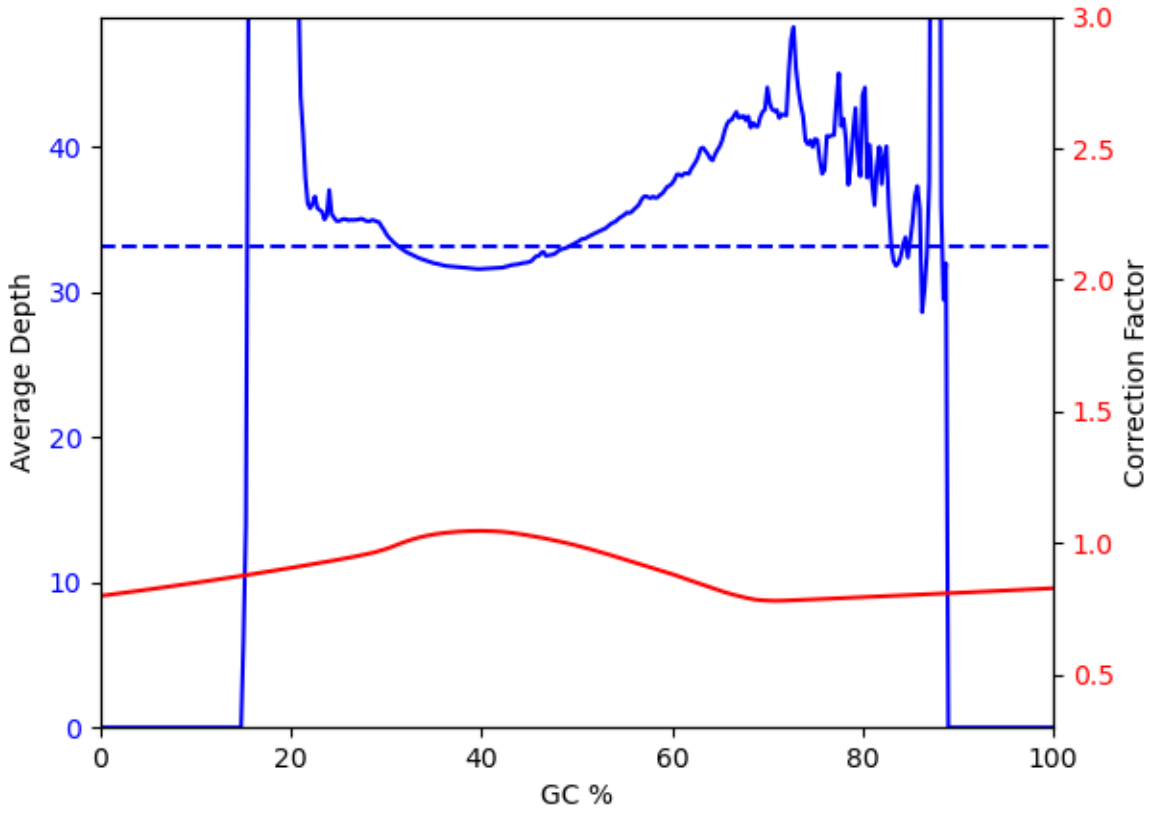

AFR
AFR
nonAFR
HLA-DRB1
DDT
TCAF1
CROCC
GSTT2
LOC105375434
APOBEC3A
MUC1
CLPS
TRIM61
FFAR1
PDXDC1
SYCE1
LIMS4
CLCNKA
DUSP22
SERHL2
TBC1D26
TPPP
EIF3C
FAM131C
IL32
APOBEC3D
HCAR2
ZNF658
UPK3B
DTX2
H2BC18
CYP2E1
RHCE
SCART1
HLA-DQA1
AJM1
SRSF8
TUBB4A
DRICH1
TERB2
CHST5
PRB1
TARP
LYSET
OR9G1
PDPK1
CELA3A
GTPBP3
AKR7A3
PCDHB15
MPHOSPH10
LOC124902325
cat_PCDHB11
GOLPH3
ANXA8L1
ERVV-1
CCDC146
ERVV-2
FCGR2A
FLG
PGA5
KLRC3
CEL
CHRFAM7A
CES1
FCGR3A
cat_PCDHB8
OR4K1
DHFR2
CFHR2
KANSL1
Family
1
2
3
4
5
6
7
8
CN


AFR
AFR
nonAFR
UGT2B15
FFAR3
LOC124900189
GSTM2
KRT34
KIAA2013
GSTM4
GSTM5
PRR20G
RPS8
ZNF77
KCTD7
SPATA45
ASCL1
EGR3
MARCKS
SUSD2
RHOB
SOX18
LOC124901482
HLA-A
KLF2
B3GALT6
ZFP36L2
TOMM6
TAGLN2
CLPSL2
ACTRT3
POU3F1
cat_OR2G6
SAA2-SAA4
LOC124903151
TEKT2
RDH14
ZNF513
RHBDL1
RBMXL2
SP8
OR4K14
FAM174C
SLC2A3
FAM170B
GPR6
CENPX
ZNF526
ADRA2C
ALYREF
GTF2F1
IGFLR1
PRR15L
PAQR4
ARHGDIA
ZNF689
ETV2
PGP
NOTUM
OR5D3P
RFNG
LIPT1
UPK1A
IGFALS
MISP3
ZNF57
MYADML2
KLRC1
ZNF497
LOC124906793
OR52E5
KLRC4
cat_PCDHB13
OR2A5
OR52E6
C17orf58
CFHR4
SIRPB1
KRT14
Family
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
CN

Supplementary Figure 7. Population-stratified gene copy number in 1000 Genomes Project. To validate population differences in gene copy number, we repeated the analysis on an unrelated subset of high-coverage Illumina data from the 1000 Genomes Project, excluding samples identified as related by published pedigree or with cryptic relationships (2nd degree or less) as identified by Somalier (Pedersen, Genome Medicine 2020). Individuals from the highly admixed ACB, ASW, and PUR populations were also excluded, leaving n=2,196.



Supplementary Figure 8. Multiple sequence alignment of novel *LRRC37A* family gene predictions compared to reference annotations. Black boxes indicate matches between the novel prediction and reference annotations. **(A)** The *NSFP1-LRRC37A2* gene prediction encodes a fusion between *NSFP1* and *LRRC37A2*. **(B)** The 5' truncated *LRRC37A/2* gene prediction has 39 novel amino acids followed by a hybrid of *LRRC37A* and *LRRC37A2*.



Supplementary Figure 9. Adjustment factor for gene copy number estimation.
The read depth of k-mers from decomposed T2T genome assembly is shown as a function of GC composition (blue). Even in a finished genome where there is no experimental or technical error, k-mer read depth is not uniform, since as GC and AT content increases so too does the number of low-complexity k-mers mapping elsewhere. Based on this, we estimated an adjustment factor required by fastCN (red line) to correct for this bias.

Table S1. Geographic origin of the samples used in this study.

Source	Sample	Sex	Trio data available	Population	Superpopulation
HPRC	HG01891	Female	1	ACB	AFR
HPRC	HG02257	Female	1	ACB	AFR
HPRC	HG02559	Female	1	ACB	AFR
HPRC	HG02622	Female	1	GWD	AFR
HPRC	HG02630	Female	1	GWD	AFR
HPRC	HG02723	Female	1	GWD	AFR
HPRC	HG02818	Female	1	GWD	AFR
HPRC	HG02886	Female	1	GWD	AFR
HPRC	HG03453	Female	1	MSL	AFR
HPRC	HG03486	Female	1	MSL	AFR
HPRC	HG03516	Female	1	ESN	AFR
HPRC	HG03540	Female	1	GWD	AFR
HPRC	NA18906	Female	1	YRI	AFR
HPRC	NA20129	Female	1	ASW	AFR
HPRC	HG00735	Female	1	PUR	AMR
HPRC	HG00741	Female	1	PUR	AMR
HPRC	HG01071	Female	1	PUR	AMR
HPRC	HG01123	Female	1	CLM	AMR
HPRC	HG01175	Female	1	PUR	AMR
HPRC	HG01361	Female	1	CLM	AMR
HPRC	HG01978	Female	1	PEL	AMR
HPRC	HG02148	Female	1	PEL	AMR
HPRC	HG00438	Female	1	CHS	EAS
HPRC	HG02080	Female	1	KHV	EAS
HPRC	HG02055	Male	1	ACB	AFR
HPRC	HG02145	Male	1	ACB	AFR
HPRC	HG02486	Male	1	ACB	AFR
HPRC	HG02572	Male	1	GWD	AFR
HPRC	HG02717	Male	1	GWD	AFR
HPRC	HG03098	Male	1	MSL	AFR
HPRC	HG03579	Male	1	MSL	AFR
HPRC	HG01106	Male	1	PUR	AMR
HPRC	HG01258	Male	1	CLM	AMR
HPRC	HG01358	Male	1	CLM	AMR
HPRC	HG01928	Male	1	PEL	AMR
HPRC	HG01952	Male	1	PEL	AMR
HPRC	HG00621	Male	1	CHS	EAS

HPRC	HG00673	Male	1	CHS	EAS
HPRC	HG002	Male	1	Ashk	EUR
HPRC	HG03492	Male	1	PJL	SAS
HGSVC	HG02587	Female	0	GWD	AFR
HGSVC	HG03125	Female	0	ESN	AFR
HGSVC	NA19238	Female	0	YRI	AFR
HGSVC	NA19983	Female	0	ASW	AFR
HGSVC	HG00732	Female	0	PUR	AMR
HGSVC	HG01114	Female	0	CLM	AMR
HGSVC	HG01352	Female	1	CLM	AMR
HGSVC	HG02106	Female	1	PEL	AMR
HGSVC	HG00513	Female	0	CHS	EAS
HGSVC	HG00864	Female	0	CDX	EAS
HGSVC	HG02018	Female	1	KHV	EAS
HGSVC	HG02059	Female	1	KHV	EAS
HGSVC	HG00171	Female	0	FIN	EUR
HGSVC	HG00268	Female	0	FIN	EUR
HGSVC	NA12329	Female	0	CEU	EUR
HGSVC	NA12878	Female	0	CEU	EUR
HGSVC	HG03683	Female	0	STU	SAS
HGSVC	NA20847	Female	0	GIH	SAS
HGSVC	HG03807	Female	1	BEB	SAS
HGSVC	HG04036	Female	1	STU	SAS
HGSVC	HG04217	Female	1	ITU	SAS
HGSVC	HG02011	Male	0	ACB	AFR
HGSVC	HG02554	Male	0	ACB	AFR
HGSVC	HG02666	Male	0	GWD	AFR
HGSVC	HG02953	Male	0	ESN	AFR
HGSVC	HG03065	Male	0	MSL	AFR
HGSVC	HG03371	Male	0	ESN	AFR
HGSVC	NA19317	Male	0	LWK	AFR
HGSVC	NA19331	Male	0	LWK	AFR
HGSVC	NA19347	Male	0	LWK	AFR
HGSVC	HG03248	Male	1	GWD	AFR
HGSVC	HG03456	Male	1	MSL	AFR
HGSVC	NA19705	Male	1	ASW	AFR
HGSVC	HG00731	Male	0	PUR	AMR
HGSVC	NA19650	Male	0	MXL	AMR
HGSVC	HG01457	Male	1	CLM	AMR

HGSVC	HG00512	Male	0	CHS	EAS
HGSVC	HG00096	Male	0	GBR	EUR
HGSVC	HG00358	Male	0	FIN	EUR
HGSVC	HG01505	Male	0	IBS	EUR
HGSVC	NA20509	Male	0	TSI	EUR
HGSVC	HG02492	Male	0	PJL	SAS
HGSVC	HG03009	Male	0	BEB	SAS
HGSVC	HG03732	Male	0	ITU	SAS
HGSVC	NA19239	Male	0	YRI	YRI

Table S2. Summary statistics of genome assemblies and segmental duplications (SDs).

sample	hap	sequence coverage (HiFi)	contig N50	asm size (contig)	asm size (RagTag; autosome)	intrachromosomal SDs (non-overlap) (bp)	interchromosomal SDs (non-overlap) (bp)
HG00096	hap1	36.36	67.92	2870960595	2870978495	122737555	89805241
HG00096	hap2	36.36	72.5	2857181257	2857196557	111843466	79851574
HG00171	hap2	36.34	72.46	3038968752	2885296041	118436823	90649938
HG00171	hap1	36.34	88.09	3111802139	2959813194	122478751	95149642
HG002	hap2	180.03	81.88	3060609068	2906409739	118489271	89634159
HG002	hap1	186.23	84.97	2958633312	2904407133	110291949	85826437
HG00268	hap1	37.02	31.32	3046055381	2892770460	114596611	84828666
HG00268	hap2	37.02	25.51	3022501768	2869678797	119170359	88896579
HG00358	hap2	69.34	93.8	2930444717	2930456217	117376045	88948149
HG00358	hap1	69.34	77.32	2873266841	2873280941	115255191	89324666
HG00438	hap2	41.83	54.94	3035735720	2882116915	119741609	85847245
HG00438	hap1	41.98	48.06	3025118465	2871102072	117003799	83659831
HG00512	hap2	29.84	39.11	2884911297	2884948297	112989953	80159717
HG00512	hap1	29.84	40.61	2920270862	2920299462	112722732	84506798
HG00513	hap2	40.91	50.23	3043508037	2893091513	114887209	85224474
HG00513	hap1	40.91	51.71	3035409491	2880384271	123432296	86006069
HG00621	hap2	40.36	50.29	3023026071	2868327268	110771232	78095704
HG00621	hap1	41.98	54.67	2905948993	2861724295	111601975	78528466
HG00673	hap2	39.95	29.08	3053585067	2898972124	126319061	91290906
HG00673	hap1	41.70	34.84	2925716157	2877715772	117003735	84491849
HG00731	hap2	73.64	41.39	2874958616	2874990216	117963455	88457169
HG00731	hap1	73.64	39	2874645509	2874679509	108694805	76970990
HG00732	hap2	92.45	57.32	3028848085	2874625821	124958058	84926171
HG00732	hap1	92.45	54.26	3078702073	2923315535	127225671	95223274
HG00735	hap2	43.78	56.47	3037795105	2884676984	112988474	86363352

HG00735	hap1	43.84	53.42	3033541617	2878732983	121710690	90255050
HG00741	hap2	39.19	41	3036701854	2882388204	117593716	88123586
HG00741	hap1	39.28	51.04	3029878036	2875655707	115328858	86776989
HG00864	hap2	33.08	52.1	3054415158	2899336169	116970108	85751838
HG00864	hap1	33.08	71.39	3103903447	2953388264	110878792	84600282
HG01071	hap1	35.65	55.59	3057222025	2902924045	121731593	87057468
HG01071	hap2	36.18	50.13	3012710110	2858328009	115256692	86141418
HG01106	hap2	47.76	47.71	3035845582	2881415376	115751782	87270422
HG01106	hap1	49.54	57.17	2927007346	2866221620	116180766	87420149
HG01114	hap2	34.64	62.17	3031686963	2878545575	123423501	90657760
HG01114	hap1	34.64	79.15	3110415371	2955865762	116609666	84504463
HG01123	hap1	39.48	44.72	3014197469	2859358782	114164939	84708425
HG01123	hap2	39.50	54.36	3012822948	2858115370	111868840	84303004
HG01175	hap2	36.95	36.54	3030735652	2877614508	113401399	86269646
HG01175	hap1	36.96	34.8	3030026811	2875617013	114510755	84937115
HG01258	hap2	36.60	56.64	3032420282	2877667767	115394432	87112245
HG01258	hap1	38.08	49.86	2915178237	2869054873	115397855	87844672
HG01352	hap1	41.82	57.22	3029661384	2875708030	119877534	88685409
HG01352	hap2	41.82	59.38	3037085992	2882012014	118349303	89134277
HG01358	hap2	38.62	48.75	3029587694	2876639609	114919127	85033621
HG01358	hap1	39.90	52.44	2932540871	2885893852	117967868	87562865
HG01361	hap2	42.97	45.12	3025313608	2870493059	112176114	81388405
HG01361	hap1	43.19	47.18	3010067136	2855680841	113256234	83519004
HG01457	hap2	58.72	52.79	3025342371	2870109196	117918205	83268540
HG01457	hap1	58.72	57.34	2935859869	2889096677	117045599	83295973
HG01505	hap1	34.57	76.96	2898558372	2898570472	118406162	80032133
HG01505	hap2	34.57	55.22	2861687509	2861703909	114915506	87505872
HG01891	hap1	37.79	57.1	3043232268	2889086648	118444684	88084684

HG01891	hap2	38.04	81.11	3022952778	2868813113	109591443	83760879
HG01928	hap2	36.02	53.72	3025961049	2870513031	109496306	80973040
HG01928	hap1	37.29	45.7	2923053911	2876497072	115654207	88768932
HG01952	hap2	42.38	54.64	3020563005	2865702725	118458918	84972741
HG01952	hap1	43.94	44.25	2913263982	2866836370	108424068	75537368
HG01978	hap1	38.62	52.81	3055071491	2899825623	121715140	87504120
HG01978	hap2	38.66	60.49	3051869401	2896874887	110982434	83174741
HG02011	hap2	39.01	44.46	2887981713	2887995313	114868127	86368271
HG02011	hap1	39.01	81.24	2887920686	2887933286	124200252	88345708
HG02018	hap1	36.13	66.57	3248423247	3093889521	113625668	82402664
HG02018	hap2	36.13	59.25	3079868788	2924842365	120947289	85868307
HG02055	hap2	40.73	35.18	3019590183	2866059285	113822299	86115959
HG02055	hap1	41.69	34.12	2950246117	2900478632	123062070	94086612
HG02059	hap2	37.07	60.42	3029461372	2875972676	112850132	86038877
HG02059	hap1	37.07	54.12	3039177748	2885728773	113243529	82110306
HG02080	hap2	34.95	20.23	3033198064	2879184161	109689510	80742575
HG02080	hap1	35.05	24.29	3024505444	2870121016	116068956	83713945
HG02106	hap2	36.20	20.96	3032195460	2875635160	116703134	85794973
HG02106	hap1	36.20	36.13	3023979451	2873228152	113208690	84195134
HG02145	hap2	37.29	24.06	3030589160	2877353519	115072197	84531096
HG02145	hap1	38.47	19.77	2937649060	2899790942	114067136	80839649
HG02148	hap2	40.51	39.94	3036136215	2882416239	112975026	86907084
HG02148	hap1	40.65	41.87	3026089995	2871958601	109784911	80365803
HG02257	hap1	36.48	57.98	3042496887	2890251979	115489477	87132085
HG02257	hap2	36.62	59.04	3031071346	2877543368	119233028	87555914
HG02486	hap2	40.92	55.07	3030662429	2877024443	113550237	84383029
HG02486	hap1	42.31	58.49	2930935689	2881198909	118840286	88270782
HG02492	hap1	35.87	69.58	2911900640	2911916840	111971916	83839435

HG02492	hap2	35.87	46.62	2865321881	2865342681	114777504	83275387
HG02554	hap2	39.60	45.37	2912820170	2912840970	119579747	85992736
HG02554	hap1	39.60	48.67	2891624287	2891647587	117899295	91753746
HG02559	hap2	42.43	59.53	3040531780	2885732205	117234638	86766873
HG02559	hap1	42.66	57.25	3024156283	2869914115	114602944	84290366
HG02572	hap2	31.38	23.13	3055382050	2901037871	120308964	87061597
HG02572	hap1	32.58	19.53	2943627449	2899217714	110110019	82988274
HG02587	hap1	38.77	80.12	3087011490	2934947733	121161067	90439213
HG02587	hap2	38.77	87.27	3039448304	2883832616	118443840	83977455
HG02622	hap2	47.27	60.04	3046105980	2887745804	117545116	88781733
HG02622	hap1	47.32	51.21	3043426064	2889148421	121394882	90079729
HG02630	hap1	49.45	29.25	3053354263	2899081593	117577862	86171815
HG02630	hap2	49.64	25.38	3041877443	2887403611	119138944	86268988
HG02666	hap1	59.41	61.61	2894727135	2894739335	114519288	83737216
HG02666	hap2	59.41	49.85	2872204363	2872216763	120032620	93132834
HG02717	hap2	46.08	43.94	3038320685	2885107687	110166628	81399103
HG02717	hap1	47.52	46.51	2946102667	2896106218	119347412	88234966
HG02723	hap1	46.89	24.83	3049492048	2895290776	115216071	85744146
HG02723	hap2	47.24	22.44	3027203092	2873326702	113958153	86480462
HG02818	hap2	37.53	19.38	3037441458	2883980318	113029117	85642980
HG02818	hap1	37.75	18.22	3019578985	2866243817	111843059	81439559
HG02886	hap2	44.93	28.88	3049134634	2890862294	122528353	91033903
HG02886	hap1	44.96	29.11	3047149239	2894103758	116380785	89280566
HG02953	hap2	38.54	57.06	2913168917	2913187117	120897957	87310654
HG02953	hap1	38.54	54.4	2897206738	2897227638	116554937	86560490
HG03009	hap2	39.57	57.22	2876335532	2876348232	121315760	88967660
HG03009	hap1	39.57	56.6	2894319270	2894333670	116272015	82075362
HG03065	hap1	40.11	58.93	2914792744	2914803644	120476036	85113824

HG03065	hap2	40.11	67.3	2885370537	2885382237	127192707	92147480
HG03098	hap2	36.28	36.98	3059522785	2906767362	122247092	88924428
HG03098	hap1	37.81	34.38	2935645968	2889674227	110230573	86926067
HG03125	hap2	27.32	20.41	3023508899	2869591274	116970666	82892295
HG03125	hap1	27.32	17.35	3029338317	2875077583	113741101	86664515
HG03248	hap1	40.97	67.74	2950635091	2901834561	119677003	85498896
HG03248	hap2	40.97	61.62	3038411883	2884037265	116088446	86956784
HG03371	hap1	40.96	60.64	2879282931	2879295131	123718145	89003318
HG03371	hap2	40.96	65.06	2867524326	2867536026	118053728	86028400
HG03453	hap1	52.45	27.05	3050441176	2896239181	122155642	84261283
HG03453	hap2	52.49	26.38	3047998942	2894381861	120450495	90978477
HG03456	hap2	36.84	59.66	3059440543	2901641755	117004656	83941081
HG03456	hap1	36.84	60.44	2924124464	2874695843	117387283	83294262
HG03486	hap1	40.01	27.18	3049588884	2894877406	119259893	89061535
HG03486	hap2	40.20	24.25	3034915351	2880752366	122325282	85990884
HG03492	hap2	34.39	18.86	3023922307	2868930450	115110135	82764282
HG03492	hap1	35.57	20.16	2923665925	2878846529	113573000	80987238
HG03516	hap1	35.87	55.48	3067004974	2913201497	125868428	86834914
HG03516	hap2	36.26	44.77	3033479640	2880199890	119698171	86519929
HG03540	hap1	49.91	34.16	3065276644	2910604628	120666228	85940555
HG03540	hap2	50.19	30.47	3048418776	2894037756	115549641	84646043
HG03579	hap2	49.42	27.01	3035143227	2882155076	118266667	87608484
HG03579	hap1	50.90	27.54	2947164001	2895630355	119142729	85145260
HG03683	hap2	40.65	67.01	3035263093	2882282552	115105918	80599057
HG03683	hap1	40.65	78.81	3107922406	2951677282	119086199	89101201
HG03732	hap1	32.01	51.67	2890553742	2890571342	113354393	81830252
HG03732	hap2	32.01	55.69	2888698034	2888717134	114295138	82344336
HG03807	hap1	37.27	38.18	3017798208	2862719201	115952295	84629092

HG03807	hap2	37.27	34.4	3023125354	2868548587	118535693	82662325
HG04036	hap2	39.27	51.42	3026896339	2870512387	118418626	88777969
HG04036	hap1	39.27	54.42	3045939572	2890955712	116631802	84583831
HG04217	hap1	35.81	35.13	3028000242	2874383248	117565539	87083593
HG04217	hap2	35.81	32.61	3018924023	2864517713	115701773	87237562
NA12329	hap2	36.74	48.28	3053001015	2899238993	113965960	83314610
NA12329	hap1	36.74	89.88	3188981990	3034820489	117753630	84758918
NA12878	hap1	59.42	39.91	3022530992	2870983483	107265287	79222251
NA12878	hap2	59.42	39.67	3025655484	2874934035	105482852	76585996
NA18906	hap2	42.87	40.08	3055692855	2901357307	123323950	85889603
NA18906	hap1	43.00	43.52	3046330261	2893225836	126286502	88573507
NA19238	hap1	134.76	19.91	3023841743	2872583333	120156101	82348663
NA19238	hap2	134.76	15.95	2999735145	2849613319	104811045	74608571
NA19239	hap2	99.10	28.92	2880637615	2880679615	123112766	85368757
NA19239	hap1	99.10	19.2	2888519604	2888560604	117993971	81178844
NA19317	hap2	63.51	81.23	2848524863	2848538663	110932207	80632548
NA19317	hap1	63.51	134.47	2900425947	2900439947	131384711	90393701
NA19331	hap2	32.91	14.12	2855822265	2855856565	113038110	80073853
NA19331	hap1	32.91	17.91	2949451642	2949484642	116836093	86768211
NA19347	hap2	55.14	62.66	2848169441	2848179541	124973991	93984323
NA19347	hap1	55.14	80.47	2891537856	2891551356	122641564	87054591
NA19650	hap1	38.95	71.92	2874900915	2874912315	120139573	90029498
NA19650	hap2	38.95	63.44	2853654479	2853668479	112703389	79760248
NA19705	hap1	38.50	91.78	2927720191	2879237138	122305814	89733871
NA19705	hap2	38.50	58.96	3035888991	2882904172	118825773	84631229
NA19983	hap2	39.60	66.05	3023924888	2870194057	121323242	94499259
NA19983	hap1	39.60	71.26	3088603339	2935530377	128668973	86325825
NA20129	hap2	38.09	21.15	3045049146	2890691497	115605669	89371156

NA20129	hap1	38.30	22.42	3029071557	2875442608	116760831	85562386
NA20509	hap1	39.76	87.55	2879055923	2879069123	117523761	87258909
NA20509	hap2	39.76	70.37	2884838749	2884854249	114355203	85076374
NA20847	hap1	28.86	40.27	3038991407	2885248572	121651999	85553992
NA20847	hap2	28.86	52.35	3021934057	2867081430	115780548	84546838

Table S3. Enrichment of low-frequency SDs.

Region	Observed rare SD (bp)	Average null (bp)	Fold	empirical p-value
chr17_p-arm	1269399	304937.277	4.162820015	0
chr22_q-arm	2026627	462012.359	4.386521184	0
chr19_p-arm	1420605	311522.177	4.560205035	0.001
chr1_q-arm	3127843	1595123.316	1.96087849	0.007
chr16_p-arm	1269170	456141.176	2.782406121	0.01
chr10_p-arm	1397331	536318.974	2.605410339	0.014
chr15_q-arm	2087522	1102660.305	1.893168722	0.024
chr8_p-arm	1281668	579474.285	2.21177718	0.026
chr18_p-arm	547451	204242.967	2.680390949	0.059
chr19_q-arm	794922	422279.301	1.882455517	0.096
chr10_q-arm	1666055	1231909.353	1.352416877	0.173
chr7_q-arm	1647341	1261772.614	1.305576759	0.189
chr16_q-arm	1023678	751031.229	1.363029872	0.206
chr12_p-arm	611093	444229.465	1.375624645	0.213
chr9_p-arm	610583	585891.111	1.04214416	0.382
chr11_p-arm	719653	672795.12	1.069646581	0.389
chr21_q-arm	448446	442163.541	1.014208451	0.391
chr14_q-arm	1187882	1189970.341	0.998245048	0.447
chr2_q-arm	1924323	1935400.698	0.994276277	0.465
chr20_p-arm	215262	338528.076	0.635876358	0.644
chr17_q-arm	535798	727722.935	0.736266475	0.669
chr13_q-arm	977004	1244791.136	0.784873841	0.688
chr6_q-arm	1165770	1468854.653	0.793659194	0.689
chr9_q-arm	1088625	1367656.897	0.795978145	0.701
chr7_p-arm	556626	800749.508	0.695131242	0.706
chr20_q-arm	277307	477947.252	0.580204194	0.727
chr2_p-arm	870919	1221572.318	0.712949194	0.776
chr5_p-arm	354738	622937.822	0.569459724	0.792
chr4_p-arm	310020	662948.453	0.467638168	0.903
chr3_p-arm	686988	1224834.491	0.560882311	0.917
chr5_q-arm	1044933	1763694.869	0.592468129	0.934
chr4_q-arm	1037587	1810520.182	0.573087785	0.954
chr3_q-arm	753173	1420209.018	0.530325459	0.957
chr11_q-arm	429284	1059402.779	0.405213209	0.97
chr6_p-arm	224885	777921.501	0.289084438	0.981
chr12_q-arm	512110	1259002.86	0.406758409	0.984
chr18_q-arm	195348	776854.674	0.251460159	0.994

chr1_p-arm	565572	1605161.046	0.352345954	0.998
chr8_q-arm	386059	1298185.792	0.297383473	0.999

Table S4. Orientation and pairwise dispersion of polymorphic and single SDs.

Strand	Distance	Polymorphic SDs		Single SDs	
		Mean count	Total count	Mean count	Total count
Inverted	Clustered (<1 Mbp)	249.7	46,445	2.3	65
Tandem	Clustered (<1 Mbp)	421.9	78,482	6.7	1,032
Inverted	Interspersed	336.4	62,573	2.6	31
Tandem	Interspersed	284.5	52,916	2.8	73
Inverted	>50 Mbp	36.9	6,870	1.8	11
Tandem	>50 Mbp	21.4	3,983	2.7	19

Table S5. Summary of copy number estimation using asm-based method.

Gene Family	Contiguously assembled samples	Number of Genes	Mean	Median	Min	Max	StDev	Variance	Dispersion	Mean (African samples)	mean (non-African samples)	p-value (FDR controlled)
ANTXR2	91	1	2.87	3	2	4	0.78	0.60	0.21	3.50	2.36	0.00
CCDC127	89	1	2.33	2	2	4	0.56	0.31	0.13	2.76	2.02	0.00
ANKRD31	91	1	3.26	3	2	4	0.74	0.55	0.17	3.83	2.86	0.00
PDLIM3	91	1	2.92	3	1	4	0.81	0.65	0.22	3.50	2.46	0.00
WDR72	91	1	2.24	2	2	4	0.54	0.30	0.13	2.63	2.00	0.00
STARD9	89	1	2.17	2	2	3	0.38	0.14	0.07	2.46	2.00	0.00
MAP3K5	91	1	2.26	2	2	4	0.47	0.22	0.10	2.53	2.02	0.00
KCNQ5	90	1	2.22	2	2	4	0.51	0.26	0.12	2.59	2.00	0.00
LGALS9	90	4	6.21	6	4	8	0.99	0.98	0.16	6.93	5.76	0.00
GSDMC	91	1	3.07	3	2	4	0.81	0.66	0.22	3.53	2.62	0.00
CHODL	91	1	2.84	3	2	4	0.75	0.56	0.20	2.37	3.20	0.00
UIMC1	90	1	2.73	3	2	4	0.76	0.58	0.21	2.24	3.06	0.00
GALNTL6	89	1	2.52	2	2	4	0.69	0.48	0.19	2.97	2.22	0.00
PRH1	89	1	2.66	2	2	4	0.78	0.61	0.23	2.21	3.02	0.00
SLC25A48	90	1	3.42	4	2	4	0.70	0.49	0.14	3.83	3.10	0.00
TBC1D3	46	15	30.33	30	23	42	3.89	15.11	0.50	33.64	28.15	0.00
ANKRD36	63	2	5.25	5	3	6	0.84	0.71	0.13	5.76	4.76	0.00
RAG1	91	1	2.32	2	2	4	0.53	0.29	0.12	2.63	2.10	0.00
RPTOR	82	1	2.34	2	2	5	0.61	0.38	0.16	2.72	2.09	0.00
PRSS2	90	1	3.22	3	2	4	0.78	0.60	0.19	3.63	2.86	0.00
YBEY	89	1	2.27	2	2	4	0.52	0.27	0.12	2.55	2.08	0.00
EPPK1	91	1	2.18	2	2	4	0.41	0.17	0.08	2.40	2.02	0.00
PHRF1	84	1	2.46	2	2	4	0.63	0.40	0.16	2.11	2.73	0.00
SUZ12	89	2	4.28	4	4	6	0.54	0.30	0.07	4.55	4.06	0.00
FRAS1	91	1	2.47	2	2	4	0.62	0.39	0.16	2.13	2.72	0.00
DMXL2	91	1	2.13	2	2	3	0.34	0.12	0.05	2.30	2.00	0.00
OPCML	87	1	2.11	2	2	4	0.36	0.13	0.06	2.33	2.00	0.00
FFAR1	91	2	3.75	4	3	4	0.44	0.19	0.05	4.00	3.60	0.00
C1orf159	88	1	2.83	3	2	4	0.76	0.58	0.20	3.24	2.54	0.00
CTNNA2	90	1	2.23	2	2	4	0.50	0.25	0.11	2.45	2.04	0.00
SCAPER	90	1	3.49	4	2	4	0.67	0.45	0.13	3.83	3.22	0.00
SHC2	81	1	2.12	2	2	4	0.37	0.13	0.06	2.33	2.00	0.00
FAM118A	91	1	2.21	2	1	4	0.51	0.26	0.12	1.97	2.40	0.00
VRK3	88	1	2.23	2	2	4	0.47	0.22	0.10	2.47	2.06	0.00
B3GALT1	91	1	2.32	2	2	4	0.56	0.31	0.13	2.63	2.14	0.00
MAMDC2	91	1	2.22	2	2	4	0.49	0.24	0.11	2.00	2.40	0.00
OSBPL6	91	1	2.11	2	1	3	0.35	0.12	0.06	2.23	1.98	0.00
ENPP7P10	78	8	15.23	15	13	19	1.15	1.32	0.09	15.81	14.76	0.00

ASTN2	89	1	2.15	2	1	4	0.41	0.17	0.08	2.28	2.00	0.00
CSMD1	83	1	3.34	3	2	4	0.70	0.49	0.15	3.67	3.07	0.00
KCNIP4	90	1	2.18	2	2	3	0.38	0.15	0.07	2.33	2.04	0.00
ROBO1	91	1	2.30	2	2	4	0.53	0.28	0.12	2.50	2.10	0.00
LOC124906470	91	1	3.32	3	2	9	1.27	1.62	0.49	3.93	2.94	0.00
DNER	91	1	2.20	2	1	4	0.48	0.23	0.10	1.97	2.34	0.00
NPEPPS	59	3	6.25	6	4	12	1.50	2.26	0.36	7.28	5.66	0.00
KSR2	91	1	2.79	3	2	4	0.71	0.50	0.18	3.17	2.60	0.00
GPAT2	70	1	2.51	2	1	4	0.70	0.49	0.19	2.17	2.76	0.01
PRR23D2	82	3	3.71	3.5	0	11	2.24	5.00	1.35	4.66	2.88	0.01
ABCA1	91	1	2.18	2	1	4	0.49	0.24	0.11	1.97	2.34	0.01
PGBD2	91	1	2.07	2	1	3	0.33	0.11	0.05	2.20	1.96	0.01
TRIM66	90	1	2.13	2	2	5	0.45	0.21	0.10	2.31	2.00	0.01
KANSL1	78	1	2.21	2	2	4	0.49	0.24	0.11	2.00	2.38	0.01
TANC1	91	1	3.20	3	2	4	0.75	0.56	0.18	3.50	2.94	0.01
THSD4	88	1	2.90	3	2	4	0.76	0.58	0.20	3.25	2.65	0.01
HP	89	1	3.00	3	2	4	0.78	0.61	0.20	2.68	3.26	0.01
DLG1	90	1	3.73	4	2	4	0.56	0.31	0.08	3.97	3.55	0.01
GRIN2D	89	1	2.37	2	2	4	0.63	0.40	0.17	2.62	2.18	0.01
SLC44A5	87	1	2.45	2	2	4	0.64	0.41	0.17	2.12	2.60	0.01
SORD	89	2	4.83	5	4	6	0.77	0.60	0.12	5.14	4.55	0.01
MUC4	91	1	2.26	2	2	4	0.47	0.22	0.10	2.40	2.10	0.01
PRIM2	88	3	7.50	7.5	6	10	0.87	0.76	0.10	7.93	7.24	0.01
AFAP1	80	1	2.36	2	2	4	0.60	0.36	0.15	2.11	2.57	0.01
NCOA3	91	1	2.34	2	2	4	0.58	0.34	0.14	2.57	2.14	0.01
BAHCC1	87	1	2.32	2	2	6	0.64	0.41	0.18	2.54	2.10	0.01
DYM	78	1	2.18	2	2	4	0.42	0.18	0.08	2.36	2.05	0.01
SKAP2	91	1	2.19	2	1	4	0.45	0.20	0.09	2.33	2.04	0.01
TBC1D32	91	1	2.12	2	2	3	0.33	0.11	0.05	2.23	2.02	0.01
CAB39L	91	1	3.01	3	2	4	0.78	0.61	0.20	2.70	3.24	0.01
CLTCL1	91	1	2.15	2	2	3	0.36	0.13	0.06	2.00	2.26	0.01
MSH2	90	1	2.18	2	1	4	0.46	0.22	0.10	2.00	2.33	0.01
LOC124907512	91	2	2.90	3	0	8	1.73	2.98	1.03	2.17	3.36	0.01
CCSER1	91	1	3.52	4	2	4	0.58	0.34	0.10	3.73	3.32	0.01
OR2T6	83	1	2.13	2	2	4	0.38	0.14	0.07	2.25	2.02	0.01
GTF2H2	75	2	4.51	5	2	7	0.96	0.93	0.21	4.93	4.20	0.01
PRAMEF1	78	7	13.38	14	10	14	0.98	0.97	0.07	12.93	13.64	0.01
CNOT2	91	1	2.14	2	2	3	0.35	0.12	0.06	2.00	2.24	0.02
DND1	91	1	2.15	2	2	3	0.36	0.13	0.06	2.00	2.24	0.02
LSAMP	91	1	2.14	2	2	4	0.38	0.15	0.07	2.00	2.26	0.02
SAR1B	91	1	2.34	2	2	4	0.56	0.32	0.14	2.13	2.52	0.02

TRIM64	56	4	7.95	8	5	10	0.88	0.78	0.10	7.47	8.23	0.02
GRAMD4	90	1	2.22	2	2	4	0.44	0.20	0.09	2.40	2.10	0.02
TRAPPC12	91	1	2.91	3	1	4	0.78	0.61	0.21	3.27	2.74	0.02
cat_AL732372.3	86	4	7.07	7	4	10	0.93	0.87	0.12	7.39	6.84	0.02
RASA3	73	1	2.48	2	2	4	0.63	0.39	0.16	2.25	2.69	0.02
DEFB107A	83	3	4.87	5	2	9	1.46	2.14	0.44	5.36	4.36	0.02
TYW1	80	2	4.53	4	4	6	0.64	0.40	0.09	4.71	4.33	0.02
F13A1	91	1	2.24	2	1	4	0.52	0.27	0.12	2.40	2.08	0.02
DEFB4A	87	3	4.48	4	2	7	1.20	1.44	0.32	4.93	4.13	0.02
ZDHHC14	91	1	3.43	4	2	4	0.69	0.47	0.14	3.67	3.22	0.02
ZNF676	91	1	3.08	3	2	4	0.78	0.61	0.20	3.33	2.84	0.02
SEC23A	91	1	2.37	2	2	4	0.57	0.33	0.14	2.17	2.54	0.02
TAF11L10	78	2	3.38	4	0	7	1.28	1.64	0.49	3.92	2.98	0.02
RIMS1	91	1	3.78	4	2	4	0.47	0.22	0.06	3.97	3.70	0.02
ZDHHC11	81	1	2.14	2	2	4	0.38	0.14	0.07	2.28	2.07	0.02
ZNF705A	78	7	11.56	11	8	18	2.11	4.46	0.39	12.46	10.85	0.02
OR4F3	87	1	2.84	3	2	5	0.87	0.76	0.27	3.17	2.60	0.03
FSTL4	90	1	2.43	2	2	4	0.56	0.32	0.13	2.24	2.60	0.03
CDH4	87	1	2.16	2	2	4	0.43	0.18	0.08	2.00	2.27	0.03
EYS	88	1	2.24	2	1	4	0.50	0.25	0.11	2.38	2.08	0.03
STRC	86	2	3.99	4	2	8	0.52	0.27	0.07	3.83	4.09	0.03
CCDC148	91	1	2.60	3	2	4	0.66	0.44	0.17	2.87	2.42	0.03
DIP2B	88	1	2.33	2	1	4	0.64	0.41	0.17	2.11	2.49	0.03
LOC105375434	88	1	2.72	3	2	4	0.71	0.50	0.19	3.00	2.54	0.03
LOC124902694	91	1	3.73	4	2	5	0.54	0.29	0.08	3.93	3.62	0.03
OTOAP1	84	2	3.70	4	2	6	0.58	0.33	0.09	3.87	3.57	0.03
cat_LINC02203	83	1	2.88	3	1	6	1.12	1.25	0.44	3.31	2.67	0.04
MRPL48	90	1	2.24	2	2	4	0.48	0.23	0.10	2.38	2.16	0.04
CLVS1	89	1	3.76	4	2	4	0.52	0.27	0.07	3.96	3.66	0.04
ZRANB3	90	1	2.38	2	2	4	0.57	0.33	0.14	2.21	2.52	0.04
RPH3AL	84	1	2.08	2	1	4	0.52	0.27	0.13	2.25	1.93	0.04
RDM1	88	2	3.78	4	2	4	0.47	0.22	0.06	3.97	3.71	0.04
DEFB106A	85	3	4.61	5	2	8	1.28	1.65	0.36	5.07	4.28	0.04
SPAG11A	85	3	4.49	5	2	7	1.20	1.44	0.32	4.89	4.19	0.04
GLIS3	90	1	2.12	2	1	4	0.39	0.15	0.07	2.21	2.02	0.04
ST8SIA5	91	1	2.12	2	2	3	0.33	0.11	0.05	2.00	2.18	0.04
EHMT1	60	1	2.45	2	2	4	0.59	0.35	0.14	2.71	2.31	0.04
PDXDC1	49	1	3.33	3	2	5	0.72	0.52	0.16	3.72	3.13	0.04
DIP2C	51	1	2.24	2	2	4	0.51	0.26	0.12	2.50	2.10	0.05
FAM90A1	29	49	46.07	49	15	78	17.59	309.57	6.72	55.80	40.17	0.05
REXO1L1P	58	15	26.81	27	12	36	5.25	27.59	1.03	28.62	25.46	0.05

CNTNAP3B	83	2	3.66	4	2	6	0.82	0.67	0.18	3.38	3.84	0.05
ANKRD62	91	1	2.53	2	2	4	0.56	0.32	0.13	2.73	2.40	0.05
FRG2P	84	5	8.61	9	7	11	0.92	0.84	0.10	8.32	8.82	0.06
SMG1	45	11	21.09	21	18	22	1.14	1.31	0.06	21.38	20.56	0.06
ARFGEF2	91	1	2.11	2	2	4	0.35	0.12	0.06	2.00	2.18	0.06
SBF2	91	1	2.19	2	1	4	0.45	0.20	0.09	2.03	2.26	0.06
BMS1	47	13	25.26	25	21	29	1.85	3.41	0.14	26.00	24.68	0.06
WDR70	91	1	2.12	2	1	4	0.42	0.17	0.08	2.20	2.00	0.06
ERICH1	90	1	2.11	2	2	4	0.35	0.12	0.06	2.00	2.18	0.06
PDPR	81	4	5.09	5	4	7	0.78	0.60	0.12	4.82	5.23	0.06
GRAP	27	2	3.26	3	2	5	0.90	0.81	0.25	3.78	2.94	0.06
MUC1	89	1	2.42	2	1	4	0.64	0.40	0.17	2.59	2.29	0.06
CBFA2T3	68	1	2.16	2	2	4	0.41	0.17	0.08	2.25	2.05	0.06
PGA3	86	2	2.87	3	1	4	0.82	0.68	0.24	2.68	3.11	0.06
MUC17	88	1	2.50	2	2	4	0.61	0.37	0.15	2.71	2.39	0.07
MUC19	89	1	2.31	2	2	4	0.58	0.33	0.14	2.14	2.46	0.08
AKAP3	91	1	2.67	3	2	4	0.67	0.45	0.17	2.87	2.56	0.08
ENOX1	90	1	2.27	2	2	4	0.51	0.27	0.12	2.13	2.41	0.08
LILRA3	88	4	7.82	8	6	16	1.08	1.16	0.15	7.93	7.76	0.08
LOC124900584	83	1	2.39	2	1	4	0.81	0.65	0.27	2.53	2.10	0.08
DEFB104A	85	3	4.53	5	2	7	1.20	1.44	0.32	4.88	4.27	0.08
CLPTM1L	91	1	2.23	2	2	4	0.47	0.22	0.10	2.10	2.34	0.08
NBPF10	79	4	9.89	10	7	14	1.27	1.62	0.16	10.32	9.66	0.09
TRIM43CP	59	4	5.88	6	2	9	1.53	2.35	0.40	5.38	6.25	0.09
ZFPM1	88	1	2.30	2	2	3	0.46	0.21	0.09	2.14	2.36	0.09
USP17L10	71	5	9.55	10	5	18	1.92	3.68	0.39	10.33	9.05	0.10
SCFD1	91	1	2.19	2	2	4	0.45	0.20	0.09	2.07	2.28	0.10
TMEM191C	85	2	4.42	4	4	8	0.66	0.44	0.10	4.33	4.51	0.10
KLRC1	91	1	3.67	4	2	4	0.52	0.27	0.07	3.53	3.76	0.10
RASA4	65	4	8.03	8	6	10	0.68	0.47	0.06	8.18	7.91	0.10
C2orf78	72	1	2.19	2	1	4	0.52	0.27	0.12	2.04	2.29	0.10
GUSBP3	66	4	6.64	7	4	9	1.30	1.68	0.25	6.26	6.91	0.10
LOC124906322	89	1	3.13	3	2	6	0.99	0.98	0.31	2.83	3.27	0.11
FAM20C	91	1	2.26	2	2	4	0.51	0.26	0.12	2.13	2.38	0.11
LOC124905373	88	3	3.18	3	1	14	1.86	3.46	1.09	3.76	2.77	0.11
KATNAL2	44	1	2.75	3	2	4	0.69	0.47	0.17	3.07	2.62	0.11
LOC124908084	61	2	3.30	3	1	5	0.95	0.91	0.28	3.55	3.09	0.11
B3GNTL1	89	2	2.16	2	2	3	0.37	0.13	0.06	2.07	2.23	0.13
ANKRD30A	77	1	2.32	2	2	4	0.55	0.30	0.13	2.19	2.43	0.13
MUC20	87	1	3.20	3	2	5	0.89	0.79	0.25	3.00	3.40	0.13
ANHX	91	1	2.04	2	2	5	0.33	0.11	0.05	2.13	2.00	0.13

ZNF10	91	1	2.04	2	2	5	0.33	0.11	0.05	2.13	2.00	0.13
ZNF268	91	1	2.04	2	2	5	0.33	0.11	0.05	2.13	2.00	0.13
CACNA1C	89	1	2.17	2	2	3	0.38	0.14	0.07	2.29	2.12	0.14
CROCC	90	1	2.29	2	2	4	0.50	0.25	0.11	2.40	2.20	0.14
DUSP22	83	2	3.73	4	2	5	0.50	0.25	0.07	3.81	3.62	0.15
CKMT1A	86	2	3.99	4	2	8	0.52	0.27	0.07	3.86	4.06	0.15
RHPN2	66	5	11.03	10.5	8	20	2.13	4.55	0.41	11.33	10.58	0.18
HSPA6	88	3	3.99	4	2	8	0.78	0.61	0.15	3.90	4.09	0.19
FKBP5	90	1	2.84	3	2	5	0.76	0.58	0.20	3.00	2.68	0.19
LOC128966677	78	1	2.03	2	0	6	1.37	1.87	0.92	2.32	1.77	0.19
LOC128966590	71	2	3.83	4	2	7	1.45	2.11	0.55	4.21	3.64	0.19
SULT1A1	90	1	2.37	2	2	5	0.63	0.39	0.17	2.47	2.33	0.19
LOC100652777	77	1	2.47	2	1	4	0.64	0.41	0.17	2.60	2.32	0.20
REXO1L3P	60	5	8.18	8	2	16	2.90	8.42	1.03	9.09	7.97	0.20
EEF2KMT	82	15	28.61	29	25	32	1.42	2.02	0.07	28.96	28.44	0.20
ALG1	84	11	20.48	20	18	24	1.28	1.65	0.08	20.82	20.24	0.21
CEP72	91	1	2.14	2	2	3	0.35	0.12	0.06	2.20	2.08	0.21
ACSL3	90	1	2.12	2	2	3	0.33	0.11	0.05	2.20	2.08	0.23
CPSF2	91	1	2.55	2	2	4	0.62	0.38	0.15	2.40	2.64	0.23
SLX1A	67	4	7.78	8	5	8	0.65	0.42	0.05	7.88	7.71	0.24
LOC124904409	83	3	5.52	6	4	7	0.67	0.45	0.08	5.37	5.58	0.24
FCGR2B	90	2	2.74	3	2	6	0.79	0.62	0.23	2.67	2.84	0.24
TTN	91	1	2.14	2	2	4	0.38	0.15	0.07	2.17	2.08	0.24
MSH3	91	1	2.41	2	2	4	0.56	0.31	0.13	2.27	2.44	0.24
ZMYM5	89	1	2.02	2	1	5	0.34	0.11	0.06	1.96	2.06	0.25
APOBEC3A	91	2	3.76	4	2	4	0.52	0.27	0.07	3.87	3.66	0.25
PSPC1	91	1	2.02	2	1	5	0.33	0.11	0.05	1.97	2.06	0.25
PTPRD	88	1	2.20	2	2	4	0.43	0.19	0.09	2.31	2.15	0.25
PPP1R26	56	3	7.30	7	5	12	1.51	2.29	0.31	7.81	7.03	0.26
PTPRN2	65	1	2.37	2	1	4	0.60	0.36	0.15	2.26	2.51	0.26
NPY4R	91	1	2.12	2	1	4	0.51	0.26	0.12	2.03	2.22	0.27
DEFA1	82	3	5.68	6	4	9	0.68	0.47	0.08	5.52	5.76	0.27
UPK3BL2	73	1	2.15	2	1	6	0.64	0.41	0.19	2.00	2.24	0.28
DRD5P1	80	2	4.00	4	2	6	0.64	0.41	0.10	3.88	4.09	0.28
CTRB2	91	1	2.46	2	0	4	0.95	0.90	0.36	2.60	2.32	0.30
RAP1A	91	1	2.16	2	2	3	0.37	0.14	0.06	2.23	2.12	0.30
PDE4DIP	76	3	6.58	7	4	8	0.87	0.75	0.11	6.44	6.74	0.30
EXD3	35	1	2.14	2	1	4	0.49	0.24	0.11	2.33	2.06	0.30
TCAF1	87	1	2.21	2	2	4	0.44	0.19	0.09	2.11	2.23	0.30
FCGR3A	89	3	4.07	4	2	8	0.82	0.68	0.17	3.97	4.13	0.30
DDT	81	1	2.20	2	2	4	0.46	0.21	0.10	2.30	2.12	0.32

GSTT2	81	1	2.20	2	2	4	0.46	0.21	0.10	2.30	2.12	0.32
RHCE	90	2	3.59	4	2	5	0.62	0.38	0.11	3.70	3.49	0.33
NBPF26	73	6	11.58	12	9	17	1.25	1.55	0.13	11.46	11.66	0.34
ANKS1B	90	1	3.69	4	2	4	0.49	0.24	0.06	3.60	3.76	0.34
EIF3C	85	2	3.81	4	2	5	0.48	0.23	0.06	3.90	3.76	0.35
ZNF140	91	1	2.13	2	2	4	0.37	0.14	0.06	2.20	2.12	0.35
POTEB2	34	12	28.35	28	24	36	2.74	7.51	0.26	29.00	27.92	0.36
GSTT4	86	2	4.22	4	4	6	0.47	0.22	0.05	4.32	4.15	0.36
LOC124906869	84	1	2.15	2	0	4	0.81	0.66	0.31	1.96	2.24	0.37
TWIST2	87	1	2.28	2	2	4	0.47	0.23	0.10	2.21	2.35	0.37
ANXA8L1	88	2	4.15	4	3	6	0.52	0.27	0.06	4.07	4.20	0.39
SLC25A18	91	1	2.57	3	2	4	0.60	0.36	0.14	2.67	2.52	0.39
POLR2J2	71	2	4.20	4	3	11	1.04	1.07	0.26	3.96	4.33	0.39
LSP1P4	70	3	6.16	6	4	9	1.06	1.12	0.18	6.34	6.10	0.40
KLHL24	90	1	2.76	3	2	4	0.72	0.52	0.19	2.86	2.66	0.40
CDK11A	85	2	3.86	4	1	4	0.44	0.19	0.05	3.86	3.89	0.41
OR4F17	91	2	5.23	5	1	10	1.67	2.78	0.53	5.50	4.98	0.41
TAFA5	90	1	2.18	2	2	3	0.38	0.15	0.07	2.21	2.12	0.44
PRAMEF10	84	7	13.27	14	7	14	1.22	1.48	0.11	13.00	13.42	0.44
ABHD17A	56	8	14.54	15	11	20	1.50	2.25	0.16	14.31	14.63	0.45
PRR23D1	88	1	2.18	2	0	4	0.74	0.54	0.25	2.07	2.25	0.45
OR4Q3	82	1	2.93	3	2	5	0.81	0.66	0.23	2.81	3.00	0.45
OR7E100P	85	4	6.78	7	6	10	0.90	0.82	0.12	6.93	6.66	0.45
SULT1A2	73	2	3.60	4	2	6	0.66	0.44	0.12	3.65	3.53	0.45
ADAMTSL3	90	1	3.57	4	2	4	0.67	0.45	0.13	3.50	3.63	0.45
PRKN	61	1	2.28	2	2	3	0.45	0.20	0.09	2.35	2.22	0.45
FCGBP	87	1	2.17	2	1	3	0.41	0.17	0.08	2.11	2.21	0.45
MRGPRX1	91	4	6.46	6	5	8	0.69	0.47	0.07	6.37	6.52	0.46
FAM25A	90	3	6.17	6	5	8	0.57	0.32	0.05	6.10	6.24	0.48
GGT1	77	5	9.27	9	5	14	2.00	4.02	0.43	9.42	9.03	0.48
GALNT9	88	1	2.39	2	1	4	0.56	0.31	0.13	2.45	2.35	0.49
RIMBP3	86	2	4.45	4	3	8	0.79	0.63	0.14	4.45	4.48	0.49
cat_PCDHB11	90	3	6.44	6	6	10	0.62	0.38	0.06	6.47	6.49	0.49
ZCWPW2	91	1	2.16	2	2	4	0.40	0.16	0.07	2.20	2.14	0.49
ZNF888	91	1	2.27	2	1	4	0.52	0.27	0.12	2.37	2.26	0.49
DLG2	90	1	2.46	2	2	4	0.62	0.39	0.16	2.53	2.39	0.49
CLEC18A	78	6	9.29	9	7	12	1.03	1.07	0.11	9.14	9.39	0.50
PSG1	90	4	7.46	8	5	9	0.80	0.63	0.08	7.47	7.55	0.50
PRAMEF11	84	7	13.27	14	7	16	1.35	1.82	0.14	13.00	13.43	0.51
OTOP1	69	4	7.88	8	6	10	0.83	0.69	0.09	8.00	7.81	0.51

ARHGEF35	79	2	4.06	4	3	7	0.69	0.47	0.12	3.96	4.15	0.52
SPTLC2	91	1	2.23	2	2	4	0.50	0.25	0.11	2.27	2.22	0.55
MYO5BP1	90	1	2.37	2	0	4	0.76	0.57	0.24	2.27	2.40	0.56
ZNG1A	82	8	16.10	16	14	19	1.01	1.03	0.06	15.96	16.16	0.56
CTRB1	90	1	2.38	2	1	4	0.57	0.33	0.14	2.34	2.42	0.57
MST1	86	2	4.23	4	3	6	0.68	0.46	0.11	4.24	4.11	0.57
TRIM48	41	9	17.71	18	13	20	1.27	1.61	0.09	17.27	17.79	0.57
GRID2	91	1	3.38	3	2	4	0.66	0.44	0.13	3.50	3.34	0.57
LPA	76	1	2.67	3	2	4	0.68	0.46	0.17	2.70	2.62	0.57
SAMD11	88	1	2.16	2	2	4	0.40	0.16	0.07	2.21	2.12	0.59
STON1	91	1	3.49	4	2	4	0.62	0.39	0.11	3.57	3.44	0.59
PTGER4P2-CDK2AP2P2	67	12	22.81	23	19	27	1.58	2.49	0.11	22.73	22.87	0.61
NBPF4	75	3	5.84	6	5	7	0.55	0.30	0.05	5.76	5.88	0.61
CRTC1	90	1	2.98	3	2	4	0.75	0.56	0.19	2.90	3.02	0.62
OR2A1	78	1	2.10	2	2	6	0.52	0.28	0.13	2.04	2.17	0.62
INTS4	65	3	6.35	6	4	8	0.67	0.45	0.07	6.45	6.35	0.64
ST6GAL1	91	1	2.33	2	2	4	0.56	0.31	0.13	2.27	2.34	0.65
PIWIL1	90	1	2.48	2	1	4	0.67	0.45	0.18	2.41	2.52	0.65
GPRIN2	90	1	2.09	2	1	5	0.57	0.33	0.16	2.03	2.16	0.67
DMBT1	87	1	3.71	4	2	4	0.53	0.28	0.07	3.74	3.63	0.69
ZNF705G	78	1	2.41	2	1	5	0.73	0.53	0.22	2.50	2.37	0.69
SYT15	86	1	2.06	2	1	4	0.47	0.22	0.11	2.10	2.09	0.70
RAB38	82	1	2.22	2	2	4	0.45	0.20	0.09	2.18	2.20	0.70
FOXD4	88	5	9.90	10	8	12	0.76	0.58	0.06	9.97	9.85	0.73
DDTL	81	1	2.62	2	2	4	0.72	0.51	0.20	2.59	2.72	0.74
NBPF1	65	8	19.34	19	16	23	1.67	2.79	0.14	19.45	19.18	0.77
PRB1	90	4	7.84	8	4	9	0.65	0.42	0.05	7.76	7.96	0.77
PLAAT2	91	1	2.44	2	1	4	0.65	0.43	0.17	2.37	2.44	0.77
SPATA31E1	83	3	6.34	6	4	10	1.11	1.23	0.19	6.23	6.39	0.78
OR4M1	82	1	2.93	3	2	5	0.81	0.66	0.23	2.85	2.93	0.78
LOC124901712	79	1	2.20	2	1	6	0.97	0.93	0.42	2.29	2.22	0.78
ARHGAP23	79	3	5.61	6	4	7	0.67	0.45	0.08	5.68	5.58	0.78
PRAMEF18	85	3	7.12	7	4	8	0.79	0.63	0.09	7.00	7.15	0.78
HIC2	68	2	4.40	4	3	6	0.60	0.36	0.08	4.36	4.42	0.81
SPDYE1	49	22	38.35	38	32	43	2.90	8.40	0.22	38.38	37.96	0.81
CCDC77	89	1	2.79	3	2	4	0.75	0.56	0.20	2.80	2.73	0.81
PKHD1	89	1	3.03	3	2	4	0.75	0.56	0.18	3.07	3.00	0.81
PSG11	88	6	11.74	12	6	14	1.06	1.11	0.09	11.83	11.85	0.82
PRR20A	10	4	6.20	6	4	8	1.48	2.18	0.35	6.00	6.40	0.84
ASCC3	91	1	2.70	3	2	4	0.71	0.50	0.18	2.63	2.72	0.84

CLPSL1	91	1	2.37	2	1	4	0.59	0.35	0.15	2.37	2.42	0.85
ENPP7P5	88	2	3.36	3	2	5	0.78	0.60	0.18	3.36	3.43	0.85
ABCG8	91	1	2.11	2	2	4	0.35	0.12	0.06	2.10	2.08	0.85
PCDH15	89	1	2.13	2	2	4	0.38	0.14	0.07	2.10	2.10	0.86
GOLGA6L24	82	4	8.39	8	5	12	1.35	1.82	0.22	8.26	8.45	0.86
NPIPA1	12	24	47.00	47	42	54	2.73	7.45	0.16	47.83	45.67	0.86
NAIP	71	2	4.62	5	2	7	0.99	0.98	0.21	4.62	4.58	0.86
OR4N2	86	1	2.95	3	2	5	0.81	0.66	0.22	2.90	2.96	0.86
GTF2IRD2	72	3	6.11	6	5	10	0.62	0.38	0.06	6.22	6.06	0.87
SHLD2	90	2	4.10	4	3	6	0.56	0.32	0.08	4.10	4.14	0.87
LOC124908083	45	5	8.38	8	4	11	1.43	2.06	0.25	8.25	8.54	0.87
KIR2DL1	87	6	12.90	13	9	14	1.24	1.54	0.12	12.82	12.88	0.88
C4A	91	2	3.77	4	2	5	0.62	0.38	0.10	3.77	3.78	0.88
SERF1A	70	2	3.41	4	1	5	0.84	0.71	0.21	3.48	3.39	0.88
SMN1	64	2	3.48	4	2	5	0.76	0.57	0.16	3.48	3.52	0.88
ZNF658	88	2	4.02	4	2	6	0.48	0.23	0.06	4.10	4.02	0.88
DYNC2I1	90	1	2.69	3	2	4	0.70	0.49	0.18	2.70	2.67	0.88
OR4K1	87	1	2.87	3	2	5	0.74	0.55	0.19	2.90	2.85	0.88
TP53TG3	37	5	9.30	9	6	12	1.39	1.94	0.21	9.20	9.05	0.88
DLC1	88	1	2.14	2	2	3	0.35	0.12	0.06	2.13	2.15	0.89
C1QTNF3	81	1	2.26	2	2	4	0.54	0.29	0.13	2.21	2.28	0.91
ASNS	74	6	11.46	11	9	14	1.02	1.05	0.09	11.40	11.43	0.91
FAM153A	84	4	7.88	8	6	12	0.88	0.78	0.10	7.96	7.81	0.91
PKD1	31	3	5.81	6	5	8	0.65	0.43	0.07	5.93	5.83	0.93
LOC124908120	85	2	3.42	3	2	6	0.89	0.79	0.23	3.43	3.50	0.95
OR4K2	89	1	2.91	3	2	5	0.78	0.61	0.21	2.90	2.92	0.96
CRYBG1	91	1	2.44	2	2	4	0.58	0.34	0.14	2.40	2.44	0.96
MUC12	88	1	2.83	3	2	4	0.71	0.51	0.18	2.79	2.78	0.96
PDGFD	91	1	2.23	2	2	4	0.47	0.22	0.10	2.23	2.26	0.99
AMY1A	44	10	11.48	11	7	19	2.74	7.51	0.65	11.23	11.39	0.99
ARL17B	72	4	7.56	7.5	6	11	1.15	1.32	0.17	7.41	7.60	0.99
AQP7	82	7	13.66	14	11	16	0.85	0.72	0.05	13.64	13.64	0.99
inferred_1_LOC128966740	60	2	3.43	4	0	6	1.21	1.47	0.43	3.29	3.44	0.99

Table S6. Iso-Seq resource of 563 million full-length cDNA (FLNC) sequence generated from 241 libraries and 67 distinct tissues.

Library	Tissue/cell type	Number of cDNA reads	Source
ENCFF861BKY	A673	1786525	ENCODE
ENCFF168MIB	A673	1378362	ENCODE
ENCFF417ALN	adrenal gland	1594860	ENCODE
ENCFF902BIU	aorta	2043939	ENCODE
ENCFF144KHH	aorta	1175321	ENCODE
ENCFF316EZQ	astrocyte	808252	ENCODE
ENCFF474GEK	astrocyte	469341	ENCODE
ENCFF827OXR	Caco-2	1733868	ENCODE
ENCFF649CYY	Caco-2	1449402	ENCODE
ENCFF548JGS	Calu3	1733596	ENCODE
ENCFF569KOA	Calu3	1720338	ENCODE
ENCFF680XXE	cardiac septum	1845269	ENCODE
ENCFF011BFA	chondrocyte	1589522	ENCODE
ENCFF342HOS	chondrocyte	1346171	ENCODE
ENCFF352CGL	chondrocyte	916864	ENCODE
ENCFF142LPL	endodermal cell	2536086	ENCODE
ENCFF561HIY	endodermal cell	2420444	ENCODE
ENCFF712CBL	endodermal cell	1511682	ENCODE
ENCFF235QXW	endodermal cell	1492674	ENCODE
ENCFF731HST	endodermal cell	1187556	ENCODE
ENCFF096UHO	endothelial cell of umbilical vein	2611657	ENCODE
ENCFF033LRZ	endothelial cell of umbilical vein	2562641	ENCODE
ENCFF694DIE	GM12878	2499375	ENCODE
ENCFF450VAU	GM12878	2115533	ENCODE
ENCFF417VHJ	GM12878	1575236	ENCODE
ENCFF688QGB	H9	3297669	ENCODE
ENCFF272VSN	H9	2837013	ENCODE
ENCFF337VWR	HCT116	1460741	ENCODE
ENCFF537NCV	heart left ventricle	2661413	ENCODE
ENCFF429JUP	heart left ventricle	1285467	ENCODE
ENCFF602MAI	heart left ventricle	1045691	ENCODE
ENCFF615FIC	heart right ventricle	1685360	ENCODE
ENCFF793PGJ	heart right ventricle	1345058	ENCODE
ENCFF425VDL	heart right ventricle	1106468	ENCODE
ENCFF483HTA	HepG2	2612690	ENCODE

ENCFF609QIM	HL-60	2878033	ENCODE
ENCFF274DYS	HL-60	2432288	ENCODE
ENCFF666JCQ	HL-60	2115302	ENCODE
ENCFF417RYS	HL-60	2042269	ENCODE
ENCFF782UMU	HL-60	1950685	ENCODE
ENCFF810XST	HL-60	1907878	ENCODE
ENCFF564TOK	HL-60	1890595	ENCODE
ENCFF032UMC	HL-60	1886923	ENCODE
ENCFF805YXK	HL-60	1868899	ENCODE
ENCFF457TIY	HL-60	1633482	ENCODE
ENCFF321PMP	HL-60	1445126	ENCODE
ENCFF407SUN	HL-60	1331716	ENCODE
ENCFF199OTG	HL-60	1231903	ENCODE
ENCFF173QRD	HL-60	1141416	ENCODE
ENCFF145QNC	HL-60	1090893	ENCODE
ENCFF260AJN	HL-60	894750	ENCODE
ENCFF197DCI	IMR-90	1234075	ENCODE
ENCFF634YSN	K562	2800443	ENCODE
ENCFF429VVB	K562	2698366	ENCODE
ENCFF696GDL	K562	2020521	ENCODE
ENCFF492BYP	kidney	901716	ENCODE
ENCFF920VXE	left cardiac atrium	769976	ENCODE
ENCFF245MBY	left colon	1312593	ENCODE
ENCFF793CMQ	left ventricle myocardium inferior	1430840	ENCODE
ENCFF624IQY	left ventricle myocardium superior	1469933	ENCODE
ENCFF552NVU	lower lobe of left lung	2602494	ENCODE
ENCFF341BSQ	lower lobe of left lung	777834	ENCODE
ENCFF250IWT	lower lobe of right lung	1900303	ENCODE
ENCFF237FMP	mammary epithelial cell	2635356	ENCODE
ENCFF617YVE	mammary epithelial cell	2517178	ENCODE
ENCFF702KLU	MCF 10A	1989087	ENCODE
ENCFF041EGI	MCF 10A	1953578	ENCODE
ENCFF887DGG	MCF-7	1742102	ENCODE
ENCFF907SZK	mesenteric fat pad	1898739	ENCODE
ENCFF156TTD	middle frontal area 46	2756340	ENCODE
ENCFF311CZO	middle frontal area 46	2719170	ENCODE
ENCFF446EFU	middle frontal area 46	2533014	ENCODE
ENCFF708BOP	middle frontal area 46	2505116	ENCODE
ENCFF785KVJ	middle frontal area 46	2197031	ENCODE

ENCFF838DFB	middle frontal area 46	2112299	ENCODE
ENCFF827DUW	middle frontal area 46	1641325	ENCODE
ENCFF206TQZ	middle frontal area 46	1002212	ENCODE
ENCFF260AWP	middle frontal area 46	521182	ENCODE
ENCFF387HPO	mucosa of descending colon	1355283	ENCODE
ENCFF249GFH	neural crest cell	1199310	ENCODE
ENCFF026VEI	neural crest cell	916165	ENCODE
ENCFF417UQV	OCI-LY7	1805339	ENCODE
ENCFF511KJB	OCI-LY7	1475971	ENCODE
ENCFF556DYU	osteocyte	2115770	ENCODE
ENCFF560XTG	osteocyte	1683994	ENCODE
ENCFF756AHG	ovary	2155346	ENCODE
ENCFF422XLS	ovary	1359776	ENCODE
ENCFF990CUL	Panc1	1687025	ENCODE
ENCFF834KTE	PC-3	2062428	ENCODE
ENCFF107YRM	PC-9	1810603	ENCODE
ENCFF860AWQ	PC-9	1720299	ENCODE
ENCFF658OZB	posterior vena cava	1350298	ENCODE
ENCFF960KBO	posterior vena cava	621106	ENCODE
ENCFF471YEK	progenitor cell of endocrine pancreas	1698651	ENCODE
ENCFF988RQM	progenitor cell of endocrine pancreas	1225562	ENCODE
ENCFF750LYC	psoas muscle	1839037	ENCODE
ENCFF630XEC	psoas muscle	744415	ENCODE
ENCFF899MTI	right cardiac atrium	2489853	ENCODE
ENCFF905RVF	right cardiac atrium	1815720	ENCODE
ENCFF722JJS	right cardiac atrium	793250	ENCODE
ENCFF242WRZ	right cardiac atrium	767088	ENCODE
ENCFF738UZJ	Right ventricle myocardium inferior	683430	ENCODE
ENCFF665LBS	Right ventricle myocardium superior	1648621	ENCODE
ENCFF743MYM	technical sample	2409856	ENCODE
ENCFF372YUA	technical sample	1731829	ENCODE
ENCFF525JUC	technical sample	1226796	ENCODE
ENCFF580BQX	type B pancreatic cell	1680369	ENCODE
ENCFF489XQJ	type B pancreatic cell	1647160	ENCODE
ENCFF934MBW	upper lobe of right lung	1716474	ENCODE
ENCFF245IPA	WTC11	3007792	ENCODE
ENCFF563QZR	WTC11	2563543	ENCODE
ENCFF370NFS	WTC11	1632302	ENCODE

cmh001658-05_IsoSeqv2-Cell1_CCS	blood	2063350	ANVIL phs002206
cmh001658-06_IsoSeqv2-Cell1_CCS	blood	1871523	ANVIL phs002206
cmh001658-04_IsoSeqv2-Cell1_CCS	blood	1796512	ANVIL phs002206
cmh001658-01_IsoSeqv2-Cell1_CCS	blood	1295479	ANVIL phs002206
cmh001807-01_IsoSeqv2_iPSC-Cell1_CCS	blood	834019	ANVIL phs002206
cmh003036-01_IsoSeqv2-IS-brain_CCS	brain	4345676	ANVIL phs002206
cmh003217-01_IsoSeqv2-IS-brain_CCS	brain	3686653	ANVIL phs002206
PBIsoSeq_cmh002013-01_brain_combined_CCS	brain	3517268	ANVIL phs002206
cmh001768-01_IsoSeqv2-IS-brain_CCS	brain	3081084	ANVIL phs002206
cmh003306-01_IsoSeqv2-Cell3_CCS	cartilage	3196670	ANVIL phs002206
cmh003629-01_IsoSeqv2-Cell3_CCS	cartilage	2724198	ANVIL phs002206
cmh001950-01_IsoSeqv2-Cell2_CCS	cartilage	2133723	ANVIL phs002206
cmh003465-01_IsoSeqv2-Cell4_CCS	cartilage	2040261	ANVIL phs002206
cmh003514-01_IsoSeqv2-Cell1_CCS	cartilage	1766213	ANVIL phs002206
cmh001818-01_IsoSeqv2-Cell3_CCS	granuloma	2368783	ANVIL phs002206
cmh001807-01_IsoSeqv2_iPSC-Neur-Cell1_CCS	ipsc	4827811	ANVIL phs002206
cmh002429-01_IsoSeqv2_iPSC-Cell2_CCS	ipsc	3733654	ANVIL phs002206
cmh001748-04_IsoSeqv2-Cell1_CCS	ipsc	3711640	ANVIL phs002206
cmh002697-01_IsoSeqv2_iPSC-Cell4_CCS	ipsc	3702185	ANVIL phs002206
cmh001866-01_IsoSeqv2-Cell1_CCS	ipsc	3676074	ANVIL phs002206
cmh002248-01_IsoSeqv2_iPSC-Cell5_CCS	ipsc	3641579	ANVIL phs002206
cmh001743-01_IsoSeqv2-Cell1_CCS	ipsc	3525088	ANVIL phs002206
cmh002189-01_IsoSeqv2_iPSC-Cell3_CCS	ipsc	3500298	ANVIL phs002206
cmh002275-01_IsoSeqv2_iPSC-Cell1_CCS	ipsc	3425947	ANVIL phs002206
cmh001761-01_IsoSeqv2_iPSC-Cell5_CCS	ipsc	3417624	ANVIL phs002206
cmh002531-01_IsoSeqv2-Cell1_CCS	ipsc	3370620	ANVIL phs002206
cmh001748-01_IsoSeqv2_iPSC-Cell1_CCS	ipsc	3346776	ANVIL phs002206
cmh002004-01_IsoSeqv2_iPSC-Cell5_CCS	ipsc	3334782	ANVIL phs002206
cmh002208-01_IsoSeqv2_iPSC-Cell2_CCS	ipsc	3330144	ANVIL phs002206
cmh002557-01_IsoSeqv2-Cell1_CCS	ipsc	3325405	ANVIL phs002206
cmh002114-01_IsoSeqv2_iPSC-Cell2_CCS	ipsc	3309362	ANVIL phs002206
cmh002971-01_IsoSeqv2_iPSC-Cell2_CCS	ipsc	3279501	ANVIL phs002206
cmh001648-01_IsoSeqv2_iPSC-Cell4_CCS	ipsc	3245072	ANVIL phs002206
cmh001807-01_IsoSeqv2-Cell1_CCS	ipsc	3202327	ANVIL phs002206

cmh002381-01_IsoSeqv2_iPSC-Cell1_CCS	ipsc	3197844	ANVIL phs002206
cmh002650-01_IsoSeqv2_iPSC-Cell4_CCS	ipsc	3184033	ANVIL phs002206
cmh001991-01_IsoSeqv2_iPSC-Cell4_CCS	ipsc	3173897	ANVIL phs002206
cmh001370_IsoSeqv2_iPSC-Cell3_CCS	ipsc	3151961	ANVIL phs002206
cmh002478-01_IsoSeqv2_iPSC-Cell3_CCS	ipsc	3150460	ANVIL phs002206
cmh002362-01_IsoSeqv2-Cell1_CCS	ipsc	3110739	ANVIL phs002206
cmh002048-01_IsoSeqv2-Cell1_CCS	ipsc	3092308	ANVIL phs002206
cmh002355-01_IsoSeqv2_iPSC-Cell3_CCS	ipsc	3045536	ANVIL phs002206
cmh002222-01_IsoSeqv2_iPSC-Cell4_CCS	ipsc	3039743	ANVIL phs002206
cmh001935-01_IsoSeqv2_iPSC-Cell1_CCS	ipsc	3030798	ANVIL phs002206
cmh002178-01_IsoSeqv2_iPSC-Cell1_CCS	ipsc	3021929	ANVIL phs002206
cmh002770-01_IsoSeqv2_iPSC-Cell1_CCS	ipsc	3020021	ANVIL phs002206
cmh002126-01_IsoSeqv2_iPSC-Cell1_CCS	ipsc	2987085	ANVIL phs002206
cmh002743-01_IsoSeqv2_iPSC-Cell5_CCS	ipsc	2966435	ANVIL phs002206
cmh002832-01_IsoSeqv2_iPSC-Cell4_CCS	ipsc	2948572	ANVIL phs002206
cmh003006-01_IsoSeqv2_iPSC-Cell3_CCS	ipsc	2942142	ANVIL phs002206
cmh002039-04_IsoSeqv2_iPSC-Cell1_CCS	ipsc	2915315	ANVIL phs002206
cmh001256_IsoSeqv2_iPSC-Cell2_CCS	ipsc	2909351	ANVIL phs002206
cmh002207-01_IsoSeqv2_iPSC-Cell1_CCS	ipsc	2880112	ANVIL phs002206
cmh002985-01_IsoSeqv2_iPSC-Cell1_CCS	ipsc	2845020	ANVIL phs002206
cmh002059-01_IsoSeqv2_iPSC-Cell1_CCS	ipsc	2841911	ANVIL phs002206
cmh002657-01_IsoSeqv2_iPSC-Cell2_CCS	ipsc	2837510	ANVIL phs002206
cmh002618-01_IsoSeqv2_iPSC-Cell1_CCS	ipsc	2821230	ANVIL phs002206
cmh001961-04_IsoSeqv2_iPSC-Cell1_CCS	ipsc	2818872	ANVIL phs002206
cmh002258-01_IsoSeqv2_iPSC-Cell5_CCS	ipsc	2810743	ANVIL phs002206
cmh002255-04_IsoSeqv2_iPSC-Cell2_CCS	ipsc	2805822	ANVIL phs002206
cmh001573-01_IsoSeqv2_iPSC-Cell3_CCS	ipsc	2777639	ANVIL phs002206
cmh002039-01_IsoSeqv2_iPSC-Cell1_CCS	ipsc	2773005	ANVIL phs002206
cmh002124-04_IsoSeqv2_iPSC-Cell1_CCS	ipsc	2770350	ANVIL phs002206
cmh001796-01_IsoSeqv2-Cell1_CCS	ipsc	2766009	ANVIL phs002206
cmh002066-01_IsoSeqv2_iPSC-Cell1_CCS	ipsc	2744386	ANVIL phs002206
cmh001982-01_IsoSeqv2_iPSC-Cell4_CCS	ipsc	2723160	ANVIL phs002206
cmh002651-01_IsoSeqv2_iPSC-Cell5_CCS	ipsc	2709700	ANVIL phs002206
cmh001996-01_IsoSeqv2_iPSC-Cell1_CCS	ipsc	2689569	ANVIL phs002206

cmh002426-01_IsoSeqv2_iPSC-Cell5_CCS	ipsc	2687080	ANVIL phs002206
cmh001866-04_IsoSeqv2-Cell1_CCS	ipsc	2672972	ANVIL phs002206
cmh002692-01_IsoSeqv2_iPSC-Cell3_CCS	ipsc	2655894	ANVIL phs002206
cmh002249-01_IsoSeqv2_iPSC-Cell1_CCS	ipsc	2631399	ANVIL phs002206
cmh002145-01_IsoSeqv2_iPSC-Cell3_CCS	ipsc	2619527	ANVIL phs002206
cmh002350-01_IsoSeqv2_iPSC-Cell1_CCS	ipsc	2602238	ANVIL phs002206
cmh003042-01_IsoSeqv2_iPSC-Cell2_CCS	ipsc	2572752	ANVIL phs002206
cmh002853-01_IsoSeqv2_iPSC-Cell5_CCS	ipsc	2571919	ANVIL phs002206
cmh001977-01_IsoSeqv2-Cell1_CCS	ipsc	2555892	ANVIL phs002206
cmh002656-01_IsoSeqv2_iPSC-Cell1_CCS	ipsc	2549544	ANVIL phs002206
cmh002183-01_IsoSeqv2_iPSC-Cell2_CCS	ipsc	2546320	ANVIL phs002206
cmh002319-01_IsoSeqv2_iPSC-Cell4_CCS	ipsc	2456991	ANVIL phs002206
cmh002550-04_IsoSeqv2_iPSC-Cell2_CCS	ipsc	2450668	ANVIL phs002206
cmh002821-01_IsoSeqv2_iPSC-Cell2_CCS	ipsc	2430615	ANVIL phs002206
cmh002208-01_IsoSeqv2_iPSC-EB-Cell3_CCS	ipsc	2413714	ANVIL phs002206
cmh002444-01_IsoSeqv2_iPSC-Cell1_CCS	ipsc	2413102	ANVIL phs002206
cmh002326-01_IsoSeqv2_iPSC-Cell2_CCS	ipsc	2387973	ANVIL phs002206
cmh001992-01_IsoSeqv2_iPSC-Neur-Cell2_CCS	ipsc	2374637	ANVIL phs002206
cmh002243-01_IsoSeqv2_iPSC-Cell4_CCS	ipsc	2373881	ANVIL phs002206
cmh002197-01_IsoSeqv2_iPSC-Cell3_CCS	ipsc	2353298	ANVIL phs002206
cmh002668-01_IsoSeqv2_iPSC-Cell2_CCS	ipsc	2344094	ANVIL phs002206
cmh002160-01_IsoSeqv2_iPSC-Cell4_CCS	ipsc	2303484	ANVIL phs002206
cmh002548-01_IsoSeqv2_iPSC-Cell4_CCS	ipsc	2301457	ANVIL phs002206
cmh002871-01_IsoSeqv2_iPSC-Cell1_CCS	ipsc	2270248	ANVIL phs002206
cmh002397-01_IsoSeqv2-Cell1_CCS	ipsc	2251955	ANVIL phs002206
cmh002006-01_IsoSeqv2_iPSC-Cell1_CCS	ipsc	2246668	ANVIL phs002206
cmh001749-01_IsoSeqv2_iPSC-Cell2_CCS	ipsc	2237111	ANVIL phs002206
cmh002264-01_IsoSeqv2_iPSC-Cell3_CCS	ipsc	2222856	ANVIL phs002206
cmh002459-01_IsoSeqv2_iPSC-Cell2_CCS	ipsc	2197797	ANVIL phs002206
cmh002268-01_IsoSeqv2_iPSC-Cell2_CCS	ipsc	2141361	ANVIL phs002206
cmh001581-01_IsoSeqv2_iPSC-Cell4_CCS	ipsc	2126667	ANVIL phs002206
cmh002017-01_IsoSeqv2_iPSC-Cell5_CCS	ipsc	2112894	ANVIL phs002206
cmh001760-01_IsoSeqv2_iPSC-Cell1_CCS	ipsc	2109227	ANVIL phs002206
cmh002497-01_IsoSeqv2_iPSC-Cell1_CCS	ipsc	2102024	ANVIL phs002206

cmh002495-01_IsoSeqv2_iPSC-Cell3_CCS	ipsc	2080871	ANVIL phs002206
cmh002173-01_IsoSeqv2_iPSC-Cell5_CCS	ipsc	2056376	ANVIL phs002206
cmh002105-01_IsoSeqv2_iPSC-Cell2_CCS	ipsc	1994116	ANVIL phs002206
cmh001968-01_IsoSeqv2_iPSC-Cell2_CCS	ipsc	1915058	ANVIL phs002206
cmh001971-01_IsoSeqv2_iPSC-Cell3_CCS	ipsc	1908723	ANVIL phs002206
cmh002430-01_IsoSeqv2_iPSC-Cell1_CCS	ipsc	1889535	ANVIL phs002206
cmh002818-01_IsoSeqv2_iPSC-Cell1_CCS	ipsc	1885651	ANVIL phs002206
cmh002193-01_IsoSeqv2-Cell1_CCS	ipsc	1660268	ANVIL phs002206
cmh000118_IsoSeqv2_iPSC-Cell1_CCS	ipsc	1460148	ANVIL phs002206
cmh001712-01_IsoSeqv2-Cell1_CCS	ipsc	1446090	ANVIL phs002206
cmh001805-01_IsoSeqv2-Cell1_CCS	ipsc	1328503	ANVIL phs002206
cmh003187-01_IsoSeqv2-Cell1_CCS	soft tissue	3073217	ANVIL phs002206
cmh001715-06_IsoSeqv2-Cell2_CCS	thymus	2815124	ANVIL phs002206
cmh001637-01_IsoSeqv2-Cell4_CCS	tonsil	1940008	ANVIL phs002206
Re-run_of_cmh003068-01C_TSC_IsoSeq_GRCh38-tsc_CCS	tsc	7832608	ANVIL phs002206
pending	thymus	263952	SRA
https://downloads.pacbcloud.com/public/dataset/Kinnex-full-length-RNA/DATA-Revio-HG002-1/2-FLNC/flnc.bam	HG002	37317270	PacBio
https://downloads-ap.pacbcloud.com/public/dataset/Melanoma2019_IsoSeq/FullLengthReads/flnc.bam	Melanoma	1872695	PacBio
https://downloads.pacbcloud.com/public/dataset/UHR_IsoSeq/FullLengthReads/flnc.bam	UHR	6374832	PacBio
SRR12638398	Dorsal Root Ganglion	47089	SRA
SRR12638397	Esophagus	43898	SRA
SRR12638394	Fetal Brain	140151	SRA
SRR12638396	GM10539	49025	SRA
SRR12638395	Skin	14045	SRA
SRR12544672	testis	606371	SRA
SRR12544673	testis	550195	SRA
SRR12638393	testis	186323	SRA

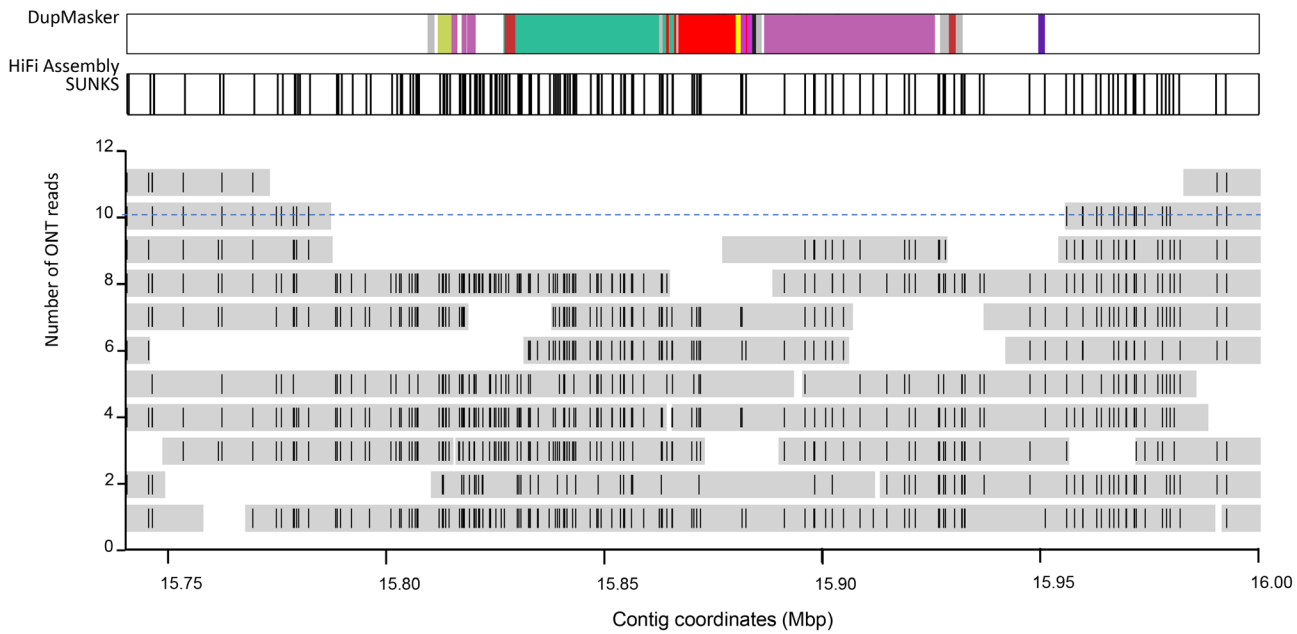
Table S7. Summary of misassemblies and collapse within human genomes.

Sample	Number of Collapse	Collapse bases	Number of Misassembly	Misassembly bases	Source
HG00621.h1	157	5605234	118	1371843	HPRC
HG00621.h2	112	2745195	104	1115476	HPRC
HG00673.h1	247	8797953	115	1245090	HPRC
HG00673.h2	138	3785684	102	1076419	HPRC
HG00735.h1	154	4867526	88	932565	HPRC
HG00735.h2	145	5624387	79	963306	HPRC
HG00741.h1	115	3821209	109	1185968	HPRC
HG00741.h2	129	4385979	110	1253159	HPRC
HG01071.h1	97	3950099	123	2282226	HPRC
HG01071.h2	82	1829159	130	1477330	HPRC
HG01106.h1	182	7349444	82	883891	HPRC
HG01106.h2	119	5135040	62	658943	HPRC
HG01123.h1	145	5212257	130	1520806	HPRC
HG01123.h2	145	4284102	107	1183128	HPRC
HG01175.h1	172	7664884	111	1233676	HPRC
HG01175.h2	107	2967816	107	1097643	HPRC
HG01258.h1	151	4404171	98	1018030	HPRC
HG01258.h2	184	4393731	131	1428052	HPRC
HG01352.h1	151	4752123	109	1141919	HGSVC
HG01352.h2	130	2634847	137	1446214	HGSVC
HG01358.h1	179	5642674	108	1210142	HPRC
HG01358.h2	166	4860239	91	1031615	HPRC
HG01361.h1	98	2479100	88	948745	HPRC
HG01361.h2	200	5838621	104	1087125	HPRC
HG01457.h1	100	1900310	70	716485	HGSVC
HG01457.h2	99	1808642	85	894389	HGSVC
HG01891.h1	124	3799048	65	711279	HPRC
HG01891.h2	143	3879915	94	1007943	HPRC
HG01928.h1	132	5542791	119	1277546	HPRC
HG01928.h2	131	4305551	128	1404510	HPRC
HG01952.h1	143	3911438	92	1170950	HPRC
HG01952.h2	82	1441238	101	1070289	HPRC
HG01978.h1	187	7727963	119	1251765	HPRC
HG01978.h2	117	3884548	113	1326865	HPRC
HG02018.h1	22	376489	176	1864079	HGSVC
HG02018.h2	9	140860	149	1514794	HGSVC
HG02055.h1	191	7046769	62	686916	HPRC

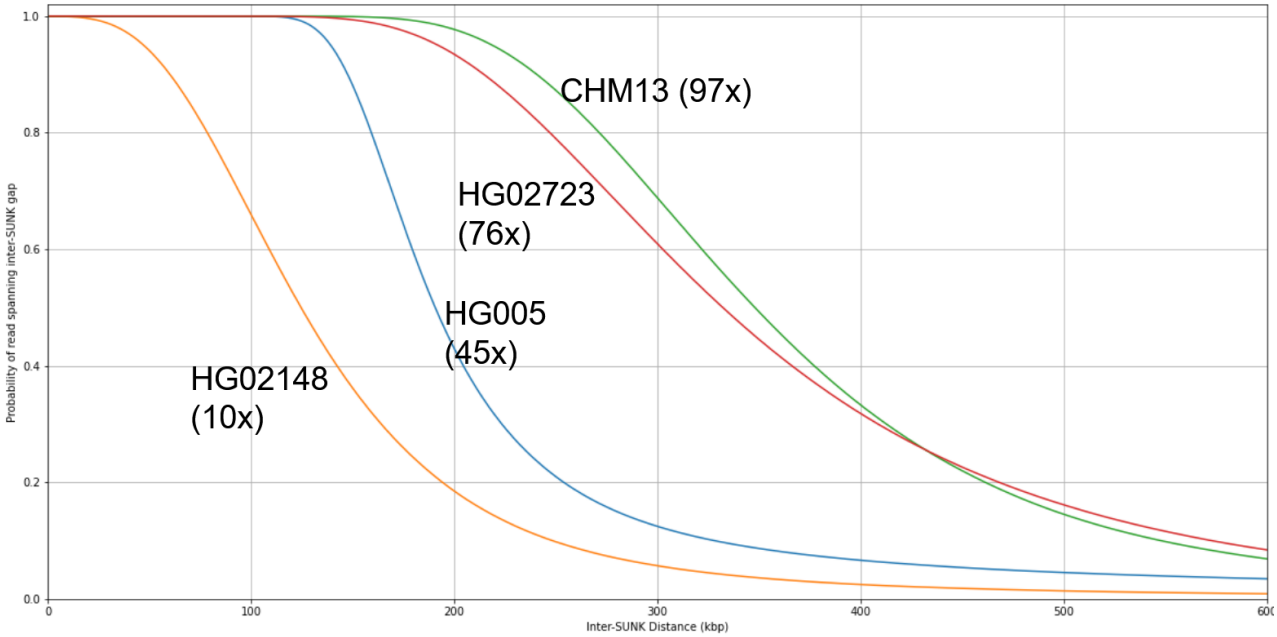
HG02055.h2	162	4995686	56	564707	HPRC
HG02059.h1	100	2301302	124	1321988	HGSVC
HG02059.h2	112	2647514	147	1538074	HGSVC
HG02080.h1	138	4132506	140	1478700	HPRC
HG02080.h2	152	5723599	129	1348465	HPRC
HG02106.h1	132	2917084	153	1567375	HGSVC
HG02106.h2	179	3636185	156	1610777	HGSVC
HG02145.h1	158	3973495	84	854903	HPRC
HG02145.h2	203	5523563	83	858512	HPRC
HG02148.h1	126	3944647	126	1302816	HPRC
HG02148.h2	168	6081572	113	1181408	HPRC
HG02257.h1	180	4781661	79	901278	HPRC
HG02257.h2	157	5643968	63	704532	HPRC
HG02486.h1	167	5234634	70	736841	HPRC
HG02486.h2	120	4058305	63	651177	HPRC
HG02572.h1	235	7178423	162	1805362	HPRC
HG02572.h2	159	4565611	179	1999804	HPRC
HG02622.h1	137	4637847	53	599752	HPRC
HG02622.h2	150	4173323	52	615966	HPRC
HG02630.h1	196	4498053	49	550310	HPRC
HG02630.h2	157	3718532	63	689737	HPRC
HG02717.h1	186	6403195	56	599612	HPRC
HG02717.h2	134	4046748	71	827790	HPRC
HG02723.h1	268	7052963	69	801801	HPRC
HG02723.h2	109	2779830	60	626456	HPRC
HG02818.h1	146	4221203	98	1021792	HPRC
HG02818.h2	153	4540616	103	1080590	HPRC
HG02886.h1	172	6389480	70	787218	HPRC
HG02886.h2	143	4575393	86	957621	HPRC
HG03098.h1	238	7284204	74	858256	HPRC
HG03098.h2	247	7336232	71	787150	HPRC
HG03125.h1	216	5388159	218	2310926	HGSVC
HG03125.h2	187	4717868	203	2146690	HGSVC
HG03248.h1	144	3930157	67	684829	HGSVC
HG03248.h2	101	2534188	54	566957	HGSVC
HG03453.h1	230	5560771	63	757887	HPRC
HG03453.h2	195	5552022	49	514539	HPRC
HG03456.h1	101	2518888	89	963840	HGSVC
HG03456.h2	106	2175516	92	998032	HGSVC

HG03486.h1	215	6218095	78	844184	HPRC
HG03486.h2	193	5019192	64	673117	HPRC
HG03516.h1	127	3333349	87	949821	HPRC
HG03516.h2	150	4190932	90	1002622	HPRC
HG03540.h1	235	5784460	63	777026	HPRC
HG03540.h2	187	6884094	47	496206	HPRC
HG03579.h1	173	5382736	54	591932	HPRC
HG03579.h2	179	4182099	44	478107	HPRC
HG03807.h1	107	1924666	125	1288965	HGSVC
HG03807.h2	104	2171744	138	1455232	HGSVC
HG04036.h1	195	3641160	130	1377967	HGSVC
HG04036.h2	211	4420741	157	1610234	HGSVC
HG04217.h1	124	2818591	121	1239080	HGSVC
HG04217.h2	126	2453813	143	1447432	HGSVC
NA12878.h1	303	7953657	396	4335609	HGSVC
NA12878.h2	313	7938108	440	4696219	HGSVC
NA18906.h1	189	5412124	64	756863	HPRC
NA18906.h2	229	5754690	56	576015	HPRC
NA19705.h1	110	2705689	79	817564	HGSVC
NA19705.h2	72	1700324	61	667387	HGSVC
NA20129.h1	104	2437506	97	1022139	HPRC
NA20129.h2	115	2914825	85	929688	HPRC

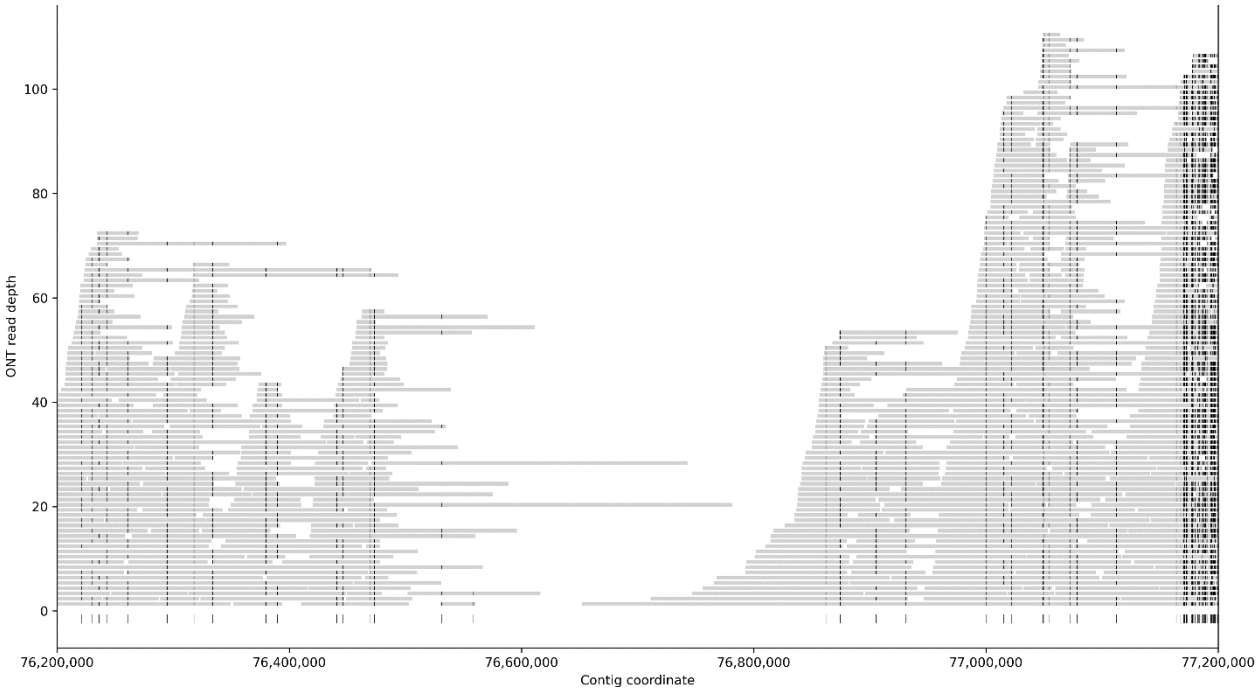
APPENDIX B. SUPPLEMENT FOR CHAPTER 3



Supplementary Figure 1. An example of a validated region within the maternal haplotype assembly of HG00733. Horizontal gray bars represent individual ONT reads, and vertical black lines indicate the position of SUNKs originally detected from the HiFi assembly. All possible assembly SUNKs are shown above, along with a DupMasker track marking segmental duplications. The dotted blue line indicates the mean genome-wide coverage of ONT sequencing data.

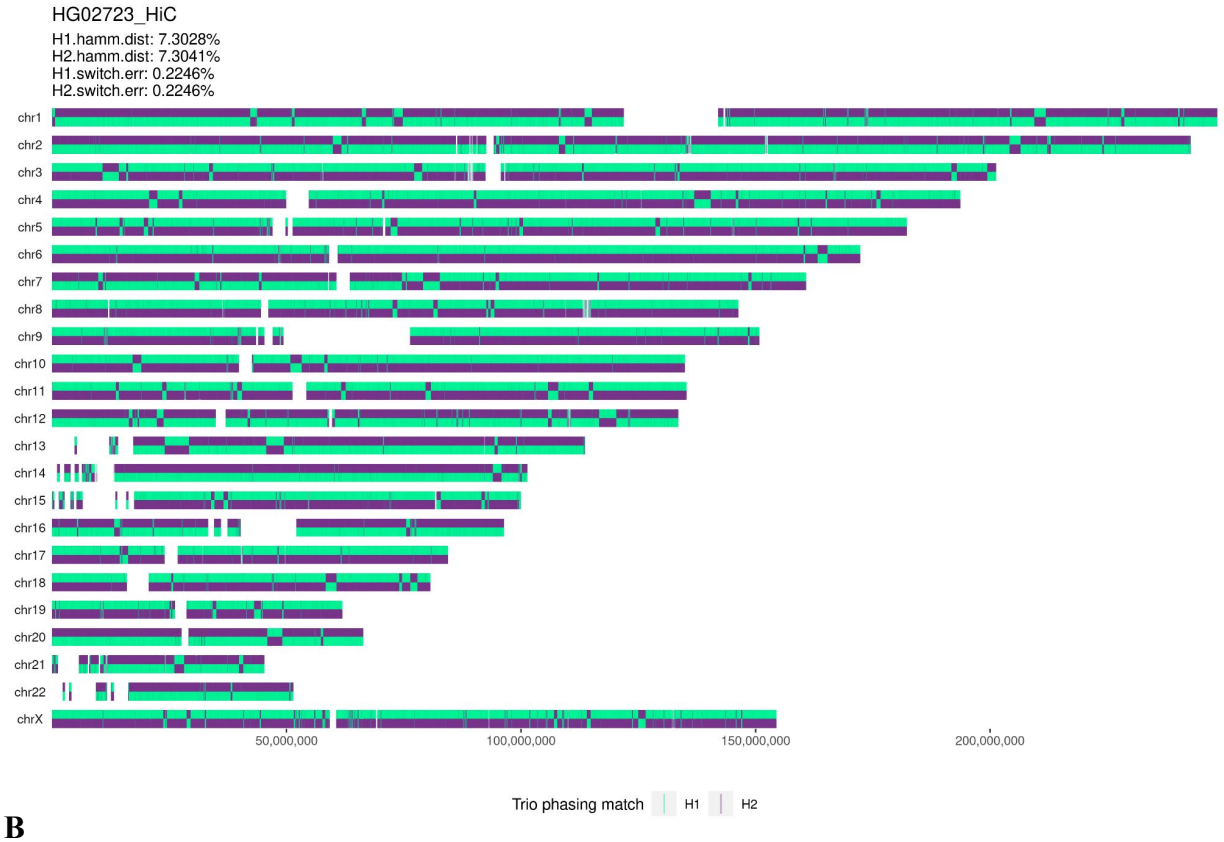


Supplementary Figure 2. Simulation of probability of spanning inter-SUNK gaps. The probability of reads spanning a given inter-SUNK distance is determined on a per-sample basis by calculating the Poisson distribution of ONT reads across the genome, adjusting for sequencing accuracy and size of the SUNK group. Haplotype-specific ONT sequencing depth is shown in parentheses for each sample.

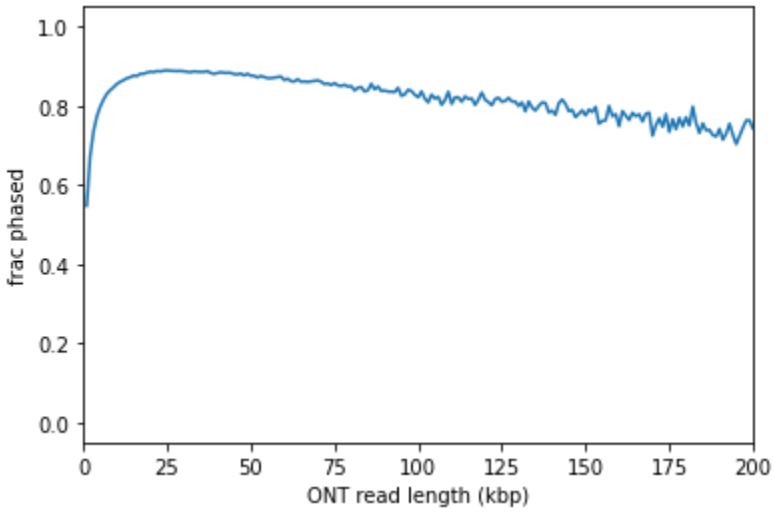


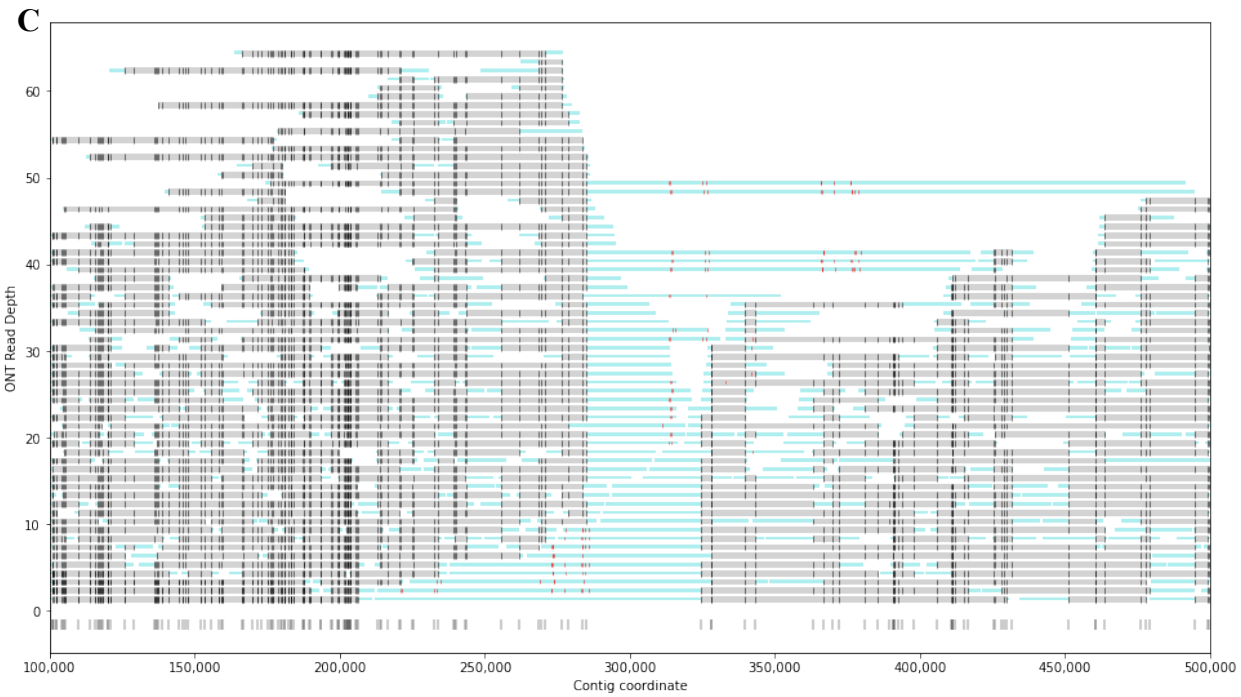
Supplementary Figure 3. Unspannable gap in T2T-CHM13 *HYDIN*. Gray bars with black tickmarks represent ONT reads and their respective SUNKs, while black tickmarks below represent all assembly SUNKs. Because the SUNK density at this recently duplicated locus is too low for ONT reads at the given length and coverage to span between SUNK groups, no determination can be made of assembly correctness.

A


HG02723_HiC
H1.hamm.dist: 7.3028%
H2.hamm.dist: 7.3041%
H1.switch.err: 0.2246%
H2.switch.err: 0.2246%
chr1
chr2
chr3
chr4
chr5
chr6
chr7
chr8
chr9
chr10
chr11
chr12
chr13
chr14
chr15
chr16
chr17
chr18
chr19
chr20
chr21
chr22
chrX
50,000,000
100,000,000
150,000,000
200,000,000
Trio phasing match
H1
H2

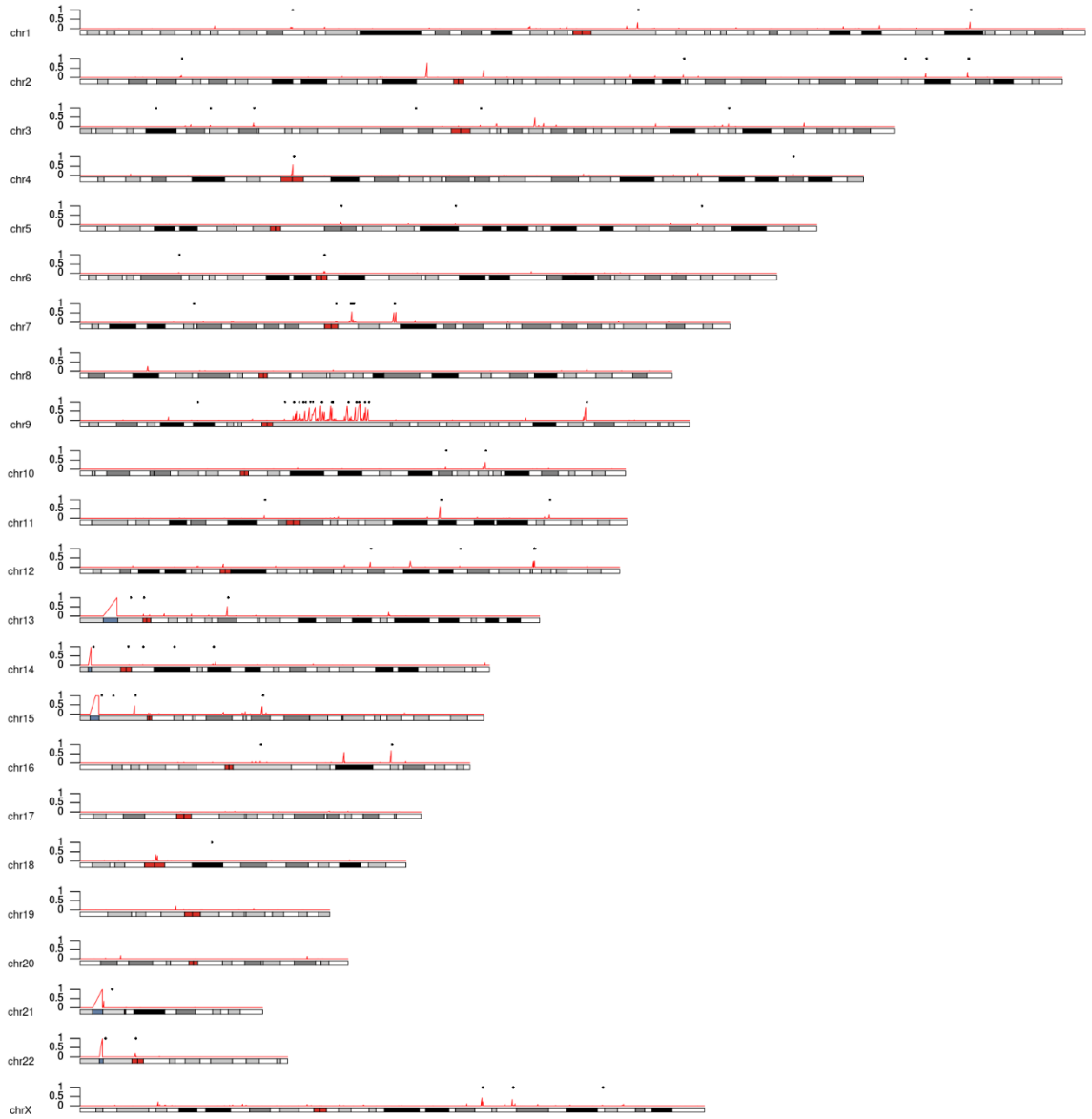
B


1.0
0.8
0.6
0.4
0.2
0.0
frac phased
0
25
50
75
100
125
150
175
200
ONT read length (kbp)



Supplementary Figure 4. Human genome sample analysis without parental phasing data. **A)** HiC-based phasing of HG02723 evaluated with parental data post hoc. **B)** Phasing of ONT data as a function of ONT read length from the same sample using custom script, HiCphaseONT. Phased ONT data can now be used for GAVISUNK. **C)** AMY test region phased with Hi-C instead of parental Illumina data.

Supplementary Figure 5. T2T-CHM13 empirical validation gaps. Each dot is an empirical validation gap, while the red line indicates the estimated probability of failing to span an inter-SUNK distance. Validation gaps with high simulated probabilities of coverage fall in the qh regions of chromosomes 1, 9, and 16, rDNA arrays, and centromeres.



Supplementary Figure 6. T2T-CHM13 simulated validation gaps. Each dot is a simulated validation gap, while the red line indicates the estimated probability of failing to span an inter-SUNK distance.

Table S1. Results of genome-wide analysis.

Sample		Inter-SUNK validation gaps	(Mbp)	Non-cenSat[a], non-rDNA gaps	(Mbp)	SegDup[b] gaps	(Mbp)	Assembly size (Gbp)	ONT Coverage >100kbp (x)
HG02723	hap1	343	51.3	59	13.2	15	1.1	3.049	26.1
	hap2	200	31.6	113	16.3	0	0	3.027	26.5
Pseudodiploid	CHM13	103	31.7	4	1.2	3	0.9	3.055	38.9
	CHM1	274	23.9	59	7.6	15	1.3	3.03	39.9

[a]Centromeric satellite [b]Segmental duplications, excluding centromeric satellite and rDNA repeats

Table S2. T2T-CHM13 validation gaps.

chr	Start	End	type	spanned in 2pass	centromeric	HOR	qh	rdna	sd	size	inter-SUNK distance	coverage_prob
chr13	5,770,469	9,174,328	gap		X			X	X	3,403,859	3,403,860	0.0%
chr21	3,118,235	5,570,653	gap		X			X	X	2,452,418	2,415,189	0.1%
chr15	2,455,480	4,666,065	gap		X			X	X	2,210,585	1,485,732	0.2%
chr22	4,792,736	5,547,731	gap		X			X	X	754,995	754,996	0.7%
chr14	2,100,640	2,741,434	gap		X			X	X	640,794	640,795	1.3%
chr9	67,720,009	69,193,786	gap		X		X			1,473,777	464,280	4.8%
chr2	85,401,198	85,758,663	gap							357,465	357,466	13.1%
chr9	59,019,771	59,473,805	gap		X		X			454,034	342,260	15.2%
chr9	61,211,373	62,255,017	gap		X		X			1,043,644	341,721	15.3%
chr9	65,681,754	66,189,625	gap		X		X			507,871	337,579	15.9%
chr9	70,290,665	71,140,231	gap		X		X			849,566	308,820	21.1%
chr9	57,365,965	58,084,879	gap		X					718,914	303,052	22.2%
chr16	76,558,669	76,862,032	gap				X		X	303,363	303,364	22.2%
chr9	56,199,250	56,701,844	gap		X		X			502,594	302,624	22.4%
chr11	88,686,368	88,976,790	gap						X	290,422	290,423	25.1%
chr7	66,882,979	67,150,397	gap	X					X	267,418	267,419	31.3%
chr9	55,308,345	55,607,226	gap		X		X			298,881	246,053	38.0%
chr9	59,486,868	59,938,062	gap		X		X			451,194	237,929	41.2%
chr15	13,270,091	13,505,792	gap		X					235,701	235,702	42.0%
chr9	60,157,254	60,986,479	gap		X		X			829,225	231,332	43.5%
chr9	70,008,102	70,263,639	gap		X		X			255,537	220,214	47.9%
chr9	53,018,210	53,231,168	gap		X		X			212,958	211,916	51.7%
chr9	53,683,734	54,320,032	gap		X		X			636,298	209,811	52.6%
chr18	18,574,097	18,783,509	gap	X	X	X				209,412	209,413	52.6%
chr1	137,325,263	137,800,970	gap		X		X			475,707	205,102	54.4%
chr18	18,920,906	19,149,218	gap	X	X	X				228,312	195,979	58.9%

chr9	52,529,084	53,005,197	gap		X		X			476,113	182,119	65.0%
chr9	66,793,601	67,101,902	gap		X		X			308,301	177,040	67.4%
chr22	13,502,864	13,675,122	gap	X	X	X				172,258	172,259	69.8%
chr1	134,909,723	135,070,393	gap		X		X			160,670	160,671	75.4%
chr9	66,328,060	66,650,392	gap		X		X			322,332	158,744	76.4%
chr9	58,359,808	58,729,528	gap		X		X			369,720	152,758	79.1%
chr6	60,120,609	60,311,031	gap	X	X	X				190,422	149,710	80.4%
chr16	44,418,847	44,567,538	gap		X		X			148,691	148,692	80.8%
chr9	54,399,656	54,549,900	gap		X		X			150,244	147,092	81.3%
chr9	62,994,431	63,350,977	gap		X		X			356,546	144,285	82.6%
chr9	54,559,671	55,161,790	gap		X		X			602,119	142,837	83.4%
chr6	60,369,327	60,511,279	gap	X	X					141,952	141,953	83.8%
chr16	43,234,272	43,371,967	gap		X		X			137,695	137,696	85.4%
chr9	49,774,584	50,797,944	gap		X		X			1,023,360	136,020	85.8%
chr7	63,044,101	63,215,298	gap	X	X	X				171,197	130,520	88.1%
chr15	16,839,461	16,968,687	gap		X	X				129,226	129,227	88.4%
chr16	45,986,713	46,236,746	gap		X		X			250,033	126,609	89.5%
chr9	56,933,686	57,067,152	gap		X		X			133,466	121,447	91.1%
chr1	141,796,002	141,973,046	gap		X		X			177,044	116,348	92.6%
chr9	58,753,272	59,014,877	gap		X		X			261,605	105,686	95.4%
chr1	137,839,586	138,079,576	gap		X		X			239,990	104,457	95.6%
chr1	141,448,454	141,603,783	gap		X		X			155,329	104,785	95.6%
chr1	122,474,798	122,644,846	gap		X	X				170,048	99,474	96.5%
chr1	141,219,074	141,346,393	gap		X		X			127,319	98,109	96.7%
chr9	62,482,094	62,621,655	gap		X		X			139,561	98,072	96.7%
chr1	136,663,057	136,794,649	gap		X		X			131,592	91,952	97.8%
chr9	67,426,424	67,487,775	gap		X		X			61,351	91,368	97.8%
chr1	131,127,896	131,386,866	gap		X		X			258,970	86,462	98.4%

chr1	139,281,444	139,439,175	gap		X		X			157,731	86,101	98.4%
chr13	15,746,009	15,834,435	gap	X	X	X				88,426	84,086	98.6%
chr1	135,754,981	135,847,615	gap		X		X			92,634	83,638	98.7%
chr1	133,355,038	133,485,526	gap		X		X			130,488	80,868	98.9%
chr1	131,577,446	131,866,142	gap		X		X			288,696	78,191	99.1%
chr1	132,881,749	133,281,499	gap		X		X			399,750	75,521	99.3%
chr1	133,506,130	133,631,988	gap		X		X			125,858	72,035	99.4%
chr1	130,964,392	131,080,644	gap		X		X			116,252	71,969	99.5%
chr1	131,395,053	131,466,274	gap		X		X			71,221	71,222	99.5%
chr1	134,236,265	134,344,667	gap		X		X			108,402	70,599	99.5%
chr1	135,916,321	136,328,230	gap		X		X			411,909	70,104	99.5%
chr16	42,060,596	42,167,997	gap		X		X			107,401	65,661	99.7%
chr1	134,027,783	134,119,790	gap		X		X			92,007	60,930	99.8%
chr13	4,184,794	4,252,422	gap		X		X			67,628	58,580	99.9%
chr16	42,676,201	42,770,543	gap		X		X			94,342	58,795	99.9%
chr16	42,959,937	43,031,535	gap		X					71,598	58,848	99.9%
chr10	40,218,299	40,275,807	gap	X	X	X				57,508	57,509	99.9%
chr16	42,597,205	42,650,752	gap		X		X			53,547	55,116	99.9%
chr15	6,802,532	6,904,099	gap	X	X					101,567	54,304	99.9%
chr15	12,764,445	12,888,384	gap	X	X					123,939	54,625	99.9%
chr1	132,007,608	132,058,713	gap		X		X			51,105	51,106	100.0%
chr1	129,881,032	129,944,544	gap		X		X			63,512	50,198	100.0%
chr1	132,518,198	132,666,764	gap		X		X			148,566	48,501	100.0%
chr1	137,104,207	137,216,545	gap		X		X			112,338	48,007	100.0%
chr1	141,980,178	142,161,194	gap		X		X			181,016	48,454	100.0%
chr1	140,365,923	140,458,629	gap		X		X			92,706	47,966	100.0%
chr1	133,825,422	133,941,734	gap		X		X			116,312	44,035	100.0%
chr1	131,476,062	131,531,441	gap		X		X			55,379	43,465	100.0%

chr1	138,774,925	138,828,835	gap		X		X			53,910	41,887	100.0%
chr1	130,152,695	130,193,004	gap		X		X			40,309	40,310	100.0%
chr1	140,657,839	140,883,893	gap		X		X			226,054	40,134	100.0%
chr1	138,437,521	138,477,201	gap		X		X			39,680	39,681	100.0%
chr1	141,674,911	141,761,578	gap		X		X			86,667	37,114	100.0%
chr1	130,009,098	130,043,786	gap		X		X			34,688	34,689	100.0%
chr1	135,289,311	135,449,393	gap		X		X			160,082	32,872	100.0%
chr3	94,067,124	94,097,836	gap	X	X	X				30,712	30,713	100.0%
chr15	7,903,291	7,931,879	gap	X	X					28,588	28,589	100.0%
chr1	140,243,201	140,316,094	gap		X		X			72,893	26,963	100.0%
chr1	140,938,639	140,973,918	gap		X		X			35,279	26,266	100.0%
chr9	50,892,754	50,933,630	gap		X		X			40,876	24,500	100.0%
chr15	12,429,187	12,451,662	gap	X	X					22,475	22,476	100.0%
chr16	52,081,511	52,123,047	gap		X		X			41,536	22,028	100.0%
chr16	51,346,344	51,361,163	gap		X		X			14,819	21,808	100.0%
chr13	175,479	225,995	gap		X		X			50,516	20,313	100.0%
chr16	44,876,894	44,915,693	gap	X	X					38,799	20,376	100.0%
chr15	7,187,424	7,204,739	gap	X	X					17,315	17,316	100.0%
chr13	319,158	345,581	gap		X					26,423	14,825	100.0%
chr13	240,979	259,584	gap	X	X					18,605	13,038	100.0%
chr1	130,088,031	130,100,191	gap		X		X			12,160	12,161	100.0%

Table S3. T2T-CHM13 simulated gaps.

Chr	Start	End	inter-sunk distance	coverage_prob	cen	rdna	sd
chr13	9,174,329	12,578,189	3,403,860	0.04%	X	X	X
chr21	5,533,424	7,948,613	2,415,189	0.09%	X	X	X
chr15	3,941,212	5,426,944	1,485,732	0.23%	X	X	X
chr22	5,547,732	6,302,728	754,996	0.73%	X	X	X
chr14	2,741,435	3,382,230	640,795	1.28%	X	X	X
chr9	59,473,806	59,816,066	342,260	15.18%	X		
chr9	61,938,522	62,280,243	341,721	15.33%	X		
chr9	66,019,333	66,356,912	337,579	15.94%	X		
chr9	68,029,402	68,338,795	309,393	20.91%	X		
chr9	124,916,859	125,222,793	305,934	21.73%			
chr16	76,862,033	77,165,397	303,364	22.16%			
chr9	56,613,182	56,915,806	302,624	22.37%	X		
chr9	62,255,018	62,547,682	292,664	24.64%	X		
chr11	88,976,791	89,267,214	290,423	25.12%			X
chr4	52,601,896	52,875,402	273,506	29.55%	X		
chr9	71,140,232	71,408,922	268,690	30.98%	X		
chr7	67,150,398	67,417,817	267,419	31.28%			
chr9	68,729,507	68,987,008	257,501	34.34%	X		
chr7	77,546,997	77,801,154	254,157	35.31%			
chr13	36,395,872	36,649,444	253,572	35.64%			
chr9	55,554,398	55,800,451	246,053	38.01%	X		
chr15	13,505,793	13,741,495	235,702	41.99%	X		
chrX	99,264,340	99,494,340	230,000	43.90%			
chr15	44,994,912	45,219,592	224,680	46.28%			
chr9	70,228,316	70,448,530	220,214	47.91%	X		
chr10	100,099,597	100,317,625	218,028	48.74%			
chr1	219,903,486	220,116,857	213,371	50.86%			
chrX	106,811,627	107,021,671	210,044	52.16%			
chr12	112,268,460	112,478,082	209,622	52.59%			
chr9	57,365,965	57,573,076	207,111	53.47%	X		
chr1	137,800,971	138,006,073	205,102	54.36%	X		
chr12	112,004,225	112,209,526	205,301	54.36%			
chr2	219,331,755	219,527,920	196,165	58.43%			
chr12	71,718,219	71,912,218	193,999	59.82%			
chr9	52,711,203	52,893,322	182,119	65.00%	X		
chr2	208,943,928	209,123,189	179,261	66.43%			
chr3	42,887,570	43,063,403	175,833	68.33%			
chr22	13,675,123	13,847,382	172,259	69.77%	X		

chr11	115,976,094	116,147,781	171,687	70.24%			
chr7	67,508,918	67,671,876	162,958	74.49%			
chr11	45,541,529	45,704,516	162,987	74.49%			
chr3	160,223,555	160,384,845	161,290	74.96%			
chr2	149,069,655	149,226,896	157,241	76.81%			
chr10	90,354,051	90,504,236	150,185	79.97%			
chr6	60,311,032	60,460,742	149,710	80.41%	X		
chr13	15,710,061	15,858,674	148,613	80.84%	X		
chr16	44,567,539	44,716,231	148,692	80.84%	X		
chr2	25,094,004	25,241,355	147,351	81.28%			
chr5	64,460,929	64,607,608	146,679	81.71%			
chr9	55,045,055	55,187,892	142,837	83.40%	X		
chr1	52,424,055	52,565,419	141,364	83.82%			
chr9	54,025,295	54,167,108	141,813	83.82%	X		
chr4	176,135,639	176,276,214	140,575	84.23%			
chr3	98,930,849	99,068,744	137,895	85.43%			
chr9	50,540,920	50,676,940	136,020	85.82%	X		
chr12	93,834,368	93,970,062	135,694	86.21%			
chr3	32,193,533	32,327,567	134,034	86.60%			
chr14	32,923,207	33,057,975	134,768	86.60%			
chr9	62,398,637	62,532,198	133,561	86.98%	X		
chr7	63,174,621	63,305,141	130,520	88.08%	X		
chr2	219,475,946	219,603,705	127,759	89.15%			
chr6	24,450,479	24,576,643	126,164	89.49%			
chr5	92,716,641	92,834,602	117,961	92.33%			
chr14	15,539,804	15,655,304	115,500	92.90%	X		
chr9	29,057,071	29,159,872	102,801	95.97%			
chr7	66,826,261	66,922,846	96,585	97.04%			X
chr18	32,527,663	32,612,685	85,022	98.48%			
chr14	23,295,951	23,373,959	78,008	99.08%			
chr3	82,972,837	83,047,932	75,095	99.28%			
chr5	153,592,857	153,668,355	75,498	99.28%			
chr3	18,751,218	18,825,422	74,204	99.33%			
chr15	8,200,146	8,273,693	73,547	99.39%	X		
chrX	129,118,246	129,191,382	73,136	99.39%			
chr2	203,860,977	203,933,382	72,405	99.44%			
chr7	28,094,692	28,167,041	72,349	99.44%			
chr14	11,883,612	11,949,729	66,117	99.68%	X		

Table S4. Sensitivity test.

Assembly error		Heterozygous errors						Homozgyous errors (inserted into both CHM1 and CHM13)						Overall	
		Error introduced into CHM13			Error introduced into CHM1			Error detection in CHM13			Error detection in CHM1				
Type	Size (kbp)	True positive rate (%)	False positive rate (%)	False negative rate (%)	True positive rate (%)	False positive rate (%)	False negative rate (%)	True positive rate (%)	False positive rate (%)	False negative rate (%)	True positive rate (%)	False positive rate (%)	False negative rate (%)	Sensitivity (%)	Precision (%)
Insertion	1	0	0	100	0	10	100	0	0	100	0	0	100	0	0
	10	100	0	0	30	0	70	90	0	10	30	0	70	62.5	100
	50	90	0	10	70	0	30	100	0	0	70	0	30	82.5	100
	100	100	0	0	80	0	20	100	0	0	80	0	20	90	100
Deletion	1	0	0	100	0	0	100	0	0	100	0	0	100	0	--
	10	0	0	100	0	0	100	10	0	90	90	0	10	25	100
	50	60	0	40	100	0	0	100	0	0	100	0	0	90	100
	100	80	0	20	100	0	0	70	0	30	100	0	0	87.5	100

All rates were quantified from 10 independent tests.

The true positive rate is defined as the detection rate of an introduced error.

The false positive rate is defined as the detection rate of a nonexistent error.

The false negative rate is defined as the detection failure rate of an introduced error.

APPENDIX C. SUPPLEMENT FOR CHAPTER 4

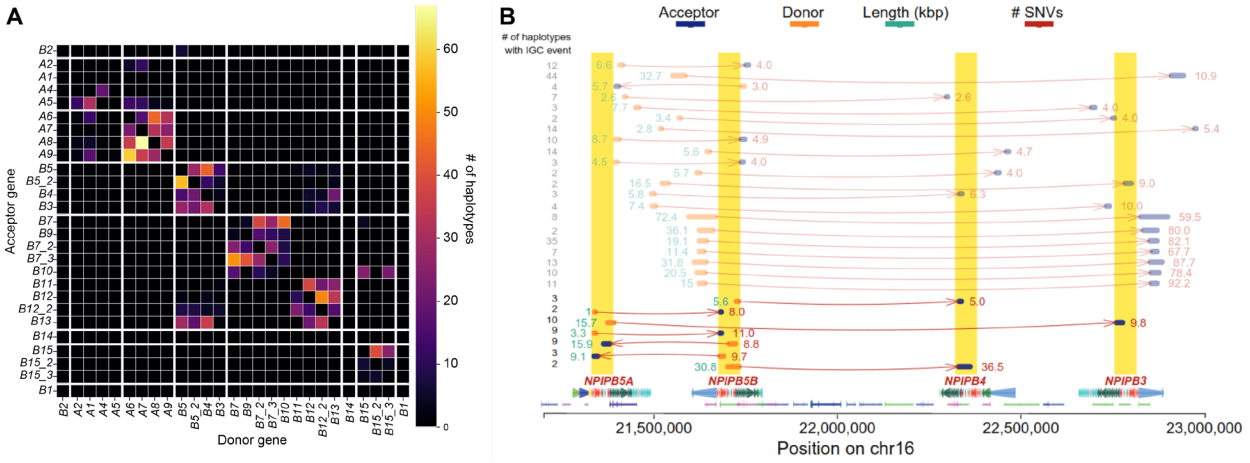


Figure S1. *NPIP* interlocus gene conversion (IGC) and structural changes. A) Counts of IGC events detected between paralogs for a subset of 94 haplotypes. X-axis labels indicate the IGC donor, and the y-axis the acceptor. **B)** IGC events at the B3-5 locus. Arrows indicate acceptors (blue) and donors (orange). The length of each event and number of SNVs are shown to the left and right, respectively.

	10									20									30												
NPIPConsensus	M	V	K	L	S	I	V	L	T	P	Q	F	L	S	H	D	Q	G	Q	L	T	K	E	L	Q	Q	H	V	K	S	V	T	C	P	C	E	Y	L	R	K
A2_2	M	V	K	L	S	I	V	L	T	P	R	F	L	S	H	D	Q	G	Q	L	T	K	E	L	Q	Q	H	V	K	S	V	T	C	P	C	E	Y	L	R	K
B345L1_1	M	V	K	L	S	I	V	L	T	P	Q	F	L	S	H	D	Q	G	Q	L	T	K	E	L	Q	Q	H	V	K	S	V	T	C	P	C	E	Y	L	R	K
B5_1	M	V	K	L	S	I	V	L	T	P	Q	F	L	S	H	D	Q	G	Q	L	T	K	E	L	Q	Q	H	V	K	S	V	T	C	P	C	E	Y	L	R	K
B12B13_1	M	V	K	L	S	I	V	L	T	P	Q	F	L	S	H	D	Q	G	Q	L	T	K	E	L	Q	Q	H	V	K	S	V	T	C	P	C	E	Y	L	R	K
B11_1	M	V	K	L	S	I	V	L	T	P	Q	F	L	S	H	D	Q	G	Q	L	T	K	E	L	Q	Q	H	V	K	S	V	T	C	P	C	E	Y	L	R	K
B35L1_1	M	V	K	L	S	I	V	L	T	P	Q	F	L	S	H	D	Q	G	Q	L	T	K	E	L	Q	Q	H	V	K	S	V	T	C	P	C	E	Y	L	R	K
B11_2	M	V	K	L	S	I	V	L	T	P	Q	F	L	S	H	D	Q	G	Q	L	T	K	E	L	Q	Q	H	V	K	S	V	T	C	P	C	E	Y	L	R	K
B3_1	M	V	K	L	S	I	V	L	T	P	Q	F	L	S	H	D	Q	G	Q	L	T	K	E	L	Q	Q	H	V	K	S	V	T	C	P	C	E	Y	L	R	K
B3_2	M	V	K	L	S	I	V	L	T	P	Q	F	L	S	H	D	Q	G	Q	L	T	K	E	L	Q	Q	H	V	K	S	V	T	C	P	C	E	Y	L	R	K
B6_1	M	V	K	L	S	I	V	L	T	P	Q	F	L	S	H	D	Q	S	Q	L	T	K	E	L	Q	Q	H	V	K	S	V	T	C	P	C	E	Y	L	R	K
B14_2	M	V	K	L	S	I	V	L	T	P	Q	F	L	S	H	D	Q	G	P	L	T	K	E	L	Q	Q	H	V	K	S	V	T	C	P	C	E	Y	L	N	Q
ACSM1	V	V	K	A	F	I	V	L	T	P	Q	F	L	S	H	D	K	D	Q	L	T	K	E	L	Q	Q	H	V	K	S	V	T	A	P	Y	K	Y	P	R	K
Mouse_ACSM1	V	V	K	A	F	I	V	L	N	P	E	F	L	S	H	D	Q	E	Q	L	I	K	E	L	Q	H	H	V	K	S	V	T	A	P	Y	K	Y	P	R	K
ACSM2A	V	V	K	A	F	V	V	L	A	S	Q	F	L	S	H	D	P	E	Q	L	T	K	E	L	Q	Q	H	V	K	S	V	T	A	P	Y	K	Y	P	R	K

Figure S2. *NPIP* start sequence derived from *ACSM1*. Multiple sequence alignment of human *NPIP* paralog start sequences compared to human *ACSM1* and *ACSM2A* and mouse *ACSM1*.

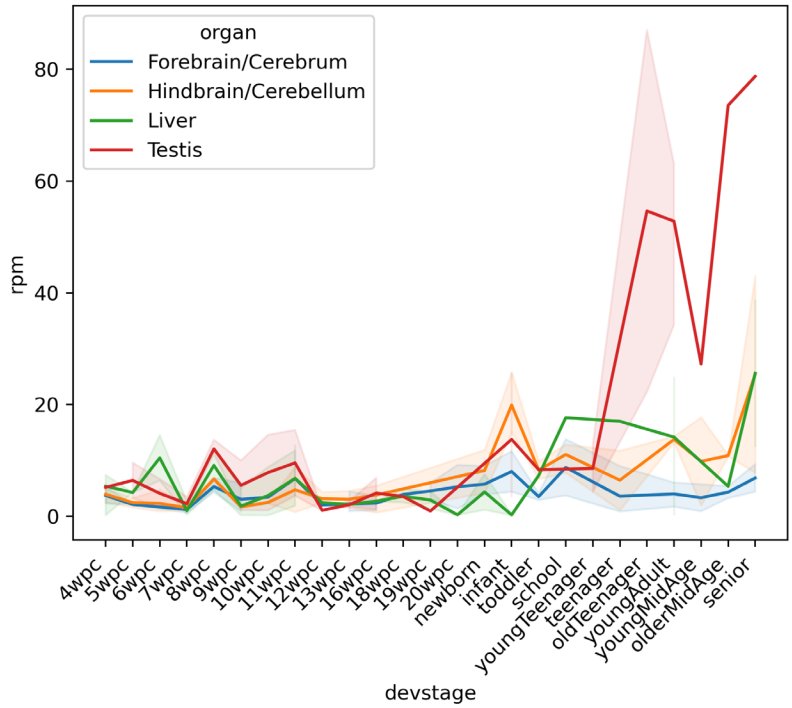


Figure S3. *NPIPB15* expression across development. Short-read expression estimates for human developmental timepoints in four tissues, using unique k-mers for paralog identity. Transparent error bands represent 95% confidence interval of replicates.

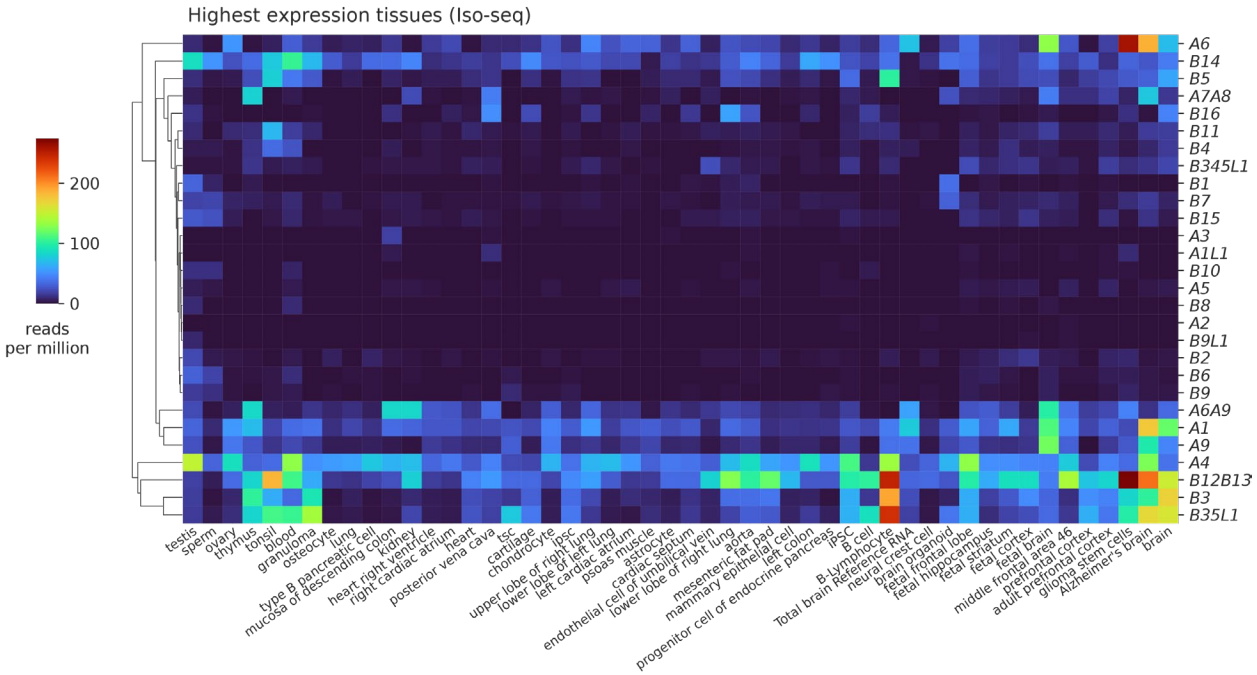


Figure S4. Variable expression of *NPIP* paralogs across tissues and cell types. Iso-Seq expression estimates for 50 tissues showing the highest *NPIP* expression, clustered with UPGMA.

Table S1. Iso-Seq data sources.

Tissue/cell type	Total reads	NPIP reads	NPIP percent	Enrichment	source	ID
fetal brain_hyb	87,228	76,768	88.01%	NPIP	Eichler	
fetal brain_hyb	231,686	57,712	24.91%	NPIP	Eichler	
testis_hyb	64,506	52,749	81.77%	NPIP	Eichler	
midfrontal cortex, snRNA-seq, exome capture	3,217,137	41,588	1.29%	exome	SRA	SRR14808731
midfrontal cortex, snRNA-seq, exome capture	2,911,790	38,480	1.32%	exome	SRA	SRR14808729
midfrontal cortex, snRNA-seq, exome capture	2,729,943	34,084	1.25%	exome	SRA	SRR14808734
midfrontal cortex, snRNA-seq, exome capture	2,445,841	32,880	1.34%	exome	SRA	SRR14808732
midfrontal cortex, snRNA-seq, exome capture	2,557,766	32,259	1.26%	exome	SRA	SRR14808730
midfrontal cortex, snRNA-seq, exome capture	2,332,661	30,964	1.33%	exome	SRA	SRR14808733
midfrontal cortex, snRNA-seq, exome capture	1,993,591	26,362	1.32%	exome	SRA	SRR14808728
midfrontal cortex, snRNA-seq, exome capture	2,012,223	25,558	1.27%	exome	SRA	SRR14808735
testis_hyb	29,419	24,867	84.53%	NPIP	Eichler	
adult brain	163,565,243	20,085	0.01%	TFs	SRA	SRR11492433
brain	3,686,653	16,990	0.46%		ANVIL phs002316	cmh003217-01_IsoSeqv2-IS-brain_CCS
midfrontal cortex, snRNA-seq, exome capture	1,556,910	15,573	1.00%	exome	SRA	SRR14808740
fetal frontal lobe	14,080,409	14,308	0.10%		SRA	SRR12660773
fetal frontal lobe	16,269,707	12,852	0.08%		SRA	SRR12660772
heart	44,292,501	12,079	0.03%		SRA	SRR16352587

HG002	37,286,400	11,621	0.03%		https://downloads.pacbcloud.com/public/dataset/Kinnex-full-length-RNA/	HG002
brain	5,549,730	11,483	0.21%		SRA	DRR481124
brain_hyb	346,542	10,873	3.14%	NPIP	Eichler	
brain	5,516,891	10,412	0.19%		SRA	DRR481120
CHM1_hyb	32,799	10,379	31.64%	NPIP	Eichler	
midfrontal cortex, snRNA-seq, exome capture	1,068,706	10,192	0.95%	exome	SRA	SRR14808736
fetal brain_hyb	16,449	9,924	60.33%	NPIP	Eichler	
ESC	52,741,436	9,720	0.02%		SRA	SRR25855036
adult prefrontal cortex	8,981,446	9,698	0.11%		SRA	SRR12660778
brain	4,585,082	9,277	0.20%		SRA	DRR481118
ESC	46,888,337	8,898	0.02%		SRA	SRR25855035
adult prefrontal cortex	15,161,873	8,437	0.06%		SRA	SRR12660776
brain	5,076,130	8,328	0.16%		SRA	DRR481115
adult brain_hyb	16,717	7,533	45.06%	NPIP	Eichler	
midfrontal cortex, snRNA-seq, exome capture	754,462	7,508	1.00%	exome	SRA	SRR14808741
ESC	46,050,903	7,440	0.02%		SRA	SRR25855038
ESC	41,921,314	7,352	0.02%		SRA	SRR25855034
ESC	39,549,999	6,871	0.02%		SRA	SRR25855037
B-Lymphocyte	4,406,914	6,519	0.15%		SRA	SRR18074968
Alzheimer's brain	4,271,005	6,120	0.14%		https://downloads.pacbcloud.com/public/dataset/Alzheimer2019_IsoSeq/	Alzheimer
brain	4,345,676	6,047	0.14%		ANVIL phs002313	cmh003036-01_IsoSeqv2-IS-brain_CCS
brain	6,262,294	5,928	0.09%		SRA	DRR481119
brain	4,639,736	5,456	0.12%		SRA	DRR481121

midfrontal cortex, snRNA-seq, exome capture	549,013	5,397	0.98%	exome	SRA	SRR14808738
ESC	32,040,262	5,279	0.02%		SRA	SRR25855040
ESC	32,555,688	5,277	0.02%		SRA	SRR25855039
brain	5,174,570	5,188	0.10%		SRA	DRR481122
CHM1_hyb	71,072	4,863	6.84%	NPIP	Eichler	
brain	6,169,651	4,598	0.07%		SRA	DRR481116
CHM1_hyb	65,501	4,366	6.67%	NPIP	Eichler	
brain	4,322,512	4,016	0.09%		SRA	DRR481117
iPSC	5,081,985	3,991	0.08%		SRA	SRR11729902
iPSC	4,588,890	3,956	0.09%		SRA	SRR11729903
midfrontal cortex, snRNA-seq, exome capture	402,408	3,827	0.95%	exome	SRA	SRR14808739
glioma stem cells	2,557,782	3,732	0.15%		SRA	SRR17260131
brain	3,081,084	3,660	0.12%		ANVIL phs002225	cmh001768-01_IsoSeqv2-IS-brain_CCS
HepG2	2,612,690	3,605	0.14%		ENCODE	ENCFF483HTA
iPSC	5,387,513	3,539	0.07%		SRA	SRR11729904
brain	5,992,000	3,474	0.06%		SRA	DRR481123
iPSC	5,500,694	3,324	0.06%		SRA	SRR11729905
adult prefrontal cortex	7,816,194	3,003	0.04%		SRA	SRR12660779
adult prefrontal cortex	7,467,067	2,983	0.04%		SRA	SRR12660775
fetal brain_hyb	4,071	2,904	71.33%	NPIP	Eichler	
sperm	9,924,497	2,902	0.03%		SRA	SRR10123831
brain	5,974,301	2,887	0.05%		SRA	DRR481125
brain	3,517,268	2,859	0.08%		ANVIL phs002321	PBIsoSeq_cmh002013-01_brain_combined_CCS

middle frontal area 46	2,756,340	2,840	0.10%		ENCODE	ENCFF156TTD
midfrontal cortex, snRNA-seq, exome capture	297,670	2,796	0.94%	exome	SRA	SRR14808737
iPSC	3,765,454	2,747	0.07%		SRA	SRR18074969
ipsc	3,641,579	2,742	0.08%		ANVIL phs002269	cmh002248-01_IsoSeqv2_iPSC-Cell5_CCS
middle frontal area 46	2,533,014	2,710	0.11%		ENCODE	ENCFF446EFU
CHM1_hyb	318,234	2,561	0.80%	NPIP	Eichler	
ipsc	3,733,654	2,377	0.06%		ANVIL phs002284	cmh002429-01_IsoSeqv2_iPSC-Cell2_CCS
middle frontal area 46	2,719,170	2,275	0.08%		ENCODE	ENCFF311CZO
UHR	6,374,832	2,214	0.03%		https://downloads.pacbcloud.com/public/dataset/UHR_IsoSeq/	UHR
brain_hyb	63,154	2,152	3.41%	NPIP	Eichler	
OCI-LY7	1,805,339	2,124	0.12%		ENCODE	ENCFF417UQV
tsc	7,832,608	2,085	0.03%		ANVIL phs002322	Re-run_of_cmh003068-01C_TSC_IsoSeq_GRCh38-tsc_CCS
brain	5,749,358	2,009	0.03%		SRA	DRR481126
iPSC	4,081,253	1,888	0.05%		SRA	SRR18074967
adult brain, fetal brain, heart, liver, pancreas, placenta	20,205,611	1,869	0.01%	TFs	SRA	SRR11492435
ipsc	3,425,947	1,867	0.05%		ANVIL phs002275	cmh002275-01_IsoSeqv2_iPSC-Cell1_CCS
thymus	2,815,124	1,837	0.07%		ANVIL phs002218	cmh001715-06_IsoSeqv2-Cell2_CCS
ipsc	3,346,776	1,789	0.05%		ANVIL phs002220	cmh001748-01_IsoSeqv2_iPSC-Cell1_CCS
tonsil	1,940,008	1,750	0.09%		ANVIL phs002211	cmh001637-01_IsoSeqv2-Cell4_CCS
middle frontal area 46	2,197,031	1,711	0.08%		ENCODE	ENCFF785KVJ

midfrontal cortex, snRNA-seq, exome capture	194,380	1,704	0.88%	exome	SRA	SRR14808742
middle frontal area 46	2,112,299	1,686	0.08%		ENCODE	ENCFF838DFB
ipsc	2,631,399	1,665	0.06%		ANVIL phs002270	cmh002249-01_IsoSeqv2_iPSC-Cell1_CCS
ipsc	3,020,021	1,574	0.05%		ANVIL phs002304	cmh002770-01_IsoSeqv2_iPSC-Cell1_CCS
fetal striatum	3,007,922	1,559	0.05%		SRA	SRR12660769
K562	2,698,366	1,558	0.06%		ENCODE	ENCFF429VVB
blood	2,063,350	1,463	0.07%		ANVIL phs002215	cmh001658-05_IsoSeqv2-Cell1_CCS
ipsc	3,500,298	1,432	0.04%		ANVIL phs002261	cmh002189-01_IsoSeqv2_iPSC-Cell3_CCS
OCI-LY7	1,475,971	1,418	0.10%		ENCODE	ENCFF511KJB
granuloma	2,368,783	1,413	0.06%		ANVIL phs002231	cmh001818-01_IsoSeqv2-Cell3_CCS
fetal striatum	2,486,524	1,398	0.06%		SRA	SRR12660777
ipsc	3,309,362	1,342	0.04%		ANVIL phs002253	cmh002114-01_IsoSeqv2_iPSC-Cell2_CCS
ipsc	3,279,501	1,332	0.04%		ANVIL phs002310	cmh002971-01_IsoSeqv2_iPSC-Cell2_CCS
blood	1,871,523	1,325	0.07%		ANVIL phs002216	cmh001658-06_IsoSeqv2-Cell1_CCS
fetal hippocampus	2,805,406	1,320	0.05%		SRA	SRR12660770
lower lobe of right lung	1,900,303	1,320	0.07%		ENCODE	ENCFF250IWT
ipsc	3,197,844	1,320	0.04%		ANVIL phs002281	cmh002381-01_IsoSeqv2_iPSC-Cell1_CCS
ipsc	3,370,620	1,311	0.04%		ANVIL phs002291	cmh002531-01_IsoSeqv2-Cell1_CCS
mesenteric fat pad	1,898,739	1,310	0.07%		ENCODE	ENCFF907SZK
MCF-7	1,742,102	1,294	0.07%		ENCODE	ENCFF887DGG

ipsc	2,430,615	1,283	0.05%		ANVIL phs002306	cmh002821-01_IsoSeqv2_iPSC-Cell2_CCS
Panc1	1,687,025	1,276	0.08%		ENCODE	ENCFF990CUL
aorta	1,175,321	1,271	0.11%		ENCODE	ENCFF144KHH
blood	1,796,512	1,266	0.07%		ANVIL phs002214	cmh001658-04_IsoSeqv2-Cell1_CCS
MCF 10A	1,989,087	1,253	0.06%		ENCODE	ENCFF702KLU
midfrontal cortex, snRNA-seq, no capture	1,480,440	1,250	0.08%		SRA	SRR14808727
ipsc	2,709,700	1,226	0.05%		ANVIL phs002297	cmh002651-01_IsoSeqv2_iPSC-Cell5_CCS
K562	2,800,443	1,221	0.04%		ENCODE	ENCFF634YSN
blood	1,295,479	1,211	0.09%		ANVIL phs002213	cmh001658-01_IsoSeqv2-Cell1_CCS
ipsc	3,184,033	1,176	0.04%		ANVIL phs002296	cmh002650-01_IsoSeqv2_iPSC-Cell4_CCS
ipsc	3,110,739	1,169	0.04%		ANVIL phs002280	cmh002362-01_IsoSeqv2-Cell1_CCS
aorta	2,043,939	1,154	0.06%		ENCODE	ENCFF902BIU
mammary epithelial cell	2,517,178	1,139	0.05%		ENCODE	ENCFF617YVE
fetal hippocampus	2,868,834	1,133	0.04%		SRA	SRR12660771
ipsc	3,417,624	1,132	0.03%		ANVIL phs002224	cmh001761-01_IsoSeqv2_iPSC-Cell5_CCS
ipsc	3,039,743	1,122	0.04%		ANVIL phs002267	cmh002222-01_IsoSeqv2_iPSC-Cell4_CCS
PC-9	1,720,299	1,114	0.06%		ENCODE	ENCFF860AWQ
middle frontal area 46	1,002,212	1,110	0.11%		ENCODE	ENCFF206TQZ
prefrontal cortex	2,885,132	1,109	0.04%		SRA	SRR23409825
PC-3	2,062,428	1,106	0.05%		ENCODE	ENCFF834KTE
endothelial cell of umbilical vein	2,562,641	1,103	0.04%		ENCODE	ENCFF033LRZ
ipsc	3,525,088	1,101	0.03%		ANVIL phs002219	cmh001743-01_IsoSeqv2-Cell1_CCS

cartilage	2,724,198	1,095	0.04%		ANVIL phs002320	cmh003629-01_IsoSeqv2-Cell3_CCS
midfrontal cortex, snRNA-seq, no capture	1,115,396	1,086	0.10%		SRA	SRR14808726
fetal cortex	2,325,960	1,073	0.05%		SRA	SRR12660774
A673	1,786,525	1,073	0.06%		ENCODE	ENCFF861BKY
blood	834,019	1,070	0.13%		ANVIL phs002228	cmh001807-01_IsoSeqv2_iPSC-Cell1_CCS
ipsc	2,619,527	1,069	0.04%		ANVIL phs002256	cmh002145-01_IsoSeqv2_iPSC-Cell3_CCS
ipsc	2,948,572	1,067	0.04%		ANVIL phs002307	cmh002832-01_IsoSeqv2_iPSC-Cell4_CCS
ovary	1,359,776	1,056	0.08%		ENCODE	ENCFF422XLS
K562	2,020,521	1,035	0.05%		ENCODE	ENCFF696GDL
ipsc	2,942,142	1,015	0.03%		ANVIL phs002312	cmh003006-01_IsoSeqv2_iPSC-Cell3_CCS
adult brain, fetal brain, heart, liver, pancreas, placenta	4,683,641	1,014	0.02%	TFs	SRA	SRR11492438
mammary epithelial cell	2,635,356	1,014	0.04%		ENCODE	ENCFF237FMP
ipsc	4,827,811	1,003	0.02%		ANVIL phs002229	cmh001807-01_IsoSeqv2_iPSC-Neur-Cell1_CCS
ipsc	3,045,536	996	0.03%		ANVIL phs002279	cmh002355-01_IsoSeqv2_iPSC-Cell3_CCS
MCF 10A	1,953,578	981	0.05%		ENCODE	ENCFF041EGI
middle frontal area 46	1,641,325	978	0.06%		ENCODE	ENCFF827DUW
ipsc	3,202,327	967	0.03%		ANVIL phs002230	cmh001807-01_IsoSeqv2-Cell1_CCS
ipsc	2,549,544	946	0.04%		ANVIL phs002298	cmh002656-01_IsoSeqv2_iPSC-Cell1_CCS
ipsc	2,572,752	945	0.04%		ANVIL phs002314	cmh003042-01_IsoSeqv2_iPSC-Cell2_CCS
PC-9	1,810,603	927	0.05%		ENCODE	ENCFF107YRM

WTC11	2,563,543	913	0.04%		ENCODE	ENCFF563QZR
ipsc	2,723,160	912	0.03%		ANVIL phs002240	cmh001982-01_IsoSeqv2_iPSC-Cell4_CCS
ipsc	2,777,639	910	0.03%		ANVIL phs002209	cmh001573-01_IsoSeqv2_iPSC-Cell3_CCS
iPSC	4,116,429	904	0.02%		SRA	SRR18130587
endothelial cell of umbilical vein	2,611,657	895	0.03%		ENCODE	ENCFF096UHO
ipsc	2,571,919	877	0.03%		ANVIL phs002308	cmh002853-01_IsoSeqv2_iPSC-Cell5_CCS
ipsc	3,173,897	877	0.03%		ANVIL phs002241	cmh001991-01_IsoSeqv2_iPSC-Cell4_CCS
ipsc	3,325,405	874	0.03%		ANVIL phs002294	cmh002557-01_IsoSeqv2-Cell1_CCS
ipsc	2,251,955	874	0.04%		ANVIL phs002282	cmh002397-01_IsoSeqv2-Cell1_CCS
ipsc	3,711,640	873	0.02%		ANVIL phs002221	cmh001748-04_IsoSeqv2-Cell1_CCS
HL-60	1,445,126	867	0.06%		ENCODE	ENCFF321PMP
upper lobe of right lung	1,716,474	863	0.05%		ENCODE	ENCFF934MBW
ipsc	1,908,723	838	0.04%		ANVIL phs002238	cmh001971-01_IsoSeqv2_iPSC-Cell3_CCS
ipsc	3,334,782	832	0.02%		ANVIL phs002244	cmh002004-01_IsoSeqv2_iPSC-Cell5_CCS
ipsc	3,021,929	826	0.03%		ANVIL phs002259	cmh002178-01_IsoSeqv2_iPSC-Cell1_CCS
ipsc	2,805,822	819	0.03%		ANVIL phs002271	cmh002255-04_IsoSeqv2_iPSC-Cell2_CCS
middle frontal area 46	2,505,116	815	0.03%		ENCODE	ENCFF708BOP
ipsc	3,150,460	809	0.03%		ANVIL phs002288	cmh002478-01_IsoSeqv2_iPSC-Cell3_CCS
Calu3	1,733,596	793	0.05%		ENCODE	ENCFF548JGS

ipsc	1,889,535	790	0.04%		ANVIL phs002285	cmh002430-01_IsoSeqv2_iPSC-Cell1_CCS
Calu3	1,720,338	787	0.05%		ENCODE	ENCFF569KOA
ovary	2,155,346	784	0.04%		ENCODE	ENCFF756AHG
ipsc	2,270,248	784	0.03%		ANVIL phs002309	cmh002871-01_IsoSeqv2_iPSC-Cell1_CCS
ipsc	2,821,230	783	0.03%		ANVIL phs002295	cmh002618-01_IsoSeqv2_iPSC-Cell1_CCS
ipsc	2,966,435	779	0.03%		ANVIL phs002303	cmh002743-01_IsoSeqv2_iPSC-Cell5_CCS
ipsc	3,030,798	776	0.03%		ANVIL phs002234	cmh001935-01_IsoSeqv2_iPSC-Cell1_CCS
Ptr lymp_hyb	30,464	770	2.53%	NPIP	Eichler	
Total brain Reference RNA	1,907,981	761	0.04%		SRA	SRR16762346
endodermal cell	2,420,444	761	0.03%		ENCODE	ENCFF561HIY
ipsc	3,702,185	743	0.02%		ANVIL phs002302	cmh002697-01_IsoSeqv2_iPSC-Cell4_CCS
Caco-2	1,733,868	742	0.04%		ENCODE	ENCFF827OXR
ipsc	2,841,911	742	0.03%		ANVIL phs002250	cmh002059-01_IsoSeqv2_iPSC-Cell1_CCS
cartilage	3,196,670	740	0.02%		ANVIL phs002317	cmh003306-01_IsoSeqv2-Cell3_CCS
ipsc	2,744,386	733	0.03%		ANVIL phs002251	cmh002066-01_IsoSeqv2_iPSC-Cell1_CCS
technical sample	2,409,856	730	0.03%		ENCODE	ENCFF743MYM
WTC11	3,007,792	726	0.02%		ENCODE	ENCFF245IPA
ipsc	3,245,072	726	0.02%		ANVIL phs002212	cmh001648-01_IsoSeqv2_iPSC-Cell4_CCS
ipsc	2,766,009	723	0.03%		ANVIL phs002226	cmh001796-01_IsoSeqv2-Cell1_CCS

ipsc	2,126,667	721	0.03%		ANVIL phs002210	cmh001581-01_IsoSeqv2_iPSC-Cell4_CCS
ipsc	2,450,668	719	0.03%		ANVIL phs002293	cmh002550-04_IsoSeqv2_iPSC-Cell2_CCS
ipsc	3,676,074	709	0.02%		ANVIL phs002232	cmh001866-01_IsoSeqv2-Cell1_CCS
ipsc	2,837,510	707	0.02%		ANVIL phs002299	cmh002657-01_IsoSeqv2_iPSC-Cell2_CCS
A673	1,378,362	706	0.05%		ENCODE	ENCFF168MIB
GM12878	1,575,236	701	0.04%		ENCODE	ENCFF417VHJ
ipsc	2,602,238	701	0.03%		ANVIL phs002278	cmh002350-01_IsoSeqv2_iPSC-Cell1_CCS
ipsc	1,994,116	695	0.03%		ANVIL phs002252	cmh002105-01_IsoSeqv2_iPSC-Cell2_CCS
ipsc	2,373,881	690	0.03%		ANVIL phs002268	cmh002243-01_IsoSeqv2_iPSC-Cell4_CCS
ipsc	2,387,973	689	0.03%		ANVIL phs002277	cmh002326-01_IsoSeqv2_iPSC-Cell2_CCS
ipsc	2,773,005	689	0.02%		ANVIL phs002247	cmh002039-01_IsoSeqv2_iPSC-Cell1_CCS
endodermal cell	1,492,674	688	0.05%		ENCODE	ENCFF235QXW
mucosa of descending colon	1,355,283	687	0.05%		ENCODE	ENCFF387HPO
ipsc	2,102,024	684	0.03%		ANVIL phs002290	cmh002497-01_IsoSeqv2_iPSC-Cell1_CCS
ipsc	2,770,350	682	0.02%		ANVIL phs002254	cmh002124-04_IsoSeqv2_iPSC-Cell1_CCS
ipsc	2,141,361	681	0.03%		ANVIL phs002274	cmh002268-01_IsoSeqv2_iPSC-Cell2_CCS
HL-60	1,141,416	674	0.06%		ENCODE	ENCFF173QRD
ipsc	2,655,894	665	0.03%		ANVIL phs002301	cmh002692-01_IsoSeqv2_iPSC-Cell3_CCS

HL-60	1,907,878	657	0.03%		ENCODE	ENCFF810XST
ipsc	2,987,085	656	0.02%		ANVIL phs002255	cmh002126-01_IsoSeqv2_iPSC-Cell1_CCS
ipsc	2,810,743	652	0.02%		ANVIL phs002272	cmh002258-01_IsoSeqv2_iPSC-Cell5_CCS
HCT116	1,460,741	645	0.04%		ENCODE	ENCFF337VWR
ipsc	2,353,298	645	0.03%		ANVIL phs002263	cmh002197-01_IsoSeqv2_iPSC-Cell3_CCS
ipsc	3,330,144	638	0.02%		ANVIL phs002265	cmh002208-01_IsoSeqv2_iPSC-Cell2_CCS
ipsc	2,845,020	625	0.02%		ANVIL phs002311	cmh002985-01_IsoSeqv2_iPSC-Cell1_CCS
cardiac septum	1,845,269	624	0.03%		ENCODE	ENCFF680XXE
ipsc	2,056,376	623	0.03%		ANVIL phs002258	cmh002173-01_IsoSeqv2_iPSC-Cell5_CCS
ipsc	2,546,320	621	0.02%		ANVIL phs002260	cmh002183-01_IsoSeqv2_iPSC-Cell2_CCS
lung	3,309,532	619	0.02%		SRA	SRR23517787
HL-60	2,432,288	619	0.03%		ENCODE	ENCFF274DYS
ipsc	3,092,308	619	0.02%		ANVIL phs002249	cmh002048-01_IsoSeqv2-Cell1_CCS
technical sample	1,731,829	617	0.04%		ENCODE	ENCFF372YUA
cartilage	2,040,261	616	0.03%		ANVIL phs002318	cmh003465-01_IsoSeqv2-Cell4_CCS
ipsc	2,344,094	613	0.03%		ANVIL phs002300	cmh002668-01_IsoSeqv2_iPSC-Cell2_CCS
HL-60	2,115,302	612	0.03%		ENCODE	ENCFF666JCQ
ipsc	2,303,484	612	0.03%		ANVIL phs002257	cmh002160-01_IsoSeqv2_iPSC-Cell4_CCS
fetal brain	352,432	610	0.17%		Eichler	SRR12524788
endodermal cell	1,511,682	600	0.04%		ENCODE	ENCFF712CBL

HL-60	1,090,893	598	0.05%		ENCODE	ENCFF145QNC
HL-60	1,868,899	584	0.03%		ENCODE	ENCFF805YXK
ipsc	2,915,315	582	0.02%		ANVIL phs002248	cmh002039-04_IsoSeqv2_iPSC-Cell1_CCS
kidney	901,716	572	0.06%		ENCODE	ENCFF492BYP
posterior vena cava	621,106	570	0.09%		ENCODE	ENCFF960KBO
ipsc	2,672,972	565	0.02%		ANVIL phs002233	cmh001866-04_IsoSeqv2-Cell1_CCS
HL-60	1,886,923	562	0.03%		ENCODE	ENCFF032UMC
right cardiac atrium	2,489,853	558	0.02%		ENCODE	ENCFF899MTI
chondrocyte	1,589,522	557	0.04%		ENCODE	ENCFF011BFA
ipsc	2,109,227	553	0.03%		ANVIL phs002223	cmh001760-01_IsoSeqv2_iPSC-Cell1_CCS
Caco-2	1,449,402	551	0.04%		ENCODE	ENCFF649CYY
HL-60	894,750	551	0.06%		ENCODE	ENCFF260AJN
H9	3,297,669	550	0.02%		ENCODE	ENCFF688QGB
ipsc	2,374,637	539	0.02%		ANVIL phs002242	cmh001992-01_IsoSeqv2_iPSC-Neur-Cell2_CCS
middle frontal area 46	521,182	538	0.10%		ENCODE	ENCFF260AWP
HL-60	1,633,482	537	0.03%		ENCODE	ENCFF457TIY
endodermal cell	2,536,086	535	0.02%		ENCODE	ENCFF142LPL
technical sample	1,226,796	532	0.04%		ENCODE	ENCFF525JUC
ipsc	3,151,961	528	0.02%		ANVIL phs002208	cmh001370_IsoSeqv2_iPSC-Cell3_CCS
ipsc	2,246,668	525	0.02%		ANVIL phs002245	cmh002006-01_IsoSeqv2_iPSC-Cell1_CCS
ipsc	2,687,080	523	0.02%		ANVIL phs002283	cmh002426-01_IsoSeqv2_iPSC-Cell5_CCS
HL-60	1,331,716	522	0.04%		ENCODE	ENCFF407SUN

ipsc	2,909,351	522	0.02%		ANVIL phs002207	cmh001256_IsoSeqv2_iPSC-Cell2_CCS
HL-60	1,890,595	521	0.03%		ENCODE	ENCFF564TOK
ipsc	2,689,569	521	0.02%		ANVIL phs002243	cmh001996-01_IsoSeqv2_iPSC-Cell1_CCS
HL-60	2,878,033	519	0.02%		ENCODE	ENCFF609QIM
IMR-90	1,234,075	515	0.04%		ENCODE	ENCFF197DCI
HL-60	1,950,685	512	0.03%		ENCODE	ENCFF782UMU
testis	550,195	508	0.09%		Eichler	SRR12544673
ipsc	2,880,112	507	0.02%		ANVIL phs002264	cmh002207-01_IsoSeqv2_iPSC-Cell1_CCS
ipsc	1,915,058	504	0.03%		ANVIL phs002237	cmh001968-01_IsoSeqv2_iPSC-Cell2_CCS
cartilage	2,133,723	503	0.02%		ANVIL phs002235	cmh001950-01_IsoSeqv2-Cell2_CCS
H9	2,837,013	499	0.02%		ENCODE	ENCFF272VSN
lower lobe of left lung	2,602,494	498	0.02%		ENCODE	ENCFF552NVU
ipsc	2,301,457	497	0.02%		ANVIL phs002292	cmh002548-01_IsoSeqv2_iPSC-Cell4_CCS
chondrocyte	1,346,171	496	0.04%		ENCODE	ENCFF342HOS
soft tissue	3,073,217	490	0.02%		ANVIL phs002315	cmh003187-01_IsoSeqv2-Cell1_CCS
psoas muscle	1,839,037	482	0.03%		ENCODE	ENCFF750LYC
ipsc	2,555,892	481	0.02%		ANVIL phs002239	cmh001977-01_IsoSeqv2-Cell1_CCS
left colon	1,312,593	478	0.04%		ENCODE	ENCFF245MBY
lower lobe of left lung	777,834	474	0.06%		ENCODE	ENCFF341BSQ
ipsc	2,456,991	471	0.02%		ANVIL phs002276	cmh002319-01_IsoSeqv2_iPSC-Cell4_CCS
HL-60	1,231,903	466	0.04%		ENCODE	ENCFF199OTG
HL-60	2,042,269	462	0.02%		ENCODE	ENCFF417RYS

CHM13	1,553,081	459	0.03%		Eichler	SRS798661
ipsc	2,413,714	456	0.02%		ANVIL phs002266	cmh002208-01_IsoSeqv2_iPSC-EB-Cell3_CCS
endodermal cell	1,187,556	455	0.04%		ENCODE	ENCFF731HST
ipsc	2,222,856	455	0.02%		ANVIL phs002273	cmh002264-01_IsoSeqv2_iPSC-Cell3_CCS
progenitor cell of endocrine pancreas	1,698,651	452	0.03%		ENCODE	ENCFF471YEK
ipsc	1,328,503	441	0.03%		ANVIL phs002227	cmh001805-01_IsoSeqv2-Cell1_CCS
right cardiac atrium	1,815,720	440	0.02%		ENCODE	ENCFF905RVF
ipsc	2,818,872	435	0.02%		ANVIL phs002236	cmh001961-04_IsoSeqv2_iPSC-Cell1_CCS
ipsc	2,080,871	421	0.02%		ANVIL phs002289	cmh002495-01_IsoSeqv2_iPSC-Cell3_CCS
ipsc	1,885,651	420	0.02%		ANVIL phs002305	cmh002818-01_IsoSeqv2_iPSC-Cell1_CCS
ipsc	1,660,268	420	0.03%		ANVIL phs002262	cmh002193-01_IsoSeqv2-Cell1_CCS
ipsc	2,237,111	417	0.02%		ANVIL phs002222	cmh001749-01_IsoSeqv2_iPSC-Cell2_CCS
ipsc	1,460,148	413	0.03%		ANVIL phs002206	cmh000118_IsoSeqv2_iPSC-Cell1_CCS
WTC11	1,632,302	411	0.03%		ENCODE	ENCFF370NFS
H1	1,285,097	399	0.03%		ENCODE	ENCFF436GKZ
cartilage	1,766,213	391	0.02%		ANVIL phs002319	cmh003514-01_IsoSeqv2-Cell1_CCS
Ptr brain_hyb	33,653	388	1.15%	NPIP	Eichler	
ipsc	2,112,894	386	0.02%		ANVIL phs002246	cmh002017-01_IsoSeqv2_iPSC-Cell5_CCS
osteocyte	2,115,770	378	0.02%		ENCODE	ENCFF556DYU
ipsc	2,197,797	367	0.02%		ANVIL phs002287	cmh002459-01_IsoSeqv2_iPSC-Cell2_CCS

heart right ventricle	1,345,058	358	0.03%		ENCODE	ENCFF793PGJ
type B pancreatic cell	1,647,160	355	0.02%		ENCODE	ENCFF489XQJ
type B pancreatic cell	1,680,369	342	0.02%		ENCODE	ENCFF580BQX
osteocyte	1,683,994	337	0.02%		ENCODE	ENCFF560XTG
GM12878	2,499,375	329	0.01%		ENCODE	ENCFF694DIE
ipsc	1,446,090	328	0.02%		ANVIL phs002217	cmh001712-01_IsoSeqv2-Cell1_CCS
heart right ventricle	1,685,360	321	0.02%		ENCODE	ENCFF615FIC
melanoma	1,872,695	320	0.02%		https://downloads.pacbcloud.com/public/dataset/Melanoma2019_IsoSeq/	Melanoma
chondrocyte	916,864	302	0.03%		ENCODE	ENCFF352CGL
heart right ventricle	1,106,468	294	0.03%		ENCODE	ENCFF425VDL
posterior vena cava	1,350,298	293	0.02%		ENCODE	ENCFF658OZB
left ventricle myocardium inferior	1,430,840	290	0.02%		ENCODE	ENCFF793CMQ
ipsc	2,413,102	285	0.01%		ANVIL phs002286	cmh002444-01_IsoSeqv2_iPSC-Cell1_CCS
adult brain, fetal brain, heart, liver, pancreas, placenta	11,863,697	282	0.00%	TFs	SRA	SRR11492437
adrenal gland	1,594,860	272	0.02%		ENCODE	ENCFF417ALN
Right ventricle myocardium superior	1,648,621	262	0.02%		ENCODE	ENCFF665LBS
left ventricle myocardium superior	1,469,933	256	0.02%		ENCODE	ENCFF624IQY
thymus	262,972	255	0.10%		Eichler	
progenitor cell of endocrine pancreas	1,225,562	251	0.02%		ENCODE	ENCFF988RQM
GM12878	2,115,533	250	0.01%		ENCODE	ENCFF450VAU
heart left ventricle	1,285,467	238	0.02%		ENCODE	ENCFF429JUP
heart left ventricle	1,045,691	228	0.02%		ENCODE	ENCFF602MAI
astrocyte	808,252	221	0.03%		ENCODE	ENCFF316EZQ

neural crest cell	1,199,310	219	0.02%		ENCODE	ENCFF249GFH
right cardiac atrium	793,250	211	0.03%		ENCODE	ENCFF722JJS
embryo_1C	2,268,330	208	0.01%		SRA	SRR17180610
brain	1,023,256	195	0.02%		SRA	SRR3476690
left cardiac atrium	769,976	193	0.03%		ENCODE	ENCFF920VXE
heart left ventricle	2,661,413	193	0.01%		ENCODE	ENCFF537NCV
B cell	485,946	181	0.04%		SRA	SRR17817858
endodermal cell	1,686,632	181	0.01%		ENCODE	ENCFF761BFK
testis	606,370	168	0.03%		Eichler	SRR12544673
endodermal cell	1,768,473	166	0.01%		ENCODE	ENCFF054KCY
H1	650,593	165	0.03%		ENCODE	ENCFF853OFP
psoas muscle	744,415	164	0.02%		ENCODE	ENCFF630XEC
neural crest cell	916,165	158	0.02%		ENCODE	ENCFF026VEI
right cardiac atrium	767,088	156	0.02%		ENCODE	ENCFF242WRZ
astrocyte	469,341	147	0.03%		ENCODE	ENCFF474GEK
brain	227,451	143	0.06%	TFs	SRA	SRR5189652
fetal osteoblast	2,368,543	143	0.01%		SRA	SRR23347363
endodermal cell	1,599,737	143	0.01%		ENCODE	ENCFF296KQK
technical sample	1,797,310	132	0.01%		ENCODE	ENCFF885YGF
technical sample	1,762,773	127	0.01%		ENCODE	ENCFF822IZD
mucosa of descending colon	717,159	124	0.02%		ENCODE	ENCFF511AVQ
heart	293,103	124	0.04%		Eichler	
HFFc6	1,114,145	119	0.01%		ENCODE	ENCFF385QZZ
H1	1,661,789	117	0.01%		ENCODE	ENCFF143ICB

embryo_4C	1,770,869	112	0.01%		SRA	SRR17180613
Right ventricle myocardium inferior	683,430	112	0.02%		ENCODE	ENCFF738UZJ
HFFc6	875,196	112	0.01%		ENCODE	ENCFF728ITF
H1	1,300,557	111	0.01%		ENCODE	ENCFF339FMQ
technical sample	1,798,001	108	0.01%		ENCODE	ENCFF705IEA
WTC11	1,974,884	107	0.01%		ENCODE	ENCFF212HLP
HFFc6	1,105,673	106	0.01%		ENCODE	ENCFF288CJF
embryo_morula	2,191,914	102	0.00%		SRA	SRR17180616
fetal osteoblast	1,656,396	101	0.01%		SRA	SRR23347361
embryo_1C	1,445,342	92	0.01%		SRA	SRR17180609
GM23338	806,340	88	0.01%		ENCODE	ENCFF251CBB
excitatory neuron	710,632	86	0.01%		ENCODE	ENCFF919JFJ
WTC11	1,751,296	85	0.00%		ENCODE	ENCFF105WIJ
fetal osteoblast	1,229,849	82	0.01%		SRA	SRR23347362
testis	47,988	80	0.17%	ampliconic genes	SRA	SRR22838397
GM23338	872,056	80	0.01%		ENCODE	ENCFF954UFG
testis	56,239	67	0.12%	ampliconic genes	SRA	SRR22838398
excitatory neuron	766,795	66	0.01%		ENCODE	ENCFF982WKN
adrenal gland	647,596	66	0.01%		ENCODE	ENCFF912HPY
WTC11	2,407,636	65	0.00%		ENCODE	ENCFF003QZT
K562	1,353,805	60	0.00%		ENCODE	ENCFF763VZC
lung	107,510	60	0.06%		Eichler	
fetal osteoblast	1,129,785	58	0.01%		SRA	SRR23347358
K562	635,981	58	0.01%		ENCODE	ENCFF694INI

embryo_8C	656,375	57	0.01%		SRA	SRR17180614
brain organoid	144,293	57	0.04%		SRA	SRR13316194
H1	900,523	57	0.01%		ENCODE	ENCFF684YOO
fetal osteoblast	1,112,165	56	0.01%		SRA	SRR23347360
HepG2	1,156,940	56	0.00%		ENCODE	ENCFF589SMB
adrenal gland	851,962	52	0.01%		ENCODE	ENCFF211SQY
endothelial cell	764,154	50	0.01%		ENCODE	ENCFF595PPR
GM12878	676,362	50	0.01%		ENCODE	ENCFF329AYV
HepG2	662,230	46	0.01%		ENCODE	ENCFF427JDY
fetal osteoblast	787,041	45	0.01%		SRA	SRR23347367
embryo_blastocyst	318,009	44	0.01%		SRA	SRR17180615
H1	239,905	44	0.02%		ENCODE	ENCFF400BQQ
GM12878	679,205	44	0.01%		ENCODE	ENCFF281TNJ
GM12878	784,800	43	0.01%		ENCODE	ENCFF475ORL
fetal osteoblast	824,888	41	0.00%		SRA	SRR23347365
human_merged_TD01673	222,157	37	0.02%		SRA	SRR17180617
endothelial cell	647,304	37	0.01%		ENCODE	ENCFF770DXN
GM12878	739,248	36	0.00%		ENCODE	ENCFF902UIT
right lobe of liver	556,003	36	0.01%		ENCODE	ENCFF306ZPP
fetal osteoblast	894,144	35	0.00%		SRA	SRR23347359
fetal osteoblast	798,867	34	0.00%		SRA	SRR23347366
embryo_4C	595,911	33	0.01%		SRA	SRR17180612
fetal osteoblast	879,251	31	0.00%		SRA	SRR23347368
fetal brain	448,133	31	0.01%		Eichler	SRR12524789

left lung	751,831	31	0.00%		ENCODE	ENCFF733RRO
right lobe of liver	679,523	31	0.00%		ENCODE	ENCFF318SKH
ovary	400,930	31	0.01%		ENCODE	ENCFF187BTK
fetal osteoblast	867,126	30	0.00%		SRA	SRR23347364
brain organoid	99,845	30	0.03%		SRA	SRR13316197
left lung	777,514	27	0.00%		ENCODE	ENCFF196WMM
CHM13	201,580	25	0.01%		Eichler	
brain organoid	60,952	21	0.03%		SRA	SRR13316196
embryo_1C	287,802	19	0.01%		SRA	SRR17180608
heart left ventricle	530,796	19	0.00%		ENCODE	ENCFF185VYD
brain organoid	26,336	17	0.06%		SRA	SRR13316199
brain organoid	49,386	15	0.03%		SRA	SRR13316195
brain organoid	27,588	14	0.05%		SRA	SRR13316198
embryo_2C	348,180	13	0.00%		SRA	SRR17180611
ovary	13,576	13	0.10%		Eichler	
testis	21,625	11	0.05%	ampliconic genes	SRA	SRR22838406
testis	25,386	6	0.02%	ampliconic genes	SRA	SRR22838405
fetal brain_hyb	93,801	4	0.00%	Human-specific duplication panel	Eichler	SRX9120454
Dorsal Root Ganglion (HSD2hyb)	47,089	2	0.00%	Human-specific duplication panel	Eichler	SRX9120450
testis_hyb	120,660	2	0.00%	Human-specific duplication panel	Eichler	SRX9120455
GM10539 (HSD2hyb)	49,025	2	0.00%	Human-specific	Eichler	SRX9120452

				duplication panel		
Esophagus (HSD2hyb)	43,898	1	0.00%	Human-specific duplication panel	Eichler	SRX9120451
muscle, brain, testes	39,145	-	0.00%		SRA	SRR5009600
muscle, brain, testes	12,624	-	0.00%		SRA	SRR5009592
muscle, brain, testes	22,869	-	0.00%		SRA	SRR5009590
testis_hyb	65,663	-	0.00%	Human-specific duplication panel	Eichler	SRX9120455
fetal brain_hyb	46,349	-	0.00%	Human-specific duplication panel	Eichler	SRX9120454
Skin (HSD2hyb)	14,045	-	0.00%	Human-specific duplication panel	Eichler	SRX9120453

Table S2. Genome assemblies.

Sample	Haplotype	Superpopulation	N50	*NPIP* paralogs	*NPIP* contigs	Project
CHM1	1	EUR	110,525,406	26	3	CHM1
GRCh38	1	EUR	145,138,636	31	3	GRCh38
HG00171	1	EUR	140,737,330	25	6	HGSVC
HG00171	2	EUR	135,792,130	27	2	HGSVC
HG00268	1	EUR	146,208,399	29	2	HGSVC
HG00268	2	EUR	135,181,456	28	2	HGSVC
HG00358	1	EUR	135,607,687	26	2	HGSVC
HG00358	2	EUR	134,783,754	25	3	HGSVC
HG00732	1	AMR	134,653,171	26	4	HGSVC
HG00732	2	AMR	134,386,536	26	3	HGSVC
HG00733	1	AMR	135,320,614	26	2	HGSVC
HG00733	2	AMR	136,679,475	27	2	HGSVC
HG01114	1	AMR	133,607,674	27	3	HGSVC
HG01114	2	AMR	135,125,707	26	2	HGSVC
HG01352	1	AMR	142,592,769	25	3	HGSVC
HG01352	2	AMR	136,151,478	26	3	HGSVC
HG01457	1	AMR	137,866,935	27	2	HGSVC
HG01457	2	AMR	133,362,811	26	2	HGSVC
HG01505	1	EUR	134,101,868	28	2	HGSVC
HG01505	2	EUR	125,504,060	27	3	HGSVC
HG01573	1	AMR	136,589,329	25	2	HGSVC
HG01573	2	AMR	136,571,735	28	2	HGSVC
HG02018	1	EAS	136,673,151	23	3	HGSVC
HG02018	2	EAS	135,348,401	26	3	HGSVC
HG02059	1	EAS	146,923,252	24	4	HGSVC
HG02059	2	EAS	152,425,962	24	2	HGSVC
HG02106	1	AMR	135,415,833	26	3	HGSVC
HG02106	2	AMR	134,516,900	26	3	HGSVC
HG02282	1	AFR	107,590,042	30	2	HGSVC
HG02282	2	AFR	125,692,346	28	3	HGSVC
HG02554	1	AFR	134,946,215	27	4	HGSVC
HG02554	2	AFR	107,135,070	32	5	HGSVC
HG02587	1	AFR	139,299,482	30	2	HGSVC
HG02587	2	AFR	136,083,123	28	2	HGSVC
HG02666	1	AFR	139,545,746	28	2	HGSVC
HG02666	2	AFR	137,190,139	26	3	HGSVC
HG02769	1	AFR	103,543,622	30	3	HGSVC
HG02769	2	AFR	134,418,817	27	3	HGSVC
HG02818	1	AFR	96,622,806	28	4	HGSVC
HG02818	2	AFR	104,520,653	28	5	HGSVC
HG02953	1	AFR	134,282,667	27	4	HGSVC
HG02953	2	AFR	109,062,915	28	2	HGSVC
HG03248	1	AFR	131,684,344	29	4	HGSVC
HG03248	2	AFR	135,822,003	30	4	HGSVC
HG03452	1	AFR	135,384,376	29	3	HGSVC
HG03452	2	AFR	121,296,586	23	3	HGSVC
HG03520	1	AFR	107,538,364	25	2	HGSVC
HG03520	2	AFR	132,474,249	27	2	HGSVC
HG03683	1	SAS	133,350,545	29	2	HGSVC
HG03683	2	SAS	154,166,241	27	2	HGSVC
HG03807	1	SAS	135,699,884	28	7	HGSVC
HG03807	2	SAS	135,340,106	26	3	HGSVC

HG04217	1	SAS	147,183,591	27	3	HGSVC
HG04217	2	SAS	146,808,645	27	4	HGSVC
NA18989	1	EAS	135,576,480	26	3	HGSVC
NA18989	2	EAS	137,631,789	27	2	HGSVC
NA19036	1	AFR	134,461,585	25	4	HGSVC
NA19036	2	AFR	153,764,006	29	4	HGSVC
NA19129	1	AFR	135,541,088	28	5	HGSVC
NA19129	2	AFR	145,947,918	28	3	HGSVC
NA19317	1	AFR	146,928,310	27	3	HGSVC
NA19317	2	AFR	147,849,691	26	3	HGSVC
NA19331	1	AFR	103,278,421	27	2	HGSVC
NA19331	2	AFR	133,868,608	28	4	HGSVC
NA19347	1	AFR	134,978,426	29	4	HGSVC
NA19347	2	AFR	155,109,183	30	4	HGSVC
NA19384	1	AFR	140,534,053	27	2	HGSVC
NA19384	2	AFR	140,056,662	28	3	HGSVC
NA19434	1	AFR	107,748,930	26	6	HGSVC
NA19434	2	AFR	130,955,708	23	5	HGSVC
NA19836	1	AFR	135,159,888	29	2	HGSVC
NA19836	2	AFR	131,390,575	27	5	HGSVC
NA19983	1	AFR	134,070,787	27	3	HGSVC
NA19983	2	AFR	154,390,956	27	2	HGSVC
NA20355	1	AFR	135,831,886	27	2	HGSVC
NA20355	2	AFR	138,003,900	28	6	HGSVC
NA21487	1	AFR	115,066,146	28	3	HGSVC
NA21487	2	AFR	133,421,945	27	3	HGSVC
HG00438	1	EAS	48,061,544	28	6	HPRC
HG00438	2	EAS	54,936,949	28	7	HPRC
HG005	1	EAS	58,303,677	32	9	HPRC
HG005	2	EAS	69,736,411	26	8	HPRC
HG00621	1	EAS	54,673,245	27	5	HPRC
HG00621	2	EAS	50,294,217	25	5	HPRC
HG00673	1	EAS	34,843,587	26	8	HPRC
HG00673	2	EAS	29,077,043	30	8	HPRC
HG00735	1	AMR	53,422,923	29	6	HPRC
HG00735	2	AMR	56,474,489	26	6	HPRC
HG00741	1	AMR	51,040,418	25	7	HPRC
HG00741	2	AMR	41,001,116	30	8	HPRC
HG01071	1	AMR	50,125,412	28	11	HPRC
HG01106	1	AMR	57,173,280	27	8	HPRC
HG01106	2	AMR	47,714,433	25	4	HPRC
HG01109	1	AMR	30,220,240	26	7	HPRC
HG01109	2	AMR	32,308,382	30	11	HPRC
HG01123	1	AMR	44,719,827	33	12	HPRC
HG01123	2	AMR	54,362,305	26	7	HPRC
HG01175	1	AMR	34,803,293	31	4	HPRC
HG01175	2	AMR	36,535,860	27	6	HPRC
HG01243	1	AMR	29,118,474	30	7	HPRC
HG01243	2	AMR	31,384,221	27	5	HPRC
HG01258	1	AMR	49,858,426	27	7	HPRC
HG01258	2	AMR	56,644,858	24	12	HPRC
HG01358	1	AMR	52,440,137	26	5	HPRC
HG01358	2	AMR	48,753,445	26	6	HPRC
HG01361	1	AMR	47,178,056	29	9	HPRC
HG01361	2	AMR	45,122,217	28	10	HPRC

HG01891	1	AFR	57,096,483	30	7	HPRC
HG01891	2	AFR	81,112,077	28	4	HPRC
HG01928	1	AMR	45,697,985	25	7	HPRC
HG01928	2	AMR	53,717,995	23	6	HPRC
HG01952	1	AMR	44,250,376	28	8	HPRC
HG01952	2	AMR	54,639,450	30	9	HPRC
HG01978	1	AMR	52,814,149	31	7	HPRC
HG01978	2	AMR	60,487,953	30	7	HPRC
HG02055	1	AFR	34,122,691	30	9	HPRC
HG02055	2	AFR	35,182,712	27	11	HPRC
HG02080	1	EAS	24,285,731	27	7	HPRC
HG02080	2	EAS	20,228,813	26	8	HPRC
HG02109	1	AFR	24,098,322	34	10	HPRC
HG02109	2	AFR	23,115,113	27	6	HPRC
HG02145	1	AFR	19,770,268	29	5	HPRC
HG02145	2	AFR	24,061,170	28	6	HPRC
HG02148	1	AMR	41,874,143	30	8	HPRC
HG02148	2	AMR	39,938,933	24	6	HPRC
HG02257	1	AFR	57,982,120	29	9	HPRC
HG02257	2	AFR	59,044,574	32	11	HPRC
HG02486	1	AFR	58,491,195	30	6	HPRC
HG02486	2	AFR	55,069,743	24	6	HPRC
HG02559	1	AFR	57,249,657	27	3	HPRC
HG02559	2	AFR	59,525,942	27	6	HPRC
HG02572	1	AFR	19,525,040	29	8	HPRC
HG02572	2	AFR	23,127,662	30	11	HPRC
HG02622	1	AFR	51,206,351	27	6	HPRC
HG02622	2	AFR	60,041,455	28	3	HPRC
HG02630	1	AFR	29,253,507	33	6	HPRC
HG02630	2	AFR	25,380,071	28	5	HPRC
HG02717	1	AFR	46,513,843	29	5	HPRC
HG02717	2	AFR	43,942,241	28	9	HPRC
HG02723	1	AFR	24,827,663	26	3	HPRC
HG02723	2	AFR	22,438,397	27	6	HPRC
HG02886	1	AFR	29,113,777	27	4	HPRC
HG02886	2	AFR	28,882,481	27	5	HPRC
HG03098	1	AFR	34,383,521	24	5	HPRC
HG03098	2	AFR	36,975,063	27	6	HPRC
HG03453	1	AFR	27,053,039	27	3	HPRC
HG03453	2	AFR	26,379,428	27	3	HPRC
HG03486	1	AFR	27,177,738	27	5	HPRC
HG03486	2	AFR	24,250,446	24	5	HPRC
HG03492	1	SAS	20,158,908	26	7	HPRC
HG03492	2	SAS	18,860,066	29	5	HPRC
HG03516	1	AFR	55,482,364	26	5	HPRC
HG03516	2	AFR	44,773,628	29	7	HPRC
HG03540	1	AFR	34,159,233	27	4	HPRC
HG03540	2	AFR	30,474,809	28	6	HPRC
HG03579	1	AFR	27,544,339	33	5	HPRC
HG03579	2	AFR	27,014,075	30	5	HPRC
NA18906	1	AFR	43,522,948	36	5	HPRC
NA18906	2	AFR	40,078,400	29	3	HPRC
NA19240	1	AFR	25,199,424	27	7	HPRC
NA19240	2	AFR	28,898,930	29	6	HPRC
NA20129	1	AFR	22,424,799	30	7	HPRC

NA20129	2	AFR	21,149,676	25	4	HPRC
NA21309	1	AFR	17,440,620	30	8	HPRC
NA21309	2	AFR	20,564,675	24	8	HPRC
CHM13	1	EUR	150,617,247	27	2	T2T-CHM13v2
HG002	1	EUR	146,793,688	28	2	T2T-HG002
HG002	2	EUR	154,349,815	27	2	T2T-HG002
PNG15	1	OCA	39,274,718	27	6	
PNG15	2	OCA	41,067,075	26	6	
PNG16	1	OCA	45,190,519	29	8	
PNG16	2	OCA	41,537,965	26	6	

Table S3. Biotinylated hybridization capture probes.

Target	length (nt)	Sequence
A1_exon3	100	GTGTCTTTCCTGAAGACTATCTTCCCGTCTCAAAATGGACATGATGGATCCACGGATGTACAGCAGAGAGCCAGGAGGTCCAACCGCCGTAGACAGGAAG
A1_exon5	108	TCTGAGGAAACTAAGCATGAAAGAACGTGAGCACGGAGAAAAGGAGAGGCAGGTGTCAGAGGCAGAGGAAAATGGGAAATTGGATATGAAAGAAATACACACCTACAT
A1_exon6	61	GGAAATGTTTCAACGTGCGCAAGCGTTGCGGCGGCGGGCAGAGGACTACTACAGATGCAAA
B5_exon2	129	GTTATCAATACTCTGGCTGACCATCATCATCGTGGGACTGACTTTGGTGGAAGTCCTTGGTTACATGTCATTATTGCGTTTCCGACAAGTTATAAAGTTGTCATTACCCTCTGGATAGTTTACCTTTGG
B14_exon2	120	GTTATCAATACTCTGGCTGACCATCGTCATCGTGAGACTGACTTTGGTGGAAGTCCTTGGATAATTATCATTATTGTGTTTCTGGGACGTTACAAATTTACCATTCTCTTCTGCACAATT
B6_exon2	140	GTTATCAATAGTCTGGCTGTCTATCGTCATCGTGAGACTGACTTTGGTGTAGGAGTTCGAGACCACCCTGGCCAACATGGCAAAACCCCATCTCCACAAAAATTGGATAATTTGATAATTATCATTATTGGGTTTCTGAG
B6_exon3	100	GTGTCTTTCCTGAAGACTATCTTCTGGTCTCGAAATGGACATGATGGATCCATGGATGTACAGCAGAGAGCCTGGAGGTCCAACCGCAGTAGACAGAAAG
B6_exon6-7	97	GAAAATGTTTCAACGTGCGCAAGAGTTGCGGCGGCGGGCAGAGGACTACCACAAATGCAAAATCCCCCCTTCTGCAAGAAAGCCTCTTTGCAACTGG

VITA

Philip C. Dishuck (born 1992) grew up in Tuscaloosa, AL, and attended Holy Spirit Catholic High School. He graduated with a B.S. in Biology from Washington & Lee University in Lexington, VA, in 2014, where he also studied circadian entrainment of the hypothalamic-pituitary-adrenal axis in the lab of Natalia Toporikova. He then worked as a research analyst and consultant at the Lewin Group in Falls Church, VA, analyzing healthcare data for federal agencies, until entering the PhD program at the University of Washington Department of Genome Sciences in 2017 and joining the lab of Evan Eichler in 2018.