Characterization of primate structural variation using diverse sequencing technologies

By

DANIELA CATALINA SOTO NEGRETE
DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Integrative Genetics and Genomics

in the

OFFICE OF GRADUATE STUDIES

of the

UNIVERSITY OF CALIFORNIA

DAVIS

Approved:

Megan Y. Dennis

C. Titus Brown

John D. McPherson

Committee in Charge

2022

A mis padres, Germán y Catalina

Characterization of primate structural variation using diverse sequencing technologies

Daniela C. Soto

University of California, Davis

Doctor of Philosophy

2022

ABSTRACT

Elucidating the genetic changes underlying the evolution of human traits remains an unfinished puzzle. Structural variants (SVs) account for more genetic differences than single-nucleotide polymorphisms between humans and our closest living relatives, chimpanzees, and are a hallmark of great ape evolution. The genomes of great apes are enriched in large interspersed segmental duplications (SDs), defined as duplications larger than 1 kbp with over 90% sequence identity, that sensitize the genome to further genomic rearrangements, including copy-number variation, via non-allelic homologous recombination. Despite their relevance, the identification and characterization of these SVs has been hindered by short reads lengths as they lack enough sequence context to discover breakpoints and cannot unequivocally be mapped to highly identical duplicates. Long-read sequencing technologies overcome these limitations by providing reads thousands of bases long, but the availability of population cohorts remains limited.

This thesis studies primate SVs and SDs characterized using diverse sequencing technologies and assesses their representation in reference genomes, variation across modern populations, their putative molecular impacts, and their roles in evolution and adaptation. We found novel SVs, including 88 deletions and 36 inversions, in two chimpanzee individuals sequenced with nanopore and optical mapping. Deletions and inversion breakpoints were depleted within topologically associated domains but enriched in differentially expressed genes between the two species. Focusing on human SDs, we identified eight Mbp of erroneously collapsed duplications in the human reference genome, impacting 48 protein coding and ten medically relevant genes, that are corrected in the first complete sequence of a human genome, T2T-CHM13. Leveraging this new reference, we identified 417 genes embedded in SDs with over 98% sequence identity (SD-98) that are near copy-number (CN) fixed in modern humans (1000 Genomes Project; 1KGP), 205 genes showing stratification between diverse modern populations (V_{ST}>95th percentile), and 22 protein-encoding

genes showing consistent Tajima's D outlier values across all humans examined. Our approach highlighted potential relevant human gene duplications, which are priority candidates for functional studies. Finally, we provide a compendium of tools and practices that we recommend be adopted by computational biologists to increase reproducibility in the field.

ACKNOWLEDGEMENTS

First, I would like to thank my Ph.D. advisor, Dr. Megan Dennis. She has been pivotal in my development as a scientist, not only because of her insightful and constructive scientific guidance, but also because she has supported me as a person. I am deeply thankful of her quick aid navigating the uncertainties associated with being an international scholar. I have a deep admiration for her commitment to scientific excellence and will strive to nurture these attributes in my future scientific career.

I am also thankful of my committee members, Dr. Titus Brown and Dr. John McPherson. Despite their busy schedules, both have been available to provide guidance when needed. Their perspectives have enriched my graduate school experience and have nicely complemented Dr. Dennis' mentoring.

I want to thank my friends and colleagues during graduate school. Moving countries was not easy, but you have made my experience far better. Thank you to all my lab mates that have made the Dennis Lab a supportive and welcoming environment. Thank you, Jose, for those *cafecitos* in Scrubs. Thanks to all my IGG cohort. Thank you, Julie, for organizing those amazing parties, and thank you, Kyle, for being our Dungeon Master. Thank you, Ellen and Noemie, for letting me pet your cats.

I want to acknowledge my grandparents, Enrique, Matilde, Guillermo, and Bedalia—the pillars of my big, loving Chilean family. I want to thank all my aunts and uncles, especially Miriam and Jaime, and my cousin Pablo, who helped me navigate paperwork-induced anxiety. Thank you, Jeff, for being my favorite coworker and cheerleader. You have made my life during this last year of graduate school so much brighter, fun, and wholesome. And thank you to my family: my dogs, my brother, Cristóbal, my sister-in-law, Daniela, and my parents, Germán and Catalina. Thank you, mom and dad, for your selfless love and for supporting my pursuing my dreams.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABBREVIATIONS

(Excluding gene names)

CNP: copy-number polymorphism

CNV: copy-number variant

HSD: human-specific duplication

IGC: interlocus gene conversion

LD: linkage disequilibrium

LRS: long-read sequencing

NAHR: non-allelic homologous recombination

ONT: Oxford Nanopore Technologies

SD: segmental duplication

SRS: short-read sequencing

SNP: single-nucleotide polymorphism

SNV: single-nucleotide variant

SV: structural variants

TAD: topologically associated domain

CHAPTER 1. Introduction

Elucidating the genetic changes underlying the evolution of human traits remains an unfinished puzzle. Genetic analyses have historically relied on single-nucleotide variants (SNVs) for the identification of species differences and selection signatures. Although complex genomic variation has long been recognized as a force underlying phenotypic diversity—e.g., transposable elements in maize (McClintock, 1931), and chromosomal inversions (Sturtevant, 1913) and duplications (Bridges, 1936) in *Drosophila*—as well as a key driver of primate evolution (Jeffrey A. Bailey & Eichler, 2006), methodological difficulties have limited the understanding of their functional and evolutionary impact. Scientists are now poised to explore this question at unprecedented resolution with the large-scale adoption of high-throughput sequencing technologies (Goodwin et al., 2016). Together, the widespread availability of high-quality reference genomes and population-level whole-genome sequencing datasets have reignited interest in studying the role of complex genomic variation in human/primate traits.

Broadly speaking, structural variants (SVs) are defined as complex genomic differences larger than 50 bp (Reviewed by Alkan, Coe, and Eichler 2011) (**Figure 1.1**). These include copy number variants (CNVs) that can change the dosage of a gene or genomic region and include deletions and duplications. Larger (>1 kbp) duplications with high sequence identity (>90%) are termed segmental duplications (SDs) or low-copy repeats (J. A. Bailey et al., 2001; Jeffrey A. Bailey et al., 2002). Other types of SVs include insertions, which can comprise novel sequence and mobile elements such as retrotransposons (for a comprehensive review see (Kazazian & Moran, 2017)), translocations, and inversions.

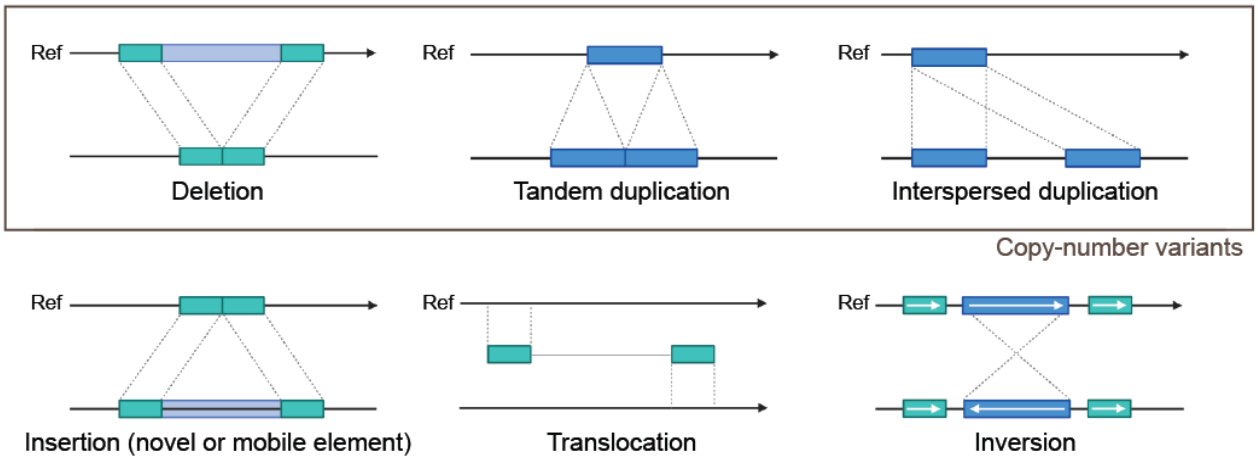


Figure 1.1. Examples of genomic structural variation. SVs exist as deletions and duplications (with the largest, most similar duplications termed segmental duplications, or SDs) that change the copy of a genomic segment (i.e., CNVs). Other types of SVs include insertions, translocations, inversions, as well as more complex events not pictured. Figure is adapted from (Alkan, Coe, et al., 2011) via "Genome Structural Variations" by BioRender.com (2022). Retrieved from https://app.biorender.com/biorender-templates.

1.1 The origins of primate structural variation

1.1.1 Hominid SDs and the "core" duplicon hypothesis

Comparative genomic analyses of great apes' species have shown their genomes to have been primarily shaped by SDs. The branch leading to African great apes (**Figure 1.2**), experienced a ~2.6-fold increase in duplication activity 8–12 million years ago (mya), concomitant with a clocklike rate of deletions and a decreased rate of single nucleotide variants (SNVs), chromosomal rearrangements, and retrotransposition activity (Tomas Marques-Bonet et al., 2009; Sudmant et al., 2013). As a consequence, hominid genomes are enriched for SDs contained in large interspersed blocks, differing from other sequenced mammals—like mice, dogs, and cows—where SDs are primarily organized in tandem (Liu et al., 2009; Nicholas et al., 2009; She et al., 2008). In humans, SDs account for 7% (218 Mbp) of the genome, according to the sequence of the first complete reference genome (Vollger, Guitart, et al., 2022).

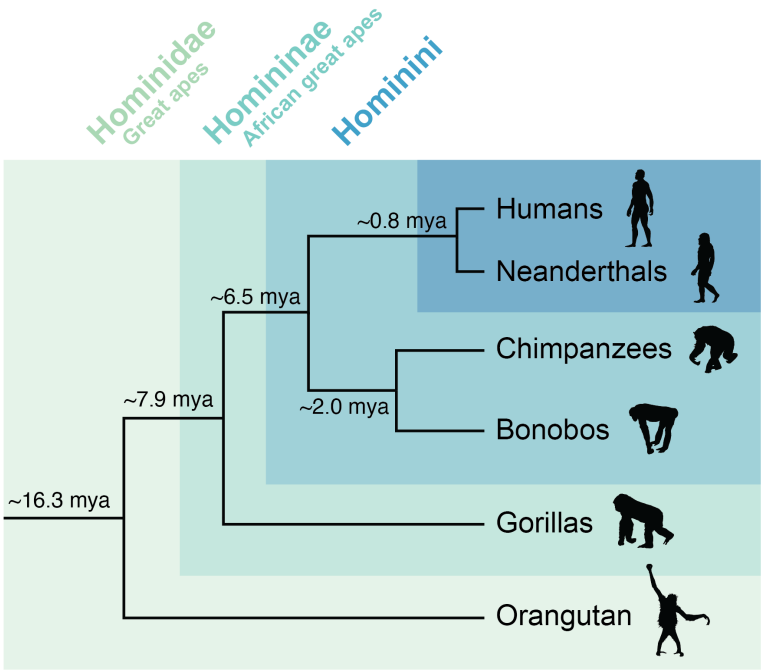


Figure 1.2. Cladogram of Hominidae family. Divergence time estimates were obtained from Sudmant et al. (Sudmant et al. 2013). Mya: million years ago.

Hominid SDs are non-randomly distributed across the genome and organized in large blocks (>250 kbp) that display a complex structure of duplications-within-duplications arranged around sequence elements known as 'core' or 'seed' duplicons (Dennis & Eichler, 2016; Jiang et al., 2007; T. Marques-Bonet & Eichler, 2009). These regions represent the focal point from which duplications accrue, with younger events located farther away from the core. In the human genome, hierarchical clustering of 437 duplicated blocks identified 24 core duplicons of ~15 kbp in size, of which fourteen were confined to one chromosome and ten were mixed in non-homologous chromosomes, mostly within subtelomeric and pericentromeric regions (Jiang et al., 2007). Evidence suggests that core duplicons have been reused independently and recurrently in different primate lineages (Cantsilieris et al., 2020; Matthew E. Johnson et al., 2006).

The core duplicons themselves are enriched for transcribed genes. Human core duplicon gene families exhibit signatures of positive selection (*NBPF*, *RGPD*, *PMS2P*, *SPATA31*, *TRIM51*, *GOLGA8*, Morpheus [*NPIP*], *TBC1D3*) (M. E. Johnson et al., 2001; Lorente-Galdos et al., 2013) and are among the most copy-number polymorphic (CNP) gene families in the human genome (e.g. *SPATA31*, Morpheus [*NPIP*], and *LRRC37A*) (Redon et al., 2006; Sharp et al., 2005). Since their original discovery, only three of these gene families have been functionally characterized (*NBPF*, *TBC1D3*, and *SPATA31*), leaving the function of most core duplicon genes unknown (Bekpen & Tautz, 2019).

Different molecular mechanisms have been proposed as the origin of primate SDs. The enrichment of *Alu* short interspersed elements (SINE)—the most abundant interspersed repeats in the human genome—at the boundaries of interstitial (euchromatic) and pericentromeric SDs suggests *Alu*-mediated origins (Jeffrey A. Bailey & Eichler, 2006; Jeffrey A. Bailey et al., 2003). As such, it has been proposed that the primate-specific 'burst' of *Alu* retrotransposition activity that occurred 35–40 mya sensitized the ancestral primate genome to *Alu*-mediated recombination events, that later propelled duplication events via non-allelic homologous recombination (NAHR). In the case of the LCR16a core duplicon, which contains the rapidly-evolving primate-specific gene family Morpheus (*NPIP*), interchromosomal and intrachromosomal expansions have been linked to the hominid-specific retrotransposon SINE-R-VNTR-*Alu* (SVA) (Damert, 2022). It has been found, however, that the association with *Alu* elements significantly decreases for younger SDs and CNVs, implying a change in the molecular mechanisms underlying SD formation, with newer events driven by other repeat classes or different molecular mechanisms such as non-homologous end-joining (NHEJ) (Kim et al., 2008). Recent evidence also suggests that the core duplicons might be subjected to genome instability through the

formation of non-B-form DNA structures (G-quadruplex) that remain highly active in the genomes of modern humans (Heft et al., 2020).

1.1.2 Molecular mechanisms contributing to structural variation

As large stretches of homologous sequences provide a substrate for recombination, SDs sensitize the genome to NAHR, resulting in genomic rearrangements such as unequal crossing-over and interlocus gene conversion (IGC), where a donor sequence overwrites an acceptor sequence (J. M. Chen et al., 2007). SDs, therefore, are often found in regions of genome instability, or 'hotspots', prone to recurrent genomic rearrangements, some of which have been associated with disease (Stankiewicz & Lupski, 2002, 2010). Core duplicons seem to be preferential sites for rearrangement hotspots (Dennis & Eichler, 2016). This suggests that human evolution has balanced the advantages conferred by duplication—a well-established driver of gene innovation (Ohno, 1970)—against genome instability and disease risk.

The rate and products of NAHR are determined by the characteristics of the impacted loci, including the size of the paralog sequence, degree of identity, distance, and orientation. However, >200 bp of sequence homology is required for efficient homologous recombination (J. M. Chen et al., 2007). CNVs (deletions, duplications) and balanced rearrangements (inversions, translocations) can be produced depending on the orientation of the paralog sequence. For example, NAHR between interchromosomal SDs would lead to deletion/duplication if the paralog sequences are directly oriented and inversions if they are invertedly oriented (Stankiewicz & Lupski, 2002) (**Figure 1.3**). Collectively, these NAHR-dependent rearrangements are known as SD-mediated structural variation.

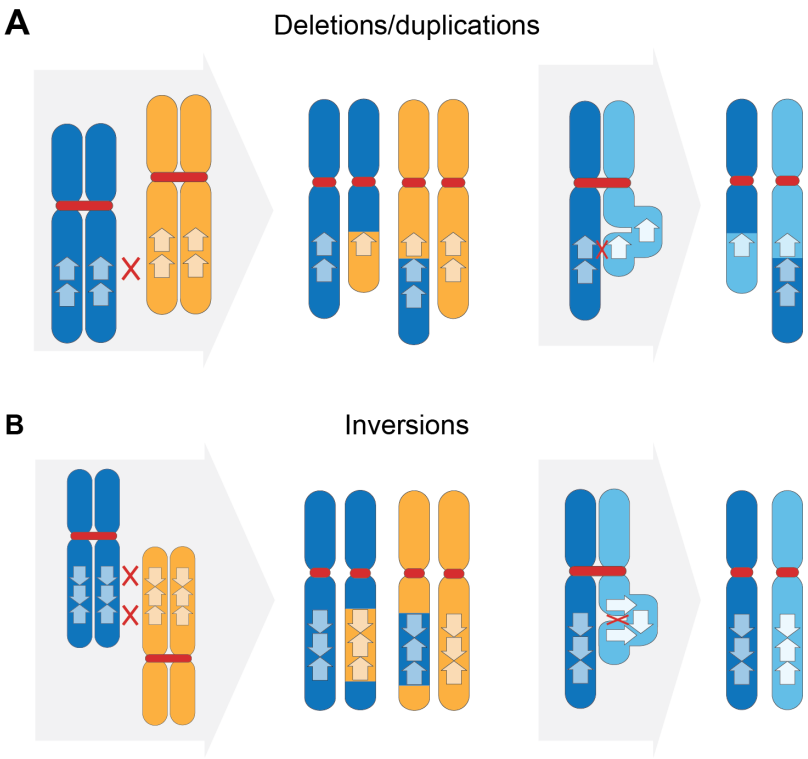


Figure 1.3. Products of meiotic NAHR. (A) Deletions/duplications mediated by directly oriented duplications in homologous chromosomes (left) or sister chromatids (right). (B) Inversions mediated by invertedly oriented duplications in homologous chromosomes (left) or intrachromatid (right).

However, SD-mediated SVs constitute only a portion of human SVs, with NAHR contribution estimated to range from 14 to 28% (Kidd et al., 2010; Korbel et al., 2007; Lam et al., 2010). Alternative DNA-repair mechanisms associated with SV formation include non-homologous end-joining (NHEJ), shrinking and expansion of variable number tandem repeats (VNTR), and retrotransposition. From these, NHEJ has been pointed to as the major source (~50%) of human SVs (Kidd et al., 2010; Korbel et al., 2007; Lam et al., 2010; Mills et al., 2011), followed by retrotransposition activity (~30%), mostly from L1 elements (Korbel et al., 2007).

SVs distribute non-randomly across the genome, clustering around 'SV hotspots', with most clusters containing variants originating from the same molecular mechanism (Ebert et al., 2021; Korbel et al., 2007; Mills et al., 2011; Sudmant, Rausch, et al., 2015). The location of SV clusters is, therefore, associated with the distribution of their source of origin, e.g., enrichment of VNTR near centromeric and pericentromeric regions and of NAHR near telomeres and SDs. Around 278 SV hotspots have been reported, preferentially located (~4-fold enrichment) near the 5 Mbp ends of chromosome arms (Ebert et al., 2021).

CNVs, in particular, have been shown to be enriched in SDs, suggesting that SD-based NAHR is a major contributor to CNV formation (Redon et al., 2006; Sharp et al., 2005). However, preferential mutational mechanisms differ among CNV types and sizes. Analyses of breakpoint signatures of 4,978 CNVs (>443 bp) showed that duplications are more likely to form from sequence-dependent mechanisms—such as NAHR, VNTR, and retrotransposition—than deletions. Although NAHR (13.5%) and VNTR (11.2%) contribute similarly to CNV formation, NAHR is primarily associated with CNVs in the largest decile and VNTR with CNVs in the smallest decile (Conrad et al., 2010).

A subgroup of CNVs known as microdeletions and microduplications have been implicated in certain diseases, collectively termed "genomic disorders" (Carvalho & Lupski, 2016; Inoue & Lupski, 2002; Stankiewicz & Lupski, 2002, 2010; Watson et al., 2014; Zhang et al., 2009). These CNVs are submicroscopic (not observable with common cytogenetic approaches) but usually large-scale (~0.1-1 Mbp), impacting multiple genes, and occur at hotspots of chromosomal rearrangements via NAHR (Inoue & Lupski, 2002). Genomic disorders range from Mendelian traits to contiguous gene syndromes. Several microdeletions and microduplications syndromes have been associated with autism, schizophrenia, and epilepsy (Dennis et al., 2017). Remarkable examples of disease-implicated hotspots include chromosomes 7q11.23 deletion (Williams-Beuren syndrome) and duplication (autism), 15q11–q13 deletion (Prader-Willi and Angelman syndromes), and 1q21.1 microdeletion (intellectual disability, schizophrenia) and duplication (autism).

1.2 Contribution of structural variation to hominid evolution and adaptation

1.2.1 Structural variation landscape across hominid genomes

The availability of high-coverage population-level short-read sequencing data provided by large-scale sequencing projects—such as the 1000 Genomes Project (1KGP) (Byrska-Bishop et al., 2022; Sudmant, Rausch, et al., 2015), the Human Genome Diversity Project (HGDP) (Almarri et al., 2020; Bergström et al., 2020), the Genome Aggregation Database (gnomAD) (Collins et al., 2020), and the UK BioBank (Halldorsson et al., 2022)—as well as the growing body of individuals sequenced with long reads of diverse backgrounds (Aganezov et al., 2022; Audano et al., 2019; Ebert et al., 2021), have allowed the comprehensive discovery of SVs in both human and primate genomes (Abel et al., 2020; Almarri et al., 2020; Audano et al., 2019; Byrska-Bishop et al., 2022; Collins et al., 2020; Ebert et al., 2021; Hehir-Kwa et al., 2016; Jakubosky et al., 2020; Sirén et al., 2021; Sudmant et al., 2013; Sudmant, Mallick, et al., 2015;

Sudmant, Rausch, et al., 2015; Yan et al., 2021) (**Table 1.1**). These surveys ratified SVs as a major source of genomic diversity within primate lineages and across human populations. Collectively, around 9% of the human genome is affected by insertions, deletions, and inversions alone (~279 Mbp) (Ebert et al., 2021), while at least 12% of the human genome (Redon et al., 2006) and ~16% of the hominid genome (Sudmant et al., 2013) is impacted by CNVs. Individually, each diploid genome harbors ~18.4 Mbp (0.6%) of SVs, accounting for more than five times as many affected base pairs as SNVs (~0.1%) (Sudmant, Rausch, et al., 2015). Per generation, at least 4.1 kbp are associated with de novo SV events, a 90-fold increase with respect to de novo SNVs (Kloosterman et al., 2015).

Table 1.1. Population cohorts of human structural variation obtained from whole-genome sequencing data.

Reference	Dataset	SV Discovery			SV Genotyping		
		Cohort	Population(s)	Platform	Cohort	Population(s)	Platform
(Sudmant, Mallick, et al., 2015)	-	236	125 populations	IL	-	-	-
(Sudmant, Rausch, et al., 2015)	1KGP (low-cov)	2,504	AFR, EUR, EAS, SAS, AMR	IL	-	-	-
(Hehir-Kwa et al., 2016)	GoNL	250	Dutch	IL	-	-	-
(Chiang et al., 2017)	GTEx	147	AFR, EUR, American Indian, Asian	IL	-	-	-
(Chaisson et al., 2019)	HGSVC	9	AFR, EAS, AMR	IL, PB, ONT, BNG	-	-	-
(Audano et al., 2019)	-	15	AFR, EUR, EAS, SAS, AMR	IL	440	AFR, EUR, EAS, SAS, AMR	Short reads (Illumina)
(Jakubosky et al., 2020)	i2QTL	719	AFR, EUR, EAS, SAS, AMR	IL	-	-	-
(Almarri et al., 2020)	HGDP	911	54 populations	IL	-	-	-
(Collins et al., 2020)	gnomAD	14,891	AFR, EUR, EAS, AMR	IL	-	-	-
(Abel et al., 2020)	CCDG	17,795	AFR, EUR, AMR	IL	-	-	-
(Ebert et al., 2021)	-	32	AFR, EUR, EAS, SAS, AMR	PB	3,202	AFR, EUR, EAS, SAS, AMR	Short reads (Illumina)
(Beyter et al., 2021)	-	3,622	Icelandics	ONT	-	-	-
(Yan et al., 2021)	-	-	-	-	2,504	AFR, EUR, EAS, SAS, AMR	Short reads (Illumina)
(Sirén et al., 2021)	-	-	-	-	5,202	AFR, EUR, EAS, SAS, AMR, MESA	Short reads (Illumina)
(Aganezov et al., 2022)	-	17	AFR, EUR, EAS, SAS, AMR	PB, ONT	-	-	-
(Byrska-Bishop et al., 2022)	1KGP (high-cov)	3,202	AFR, EUR, EAS, SAS, AMR	IL	-	-	-
(Halldorsson et al., 2022)	UK BioBank	150,119	British Irish, AFR, SAS	IL	-	-	-

1KGP: 1000 Genome Project. HGDP: Human Genome Diversity Project. gnomAD: Genome Aggregation Database. i2QTL: Integrated iPSC QTL. GoNL: Genome of the Netherlands Project. GTEx: Genotype-Tissue Expression Project. MESA: Multi-Ethnic Study of Atherosclerosis. AFR: African. EUR: European, EAS: East Asian. SAS: South

Asian. AMR: American. IL: Illumina short reads. PB: PacBio long-reads. ONT: Oxford Nanopore Technologies long reads. BNG: Bionano Genomics.

Albeit differences in distribution and affected sequence, SVs and SNVs share population genetic properties and global distribution patterns. Frequency-wise, most variants are rare and those with higher allele frequencies are shared among the five human continental groups. All SV classes can broadly recapitulate SNV-derived ancestries (Sudmant, Rausch, et al., 2015), including CNVs (Jakobsson et al., 2008). In concordance with SNVs, individuals of African ancestry exhibit more heterozygous SVs than other populations (Sudmant, Rausch, et al., 2015).

SVs and surrounding SNVs also display similar distributions of linkage disequilibrium (LD) (Hinds et al., 2006; Devin P. Locke et al., 2006; McCarroll et al., 2006), suggesting a shared evolutionary history. This has enabled appraising SVs via tagging SNVs (Beyter et al., 2021; Conrad et al., 2010; Hehir-Kwa et al., 2016; Marie Saitou et al., 2021; Yan et al., 2021). CNVs within duplicated-rich regions, however, show a weaker correlation with surrounding SNVs than those situated in less complex regions (Devin P. Locke et al., 2006; Sudmant, Mallick, et al., 2015), likely due to methodological difficulties in SNV detection within large duplications as well as haplotype-disruptive recurrence and interlocus gene conversion (Marie Saitou & Gokcumen, 2019b) (**Figure 1.4**).

Multicopy CNVs (mCNVs), also known as multiallelic CNVs, are particularly challenging for LD analyses, as the duplicated segment might not exist at the same locus of origin. Whole-genome shotgun sequencing-based approaches estimate that 73% of CNVs (>1% allele frequency) are in medium to strong LD ($r^2 > 0.6$) with nearby SNVs (Sudmant, Mallick, et al., 2015), while microarray-based approaches estimate that 40% of common mCNVs are in LD with nearby SNVs (Campbell et al., 2011). Similarly, NAHR-derived inversions also display a lack of LD with surrounding SNVs (Giner-Delgado et al., 2019). Considering that most genome-wide association studies and population genetics analyses depend on the underlying LD architecture of the genome, the lack of linkage information has hindered genotype–phenotype studies and selection scans (Marie Saitou & Gokcumen, 2019b).

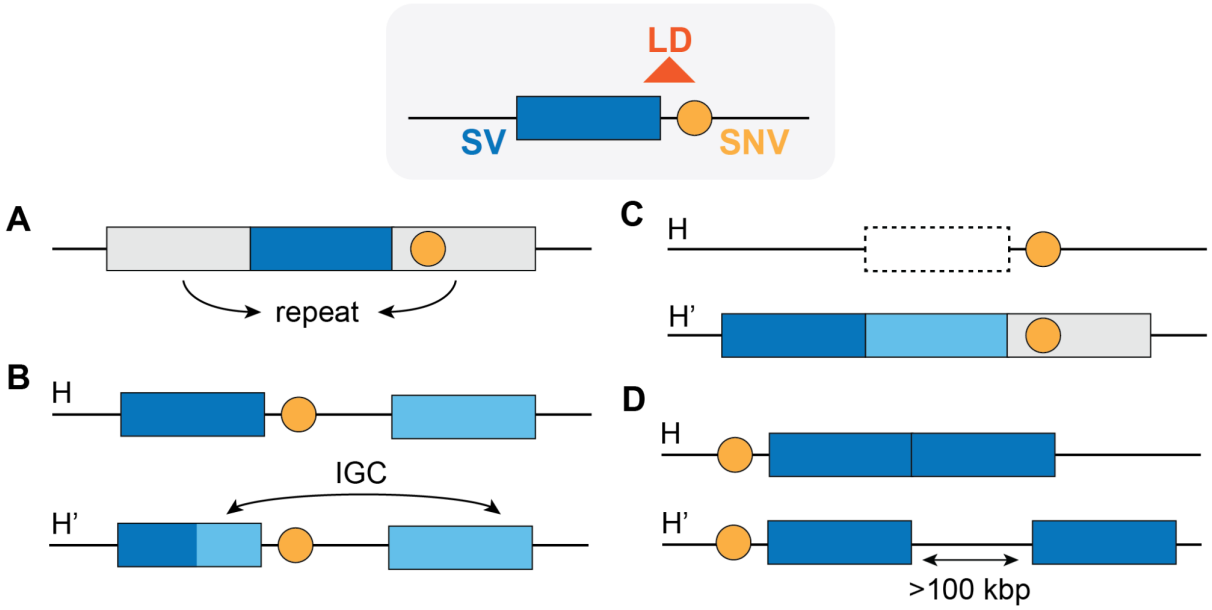


Figure 1.4. Difficulties appraising haplotypes between SVs and neighboring SNVs. (A) Neighboring SNVs are difficult to detect when an SV is embedded in repeat-rich regions. (B, C) Haplotypes can be disrupted by (B) interlocus gene conversion (IGC) and (C) recurrent deletions (H) and duplication (H'). (D) Multicopy CNVs can be in the same locus (H) or several kilobases apart (H').

Despite methodological difficulties, the function and disease implication of some SVs have been inferred based on strong LD with surrounding disease-implicated SNVs or performing association tests with phenotype cohorts (Aguirre et al., 2019; Beyter et al., 2021). Linkage information, in particular, shows that SVs are 1.5 times more likely to be in strong LD with genome-wide association study (GWAS) hits than SNVs (Sudmant, Rausch, et al., 2015). Similarly, SVs are ~50 times more likely than SNVs to be the lead cause of eQTL signals, with large SVs having larger effect sizes. Estimates based on expression data from 613 individuals from the GTEx project predict that common SVs are causal of 2.66% of eQTLs, which represents a 10.5-fold enrichment compared to SNVs, considering their relative abundance in the genome (Scott et al., 2021). Per genome, SVs are predicted to account for 17.2% of strongly deleterious variants, with rare SVs being 841 times more likely to be deleterious than rare SNVs (Abel et al., 2020). Thus, although less abundant, SVs disproportionately impact function.

1.2.2 Natural selection of human structural variation

Over evolutionary timescales, SVs are subjected to stronger selective pressures than SNVs. The majority of SV hotspots develop in gene-poor regions, evolving under relaxed negative selection or neutrality (Lin & Gokcumen, 2019). For example, it has been proposed that relaxation of negative selection allowed for extensive gain in copy

number of olfactory genes in the primate lineage, with negligible fitness implications for modern humans (Young et al., 2008). Conversely, functionally relevant sites—including coding regions, regulatory elements (enhancers, promoters), and topologically associated domains (TAD) boundaries—have been found to be both depleted in SVs and enriched in rare SVs (Beyter et al., 2021), a signature consistent with purifying selection. Among CNVs, deletions show stronger selective pressures than duplications, as they can severely impact the function by fully or partially ablating transcripts, regulatory elements, and TAD boundaries. Consequently, deletions are significantly depleted within functional elements in humans (Devin P. Locke et al., 2006; Mills et al., 2011) and other non-human primates (Fudenberg & Pollard, 2019; Soto et al., 2020).

Nonetheless, several examples of adaptive SVs under positive or balancing selection have been described in the literature, mostly implicated in local adaptation to dietary changes, environmental changes (*e.g.*, pigmentation, thermoregulation, xenobiotic), and resistance to infectious diseases (Edward J. Hollox et al., 2022; Marie Saitou & Gokcumen, 2019b) (**Table 1.2**). Positive selection of copy number gains in mCNVs has been associated with gene dosage effect (Handsaker et al., 2015). This is the case of the β-defensin genes, where copy number gains lead to greater protein expression on the mucosal surface and higher antimicrobial activity (E. J. Hollox et al., 2003). Other immune-related loci rich in common CNVs, such as the major histocompatibility complex, are thought to maintain their genetic diversity through the action of balancing and diversifying selection (Lin & Gokcumen, 2019).

Some deletions have been maintained polymorphic by the action of balancing selection for thousands of years (Aqil et al., 2022), even before the divergence of modern humans and Neanderthals estimated ~800 kya (Gómez-Robles, 2019). A well-known example of this phenomenon is a common 32 kbp deletion impacting genes *LCE3B* and *LCE3C* associated with psoriasis. This deletion emerged in the common ancestor with Neanderthals and was maintained through balancing selection, likely due to increased effectiveness of the acquired immunity system, albeit higher susceptibility to autoimmune disorders (Pajic et al., 2016). Interestingly, genes *GSTM1* and *UGT2B17* are polymorphic in humans and chimpanzees, suggesting inter-species balancing selection. However, further analyses revealed that *GSTM1* deletion evolved recurrently in both lineages (M. Saitou, Satta, & Gokcumen, 2018; M. Saitou, Satta, Gokcumen, et al., 2018), while the evolutionary history of *UGT2B17* remains unknown.

Table 1.2. Examples of large-scale SVs and whole-gene CNVs exhibiting signatures of natural selection in human populations.

Gene/locus	Region	Variant	Selection type	Category	Putative trait	References
GSTM1	1p13.3	Deletion	Positive (East Asian)	Metabolism	Xenobiotic metabolism	(M. Saitou, Satta, & Gokcumen, 2018)
Amylase (*AMY1* / AMY2)	1p21.1	mCNV	Positive	Diet	Adaptation to high-starch diet	(Pajic et al., 2019)
LCEB, *LCEC*	1q21.3	Deletion	Balancing	Immune response / Pigmentation	Psoriasis / Natural vaccination	(Pajic et al., 2016)
UGT2B17	4q13.2	Deletion	Balancing (European); Positive (East Asian)	Metabolism	Xenobiotic metabolism	(Xue et al., 2008)
Glycophorin (*GYPA* / *GYPB* / *GYPE*)	4q31.2	Complex duplication (*GYPB-GYPA* gene fusion)	Positive (East African)	Immune response	Resistance to malaria infection	(Leffler et al., 2017)
TCAF1 / *TCAF2*	7q35	Non-duplicated haplogroup	Positive (Archaics)	Diet / Thermoregulation	Unknown	(Hsieh et al., 2021)
ORM1	9132	"Runaway" duplication	Positive (European)	Immune response	Unknown	(Handsaker et al., 2015)
HERC2	15q13.1	Duplication	Negative (European)	Pigmentation	Unknown	(Marie Saitou & Gokcumen, 2019a)
BOLA2	16p11.2	mCNV	Positive	Diet	Protection against iron deficiency	(Giannuzzi et al., 2019)
α-Globin (*HBA1*/*HBA2*)	16p13.3	Deletion	Balancing (East African)	Immune response	Resistance to malaria infection	(Williams et al., 2005)
HPR	16q22.2	"Runaway" duplication	Positive (African)	Immune response	Resistance to trypanosomiasis infection	(Handsaker et al., 2015; Hardwick et al., 2014)
KANSL1	17q21.31	Inversion, duplication	Positive (European)	Fecundity	Increased fertility	(Stefansson et al., 2005)
SIGLEC14 / *SIGLEC5*	19q13.41	Deletion (gene fusion)	Positive	Immune response	Reduced risk of chronic obstructive pulmonary disease	(Angata et al., 2013; Yamanaka et al., 2009)
GSTT1 / *GSTT1P1*	22q11.23	Deletion (gene fusion)	Balancing (African)	Diet	Xenobiotic metabolism	(Lin et al., 2015)
APOBEC3B	22q13.1	Deletion	Positive	Immune response	Unknown	(Kidd et al., 2007)

An example of a positively selected inversion is the 17q21.31 900-kbp inversion polymorphism. The locus harbors two main distinct haplogroups, H1 (direct) and H2 (inverted), with little evidence of recombination for the last ~3 million years (Stefansson et al., 2005). The H2 haplogroup, although rare in Africans and Asians, is prevalent among Europeans (~20%), indicative of positive selection thought to confer increased fertility in females (Stefansson et al.,

2005). Both H1 and H2 have evolved independently and experienced complex rearrangements, with recurrent partial duplications of *KANSL1* (Steinberg et al., 2012), a haploinsufficient gene identified as the main genetic cause of the 17q21.31 microdeletion syndrome (also known as Koolen-De Vries syndrome) (Moreno-Igoa et al., 2015).

SVs involved in local adaptation—the genetic changes experienced by a population to adapt to local environmental conditions (Rees et al., 2020)—are prime targets of positive selection. The identification of most adaptive SVs has relied on genome-wide scans of population stratification (Conrad et al., 2010; Redon et al., 2006; Marie Saitou et al., 2021; Sudmant et al., 2010; Yan et al., 2021), as allele frequency differences between populations are robust to haplotype-disruptive recurrence and IGC. Stratified SNVs are frequently identified using the fixation index (F_{ST}). For mCNVs, the statistic V_{ST} (Redon et al., 2006) has been adapted from F_{ST} to account for multiple copy numbers. One of the most well-studied adaptive CNVs in humans is the amylase genes, involved in starch digestion in mammals. The copy number of the salivary amylase gene, *AMY1*, has been found to be positively correlated with dietary starch consumption in humans (Perry et al., 2007) and several starch-consuming mammals such as dogs (Pajic et al., 2019), evidencing positive selection. Although *AMY1* copy number has a dosage effect on salivary amylase production but accounts for a small portion of the variability observed among individuals (Carpenter et al., 2017). Interestingly, some adaptive SVs in out-of-Africa populations have been introgressed from archaic genomes (Hsieh et al., 2019; Yan et al., 2021). Among Melanesians, for example, 19 positively selected CNVs at chromosomes 16p11.2 and 8p21.3 have been likely introgressed from Denisovans and Neanderthals, respectively (Hsieh et al., 2019).

Some adaptive CNVs display a unique expansion pattern, where unusually high copy numbers are seen in one population, remaining low in the rest, a pattern termed ‘runaway duplications’ (Almarri et al., 2020; Handsaker et al., 2015). This is the case of *HPR*, encoding the haptoglobin-related protein which confers defense against trypanosome infection, which shows a copy-number increase in African populations consistent with the geographic distribution of the infection (Almarri et al., 2020; Handsaker et al., 2015; Hardwick et al., 2014). Other identified runaway duplications include the expansion of *ORM1* in Europeans (Handsaker et al., 2015), a private expansion downstream of *TNFRSF1B* in the Biaka group, an expansion upstream of the olfactory receptor *OR7D2* in East Asians, and expansions in medically relevant genes *HCAR2* in the Kalash group and *SULT1A1* in Oceanians (Almarri et al., 2020).

1.2.3 Complex variation and the evolution of uniquely human traits

While population-specific changes in allele frequency indicate adaptive variation, SVs found exclusively in humans are candidates for evolutionary relevant changes underlying uniquely human traits (Majesta O'Bleness et al., 2012). Gene loss caused by fixation of lineage-specific deletions has been proposed as a common and rapid local adaptation mechanism, often associated with immune response and pathogen resistance (Olson, 1999). Recent surveys of great ape genomes have identified 13.54 Mbp of human fixed deletions, containing 86 putative gene losses, 40 of which were human-specific, including known lost genes *SIGLEC13* and *CLECM4* (Sudmant et al., 2013). Conversely, human-specific segmental duplications (HSDs)—large duplication events (>1 kbp) that originated after the split between humans and chimpanzees from a common ancestor dated ~6 mya (Besenbacher et al., 2019)—and human-specific expansions (HSEs)—great ape gene duplications that reached higher copy numbers uniquely in humans—have also been pointed to as prime targets for the evolution of uniquely human traits. Direct comparisons of human and chimpanzee genomes show that HSDs are a major source of genetic differences between our species, impacting more than twice as many base pairs (~2.7%) than SNVs (~1.2%) (Z. Cheng et al., 2005). Thirty-three gene families have been identified within the largest HSDs regions (>20 kbp), several of which overlap known hotspots of genomic rearrangements associated with autism, schizophrenia, and epilepsy (Dennis et al., 2017). On average, HSDs and HSEs display higher sequence identity than most duplicated genes (97.0% and 98.7%), consistent with diverging time estimates between humans and chimpanzees (Sudmant et al., 2010).

Several HSDs and HSEs genes have been associated with brain development or neurodevelopmental disorders. Ancestral paralogs GPRIN2 (L. T. Chen et al., 1999) and SRGAP2 (Guerrier et al., 2009) have been implicated in neurite outgrowth and branching. Human-specific paralogs SRGAP2C (Charrier et al., 2012; Dennis et al., 2012), ARHGAP11B (Florio et al., 2016), and human-specific expansions NOTCH2NL (Fiddes et al., 2018, 2019; Suzuki et al., 2018) and TBC1D3 (Ju et al., 2016) affect neurodevelopment in animal models and may have contributed to neocortex expansion in the human lineage (**Table 1.3**). Human-specific HYDIN2 gene emerged from an incomplete duplication of ancestral HYDIN, likely adopting a new promoter that increased its expression in neural tissue (Dougherty et al., 2017) and has been associated with micro and macrocephaly (Brunetti-Pierri et al., 2008).

Table 1.3. Human-specific duplicated and expanded genes with evidence of functional impact in human brain evolution.

Gene	Region	Model	Functional impact	References (initial)
SRGAP2C	1q21.1	Embryonic mouse cortex	Neoteny during spine maturation	(Charrier et al., 2012; Dennis et al., 2012)
ARHGAP11B	15q13.3	Embryonic mouse cortex	Basal progenitor amplification	(Florio et al., 2015)
NOTCH2NL	1q21.1	Embryonic mouse cortex	Increase in neural progenitor proliferation and delayed neurogenesis	(Fiddes et al., 2018; Florio et al., 2018; Suzuki et al., 2018)
TBC1D3	17q12	Embryonic mouse cortex	Expansion of basal progenitors and cortical folding	(Ju et al., 2016)

However, not all duplicated paralogs retain or acquire a function (known as neofunctionalization), but instead undergo subfunctionalization or pseudogenization. Most duplicated genes, after a brief period of functional redundancy and relaxed selection, will accrue deleterious mutations and go the road of pseudogenization (Lynch & Conery, 2000). Relocation of duplicated copies in divergent epigenetic contexts and expression patterns might save them from pseudogenization and foster neofunctionalization (Rodin et al., 2005). A comparison of cross-tissue expression data from 75 HSD genes between humans and chimpanzees found that human-specific paralogs broadly exhibit patterns consistent with both relaxed selection and neofunctionalization (Shew et al., 2021).

Pseudogenes can cause disease by exchanging deleterious variants with functional paralogs via IGC. This is the case of *SMN2*, a nonfunctional HSD paralog of *SMN1*, which encodes the survival motor neuron (SMN) protein involved in the maintenance of motor neurons (Rochette et al., 2001). Unidirectional variant exchange via IGC causes *SMN2* to "overwrite" functional *SMN1* leading to the most common form of spinal muscular atrophy (Larson et al., 2015). Conversely, IGC events can also "rescue" non-functional HSD paralogs from pseudogenization, which was the case of *NOTCH2NL* (Fiddes et al., 2018; Suzuki et al., 2018).

These examples showcase that a subset of human-specific genes plays a relevant role in human neurodevelopment, evolution, and disease. However, the function and disease implication of most human duplicated genes remains to be assayed.

1.3 Challenges and opportunities in complex variation characterization

1.3.1 Short-read sequencing technologies underperform in complex regions

The study of complex variation has faced several methodological challenges caused by the complex architecture of primate SDs. Structural variants and duplicated regions have been historically difficult to assay using SRS technologies, the most widely available sequencing technology with thousands of whole-genome DNA samples sequenced in the public domain (Abel et al., 2020; Bergström et al., 2020; Byrska-Bishop et al., 2022; Karczewski et al., 2020). Short read length (~50-300 bp) poses challenges for (i) the assembly of large repeats and SDs, (ii) read mapping to repeat-rich regions, (iii) resolving SVs, and (iv) phasing haplotypes (Alkan, Sajjadian, et al., 2011; Chaisson, Wilson, et al., 2015). Since the emergence of SRS technologies, *de novo* assemblies have suffered from gaps preferentially in nearly identical SDs, satellite DNA, and other repeat-rich regions (Chaisson, Wilson, et al., 2015; Treangen & Salzberg, 2011), in addition to AT- and GC-rich regions that suffer from low sequence coverage in sequencing by synthesis approaches (Goodwin et al., 2016). SDs, in particular, tend to be either collapsed (missing copies) or misassembled (distinct paralogs assembled as a unique locus) (Eichler, 2001). Errors in the representation of SDs in reference genomes give origin to false positive heterozygous calls that confound downstream genetic analyses and lead to departure from Hardy-Weinberg equilibrium (Aganezov et al., 2022) (**Figure 1.5**).

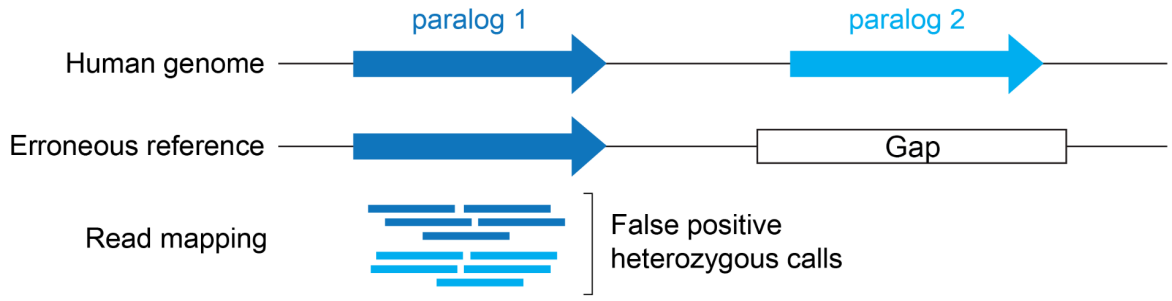


Figure 1.5. False positive heterozygous calls originated from missing copies in the reference. Bars represent a sample's short reads coming from the paralog present in the reference (dark blue) and the missing paralog (light blue).

However, when SDs are represented correctly, they are consistently tricky to assay using SRS technologies, as ambiguous mapping of reads from duplicated sequences prevents identifying true variation. These regions have been termed unmappable, inaccessible, "dark" or "camouflaged" (Ebbert et al., 2019) (**Figure 1.6**). HSD genes are particularly challenging as ancestral genes and their human-specific duplicate counterparts share on average ~99%

sequence identity and most of them exhibit copy-number polymorphisms in modern humans (Dennis et al., 2017), being ignored in most genetic analyses (Hartasánchez et al., 2018; Havrilla et al., 2019).

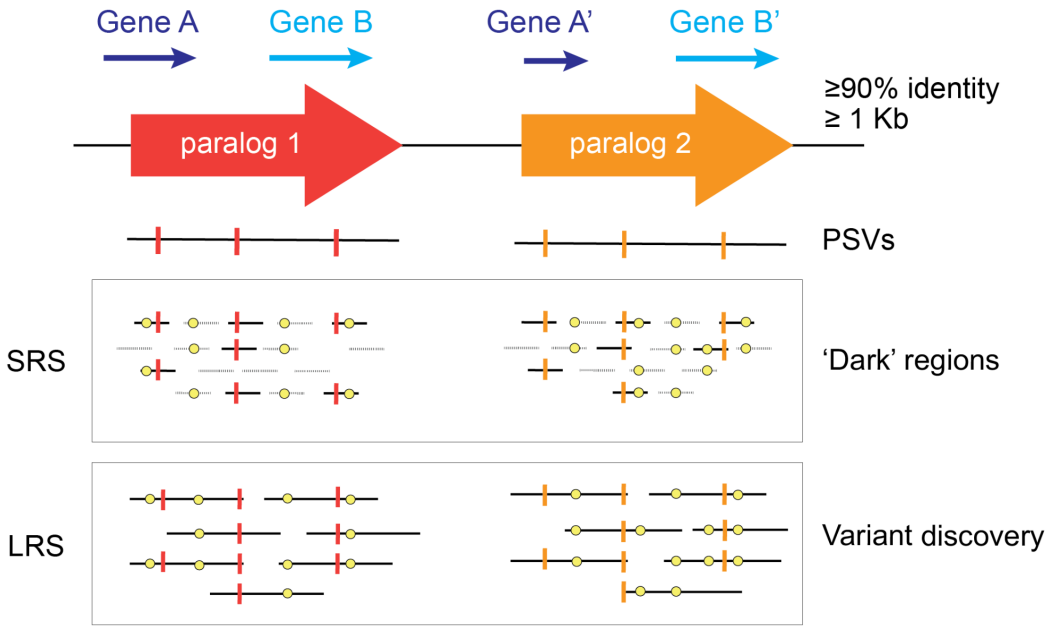


Figure 1.6. Differences in mappability between short and long reads in duplicated genes. Paralog-specific variants (PSVs) (vertical lines) enable discerning between paralogs and detecting polymorphic variation (yellow dots). Reads that do not carry PSVs (dashed lines) are unmappable in duplicated regions. SRS: short-read sequencing. LRS: long-read sequencing.

SRS technologies have also shown unequal ascertainment of SVs depending on each type. Deletions are often easier to detect, although not if they are embedded in SDs. Duplications and multicopy CNVs can be detected using read-depth signatures (Alkan, Coe, et al., 2011), but they can lack breakpoint resolution, location of the duplicated sequence, and paralog specificity (**Figure 1.7**). Non-reference unique insertions often go undetected in SRS samples (Almarri et al., 2020). Despite SRS challenges, to leverage the increasing amount of SRS databases, 'ensemble' algorithms, where a combination of tools is used to detect SVs (Ho et al., 2019), can discover thousands of SVs in large SRS databases (Abel et al., 2020; Almarri et al., 2020; Byrska-Bishop et al., 2022; Collins et al., 2020).

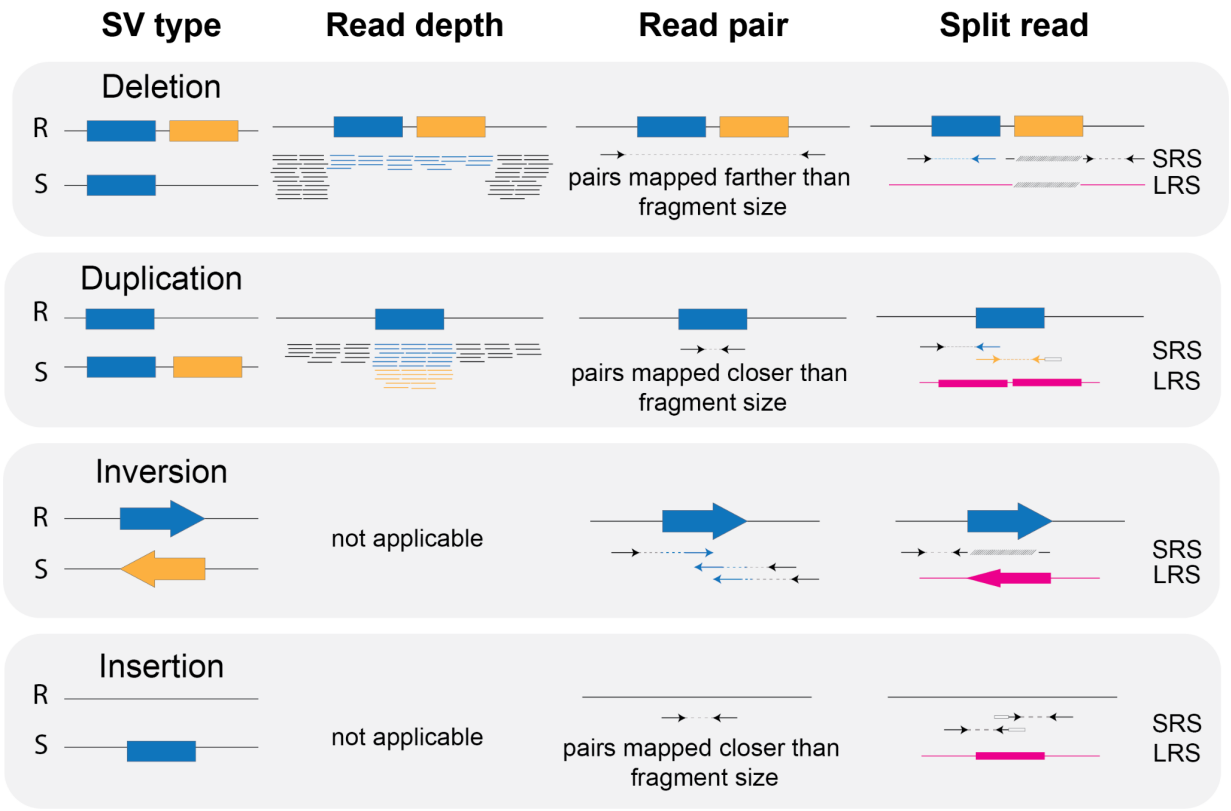


Figure 1.7. Short- and long-read SV discovery signals. R: Reference. S: Sample. SRS: short-read sequencing. LRS: long-read sequencing. Dashed line connects two pairs from the same short-read sequencing DNA fragment. Pink shapes represent long reads.

1.3.2 Long-read sequencing overcomes the limitations of short reads

In recent years, LRS technologies have overcome many of the limitations of SRS (Goodwin et al., 2016; Mantere et al., 2019; Sedlazeck, Lee, et al., 2018). Pacific Biosciences (PacBio) Single-Molecule Real-Time (SMRT) and Oxford Nanopore Technologies (ONT) can produce reads tens to hundreds of kilobases long, averaging ~10 kb. The first wave of LRS datasets enabled high-quality *de novo* assemblies of human individuals (M. Jain et al., 2018; Seo et al., 2016; Shafin et al., 2020; Shi et al., 2016; Wenger et al., 2019) and other non-human primates (Gordon et al., 2016; Kronenberg et al., 2018; Mao et al., 2021; Warren et al., 2020). Local assembly of the haploid human cell line CHM1 (Taillon-Miller et al., 1997) using bacterial artificial chromosome clones (CH17) has allowed the local reconstruction of misassembled regions of the human genome (Chaisson, Huddleston, et al., 2015; Huddleston et al., 2014; Vollger, Dishuck, et al., 2019; Vollger, Logsdon, et al., 2019).

A major achievement of LRS has been the completion of the first human genome sequence, T2T-CHM13 (Nurk et al., 2022). The new assembly finished the sequence of the missing 8% of the genome corresponding to repeat-rich regions including centromeres, telomeres, and the petite arms of the autosomal acrocentric chromosomes (13, 14, 15, 21, and 22). Additionally, T2T-CHM13 fixed euchromatic gaps and misassemblies, incorporating 51 Mbp of SDs (Vollger, Guitart, et al., 2022) and resolving ~8 Mbp of collapsed duplications compared to the previous reference genome, GRCh38, including previously missing HSD genes *GPRIN2B* and *DUSP22B* (Aganezov et al., 2022). One genome, however, is not enough to represent the full genetic diversity of modern humans, leading to the ongoing Human Pangenome Reference Consortium (Wang et al., 2022), which will deliver 350 diploid telomere-to-telomere human genomes in the next decade. The Vertebrate Genome Project is also leveraging LRS technologies to build high-quality reference genomes of over 66,000 extant vertebrate species (Rhie et al., 2021), enabling high-resolution comparative genomics.

LRS technologies have dramatically increased SV discovery (**Figure 1.7**). A combination of long-read and -range sequencing technologies—including PacBio, ONT, Illumina, 10X Genomics linked reads, Bionano Genomics optical mapping, Strand-Seq, and Hi-C—identified 27,622 SVs (≥50 bp) per genome, representing a 7-fold increase in SV discovery respect to SRS (Chaisson et al., 2019), a similar finding was obtained by ONT reads alone (22,636 SVs per genome) (Beyter et al., 2021). LRS population-level cohorts, ranging from dozens (PacBio) to thousands of individuals (ONT), have identified >100,000 SVs in modern humans (Audano et al., 2019; Beyter et al., 2021; Ebert et al., 2021; Nurk et al., 2022), which have been genotyped in SRS datasets to assess their functional and evolutionary impact (Yan et al., 2021). In addition to direct mapping approaches, LRS technologies are enabling diploid assemblies that better represent heterozygous SVs, and theoretically are the most comprehensive approach for SV discovery (Mahmoud et al., 2019).

Long reads have also shown improved mappability in “dark” regions of the human genome (Ebbert et al., 2019) (**Figure 1.6**). However, their original high-error rate (~10-15%) hindered their implementation in SNV and indel calling. Variant discovery using ONT MinION reads in a human individual (NA12878) yielded an overall accuracy of 91.40% (M. Jain et al., 2018). However, PacBio circular consensus sequencing (CCS) can yield high fidelity (HiFi) reads, averaging a base accuracy of 99.8% and variant calling precision and recall over 99.4% (Wenger et al., 2019), enabling routine discovery of SNVs and indels (insertions/deletions < 50 bp) in duplicated regions.

Together, the incoming influx of human and non-human primate genomes sequenced with LRS in combination with large-scale SRS datasets, are ushering in a new era in genomics, promising to fully unveil the functional and evolutionary impact of complex genomic variation in primate-specific traits and diseases.

1.4 Goals of this dissertation

The overall goal of this dissertation is to leverage both short- and long-read sequencing technologies to discover and characterize complex genomic variation in great apes—including SVs and SDs—and their evolutionary and functional impact. In chapter 2, I describe the identification of SVs in two new chimpanzee individuals, using a combination of two long-range technologies, Nanopore reads and Bionano optical mapping, integrated with Illumina short reads. After discovery, we assessed the functional impact of large-scale genomic variation in the gene sequence, gene expression, and genome organization using chromatin conformation capture (Hi-C). In chapter 3, I focus on my contribution to the Telomere-to-Telomere (T2T) Consortium, the international consortium that achieved the first ever complete sequence of a human genome. Here, we evaluated how a complete reference genome improves the analysis of genetic variation, including improvements in variant detection, and population and clinical genomics analyses. We systematically surveyed misrepresented duplicated sequences and showed how the new reference corrects duplication errors and removes erroneous variant calls that confound population and medical genetic analyses. In chapter 4, I describe the analysis of nearly-identical SDs using a complete human reference genome. Here, we identified genes within evolutionarily recent SDs, and analyzed their expression, copy number diversity, and population stratification, to identify priority candidates for functional studies. Finally, in chapter 5, I share our review of some of the main tools for computational biology research, suggesting a framework and a toolbox to conduct computational biology research, with the goal of promoting reproducibility and sustainability in computational biology research.

CHAPTER 2. Identification of structural variation in chimpanzees using optical mapping and nanopore sequencing

Chapter 2 is adapted with minimal modification from

Soto DC*, Shew C*, Mastoras M, Schmidt JM, Sahasrabudhe R, Kaya G, et al. Identification of Structural Variation in Chimpanzees Using Optical Mapping and Nanopore Sequencing. *Genes*. 2020;11: 276.

First authorship is shared between DCS and CS. DCS performed the structural variation identification, genotyping, filtering, comparison, and annotation of impacted genes.

2.1 Abstract

Recent efforts to comprehensively characterize great ape genetic diversity using short-read sequencing and single-nucleotide variants have led to important discoveries related to selection within species, demographic history, and lineage-specific traits. Structural variants (SVs), including deletions and inversions, comprise a larger proportion of genetic differences between and within species, making them an important yet understudied source of trait divergence. Here, we used a combination of long-read and -range sequencing approaches to characterize the structural variant landscape of two additional *Pan troglodytes verus* individuals, one of whom carries 13% admixture from *Pan troglodytes troglodytes*. We performed optical mapping of both individuals followed by nanopore sequencing of one individual. Filtering for larger variants (>10 kbp) and combined with genotyping of SVs using short-read data from the Great Ape Genome Project, we identified 425 deletions and 59 inversions, of which 88 and 36, respectively, were novel. Compared with gene expression in humans, we found a significant enrichment of chimpanzee genes with differential expression in lymphoblastoid cell lines and induced pluripotent stem cells, both within deletions and near inversion breakpoints. We examined chromatin-conformation maps from human and chimpanzee using these same cell types and observed alterations in genomic interactions at SV breakpoints. Finally, we focused on 56 genes impacted by SVs in >90% of chimpanzees and absent in humans and gorillas, which may contribute to chimpanzee-specific features. Sequencing a greater set of individuals from diverse subspecies will be critical to establish the complete landscape of genetic variation in chimpanzees.

2.2 Introduction

Great apes have considerable phenotypic diversity despite being closely related species. For humans and chimpanzees, with only ~5 to 9 million years of independent evolution (Langergraber et al., 2012; Patterson, Richter, et al., 2006), significant effort has gone into understanding the underlying genetic and molecular differences contributing to species differences, often with the primary focus on human-unique features (Majesta O'Bleness et al., 2012). Direct comparison of protein-encoding genes has identified exciting candidates, but these only account for a minor proportion of species differences (A. Varki, 2005). Recent analysis of Illumina short-read sequencing has allowed identification and genotyping of single-nucleotide variants (SNVs) at the genome scale, which have been used to address questions related to the demographic history and genetic adaptations of each species, and lineage-specific traits (Prado-Martinez et al., 2013). Further, transcriptome and epigenome comparisons of immortalized cell lines and tissues have revealed many thousands of individual genes and putative cis-acting regulatory elements that contribute to species differences in gene regulation (Brawand et al., 2011; Eres et al., 2019; Gallego Romero et al., 2015; Khan et al., 2013; McLean et al., 2011; Pollen et al., 2019; Prescott et al., 2015; Zhou et al., 2014), though often with varied results and reproducibility across studies.

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

Genomics (BNG) and Hi-C) of one individual, Clint, a male representing the subspecies *Pan troglodytes verus*, significantly increasing the number of known SVs (Kronenberg et al., 2018).

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

2.3 Results

2.3.1 Large-Scale SV Discovery and Genotyping in Chimpanzee

To date, one western chimpanzee individual (Clint) comprising the reference genome (panTro6) has been subject to hybrid long-read sequencing for genome assembly and SV discovery (Kronenberg et al., 2018). We sought to expand SV discovery via long-read sequencing to two additional chimpanzee individuals (AG18359 and S003641) for which renewable LCLs and functional genomic information, including RNA-Seq and ChIP-Seq data (Khan et al., 2013; McVicker et al., 2013; Zhou et al., 2014), are available. To begin, we performed Illumina short-read sequencing (~30× coverage) of both individuals to confirm ancestry via SNV detection followed by comparisons of population-specific genetic markers and PC analysis with chimpanzees from the GAGP (Prado-Martinez et al., 2013) (**Figure S2.1**). From this, we determined AG18359 to be a female western chimpanzee (*Pan troglodytes verus*) and S003641 to be a male western chimpanzee with some central chimpanzee ancestry (*Pan troglodytes verus* × *Pan troglodytes troglodytes*). Notably, ~13% of the ancestry of this individual is assigned to the central-chimpanzee population, similar to one individual (Donald) that was sequenced as part of the GAGP.

To discover potentially novel chimpanzee SVs, we assayed AG18359 gDNA using ONT PromethION (29×) and BNG optical mapping (116×) (**Table S2.1**). To compare SV discovery of two individuals on the same platform, we also

subjected S003641 to BNG optical mapping (70×). As it is the most accurate and well-annotated primate assembly, we mapped our sequence data to the human reference genome (GRCh38). We excluded SDs and insertions from our analysis of SVs due to challenges in their discovery and validation (Alkan, Coe, et al., 2011). Focusing exclusively on deletions and inversions, we discovered 49,579 deletions and 560 inversions using ONT and 4,790 deletions, and 280 inversions using BNG from AG18359. Similarly, we identified 5,407 deletions and 207 inversions using BNG from S003641. For comparative purposes, we also mapped the AG18359 ONT sequence data to the most recent chimpanzee reference genome (panTro6) and discovered fewer events (7,895 deletions and 142 inversions) suggesting that a significant proportion of SVs identified via mapping to the human reference represented species differences.

As the primary goal of our study was to identify species differences, we moved forward with SVs identified using the human reference genome. We next compared SV discovery across our two platforms. Although ONT had higher sensitivity to discover smaller variants, down to 50 bp, there was a higher chance of detecting false positives and errors at this resolution **(Figure S2.2A**). To properly compare across technologies, we filtered for large SVs (≥10 kbp) and compared similarities by consolidating variants with more than 50% reciprocal overlap. We found a comparable number of deletions in our three call sets (586, 586, and 666 events in AG18359 ONT, AG18359 BNG, and S003641 BNG, respectively) with 138 deletions found by all three call sets (**Table S2.3**, **Figure S2.2B**). Out of the 586 deletions found in the AG18359 ONT call set, 381 were uniquely discovered using this technology, while BNG contributed another 553 deletions, out of which 307 (55.5%) had support from both individuals. As such, deletion call sets from the same technology exhibited a greater overlap than comparing calls from different technologies of the same individual. We also found a comparable number of inversions across all three call sets (243, 269, and 207 variants in AG18359 ONT, AG18359 BNG, and S003641 BNG, respectively) (**Figure S2.2B**), of which 34 variants were shared among them all. Again, the most overlap for inversions was identified between different individuals assayed using the same BNG technology, representing 80 shared out of the total 274 unique variants.

In order to narrow in on a higher-confidence set of SVs, we subsequently performed genotyping of this discovery set using short-read Illumina data from GAGP (>20-fold coverage) of all four chimpanzee subspecies (n = 25) (**Table S2.2**) using SVTyper (Chiang et al., 2015). We also compared our discovered SVs with previously-reported datasets from three recent whole-genome SV screens of chimpanzees (Catacchio et al., 2018; Kronenberg et al., 2018; Sudmant et al., 2013), each using diverse genomic methods for discovery (**Table S2.5** and **Table S2.6**). From this, we identified

425 deletions and 59 inversions that had support from short-read genotyping and/or intersecting with a previously-discovered SV (**Table S2.7** and **Table S2.8**). In all, our discovery approach using ONT and BNG data achieved 88 novel deletions and 36 novel inversions when compared with the most recent genome-assembly alignment (Catacchio et al., 2018; Kronenberg et al., 2018) and read-depth (Sudmant et al., 2013) approaches (**Figure 2.1A** and **Figure 2.1B**).

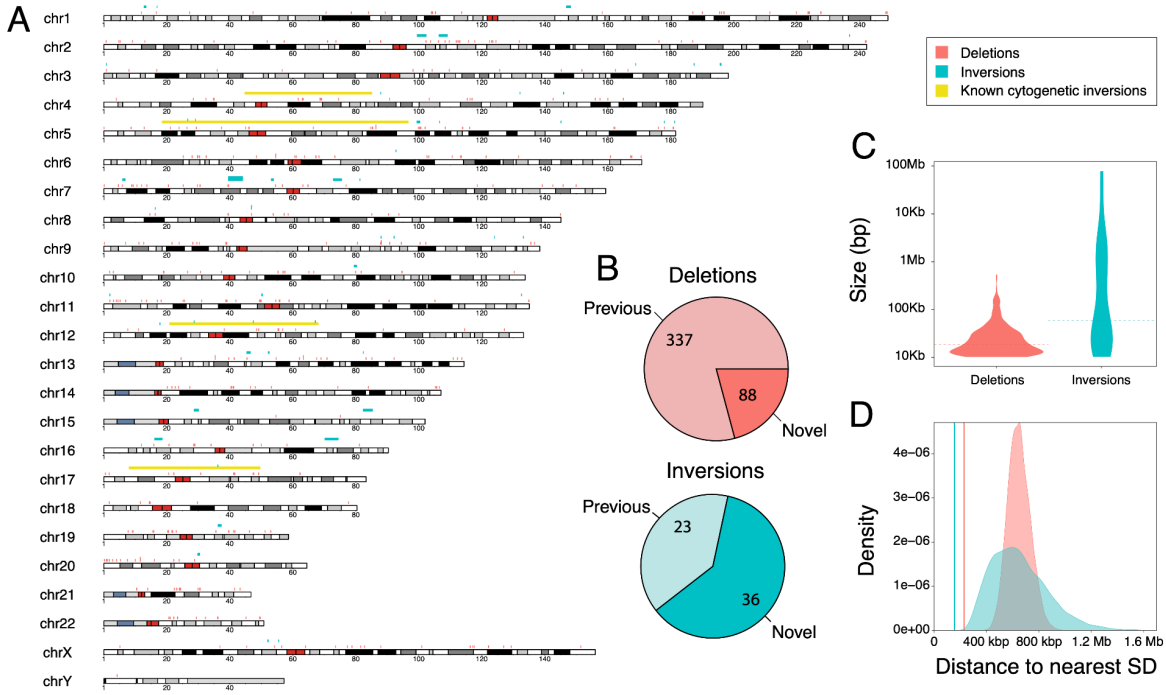


Figure 2.1. Genomic features of identified SVs. (A) Deletions (red), inversions (cyan), and large-scale cytogenetic inversions (yellow) are interspersed across all 24 human orthologous chromosomes, depicted as ideograms. (B) Novel variants in our dataset are defined as lacking 50% reciprocal overlap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

2.3.2 Genomic features of identified SVs

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

Sawatzki, Szamalek, et al., 2005; Nickerson & Nelson, 1998; Shimada et al., 2005; Szamalek et al., 2006)–64]. The majority of novel inversions identified in our study tended to be smaller (57 kbp mean length), perhaps influenced by strict size cutoffs (>100 kbp) used in previous studies (Catacchio et al., 2018). The distribution of SVs across the human genome (**Figure 2.1A** and **Figure S2.3**) was relatively uniform for deletions, which were found on all 24 chromosomes. The greatest number of events were identified in chromosome 2 (n = 34); however, when normalizing by the total number of bases, chromosomes 19 (0.34 deletions per Mbp) and 21 (0.32 deletions per Mbp) exhibited the highest number of deletions (**Figure S2.3**). Inversions, on the other hand, were found on 19 chromosomes, with chromosome 5 exhibiting the greatest number of variants (n = 8), and chromosomes 5, 7, and 12 displaying the greatest number of inversions per chromosome size (0.04 inversions per Mb). Further, we found that SV breakpoints of both deletions and inversions were non-randomly distributed across the human genome near SDs (**Figure 2.1D**, empirical p-value = 1×10^{-4}), similar to previously reported results for distribution of SDs in primate genomes (Z. Cheng et al., 2005; Dennis et al., 2017; Tomas Marques-Bonet et al., 2009; Sudmant et al., 2013). This observed clustering may be accounted for by SD-mediated deletions and inversions that can be created via non-allelic homologous recombination (Carvalho & Lupski, 2016).

2.3.3 Genes impacted by SVs

To evaluate the functional impact of our high-confidence set of SVs, we retrieved all annotated transcribed features within deletions (±2.5 kbp) and at inversion breakpoints (±50 kbp) (Tables S2.9 and S2.10). Deletions overlapped with 592 genes, out of which 162 were protein-encoding genes (**Figure 2.2A**). To further refine the impact of SVs and gene function, we focused on protein-encoding genes and used Ensembl Variant Effect Predictor (VEP) to predict functional impact. VEP annotated 80 protein-encoding genes as highly impacted by deletions (i.e., feature ablation or truncation), out of which 54 have been previously classified as loss of function (LoF) tolerant (pLI ≤ 0.1) by the Exome Aggregation Consortium (Lek et al., 2016; Samocha et al., 2014) (**Figure 2.2B**). Also, three genes (ATXN2L, SH2B1, and IL27), which all reside within the same ~500 kbp "deletion" mapped to human chromosome 16p11.2, were classified as LoF intolerant (pLI ≥ 0.9). A search through the chimpanzee reference (panTro6) found ATXN21 and SH2B1 residing on an uncharacterized chimpanzee chromosome Un_NW_019937196v1, suggesting that these genes have been translocated to a new genomic locus. This is likely the case for other genes with predicted high-variant effect and LoF intolerance. Focusing on inversions, we found breakpoints overlapping with 342 transcribed

elements of which 64 genes were within 2.5 kbp of breakpoints, including 95 and 21 protein-encoding genes, respectively (**Figure 2.2A**). No highly impacted genes, as predicted by VEP, were found in this dataset. Using pLI scores, we identified 9 genes either modified or overlapped by inversions classified as loss-of-function intolerant in humans (**Figure 2.2B**).

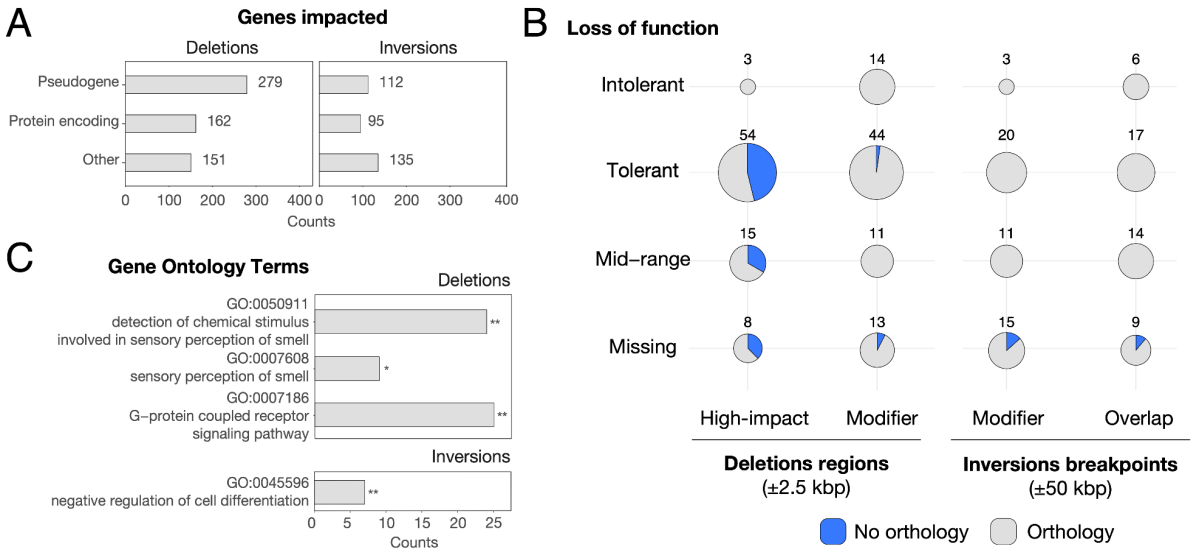


Figure 2.2. Description of genes overlapping identified SVs. (A) Categories of genes overlapping deletion regions ±2.5 kbp and inversion breakpoints ±50 kbp as defined by ENSEMBL biotypes. (B) Number of protein-encoding genes classified as LoF tolerant (pLI ≤ 0.1), intolerant (pLI ≥ 0.9) and middle range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

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

Finally, we explored functional annotations of genes impacted by SVs. We found 208 transcribed elements impacted by deletions with known GO annotations as reported by DAVID (Huang et al., 2009a, 2009b) (**Figure 2.2C**). Compared to the complete set of human GO annotations, this gene list displays an overrepresentation of genes associated with sensory perception of smell (GO:0050911, q-value = 8.7×10^{-11} and GO:0007608, q-value = 3.3×10^{-2}). We also found an overrepresentation of deletion-impacted genes involved in the G-protein coupled receptor signaling pathway (GO:0007186, q-value = 5×10^{-5}). Notably, both ontologies are primarily driven by known copy-number polymorphism that exists among olfactory-receptor genes (Nozawa et al., 2007). Inversions contained 140 genes with known GO functional annotation exhibiting an overrepresentation of regulation of cell differentiation (GO: 0045596, q-value = 1.2×10^{-4}).

2.3.4 SVs and Gene Regulation

To understand if variants might affect gene regulation, we leveraged existing RNA-seq datasets generated from chimpanzee and human LCLs (Khan et al., 2013) and iPSCs (Pavlovic et al., 2018). From 55,461 human–chimpanzee orthologous transcribed features, we identified 6,565 and 8,946 genes in LCLs and iPSCs, respectively, as significantly DE between the two species (**Table S2.11** and **Table S2.12**). Among genes for which human-chimpanzee orthology was assigned that directly intersected SVs (N = 397 in deletions ±2.5 kb; N = 61 for inversion breakpoints ±2.5 kb), roughly half were significantly DE (57/135 LCL and 60/129 iPSC tested genes in deletions; 25/37 LCL and 22/36 iPSC tested genes in inversion breakpoints) (**Table S2.9** and **Table S2.10**). We report a significant enrichment of DE genes from both cell types within (±2.5 kb; permutation test empirical p-value < 0.04) and near (±50 kb; p-value < 0.01) deletions and near (±50 kbp; p-value < 0.002) inversion breakpoints. DE gene enrichment was only significant within (±2.5 kbp) inversion breakpoints in LCLs (**Figure 2.3** and **Figure S2.4**).

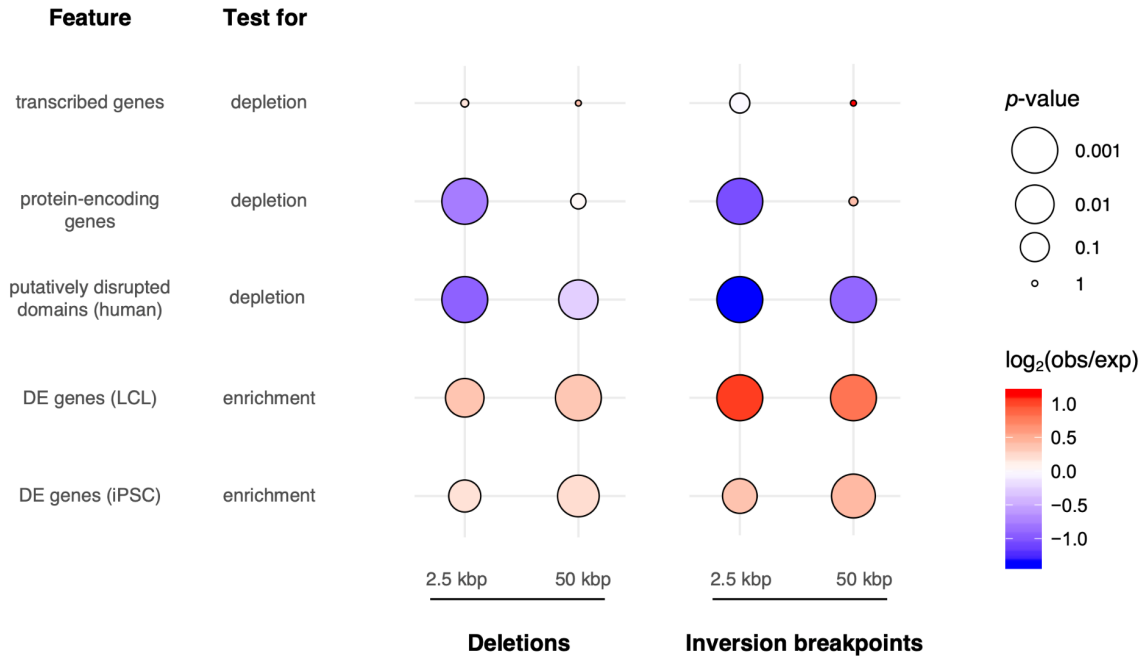


Figure 2.3. Enrichment and depletion tests of SVs with genomic features. Both deletions and duplications were tested within 2.5 kbp (resolution of the SV calls) and 50 kbp. All annotated genes (GENCODE v27) and protein-encoding genes were tested for depletion of SVs (top two rows) via permutation testing.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

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

in **Figure 2.4A**, **Figure S2.5**, **Figure S2.7**, and **Figure S2.8**. Notably, chromatin structure was also apparently altered by variants near but not directly intersecting identified TAD boundaries (**Figure 2.4B** and **Figure S2.6**).

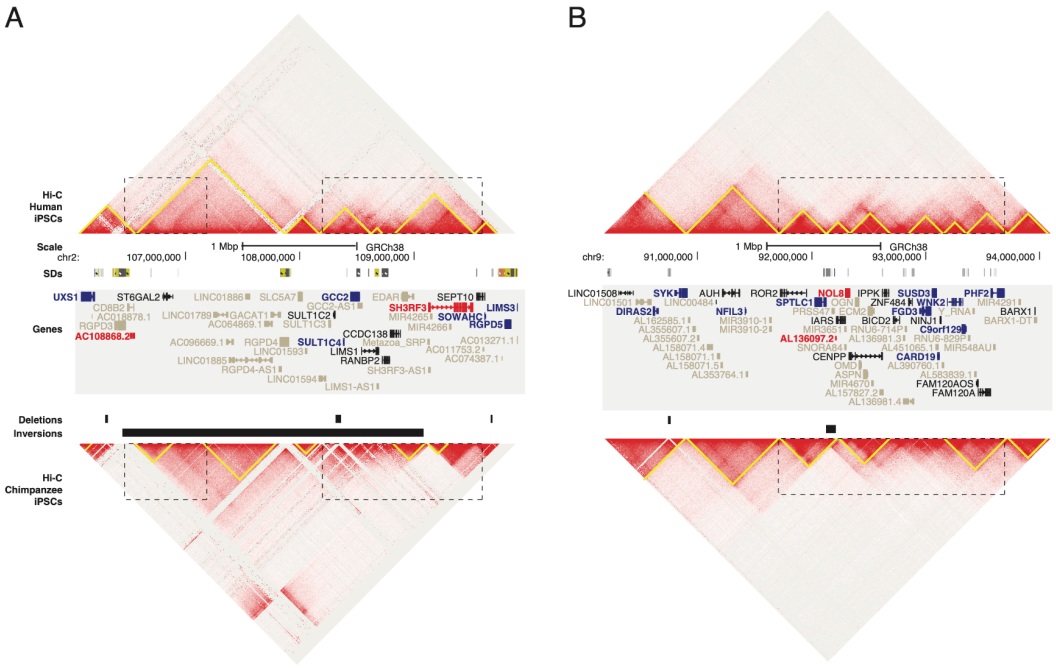


Figure 2.4. Genome organization of human and chimpanzee across regions with identified SVs. The Hi-C genomic landscape of human (top) and chimpanzee (bottom) are depicted for iPSCs using Juicebox for (**A**) chromosome 2q12.2–q13 (chr2:106,095,001-109,905,000, GRCh38) and (**B**) chromosome 9q22.2–q22.32 (chr9:90,200,001-94,010,000, GRCh38). Predicted TADs (yellow triangles) were compared between species, noting differences at SVs (dotted boxes) including deletions and inversions. SDs are depicted as colored bars, taken from the UCSC Genome Browser track. Genes showing significant DE in chimpanzee versus humans are colored as red (up in chimpanzee) or blue (down in chimpanzee). Genes not included in the DE analysis are in gray (Tables S2.11 and S2.12).

To examine chromatin structure of PDTs, we generated orthologous Hi-C maps from human and chimpanzee LCLs and iPSCs (Eres et al., 2019) against the human reference (GRCh38) and directly compared differences in domain boundaries between species. Overall, domain calls were similar between species (MoC 0.75 and 0.79 for LCLs and iPSCs, respectively (Zufferey et al., 2018)). We examined chimpanzee PDTs and identified more chimpanzee-unique boundaries than genome-wide boundaries (30.5% (18/59) versus 24.9% (1424/5714)). Similarly, for iPSCs we found 22.0% (13/59) of boundaries in PDTs were not shared with humans, compared to 14.9% genome-wide boundaries (868/5834). These numbers suggest that TAD-altering SVs may impact chromatin structure in chimpanzees.

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

2.3.5 Genes Showing Signatures of Natural Selection

Recent efforts to sequence diverse great ape genomes have led to identification of signatures of natural selection using SNV data that may help to explain features unique to chimpanzee species and subspecies (Cagan et al., 2016; de Manuel et al., 2016; Prado-Martinez et al., 2013; Schmidt et al., 2019). To understand if our identified SVs might impact the outcome of such studies or explain signatures of selection previously identified, we compared our map of SVs with a recent study of natural selection in multiple genomes of the four chimpanzee subspecies (*Pan troglodytes verus*, *troglodytes*, *ellioti*, and *schweinfurthii*) mapped to the human reference genome (Cagan et al., 2016). In this study, among several other tests, the Hudson–Kreitman–Aguade (HKA) test (Hudson et al., 1987) was used to identify the top 200 genes showing the strongest signatures of long-term balancing selection and positive selection in each subspecies. Intersecting this set of genes with our complete list of genes residing within or near deletions (**Table S2.9**), we determined that of the 592 genes putatively disrupted by a deletion, 54 show strong signatures of natural selection using the HKA test (32 for positive and 22 for balancing selection). For inversions, of the 342 genes at or near inversion breakpoints, six show strong signatures of natural selection (five for positive, one for balancing) (**Table S2.10**). Of all the genes affected by SVs and with strong signatures of natural selection, nine have evidence of DE in

either LCLs or iPSCs, including two protein-encoding genes showing signatures of balancing selection: *INPP4B*, which carries a deletion upstream of the transcription-start site and is upregulated in chimpanzee LCLs, and *HLA-F,* which is completely deleted and is upregulated in chimpanzee LCLs and downregulated in iPSCs. The possibility that these deletions generated beneficial expression changes that became strongly affected by natural selection makes these genes interesting candidates for follow up.

2.3.6 Genes Impacted by Chimpanzee-specific SVs

To hone in on SVs unique and universal to chimpanzees that may contribute to species-specific features, we consolidated the complete dataset of our newly discovered SVs and those previously published (Catacchio et al., 2018; Kronenberg et al., 2018; Sudmant et al., 2013). Filtering for only those with positive genotypes in >90% of chimpanzee individuals genotyped but found in neither humans (n = 8) nor gorillas (n = 8), we identified 209 deletions and 18 inversions. This set ranged in size from 10 kbp to 526 kbp for deletions and 12 kbp to 78 Mbp for inversions (including the four large-scale cytogenetic events). Again, due to the olfactory receptors at these loci, GO analysis shows that the genes contained within these SVs were overrepresented for the detection of chemical stimuli involved in sensory perception of smell (GO:0050911, q-value 4.1×10^{-2}). Focusing on genes with a higher likelihood of being functionally impacted by SVs, we identified 56 protein-encoding genes with a high-impact VEP score (deletions) or within 2.5 kbp of a breakpoint (inversions) (**Table 2.1**). Of the 35 genes queried in our cross-species RNA-seq comparisons, 13 exhibited significant DE in chimpanzee versus human in LCLs and/or iPSCs, including *APOL4*, *CAST*, *CLN3*, *EFCAB13*, *EIF3C*, *IL18R1*, *NPIPB8*, *NPIPB9*, *NUPR1*, *RABEP2*, *SGF29*, SLC01B3, and SULT1A1. Additionally, six genes showed strong signatures of positive selection (*APOBR*, *IL27*, and *TUFM* at human chromosome 16p11.2 and *OR10H1* and *OR10H5* at human chromosome 19p13.12) or balancing selection (CLC at human chromosome 19q13.2). In all, this list of genes represents exciting candidates putatively implicated in chimpanzee-specific traits.

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

Gene	ENSEMBL ID	SV type	Description
APOBR	**ENSG00000184730**	**deletion**	**Apolipoprotein B receptor**
APOL1	ENSG00000100342	deletion	Apolipoprotein L1
*APOL4**	ENSG00000100336	deletion	Apolipoprotein L4

ATP2A1	ENSG00000196296	deletion	Sarcoplasmic/endoplasmic reticulum calcium ATPase 1
ATXN2L	ENSG00000168488	deletion	Ataxin 2 like
CARD18	ENSG00000255501	deletion	Caspase recruitment domain family member 18
*CAST**	ENSG00000153113	inversion	Calpastatin
CD19	ENSG00000177455	deletion	CD19 Molecule
CEACAM21	ENSG00000007129	deletion	CEA Cell Adhesion Molecule 21
CFHR2	ENSG00000080910	deletion	Complement Factor H Related 2
CFHR4	ENSG00000134365	deletion	Complement Factor H Related 4
CLC	**ENSG00000105205**	**deletion**	**Charcot-Leyden crystal Galectin**
*CLN3**	ENSG00000188603	deletion	CLN3 Lysosomal/Endosomal Transmembrane Protein, Battenin
CMPK1	ENSG00000162368	deletion	Cytidine/Uridine Monophosphate Kinase 1
CROCC	ENSG00000058453	inversion	Ciliary Rootlet Coiled-Coil, Rootletin
CYP2C18	ENSG00000108242	deletion	Cytochrome P450 Family 2 Subfamily C Member 18
DEFB128	ENSG00000185982	deletion	Defensin Beta 128
*EFCAB13**	ENSG00000178852	deletion	EF-Hand Calcium Binding Domain 13
*EIF3C**	ENSG00000184110	deletion	Eukaryotic Translation Initiation Factor 3 Subunit C
*IL18R1**	ENSG00000115604	inversion	Interleukin 18 Receptor 1
IL1RL1	ENSG00000115602	inversion	Interleukin 1 Receptor Like 1
IL27	**ENSG00000197272**	**deletion**	**Interleukin 27**
IL36B	ENSG00000136696	deletion	Interleukin 36B
IL37	ENSG00000125571	deletion	Interleukin 37
KRTAP19-6	ENSG00000186925	deletion	Keratin Associated Protein 19-6
KRTAP19-7	ENSG00000244362	deletion	Keratin Associated Protein 19-7
LCN10	ENSG00000187922	deletion	Lipocalin 10
LCN6	ENSG00000267206	deletion	Lipocalin 6
LGALS14	ENSG00000006659	deletion	Galectin 14
MERTK	ENSG00000153208	deletion	MER Proto-Oncogene, Tyrosine Kinase
*NPIPB8**	ENSG00000255524	deletion	Nuclear Pore Complex Interacting Protein Family Member B8
*NPIPB9**	ENSG00000196993	deletion	Nuclear Pore Complex Interacting Protein Family Member B9
*NUPR1**	ENSG00000176046	deletion	Nuclear Protein 1, Transcriptional Regulator
OBP2A	ENSG00000122136	deletion	Odorant Binding Protein 2A
OR10H1	**ENSG00000186723**	**deletion**	**Olfactory Receptor Family 10 Subfamily H Member 1**
OR10H5	**ENSG00000172519**	**deletion**	**Olfactory Receptor Family 10 Subfamily H Member 5**
OR2T33	ENSG00000177212	deletion	Olfactory Receptor Family 2 Subfamily T Member 33
OR6C2	ENSG00000179695	deletion	Olfactory Receptor Family 6 Subfamily C Member 2
OR6C3	ENSG00000205329	deletion	Olfactory Receptor Family 6 Subfamily C Member 3
OR6C65	ENSG00000205328	deletion	Olfactory Receptor Family 6 Subfamily C Member 65
OR6C70	ENSG00000184954	deletion	Olfactory Receptor Family 6 Subfamily C Member 70
OR6C75	ENSG00000187857	deletion	Olfactory Receptor Family 6 Subfamily C Member 75
OR6C76	ENSG00000185821	deletion	Olfactory Receptor Family 6 Subfamily C Member 76

POU6F2	ENSG00000106536	deletion	POU Class 6 Homeobox 2
*RABEP2**	ENSG00000177548	deletion	Rabaptin, RAB GTPase Binding Effector Protein 2
RACK1	ENSG00000204628	inversion	Receptor For Activated C Kinase 1
*SGF29**	ENSG00000176476	deletion	SAGA Complex Associated Factor 29
SH2B1	ENSG00000178188	deletion	SH2B Adaptor Protein 1
SLC35G4	ENSG00000236396	deletion	Solute Carrier Family 35 Member G4
*SLCO1B3**	ENSG00000111700	inversion	Solute Carrier Organic Anion Transporter Family Member 1B3
*SULT1A1**	ENSG00000196502	deletion	Sulfotransferase Family 1A Member 1
SULT1A2	ENSG00000197165	deletion	Sulfotransferase Family 1A Member 2
TUFM	**ENSG00000178952**	**deletion**	**Tumor Protein P53**
YAE1D1	ENSG00000241127	deletion	YAE1 Maturation Factor Of ABCE1
AC011604.2	ENSG00000257046	inversion	Uncharacterized
AL355987.1	ENSG00000204003	deletion	Uncharacterized

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

2.4 Discussion

Most extensive SV analyses using comparative genomic approaches have used a single genome from one chimpanzee individual of the subspecies *Pan troglodytes verus* (i.e., Clint) (Catacchio et al., 2018; Chimpanzee Sequencing and Analysis Consortium, 2005; Feuk et al., 2005; Kronenberg et al., 2018; Tomas Marques-Bonet et al., 2009; Newman et al., 2005). Here, we performed long-read sequencing of two additional individuals of the same subspecies, one of which carried admixture with *Pan troglodytes troglodytes*, using two orthogonal technologies: optical mapping and nanopore sequencing. To our knowledge, this represents the first nanopore sequence of a chimpanzee genome. From this, we discovered over 60,000 deletions and over 500 inversions (≥50 bp) when compared with the human reference (GRCh38), on the same scale as found in a recent comparison of the new chimpanzee assembly using a hybrid assembly approach (panTro6) (Kronenberg et al., 2018). As expected, ONT sequencing was capable of detecting significantly more SVs, down to 50 bp with higher resolution at breakpoints (**Figure S2.2A**), compared to our BNG datasets. Many of the bioinformatically-identified SVs were redundant within and across technologies, which required additional filtering. To determine a higher-confidence set of SVs, we limited our analysis to variants ≥10 kbp in size with short-read Illumina sequencing evidence of the variant using SVtyper, a genotyping approach. Though the genotyping step significantly increased our confidence in variant calls, it also reduced the number of variants we

identified (from 1,838 to 858 deletions and from 719 to 253 inversions), particularly for inversions, which are difficult to detect/genotype using short-read data. Additionally, our strict size cutoff limited our ability to discover transposable elements, which has been shown to represent a significant proportion of lineage divergence between chimpanzees and humans (Yohn et al., 2005). Furthermore, due to the uncertainty of the BNG breakpoints, most SVs discovered using only this approach were largely filtered from our subsequent analyses due to an inability to accurately genotype events. Nevertheless, our approach led to the discovery of 88 novel deletions and 36 novel inversions when compared to recent genome-wide scans. We note that we also excluded SDs and insertions from our analysis due to difficulties in discovery and subsequent validations using standard short-read genotyping approaches (Chander et al., 2019). As improved hybrid-based methods combining long- and short-read data are developed to more accurately identify SVs and their breakpoints, it will be a worthwhile endeavor to return to our dataset to discover additional SVs.

Our results implicated chimpanzee SVs in potentially impacting gene regulation and chromatin organization. It has been established that TAD structures are evolutionarily conserved (Dixon et al., 2012; Rao et al., 2014), and recent work finds that deletions altering TAD boundaries in humans are under purifying selection (Fudenberg & Pollard, 2019; Huynh & Hormozdiari, 2019). TAD structure is also conserved across apes, as evidenced by the incidence of gibbon–human synteny breaks at domain boundaries (Lazar et al., 2018). Similarly, we find a depletion of PDTs generated by deletions in chimpanzees, as well as an expected but previously unreported reduction of inversions altering TADs. Taken together, the paucity of SVs altering domain boundaries suggests such variants in chimpanzee experience strong negative selection, as observed in other species, perhaps due to conserved roles of TADs in modulating gene regulation. Despite the overall depletion of SVs at TAD boundaries, we did find an increased incidence of species-specific domain boundaries and significant enrichment of DE genes near SVs in the two cell types queried in this study, concordant with previous findings assessing the impact of deletions and duplications on differential gene expression in primate LCLs (Iskow et al., 2012). Our analyses are subject to some limitations. Domain calling is highly sensitive to input parameters, but the pairs of Hi-C maps were subject to the same analysis and highly correlated at a variety of resolutions tested (MoC>0.7 at 100 kbp, 50 kbp, 25 kbp, and 10 kbp for iPSCs; 100 kbp and 50 kbp for LCLs) allowing for an assessment of genome-wide domain differences. Though the number of aligned reads were normalized to comparable levels, relative read depth is likely to vary across the genome due to differences in mappability. This is particularly likely at SV loci, where deletions and SDs generate discontinuities in

the Hi-C matrix. As such, these domain calls should be interpreted primarily as a means of identifying regions of putatively disrupted chromatin structure.

Notably, many of the genes near SVs were not DE; however, it is plausible that these non-DE genes either remain connected to their regulatory elements or their associated elements are specific to cell types not assayed. Further, while it has been reported that topology-altering SVs can have little effect on gene expression (Ghavi-Helm et al., 2019), or that expression is not globally altered by loss of TADs (Rao et al., 2017), it could still be the case that expression-altering SVs are frequently subject to negative selection. For instance, TAD- and expression-altering SVs reported in humans are typically de novo and pathogenic (Franke et al., 2016; Lupiáñez et al., 2015). Regardless, our findings are concordant with those of (Kronenberg et al., 2018), who reported an enrichment of human–chimpanzee cortical organoid DE genes near fixed human-specific SVs. While they find an enrichment for downregulated genes at insertions and deletions and upregulated genes at SDs, their analysis produced a much smaller set of DE genes (785 across both cell types using single-cell RNA-seq) and a much larger set of variants (17,789). These findings are also in line with reports that SVs underlie many human expression quantitative trait loci (Chiang et al., 2017). However, considering the currently incomplete understanding of the relationship between gene regulation and three-dimensional chromatin structure, we emphasize that functional studies are necessary to causally implicate SVs in gene expression differences within or between species.

In addition to using Illumina genotyping of our identified SVs to filter out putatively false positive variants, we also used this information to query SV differences across subspecies. In our high-confidence set of SVs, we identified one novel deletion in chimpanzees (human chromosome 6q11.1; chr6:60,639,753-60,662,981, GRCh38) from our BNG data of the western individual carrying substantial central ancestry (S003641) that was also found uniquely in central chimpanzees (n = 4). Considering the relatively low ancestry contribution of this individual assigned to the central-chimpanzee population (~13%), this highlights the importance of sequencing more diverse individuals to identify additional subspecies-specific SVs to better survey the complete variant landscape. Using these same genotypes, we also focused on a set of genes universally impacted by SVs across all chimpanzees tested, but not detected in the other great apes studied (humans and gorillas), since these genes may putatively contribute to species-specific traits (Table 2.1). One example, *APOL4*, encoding Apolipoprotein L4, was completely deleted in all chimpanzees tested (n = 25) and also shown to be downregulated in both LCLs and iPSCs in chimpanzees when compared with humans. This gene

is a member of a tandemly-duplicated family that has experienced a recent expansion in the primate lineage (Monajemi et al., 2002) and may play a role in lipid trafficking throughout the body. Human polymorphism at this locus has been shown to be associated with schizophrenia (Takahashi et al., 2008). Several identified genes also exhibited signatures of natural selection. One example region putatively under balancing selection includes two deletions impacting the primate-expanded galectin gene cluster, a family of proteins that specifically bind β-galactoside sugars and are important in modulating immune response through interactions with T cells (Balogh et al., 2019). Both deletions (10 kbp and 35 kbp in size, respectively) are found homozygously in all chimpanzees tested (n = 25), and thus are likely not the target of balancing selection, but they completely ablated *CLC* (or *LGALS10*) and *LGALS14*, as well as the downstream region of *LGALS13* (**Figure S2.9**). Two of these genes (*LGALS13* and *LGALS14*), expressed exclusively in human placenta (Nandor Gabor Than et al., 2009), are important drivers of maternal adaptive immune response, with reductions in expression of either gene shown to be associated with an increased risk of preeclampsia (Nándor Gábor Than et al., 2014). Although the mechanisms are unclear, it is notable that other immune-related genes with connections to preeclampsia also exhibit signatures of balancing selection in humans (Andrés et al., 2010; Tan et al., 2005; Wedenoja et al., 2019). It is possible that deletions impacting this gene cluster may contribute to pregnancy-related outcomes in chimpanzees that could be subject to natural selective pressures.

2.5 Methods

2.5.1 Cell line Growth and DNA Extraction

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

resuspended in Cell Buffer and embedded in an agarose plug. The plug was treated with Proteinase K for 18 hours followed by RNase A digestion for one hour. After extensive washing, the plug was melted, agarose was digested, and drop dialysis was performed to clean the DNA. A Qubit dsDNA BR Assay kit (Thermo Fisher Scientific) was used to quantify the DNA. All sequence data generated as part of this project are available for download at the European Nucleotide Archive (accession number PRJEB36949).

2.5.2 Determination of Chimpanzee Subspecies

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

2.5.3 ONT Promethion Library Preparation and Sequencing

gDNA was sheared to an average size of 50 kbp using a Megaruptor instrument (Diagenode, Denville, NJ) and then verified on a Pippin Pulse gel. A sequencing library was prepared starting with 2 µg of sheared DNA using the ligation sequencing kit SQK-LSK109 (ONT, Oxford, UK) following the instructions of the manufacturer with the exception

of extended incubation times for DNA damage repair, end repair, ligation, and bead elutions. Thirty femtomoles of the final library were loaded on PromethION R9.4.1 flow cell (ONT, Oxford, UK) and the data were collected for 64 hours. Basecalling was performed live on the compute module using MinKNOW v2.1 (Oxford Nanopore Technologies, Oxford, UK). Details of the dataset can be found in Table S2.1.

2.5.4 BNG Saphyr Library Preparation and Sequencing

AG18359 and S003641 were sequenced at the McDonnell Genome Institute at Washington University and the UC Davis Genome Center DNA and Expression Analysis Core, respectively. A total of 750 ng of HMW gDNA was labeled with DLE-1 enzyme, followed by proteinase digestion and a membrane clean-up step using the BNG Prep DLS DNA Labeling Kit (#80005). After overnight staining with an intercalating dye, the labeled DNA was loaded onto a Saphyr Chip G2.3 (BNG #20366) and run on the Saphyr system (BNG #60325) using the Saphyr Instrument Control Software (ICS, version 3.1) to maximize throughput of molecules. Raw images of DNA were converted into digital molecules files using Saphyr ICS version 3.1. Details of both datasets can be found in Table S2.1.

2.5.5 Detection of SVs

To detect SVs, ONT long-reads were mapped to the human (GRCh38, no alternative haplotypes) and the chimpanzee reference genome (panTro6) using minimap2 (v2.17-r941) and SVs were identified using Sniffles (v1.0.11) with "--genotype" flag and default parameters. Large SVs were identified from BNG opticals maps using Bionano Solve (v3.5) (Hastie et al., 2017) de novo genome assembly and SV-discovery pipeline using human GRCh38 as the reference. The SV file in SMAP format was converted to VCF format using the smap_to_vcf_v2.py script contained in Solve software (v3.4.1). Only the variants with "PASS" filter were considered in the analysis and homozygous reference calls were removed. SV size selection and filtering were performed with the bcftools (v1.9) view using the filter "INFO/SVLEN ≥ 10000 || INFO/SVLEN < −10,000" for both ONT and BNG datasets. To compare overlap between the SVs discovered by each method, we obtained 50% reciprocal overlap between features using bedtools intersect (v2.29.0) with flags "−f 0.5 −F 0.5." Deletions and inversions were retrieved from the SVTYPE tag and processed separately in downstream analyses.

2.5.6 Genotyping and Filtering of SVs

Variants for each callset were genotyped independently using previously published Illumina data from 25 chimpanzees from all four subspecies, as well as eight gorillas and eight humans. SNV genotypes from non-human primates were retrieved from the GAGP (Prado-Martinez et al., 2013) and human SNV genotypes were obtained from the Simons Genome Diversity Project (Mallick et al., 2016) (Table S2.2). Reads were mapped to the human reference (GRCh38) using BWA MEM (0.7.17−r1188) (H. Li, 2013) and subsequently merged and sorted with samtools (v1.9) for each individual. Large inversions and deletions (>10 kbp) were genotyped with SVtyper (v.0.7.1) (Chiang et al., 2015). Genotype information was retrieved using bedtools query (v2.29.0). To assess whether a variant was novel to this study, calls were compared to previously reported deletions and inversions larger than 10 kbp found in any great ape or any variant discovered in chimpanzee (Catacchio et al., 2018; Kronenberg et al., 2018; Sudmant et al., 2013) using bedtools intersect (v2.29.0) with 50% reciprocal overlap. SVs that were either (1) genotyped in one chimpanzee individual (1/1 or 0/1) or (2) reported as discovered in chimpanzee in previous studies, were selected to generate a higher confidence set (filter 1). This dataset was further refined by collapsing calls within the dataset with 50% reciprocal overlap. All novel calls were visually inspected in Integrative Genome Browser for ONT calls (Robinson et al., 2011) and Bionano Access for BNG calls. Also, SVs present in ≥90% of the chimpanzee individuals (22 or more) as well as absent in outgroups (human and gorilla) were included in the likely chimpanzee-specific dataset (filter 2). In (Kronenberg et al., 2018), eight chimpanzee individuals were genotyped; as such, variants with evidence in seven or more individuals were also included in the chimpanzee-specific dataset. The distribution of high-confidence calls across the human reference (GRCh38) was plotted using the R package Karyoplotter (Gel & Serra, 2017).

2.5.7 Annotation of Impacted Genes

Genes impacted by SVs were obtained by intersecting Gencode v27 genomics features annotation file to deletion coordinates ±2.5 kbp and inversion breakpoints (considered as estimated breakpoints ±2.5 kbp and ±50 kbp) using bedtools intersect (v2.29.0). The impact of the SVs on the function of the gene was predicted using Ensembl Variant Effect Predictor (VEP) (McLaren et al., 2016) with the Gencode v27 GTF file. The probability of loss of function intolerance score (pLI) was obtained from the gene constraints scores table in the Exome Aggregation Consortium database (Karczewski et al., 2019). Gene ontology (GO) annotations and overrepresented terms were retrieved for

each gene using DAVID (Huang et al., 2009a, 2009b) and by selecting terms at a 5% false-discovery rate (FDR). Genes previously identified as showing signatures of positive and balancing selection in chimpanzees were retrieved from previously published data (Cagan et al., 2016) and intersected with the set of genes impacted by SVs.

2.5.8 Differential Gene Expression

We obtained previously-published RNA-seq data from chimpanzee and human LCLs (Khan et al., 2013) and induced pluripotent stem cells (iPSCs) (Pavlovic et al., 2018). Raw data were trimmed using TrimGalore (v0.6.0) with the following parameters: "-q 20 --phred33 --length 20". Transcripts per million (TPM) values were estimated using Salmon (v0.14.1) (Patro et al., 2017) with the "--validateMappings" flag for all transcripts in GENCODE v27 and chimpanzee transcriptome published by (Kronenberg et al., 2018), which was based on a combination of orthologous genes identified via comparisons of human GENCODE v27 and novel transcripts identified through PacBio isoSeq of iPSCs. The R package tximport (Soneson et al., 2015) was used to estimate gene-level counts from TPM values using the setting 'countsFromAbundance = "lengthScaledTPM"' for 55,461 annotated genes with equivalent identifiers in the two transcriptomes. Differential expression analysis was conducted with limma-voom (Law et al., 2014; Ritchie et al., 2015). Genes with fewer than 1 count per million across all samples were filtered from the analysis, and a model accounting for species and sex was implemented. Differentially-expressed (DE) genes were called at a 5% FDR.

2.5.9 Topologically Associated Domain (TAD) Analyses

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

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

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

2.5.10 Permutation Analyses

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

2.6 Acknowledgments

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

CHAPTER 3. A complete reference genome improves analysis of human genetic variation

Chapter 3 is adapted with minimal modification from

Aganezov S*, Yan SM*, Soto DC*, Kirsche M*, Zarate S*, Avdeyev P, et al. A complete reference genome improves analysis of human genetic variation. *Science*. 2022;376: eabl3533.

First authorship is shared between SA, SMY, DCS, MK, and SZ. DCS performed the genome-wide analysis of collapsed duplications (section 1) and medically relevant genes impacted by errors in the reference genome GRCh38 (section 5).

3.1 Abstract

Compared to its predecessors, the Telomere-to-Telomere CHM13 genome adds nearly 200 Mbp of sequence, corrects thousands of structural errors, and unlocks the most complex regions of the human genome to clinical and functional study. Here we demonstrate how this reference universally improves read mapping and variant calling for 3,202 and 17 globally diverse samples sequenced with short and long reads, respectively. We identify hundreds of thousands of variants per sample in previously unresolved regions, showcasing the promise of the T2T-CHM13 reference for evolutionary and biomedical discovery. Simultaneously, this reference eliminates tens of thousands of spurious variants per sample, including up to a 12-fold reduction of false positives in 269 medically relevant genes. The improvement in variant discovery coupled with population and functional genomic resources position T2T-CHM13 to replace GRCh38 as the prevailing reference for human genetics.

3.2 Introduction

For the past twenty years, the human reference genome (GRCh38) has served as the bedrock of human genetics and genomics (International Human Genome Sequencing Consortium, 2004; Lander et al., 2001; Schneider et al., 2017). One of the central applications of the human reference genome, and of reference genomes in general, has been to serve as a substrate for clinical, comparative, and population genomic analyses. More than one million human genomes have been sequenced to study genetic diversity and clinical relationships, and nearly all of them have been analyzed

by aligning the sequencing reads from the donors to the reference genome, e.g. (Karczewski et al., 2020; Stephens et al., 2015; Sudmant, Rausch, et al., 2015). Even when donor genomes are assembled de novo, independent of any reference, the assembled sequences are almost always compared to a reference genome to characterize variation by leveraging deep catalogs of available annotations (Seo et al., 2016; Shafin et al., 2020). Consequently, human genetics and genomics benefit from the availability of a high-quality reference genome, ideally without gaps or errors that may obscure important variation and regulatory relationships.

The current human reference genome, GRCh38, is used for countless applications, with rich resources available to visualize and annotate the sequence across cell types and disease states (ENCODE Project Consortium et al., 2020; GTEx Consortium, 2020; Navarro Gonzalez et al., 2021; Schneider et al., 2017; Taliun et al., 2021). However, despite decades of effort to construct and refine its sequence, the human reference genome still suffers from several major limitations that hinder comprehensive analysis. Most immediately, GRCh38 contains more than 100 million nucleotides that either remain entirely unresolved (currently represented as 'N' characters), such as the p-arms of the acrocentric chromosomes, or are substituted with artificial models, such as the centromeric satellite arrays (Miga et al., 2014). Furthermore, GRCh38 possesses 11.5 Mbp of unplaced and unlocalized sequences that are represented separately from the primary chromosomes (Church et al., 2015; Schneider et al., 2017). These sequences are difficult to study, and many genomic analyses exclude them to avoid identifying false variants and false regulatory relationships (Karczewski et al., 2020). Relatedly, artifacts such as an apparent imbalance between insertions and deletions (indels) have been attributed to systematic misassemblies in GRCh38 (Audano et al., 2019; Chaisson, Huddleston, et al., 2015; Chaisson et al., 2019). Overall, these errors and omissions in GRCh38 introduce biases in genomic analyses, particularly in centromeres, satellites, and other complex regions.

Another major concern regards the influence of the reference genome on the analysis of variation across large cohorts for population and clinical genomics. Several studies, such as the 1000 Genomes Project (1KGP) (The 1000 Genomes Project Consortium, 2015) and gnomAD (Karczewski et al., 2020), have provided information about the extent of genetic diversity within and between human populations. Many analyses of Mendelian and complex diseases use these catalogs of single nucleotide variants (SNVs), small indels, and structural variants (SVs) to rank and prioritize potential causal variants on the basis of allele frequencies (AFs) and other evidence (Gulko et al., 2015; Kircher et al., 2014; Yandell et al., 2011). When evaluating these resources, the overall quality and representativeness of the human

reference genome are important, if often overlooked, factors. Any gaps or errors in the sequence could obscure variation and its contribution to human phenotypes and disease. In addition to omissions such as centromeric sequences or acrocentric chromosome arms, the current reference genome possesses other errors and biases, including within genes of known medical relevance (Miller et al., 2021; Wagner et al., 2021). Furthermore, GRCh38 was assembled from multiple donors with clone-based sequencing, which creates an excess of artificial haplotype structures that can subtly bias analyses (Green et al., 2010; Lander et al., 2001). Over the years, there have been attempts to replace certain rare alleles with more common alleles, but hundreds of thousands of artificial haplotypes and rare alleles remain to this day (Ballouz et al., 2019; Schneider et al., 2017; Zerbino et al., 2020). Increasing the continuity, quality, and representativeness of the reference genome is therefore crucial for improving genetic diagnosis, as well as for understanding the complex relationship between genetic and phenotypic variation.

The Telomere-to-Telomere (T2T) CHM13 genome addresses many of the limitations of the current reference (Nurk et al., 2022). Specifically, the T2T-CHM13v1.0 assembly adds nearly 200 Mbp of sequence and resolves errors present in GRCh38. Here we demonstrate the impact of the T2T-CHM13 reference on variant discovery and genotyping in a globally diverse cohort. This includes all 3,202 samples from the recently expanded 1KGP sequenced with short reads (Byrska-Bishop et al., 2022) along with 17 samples from diverse populations sequenced with long reads (Nurk et al., 2022; Shafin et al., 2020; Zook et al., 2020). Our analysis reveals more than two million variants within previously unresolved regions of the genome, genome-wide improvements in structural variant discovery, and enhancement in variant calling accuracy across 622 medically relevant genes. In summary, our work demonstrates universal improvements in read mapping and variant calling, broadening the horizon for future genomic studies.

3.3 Results

3.3.1 Structural comparisons of GRCh38 and T2T-CHM13

3.3.1.1 Introducing the T2T-CHM13 genome

The T2T-CHM13 reference genome was primarily assembled from Pacific Biosciences (PacBio) High Fidelity (HiFi) reads augmented with Oxford Nanopore Technology (ONT) reads to close gaps and resolve complex repeats (Nurk et al., 2022). The resulting T2T-CHM13v1.0 assembly was subsequently validated and polished, with a consensus accuracy estimated to be between Phred Q67 and Q73 (McCartney et al., 2022; Nurk et al., 2022) and with only three

minor known structural defects detected (McCartney et al., 2022). The assembly is highly contiguous, with only five unresolved regions from the most highly repetitive ribosomal DNA (rDNA) arrays, representing only 9.9 Mbp of sequence out of >3.0 Gbp of fully resolved sequence. The version 1.0 assembly adds or revises 229 Mbp of sequence compared to GRCh38, defined as regions of the T2T-CHM13 assembly that do not linearly align to GRCh38 over a 1 Mbp interval (i.e., are "non-syntenic"). Furthermore, 189 Mbp of sequence are not covered by any primary alignments from GRCh38 and are resolved in the T2T-CHM13 assembly. A summary diagram of the syntenic/non-syntenic regions and their associated annotations are presented for chromosomes 1 and 21 (**Figure 3.1A**), along with a detailed report for all chromosomes (**Figures S3.1-3.4**). Note that the subsequent T2T-CHM13v1.1 assembly (Nurk et al., 2022) further resolves the rDNA regions using model sequences for some array elements, although for this study we analyze the v1.0 assembly, which does not contain these representations.

The bulk of the non-syntenic sequence within T2T-CHM13 comprises centromeric satellites (190 Mbp) (Altemose et al., 2022) and copies of segmental duplications (218 Mbp) (Vollger, Guitart, et al., 2022). These sequences could prove challenging for variant analysis, especially for variants identified using short-read sequencing. However, compared to GRCh38, we report an overall increase in unique sequence, defined as k-length strings (k-mers) found only once in the genome (e.g., 14.9 Mbp of added unique sequence when considering 50-mers, 23.5 Mbp for 100-mers, and 39.5 Mbp for 300-mers). These sequences delineate regions of confident mapping for short paired-end reads or longer reads, including in previously unrepresented portions of the genome (**Figure 3.1B** and **Figure S3.5** and **Figure S3.6**).

In highly repetitive regions, more than 106 Mbp of additional sequence was identified in T2T-CHM13 that requires reads of more than 300 bp to uniquely map compared to GRCh38. Concomitantly, T2T-CHM13 possesses fewer exactly duplicated sequences (≥5 kbp) shared across chromosomes (excluding sequence pairs within centromeres) than GRCh38 (**Figure S3.7** and **Figure S3.8**). Specifically, GRCh38 possessed 28 large shared interchromosomal sequences, primarily consisting of pairs of sub-telomeric sequences, with an additional 42 pairs involving at least one unplaced contig. All of these identical sequence pairs, save one between two subtelomeres, are non-identical in T2T-CHM13, as small but important differences between repetitive elements have now been resolved (Hoyt et al., 2022; Nurk et al., 2022).

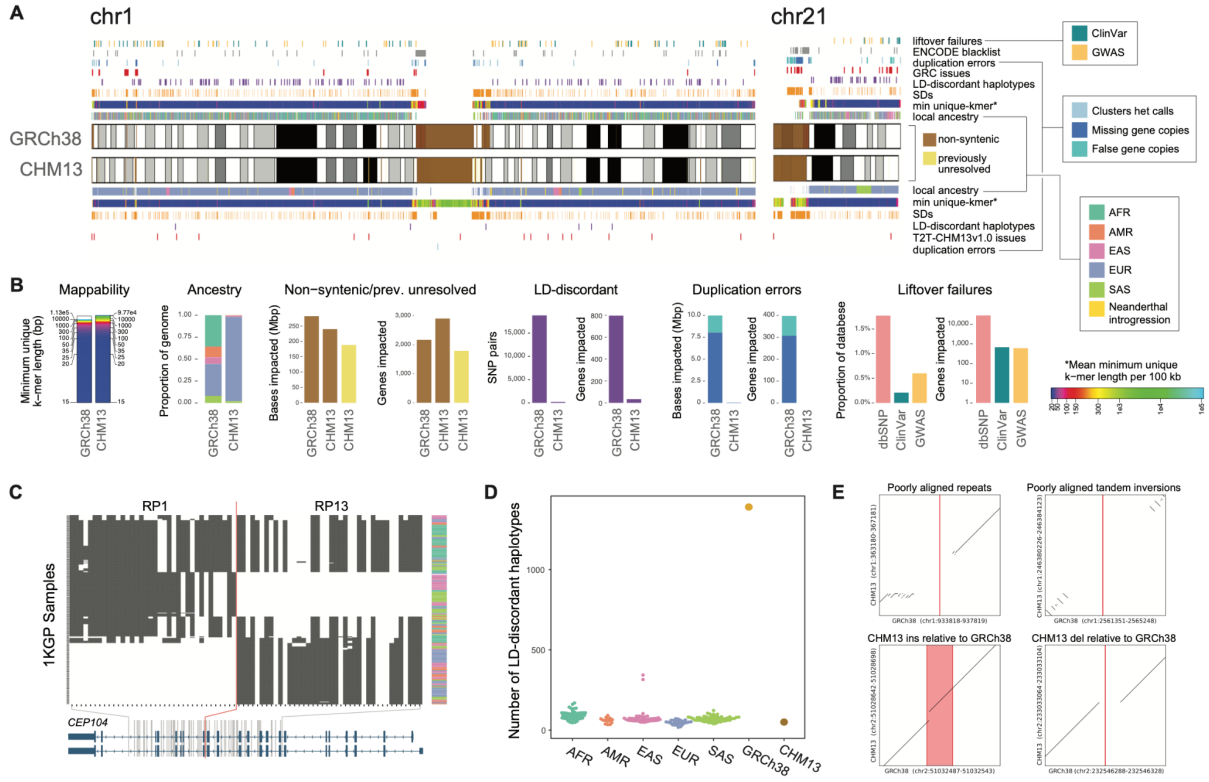


Figure 3.1. Genomic comparisons of human assemblies GRCh38 and T2T-CHM13. (A) Overview of annotations available for GRCh38 and T2T-CHM13 Chromosomes 1 and 21 with colors indicated in legends, which are also used in B-D. Colors for mean minimum (min) unique *k*-mers are defined in the legend with indicated asterisk. Cytobands are pictured as gray bands with red bands representing centromeric regions within ideograms. Complete annotations of all chromosomes can be found in **Figures S3.1-3.4**. Local ancestry is denoted using 1KGP superpopulation abbreviations (AFR: African, AMR: Admixed American, EAS: East Asian, EUR: European, SAS: South Asian). (B) Summary of the number of bases and/or genes annotated by different features for the assemblies with colors indicated in the legends shown in A. Note, dbSNP liftover failures (pink) are not annotated in A. (C) Example of a clone boundary (red line) where GRCh38 possesses a combination of alleles that segregate in negative LD within the 1KGP sample (which we term as an "LD-discordant haplotype"). SNPs are depicted in columns, while phased 1KGP samples are depicted in rows. White indicates reference allele genotypes, while black indicates alternative allele genotypes. Superpopulation ancestry of each sample is indicated in the rightmost column with colors indicated in local ancestry legend shown in A. *CEP104* splice isoforms (blue) are depicted at the bottom. (D) Tally of such LD-discordant haplotypes in a selection of 1KGP individuals, colored by population, as well as GRCh38 and T2T-CHM13. (E) Examples of variants that cannot be lifted over to T2T-CHM13 because of structural differences between the genomes. The position of the reference allele in GRCh38 is shown in red.

3.3.1.2 T2T-CHM13 accurately represents the haplotype structure of human genomes

The human reference genome serves as the standard to which other genomes are compared and is typically perceived as a haploid representation of an arbitrary genome from the population (Ballouz et al., 2019). In contrast with T2T-CHM13, which derives from a single homozygous complete hydatidiform mole, the Human Genome Project constructed the current reference genome via the tiling of sequences obtained from bacterial artificial chromosomes (BACs) and other clones with lengths ranging from ~50–150 kbp (Green et al., 2010), which derived from multiple

donor individuals. GRCh38 and its predecessors thus comprise mosaics of many haplotypes, albeit with a single library (RP11) contributing the majority (Green et al., 2010).

To further characterize this aspect of GRCh38 and its implications for population studies, we performed local ancestry inference for both GRCh38 and T2T-CHM13 through comparison to haplotypes from the 1KGP (Methods are available as supplementary materials) (**Figure 3.1A** and **Figure S3.2** and **Figure S3.9**). Continental superpopulation-level ancestry was inferred for 72.9% of GRCh38 clones based on majority votes of nearest-neighbor haplotypes. For the remaining 27.1% of clones, no single superpopulation achieved a majority of nearest neighbors, and ancestry thus remained ambiguous. This ambiguity occurs primarily for short clones with few informative SNPs (**Figure S3.10**), but also for some longer clones with potential admixed ancestry.

In accordance with Green et al., we inferred that library RP11, which comprises 72.6% of the genome, is derived from an individual of admixed African-American ancestry, with 56.0% and 28.1% of its component clones assigned to African and European local ancestries, respectively. The second most abundant library, CTD (5.5% of the genome), consists of clones of predominantly (86.3%) East Asian local ancestries, while the remaining libraries are derived from individuals of predominantly European ancestries. In contrast, CHM13 exhibits European ancestries nearly genome-wide (**Figure S3.11**). In addition, GRCh38 and T2T-CHM13 harbor 26.7 Mbp and 51.0 Mbp, respectively, of putative Neanderthal-introgressed sequences that originated from ancient interbreeding between the two hominin groups approximately 60 thousand years ago(Green et al., 2010). The excess of introgressed sequence in CHM13, even when restricting to the genomic intervals of GRCh38 clones with confident ancestry assignments, is consistent with its greater proportion of non-African ancestry.

We hypothesized that the mosaic nature of GRCh38 would generate abnormal haplotype structures at the boundaries of clones used for its construction, producing combinations of alleles that are rare or absent from the human population. Indeed, some previous patches of the reference genome sought to correct abnormal haplotype structures wherever noticed due to their impacts on genes of clinical importance (e.g., ABO and SLC39A4)(Schneider et al., 2017). Such artificial haplotypes would mimic rare recombinant haplotypes private to any given sample, but at an abundance and genomic scale unrepresentative of any living human. To test this hypothesis, we identified pairs of common (minor allele frequency [MAF] > 10%) autosomal SNP alleles always observed on the same haplotype (i.e.,

segregate in perfect [$R^2 = 1$] linkage disequilibrium [LD]) in the 2,504 unrelated individuals of the 1KGP and queried the allelic states of these SNPs in both GRCh38 and T2T-CHM13 (Materials and Methods Are Available as Supplementary Materials).

In accordance with our expectations, we identified numerous haplotype transitions in GRCh38 absent from the 1KGP samples, with 18,813 pairs of LD-discordant SNP alleles (i.e., in perfect negative LD) distributed in 1,390 narrow non-overlapping clusters (median length = 3,703 bp) throughout the genome (**Figure 3.1C**). Such rare haplotype transitions are comparatively scarce in T2T-CHM13, with only 209 pairs of common high-LD SNPs (50 non-overlapping clusters) possessing allelic combinations absent from the 1KGP sample (**Figure 3.1D**). Using a leave-one-out analysis, we confirmed that T2T-CHM13 possesses a similar number of LD-discordant haplotypes as phased "haploid" samples from 1KGP, whereas GRCh38 vastly exceeds this range (**Figure 3.1E**). By intersecting the GRCh38 results with the tiling path of BAC clones, we found that 88.9% (16,733 of 18,813) of discordant SNP pairs straddle the documented boundaries of adjacent clones (**Figure S3.12**). Of these, 45.9% (7,686 of 16,733) of the clone pairs derived from different BAC libraries, whereas the remainder likely largely reflects random sampling of distinct homologous chromosomes from the same donor individual. Thus, our analysis suggests that T2T-CHM13 accurately reflects haplotype patterns observed in contemporary human populations, whereas GRCh38 does not.

3.3.1.3 T2T-CHM13 corrects genomic collapsed duplications and falsely-duplicated regions

Genome assemblies often suffer from errors in complex genomic regions such as segmental duplications (SDs). In the case of GRCh38, targeted sequencing of BAC clones has been performed to fix many such loci (Chaisson, Huddleston, et al., 2015; Dennis et al., 2017; Huddleston et al., 2014; M. O'Bleness et al., 2014; Schneider et al., 2017; Steinberg et al., 2014), but problems persist. To systematically identify errors in GRCh38 that could produce spurious variant calls, we leveraged the fact that T2T-CHM13 is an effectively haploid cell line that should produce only homozygous variants when its sequence is aligned to GRCh38. Thus, any apparent heterozygous variant can be attributed to mutations accrued in the cell line, sequencing errors, or read mapping errors. In the last case, assembly errors or copy number polymorphism of SDs produce contiguous stretches of heterozygous variants (H. Cheng et al., 2021), which confound the accurate detection of paralog-specific variants (PSVs). Mapping PacBio HiFi reads from the CHM13 cell line (Nurk et al., 2022) as well as Illumina-like simulated reads (150 bp) obtained from the T2T-CHM13 reference to GRCh38, we identified 368,574 heterozygous SNVs within the autosomes and Chromosome X, of which 56,413

(15.3%) were shared between datasets. This evidence shows that each technology is distinctively informative due to differences in mappability (**Figure S3.13** and **Table S3.1**).

To home in on variants deriving from collapsed duplications, we delineated 'clusters' of heterozygous calls (Methods are available as supplementary materials) and identified 908 putative problematic regions (541 supported by both technologies) comprising 20.8 Mbp (**Figure 3.1** and **Figure S3.13**). Many of these loci intersected SD- (668/908; 73.6%) and centromere-associated regions (542/908; 59.7%) (Altemose et al., 2022) as well as known GRCh38 issues (341/908; 37.55%). Variants flagged as excessively heterozygous in the population by gnomAD (Karczewski et al., 2020) were significantly enriched in these regions (10,000 permutations, empirical p-value = $1x10^{-4}$), representing 23.6% (87,005/368,574) of our discovered CHM13 heterozygous variants, suggesting that these spurious variants arise in genome screens and represent false positives (**Figure 3.1A** and **Figure S3.1** and **Figure S3.3**).

We next 'lifted over' (i.e., converted the coordinates of) 821 of these 908 putative problematic regions to the T2T-CHM13 assembly and used human copy number estimates (n=268 individuals from the Simons Genome Diversity Project (SGDP)) (Mallick et al., 2016; Vollger, Guitart, et al., 2022) to conservatively identify 203 loci (8.04 Mbp) evidencing missing copies in GRCh38 (**Figure S3.14**). These regions impact 308 gene features, with 14 of the total 48 protein-coding genes fully contained within a problematic region, indicating that complete gene homologs are hidden from GRCh38-based population analyses of variation. Examples include DUSP22, a gene involved in immune regulation(J.-P. Li et al., 2014), as well as KMT2C, a gene implicated in Kleefstra syndrome 2(*OMIM Entry - #617768 - KLEEFSTRA SYNDROME 2; KLEFS2*, n.d.) (**Figure S3.15**). Additionally, we identified 30 SNPs within problematic regions with known phenotype associations from the GWAS Catalog(Buniello et al., 2019). Finally, we evaluated the status of these regions in the T2T-CHM13 reference by following a similar approach to obtain 9,193 heterozygous variants clustered in 11 regions—none of which overlapped GRCh38 problematic regions (**Table S3.2**). As a result, we are now able to call variants in these 48 previously-inaccessible protein-coding genes. We did identify one putative collapsed duplication in T2T-CHM13, based on the presence of a heterozygous variant cluster and reduced copy number in T2T-CHM13, localized to an rDNA array corrected in the most recent version of T2T-CHM13v1.1 (Nurk et al., 2022).

Conversely, the T2T-CHM13 reference also corrects regions falsely portrayed as duplicated in GRCh38. Specifically, we identified 12 regions affecting 1.2 Mbp and 74 genes (including 22 protein-coding genes) with duplications private to GRCh38 and not found in T2T-CHM13 or the 268 genomes from SGDP (Mallick et al., 2016) (**Figure S3.14** and **Table S3.3**). In contrast, only five regions affecting 160 kbp have duplications in T2T-CHM13 that are not in GRCh38 or the SGDP, suggesting that genuine rare variation cannot explain the excess of private duplications in GRCh38. Indeed, upon inspecting the CHM13 data, we deemed that these five loci are true duplications with support from mapped HiFi reads (McCartney et al., 2022).

The five largest duplications in GRCh38, affecting 15 protein-coding genes on the q-arm of Chromosome 21, involve BAC clones with sequence misplaced between gaps on the heterochromatic p-arm of the same chromosome. Based on admixture mapping, the Genome Reference Consortium (GRC)—an international team of researchers that has maintained and improved the reference genome and related resources since its initial publication—determined that these five clones were incorrectly localized to the acrocentric short arm and should not have been added to GRCh38 (Materials and Methods Are Available as Supplementary Materials). Of the seven false duplications outside Chromosome 21, two occur in short contigs between gaps, two occur adjacent to a gap, two occur on unlocalized "random" contigs, and one occurs as a tandem duplication (**Table S3.4**). We provide an exhaustive list of falsely duplicated gene pairs corrected in T2T-CHM13 (**Table S3.5**). Thus, T2T-CHM13 authoritatively corrects many false duplications, improving variant calling for short- and long-read technologies, including in medically relevant genes.

3.3.1.4 Liftover of clinically relevant and trait-associated variation from GRCh38 to T2T-CHM13

In transitioning to a different reference genome, it is imperative to document the locations of known genetic variation of biological and clinical relevance respective to the updated coordinate system. To this end, we sought to lift over 802,674 unique variants in the ClinVar database and 736,178,420 variants from the NCBI dbSNP database (including 151,876 NHGRI-EBI GWAS Catalog variants) from the GRCh38 reference to the T2T-CHM13 reference. Liftover was successful for 800,942 (99.8%) ClinVar variants, 723,117,125 (98.2%) NCBI dbSNP variants, and 150,962 (99.4%) GWAS Catalog SNPs (**Table S3.6**). We provide these lifted-over datasets as a resource for the scientific community within the UCSC Genome Browser and the NHGRI AnVIL, along with lists of all variants that failed liftover and the associated reasons (**Figure 3.1A**, **Figure 3.1B**, **Figure S3.1**, and **Figure S3.4**). Critically, this resource includes 138,319 of 138,927 (99.6%) ClinVar variants annotated as "pathogenic" or "likely pathogenic."

Of the 1,732 ClinVar variants that failed to lift over, 1,186 overlap documented insertions or deletions that distinguish the GRCh38 and T2T-CHM13 assemblies. The remaining 546 variants (< 0.1% of all variants) lie within regions of poor alignment between the GRCh38 and T2T-CHM13 assemblies (**Figure 3.1E**). The modes of liftover failure for variants in dbSNP and the GWAS Catalog follow similar distributions (**Table S3.6**). In all, these annotated variants offer a resource to enable researchers to interpret genetic results using the T2T-CHM13 assembly.

3.3.2 T2T-CHM13 improves analysis of global genetic diversity based on 3,202 short-read samples from the 1KGP dataset

3.3.2.1 T2T-CHM13 improves mapping of 3,202 short-read samples from the 1KGP dataset

To investigate how the T2T-CHM13 assembly impacts short-read variant calling, we realigned and reprocessed all 3,202 samples from the 1KGP cohort (Byrska-Bishop et al., 2022) using the NHGRI AnVIL Platform (Schatz et al., 2021) (**Figure S3.16** and **Figure S3.17**). In this collection, each sample is sequenced to at least 30× coverage with paired-end Illumina sequencing, with samples from 26 diverse populations across five major continental superpopulations (**Figure S3.18**). Though most samples are unrelated, the expanded collection includes 602 complete trios that we use to estimate the rate of false variants below based on discordance with Mendelian expectations. We matched the analysis pipeline for GRCh38 (Byrska-Bishop et al., 2022) as closely as possible so that any major differences would be attributable to the reference genome rather than technical differences in the analysis software (Methods are available as supplementary materials).

On average, BWA-MEM (H. Li, 2013) maps an additional 7.4×10^6 (0.97%) of properly paired reads map to T2T-CHM13 compared to GRCh38 , even when considering the alternative (ALT) and decoy sequences used in the original analysis (**Figure S3.19**). Interestingly, even though more reads align to T2T-CHM13, the subsequent per-read mismatch rate is 20% to 25% lower across all continental populations. African samples continue to present the highest mismatch rate (**Figure 3.2A**), as the observed mismatch rate includes both genuine sequencing errors, which are largely consistent across all samples, and any true biological differences between the read and the reference genome, which vary substantially based on the ancestry of the sample. Relatedly, T2T-CHM13 improved other mapping characteristics, including reducing the number of mis-oriented read pairs (**Figure 3.2A**). Finally, by considering the alignment coverage across 500 bp bins across the respective genomes, we observed improvement in coverage

uniformity within every sample's genome when using T2T-CHM13 rather than GRCh38. For example, within gene regions, we noted a 4-fold decrease in the standard deviation of the coverage (**Figure 3.2A**) and similar improvements in other types of genomic regions among all population groups (**Figure S3.20**). Overall, these improvements in error rates, mapping characteristics, and coverage uniformity demonstrate the superiority of T2T-CHM13 as a reference genome for short-read alignment across all populations.

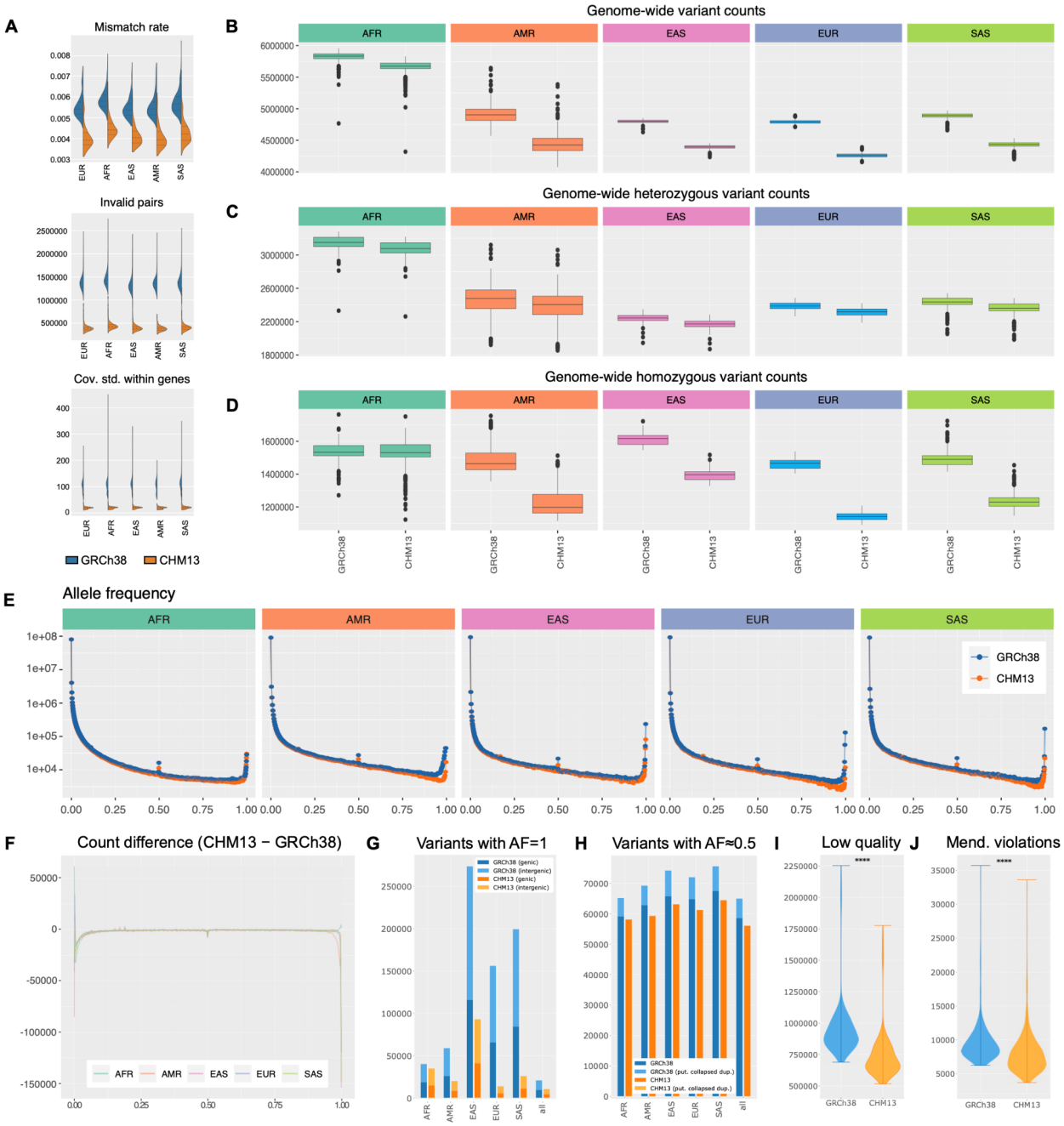


Figure 3.2. Improvements to Short-Read Mapping and Variant Calling. (A) Summary of alignment characteristics aligning to CHM13 instead of GRCh38. (B) Boxplot of overall number of variants found in each person across superpopulations, with colors indicated in **Figure 3.1A** legend. (C) Boxplot of the number of heterozygous variants found in each person across superpopulations. (D) Boxplot of the number of homozygous variants found in each person across superpopulations. (E) Allele frequency distribution of each superpopulation relative to CHM13 and GRCh38. (F) Change in allele frequency distribution. (G) Number of variants with allele frequency equal to 100%, both within protein-coding genes and without. (H) Number of variants with allele frequency equal to 50%, both within putative collapsed duplications and without. (I) Violin plot of the number of low-quality variants found when aligning to GRCh38 and CHM13. (J) Violin plot of the number of Mendelian violations found when aligning to GRCh38 and CHM13.

3.3.2.2 T2T-CHM13 improves variant calling across populations

From these alignments, we next generated SNV and small indel variant calls with the GATK Haplotype Caller, which uses a joint genotyping approach to optimize accuracy across large populations (Poplin et al., 2018). Again, we matched the pipeline used in the prior 1KGP study, albeit with updated versions of some analysis tools, to minimize software discrepancies and attribute differences to changes in the reference genome. Across all samples, we identified 126,591,489 high-quality ("PASS") variants relative to T2T-CHM13 (per-sample mean: 4,717,525; median: 4,419,802) compared to 125,484,020 variants relative to GRCh38 (per-sample mean: 5,101,897; median: 4,867,871), additionally noting a decrease in the number of called variants per individual genome (**Figure 3.2B**, **Figure S3.21**). We performed all subsequent analyses using these high-quality variants, as the PASS filter successfully removed spurious variants (**Figure S3.22**), particularly in complex regions (**Figure S3.23**).

As with the improvement to the per-read mismatch rate, we attribute the reduction in the number of per-sample variant calls to improvements in the number of rare alleles, consensus errors, and structural errors in T2T-CHM13. This conclusion is supported by the observation that the number of heterozygous variants per sample is more similar (**Figure 3.2C**, **Figure S3.24**) across reference genomes in contrast to homozygous variants (**Figure 3.2D**, **Figure S3.24**). This discrepancy is especially pronounced in non-African samples, which have on average 200,000 to 300,000 more homozygous variants relative to GRCh38 than T2T-CHM13, likely because ~70% of the GRCh38 sequence comes from an individual with African-American ancestry, and African populations are enriched for rare and private variants (The 1000 Genomes Project Consortium, 2015).

Further investigating this relationship, we computed the AFs of variants from unrelated samples from each of the five continental superpopulations (**Figure 3.2E**). Though the distributions were nearly equivalent over most of the AF spectrum, we observed substantial differences for rare alleles (AF < 0.05), intermediate-frequency alleles, including

errors where nearly all individuals are heterozygous (AF ≈ 0.5), and fixed/nearly-fixed alleles (AF > 0.95). The most prominent difference in AF distributions affected fixed or nearly-fixed alleles in each assembly, for which all non-African superpopulations showed an excess of ~150,000 variants in GRCh38, while the African superpopulation showed an excess of 2,364 variants in T2T-CHM13 (**Figure 3.2F**). This observation is driven by a decrease in the number of completely fixed variants (100% AF) relative to GRCh38 (**Figure 3.2G**). Such variants represent positions where the reference genome itself is the only sample observed to possess the corresponding allele. These alleles arise either because of genuine private variants in one of the GRCh38 donors, or from sequencing errors in the reference genome itself, and result in 100% of other individuals possessing two copies of the alternative allele. As a result, these 'variants' will not be reported at all if the same reads are mapped to a different genome that does not have these private alleles. Interestingly, the number of such private 'singleton' variants in T2T-CHM13 lies squarely within the observed range of singleton counts among 1KGP samples, adjusting for the difference in ploidy (**Figure S3.25**). In addition to the lower rate of private variation compared to GRCh38, T2T-CHM13 possesses fewer ultra-rare variants, effectively reducing the number of 'nearly fixed' alleles in population data such as 1KGP.

Finally, the reduction in AF ≈ 0.5 variants is largely explained by the corrections to collapsed SDs (**Table S3.1**), as these regions are highly enriched for heterozygous PSVs in nearly all individuals caused by the false pileup of reads from the duplicated regions to a single location (**Figure 3.2H**). Collectively, the decrease in variants with AF = 1 and AF ≈ 0.5 largely explains the decrease in the overall number of variants observed per sample and across the entire population for T2T-CHM13.

Informed by these results, we considered the feasibility of calling variants using the T2T-CHM13 reference and then lifting over the results to GRCh38 for further analyses. Using a liftover tool to transform a variant call set for a single sample into a call set with respect to GRCh38 requires special handling to account for variants for which the two references have different alleles. Specifically, if one of the reference alleles is not present in the sample, it will be necessary to genotype the site against the T2T-CHM13 reference. Although this issue is less of a concern for large datasets like the 1KGP, even these large samples will contain a small number of variants that become invisible when switching reference genomes (**Figure S3.26**). In addition, differences in variant representation, especially in regions of low complexity, may cause lifted variant sets to differ from those called against the target reference.

3.3.2.3 Reduction of Mendelian discordant variants

As further quality control for the variant calls, we performed a Mendelian concordance analysis using the 602 trios represented in the 1KGP cohort. We observed a statistically significant decrease in both the number of low-quality variants (median: 890,701 (GRCh38) vs. 682,609 (T2T-CHM13), p-value = 4.943×10^{-96}, Wilcoxon signed-rank test) (**Figure 3.2I**) and the number of Mendelian-discordant variants (*i.e.*, variants found in children but not their parents, or homozygous parental variants not observed in their children (median: 8,879 (GRCh38) vs. 7,484 (T2T-CHM13), p-value = 7.346×10^{-96}, Wilcoxon signed-rank test) (**Figure 3.2J**) when aligned to T2T-CHM13 as compared to GRCh38. In addition to providing an estimate of the error rate for variant calls in this callset, this improvement has broad implications for clinical genetics analyses of de novo or somatic mutations, which have been implicated as causes of autism spectrum disorders (Iossifov et al., 2014) and many forms of cancer (Alexandrov et al., 2013).

3.3.3 T2T-CHM13 improves structural variant analysis of 17 diverse long-read samples

3.3.3.1 T2T-CHM13 improves mapping of 17 long-read samples

Next, we investigated the effects of using T2T-CHM13 as a reference genome for alignment and large SV calling from both PacBio HiFi and ONT long reads. To this end, we aligned reads and called SVs in 17 samples of diverse ancestries from the Human Pangenome Reference Consortium (HPRC+) (Nurk et al., 2022) and the Genome in a Bottle Consortium (GIAB) (Zook et al., 2020), including two trios (**Table S3.7**). All of these samples had HiFi data available, and fourteen had also been sequenced with ONT (**Figure 3.3A**), with mean read lengths of 18.1 kbp and 21.9 kbp and read N50 values of 18.3 kbp and 44.9 kbp, respectively (**Figure S3.27**).

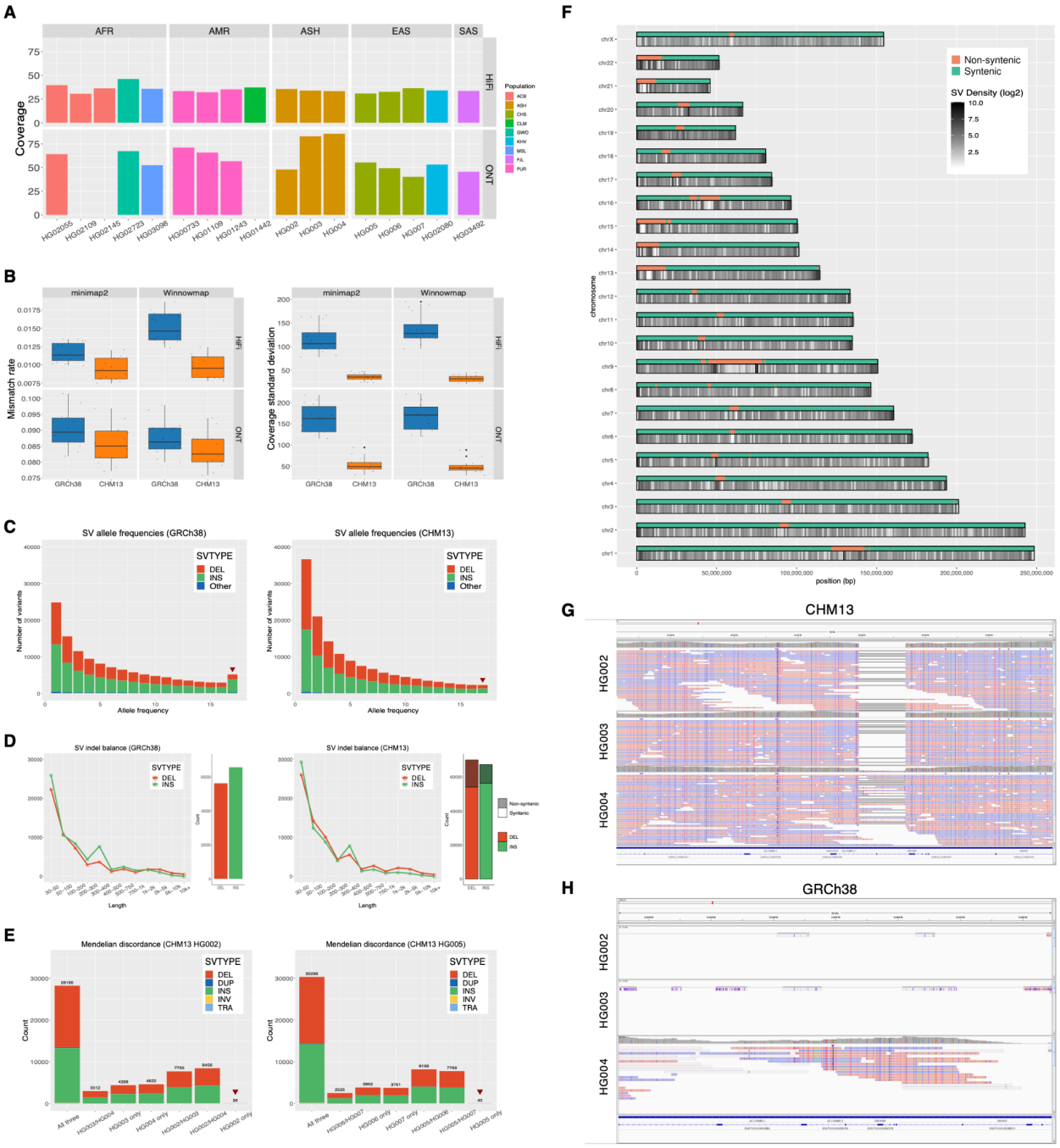


Figure 3.3. Improvements to Long-Read Alignment and SV Calling in CHM13. (A) The coverage, ancestry, and sequencing platforms available for the 17 samples sequenced with long reads (headers: AFR: African, AMR: Admixed American, ASH: Ashkenazi, EAS: East Asian, SAS: South Asian; populations: ACB: African Caribbean in Barbados, ASH: Ashkenazi, CHS: Southern Han Chinese, CLM: Colombian in Medellin, Colombia, GWD: Gambian in Western Division, The Gambia, KHV: Kinh in Ho Chi Minh City, Vietnam, MSL: Mende in Sierra Leone, PJL: Punjabi in Lahore, Pakistan, PUR: Puerto Rican in Puerto Rico). (B) The genome-wide mapping error rate and the standard deviation of the coverage for CHM13 (orange) and GRCh38 (blue). The standard deviation was computed across each 500bp bin of the genome. (C) The allele frequency of SVs derived from HiFi data in CHM13 and GRCh38 among the 17-sample cohort. The red arrows indicate fixed (100% frequency) variants. (D) The balance of insertions (INS) vs. deletion (DEL) calls in the 17-sample cohort in CHM13 and GRCh38. Variants in CHM13 are stratified by whether

or not they intersect regions which are non-syntenic with GRCh38. (E) The SV calls in CHM13 for two trios: a trio of Ashkenazi ancestry (child HG002, and parents HG003 (46XY), and HG004 (46XX), and a trio of Han Chinese ancestry (child HG005, and parents HG006 (46XY) and HG007 (46XX)). The red arrows indicate child-only, or candidate *de novo*, variants (DEL: Deletion, DUP: Duplication, INS: Insertion, INV: Inversion, TRA: Translocation). (F) The density of SVs called from HiFi data in the 17-sample cohort across CHM13. (G) Alignments of HiFi reads in the HG002 trio to CHM13 showing a deletion spanning an exon of the transcript AC134980.2 viewed using the Integrative Genomic Viewer (IGV). Pink horizontal rectangles indicate reads aligned to the forward strand; blue horizontal rectangles indicate reads aligned to the reverse strand. Thin black lines indicate split-read alignments. Small vertical rectangles indicate SNVs (H) Alignments of HiFi reads in the HG002 trio to the same region of GRCh38 as shown in (g), showing much poorer mapping to GRCh38 than to CHM13, viewed using IGV with colors same as (G).

In line with our short-read results, aligning long reads to T2T-CHM13 compared to GRCh38 did not substantially change the number of reads mapped with either Winnowmap (C. Jain et al., 2020) or minimap2 (H. Li, 2018) because most of the previously unresolved sequence in T2T-CHM13 represents additional copies of SDs or satellite repeats already partially represented in GRCh38 (**Figure S3.28**). However, aligning to T2T-CHM13 reduced the observed mismatch rate per mapped read by 5% to 40% across the four combinations of sequencing technologies and aligners because GRCh38 has more rare alleles. T2T-CHM13 also corrects structural errors in GRCh38 and is a complete assembly of the genome, which facilitates accurate alignment, similar to what we observed for short reads (**Figure 3.3B**). Relatedly, we find that previously reported African-specific (Sherman et al., 2019) and Icelandic-specific (Beyter et al., 2021) sequences at least 1 kbp in length align with substantially greater identity and completeness to T2T-CHM13 compared to GRCh38 (Materials and Methods Are Available as Supplementary Materials) (**Figure S3.29** and **Figure S3.30**).

To study coverage uniformity, we next measured the average coverage across each 500-bp bin on a per-sample basis and computed the standard deviation of the coverage. Across all aligners and technologies, the median standard deviation of the per-bin coverage was reduced by more than a factor of three, indicating more stable mapping to T2T-CHM13 (**Figure 3.3B**). This difference in coverage uniformity was pronounced in satellite repeats and other regions of GRCh38 that are non-syntenic with T2T-CHM13 (**Figure S3.31** and **Figure S3.32**). This coverage uniformity will broadly improve variant calling and other long-read-based analyses.

3.3.3.2 T2T-CHM13 improves SV imbalances on GRCh38

We next used our optimized SV-calling pipeline, including Sniffles (Sedlazeck, Rescheneder, et al., 2018), Iris, and Jasmine (Kirsche et al., 2021), to call SVs in all 17 samples (**Figure S3.33** and **Figure S3.34**) and consolidate them into a cohort-level callset in each reference from HiFi data. From these results, we observe a reduction from 5,147 to

2,260 SVs that are homozygous in all 17 individuals when calling variants relative to T2T-CHM13 instead of GRCh38 (**Figure 3.3C**). Previous studies (Audano et al., 2019; Chaisson, Huddleston, et al., 2015) have noted the excess of such SV calls when using GRCh38 as a reference and attributed them to structural errors. Here we find that using a complete and accurate reference genome naturally reduces the number of such variants. In addition, the number of indels is more balanced when calling against T2T-CHM13, whereas GRCh38 exhibited a bias towards insertions caused by missing or incomplete sequence (**Figure 3.3D**), such as incorrectly collapsed tandem repeats (Chaisson, Huddleston, et al., 2015). With respect to T2T-CHM13, we observe a small bias towards deletions, which likely results from the challenges in calling insertions with mapping-based methods and in representing SVs within repeats, as this difference is especially prominent in highly repetitive regions such as centromeres and satellite repeats (**Figure S3.35**). The variants we observe relative to T2T-CHM13 are enriched in the centromeres and sub-telomeric sequences, likely because of a combination of repetitive sequence and greater recombination rates (Audano et al., 2019). We observe similar trends among SVs unique to single samples (**Figure S3.36**).

We also observe similar improvements in the insertion/deletion balance for large SVs (>500 bp) detected by Bionano optical mapping data in HG002 against the T2T-CHM13 reference, with an increase in deletions (1,199 vs. 1,379) and a decrease in insertions (2,771 vs. 1,431) with GRCh38 and T2T-CHM13, respectively (**Figure S3.37**). Using the T2T-CHM13 reference for Bionano optical mapping also improves SV calling around gaps in GRCh38 that are closed in T2T-CHM13 (**Figure S3.38**), suggesting that T2T-CHM13 offers improved indel balance compared to GRCh38 across multiple SV-calling methods.

3.3.3.3 De novo SV analysis within trios

To investigate the impacts of the T2T-CHM13 reference on our ability to accurately detect de novo variants, we called SVs in both of our trio datasets using a combination of HiFi and ONT data and identified SVs only present in the child of the trio and supported by both technologies—approximately 40 variants per trio (**Figure 3.3E**). Manual inspection revealed a few variants in each trio strongly supported with consistent coverage and alignment breakpoints, while the other candidates exhibited less reliable alignments as noted in previous reports (Kirsche et al., 2021). In HG002, we detected six strongly-supported candidate de novo SVs that had been previously reported (Kirsche et al., 2021; Zook et al., 2020). In HG005, we detected a 1,571 bp deletion at chr17:49401990 in T2T-CHM13 supported as a candidate

de novo SV relative to both T2T-CHM13 and GRCh38 (**Figure S3.39**). This demonstrates the ability of T2T-CHM13 to be used as a reference genome for de novo SV analysis.

3.3.3.4 T2T-CHM13 enables the discovery of additional SVs within previously unresolved sequences

The improved accuracy and completeness of the T2T-CHM13 genome help resolve complex genomic regions. Within non-syntenic regions, we identified a total of 27,055 SVs (**Figure 3.3D**), the majority of which were deletions (15,998) and insertions (10,912). 22,362 of these SVs (82.7%: 8,903 insertions, 13,334 deletions) overlap previously unresolved sequences in T2T-CHM13, while the remaining SVs are now accessible because of the accuracy of the T2T-CHM13 reference. The AF and size distributions for these variants mirror the characteristics of the syntenic regions, with rare variants (**Figure S3.40**) and smaller (30–50 bp) indels (**Figure S3.41**) being the most abundant. However, we also note some non-syntenic regions with few or zero SVs identified. While many of these regions lie at the interiors of p-arms of acrocentric centromeres, which are gaps in T2T-CHM13v1.0 that have been resolved in later versions of the assembly, we also noticed depletions of SVs in a few other highly repetitive regions, such as the resolved human satellite array on Chromosome 9 (**Figure 3.3F**). We largely attribute the reduction in variant density to the low mappability of these complex and repetitive regions. Future improvements in read lengths and alignment algorithms are needed to further resolve such loci.

Within syntenic regions, we also note improvements to alignment and variant calling accuracy, including the identification of variant calls not previously observed within homologous regions of GRCh38. For example, in T2T-CHM13, we observe a deletion in all of the samples of the HG002 trio in an exon of the olfactory receptor gene AC134980.2 (**Figure 3.3G**), while the reads from those samples largely fail to align to the corresponding region of GRCh38 (**Figure 3.3H**). Meanwhile, reads from African samples (**Figure S3.42**) align to both references at this locus. The difference in alignment among different samples is likely due to the region being highly polymorphic for copy number variation; GRCh38 contains a reasonable representation of that region for the tested African samples, while the homologous region in T2T-CHM13 more closely resembles European samples (**Figure S3.43**). This highlights the need for T2T reference genomes for as many diverse individuals as possible to account for common haplotype diversity.

3.3.4 Variation within previously unresolved regions of the genome

3.3.4.1 T2T-CHM13 enables variant calling in previously unresolved and corrected regions of the genome

The T2T-CHM13 genome contains 229 Mbp of sequence that is non-syntenic to GRCh38, which intersects 207 protein-coding genes (**Figure 3.4A** and **Table 3.1**). Within these regions, we report 3,692,439 PASS variants across all 1KGP samples from short reads (**Figure 3.4B** and **Table 3.1**). Comparing variants called in a subset of 14 HPRC+ samples with Illumina, HiFi, and ONT data, we found that 73–78% of the Illumina-discovered SNVs are concordant with variants identified with PacBio HiFi long-read data using the PEPPER-Margin-DeepVariant algorithm (51,306–74,122 matching SNVs and genotypes per sample) (Shafin et al., 2021). Long reads discover over ten times more SNVs per sample than short reads in these regions, with 447,742–615,085 (41–43%) of SNVs matching between HiFi and ONT with PEPPER-Margin-DeepVariant. In non-syntenic regions, 97% of the SNVs called by HiFi fall in centromeric regions of CHM13, so we stratified concordance by type of satellite repeat within the centromere. We found that non-satellites in centric transition regions and monomeric satellites had higher concordance between HiFi and ONT, with >99% concordance in a few regions, but some as low as 50%. HSAT regions, which pose some of the greatest challenges for read mapping and harbor abundant structural variation, exhibit the lowest rates of concordance between the platforms.

Table 3.1. Overview of non-syntenic and previously unresolved regions and their respective variant counts.

	Nonsyntenic	**Previously unresolved**
Summary		
Total span (bp) (excluding Ns)	240,044,315 (228,569,315)	189,036,735 (177,561,735)
Unique span (100mers)	65,471,195	40,205,401
Protein-coding genes	207	207
Entire region		
1KGP SNVs + indels (within genes)	3,692,439 (138,829)	2,370,384 (52,567)
Short-read SNVs per sample	65,931 to 101,161	35,506 to 56,489
Long-read SNVs per sample	1,178,371 to 1,467,243	957,629 to 1,197,463
Short-read SNVs confirmed by long reads	73 to 78%	64 to 69%
Long-read SNVs identified in short reads	4 to 5%	3%
SNVs concordant between long reads	41 to 43%	38 to 40%
High-confidence regions (excluding coverage abnormalities)		
High-confidence region bases	13,683,528	2,987,935
Short-read SNVs confirmed by long reads in high-confidence regions	95 to 96%	84 to 88%
Long-read SNVs identified in short reads in high-confidence regions	60 to 63%	39 to 46%
SNVs concordant between long reads in high-confidence regions	91 to 95%	81 to 90%

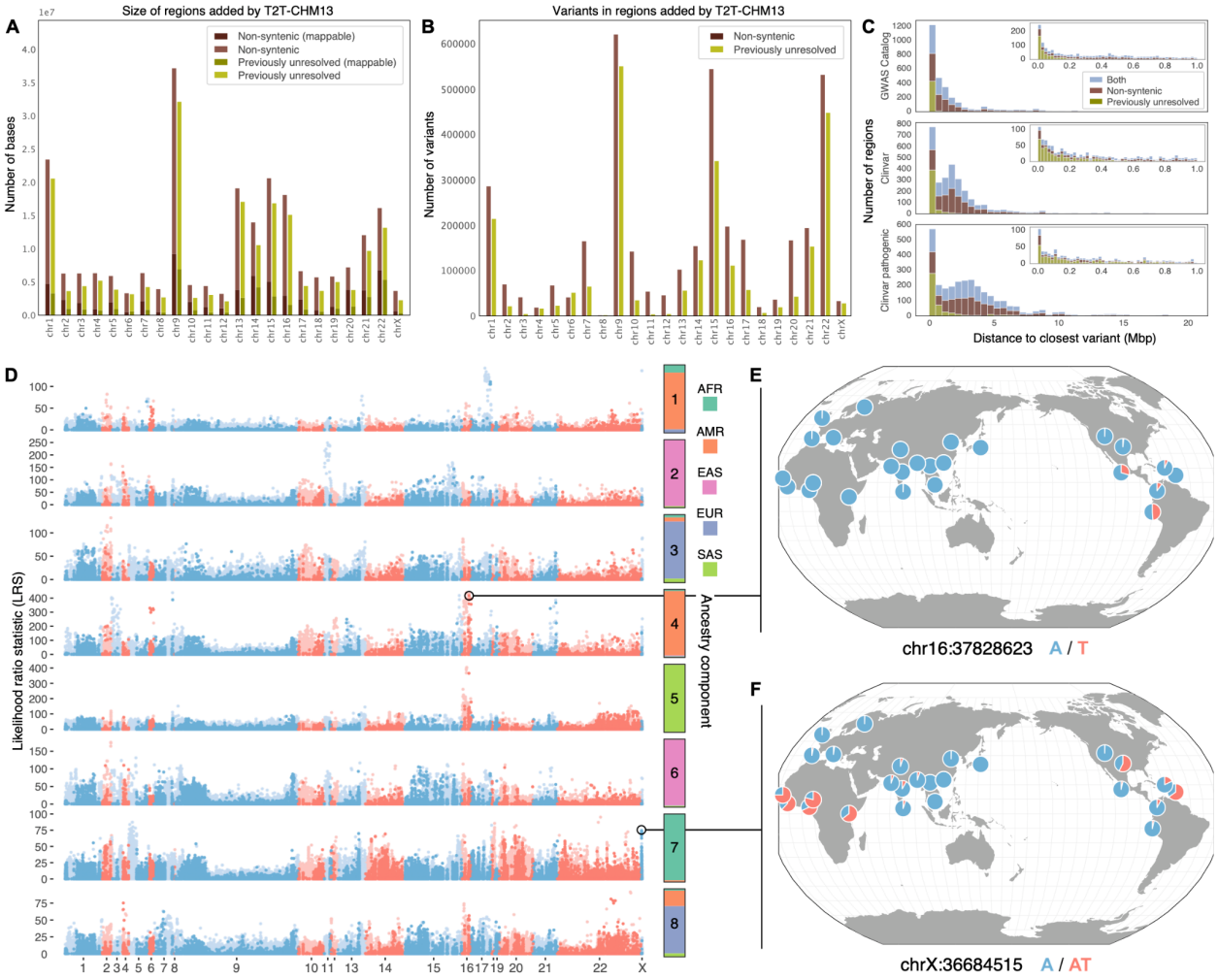


Figure 3.4. Characterization of variants within regions of the genome resolved by T2T-CHM13. (A) Number of bases added in non-syntenic and previously unresolved regions by chromosome, along with how many variants for each respective region are mappable (have contiguous unique 100mers). (B) Number of variants in non-syntenic and previously unresolved regions by chromosome. (C) Distance from each previously-unresolved-only, non-syntenic-only, or overlapping region to the closest Clinvar or GWAS Catalog variant. Insets are zoomed to 1 Mbp. (D) Scan for variants in non-syntenic (light blue and red) and previously unresolved (dark blue and red) regions that exhibit extreme patterns of allele frequency differentiation. Allele frequency outliers were identified for each of eight ancestry components, colored by the superpopulation membership of the corresponding 1KGP samples. Large values of the likelihood ratio statistic (LRS) denote variants for which AF differences in the corresponding ancestry component exceeds that of a null model based on genome-wide covariances in allele frequencies. (E, F) Population-specific allele frequencies of two highly differentiated variants in previously unresolved regions.

We further define conservative high-confidence regions by excluding regions with abnormal coverage in any long-read sample (i.e., coverage outside of 1.5× the interquartile range). This effectively excludes difficult-to-map regions with excessively repetitive alignments as well as copy number variable regions. After excluding abnormal coverage from non-syntenic regions, 14 Mbp remain, and SNVs from HiFi and ONT long reads are 91–95% concordant (21,835–28,237 variants). 95–96% (14,575–18,949) of short-read SNVs are found in HiFi long-read calls, though 37–

40% of HiFi SNVs are still missing from the short-read calls due to poorer mappability of the short-reads (**Table S3.8**). While many non-syntenic regions will require further method development (e.g., pangenome references (Miga & Wang, 2021)) to achieve accurate variant calls, the concordance of long- and short-read calls for tens of thousands of variants highlights previously unresolved sequences that are immediately accessible to both technologies.

As these broadly-defined non-syntenic regions include inversions and other structural changes between GRCh38 and T2T-CHM13 that do not necessarily alter many of the variants contained within, we also considered a narrower class of 'previously unresolved' sequences, representing segments of the T2T-CHM13 genome that do not align to GRCh38 with Winnowmap (C. Jain et al., 2020). Within these previously unresolved sequences, which span a total of 189 Mbp (**Figure 3.4A**, **Table 3.1**, and **Figure S3.44**), we report a total of 2,370,384 PASS variants in 1KGP samples based on short reads, intersecting 207 protein-coding genes (**Figure 3.4B**, **Table 3.1**, and **Figure S3.45**). We note that this set of 207 genes is distinct from the 207 genes that intersected with the non-syntenic regions, and these two sets together comprise 329 unique genes. Because these previously unresolved sequences are enriched for highly repetitive sequences, concordance is slightly lower, such that 64–69% of the SNVs in each sample match variants found in PacBio HiFi long-read data from the same samples (24,371–36,501 matching SNVs and genotypes per sample), and 339,783–473,074 (38–40%) of SNVs match between HiFi and ONT. When removing difficult-to-map and copy-number-variable regions as above, 3 Mbp of high-confidence regions remain. Within high-confidence regions, 84–88% of short-read SNVs in each sample match variants found in each sample's PacBio HiFi long-read data (2,938–3,811 matching SNVs and genotypes per sample), and 5,544–8,298 (81-90%) of SNVs match between HiFi and ONT (**Table S3.8**). While these previously unresolved regions are more challenging than non-syntenic regions, thousands of variants can still be called concordantly with short and long reads.

We noted homology between GRCh38 collapsed duplications and many T2T-CHM13 non-syntenic and/or previously unresolved regions (137 regions comprising 6.8 Mbp), indicating that the T2T-CHM13 assembly corrects these sequences through the deconvolution of nearly identical repeats. Comparing total variants identified in the 1KGP dataset, we observed a significant decrease in variant densities of 41 protein-coding genes intersecting with GRCh38 collapsed duplications in T2T-CHM13 (mean: 27 variants per kbp) compared with GRCh38 (mean: 46 variants per kbp; p-value = 6.906×10^{-8}, Wilcoxon signed-rank test) (**Figure S3.46**). Besides differences in local ancestries between the references, these higher variant densities in GRCh38 in part represent PSVs or mis-assigned alleles from

missing paralogs (Hartasánchez et al., 2018). Conversely, 1KGP variants were significantly increased in 32 protein-coding genes contained within GRCh38 false duplications using the T2T-CHM13 reference genome (mean values of 48 variants per kbp in T2T-CHM13 vs. 12 variants per kbp in GRCh38; p-value = 4.657×10^{-10}, Wilcoxon signed-rank test).

To assess whether these corrected complex regions in T2T-CHM13 accurately reveal variation, we evaluated the concordance of variants generated from short-read Illumina and PacBio HiFi sequencing datasets of two trios from the GIAB consortium and the Personal Genome Project (Ball et al., 2012) and observed similar recall for Illumina data in T2T-CHM13 (20.1–28.3%) and GRCh38 (21.5–25.4%), but with improved precision in the variants identified (98.1–99.7% in T2T-CHM13 vs. 64.3–67.3% in GRCh38) in a subset of the GRCh38 collapsed duplications (copy number < 10; ~910 kbp) (**Table S3.9**). Corrected false duplications (1.2 Mbp) exhibited 50-fold improved recall for Illumina data compared with HiFi in T2T-CHM13 (57.4–68.3%) vs. GRCh38 (1.1–1.8%), as well as improved precision in T2T-CHM13 (98.5–99.3%) vs. GRCh38 (76.5–95.8%) (**Table S3.9**). These improvements show that variants can be discovered and genotyped in regions corrected by the T2T-CHM13 assembly.

4.3.4.2 Phenotypic associations and evolutionary signatures within non-syntenic T2T-CHM13 regions

Sequences in the T2T-CHM13 assembly that are non-syntenic with GRCh38 offer opportunities for future genetic studies. Several such loci lie in close proximity to variation that has been implicated in complex phenotypes or disease, supporting their potential biomedical importance. These include 8 loci occurring within 10 kbp of GWAS hits and 19 loci within 10 kbp of ClinVar pathogenic variants (**Figure 3.4C**). In addition, 113 of 22,474 GWAS hits (representing 0.5% of all variants in the studies we tested) segregate in LD ($R^2 \geq 0.5$) with variants in non-syntenic regions, thereby expanding the catalog of potential causal variants for these GWAS phenotypes (Buniello et al., 2019) (**Figure S3.47** and **Table S3.10**).

Using short-read-based genotypes generated from the 1KGP cohort, we also searched for variants within non-syntenic regions that exhibit large differences in AF between populations—a signature that can reflect historical positive selection or demographic forces within these previously inaccessible regions of the genome. To study these signatures, we applied Ohana (J. Y. Cheng et al., 2021), a method that models individuals as possessing ancestry from k components and tests for ancestry component-specific frequency outliers. Focusing on continental-scale patterns (k =

8; **Figure S3.48**), we identified 5,154 unique SNVs and indels across all ancestry components that exhibited strong deviation from genome-wide patterns of AF (99.9th percentile of distribution for each ancestry component; **Figure 3.4D**). These included 814 variants overlapping with annotated genes and 195 variants that intersected annotated exons.

We first focused on the 3,038 highly differentiated non-syntenic variants that lift over from T2T-CHM13 to GRCh38. These successful liftovers allowed us to make direct comparisons to selection results, generated with identical methods, using 1KGP Phase 3 data aligned to GRCh38 (**Figure S3.49**) (Methods are available as supplementary materials) (Yan et al., 2021). For 41.3% of the lifted over variants, we found GRCh38 variants within a 2 kbp window that possessed similar or higher likelihood ratio statistics for the same ancestry component, indicating that these loci were possible to identify in scans of GRCh38 (**Figure S3.50**). The remaining 58.7% of lifted over variants may represent regions of the genome where differences in the T2T-CHM13 and 1KGP Phase 3 variant calling or filtering procedures lead to discrepancies in AFs between these two datasets. They may also indicate regions whose more accurate representation in T2T-CHM13 improves variant calling enough to resolve previously unknown signatures of AF differentiation (**Figure S3.51**). We then investigated the 943 variants that could not be lifted over from T2T-CHM13 to GRCh38 and were located in both previously unresolved sequences and regions deemed mappable from unique 100-mer analysis. Some of these variants overlap with genes, including several annotated with RNA transcripts in regions not present in the GRCh38 assembly (**Figure 3.4D** and **Table S3.11**).

We highlight two loci that exhibit some of the strongest allele frequency differentiation observed across ancestry components. The first locus, located in a centromeric alpha satellite on Chromosome 16, contains variants that reach intermediate allele frequency in the ancestry component corresponding to the Peruvian in Lima, Peru (PEL) and other Admixed American populations of 1KGP (AFs: 0.49 in PEL; 0.20 in CLM [Colombian in Medellin, Colombia] and MXL [Mexican Ancestry in Los Angeles, California]; absent or nearly absent elsewhere; **Figure 3.4E** and **Figure S3.52** and **Figure S3.53**). Variants at the second locus, located in a previously unresolved T2T-CHM13 sequence on the X chromosome that contains a multi-kbp imperfect AT tandem repeat, exhibit high AFs in the ancestry component corresponding to African populations of 1KGP and low AFs in other populations (AFs: 0.67 in African populations and 0.014 in European populations; **Figure 3.4F** and **Figure S3.54** and **Figure S3.55**. The variant at this locus with

the strongest signature of frequency differentiation also lies within 10 kbp of two pseudogenes, *MOB1AP2*-201 (MOB kinase activator 1A pseudogene 2) and *BX842568.1*-201 (ferritin, heavy polypeptide-like 17 pseudogenes).

We note that due to the repetitive nature of the sequences in which they reside, many of the loci that we highlight here remain challenging to genotype with short reads, and individual variant calls remain uncertain. Nevertheless, patterns of AF differentiation across populations are relatively robust to such challenges and can still serve as proxies for more complex SVs whose sequences cannot be resolved by short reads alone. The presence of population-specific signatures at these loci highlights the potential for T2T-CHM13 to reveal evolutionary signals in previously unresolved regions of the genome.

3.3.5 Impact of T2T-CHM13 on clinical genomics

3.3.5.1 Variants of potential clinical relevance in T2T-CHM13

A deleterious variant in a reference genome can mislead the interpretation of a clinical variant identified in a patient because it may not be flagged as such using standard analysis tools. The GRCh38 reference genome is known to contain such variants that likely affect gene expression, protein structure, or protein function (Ballouz et al., 2019), though systematic efforts have sought to identify and remove these alleles (Schneider et al., 2017). To determine the existence and location of loss-of-function variants in T2T-CHM13, we aligned the assembly to GRCh38 using dipcall (H. Li et al., 2018) to identify and functionally annotate nucleotide differences (McLaren et al., 2016) (**Figure 3.5A**). This analysis identified 210 putative loss-of-function variants (defined as variants that affect protein-coding regions and predicted splice sites) impacting 189 genes, 31 of which are clinically relevant (Wagner et al., 2021). These results are in line with work showing that the average diploid human genome contains ~450 putative loss-of-function variants impacting ~200–300 genes when low-coverage Illumina sequencing is applied (before stringent filtering) (MacArthur et al., 2012).

Of these 210 variants, 158 have been identified in at least one individual from 1KGP, with most variants relatively common in human populations (median AF of 0.47), suggesting that they are functionally tolerated. The remaining variants not found in 1KGP individuals comprise larger indels, which are more difficult to identify with 1KGP Illumina data, as well as alleles that are rare or unique to CHM13. We curated the ten variants impacting medically relevant genes and found seven that likely derived from duplicate paralogs: a 100-bp insertion also found in long reads

of HG002, a stop gain in a final exon in one gnomAD sample, and an insertion in a homopolymer in a variable-number tandem repeat in *CEL*, which may be an error in the assembly. Understanding that the T2T-CHM13 assembly represents a human genome harboring potentially functional or rare variants that in turn would affect the ability to call variants at those sites, we have made available the full list of putative loss-of-function variants to aid in the interpretation of sequencing results (**Table S3.12**).

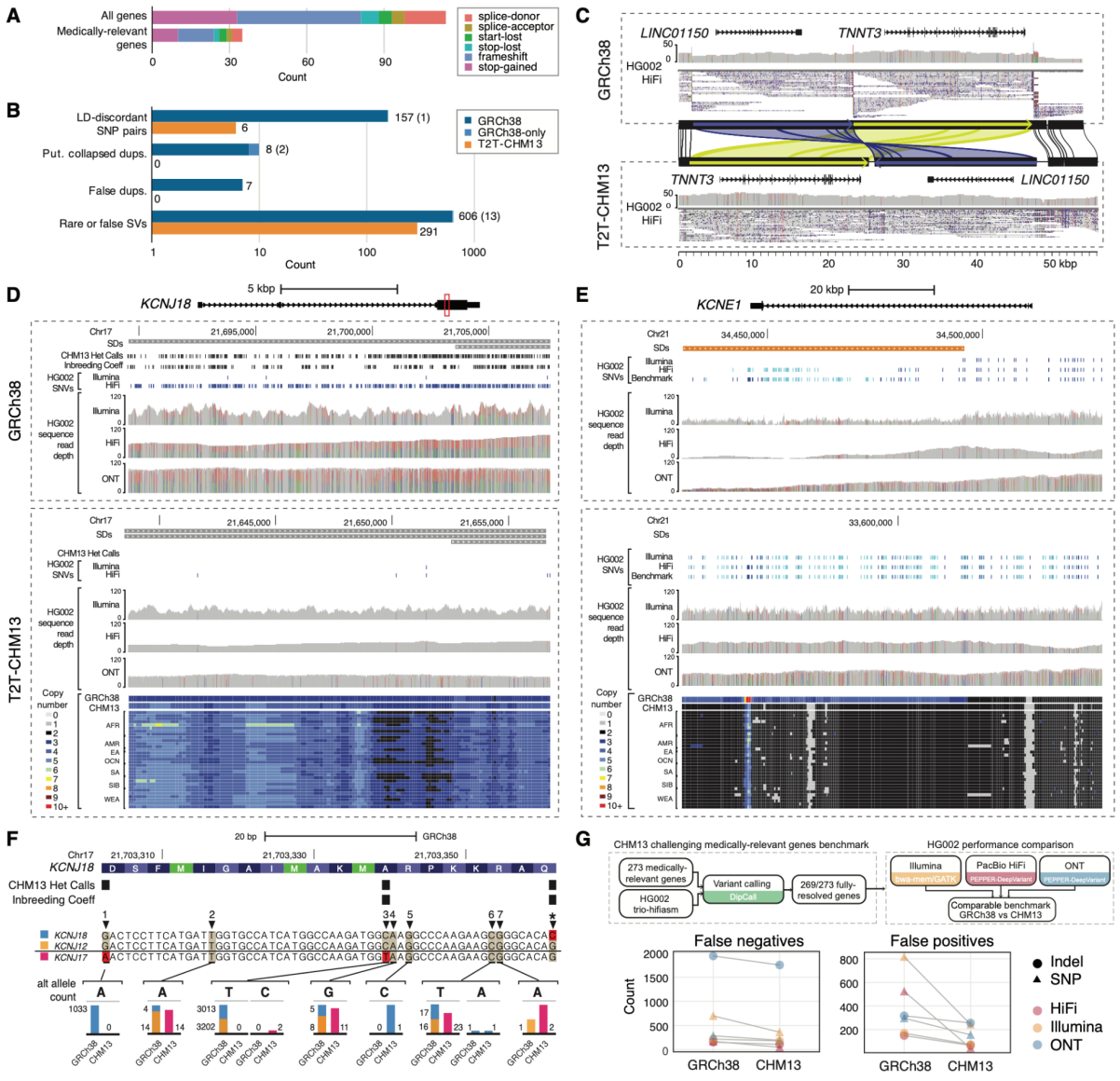


Figure 3.5. T2T-CHM13 Improves Clinical Genomics Variant Calling. (A) Numbers of potential loss-of-function mutations in the T2T-CHM13 reference. (B) The counts of medically-relevant genes impacted by genomic features and variation in GRCh38 (blue) and CHM13 (orange) are depicted as bar plots on logarithmic scale. Light blue indicates genes impacted in GRCh38 where homologous genes were not identified in T2T-CHM13 due to inability to lift over, with counts included in parentheses. (C) An example erroneous GRCh38 complex SV corrected in T2T-

CHM13 impacting *TNNT3* and *LINC01150*, displayed by sequence comparison using miropeats (Parsons, 1995) with homologous regions colored in green and blue, respectively. HG002 PacBio HiFi data is displayed showing read coverages and mappings from IGV, with allele fractions of variant sites colored (red=T; green=A; blue=C; black=G) within histograms of read depth (0–50). Snapshots of regions using IGV and UCSC Genome Browser representing (D) a collapsed duplication in GRCh38 corrected in T2T-CHM13 impacting *KCNJ18* and (E) a false duplication in GRCh38 impacting most of *KCNE1*. SDs depicted on top are colored by sequence similarity to paralog (gray: 90–98%; orange: >99%). Read mappings and variants from HG002 Illumina, PacBio HiFi, and ONT (mappings only), with homozygous (light blue) and heterozygous (dark blue) variants depicted as dashes. Colors within histograms of read depth (0–120) are the same as described in C. Copy-number estimates, displayed as colors indicated in legends, across k-merized versions of the GRCh38 and T2T-CHM13 references as well as representative examples of the SGDP individuals. (F) An example CDS region of *KCNJ18* (highlighted as a red box in D), with amino acids colored in alternating shades of blue and potential start codons (methionines) labeled in green using the UCSC Genome Browser codon-coloring scheme. Alignments of *KCNJ18* (blue), *KCNJ12* (orange), and *KCNJ17* (pink) along with allele counts of variants in each gene identified on GRCh38 and T2T-CHM13 are shown as bar plots (to approximate scale per variant), with examples 1–7 described in **Table S3.14**. (G) Schematic depicts a benchmark for 269 challenging medically relevant genes for HG002. The number of variant-calling errors from three sequencing technologies on each reference is plotted.

3.3.5.2 T2T-CHM13 improves variant calling for medically relevant genes

We sought to understand how the transition from GRCh38 to the T2T-CHM13 reference might impact variants identified in a previously compiled (Wagner et al., 2021) set of 4,964 medically relevant genes residing on human autosomes and Chromosome X (representing 4,924 genes in T2T-CHM13 via liftover; **Table S3.13**). Of these genes, 28 map to previously unresolved and/or non-syntenic regions of T2T-CHM13. We found over twice as many medically relevant genes impacted by rare or erroneous structural alleles on GRCh38 (n=756 including 14 with no T2T-CHM13 liftover) compared to T2T-CHM13 (n=306) (**Figure 3.5B**), of which 622 genes appear corrected in T2T-CHM13. This includes 116 genes falling in regions previously flagged as erroneous in GRCh38 by the GRC. The majority (82%) of impacted clinically relevant genes in GRCh38 overlap SVs that exist in all 13 HiFi-sequenced individuals, likely representing rare alleles or errors in the reference (see above), including 13 of the 14 genes with no T2T-CHM13 liftover.

One example of a resolved gene structure involves *TNNT3*, which encodes Troponin T3, fast skeletal type, and is implicated in forms of arthrogryposis (Sung et al., 2003). When calling SVs with respect to GRCh38, TNNT3 was previously postulated to be impacted by a complex structural rearrangement in all individuals, consisting of a 24-kbp inversion and 22-kbp upstream deletion, which also ablates *LINC01150* (**Figure 3.5C**). The GRC determined that a problem existed with the GRCh38 reference in this region (GRC issue HG-28). Analysis of this region in T2T-CHM13 instead shows a complex rearrangement with the 22 kbp region upstream of *TNNT3* inversely transposed in the T2T-CHM13 assembly to the proximal side of the gene. Besides potentially affecting interpretations of gene regulation,

this structural correction of the reference places TNNT3 >20 kbp closer to its genetically-linked partner *TNNI2* (Sheng & Jin, 2016). Other genes have VNTRs that are collapsed in GRCh38, such as one expanded by 17 kbp in most individuals in the medically relevant gene *GPI*. *MUC3A* was also flagged with a whole-gene amplification in all individuals, which we identified as residing within a falsely-collapsed SD in GRCh38, further evidencing that finding (**Figure 3.1A**).

Seventeen medically relevant genes reside within erroneous duplicated and putative collapsed regions in GRCh38 (**Table S3.1** and **Table S3.3**), including *KCNE1* (false duplication) and *KCNJ18* (collapsed duplication) (**Figure 3.5D** and **Figure 3.5E**). For these genes, we show that a significant skew in total variant density occurs in GRCh38 (58 variants per kbp for eight genes in collapsed duplications and 21 variants per kbp for seven genes in false duplications; p-values = 5.684×10^{-3} and 6.195×10^{-4}, respectively, Mann-Whitney U test) versus the rest of the 4,909 medically-relevant gene set (40 variants per kbp) that largely disappears in T2T-CHM13 (40 variants per kbp in collapsed duplications and 47 variants per kbp in false duplications versus 41 variants per kbp for the remaining gene set; p-values = 0.8778 and 0.0219, respectively) (**Figure S3.45**). Examining *KCNE1*, we find that coverage is much lower than normal on GRCh38 for short and long reads and that most variants are missed because many reads incorrectly map to a likely false duplication (*KCNE1B* on the p-arm of chromosome 21). The kmer-based copy number of this region in all 266 SGDP genomes supports the T2T-CHM13 copy number, and that this region was not duplicated in GRCh37 (Wagner et al., 2021). As for *KCNJ18*, which resides within a GRCh38 collapsed duplication at chromosome 17p11.2 (Ryan et al., 2010), we find increased coverage and variants within HG002 using short- and long-read sequences in GRCh38 relative to T2T-CHM13.

To verify if the additional variants identified using GRCh38 are false heterozygous calls from PSVs derived from missing duplicate paralogs, we compared the distributions of minor-allele frequencies across the 49-kbp SD. We observed a shift in SNV proportions, with a relative decrease in intermediate-frequency alleles and a relative increase in rare alleles for *KCNJ18* and *KCNJ12* (another collapsed duplication residing distally at chromosome 17p11.2) in T2T-CHM13 compared with GRCh38 (p-value = 8.885×10^{-2} and 3.102×10^{-2}, respectively; Mann-Whitney U test) (**Figure S3.56**). We matched the homologous positions of discovered alternative alleles in GRCh38 and T2T-CHM13 across the three paralogs—including the previously missing paralog located in a centromere-associated region on chromosome 17p *KCNJ17* denoted *KCNJ18*-1 in T2T-CHM13—and observed that even true variants (i.e., non-PSVs)

had discordant allele counts in *KCNJ18* and *KCNJ12* between the two references (**Figure 3.5F**, **Table S3.14**). Considering that rare variants of *KCNJ18* contribute to muscle channelopathy-thyrotoxic periodic paralysis (Ryan et al., 2010) including nine "pathogenic" or "likely pathogenic" variants in ClinVar, increased sensitivity to discover variants in patients using T2T-CHM13 would have a significant clinical impact. In summary, the improved representation of this gene and other collapsed duplications in T2T-CHM13 not only eliminates false positives but also improves detection and genotyping of true variants.

3.3.5.3 Clinical gene benchmark demonstrates T2T-CHM13 reduces errors across technologies

Finally, to determine how the T2T-CHM13 genome improved the ability to assay variation broadly, we used a curated diploid assembly to develop a benchmark for 269 challenging medically-relevant genes in GIAB Ashkenazi son HG002 (Wagner et al., 2021), with comparable benchmark regions on GRCh38 and T2T-CHM13. We tested three short- and long-read variant callsets against this benchmark: Illumina-BWAMEM-GATK, HiFi-PEPPER-DeepVariant, and ONT-PEPPER-DeepVariant. Counts of both false positives and false negatives substantially decrease for all three callsets when using T2T-CHM13 as a reference instead of GRCh38 (**Figure 3.5G** and **Table S3.15**). The number of false positives for HiFi decreases by a factor of 12 in these genes, primarily due to the addition of missing sequences similar to *KMT2C* (**Figure S3.15**) and removal of false duplications of *CBS*, *CRYAA*, *H19*, and *KCNE1* (**Figure 3.5G**). As demonstrated above, T2T-CHM13 better represents these genes and others for a diverse set of individuals, so performance should be higher across diverse ancestries. Furthermore the number of true positives decreases by a much smaller fraction than the errors (~14%) due to a reduction of true homozygous variants caused by T2T-CHM13 possessing fewer ultra-rare and private alleles (**Figure 3.2G**). This benchmarking demonstrates concrete performance gains in specific medically relevant genes resulting from the highly accurate assembly of a single genome.

3.4 Discussion

Difficult regions of the human reference genome, ranging from collapsed duplications to missing sequences, have remained unresolved for decades. The assumptions that most genomic analyses make about the correctness of the reference genome have contributed to spurious clinical findings and mistaken disease associations (Gürünlüoğlu et al., 2020; Khalilipour et al., 2018; Lalrohlui et al., 2021; Munchel et al., 2015). Here, we identify variation in difficult-

to-resolve regions and show that the T2T-CHM13 reference genome universally improves genomic analyses for all populations by correcting major structural defects and adding sequences that were absent from GRCh38. In particular, we show that the T2T-CHM13 assembly (1) revealed millions of additional variants and the existence of additional copies of medically relevant genes (e.g., *KCNJ17*) within the 240 Mbp and 189 Mbp of non-syntenic and previously unresolved sequence, respectively; (2) eliminated tens of thousands of spurious variants and incorrect genotypes per samples, including within medically relevant genes (e.g., *KCNJ18*) by expanding 203 loci (8.04 Mbp) that were collapsed in GRCh38; (3) improved genotyping by eliminating 12 loci (1.2 Mbp) that were duplicated in GRCh38; and (4) yielded more comprehensive SV calling genome-wide, with an improved insertion/deletion balance, by correcting collapsed tandem repeats. Overall, the T2T-CHM13 assembly reduced false positive and false negative SNVs from short and long reads by as much as 12-fold in challenging, medically relevant genes. The T2T-CHM13 reference also accurately represents the haplotype structure of human genomes, eliminating 1,390 artificial recombinant haplotypes in GRCh38 that occurred as artifacts of BAC clone boundaries. These improvements will broadly enable future discoveries and refine analyses across all of human genetics and genomics.

Given these advances, we advocate for a rapid transition to the T2T-CHM13 genome as a reference. While we appreciate that transitioning institutional databases, pipelines, and clinical knowledge from GRCh38 to T2T-CHM13 will require substantial bioinformatics and clinical effort, we provide several resources to advance this goal. On a practical level, improvements to large genomic regions, such as entire p-arms of the acrocentric chromosomes, and the discovery of clinically relevant genes and disease-causing variants justify the labor and cost required to incorporate T2T-CHM13 into basic science and clinical genomic studies. On a technical level, T2T-CHM13 simplifies genome analysis and interpretation because it consists of 23 complete linear sequences and is free of “patch”, unplaced, or unlocalized sequences. Many of the corrections introduced by T2T-CHM13 were previously noted and addressed by the GRC as ‘fix patches’, but few studies use these existing resources. The reduced contig set of T2T-CHM13 also facilitates interpretation and is directly compatible with the most commonly used analysis tools. To promote this transition, we provide variant calls and several other annotations for the T2T-CHM13 genome within the UCSC Genome Browser and the NHGRI AnVIL as a resource for the human genomics and medical communities.

Finally, our work underscores the need for additional T2T genomes. Most urgently, the CHM13 genome lacks a Y chromosome, so our analysis relied on the incomplete representation of Chromosome Y from GRCh38. A T2T

representation of the Y chromosome should further improve mapping and variant analysis, especially with respect to variants on the Y chromosome itself. Furthermore, many of the previously unresolved regions in T2T-CHM13 are present in all human genomes and enable variant calling with traditional methods from short and/or long reads. However, many previously unresolved regions identified in the T2T-CHM13 genome exhibit substantial variation within and between populations, including satellite DNA (Altemose et al., 2022) and SDs that are polymorphic in copy number and structure (Vollger, Guitart, et al., 2022). Relatedly, the T2T-CHM13 reference provides a basis for calling millions of variants that were previously hidden, but many of these variants are challenging to resolve accurately with current sequencing technologies and analysis algorithms. Robust variant calling in these regions will require many hundreds or thousands of diverse haplotype-resolved T2T assemblies to construct a pangenome reference, such as the effort now underway by the Human Pangenome Reference Consortium (Miga & Wang, 2021). These assemblies will then motivate further development of methods for discovering, representing, comparing, and interpreting complex variation, as well as benchmarks to evaluate their respective performances (Eizenga et al., 2020; Pritt et al., 2018).

Through our detailed assessment of variant calling across global population samples, our study showcases T2T-CHM13 as a preeminent reference for human genetics. The annotation resources provided herein will help facilitate this transition, expanding knowledge of human genetic diversity by revealing hidden functional variation.

3.5 Methods Summary

Haplotype structure: We examined the impact of the fact that GRCh38 comprises a mosaic of clones derived from multiple donor individuals on its haplotype structure. To this end, we searched for "LD-discordant" SNP pairs, defined as common (>10% minor allele frequency) SNPs that segregate in perfect LD ($R^2 = 1$) in the 1KGP sample, but for which GRCh38 possesses a pair of alleles that are never observed together on a single phased haplotype among 1KGP samples (*i.e.*, alleles in perfect negative LD). We then compared these results to the same analysis applied to each 1KGP sample using a leave-one-out strategy.

Duplication errors: We flagged putatively collapsed duplications as regions >5 kbp containing clusters of heterozygous variants identified from two CHM13 datasets (simulated Illumina-like reads from T2T-CHM13 reference v1.0 including the GRCh38 Y chromosome and PacBio HiFi reads (Vollger, Logsdon, et al., 2019)) mapping against

GRCh38 and T2T-CHM13 references. False duplications were identified as regions, converted to T2T-CHM13 coordinates, with median read-depth copy numbers (Vollger, Guitart, et al., 2022) lower in kmerized GRCh38 compared to kmerized T2T-CHM13 and 88% of SGDP individuals. Alternatively, false duplications were identified as regions >3 kbp with copy numbers greater in kmerized GRCh38 compared to kmerized T2T-CHM13 and 99% of SGDP individuals using a genomewide sliding-window approach.

Liftover of resources from GRCh38 to T2T-CHM13: Using the GATK release 4.1.9 (Van der Auwera & O'Connor, 2020) LiftoverVcf (Picard) tool, we lifted dbSNP build 154 (Sherry et al., 1999), the March 8, 2021 release of Clinvar (Landrum et al., 2018), and GWAS Catalog v1.0 (Buniello et al., 2019) from the GRCh38 assembly to the T2T-CHM13 assembly. Initial liftover was done with default LiftoverVcf parameters. A secondary round of liftover was performed to recover variants with swapped reference and alternative alleles between GRCh38 and T2T-CHM13. We cataloged variants that failed to lift over because they overlap an indel that distinguishes T2T-CHM13 and GRCh38 based on results from dipcall.

Short-read variant calling: To evaluate short-read small-variant calling between GRCh38 and T2T-CHM13, we used the NHGRI AnVIL (Schatz et al., 2021) to align all 3,202 1KGP samples to CHM13 with BWA-MEM (H. Li, 2013) and performed variant calling with GATK HaplotypeCaller (Van der Auwera et al., 2013) using a workflow modeled on the one developed by the NYGC for 1KGP analysis performed on GRCh38 (Byrska-Bishop et al., 2022). As in the NYGC analysis, we recalibrated the variant calls with GATK VariantRecalibrator. We analyzed coverage statistics using samtools and allele frequency using bedtools. To identify Mendelian-discordant variants, we used GATK VariantEval.

Long-read variant calling: To compare long-read mapping and large structural variant (SV) calling between T2T-CHM13 and GRCh38, we aligned HiFi and ONT data from 17 samples of diverse ancestry to each reference with both Winnowmap (C. Jain et al., 2020) and minimap2 (H. Li, 2018) and called SVs with Sniffles (Sedlazeck, Rescheneder, et al., 2018). Variant calls were refined with Iris, and HiFi-derived calls from both aligners were merged with Jasmine (Kirsche et al., 2021); the resulting sets of 124,566 SVs in GRCh38 and 141,193 SVs in CHM13 to compute allele frequencies and other cohort-level statistics. In addition, we constructed trio-level callsets for two trios - the HG002

and HG005 trios from the Genome-in-a-Bottle Consortium - to compare Mendelian discordance rates between the two references.

Concordance of variants analysis across sequencing type: To evaluate the variant calls in non-syntenic regions, we derived concordance between variant calls generated with HiFi, ONT, and Illumina reads. For each sample, we used bcftools to filter the non-PASS variants, indels, and non-autosomal variants from each callset. We then used hap.py (Krusche et al., 2019) to derive the precision, recall, and F1-score between each variant call set to determine how many variants are common between each pair of sets.

Allele frequency differentiation of non-syntenic variants: Using short-read-based variant calls within T2T-CHM13 non-syntenic regions, we searched for variants with signatures of extreme allele frequency (AF) differentiation across human populations. We performed this analysis with Ohana (J. Y. Cheng et al., 2021), a method that infers admixture components for each sample and quantifies frequency variation among the components. For outlier non-syntenic variants with extreme patterns of AF differentiation, we used liftover to compare our results to previous results generated with 1KGP Phase 3 data aligned to GRCh38 (Yan et al., 2021).

T2T-CHM13 dipcall and VEP: VEP (McLaren et al., 2016) (version 102.0) was used to annotate variants generated by dipcall (H. Li et al., 2018) when aligning the T2T-CHM13 reference genome (chm13_v1.0_plus38Y.fa) to the GRCh38 reference genome (hg38.no_alt.fa). VCF files were annotated without the --filter_common and --canonical flags. CADD (Rentzsch et al., 2021) v1.6 and raw SpliceAI (Jaganathan et al., 2019) scores were added using both the CADD and SpliceAI plugins. Variants were filtered based on predicted HIGH functional impact.

HG002 medically-relevant genes benchmark: To evaluate variant call accuracy when using T2T-CHM13 vs. GRCh38 as a reference, we developed equivalent small variant benchmarks for GIAB sample HG002 in 269 challenging, medically relevant genes. Methods were adapted from a companion manuscript that describes a curated benchmark for these genes created by using variants generated by dipcall (H. Li et al., 2018) when aligning a trio-based hifiasm assembly to GRCh37 and GRCh38 (Wagner et al., 2021).

3.6 Selected Methods

Complete methods are available as Supplementary Material in Aganezov et al. (Aganezov et al., 2022).

3.6.1 Identification of collapsed duplications in GRCh38

We simulated Illumina-like reads (400 million PE 150 bp reads) from T2T-CHM13 reference v1.0 including the GRCh38 Y chromosome using Mason (https://github.com/seqan/seqan/tree/master/apps/mason2) and aligned them to GRCh38 (no alt or decoy contigs) and T2T-CHM13 v1.0 including the GRCh38 Y chromosome using BWA-MEM (H. Li, 2013) (**Fig. S3.13**). Likewise, previously published CHM13 PacBio HiFi reads (~24X, SRA: SRX5633451) (Vollger, Logsdon, et al., 2019) were aligned to GRCh38 using minimap2 (H. Li, 2018) with the -ax map-pb setting. We called SNVs in both datasets with GATK v4.1.8.1 (Poplin et al., 2018) using minimum MAPQ 30, ploidy 2 and otherwise default parameters. Only PASS variants were used for downstream analyses. Heterozygous variants called by each platform were merged into one multi-sample VCF file with bcftools merge, and the number of heterozygous variants per kbp was calculated using bedtools coverage. For both references, we first defined problematic regions as regions ≥2 kbp with ≥2 heterozygous calls in the CHM13 sample. From this, we connected regions separated by ≤5 kbp, and then filtered for regions ≥5 kbp in size.

Focusing on GRCh38-derived problematic regions, we intersected them with previously published RepeatMasker and SD annotations obtained from UCSC Table Browser, as well as known GRC issues (Nurk et al., 2022) (ftp://ftp.ncbi.nlm.nih.gov/pub/grc/human/GRC/Issue_Mapping/). For each region, we determined association with SDs (Vollger, Guitart, et al., 2022), centromeres (Altemose et al., 2022), and non-syntenic and previously unresolved regions in T2T-CHM13 reference (Nurk et al., 2022), using combined lifted coordinates obtained from all minimap2 hits and UCSC LiftOver (Hinrichs et al., 2006).

Additionally, we obtained variants flagged with excess of heterozygosity by the gnomAD database (InbreedingCoeff in the FILTER field), defined as variants with an inbreeding coefficient < -0.3, but were not filtered due to low read depth, genotype quality, or minor-allele fraction (Karczewski et al., 2020). Empirical enrichment of variants with excessive heterozygosity within problematic regions was obtained by calculating the number of variants in 10,000 randomly sampled regions of the genome using bedtools shuffle. The empirical p-value was calculated as

(M+1)/(N+1), where M is the number of iterations yielding a number of features greater than observed and N is the number of iterations.

We identified each homologous GRCh38 problematic region in T2T-CHM13 using the following approach: (1) coordinates obtained by UCSC LiftOver using available chain (Nurk et al., 2022) if the size of the lifted region was within 80-120% of the original size, (2) Minimap2 longest hit if the size of the lifted region was within 80-120% of the original size, and closest hit when more than two options were available, and (3) manual selection and curation of remaining coordinates. To assess functional impact, these likely problematic regions were intersected with all gene features in Gencode v35, as well as a curated list of medically relevant genes (Wagner et al., 2021).

Using available read-depth copy number estimates in T2T-CHM13 (Vollger, Guitart, et al., 2022), we obtained the overall copy number of the lifted regions as the median window copy number for a "k-merized" version of GRCh38 and T2T-CHM13 references, as well as 268 individuals from the SGDP dataset (excluding sample LP6005442-DNA_A08). Regions where copy number in GRCh38 was lower than T2T-CHM13 and also nearly all SGDP individuals (allowing for one individual with lower copy number) were considered putative collapsed duplications in the GRCh38 reference. Additionally, we intersected lifted coordinates with T2T-CHM13 SDs (Vollger, Guitart, et al., 2022) and centromere annotations (Altemose et al., 2022).

The same analysis was performed to identify putative collapsed duplications in T2T-CHM13, but without the need to liftover homologous coordinates.

3.6.2 Identification of medically relevant genes with impacted variant discovery

Reference artifacts and errors were intersected with a previously curated list of medically relevant genes (Wagner et al., 2021). GRCh38 coordinates for genes were lifted over using Picard v2.25.0's LiftOverIntervalList tool to identify locations in T2T-CHM13v1.0. For SVs, we intersected gene coordinates after expanding breakpoints for each variant ± 5 kbp.

TNNT3 analysis: Genomic sequencing containing TNNT3 implicated with SVs in GRCh38 were extracted from each reference (chr11:1,892,362-1,946,566, GRCh38; chr11:1,978,257-2,034,355, T2T-CHM13v1.0) and homologous regions compared using miropeats -s 400(Parsons, 1995).

KCNJ18 analysis: SDs containing KCNJ18 were identified in GRCh38 (chr17:21,687,227-21,736,311; CDS: chr17:21,702,787-21,704,088) and T2T-CHM13v1.0 (chr17:21,636,382-21,685,461; CDS: chr17:21,651,942-21,653,243) and paralogs pinpointed using the UCSC Genome Browser SD annotations(Jeffrey A. Bailey et al., 2002, 2001; Numanagic et al., 2018). With KCNJ18, Genome coordinates of SDs containing KCNJ12 (GRCh38: chr17:21,399,778-21,450,415 (whole) and chr17:21,415,343-21,416,644 (CDS); T2T-CHM13v1.0: chr17:21,348,977-21,399,639 (whole) and chr17:21,364,558-21,365,859 (CDS)) and KCNJ17 (T2T-CHM13v1.0: chr17:22,634,421-22,683,415 (whole) and chr17:22,666,582-22,667,880 (CDS)) were used to extract 1KGP AFs from GRCh38-called and T2T-CHM13-called variants intersecting each locus. Histograms of the minor-allele frequencies were plotted and distributions compared using a Mann-Whitney U test (wilcox.test R function, with unpaired and two-sided settings). Finally, direct comparison of AFs for variants falling within the CDS were compared using 1KGP datasets from the NYGC (GRCh38), produced here (described above for T2T-CHM13), and the liftover of T2T-CHM13 to GRCh38 (described above).

3.7 Acknowledgements

We would like to thank M. Zody, B. Grüning, H. Li, S. Langley, C. Langley, G. Van der Auwera, V. Schneider, S. Salzberg, B. Langmead, A. Battle and several of their lab members for helpful discussions. Certain commercial equipment, instruments, or materials are identified to specify adequate experimental conditions or reported results. Such identification does not imply recommendation or endorsement by the National Institute of Standards and Technology, nor does it imply that the equipment, instruments, or materials identified are necessarily the best available for the purpose. This work utilized the computational resources of the NIH HPC Biowulf cluster (https://hpc.nih.gov) and the Maryland Advanced Research Computing Center (https://www.marcc.jhu.edu/).

CHAPTER 4. Population diversity and selection of recent gene duplications detected using a complete human genome sequence

Chapter 4 is an ongoing research project led by co-first authors Daniela C. Soto and Aarthi Sekar.

DCS performed genome-wide analysis of nearly-identical duplicated genes, including identification, expression and ontology analysis, copy-number variation and copy-number stratification, and benchmarked SNV-calling performance of long-reads.

4.1 Abstract

Human-specific duplicated genes (HSDs) are strong candidates for neurodevelopmental traits and diseases unique to our species. Assessment of the recently published telomere-to-telomere (T2T) complete genome (T2T-CHM13) identified 417 genes embedded in recent segmental duplications (>98% sequence identity; SD-98) with near fixed copy-number (CN=2) in the 1000 Genomes Project (1KGP), with 347 having evidence of expression in the fetal neocortex. This list includes genes with known roles in neurodevelopment, such as *ARHGAP11B*, as well as a number of other uncharacterized genes. Comparing CN across 1KGP populations, we also identified 205 genes that show stratification (V_{ST}>95th percentile), including those with previous evidence (e.g., *KANSL1*) and interesting new candidates (e.g., *NPY4R*, previously implicated in body-mass index, and *TPTE,* a gene expressed exclusively in testis). Examining our ability to detect SNVs and indels across SD-98, we compared short- and long-read sequence data from eight individuals from the Human Pangenome Reference Consortium (HPRC+) and observed a recall of <0.1 resulting in a depletion of total variants identified from 1KGP across these regions (12 variant/kbp) versus the unduplicated genome (38 variant/kbp). Using 1KGP variants in accessible regions (representing ~10% of SD-98), we assessed signatures of natural selection by calculating Tajima's D and identified 22 protein-encoding genes showing consistent outlier values, including the *SPDYE3* and *PMS2P1* locus, which are CN fixed. Our approach highlights potential evolutionarily relevant human gene duplications, which will become priority candidates for future functional studies.

4.2 Introduction

Significant phenotypic features distinguish modern humans from closely related great apes. Examples of anatomical, social, physiological, and behavioral traits that distinguish humans from their closest primate relatives include small canine teeth, reduced hair cover, elongated thumbs, language, bipedalism, and advanced tool usage (Carroll, 2003; Pääbo, 2014; Ajit Varki & Altheide, 2005). Perhaps one of the most well-studied innovations in modern humans relates to changes in neuroanatomy (e.g., expanded neocortex and greater complexity of neural connectome), which has led to the acquisition of novel cognitive features such as reading and language (Sousa et al., 2017). A number of genes have been implicated as contributing to such traits, with a remarkable proportion representing duplicated genes (Sudmant et al., 2013). Examples include *SRGAP2* (Charrier et al., 2012; Dennis et al., 2012), *NOTCH2NLC* (Fiddes et al., 2018; Florio et al., 2018; Suzuki et al., 2018), *ARHGAP11B* (Florio et al., 2015, 2016; Namba et al., 2020), and *TBC1D3* (Ju et al., 2016). This may be unsurprising considering segmental duplications (SDs; genomic regions >1 kbp in length that share high sequence identity [>90%]) account for greater genetic divergence across species and diversity across humans compared with single-nucleotide variants (SNVs) (Sudmant et al., 2013) and are enriched for transcribed genes (Jeffrey A. Bailey et al., 2001).

Duplications arising uniquely within the human lineage, or human-specific SDs (HSDs), are of particular interest since they can give rise to new genes with altered functions (Dennis & Eichler, 2016). Recent comparisons of great ape genomes have implicated over 200 HSD genes (Dennis et al., 2017; Sudmant et al., 2013, 2010), uncovering unexpected connections with neurodevelopment. Compared to other species-specific SDs, human duplicated genes are enriched for neurological functions (Sudmant et al., 2010) and tend to reside at genomic hotspots—or regions prone to recurrent deletions and duplications due to errors in meiosis—associated with neurodevelopmental disorders such as autism, epilepsy and intellectual disability (Dennis & Eichler, 2016). As such, HSD genes have high potential for contributing to neural features and disorders unique to humans.

Despite the clear importance of SDs in human traits and disorders, they have remained largely ignored in genome-wide screens due to difficulty in calling SNVs across nearly-identical paralogs (Amemiya et al., 2019; Cabanski et al., 2013; Derrien et al., 2012; Ebbert et al., 2019; H. Lee & Schatz, 2012). This is highlighted in a recent assessment of the 1000 Genomes Project (1KGP; Phase 3) short-read sequence data that found ~45% of SDs in the most recent

human reference assembly (GRCh38) to be inaccessible using standard mappability and coverage filters (Zheng-Bradley et al., 2017). As a result, almost a half of the SDs remain unexplored across millions of sequenced human genomes that could contribute to missing genetic risk and heritability of traits/diseases. Faced with these limitations, most studies characterizing SDs focus on copy-number (CN) variation (Conrad et al., 2010; Redon et al., 2006; Sudmant et al., 2010) or GWAS-associated haplotypes in linkage disequilibrium (LD) with SD regions (Beyter et al., 2021; Conrad et al., 2010; Hehir-Kwa et al., 2016; Marie Saitou et al., 2021; Yan et al., 2021; Zhao et al., 2017).

More recently, the new gapless T2T-CHM13 genome has enabled a more complete picture of SDs (Nurk et al., 2022; Vollger, Guitart, et al., 2022), which are historically missing or incorrect in current references due to high sequence identities, by expanding by 238 Mbp of sequence that was previously missing from GRCh38, of which ~45 Mbp comprises non-satellite SDs. From this, studies have begun to delve into unique genomic features of SDs, including increased incidence of interlocus gene conversion across duplicate paralogs and an overall higher mutation rate compared with adjacent "unique" regions (Vollger, DeWitt, et al., 2022). In particular, the new assembly corrects >8 Mbp of collapsed duplications (Aganezov et al., 2022), including 48 known protein-encoding genes, leading to missing copies of likely functional paralogs of *GPRIN2* (Vollger, Guitart, et al., 2022) and *DUSP22* (Aganezov et al., 2022), both previously identified as HSDs (Dennis et al., 2017). Results of variant discovery comparisons between Illumina and HiFi across collapsed and expanded SD regions suggests that, although sensitivity is high for a subset of variants within these regions, we are still missing a majority of SNVs using short-read data.

Here, using the new T2T-CHM13 genome, we assess if new genome-wide insights can be made with this complete genome. First, we show the extent by which SDs have been overlooked in human genetic analyses. We aimed to bridge this gap by using a combination of paralog-specific CN estimates and high-confidence short-read variants to identify signatures of selection in recent human gene duplications, shedding light on putative functional genes playing a role in human evolution.

4.3 Results

4.3.1 Identification of human gene duplications

The recently published T2T-CHM13 genome represents a complete sequence of all autosomes and chromosome X (Nurk et al., 2022), resolving all previously-collapsed duplications and gap sequences (Aganezov et al., 2022; Vollger,

Guitart, et al., 2022). Nearly-identical SDs enriched for human-specific duplications—here defined as SDs with >98% of sequence identity (SD-98)—encompass 97.8 Mbp of autosomal sequence (**Fig. S4.1A**). These regions overlap with 5,433 gene features (T2T-CHM13 gene IDs), including 885 protein-encoding genes and 1,157 unprocessed pseudogenes arising from duplication events (**Table S4.1**, **Fig. S4.1B**). Out of this, 4,746 (including 515 protein-encoding genes and 1,069 unprocessed pseudogenes) were fully contained within a recent duplication (≥99% covered). We note that this approach more permissively permits inclusion of both human unique (Dennis et al., 2017; Sudmant et al., 2010) and recently expanded duplications that may also exist with paralogs in other great apes, and previously characterized genes known to play a role in neurodevelopment (e.g., *ARHGAP11B* (Florio et al., 2015), *SRGAP2C* (Charrier et al., 2012; Dennis et al., 2012), and *NOTCH2NL* (Fiddes et al., 2018; Suzuki et al., 2018)), disease (e.g., *SMN1* and *SMN2* (Larson et al., 2015) and *KANSL1* (Moreno-Igoa et al., 2015)), and adaptation (e.g., amylase genes (Perry et al., 2007)).

To further examine the functions of SD-98 genes, we identified overrepresented gene ontology (GO) terms, finding a significant enrichment (FDR≤0.1) of genes associated with immune response (GO:0006955, GO:0002250, GO:0002377), Golgi organization (GO:0007030), protein degradation (GO:0006511, GO:0016579), and regulation of cell differentiation (GO:0045596) (**Fig. S4.1C** and **Fig. S4.1D**). Remapping expression data from human fetal brain (Fietz et al., 2012; Florio et al., 2015) and lymphoblastoid cell lines (LCLs) (Pickrell et al., 2010) to a T2T-CHM13 transcriptome reference, we found evidence of expression (mean TPM ≥1) in ~45% (2,435/5,433) of SD-98 genes in at least one dataset, including 601 protein-encoding genes (**Table S4.1**). In particular, ~43% (2,358/5,433) and ~16% (872/5,433) of SD-98 genes were expressed in at least one brain dataset or LCLs, respectively. As expected, protein-encoding genes are significantly overrepresented among expressed genes, with 601 out of 885 protein-encoding genes showing expression (hypergeometric test, p-value<$1x10^{-4}$).

Expression similarity among datasets was primarily driven by study rather than tissue or cell type (**Fig. S4.2B**). Nonetheless, we identified 486 (~9%) genes expressed in all examined tissues and cell populations, including LCL and brain samples (**Fig. S4.2A**). Conversely, 127 (~2%) SD-98 genes had evidence of expression in all brain tissues but not LCL, suggesting a specific role in brain development. In fact, this list is significantly enriched for genes associated with neurogenesis, including neuron differentiation (GO:0021953, GO:0021879, GO:0021889,

GO:0010769), projection (GO:0097485, GO:0050770, GO:0007411), migration (GO:0001764, GO:2001222) and generation (GO:0021872).

4.3.2 Phenotype and disease associations of duplicated genes

Due to difficulties mapping short reads to highly identical regions, associated variants and genes across SD-98 regions are depleted in existing genome-wide (GW) studies of phenotypes and diseases, including GWAS catalog (SD-98: 0.29 variants/100kbp; GW: 1.5 variants/100 kbp), ClinVar (SD-98: 20.81 variants/100 kbp; GW: 9.95 variants/100 kbp), and GTEx expression quantitative trait loci (eQTL) databases (SD-98: 398.7 variants/100 kbp; GW: 70.14 variants/100 kbp) (**Fig. S4.3A**).

Nevertheless, we sought to understand if any connections with disease and/or traits of recently duplicated genes existed in databases. SD-98 regions overlap 60 genes containing variants associated with a trait in genome-wide association studies (GWAS catalog v1.0) (**Table S4.1)**. Using the probability of loss-of-function intolerance (pLI) (Lek et al., 2016) and the loss-of-function observed/expected upper fraction (LOEUF) (Karczewski et al., 2020), we identified 78 genes intolerant to loss-of-function overlapping with SD-98 regions (pLI ≥ 0.9 or LOEUF < 0.35, based on (Leblond et al., 2021)), with 69 flagged by both approaches (**Table S4.1**, **Fig. S4.4**). According to an operative list of genes implicated in neurodevelopmental disorders (NDD) (Leblond et al., 2021), 44 SD-98 genes are classified as high-confidence NDD genes, including haploinsufficient genes *KANSL1* and *MEF2C*. Most of these genes display a small overlap with SD-98 regions (mean gene overlap of ~10%), except for *DDX11* which is fully duplicated (**Table S4.1)**. In all, 486 protein-encoding genes had no phenotype association in the surveyed gene-disease databases, including survival of motor neuron genes *SMN1* and *SMN2* implicated in spinal muscular atrophy (Larson et al., 2015), suggesting underassessment of disease/trait association of recently duplicated genes.

4.3.3 CN diversity of human duplicated genes

We used a paralog-specific approach based on read depth at unique *k*-mers (Shen & Kidd, 2020) to characterize the CN diversity of recent duplicates in modern humans. We obtained CN estimates in 2,504 unrelated individuals from the 1KGP for SD-98 regions overlapping autosomal protein-encoding genes (n=917 regions) and unprocessed pseudogenes (n=1,177 regions), as functional duplicates are sometimes incorrectly annotated as pseudogenes (**Fig S4.5**, **Table S4.1**, **Table S4.2**). The most frequent CN exhibited was two (median average CN = 1.9), as expected

from paralog-specific CN detection (**Fig. S4.6A**). While both protein-encoding genes and pseudogenes exhibited similar levels of CN variability (average sd = 4.7 and 4.6, respectively), we observed a better correlation between mean CN and standard deviation in protein-encoding genes (r^2=0.31) than in pseudogenes (r^2=0.05), mostly driven by high CN and highly-variable protein-encoding genes such as *USP17L11* (**Fig. S4.6B**).

Narrowing in on potentially functional genes, we identified 445 regions comprising 417 genes, including 252 protein-encoding, that were CN fixed in modern humans (≥98% of individuals with CN = 2) (**Fig. 4.1A**). Of these fixed genes, 346 (77.8%) exhibit expression in at least one brain tissue assayed (**Table S4.1**). This dataset is enriched for GO terms associated with GTPase activity (GO:0043547, GO:0090630), Golgi organization (Golgi organization), protein modification (GO:0016579, GO:0000413), and protein import to nucleus (GO:0006607) (**Fig. S4.6C**). Out of these, 164 genes (43 protein-encoding) represented complete duplications (**Fig. S4.6B**), such as the NDD-associated gene *DDX11* (CN=2 in 99.84% of 1KGP genomes). We noticed that only 49 out of 252 CN fixed protein-encoding genes were classified as loss-of-function intolerant by pLI or LOEUF scores. However, lack of CN variation at these loci suggests that some of these genes are functional in modern humans.

Of these constrained genes, 317 also exist at CN=2 among four archaic genomes, one Denisova and three Neanderthal genomes (**Fig. 4.1A**). The list contains several genes identified as human-specific duplications dated before the divergence of *Homo sapiens* and *Homo neanderthalis*, such as paralogs *GPR89B*, *PDZK1P*, *HYDIN2*, *CD8B2*, and *ARHGAP11B* (Dennis et al., 2017), the latter implicated in the expansion of the human neocortex (Florio et al., 2015, 2016; Namba et al., 2020). Of these, 261 are expressed in at least one brain tissue and 22 are known to be implicated in NDD, such as *DDX11*, *DPP6*, and *KATNAL2* (**Table S4.1**). Conversely, we identified two adjacent genes in chromosome 10q11.22, *AGAP4* and *PARGP1*, exhibiting near fixation in modern humans (CN = 2 in ≥90%) but seemingly absent from all archaic genomes surveyed. Interestingly, *AGAP4* paralogs, *AGAP5* and *AGAP6*, are CN fixed in both modern and archaic genomes. Additionally, we identified 242 genes that do not display CN losses in any individual, including haploinsufficient gene *KANSL1*.

Considering SDs are enriched for CN polymorphism, we next sought to characterize population differences in modern human using the statistic V_{ST} (Redon et al., 2006) to identify pairwise population stratification in four continental superpopulations (European [EUR], East Asian [EAS], South Asian [SAS], and American [AMR]) with respect to

Africans (AFR). As population differences are less impacted by haplotype-disruptive recurrence, the genes displaying high V_{ST} values are candidates for adaptive CNVs. V_{ST} values exhibited different standard deviation per superpopulation, with AFR-EAS having the most dispersion (AFR-EUR sd=0.1; AFR-EAS sd=0.12; AFR-SAS sd=0.088; AFR-AMR sd=0.09) (**Fig. 4.1B, Fig. S4.7)**. We identified a total of 224 SD-98 regions (205 genes) above the 95th percentile per pairwise comparison, including 66 protein-encoding and 139 pseudogenes (**Fig. 4.1A**).

While CN differences between population-stratified regions were small (median CN difference across all stratified regions: 0.77), possibly related to smaller CNVs within genotyped regions, a subset of the stratified regions was affected by sizable CN differences that suggest extra or missing copies of the genotyped region in a population. Among these, the gene *TPTE* exhibits a significant copy-number decrease in Africans compared to all other superpopulations, suggesting either an African-specific deletion or an out-of-Africa duplication (**Fig. 4.1B** and **Fig4.1C**). Interestingly, the Peruvian population (PEL) lacks copy numbers lower than one at this locus. On the other hand, the CN distribution of protein-encoding gene *TBC1D3I* is compatible with an out-of-African copy-number lost, with a mean of two copies in African populations and one in all non-African populations, including three V_{ST} pairwise comparisons (AFR-EUR, AFR-SAS, AFR-EAS) within the 95th percentile (**Fig. S4.8**).

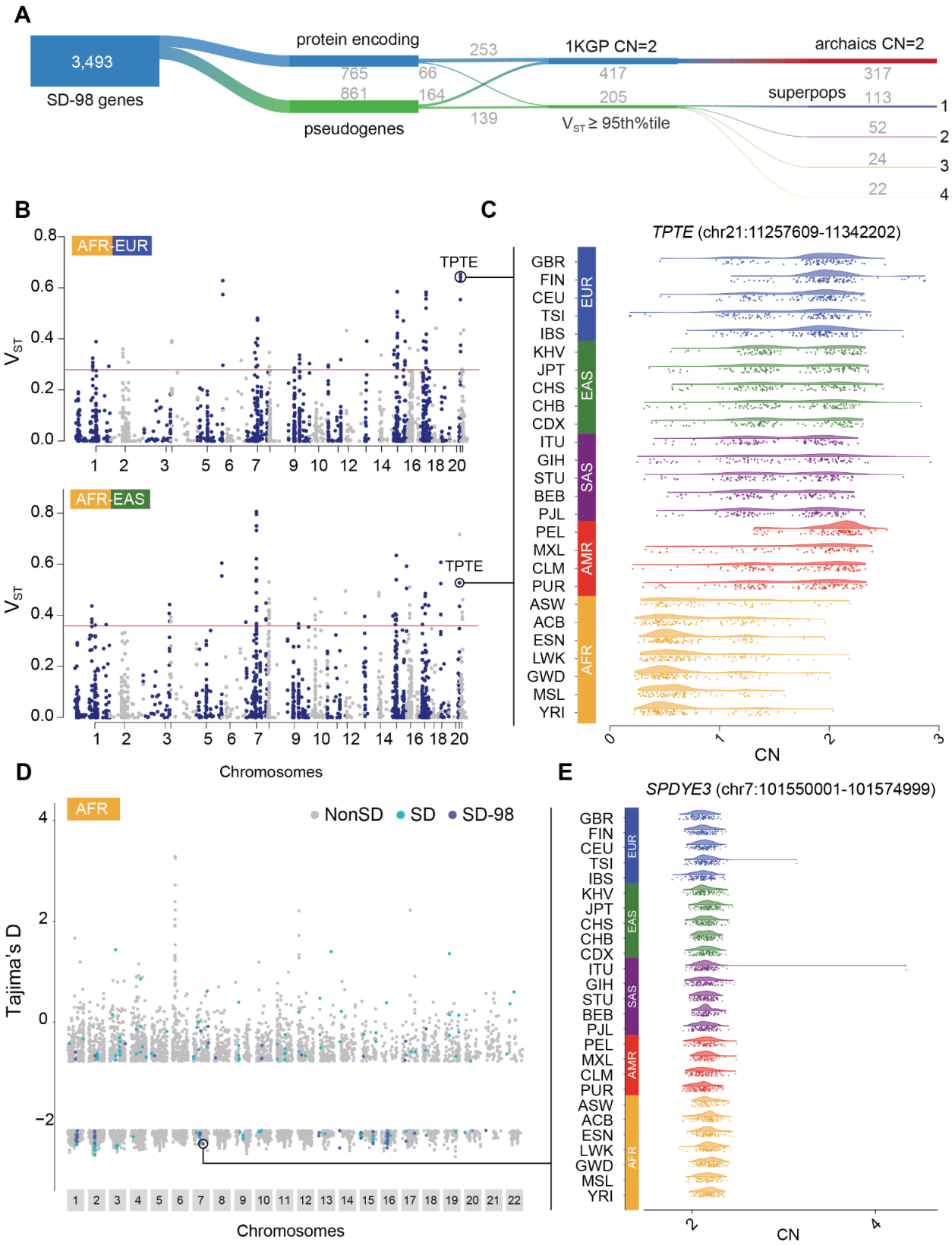

A
3,493
SD-98 genes
protein encoding
765
861
pseudogenes
253
66
164
139
1KGP CN=2
417
205
V_{ST} ≥ 95th%tile
archaics CN=2
317
superpops
113
52
24
22
1
2
3
4
B
AFR-EUR
AFR-EAS
TPTE
V_{ST}
0.0
0.2
0.4
0.6
0.8
1 2 3 5 6 7 9 10 12 14 16 18 20
Chromosomes
C
TPTE (chr21:11257609-11342202)
GBR
FIN
CEU
TSI
IBS
KHV
JPT
CHS
CHB
CDX
ITU
GIH
STU
BEB
PJL
PEL
MXL
CLM
PUR
ASW
ACB
ESN
LWK
GWD
MSL
YRI
EUR
EAS
SAS
AMR
AFR
0 1 2 3
CN
D
AFR
NonSD
SD
SD-98
Tajima's D
4
2
0
−2
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22
Chromosomes
E
SPDYE3 (chr7:101550001-101574999)
2 4
CN

Figure 4.1. Selection signatures in human duplicated genes. (A) Counts of genes (ENSEMBL gene IDs) genotyped by gene biotype, CN fixation in modern humans or archaic genomes, and number of superpopulation displaying stratification ($V_{ST} \geq$ 95th percentile). (B) Distribution of V_{ST} values per chromosome for AFR-EUR and AFR-EAS pairwise comparison (other pairwise comparison in **Fig. S4.7**). Red line indicates the 95th percentile of V_{ST} distribution. (C) Copy-number distribution of *TPTE* gene across all populations assayed, colored by superpopulation. Coordinates correspond to the CN-genotyped region. (D) Distribution of Tajima's D values in 25-kbp windows for the African superpopulation with colors highlighting regions as stated in the legend. (E) CN of *SPDYE3* gene. Coordinates correspond to Tajima's D window overlapping the gene.

Surveying genes with known or putative evolutionary relevance, we observed a significant copy-number increase in the gene *KANSL1* in Europeans in agreement with a previously identified inversion-duplication exhibiting signatures of positive selection in Europeans associated with increased fertility (Stefansson et al., 2005). Interestingly, South Asians populations exhibited a similar distribution than Europeans and were also stratified compared to Africans (**Fig. S4.9)**. Analysis of the copy number of gene *NOTCH2NLR* shows that a significant proportion of individuals (~21%) lacks the gene (CN < 0.5), as described before (Fiddes et al., 2018). Nonetheless, we identified a significant copy-number stratification in both Europeans and South Asians, where there is a higher proportion of copy-number two individuals than in African populations (**Fig. S4.10)**. A human-specific paralog, *NPY4R2*, shows signatures of population stratification exclusively in East Asians, exhibiting a decrease in copy-number that suggests a partial copy-number loss in the genotyped region (**Fig. S4.11)**. Concomitantly, the ancestral paralog *NPY4R*, which is spanned by a CNV implicated in body-mass index (Shebanits et al., 2018), displays a slight increase in copy-number in East Asians (although not significant).

4.3.4 SNV discovery and diversity across duplicated regions

Despite improved representation of duplicated genes in T2T-CHM13, genomic assessment of these regions remains challenging using short-read Illumina data. Duplicated genes are significantly depleted for SNVs in the 1KGP (The 1000 Genomes Project Consortium, 2015) using T2T-CHM13 (SD-98: 11.79 SNVs/kbp; GW: 37.49 SNVs/kbp) (**Fig. S4.3B**). The autosomal 2.4 Gbp in T2T-CHM13 accessible for accurate Illumina SNV calling (Aganezov et al., 2022)—determined using read depth, mapping quality, and base quality metrics—includes only 37.95% and 10.86% of SD and SD-98, respectively (**Table 4.1**).

Table 4.1: Autosomal regions accessible to short reads in T2T-CHM13 (v1.0).

Region	Total size (bp)	Accessible (bp)	Accessible (%)	Non-accessible (bp)	Non-accessible (%)
Non-SD	2,550,809,796	2,439,661,620	95.643	111,148,176	4.36
SD	179,849,378	68,258,070	37.953	111,591,308	62.05
SD-98	97,797,568	10,623,540	10.863	87,174,028	89.14
CenSat	198,056,319	15,285,809	7.718	182,770,510	92.28

CenSat: centromeric satellites. Non-SD excludes SDs and CenSat regions.

To evaluate our ability to detect variants within duplications, we compared SNVs discovered using Illumina short-read and PacBio HiFi long-read data across eight 1KGP individuals included in the Human Pangenome Reference Consortium (HPRC+) (Aganezov et al., 2022; Wang et al., 2022). Overall, more variants were discovered in all genomic-region categories using HiFi compared with Illumina data (**Table S4.4**). While no differences in density (SNV sites within 1-kbp non-overlapping windows) existed between data types in non-duplicated and T2T-CHM13 accessible regions (Aganezov et al., 2022), respectively, we observed reduced mean variant density from short-read (SD: 1; SD-98: 0) versus long-read data in duplicated regions (SD: 5; SD-98: 5) (**Fig. S4.12**).

Using HiFi-discovered variants as truth, we next assessed variant accuracy and found that 99.5% of SNVs matched between technologies in non-SD, which decreased to 88.6% and 81.7 % in SD and SD-98, respectively. When considering only T2T-CHM13 short-read accessible regions, SNV precision increased in the three regions assayed to 99.7%, 96.1%, and 94.2% for non-SD, SD, and SD-98 (**Table S4.4, Fig. S4.13**). On the other hand, sensitivity—measured as the proportion of HiFi-discovered SNVs also detected using Illumina data—experienced a pronounced decrease of 24.5% in SD and 0.85% in SD-98 compared to 87.6% in Non-SD regions. When considering only T2T-CHM13 short-read accessible regions, however, sensitivity improved to 91.7%, 72.5%, and 57.8%, respectively. Overall, these results indicate that existing variants identified across duplicated regions from Illumina data are generally accurate, particularly in defined accessible regions, but not comprehensive.

Leveraging our ability to assay variation across SD accessible regions, we calculated Tajima's D (Tajima, 1989) using T2T-CHM13 variant data for 2,504 unrelated individuals from the 1KGP representing five superpopulations (Aganezov et al., 2022) and compared the distribution of D statistics across 25-kbp windows intersecting accessible

regions (**Fig. 4.1D**). Recent studies have highlighted potential discrepancies of evolutionary constraints experienced between duplicated and non-duplicated genomic spaces (Hartasánchez et al., 2018). As such, we also conservatively identified outlier thresholds for each superpopulation considering only SD-98 windows instead of genome-wide (total windows in 95th percentile: 35 and 5th percentile: 27) (**Table S4.5**). GO analysis of the 15 protein-encoding genes with significantly low D values, indicative of putative directional selection, showed enrichment of genes involved in immunoglobulin complex and membrane related functions, while genes with significantly high D values, indicative of possible balancing selection, were enriched for transcription regulation processes. This higher-confident set of windows overlapped with 94 genes (22 protein encoding), including two genes in chromosome 7q22.1, *SPDYE3* and *PMS2P1*, with significantly lower D values across all five superpopulations, signature of directional selection acting on this locus or recent population expansion. Interestingly, these genes were also CN fixed in modern humans and archaic genomes, except for the same two individuals that have extra copies of the locus (**Table S4.1, Fig. 4.1E**). We also identified seven additional fixed protein encoding genes displaying significantly low Tajima's D values, including *RHCE*, *TIMM23B*, *MRPL45*, *LRRC37A3*, *CDH12*, *POM121C*, and *CNTNAP3*, as well as a readthrough transcript between *TIMM23B* and *AGAP6* genes. While precision of SNV detection in Tajima's D outliers regions was high (mean precision = 0.95), sensitivity was low (mean sensitivity = 0.61), indicating that to better characterize these duplicated loci, long-read sequencing variant detection is necessary.

4.4 Discussion

Most genetic analyses and selection scans using the human genome have systematically skipped SDs due to the difficulties in assaying these regions with short reads. As Illumina short-read sequencing has remained the most widely used sequencing platform, SDs are often excluded from genome-wide analyses. This includes the absence of population-level genomic variants within SDs, as well as their misrepresentation in human reference genomes. Recently, the new complete T2T-CHM13 genome fixed all the errors in SD space. Leveraging this new resource, we first examined evolutionarily recent SDs sharing 98% sequence identity with another homolog elsewhere in the genome. These duplications are enriched for human-specific expansions that likely occurred after divergence from a recent common ancestor of humans and chimpanzees around 6 mya. Unlike previous studies that focused on human-specific duplications of previously single-copy genes (Dennis et al., 2017), this approach also incorporates expansions of already existing duplicate gene families. Notably, we do not differentiate between conserved ancestral and human-

specific derived paralogs, which requires comparisons of synteny with primate orthologs that remains challenging without comparable T2T great ape genomes. While the function of most duplicated paralogs remains unknown, some remarkable examples show gene expansions, such as *NOTCH2NL* (Fiddes et al., 2018; Florio et al., 2018; Suzuki et al., 2018) and *TBC1D3* (Ju et al., 2016), and gene duplications, such as *SRGAP2C* (Charrier et al., 2012; Dennis et al., 2012), and *ARHGAP11* (Florio et al., 2015), to be implicated in the expansion of the neocortex during human evolution. Importantly, not all the duplicate gene families identified from our analysis are human specific. Despite their high sequence identity (>98%), some gene paralogs might appear "younger" due to the action of interlocus gene conversion, which leads to the concerted evolution of the duplicated paralogs. The full extent to which gene duplications are impacted by interlocus gene conversion remains underassayed, however, a lower bound indicates that it accounts for at least 2.7% of SNVs within SDs (Dumont, 2015). More recent estimates based on phased haplotypes propose that ~33.8% of SDs show evidence of interlocus gene conversion in at least one individual assayed (n=102) (Vollger, DeWitt, et al., 2022). Alternatively, some paralogs may remain similar, at least at the coding level, if they are functional and evolving under selective constraints, or due to random drift leading to incomplete lineage sorting.

As SD are prone to CNVs via non-allelic homologous recombination, selection of duplicated genes act on the number of copies of a gene rather than single-nucleotide differences. Some emblematic cases of adaptive SVs have been reported in the human lineage, including a positive correlation between the copy number of salivary amylase gene *AMY1* and starch-rich diets (Pajic et al., 2019), and the 'runaway duplication' of the *HPR* gene, which confers defense against trypanosome infection, whose CN correlates with the geographic distribution of infections (Almarri et al., 2020; Handsaker et al., 2015; Hardwick et al., 2014). Assessment of CN of duplicated genes has been done previously at the gene family level by using read-depth of multimapping short reads (Dennis et al., 2017; Hsieh et al., 2021). This approach is robust as it does not depend on the correct assignment of short reads to their respective paralogs. To achieve paralog specificity, however, short reads need to be mapped to regions displaying nucleotide differences that distinguish each paralog unequivocally. This approach is computationally implemented as CN estimation using read-depth at unique *k*-mers and is sensitive to the accuracy of the reference genome, as missing paralogs will result in inaccurate CN estimates. Previous studies have identified paralog-specific CNs in human reference GRCh37/hg19 (Sudmant et al., 2010) and GRCh38/hg38 (Dennis et al., 2017; Shen & Kidd, 2020). Here we leveraged a published tool, QuicK-mer2 (Shen & Kidd, 2020), to obtain CN estimates in windows of 500 unique *k*-mers using a complete

human reference genome, T2T-CHM13. The resolution of the approach, however, depends on the number of unique *k*-mers at a certain locus, which are often scarce in the most recent duplications.

As we were interested in larger duplication events rather than small structural variants, we genotyped the median CN of SD-98 regions overlapping genes. Only if a gene was fully covered by a SD-98 region, our approach captured the overall CN of the gene. Otherwise, we genotyped a smaller CNV within the gene. While the median is more resilient to outliers that other summary metrics, CN at certain loci remained noisy. In some genotyped regions, we noticed a shift in the CN distribution, averaging above or below the expected CN two for our diploid genomes. For V_{ST} analyses, however, we used the raw CNs, which allowed us to robustly detect differences between populations regardless of distribution shifts. As such, this approach yielded stratification signatures of both full copy-number gains and losses as well as smaller CN differences, which needed to be considered when analyzing putative signatures of population stratification.

In agreement with previous findings (Aganezov et al., 2022), we found that small variants detected with short reads in duplicated regions are accurate when benchmarked against variants detected with high-fidelity long reads, but are not comprehensive enough to fully characterize the variation landscape of SDs. Nonetheless, we aimed to leverage the accuracy of identified variants to assess departure from neutrality in SD regions using the Tajima's D statistic, which is sensitive to genotyping accuracy but less so to variant recall. To avoid confounding results associated with low variant recall, we used only short-read accessible regions and discarded windows with less than five SNVs for D calculations. Notably, to the best of our knowledge, this is one of the few attempts to recover high-quality short-read SNVs in SDs for a selection scan, as most of these regions are excluded upfront. To establish Tajima's D outliers, we conservatively used an empirical distribution based on SD-98 windows, avoiding cross comparison between duplicated and unduplicated regions of the genome. Recent studies have highlighted duplication-specific molecular mechanisms that lead to differential mutational rate and impact their site frequency spectra (SFS). In particular, SDs are shown to have a higher mutational rate, partially explained by the action of IGC (Vollger, DeWitt, et al., 2022). Additionally, computational simulations of duplicated regions evolving under different rates of IGC have shown statistically significant differences in SFS-based tests between single-copy and duplicated genes, suggesting they should be characterized separately (Hartasánchez et al., 2018).

Our analyses of genetic databases and expression data narrow in on a subset of human duplicated genes with the propensity to be functional and, as such, could be relevant for human evolution. We found almost half of genes within SD-98 regions showing expression in LCL and brain tissues. While promising, to more comprehensively catalog putative gene functions requires assaying additional cell lines and/or tissues. Another line of evidence highlighting putatively functional duplicated genes is CN fixation in modern humans. Our paralog-specific approach allowed us to investigate this question by analyzing the CN of ancestral andderived paralogs independentlyThis is relevant, as evolutionary theory suggests that after the duplication event the new paralog would often go the route of pseudogenization (Lynch & Conery, 2000), which translates into missing gene copies in non-disease cohorts and CN variation among modern humans. Nonetheless, in our CN analysis we purposely included unprocessed pseudogenes arising from duplication events in addition to protein-encoding, as pseudogenes annotations in the reference genome are often based on the acquisition of predicted loss-of-function mutations in the duplicated copy and not functional assessment. In fact, evidence suggests that truncated gene copies can have a relevant function in human evolution, as it is the case of *SRGAP2C* (Charrier et al., 2012; Dennis et al., 2012). Importantly, only a fifth of CN fixed genes were also classified as intolerant to loss-of-function mutations using the pLI and LOEUF scores. A likely scenario is that human duplicated genes are underassayed in these scores as they rely on accurate identification of SNPs and indels that induce a stop codon, frameshift, or splice-site disruption derived from short-read sequence data.

To further explore functional human duplicated genes, we searched for selection signatures associated with CNVs using V_{ST} to detect significant CN population differences. As population stratification is less sensitive than other selection signatures to loss of linkage disequilibrium, it has been widely used to detect adaptive structural and CN variants, often impacted by haplotype-disruptive recurrence caused by NAHR (Marie Saitou & Gokcumen, 2019b). Thus, genes exhibiting outlier V_{ST} values are strong candidates for duplicated genes involved in local adaptation.

Our approach highlighted several duplicated genes with interesting CN properties and selection signatures. The gene *DDX11,* which is fully encompassed within SD-98 space, is a known developmental gene associated with chromatin structure and DNA repair implicated in Warsaw breakage syndrome (Santos et al., 2021). Our CN analysis of *DDX11* showed that it is CN fixed in both modern and archaic humans. This is consistent with finding that loss of *DDX11* causes replication stress (Jegadesan & Branzei, 2021). CN estimates also showed that *AGAP4* and a nearby transcribed unprocessed pseudogene, *PARGP1*, were nearly CN fixed in modern humans but absent in archaic genomes,

suggesting a *Homo sapiens* specific expansion of this gene family. *PARGP1* has been identified as an enhancer RNA whose target gene is *AGAP4*, with the mRNA levels of both genes positively correlated (Ang et al., 2021). *PARGP1* has also been associated with prostate cancer, where low expression of *PARGP1*, and thus of *AGAP4*, have been linked to better survival rates. These findings are also consistent with a lack of CNVs at this locus, as gene dosage changes of these genes might play a role in disease.

Population differences in CN showed that the gene *TPTE* exhibits significantly lower CN in all populations of African ancestry compared to those with out-of-Africa ancestry. *TPTE*, or transmembrane phosphatase with tensin homology (also known as *PTEN2*), is a membrane-associated phosphatase located on chromosome 21p11.2 (Guipponi et al., 2000), which shares ~96% sequence identity with the protein-encoding gene *TPTE2* on chromosome 13q12.11 (also known as *TPIP*) (Walker et al., 2001). Several pseudogenes of these genes exist on acrocentric chromosomes 13, 15, 21, 22 and Y (Tapparel et al., 2003). Only one copy of protein-encoding *Tpte* exists in mice exhibiting conserved synteny with human *TPTE2* on chromosome 13q14.2-q21 (Guipponi et al., 2001), implying this region represents the ancestral copy from which *TPTE* emerged through a duplication event. In our analysis, *TPTE* was fully encompassed by an SD sharing >99% sequence identity with the pseudogene *TPTE2P4* on chromosome 13p11.2. Our CN estimates also showed that *TPTE* and *TPTE2P4* are absent from all archaic genomes surveyed, and *TPTE* is seemingly absent from the chimpanzee reference genome (panTro6). Remarkably, this duplication does not seem to be frequent in any of the African populations assayed, suggesting a prominence of a homozygous deletion haplotype in Africa due to either an African-specific deletion or an out-of-Africa duplication. Although its function is unknown, *TPTE* is expressed exclusively in testis, more specifically, in secondary spermatocytes (Wu et al., 2001) and could contribute to fertility.

The analysis of Tajima's D estimator, although not comprehensive, highlighted regions showing signatures of balancing or directional selection. The locus encompassing genes *SPDYE3* and *PMS2P1* on chromosome 7q22.1 showed significantly low D values across all humans queried. Low Tajima's D values indicate an excess of rare variants in the locus, which can be due to several factors, including a purifying selection or a selective sweep. We also found the regions to be CN fixed, with only two individuals among the 1KGP cohort genotyped for extra copies for both genes *SPDYE3* and *PMS2P1*. Together, D values and CN fixation are concordant with a functional region under

selective constraints. This and other loci highlighted in this study, will become priority candidates for long-read based population genetics analyses, functional studies, and disease associations.

4.5 Methods

4.5.1 Overall assessment of SD-98 regions

SD regions were extracted from previously annotated SDs (Vollger, Guitart, et al., 2022) and subsequently merged using bedtools merge (Quinlan & Hall, 2010). SD-98 were defined as a subset of SD with ≥98% sequence identity to another locus in the T2T-CHM13 genome. Genome coordinates of unique (Non-SD) regions excluded SD as well as annotated centromeric satellites (Altemose et al., 2022), while keeping pericentromeric SDs. Gene coordinates were obtained from T2T-CHM13 CAT/Liftoff annotations (v4) (Nurk et al., 2022). Overlap between SD-98 regions and genes were obtained using bedtools intersect (Quinlan & Hall, 2010). Overall numbers of distinct gene features overlapping SD-98 were counted using the assigned T2T-CHM13 gene ID.

4.5.2. Transcriptomics analysis

RNA-seq data were obtained for the developing brain (Fietz et al., 2012; Florio et al., 2015) and LCLs (Pickrell et al., 2010). Transcripts were quantified with Salmon v1.8.0 (Patro et al., 2017) with the flags "--validateMappings --gcBias", the T2T-CHM13 v2.0 CAT/Liftoff transcriptome, and the CHM13v2.0 assembly as decoy sequence. All identical transcripts were removed from the transcriptome prior to index construction. Transcripts per million (TPM) values were summed to the gene level using tximport (Soneson et al. 2015).

4.5.3 Depletion analysis

Databases of genetic analyses were obtained from GWAS Catalog v1.0 (mapped to GRCh38.p12) (Buniello et al., 2019), ClinVar (rel. 20200310) (Landrum et al., 2018), and GTEx v8 single-tissue eQTL (dbGaP Accession phs000424.v8.p2; mapped to GRCh38, excluding chromosome Y), as well as from biallelic-SNPs from 1KGP individuals mapped to T2T-CHM13 (v1.0) (Aganezov et al., 2022). Empirical distributions were generated by intersecting each dataset with randomly sampled regions of identical size to SD and SD-98 generated with bedtools shuffle -noOverlapping -maxTries 10000 -f 0.1. For depletion analyses in T2T-CHM13, centromeric satellites were also excluded using the flag -excl. One-tailed empirical *p*-values were calculated as: *p*-value = (M + 1) / (N + 1),

where M is the number of iterations yielding a number of features less than (depletion) observed and N is the number of iterations. Empirical *p*-values were calculated using 10,000 permutations.

4.5.4 Phenotype and disease associations

Gene-disease associations were obtained from GWAS catalog v1.0 (Buniello et al., 2019) and gnomad.v2.1.1 probability of loss of function intolerance scores (pLI) (Lek et al., 2016) and loss-of-function observed/expected upper fraction (LOEUF) (Karczewski et al., 2020), and intersect with SD-08 genes using gene symbols. High-confidence NDD-implicated genes were annotated with GeneTrek (https://genetrek.pasteur.fr/) (Leblond et al., 2021).

4.5.5 Paralog-specific copy-number genotyping

CN variant calls were obtained using QuicK-mer2 (Shen & Kidd, 2020). 1KGP 30× high-coverage Illumina reads in cram format (Byrska-Bishop et al., 2022) and four archaic genomes (including Altai Neanderthal [PRJEB1265] (Prüfer et al., 2014), Vindija Neanderthal [PRJEB21157] (Prüfer et al., 2017), Mezmaiskaya Neanderthal [PRJEB1757] (Prüfer et al., 2017, 2014), and Denisova [PRJEB3092] (Meyer et al., 2012)), were used as input for QuicK-mer2, using T2T-CHM13 (v1.0) as reference (Nurk et al., 2022). The resulting bed files containing CN estimates were converted into bed9 format using a custom python script for visualization in the UCSC genome browser. We genotyped CN in SD-98 regions overlapping protein-encoding and unprocessed pseudogenes (as defined in ENSEMBL biotypes) as the mean CN across the region of interest for each sample using a custom python script. Distinct gene features displaying CN fixation were counted using unique ENSEMBL gene IDs. CN-dotplots generated using the R package ggplot2 are available as an interactive Shiny web application in https://dcsoto.shinyapps.io/shinycn.

4.5.6 Copy-number stratification

CN differences between populations were calculated using the statistics V_{ST} (Redon et al., 2006), calculated as $V_{ST} = (V_T - V_S) / V_T$, where V_T is the total variance between two superpopulations var(pop1, pop2), and V_S is the weighted mean of the variance within each superpopulation, calculated as $V_S = [\ var(pop1) \cdot n_{pop1} + var(pop2) \cdot n_{pop2}\ ] / (n_{pop1} + n_{pop2})$. V_{ST} calculations were implemented in a custom R script. Distinct gene features displaying CN stratification were counted using unique ENSEMBL gene IDs.

4.5.7 Illumina SNV discovery benchmarking

Concordance between SNVs discovered with PacBio HiFi and Illumina sequencing were obtained for eight individuals of the 1KGP and HPRC+ datasets mapped to T2T-CHM13 (v1.0) (Aganezov et al., 2022), including individuals HG01109, HG01243, HG02055, HG02080, HG02145, HG02723, HG03098, and HG03492. Biallelic SNVs were selected using bcftools view (Danecek et al., 2021). Concordance between platforms, measured as precision and sensitivity, was obtained with rtg-tools vcfeval (Cleary et al., 2015) for autosomal Non-SDs, SDs, and SD-98 regions, using PacBio HiFi variants as a truth-set. Short-read accessible regions were obtained from Aganezov et al. (Aganezov et al., 2022).

4.5.8 Tajima's D calculation

Tajima's D values were obtained within 25-kbp using the software vcftools. Windows with less than 5 SNVs were removed from the analysis. To define short-read sequencing accessible Tajima's D values, 25-kbp windows were intersected with a previously published short-read combined accessibility mask.

4.5.9 GO overrepresentation

GO terms overrepresented in SD-98 genes were obtained using the R package clusterProfiler and the DAVID database (enrichDAVID function). Terms overrepresented in CN fixed genes were obtained using the ego function in R package clusterProfiler.

CHAPTER 5. Tools for (better) computational biology

Chapter 5 is a shared co-first authorship between Daniela C. Soto and Dr. Benjamín Sánchez Barja.

DCS wrote the abstract, introduction, section "personal research", part of section "collaboration", final words, and part of case studies, as well as designed figures.

5.1 Abstract

As biotechnological and biomedical research are increasingly fed by the insights arising from computation, the conversation about good practices in computational biology becomes more and more prominent. An increasing body of literature has addressed practices for shareable, reproducible, and sustainable computational research, from high-level principles for data and software stewardship to deep dives into version control or software automation. However, implementing these practices relies on incorporating the right tools into our daily routines, considering the type, scope, and stage of the research project. Here we provide a compendium of relevant tools for computational biology research, emphasizing their time and place within a continuum that traverses personal, collaborative, and community practices. This compendium will serve as a starting point and guide to help navigate the ongoing influx of tools and how to best incorporate them into a computational biologist's working routine, enabling reproducible biomedical and biotechnological research in the long term.

5.2 Introduction

Since Margaret Dayhoff pioneered the field of bioinformatics in the sixties, the application of computational tools in the field of biology has vastly grown in scope and impact. At present, biotechnological and biomedical research are routinely fed by the insights arising from novel computational approaches, machine learning algorithms, and mathematical models. The ever-increasing amount of biological data and the exponential growth in computing power will amplify this trend in the years to come.

The use of computing to address biological questions encompasses a wide array of applications usually grouped under the terms "computational biology" and "bioinformatics." Although distinct definitions have been delineated for each one (Huerta et al., 2000; Luscombe et al., 2001), here we will consider both under the umbrella term "computational

biology," alluding to any application that involves the intersection of computing and biological data. As such, a computational biologist can be a data analyst, a data engineer, a statistician, a mathematical modeler, a software developer, and many other roles. In praxis, the modern computational biologist will be a "scientist of many hats," taking on several of the duties listed above. But first and foremost, we will consider a computational biologist as a scientist whose ultimate goal is to answer a biological question or address a need in the life sciences by means of computation.

Scientific computing requires following specific principles to enable shareable, reproducible, and sustainable outputs. Computing-heavy disciplines, such as software engineering and business analytics, have adopted protocols addressing the need for collaboration, visualization, project management, and strengthening of online communities. However, as a highly interdisciplinary and evolving field, computational biology has yet to acquire a set of universal "best practices." Since most computational biologists come from diverse backgrounds and rely on self-study rather than formal education (Pinto et al., 2018), the absence of guidelines may lead many computational biologists astray, using methods that hinder reproducibility and collaboration, such as unreproducible computational workflows or closed-source software, retarding biomedical and biotechnological research.

In recent years, this "guidelines gap" has been addressed by the establishment of FAIR principles—Findability, Accessibility, Interoperability, and Reusability—in 2016 (Wilkinson et al., 2016). Originally developed for data stewardship, FAIR principles have been proposed as universal guidelines for all research-related outputs (G. Lee et al., 2021). However, translating these high-level principles into day-to-day practices requires additional nuances based on the type of research, the size and scope of the project, and the researcher's experience. To address the need for FAIR scientific software, for example, the framework ADVerTS (availability of software, documenting software, version control, testing, and support) has been proposed as a set of "barely sufficient" practices (G. Lee et al., 2021). More broadly, reviews exist covering general topics for bench scientists new to computational biology—such as programming and project organization (Carey & Papin, 2018; Grüning et al., 2019; Loman & Watson, n.d.; G. Wilson et al., 2014, 2017)—to detailed descriptions for the more seasoned data scientist—such as workflow automation (Reiter et al., 2021), software library development (Yurkovich et al., 2017), software version control with the cloud service GitHub (Perez-Riverol et al., 2016), and interactive data science notebooks with Jupyter (Rule et al., 2018).

Although the above reviews are immensely helpful, an overview of tools for better computational biology is missing. Indeed, guiding principles and general advice are key to establishing a behavior roadmap but their implementation is enabled by incorporating the right tools into our daily working routine. Tool selection has many components, such as availability, suitability, and personal preference; although the latter is left to the reader, here we will shed light on the first two. We premise that good practices in computational biology lie within a continuum that traverses three levels: personal (you), collaboration (your group), and community (your field) (**Figure 5.1**). Each of these levels has a different set of requirements and challenges, as well as a specific set of tools that can be used to address them. Here, we compiled a curated list of these tools, emphasizing their time and place in a computational biology research project. Committed to practicality, we illustrated the utility of these tools in case studies covering a wide spectrum of research topics that computational biologists can use to model their own practices, modifying them to suit their own needs and preferences.

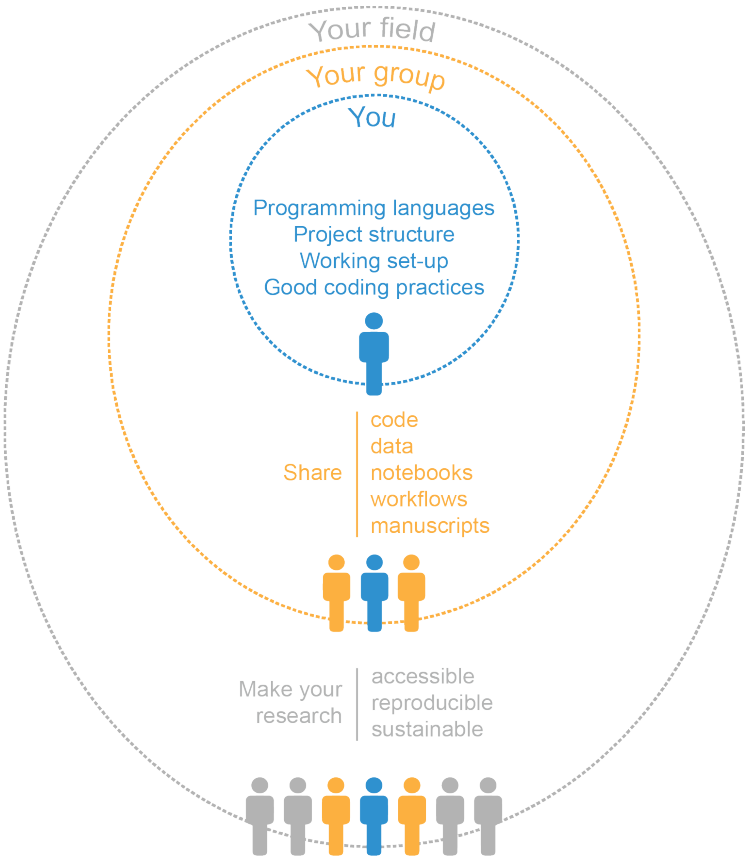


Figure 5.1. Schematic of the three "levels" of computational biology.

5.3 Level 1: Personal Research

The computational biology journey begins with you and the set of skills, tools, and practices that you have in place to conduct your research. Taking the time to optimally establish these building blocks will have high payoffs later when you find yourself going back to previous analyses. Consider that your most important collaborator is your future self, be it tomorrow or several years from now. We devised a framework involving four main sequential steps to kickstart any computational biology project (**Table 5.1**).

Table 5.1. Steps involved in starting a computational biology project.

Step	Use case	Common tools
Choose your programming languages	Interacting with a Unix/Linux HPC	• Shell/Bash
	Data analysis	• Python, R
	Scripts and programs	• Interpreted: Python, R, Perl, MATLAB, Julia • Compiled: C/C++, Rust
	Workflows	• Linux-based: Shell scripts, GNU Make • Workflow management systems: Snakemake (Python), Nextflow (Groovy) • Workflow specifications: CWL, WDL
Choose your project structure	Project structure	• Templates: Cookiecutter Data Science, rr-init • Workflows: Snakemake workflow template
	Virtual environment managers	• Language-specific: virtualenv (Python), renv (R) • Language agnostic: Conda
	Package managers	• Language-specific: pip (Python), Bioconductor (R), R Studio package manager (R) • Language-agnostic: Conda
Choose your working set-up	Text editors	• Desktop applications: Sublime, Visual Studio Code, Notepad++ • Command line: Vim, GNU Emacs
	IDEs	• For Python: JupyterLab, JetBrains/PyCharm, Spyder • For R: R Studio
	Notebooks	• Jupyter (Python, R), R Markdown (R)
Choose good coding practices	Coding style	• Styling guides: PEP-8 (Python), Google (Python, R) • Automatic code formatting: Black (Python), Snakefmt (Snakemake)
	Literate programming	• Markdown • R Markdown
	Version control	• Version control system: Git • Code repositories: GitHub, GitLab, Bitbucket • Git GUIs: GitHub Desktop, GitKraken

5.3.1 Choose your programming languages

Different programming languages serve distinctive purposes and have unique idiosyncrasies. As such, choosing a programming language for a specific project depends on your research goals, personal preferences, and skill sets.

Additionally, communities usually favor the usage and training of some programming languages over others; utilizing such languages may facilitate integrating your work within the existing ecosystem.

Interacting with high-performance computing (HPC) clusters has become a hallmark for the data-intensive discipline of computational biology. HPC infrastructures commonly use Unix/Linux distributions as their operating system. To interact with these platforms, a command-line interpreter known as the shell must be used. There are multiple versions of shells, with Bash (https://www.gnu.org/software/bash/) being one of the most widely adopted. In addition to providing an interface, the shell is also a scripting language that allows manipulating files and executing programs through shell scripts. Unix/Linux operating systems have other interesting perks, such as powerful, fast commands for searching and manipulating files (e.g., sed, grep, or join) as well as the language AWK, which can perform quick text processing and arithmetic operations.

One of the most common tasks of any computational biologist is data analysis, which usually involves data cleaning, exploration, manipulation, and visualization. Currently, Python (https://www.python.org/) is the most widely used programming language for data analysis (Kaggle, 2021; Stack Overflow, 2021). Python is also a popular language among computational biologists, a trend that will likely continue as machine learning and deep learning are more widely adopted in biological research. Python usage has been facilitated by the availability of packages for biological data analysis accessible through package managers such as pip (https://pip.pypa.io/) or Conda (https://docs.conda.io/). Likewise, R (https://www.r-project.org/) is another prominent language in the field. Arguably, one of the main strengths of R is its wide array of tools for statistical analysis. Of particular interest is the Bioconductor repository (https://www.bioconductor.org/), where many gold-standard tools for biological data analysis have been published and can be installed using BiocManager. R usage in data science has deeply benefited from the Tidyverse packages (Wickham et al., 2019) and surrounding community, increasing the readability of the R syntax for both data manipulation via dplyr and visualization via ggplot2.

Computational biologists often must code their own sets of instructions for processing data using scripts or tools. In computational biology, a script often refers to a lightweight single-file program written in an interpreted programming language and developed to perform a specific task. Scripts are quick to edit and can be run interactively but at the expense of computational performance. To automate instructions in HPC clusters, shell scripts are commonly used.

For other purposes, the most widely used scripting languages are Python and R, but Perl (https://www.perl.org/), MATLAB (https://www.mathworks.com/), and Julia (https://julialang.org/) are preferred by some researchers for bioinformatics, systems biology, and statistics, respectively. A computational biology tool, on the other hand, is a more complex program designed to tackle computationally intensive problems like developing new algorithms. Several tools devised for data-intensive biology have been written in compiled languages such as C/C++ (https://cplusplus.com/). In recent years, however, scientists have been turning to Rust (https://www.rust-lang.org/) due to its speed, memory safety, and active community (Perkel, 2020). When computational performance is less of a concern, Python and R are suitable alternatives for computational biology tool development.

Biological data processing is rarely a one-step process. To go from raw data to useful insights, several steps need to be taken in a specific order, accompanied by a plethora of decisions regarding parameters. Computational biologists have addressed this need by embracing workflow management systems to automate data analysis pipelines. A pipeline can be a shell script where commands are written sequentially, using shell variables and scripting syntax when needed. Although effective, this approach provides little control over the workflow and lacks features to run isolated parts of the pipeline or track changes in input and output files. To overcome these limitations, a shell script can be upgraded using the GNU Make (https://www.gnu.org/software/make/) program, which was originally designed to automate compilation and installation of software but is flexible enough to build workflows. More sophisticated bioinformatics workflow managers have also been developed such as Snakemake (https://snakemake.github.io/) based on Python (Mölder et al., 2021) and Nextflow (https://www.nextflow.io/) based on Groovy (a programming language for the Java virtual machine) (Di Tommaso et al., 2017). These tools offer support for software reproducibility using environment managers and software containers, as well as allow for easy scaling of pipelines to both traditional HPC and modern cloud environments. Alternatively, there are available declarative standards to define workflows in a portable and human-readable manner such as the Common Workflow Language (CWL) (https://www.commonwl.org/) and Workflow Description Language (WDL, pronounced “widdle”) (https://openwdl.org/), used by the cloud computing platform AnVIL (https://anvilproject.org/) (Schatz et al., 2022). Although these are not executable, they can be run in CWL- or DWL-enabled engines such as Cromwell (Voss et al., 2017).

5.3.2 Choose your project structure

The next step after choosing your programming languages but before starting coding is to develop an organized project structure. The project design should be intentional and tailored to the present and future needs of your project—remember to be kind to your future self! A computational biology project requires, at the very least, a folder structure that supports code, data, and documentation. Although tempting, cramming various file types into one unique folder is unsustainable. Instead, separate files into different folders and subfolders, if needed. To simplify this process, base your project structure on research templates available off-the-rack. For data science projects, the Python package Cookiecutter Data Science (https://drivendata.github.io/cookiecutter-data-science/) decreases the effort to minimal. Running the package prompts a questionnaire in the terminal where you can input the project name, authors, and other basic information. Then, the program generates a folder structure to store data—raw and processed—separate from notebooks and source code, as well as pre-made files for documentation such as a readme, a docs folder, and a license. Similarly, the Reproducible Research Project Initialization (rr-init) offers a template folder structure that can be modified by the user (https://github.com/Reproducible-Science-Curriculum/rr-init/). Although rr-init is slightly simpler, both follow an akin philosophy aimed at research correctness and reproducibility (Noble, 2009). For standard-compliant snakemake-workflows, where each workflow is in a dedicated folder divided into subfolders (https://snakemake.readthedocs.io/en/stable/index.html), we advise following the Snakemake Workflow Template (https://github.com/snakemake-workflows/snakemake-workflow-template/). In all cases, the folder must be initialized as a git repo for version control.

The software and dependencies needed to execute a tool or workflow are also part of the project structure itself. The intricacies of software installation and dependency management should not be underestimated. Fortunately, package and virtual environment managers significantly reduce this burden. A package manager is a system that automates the installation, upgrading, configuration, and removal of community-developed programs. A virtual environment manager is a tool that generates isolated environments where programs and dependencies are installed independently from other environments or the default operating system. Once a virtual environment is activated, a package manager can be used to install third-party programs. We believe that every computational biology project must start with its own virtual environment to boost reproducibility: environments save the project's dependencies and can restore them at will so the code can be run on any other computer. There are multiple options for both package and virtual

environment management—some language-specific and some language-agnostic. If you are working with Python, you can initialize a Python environment using virtualenv (https://virtualenv.pypa.io/) (where different Python versions can be installed). Inside the environment, you can use the Python package manager pip (https://pip.pypa.io/) to import Python code from the Python Package Index (PyPI) repository, cloud-based repositories, or locally. For the R language, R-specific environments can be created using renv (https://rstudio.github.io/renv/), where packages can be installed from the Comprehensive R Archive Network (CRAN) and CRAN-like repositories using the install.packages function, or from the Bioconductor repository using BiocManager. Additionally, RStudio Package Manager (https://www.rstudio.com/products/package-manager/) offers package management for third-party code available in R and Python repositories, as well as locally. Conda—a language-agnostic alternative—supports program installation from the Anaconda repository (https://repo.anaconda.com/), which contains the channel Bioconda (https://bioconda.github.io/) specifically tailored to bioinformatics software. Python dependencies can also be installed via pip inside a Conda environment. Conda is particularly helpful when working with third-party code in various languages—a common predicament in computational biology. The Conda package and environment manager is included in both the Anaconda and Miniconda distributions. The latter is a minimal version of Anaconda, containing only Conda, Python, and a few useful packages.

5.3.3 Choose your working set-up

Before coding, a more practical question needs to be answered first: Where to code? The simplest tools available for this purpose are text editors. Since writing code is ultimately writing text, any tool where characters can be typed fulfills this purpose. However, coding can be streamlined by additional features—including syntax highlight, indentation, and auto-completion—available in code editors such as Sublime (https://www.sublimetext.com/), Visual Studio Code (https://code.visualstudio.com/), and Notepad++ (https://notepad-plus-plus.org/) (Windows only), and command-line text editors such as Vim (https://www.vim.org/) and Emacs (https://www.gnu.org/software/emacs/). These tools share the advantage of being language agnostic, which is handy for the polyglot computational biologist.

In addition to text editors, integrated development environments (IDEs) are also popular options for coding. In their essence, IDEs are supercharged text editors comprising a code editor (with syntax highlight, indentation, and suggestions), a debugger, a folder structure, and a way to execute your code (a compiler or interpreter). Some IDEs

are not language-agnostic, often only allowing code in one language. The array of features also comes at a cost—IDEs typically use more memory. For Python, JupyterLab (https://jupyter.org/), Spyder (https://www.spyder-ide.org/), and PyCharm (https://www.jetbrains.com/pycharm/) are popular options, while for R, RStudio (https://www.rstudio.com/) is the gold standard. Notably, the differences between an IDE and a code editor are somewhat blurry, particularly when employing plugins with a code editor.

In recent years, data science notebooks have acquired relevance in computational biology research. A notebook is an interactive application that combines live code (read-print-eval loop or REPL), narrative, equations, and visualizations, internally stored using a format called JavaScript Object Notation (JSON). Common notebooks use interpreted languages such as Python or R, and narrative usually uses Markdown—a lightweight markup language (https://www.markdownguide.org/). Data analysis greatly benefits from using notebooks instead of plain text editors or even IDEs. The combination of visuals and texts allows researchers to tell compelling stories about their data, and the interactivity of its code enables quick testing of different strategies. Jupyter (https://jupyter.org/) is a popular web-based interactive notebook developed originally for Python coding but also accepts R and other programming languages upon installation of their kernels—the computing engine that executes the notebook's live code under the hood. Jupyter notebook can also be executed in the cloud using platforms such as Google CoLaboratory (CoLab) (https://colab.research.google.com/) and Amazon WebServices, taking advantage of the current trend of cloud computing. In addition, RStudio allows the generation of R-based notebooks known as R Markdown (https://rmarkdown.rstudio.com/), which is especially well suited for generating data analysis reports.

5.3.4 Choose good coding practices

With the foundation in place, the next step is to start writing code. Coding, however, requires good practices to ensure correctness, sustainability, and reproducibility for you, your future self, your collaborators, and the whole community. First and foremost, you need to make sure your code works correctly. In computational biology, correctness implies biological and statistical soundness. Although both are topics beyond the scope of this manuscript, a useful approach to evaluate biological correctness is to design positive and negative controls in your program, analysis, or workflow. In scientific experimentation, a positive control is a control group that is expected to produce results; a negative control is expected to produce no results. The same approach can be applied to computation, using input data whose output is

previously known. Biological soundness can also be tested by quickly assessing expected orders of magnitude in both intermediate and final files. These checks can be packaged in unit testing.

In addition to correctly functioning code, code appearance, also known as coding style, is important. Code style includes a series of small, ubiquitous decisions regarding where and how to add comments; indentation and white-space usage; variable, function, and class naming; and overall code organization. Although, as in writing, personality and preference differences dictate how you code, coding style rules facilitate collaboration with your future self and others. Indeed, as we sometimes have trouble reading our own handwriting, we can also struggle reading our own code if we disregard guidelines. At the very least, aim to follow internal consistency in writing code. Even better, consider following any of the multiple published coding-style guides such as those from software development teams. Google, for example, has guidelines for Python, R, Shell, C++, and HTML/CSS (https://github.com/google/styleguide/). Guidelines for Python are available as part of the Python Enhancement Proposal (PEP), known as PEP 8 (https://peps.python.org/pep-0008/). To facilitate compliance, tools called linters can be incorporated into most code editors and IDEs to flag stylistic errors in your code based on a given style guide. Furthermore, many editors and tools perform automatic code formatting (e.g., Black [https://black.readthedocs.io/] that formats Python code to be PEP 8 compliant), which can greatly facilitate stylistic coherence in a collaborative project. In the case of Snakemake files, stylistic errors can be flagged using the Snakemake's linter functionality or with the tool Snakefmt (https://github.com/snakemake/snakefmt/), based on Black.

On the matter of code styling, two topics merit additional attention: variable naming and comments. Variable names should be descriptive enough to convey information about the variable, function, or class content and use. The goal is to produce self-documented code that reads close to plain English. To do so, multi-word variable names should be used if necessary. In such cases, the most common conventions include Camel Case, where the second and subsequent words are capitalized (camelCase); Pascal Case, where all words are capitalized (PascalCase); and Snake Case, where words are separated by underscores (snake_case). Notably, these conventions can be used in the same coding style to differentiate variables, functions, and classes. For example, PEP-8 recommends Snake Case for functions and variables and Pascal Case for class names. As most modern code editors and IDEs provide autocompletion of variable, function, and class names, it is no longer a valid excuse to use cryptic one-character variable names (e.g., x, y, z) to save a few keystrokes.

In addition to mastering variable naming, code comments—explanatory human-readable statements not evaluated by the program—are necessary to enhance the code's readability. No matter how beautiful and well-organized your code is, high-level code decisions will not be obvious unless stated. As a corollary, code explanations that can be deduced from the syntax itself should be omitted. Comments can span a single line or several lines, and can be found in three strategic parts: at the top of the program file (header comment), which describes what the code accomplishes and sometimes the code's author/date; above every function (function header), which contains the purpose and behavior of the function; and in line, next to difficult code with behavior that is not obvious or warrants a remark.

Code-styling rules also apply to data science notebooks. However, when writing notebooks, you must also engage in literate programming—a programming paradigm where the code is accompanied by a human-readable explanation of its logic and purpose. In other words, notebooks must tell a story about the analysis, connecting the dots between the code, the results, and the figures. Human-readable language is often written in Markdown when working in Jupyter, or R Markdown when working in R. Little has been written about good practices for literate programming, but our suggested good practices are to include the purpose and interpretation of results for each section of code.

When working with a sizable codebase, we advise modular programming—the practice of subdividing a computer program into independent and interchangeable sub-programs, each one tackling a specific functionality. Modularity enhances code readability and reusability, as well as expedites testing and maintenance. In practice, modularity can be implemented at different levels, from using functions within a single-file program to separating functionalities into different files in a more complex tool. In Python, subdivisions are defined as follows: modules are a collection of functions and global variables, packages are a collection of modules, libraries are a collection of packages, and frameworks are a collection of libraries. Modules are files with .py extension, while packages are folders that contain several .py files, including one called _init_.py which can be empty or not and allows the Python interpreter to recognize a package.

Finally, use version control. Version control entails tracking and managing changes in the code. A popular version-control system is Git (https://git-scm.com/), which requires a folder to be initiated as a Git repository, after which changes to any of the files inside would be tracked. File modifications must be staged (using git add) and then committed (using git commit). The commit will serve as a screenshot of your project at that time and stage, which you

can review or recover later (using git checkout). Additionally, version control allows you to try new functions in branches (using git branch and git checkout)—independent carbon copies of the main original branch (known as main) that you can optionally merge back to the original copy. Currently, there are multiple hosting services that provide online storage of Git repositories, such as GitHub (https://github.com/), GitLab (https://about.gitlab.com/), or Bitbucket (https://bitbucket.org/), that users can navigate using the web browser or via a graphic user interface (GUI) such as GitHub Desktop (https://desktop.github.com/) or GitKraken (https://www.gitkraken.com). These platforms allow for a Git repository to be stored online by creating a copy of the repository known as the remote, providing the additional benefit of backing up your code in the cloud.

5.4 Level 2: Collaboration

Collaboration is a key aspect of scientific research, but it is especially relevant in computational biology, where interdisciplinary knowledge is often needed. Although collaborators can have a wide range of involvement with your project, here we will consider individuals that share a direct relationship with you and your research. Each type of collaboration requires its own set of good practices.

5.4.1 Share code

Sharing code is one of the most common practices in software development, where large teams work together to develop complex functions and scripts. Although computational biology projects usually involve smaller teams, proper sharing code remains essential (**Table 5.2**). Git-based hosting services, such as GitHub, GitLab, and Bitbucket , allow sharing the remote with collaborators, which becomes the official version of the repository. The key advantage of using a remote is that there will be no direct interaction between different local copies of the repository, also known as clones; instead, each clone will interact with the remote exclusively, updating only if no conflicts between the two exist. This way, if a collaborator updates the remote repository, other collaborators will not be able to send their changes until they update their local copy.

Table 5.2. Tools for collaborative research.

Goal	Tools
Share code	• Hosting services: GitHub, GitLab, Bitbucket. • Git branching strategies: GitHub flow. • Tests: correctness (e.g. pytest, testthat), style (e.g. flake8), vulnerabilities (e.g. Safety), coverage (e.g. codecov). • Continuous integration: tox, Travis CI, Circle CI, GitHub Actions. • Code reviews: Github, Crucible, Upsource.
Share data	• FAIR principles: FAIRshake. • Tidy data. • Data version control (DVC).
Share data science notebooks	• Static: GitHub, GitLab, NBviewer. • Interactive: Binder, Google CoLab. • Comparative: nbdime, ReviewNB.
Share workflows	• General hosting services: GitHub, GitLab, Bitbucket. • Dedicated workflow repositories: Snakemake Workflow Catalog, WorkflowHub.
Share manuscripts	• General-purpose word processors. • Online applications supporting Markup Languages: Overleaf (LaTeX), Manubot (Markdown + GitHub).

To guarantee that different collaborators can work simultaneously in the same repository, it is best to implement a branching strategy in the repository. In a small team, the most common strategy is to have a single main branch and generate branches from it that each different developer can work on. Then, whenever the developer is ready, they can request to combine—or merge—the changes from their branch into the main branch. This occurs via a process known as pull request (PR). Once a PR has been opened, collaborators can review, approve, and subsequently merge it into the main branch, preserving the commit history. This branching strategy is sometimes referred to as GitHub Flow (https://docs.github.com/en/get-started/quickstart/github-flow/) and will suffice for most projects.

Using Git hosting services for collaboration has many additional benefits. The commit history both shows what was done at each point in time but also specifies the collaborator who made the changes; this allows users to take responsibility for their changes so that if, for example, a bug was introduced, commands such as git blame can pinpoint the cause. To ensure bugs can be easily tracked, descriptive commit messages that follow a predefined standard (e.g. semantic commit messages [https://sparkbox.com/foundry/semantic_commit_messages/]) are recommended. Git hosting services can be accessed interactively online or from the terminal with tools such as GitHub CLI (https://cli.github.com/). Finally, Git hosting services also allow collaborators to open issues (https://docs.github.com/en/issues/tracking-your-work-with-issues/about-issues/) for listing pending tasks and/or

asking questions, acting as an open forum for development discussions, which has the advantage of remaining accessible for the future (as opposed to closed email discussions).

Another important concept to consider when developing code with collaborators, is to develop tests, meaning scripts that will run to find errors in the code (**Table 5.2**). Tests can be executed at different levels, from the individual units/components to the system/software as a whole. Unit tests, in particular, are used to determine if specific modules/functions work as intended within the codebase so that if later the function grows in scope, its proper basic functioning is ensured. For instance, if a function was defined for adding numbers, a simple test would be to assess if the function outputs 13 when the inputs 6 and 7 are provided. Besides unit tests, computational biology projects can benefit from implementing integration tests to evaluate the correct interaction between different modules and smoke tests to indicate if any core functionality has been impacted. Test runners, such as pytest (https://docs.pytest.org/) for Python and testthat (https://testthat.r-lib.org/) for R, exist to facilitate incorporating tests to the codebase. It is good practice to develop tests at the same time you develop code, as adding tests a posteriori is significantly harder. It is an even better practice to test every single step of the code (from data loading to figure plotting), a concept known in software development as end-to-end testing.

Going beyond testing correctness, flake8 (https://flake8.pycqa.org/) will test styling preferences (for complying with PEP8), Safety (https://pyup.io/safety/) will test for vulnerabilities among the software's dependencies, and Codecov (https://about.codecov.io/) will test coverage, or the percentage of the codebase tested. As a rule of thumb for testing coverage, the more lines of code tested, the more reliable the software will be. Different types of tests can be funneled into a single testing pipeline—in a process known as continuous integration (CI)—that can be tuned to run locally whenever commits are made, or online whenever a pull request is opened and/or merged. When running locally, an environment manager/command-line tool, such as tox (https://tox.wiki/), can help to ensure all tests are executed under different Python versions. Different tools, such as Travis CI (https://www.travis-ci.com/) or Circle CI (https://circleci.com/), can be used to set up the CI cycle online. More recently, GitHub Actions (https://github.com/features/actions) was developed to run integrations directly from GitHub.

Having tests is a great way to ensure that code fulfills a certain level of correctness and styling. However, it is no replacement for human assessment to determine if the code is correct, necessary, and useful. Therefore, peer code

review is essential whenever developing code in collaboration. While tools, such as Crucible (https://www.atlassian.com/software/crucible) and Upsource (https://www.jetbrains.com/upsource/), exist for making in-line reviews of each file, the most common approach is for you and/or others to directly review the code using the online review tools provided by various hosting services. In the case of GitHub, this not only allows the reviewer to open a comment in any line of the code, which creates a thread for the original author to reply but also to suggest changes that can be approved or dismissed. Reviewers can assess many features of the code, from functionality to documentation, while also following good practices, such as using constructive phrasing. We advise following Google's code review guidelines (https://google.github.io/eng-practices/review/reviewer/). Broad advice on how to code review is outside of the scope of this review but presented in detail elsewhere (Hauer, 2018).

5.4.2 Share data

The practices of sharing data are similar to sharing code: we should store our datasets, and any changes to them, in a repository and ensure they comply with standards by testing their quality. However, since data has a more consistent structure than code, often existing in standard formats, we should consider additional criteria when sharing it with collaborators (and later with the community) (**Table 5.2**). The main set of guidelines that represent these criteria was outlined in what is known as the FAIR principles (Wilkinson et al., 2016): data should be Findable (easy to locate online); Accessible (easy to access once found); Interoperable (easy to integrate with other data/applications/workflows/etc); and Reusable (presented in a way that allows for others to use it for the same or different purposes). Tools like FAIRshake (https://fairshake.cloud/) can be used to determine if data fits FAIR criteria.

For making data findable, research repositories such as Zenodo (https://zenodo.org/), Figshare (https://figshare.com/), Open Science Framework (OSF) (https://osf.io) allow you to assign a digital object identifier (DOI) to any group of files you upload, including data and/or code. Alternatively, regular code repositories like GitHub can be used instead, as you can employ commits and/or releases to identify specific versions of the data, in combination with extensions for Large File Storage (LFS), such as git LFS (https://git-lfs.github.com/), in the case of data files larger than 100 MB. GitHub can also integrate with Zenodo to automatically archive repositories and assign them a DOI. A final alternative is the Data Version Control (DVC) initiative (https://dvc.org/), which is especially useful when performing machine learning, as it can keep track of data, machine learning models, and even scoring metrics.

For making data accessible, we encourage you as much as possible to make your repositories open access. In cases in which you or your collaborators prefer some restrictions, you can create guest accounts to provide access to private repositories. For making data interoperable, distinctions between raw and clean data have been made (Noble, 2009), with raw data being the files that came out of the measuring device, and clean data representing the files that are ready to be used for any computational analysis. An important characteristic that clean data should have is to be tidy, which is reviewed in detail elsewhere (Wickham, 2014). Finally, for making data reusable, thorough documentation of the data is required, including experimental design, measurement units, and possible sources of error.

5.4.3 Share data science notebooks

Jupyter Notebooks have become a fundamental tool for data analysis, which can be shared with collaborators using either static or interactive options (**Table 5.2**). The former shares computational notebooks as rendered text, written internally in HTML. Static notebooks are a good option when you want to avoid any modifications and can work as an archive of past analyses, although interacting with its content is cumbersome—the file must be downloaded and run in a local Jupyter installation. Git-based code repositories, such as GitHub and GitLab, automatically render notebooks that can be later shared using the repository's URL. To facilitate this process, Project Jupyter provides a web application called NBviewer (https://nbviewer.org/), where you can paste a Jupyter Notebook's URL, publicly hosted on GitHub or elsewhere, and renders the file into a static HTML web page with a stable link.

Interactive notebooks, on the other hand, not only render the file but also allow collaborators to fully interact with it, tinkering with parameters or trying new input data—no installation required. Binder Project (https://mybinder.org/) enables users to fully interact with any notebook within a publicly-hosted Git-based repository via a Jupyter Notebook interface, although changes will not be saved to the original file. The platform supports Python and R, among other languages, and additional packages required to run the analysis need to be specified in a configuration file within the repository. Similarly, Jupyter Notebooks can be run interactively using Google CoLab by anyone with a Google account. Notebooks can be updated locally, from any public GitHub repository, or from Google Drive. As an added bonus, Google CoLab notebooks can be edited by multiple developers in real-time. In both cases, the machines provided by these services are comparable to a modern laptop, hence these tools may not be suitable for computing-intensive tasks.

Notebooks should be treated like any other piece of code: updates from different collaborators should be managed with version control in a platform such as GitHub. The problem, however, is that git and other version control systems use line-based differences that are not very well suited for the internal JSON representation of Jupyter notebooks. The extension nbdime (https://nbdime.readthedocs.io/) can be installed locally to enable content-aware diffing and merging. Additionally, NBreview (https://www.reviewnb.com/) can be integrated with GitHub to enable content-aware diffing, displaying the old and new versions of a notebook in parallel to facilitate code review.

5.4.4 Share computational workflows

Computational biology projects often demand using multi-step analyses with dozens of third-party software and dependencies. Although these steps can be described in the documentation, complex workflows are better shared as stand-alone code that can be easily run with minimal file manipulation from collaborators. Doing so can safeguard the reproducibility and replicability of the analysis, leading to better science and fewer challenges downstream (**Table 5.2**).

The simplest way to share a pipeline is through a shell script that receives input files via the command line; however, shell scripts offer little control over the overall workflow and cannot re-run specific parts of the pipeline. To address these issues, pipelines are better shared using a workflow automation system. Theoretically, all of the instructions regarding the workflow could be written in the main pipeline file: the .smk file (or Snakefile) in Snakemake; the .nf file in Nextflow; the .cwl file in CWL; and the .wdl fuke in WDL. However, to ensure reproducibility, it is a good practice to share complete pipelines, meaning folder structures, additional files, and software specifications, as well as custom scripts developed for the analysis. These files can be shared using the same tools as code, namely GitHub or any other Git hosting service. Alternatively, they can be uploaded to specialized hosting services for workflows, like Snakemake Workflow Catalog (https://snakemake.github.io/snakemake-workflow-catalog) or WorkflowHub (https://workflowhub.eu/) (currently in beta).

When sharing workflows, consider that sharing software version (known as dependency pinning) is necessary for your collaborators to reproduce your pipeline using their own computing setup. Conda environments, for example, can be easily created from an environment file (in YAML language), which can be exported from an existing environment. Notably, Snakemake and Nexflow can be configured to automatically build isolated environments for each rule or

step, enabling the running of different versions of a program within the same pipeline, which is especially helpful when requiring both Python 2 and 3, for example.

5.4.5 Share scientific manuscripts

Writing articles is the primary way we share our research with the scientific community at large. However, writing manuscripts collaboratively comes with its challenges when using classical word processing tools, often resulting in files with different names, jumping from one email inbox to another, and contradictory final versions. The tools we suggest will help to avoid these issues (**Table 5.2**). Besides the well-known word processors, that display text as it would appear as a printout (known as What-You-See-Is-What-You-Get, or WYSIWYG), text editors are a viable option to write manuscripts when combined with a markup language—a human-readable computer language that uses tags to delineate formatting elements in a document that will be later rendered. Since the formatting process is internally handled by the application, styling elements (e.g., headers, text formatting, and equations) are easily written in text, achieving greater consistency than word processors. Disciplines such as statistics and mathematics have historically used the markup language LaTeX for writing articles. This language has simple and specific syntax for mathematical constructs making it a popular choice for papers with many equations. To aid collaborative writing, platforms like Overleaf (https://www.overleaf.com) provide online LaTeX editors, supporting features like real-time editing. An emerging trend in collaborative writing uses the lightweight markup language Markdown within the GitHub infrastructure. The software Manubot (https://manubot.org) provides a set of functionalities to write scholarly articles within a GitHub repository, leveraging all the advantages of Git version control and the GitHub hosting platform, such as cloud storage, version control, issues and discussions (Himmelstein et al., 2019). Manubot, in particular, accepts citations using manuscript identifiers and automatically renders the article in PDF, HTML, and DOC formats. As a drawback, it requires technical expertise in Git and familiarity with GitHub; as an upside, its reliable infrastructure scales well to large and open collaborative projects. The document you are reading now was fully written using Manubot!

5.5. Level 3: Community

The third and final step of this journey is presenting your research to the community. Your main goal should be to share and maintain an open and reproducible project that can sustain community engagement over time. In this section,

we will distinguish three sub-goals to make your research: (1) accessible, (2) reproducible, and (3) sustainable. The latter is especially relevant when your research involves developing code that will be used by others in the future (e.g., a tool or workflow), but we believe that our recommendations are relevant to any computational biology project.

5.5.1. Make your research accessible

Making your research accessible includes ensuring that anyone can access your research long after your paper is published. It is extremely frustrating for any researcher to look for software or a set of scripts from a paper published a few years ago, only to find a "404 error" when accessing the source weblink. Equally frustrating is when authors offer code as "available upon reasonable request," as this often leads to dead-ends and unavailable code.

There are three main ways to publish accompanying code: the supplementary material of the manuscript, privately-owned domains, or uploaded to public repositories. Publishing code as supplementary material has low accessibility for non-open access papers. Moreover, the code will remain completely static and cannot be updated with new features or to correct errors. Making code available via privately-owned domains lacks sustainability, as it requires maintenance of the domain. Therefore, in addition to providing the code as supplementary material and/or via private domains, we recommend uploading it to public repositories (such as GitHuB or GitLab) and archive with a DOI (using Zenodo or figshare), enabling open access and sustainability over time.

When publishing your code in a public repository, two files are fundamental to include: A readme file and a license. A readme file (https://www.makeareadme.com/) introduces users to the code (**Table 5.3**) and should include a description of its main intended use, an overview of the installation, the most commonly-used commands, contact information of the developers, and acknowledgments, if appropriate. We recommend keeping the readme file short and written in a markup language such as Markdown or reStructuredText (https://docutils.sourceforge.io/rst.html) that will render automatically on the repository's landing page, below the repository file structure.

Table 5.3. Tools for making your research accessible.

Goal	Tool options	Additional remarks
Publish your code	• GitHub • GitLab • Bitbucket	All three options allow you to host your repository online for free. Choose whichever is more common in your own field.
Introduce your code	• README file: First file that shows up in a repository.	Provide a landing page to any repository with a short overview of the code (installation, usage, acknowledgments, etc).
Share your code	• Several licensing options.	Indicate with a license file what restrictions apply when using your code. If you don't include this, you will lose many users.
Archive your code	• GitHub Releases. • Archive with DOI: Zenodo, figshare, OSF	Share progressive stable versions of your code as you develop it. Use semantic versioning for assigning standard identifiers to your releases.
Publish a tool	• PyPI: Python. • CRAN: R. • Bioconductor: R. • Bioconda: Language-agnostic.	Produce a package easy to install and use. Especially useful if you think you could have a user base that will run the same analysis as you on other datasets and/or conditions.
Publish an interactive web app	• Dash: Python. • R-Shiny: R.	Provide easy and interactive data exploration to your users. Especially useful if you have large datasets that can be explored in different ways.

Adding a license to a repository is also a crucial step (**Table 5.3**). Licenses indicate how the code can be used: Is it free to use for any application? Can users modify the code as they please? Does it come with a warranty that it will work? Can it be used for profit? If no license information is provided, researchers might assume that the code is free to use but copyright law in fact prohibits use without explicit permission by the copyright holder (https://opensource.guide/legal/). Many options exist for licensing code (https://choosealicense.com/licenses/), from permissive licenses that allow any kind of use with few or no conditions, like the Unlicense and MIT licenses, to more restrictive licenses that enforce disclosing the source and requiring that any adaptation of the code uses the same license, like the GNU licenses. When deciding on a license, as a rule of thumb, consider that the more requirements you add, the fewer potential users you will have, but the more credit you will receive when users utilize your code for their own needs. Academic researchers must also consider what open-source licenses their university supports, as in many cases it will be the university that owns the copyrights.

As a computational biologist, you will likely continue lines of work from scripts or software you have already published. For instance, you could improve the performance of a given function or add a new set of features entirely. Therefore, you should not only be interested in making your code accessible but also in having different versions available. Creating and archiving successive releases of your code (**Table 5.3**) allows the organization of different

versions of your code as you develop them. GitHub Releases is one way to maintain versions with minimal effort. Research repositories, such as Zenodo, figshare, or OSF, not only store your code, notebooks, and data, but also provide a DOI for each version allowing it to be included as a citation in a manuscript. This is especially useful when the publication is not available yet or the current version of the code differs widely from what was published. Research repositories can be combined with code repositories; for example, GitHub has a Zenodo integration that will trigger a new archived version every time a new version is released. Regardless of the solution, we recommend keeping logical order to the releases, using a standard such as semantic versioning (Preston-Werner, n.d.).

In most cases, providing your code as an organized set of scripts and/or notebooks is sufficient for anyone to consult if they wish to reproduce and/or re-utilize it. However, if your code might be used routinely by other researchers, for instance for studying other organisms or other experimental conditions, consider packaging your code as a tool (**Table 5.3**) and publishing through a software repository such as Bioconda, PyPI if written for Python, or CRAN and Bioconductor if written for R. These increase your possible user base, as published packages are searchable and can be installed locally with minimal effort.

To increase the accessibility of results to users, an interactive web app or data dashboard can be developed (**Table 5.3**). Such apps allow users to interact with data by displaying different sets of variables or changing parameter settings (e.g., the significance of a statistical test). Common options for this goal are Dash (https://plotly.com/dash) for Python, R, and Julia, and Shiny (https://shiny.rstudio.com/) for R. Both platforms can include interactive graphics generated with plotly data visualization libraries (https://plotly.com/).

5.5.2. Make your research reproducible

In addition to having accessible code/data, you also need to ensure anyone can execute your code and obtain the same results. This is especially relevant in computational biology where users will come from different backgrounds and experience. A cornerstone for reproducibility is documentation that explains how the code functions and how to practically achieve the same results. We have distinguished four levels of documentation (Procida, 2017):

- Tutorials: A group of lessons that teach the reader how to become a user of your code;
- How-to guides: A set of documents that clarify how to solve common problems/tasks;

- Explanations: Discussions that clarify particular topics related to your code;
- References: Technical descriptions of your code's variables/classes/functions.

The extent of required documentation will depend on the number of expected users and, relatedly, can affect how many users you attract. If you foresee that your code has little usability outside of your own research, documenting each function using docstrings—a string specified before a module, function, class, or method to document its function—might be sufficient. However, if you aim for a broader user base, you might want to add a tutorial for beginners, how-to guides for frequently used routines, and explanations for clarifying the science behind your code, which can be re-used in a manuscript. To publish comprehensive documentation online, consider using (1) a standard documentation language such as reStructuredText or Markdown, and (2) a documentation platform such as Readthedocs (https://readthedocs.org/), Gitbook (https://www.gitbook.com/), Bookdown (https://bookdown.org/), or HackMD (https://hackmd.io) (**Table 5.4**). Alternatively, you can use a service like GitHub Pages (https://pages.github.com/) to host the documentation on a dedicated website.

Another key aspect of reproducibility is software and dependencies installation. To facilitate this process, you can (1) provide configuration instructions, (2) share dependencies with a virtual environment manager, or (3) share a runtime environment as a container. When setting up software from instructions, it is necessary to ensure the user follows a series of sequential commands in a specific order. To automate this process, Linux systems provide the tool GNU Make. Virtual environment managers handle dependencies and facilitate software installation by building virtual environments from requirements files. To achieve repeatable environments, however, it is recommended to include the specific version of software and libraries, a practice known as dependency pinning. Tools such as pip-tools (https://github.com/jazzband/pip-tools) allow defining different Python environments for a single project depending on the type of user (e.g., end-user versus developer).

Table 5.4. Tools for making your research reproducible (tool names in bold).

Goal	Tool options	Additional remarks
Document your code	• Readthedocs: Uses reStructuredText. • Gitbook: Uses Markdown. • Bookdown: Uses R Markdown. • HackMD: Uses Markdown. • Github Pages: Separate website.	Comprehensive documentation: from tutorials and how-to guides all the way down to function documentation based on all compiled docstrings.
Reproducible environments	• Virtual environment managers: See **Table 5.1** • pip-tools: Administer several environments in a single project.	As a recommendation, try having the minimum number of dependencies needed to reproduce your results.
Reproducible software	• Docker • Singularity	Package your research as a container ready to run on any computer.
Reproducible commands	• Make	Build a program by following a series of steps in a single Makefile.
Reproducible workflows	• Workflow management systems: See Table 5.1	Run a pipeline of commands on NGS data in a reproducible way.
Reproducible notebooks	• Interactive notebooks: See Table 5.1	Make your notebooks interactive and reproducible.

Beyond dependency trackers, we recommend ensuring your tool functions as expected across computing infrastructures, even between two different operating systems (e.g., Mac and Windows). This can be achieved via containerization, also known as lightweight virtualization (**Table 5.4**). Containers are standardized software that packages an entire runtime environment, meaning everything needed to run your tool: code, dependencies, system libraries and binaries, environmental variables, settings, etc. Instructions for deploying containers are stored in read-only templates called images. Two free tools available for creating containers from images are Docker (https://www.docker.com) and Singularity (http://sylabs.io/). While Docker is the most popular framework for containerization (Stack Overflow, 2021), HPC clusters with shared file systems favor Singularity due to security issues. In most cases, this is not a problem, since Singularity is compatible with all Docker images.

5.5.3. Make your research sustainable

Now that your research can be accessed and reproduced by anyone, the final step is to sustain this over time—also known as code maintenance. This is especially relevant if you continue to develop tools by integrating new features requested by users, which can foster a strong community over time. However, even in the case in which your research is a self-contained project, it is important to ensure that the user community can contact you, in case bugs are

discovered or parts of your code malfunction due to dependency updates (part of the "software rot" phenomenon). In the following section, we review useful techniques for making your code/software/research sustainable over time.

You can employ a variety of tools to communicate with users, depending on the size of your user base and the scope of questions received (**Table 5.5**). For smaller projects, a single-channel solution like Gitter (https://gitter.im/) offers a simple way for anyone in the community to ask questions and the developers to answer in threads. For larger projects, however, it could become unmanageable to have all discussions in the same channel, so a multiple-channel solution (i.e., forums), such as Google groups (https://groups.google.com/), is better suited. GitHub also allows issues to be opened, where collaborators or users can inform developers about bugs or ask questions. Additionally, GitHub recently introduced Discussions to maintain questions organized in different threads.

Table 5.5. Tools for making your research sustainable (tool names in bold).

Goal	Tool options	Additional remarks
Tell users how to contact you	• Specific/shorter questions: Gitter. • Larger issues / how-to's: Google groups, GitHub Discussions.	Provide ways for users to contact you for questions, requests, etc. Remember to visit them periodically!
Track to-do's in your research	• Github Issues.	Detail specific pending to-do's in your research / allow others to request changes and/or highlight bugs.
Encourage user contributions	• Contribution guidelines: How to open issues / contribute code. • Github Wikis: More specific how-to guides.	Provide as much information as you can to guide your users. You can also include administrator guidelines.
Foster a respectful community	• Smaller projects: Contributor Covenant.. • Larger projects: Citizen Code of Conduct.	Essential when you would like researchers to contribute code.
Branch your repo sustainably	• Gitflow.	Useful when several developers contribute code to the project. Allows users to get access to stable versions of your research in an ongoing project.
Keep track of your issues	• Kanban flowcharts: Github Projects, GitKraken Boards. • Scrum practices: Zenhub, Jira.	Keep track of your pending tasks in different projects with Agile software development practices. Especially useful if your research is split in many different repositories, each with multiple features/fixes to do.
Automate your repo	• bump2version: Easier releasing. • Danger-CI: Easier reviewing.	Do less, script more!

Now that users know where to contact you, ensure you have developed contribution guidelines (**Table 5.5**), detailing how users should (1) open issues and (2) contribute with their own code changes via PRs. These guidelines are intended for new users/contributors, so should be written in the style of a how-to guide; however, they may also include additional instructions for the main developers or the administrator of the repository. Alternatively, the detailed

guidelines can be included in a supplemental wiki, which hosting services offer as part of the repository. Equally important is a code of conduct (**Table 5.5**), which includes expectations on how users should behave in the repository and consequences when someone does not comply, promoting a respectful community. Several code-of-conduct templates exist, such as the Contributor Covenant (https://www.contributor-covenant.org/) for smaller projects and the Citizen Code of Conduct (https://github.com/stumpsyn/policies) for larger projects.

Finally, consistent development and maintenance of your software as it grows in scope and number of users will ensure the sustainability of your project. Tools that aid in this include:

1. Branching System: When many developers are involved in a project, more advanced branching methods, such as GitFlow (Driessen, 2010), ensure that users can access functional versions of your code while you work on it. (**Table 5.5**). Briefly, GitFlow includes two branches with an infinite lifetime: the main and the development (often named as devel). New branches will be based on the development branch, leaving the main one for stable versions of the code. Every time the development branch is merged into the main branch, a version release is created.
2. Project Management: Tools exist to track, organize, and prioritize user issues (**Table 5.5**), all based on Agile principles (https://agilemanifesto.org/). The simplest approach is implementing a Kanban board (as found in GitHub Projects or GitKraken Boards), where issues are organized in columns that clearly lay out the current state of a given task. For larger projects comprising multiple collaborators and/or repositories, a more structured approach, such as a Scrum framework (https://www.scrum.org/) (as implemented by Zenhub [https://www.zenhub.com/] and Jira [https://www.atlassian.com/software/jira]), allows you to prioritize issues by setting milestones and estimating difficulties.
3. Additional Automation: As your project develops, you will find that many aspects can be automated to improve efficiency. bump2version (https://github.com/c4urself/bump2version) ensures all sections of your code get updated with the new release. Danger-CI (http://danger.systems/ruby/) and git hooks (https://git-scm.com/book/en/v2/Customizing-Git-Git-Hooks) ensure contributors comply with certain standards in their pull requests. If you are no longer actively maintaining a project, you can use CI (e.g. GitHub Actions [https://github.com/features/actions]) to schedule regular tests to discover if your tool/code starts malfunctioning due to software rot and/or dependency issues. Finally, we advise against

implementing too many automation tools at the start of a project but adding them as needed. If you find yourself routinely performing a task, consider automating it.

5.6 Case Studies

We will now exemplify the effective use of the introduced tools by presenting three different computational biology projects from the literature (**Figure 5.2**). Note that our list of projects is not meant to be comprehensive, but rather is intended to be a short overview of how projects in computational biology benefit from robust tools and software development practices. Additionally, it will be evident that there is considerable redundancy in chosen tools across case studies. For instance, all projects include an environment manager such as Conda, and a version control system like Git. This redundancy is intentional as it highlights the ubiquity of some tools.

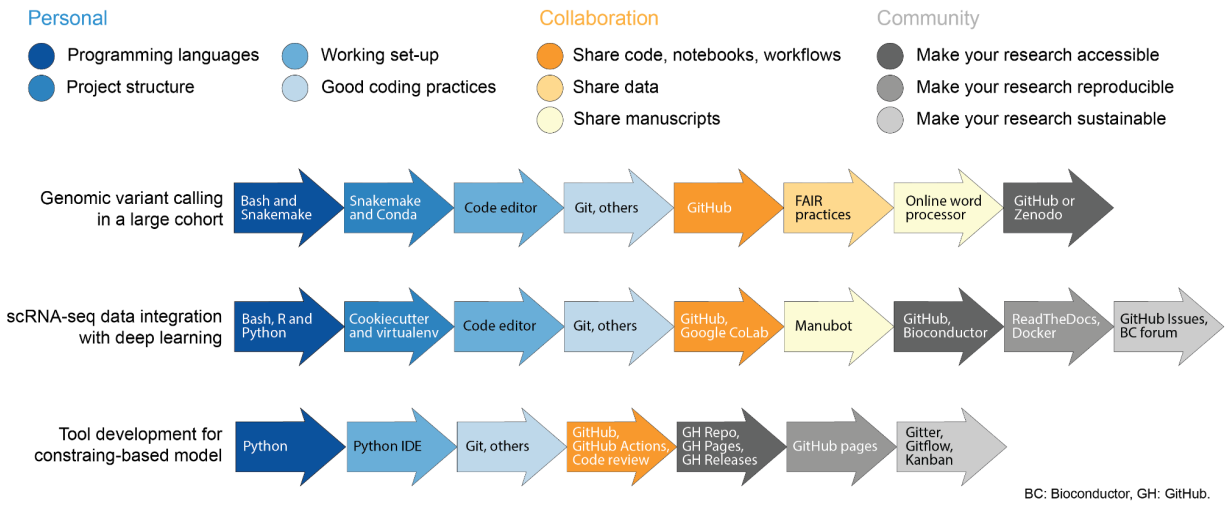


Figure 5.2. Examples of computational biology projects.

Case study 1: Genomic variant detection in a large cohort

The availability and affordability of NGS allow for the routine assessment of dozens to thousands of genomes. Resequencing experiments enable the discovery and genotyping of genomic variation within large cohorts to answer key questions regarding population history and susceptibility to disease. For this example, let's consider a project including whole-genome Illumina sequencing and variant identification in thousands of individuals such as Aganezov et al. (Aganezov et al., 2022). Herein, the challenge resides in applying a multi-step variant-calling pipeline to many samples in a reproducible manner.

In this project, the authors utilized the AnVIL cloud computing platform, which provides scalability and flexibility by running in a cloud environment and streamlines collaboration by allowing researchers to access the same tools and datasets from a centralized place. Importantly, tools like AnVIL allow the secure sharing of protected human datasets, which is paramount in human genomics studies. If you use AnVIL, then the pipeline must be written in WDL. Alternatively, a project of this nature can be written in Snakemake, employing Python to parse sample names and perform other data handling operations, and following the Snakemake workflow template for folder structure. A Conda environment can hold all necessary software since a wide array of software designed for genomic analyses is available via the Bioconda repository. Coding the workflow can be done in any text editor that offers easy integration with Git tracking and hosting, such as Visual Studio Code. For code styling, you can run Snakefmt to follow best practices.

A project of this magnitude usually requires collaborators from other research groups. The pipelines and scripts can be shared using a GitHub repository. If privacy is a concern, the repository can be set as private and made public in later stages of the project. To write the manuscript, a general-purpose word processor would suffice. Considering that these types of data are a valuable resource for the community, FAIR principles for data sharing should be followed. In addition to uploading the raw data to a public repository like the European Nucleotide Archive (ENA) or the National Center for Biotechnology Information (NCBI), we encourage open sharing of your code and notebooks in a GitHub repository archived in Zenodo with a DOI.

Case study 2: Single-cell (sc)RNA-seq data integration

scRNA-seq is a rapidly evolving technology that has enabled the study of cell heterogeneity and developmental changes of a cell lineage, otherwise intractable with bulk RNA-seq. Current scRNA-seq experiments deliver the transcriptomic profiles of thousands to millions of cells (Svensson et al., 2018), making them a suitable target for machine- or deep-learning approaches. Among the many challenges imposed by this technology, integration of scRNA-seq datasets is key, especially in case-control studies where cell types should be functionally matched across datasets before evaluating differences across conditions. For this case study, we will consider the development of an unsupervised deep-learning method for data integration as described in Johansen and Quon (Johansen & Quon, 2019).

This kind of project often uses a combination of Python, R, and shell scripting. Python can be used to write and train deep-learning models with TensorFlow (https://www.tensorflow.org/) or PyTorch (https://www.pytorch.org) libraries.

R enables straightforward data pre-processing with tools such as Seurat (https://satijalab.org/seurat/) (Hao et al., 2021). Shell scripting can process large-scale raw data files in HPC clusters. Additionally, we advise using Python's reticulate library (https://rstudio.github.io/reticulate/) to incorporate Python tools into the existing R ecosystem. To set up your working directory, we recommend a structure like Cookiecutter Data Science, which includes separate folders for trained models and other components of a deep-learning project. To establish a software environment, Python virtual environments, such as virtualenv, work well with Tensorflow and PyTorch. Coding can be performed in any general-purpose text editor, such as Visual Studio Code, where updates can be easily pushed/pulled to/from GitHub. As a good practice, maintain modular, properly commented code and name files with data stamps and model parameters to facilitate revisiting projects. Additionally, take advantage of tools such as TensorBoard (https://www.tensorflow.org/tensorboard/) to diagnose, visualize, and experiment with your models.

When working with collaborators, code should be shared through a Git hosting service like GitHub. When multiple users need to edit the code in real-time, Google CoLab offers interactive coding and GPU access. In addition to the code repository, a Manubot can be created to write the manuscript collaboratively. To make your tool accessible to a larger community, publish it to a public GitHub and include a readme and an appropriate license file. Considering that most users in the field use R, you can go one step further and share your code as a Bioconductor package, making sure your method can be called directly in R and that interacts with standard data structures in the field. For better reproducibility, document your method including example tutorials in a platform like ReadTheDocs, and share the software environment needed to deploy the models as a Docker image. GitHub issues and Bioconductor forums (https://support.bioconductor.org/) are suitable platforms to promptly reply to users' questions, bug reports, and requests for code enhancements.

Case study 3: Tool development for constraint-based modeling

The last case study we will present is related to constraint-based modeling; a common approach used for simulating cellular metabolism. In this approach, the metabolic network of a given organism is inferred from its genome and/or literature and converted to a matrix that contains the reaction's stoichiometry. Using a few simple assumptions, this matrix can be used to perform simulations under different experimental conditions to obtain additional insight into cellular physiology (Bordbar et al., 2014). Several tools have been developed for working with these types of models.

Here, we will consider cobrapy (Ebrahim et al., 2013), a community tool for reading/writing constrained-based models and performing basic simulation operations.

A tool of this nature is especially useful if developed in Python, as it should ideally be presented as a package that can be easily installed with pip. The use of an IDE is ideal for this case, as it will provide additional features for testing changes in the tool. Practices that for other case studies were useful now become essential, like complying with coding style and using version control, as hundreds of people will likely read your code. Furthermore, the code should be (1) available via a hosting service such as GitHub, (2) tested with a continuous development tool such as GitHub Actions, (3) manually reviewed by collaborators to ensure correctness, (4) released following semantic versioning standards, and (5) documented with a companion documentation website, rich with tutorials and how-to guides. As a branching strategy, Gitflow is probably the best suited, as it allows all changes to existing code in a development branch and stable releases in the main branch.

Finally, due to the large scope of this project, additional considerations must be made to maintain a healthy user base. Offer a place for users to raise questions, such as Gitter, Google groups, or GitHub Discussions, and make sure to reply to new questions often. Guidelines should also be provided for everything, including how to: open issues with example templates, contribute using pull-request templates, communicate within the community via a code of conduct, and perform other routine tasks with development guidelines and/or wikis. Addressing issues routinely and quickly is also essential in a project of this nature to avoid giving the impression of a stagnant project. Additional tools such as a Kanban flowchart with the help of GitHub Projects will help prioritize issues, or Jira or Zenhub if several repositories require joint coordination.

5.7 Final words

Good practices in computational biology have gained the spotlight among researchers thanks to several guiding principles published, as well as the increasing usage of Git-based repositories and workflow managers. This review adds to the existing literature by introducing a comprehensive list of good practices and associated tools that can be applied to any computational biology project, regardless of the specific subfield or the experience of the researcher.

We are aware that the many tools and practices introduced in this study and their ever-changing nature may seem overwhelming, especially for someone new to the field. To overcome this, we encourage you to implement only a few practices and tools first, starting from your personal research, and expanding your repertoire over time. More important than any specific tool is keeping a mindset of striving for reproducibility. We also note that our highlighted list of tools is not comprehensive, with many new tools being released. Updated reviews will be essential to help new computational biologists enter the field as well as to keep experienced computational biologists up to date with the latest trends.

The consequences of not following good computational practices are often not seen immediately but become evident and detrimental to project progress over time. As with all scientific endeavors, computational biology heavily relies on previous knowledge; as such, the good practices we adopt serve as building blocks for the overall reproducibility of the field, propelling novel and exciting future discoveries.

5.8 Acknowledgments

We would like to thank Nelson Johansen for his insights on the scRNA-seq data integration case study, Dr. Elizabeth Georgian for proofreading and editing the manuscript, and Drs. Michael C. Schatz and C. Titus Brown for their comments on this manuscript's draft.

CHAPTER 6. Summary and future directions

6.1 Summary of the presented work

The aim of this dissertation was to leverage diverse sequencing technologies, including short- (SRS) and long-read sequencing (LRS), to identify and characterize complex genomic variation—including segmental duplications (SDs) and structural variation (SVs)—in great apes' genomes, focusing on humans and chimpanzees.

In chapter 2, using nanopore LRS and optical mapping, we identified novel deletions and inversions in two non-previously sequenced chimpanzee genomes, and assessed their impact on gene regulation and chromatin organization by integrating SVs with expression data and chromatin conformation capture. The approach highlighted the power of LRS to identify novel SVs, even with a small sample set. Second, it added to the body of literature that shows the significant impact that SVs have on gene regulation by altering chromatin structure. Our approach supported the hypothesis that topologically associated domains (TAD) boundaries are evolving under purifying selection which removes deletions and inversion breakpoints impacting domain boundaries.

In chapter 3, we showed how the first ever complete sequence of a human genome, T2T-CHM13, improves the analysis of genetic variation by allowing accurate SNV and SV detection with SRS and LRS. This was achieved by (i) correcting previously erroneous gene sequences, (ii) removing of artificial haplotypes originated from assembling multiple individuals, and (iii) adding >200 Mbp of additional sequence previously missing from the reference genome. We comprehensively characterized collapsed duplications in GRCh38, i.e., regions of the genome with missing gene copies, and highlighted the impact of reference errors in variant detection within genes of biomedical significance. In all, we show how this genome allows for better functional, population, and biomedical genomic studies, and recommend T2T-CHM13 adoption by the scientific community.

In chapter 4, we leveraged the new T2T-CHM13 reference to study nearly-identical SDs— here defined as SDs with sequence identity over 98% (SD-98)—some of which were erroneously represented in GRCh38. Genes within SD-98 have been systematically excluded from genetic analyses due to the difficulties of mapping short reads to duplicated regions. Thus, the functions and evolutionary impact of most SD-98 genes remains unknown. While most population genetic scans skip SDs, we used copy-number (CN) at unique *k*-mers and high-quality SNVs, to obtain signatures of

selection in recent duplications. Our approach highlighted genes exhibiting (i) evidence of expression in LCL and human developing brain, (ii) CN fixation in modern human, (iii) CN population stratification, and (iv) putative evidence of balancing or directional selection as measured by the excess of rare SNVs, all of which are priority candidates for future functional studies and evolutionary analyses.

Finally, in chapter 5, we summarized tools and practices we recommend for sustainable computational biology research. As biology becomes a data-intensive scientific discipline, ushered in by the availability of SRS and LRS datasets as well as the many *-seqs* and *-omics*, the routine acquisition of tools and techniques that enhance reproducibility at the personal, collaborative, and community levels will become increasingly necessary. We provided a compendium of tools used in software engineering and data science that will aid computational biologists to streamline their research.

6.2 Future directions

Although we identify novel SVs in chimpanzees using LRS approaches, our study highlighted the need for more LRS datasets of all four chimpanzee subspecies. While this study included the first LRS of an individual carrying admixture with central chimpanzee, the genetic diversity of this species remains understudied. Characterizing the full spectrum of genetic variation across all subspecies will aid conservation efforts and allow us to better understand the evolutionary history of our closest living relatives.

The complete sequence of a human genome, T2T-CHM13, will enable a myriad of new biological insights, especially in previously missing sequences, including centromeres, telomeres, rDNA, and highly identical SDs. In this work, we achieved novel insights by characterizing SD-98 genes. However, our approach was not comprehensive due to the limitations of SRS. To fully characterize the variation landscape of these regions, LRS variant discovery is necessary. The Human Pangenome Reference Consortium is addressing this need by providing haplotype-resolved nearly complete assemblies of diverse humans, from which phased SNVs can be obtained. As LRS becomes more affordable, we anticipate more genomes at population scale to enable comprehensive population genetics scans. This will lead to the additional challenge of applying traditional population genetics tests to duplicated regions evolving under interlocus gene conversion. Further, improved methods to extract variation across duplications from LRS and assemblies, especially across copy number polymorphic regions, are required.

References

Abel, H. J., Larson, D. E., Regier, A. A., Chiang, C., Das, I., Kanchi, K. L., Layer, R. M., Neale, B. M., Salerno, W. J., Reeves, C., Buyske, S., NHGRI Centers for Common Disease Genomics, Matise, T. C., Muzny, D. M., Zody, M. C., Lander, E. S., Dutcher, S. K., Stitziel, N. O., & Hall, I. M. (2020). Mapping and characterization of structural variation in 17,795 human genomes. *Nature*, *583*(7814), 83–89. https://doi.org/10.1038/s41586-020-2371-0

Aganezov, S., Yan, S. M., Soto, D. C., Kirsche, M., Zarate, S., Avdeyev, P., Taylor, D. J., Shafin, K., Shumate, A., Xiao, C., Wagner, J., McDaniel, J., Olson, N. D., Sauria, M. E. G., Vollger, M. R., Rhie, A., Meredith, M., Martin, S., Lee, J., … Schatz, M. C. (2022). A complete reference genome improves analysis of human genetic variation. *Science*, *376*(6588), eabl3533. https://doi.org/10.1126/science.abl3533

Aguirre, M., Rivas, M. A., & Priest, J. (2019). Phenome-wide Burden of Copy-Number Variation in the UK Biobank. *American Journal of Human Genetics*, *105*(2), 373–383. https://doi.org/10.1016/j.ajhg.2019.07.001

Alexander, D. H., Novembre, J., & Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Research*, *19*(9), 1655–1664. https://doi.org/10.1101/gr.094052.109

Alexandrov, L. B., Nik-Zainal, S., Wedge, D. C., Aparicio, S. A. J. R., Behjati, S., Biankin, A. V., Bignell, G. R., Bolli, N., Borg, A., Børresen-Dale, A.-L., Boyault, S., Burkhardt, B., Butler, A. P., Caldas, C., Davies, H. R., Desmedt, C., Eils, R., Eyfjörd, J. E., Foekens, J. A., … Stratton, M. R. (2013). Signatures of mutational processes in human cancer. *Nature*, *500*(7463), 415–421. https://doi.org/10.1038/nature12477

Alkan, C., Coe, B. P., & Eichler, E. E. (2011). Genome structural variation discovery and genotyping. *Nature Reviews. Genetics*, *12*(5), 363–376. https://doi.org/10.1038/nrg2958

Alkan, C., Sajjadian, S., & Eichler, E. E. (2011). Limitations of next-generation genome sequence assembly. *Nature Methods*, *8*(1), 61–65. https://doi.org/10.1038/nmeth.1527

Almarri, M. A., Bergström, A., Prado-Martinez, J., Yang, F., Fu, B., Dunham, A. S., Chen, Y., Hurles, M. E., Tyler-Smith, C., & Xue, Y. (2020). Population Structure, Stratification, and Introgression of Human Structural Variation. *Cell*, *182*(1), 189-199.e15. https://doi.org/10.1016/j.cell.2020.05.024

Altemose, N., Logsdon, G. A., Bzikadze, A. V., Sidhwani, P., Langley, S. A., Caldas, G. V., Hoyt, S. J., Uralsky, L., Ryabov, F. D., Shew, C. J., Sauria, M. E. G., Borchers, M., Gershman, A., Mikheenko, A., Shepelev, V. A.,

Dvorkina, T., Kunyavskaya, O., Vollger, M. R., Rhie, A., … Miga, K. H. (2022). Complete genomic and epigenetic maps of human centromeres. *Science*, *376*(6588), eabl4178. https://doi.org/10.1126/science.abl4178

Amemiya, H. M., Kundaje, A., & Boyle, A. P. (2019). The ENCODE Blacklist: Identification of Problematic Regions of the Genome. *Scientific Reports*, *9*(1), 9354. https://doi.org/10.1038/s41598-019-45839-z

Andrés, A. M., Dennis, M. Y., Kretzschmar, W. W., Cannons, J. L., Lee-Lin, S.-Q., Hurle, B., NISC Comparative Sequencing Program, Schwartzberg, P. L., Williamson, S. H., Bustamante, C. D., Nielsen, R., Clark, A. G., & Green, E. D. (2010). Balancing selection maintains a form of ERAP2 that undergoes nonsense-mediated decay and affects antigen presentation. *PLoS Genetics*, *6*(10), e1001157. https://doi.org/10.1371/journal.pgen.1001157

Ang, X., Xu, Z., Zhou, Q., Zhang, Z., Ma, L., Zhang, X., Zhou, F., & Chen, W. (2021). PARGP1, a specific enhancer RNA associated with biochemical recurrence of prostate cancer. *All Life*, *14*(1), 774–781. https://doi.org/10.1080/26895293.2021.1969292

Angata, T., Ishii, T., Motegi, T., Oka, R., Taylor, R. E., Soto, P. C., Chang, Y.-C., Secundino, I., Gao, C.-X., Ohtsubo, K., Kitazume, S., Nizet, V., Varki, A., Gemma, A., Kida, K., & Taniguchi, N. (2013). Loss of Siglec-14 reduces the risk of chronic obstructive pulmonary disease exacerbation. *Cellular and Molecular Life Sciences: CMLS*, *70*(17), 3199–3210. https://doi.org/10.1007/s00018-013-1311-7

Aqil, A., Speidel, L., Pavlidis, P., & Gokcumen, O. (2022). Balancing selection on genomic deletion polymorphisms in humans. In *bioRxiv* (p. 2022.04.28.489864). https://doi.org/10.1101/2022.04.28.489864

Audano, P. A., Sulovari, A., Graves-Lindsay, T. A., Cantsilieris, S., Sorensen, M., Welch, A. E., Dougherty, M. L., Nelson, B. J., Shah, A., Dutcher, S. K., Warren, W. C., Magrini, V., McGrath, S. D., Li, Y. I., Wilson, R. K., & Eichler, E. E. (2019). Characterizing the Major Structural Variant Alleles of the Human Genome. *Cell*, *0*(0). https://doi.org/10.1016/j.cell.2018.12.019

Bailey, J. A., Yavor, A. M., Massa, H. F., Trask, B. J., & Eichler, E. E. (2001). Segmental duplications: organization and impact within the current human genome project assembly. *Genome Research*, *11*(6), 1005–1017. https://doi.org/10.1101/gr.gr-1871r

Bailey, Jeffrey A., & Eichler, E. E. (2006). Primate segmental duplications: crucibles of evolution, diversity and disease. *Nature Reviews. Genetics*, *7*, 552. https://doi.org/10.1038/nrg1895

Bailey, Jeffrey A., Gu, Z., Clark, R. A., Reinert, K., Samonte, R. V., Schwartz, S., Adams, M. D., Myers, E. W., Li, P. W., & Eichler, E. E. (2002). Recent segmental duplications in the human genome. *Science*, *297*(5583), 1003–1007. https://doi.org/10.1126/science.1072047

Bailey, Jeffrey A., Liu, G., & Eichler, E. E. (2003). An Alu transposition model for the origin and expansion of human segmental duplications. *American Journal of Human Genetics*, *73*(4), 823–834. https://doi.org/10.1086/378594

Bailey, Jeffrey A., Yavor, A. M., Massa, H. F., Trask, B. J., & Eichler, E. E. (2001). Segmental Duplications: Organization and Impact Within the Current Human Genome Project Assembly. *Genome Research*, *11*(6), 1005–1017. https://doi.org/10.1101/gr.187101

Ball, M. P., Thakuria, J. V., Zaranek, A. W., Clegg, T., Rosenbaum, A. M., Wu, X., Angrist, M., Bhak, J., Bobe, J., Callow, M. J., Cano, C., Chou, M. F., Chung, W. K., Douglas, S. M., Estep, P. W., Gore, A., Hulick, P., Labarga, A., Lee, J.-H., … Church, G. M. (2012). A public resource facilitating clinical use of genomes. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(30), 11920–11927. https://doi.org/10.1073/pnas.1201904109

Ballouz, S., Dobin, A., & Gillis, J. A. (2019). Is it time to change the reference genome? *Genome Biology*, *20*(1), 159. https://doi.org/10.1186/s13059-019-1774-4

Balogh, A., Toth, E., Romero, R., Parej, K., Csala, D., Szenasi, N. L., Hajdu, I., Juhasz, K., Kovacs, A. F., Meiri, H., Hupuczi, P., Tarca, A. L., Hassan, S. S., Erez, O., Zavodszky, P., Matko, J., Papp, Z., Rossi, S. W., Hahn, S., … Than, N. G. (2019). Placental Galectins Are Key Players in Regulating the Maternal Adaptive Immune Response. *Frontiers in Immunology*, *10*, 1240. https://doi.org/10.3389/fimmu.2019.01240

Bekpen, C., & Tautz, D. (2019). Human core duplicon gene families: game changers or game players? *Briefings in Functional Genomics*, *18*(6), 402–411. https://doi.org/10.1093/bfgp/elz016

Bergström, A., McCarthy, S. A., Hui, R., Almarri, M. A., Ayub, Q., Danecek, P., Chen, Y., Felkel, S., Hallast, P., Kamm, J., Blanché, H., Deleuze, J.-F., Cann, H., Mallick, S., Reich, D., Sandhu, M. S., Skoglund, P., Scally, A., Xue, Y., … Tyler-Smith, C. (2020). Insights into human genetic variation and population history from 929 diverse genomes. *Science*, *367*(6484). https://doi.org/10.1126/science.aay5012

Besenbacher, S., Hvilsom, C., Marques-Bonet, T., Mailund, T., & Schierup, M. H. (2019). Direct estimation of mutations in great apes reconciles phylogenetic dating. *Nature Ecology & Evolution*.

https://doi.org/10.1038/s41559-018-0778-x

Beyter, D., Ingimundardottir, H., Oddsson, A., Eggertsson, H. P., Bjornsson, E., Jonsson, H., Atlason, B. A., Kristmundsdottir, S., Mehringer, S., Hardarson, M. T., Gudjonsson, S. A., Magnusdottir, D. N., Jonasdottir, A., Jonasdottir, A., Kristjansson, R. P., Sverrisson, S. T., Holley, G., Palsson, G., Stefansson, O. A., … Stefansson, K. (2021). Long-read sequencing of 3,622 Icelanders provides insight into the role of structural variants in human diseases and other traits. *Nature Genetics*, *53*(6), 779–786. https://doi.org/10.1038/s41588-021-00865-4

Bordbar, A., Monk, J. M., King, Z. A., & Palsson, B. O. (2014). Constraint-based models predict metabolic and associated cellular functions. *Nature Reviews. Genetics*, *15*(2), 107–120. https://doi.org/10.1038/nrg3643

Brawand, D., Soumillon, M., Necsulea, A., Julien, P., Csardi, G., Harrigan, P., Weier, M., Liechti, A., Aximu-Petri, A., Kircher, M., Albert, F. W., Zeller, U., Khaitovich, P., Grutzner, F., Bergmann, S., Nielsen, R., Paabo, S., & Kaessmann, H. (2011). The evolution of gene expression levels in mammalian organs. *Nature*, *478*(7369), 343–348. https://doi.org/10.1038/nature10532

Bridges, C. B. (1936). The bar "gene" a duplication. *Science*, *83*(2148), 210–211. https://doi.org/10.1126/science.83.2148.210

Brunetti-Pierri, N., Berg, J. S., Scaglia, F., Belmont, J., Bacino, C. A., Sahoo, T., Lalani, S. R., Graham, B., Lee, B., Shinawi, M., Shen, J., Kang, S.-H. L., Pursley, A., Lotze, T., Kennedy, G., Lansky-Shafer, S., Weaver, C., Roeder, E. R., Grebe, T. A., … Patel, A. (2008). Recurrent reciprocal 1q21.1 deletions and duplications associated with microcephaly or macrocephaly and developmental and behavioral abnormalities. *Nature Genetics*, *40*(12), 1466–1471. https://doi.org/10.1038/ng.279

Buniello, A., MacArthur, J. A. L., Cerezo, M., Harris, L. W., Hayhurst, J., Malangone, C., McMahon, A., Morales, J., Mountjoy, E., Sollis, E., Suveges, D., Vrousgou, O., Whetzel, P. L., Amode, R., Guillen, J. A., Riat, H. S., Trevanion, S. J., Hall, P., Junkins, H., … Parkinson, H. (2019). The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Research*, *47*(D1), D1005–D1012. https://doi.org/10.1093/nar/gky1120

Byrska-Bishop, M., Evani, U. S., Zhao, X., Basile, A. O., Abel, H. J., Regier, A. A., Corvelo, A., Clarke, W. E., Musunuri, R., Nagulapalli, K., Fairley, S., Runnels, A., Winterkorn, L., Lowy, E., Human Genome Structural Variation Consortium, Paul Flicek, Germer, S., Brand, H., Hall, I. M., … Zody, M. C. (2022). High-coverage

whole-genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. *Cell*, *185*(18), 3426-3440.e19. https://doi.org/10.1016/j.cell.2022.08.004

Cabanski, C. R., Wilkerson, M. D., Soloway, M., Parker, J. S., Liu, J., Prins, J. F., Marron, J. S., Perou, C. M., & Hayes, D. N. (2013). BlackOPs: increasing confidence in variant detection through mappability filtering. *Nucleic Acids Research*, *41*(19), e178. https://doi.org/10.1093/nar/gkt692

Cagan, A., Theunert, C., Laayouni, H., Santpere, G., Pybus, M., Casals, F., Prüfer, K., Navarro, A., Marques-Bonet, T., Bertranpetit, J., & Andrés, A. M. (2016). Natural Selection in the Great Apes. *Molecular Biology and Evolution*, *33*(12), 3268–3283. https://doi.org/10.1093/molbev/msw215

Campbell, C. D., Sampas, N., Tsalenko, A., Sudmant, P. H., Kidd, J. M., Malig, M., Vu, T. H., Vives, L., Tsang, P., Bruhn, L., & Eichler, E. E. (2011). Population-genetic properties of differentiated human copy-number polymorphisms. *American Journal of Human Genetics*, *88*(3), 317–332. https://doi.org/10.1016/j.ajhg.2011.02.004

Cantsilieris, S., Sunkin, S. M., Johnson, M. E., Anaclerio, F., Huddleston, J., Baker, C., Dougherty, M. L., Underwood, J. G., Sulovari, A., Hsieh, P., Mao, Y., Catacchio, C. R., Malig, M., Welch, A. E., Sorensen, M., Munson, K. M., Jiang, W., Girirajan, S., Ventura, M., … Eichler, E. E. (2020). An evolutionary driver of interspersed segmental duplications in primates. *Genome Biology*, *21*(1), 202. https://doi.org/10.1186/s13059-020-02074-4

Carey, M. A., & Papin, J. A. (2018). Ten simple rules for biologists learning to program. *PLoS Computational Biology*, *14*(1), e1005871. https://doi.org/10.1371/journal.pcbi.1005871

Carpenter, D., Mitchell, L. M., & Armour, J. A. L. (2017). Copy number variation of human AMY1 is a minor contributor to variation in salivary amylase expression and activity. *Human Genomics*, *11*(1), 2. https://doi.org/10.1186/s40246-017-0097-3

Carroll, S. B. (2003). Genetics and the making of Homo sapiens. *Nature*, *422*(6934), 849–857. https://doi.org/10.1038/nature01495

Carvalho, C. M. B., & Lupski, J. R. (2016). Mechanisms underlying structural variant formation in genomic disorders. *Nature Reviews. Genetics*, *17*(4), 224–238. https://doi.org/10.1038/nrg.2015.25

Catacchio, C. R., Maggiolini, F. A. M., D'Addabbo, P., Bitonto, M., Capozzi, O., Lepore Signorile, M., Miroballo, M., Archidiacono, N., Eichler, E. E., Ventura, M., & Antonacci, F. (2018). Inversion variants in human and

primate genomes. *Genome Research*, *28*(6), 910–920. https://doi.org/10.1101/gr.234831.118

Chaisson, M. J. P., Huddleston, J., Dennis, M. Y., Sudmant, P. H., Malig, M., Hormozdiari, F., Antonacci, F., Surti, U., Sandstrom, R., Boitano, M., Landolin, J. M., Stamatoyannopoulos, J. A., Hunkapiller, M. W., Korlach, J., & Eichler, E. E. (2015). Resolving the complexity of the human genome using single-molecule sequencing. *Nature*, *517*(7536), 608–611. https://doi.org/10.1038/nature13907

Chaisson, M. J. P., Sanders, A. D., Zhao, X., Malhotra, A., Porubsky, D., Rausch, T., Gardner, E. J., Rodriguez, O. L., Guo, L., Collins, R. L., Fan, X., Wen, J., Handsaker, R. E., Fairley, S., Kronenberg, Z. N., Kong, X., Hormozdiari, F., Lee, D., Wenger, A. M., … Lee, C. (2019). Multi-platform discovery of haplotype-resolved structural variation in human genomes. *Nature Communications*, *10*(1), 1784. https://doi.org/10.1038/s41467-018-08148-z

Chaisson, M. J. P., Wilson, R. K., & Eichler, E. E. (2015). Genetic variation and the de novo assembly of human genomes. *Nature Reviews. Genetics*, *16*(11), 627–640. https://doi.org/10.1038/nrg3933

Chander, V., Gibbs, R. A., & Sedlazeck, F. J. (2019). Evaluation of computational genotyping of structural variation for clinical diagnoses. *GigaScience*, *8*(9). https://doi.org/10.1093/gigascience/giz110

Charrier, C., Joshi, K., Coutinho-Budd, J., Kim, J.-E., Lambert, N., de Marchena, J., Jin, W.-L., Vanderhaeghen, P., Ghosh, A., Sassa, T., & Polleux, F. (2012). Inhibition of SRGAP2 function by its human-specific paralogs induces neoteny during spine maturation. *Cell*, *149*(4), 923–935. https://doi.org/10.1016/j.cell.2012.03.034

Chen, J. M., Cooper, D. N., Chuzhanova, N., Férec, C., & Patrinos, G. P. (2007). Gene conversion: mechanisms, evolution and human disease. *Nature Reviews. Genetics*, *8*(10), 762–775. https://doi.org/10.1038/nrg2193

Chen, L. T., Gilman, A. G., & Kozasa, T. (1999). A candidate target for G protein action in brain. *The Journal of Biological Chemistry*, *274*(38), 26931–26938. https://doi.org/10.1074/jbc.274.38.26931

Cheng, H., Concepcion, G. T., Feng, X., Zhang, H., & Li, H. (2021). Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nature Methods*, *18*(2), 170–175. https://doi.org/10.1038/s41592-020-01056-5

Cheng, J. Y., Stern, A. J., Racimo, F., & Nielsen, R. (2021). Detecting selection in multiple populations by modeling ancestral admixture components. *Mol. Biol. Evol*. https://doi.org/10.1093/molbev/msab294

Cheng, Z., Ventura, M., She, X., Khaitovich, P., Graves, T., Osoegawa, K., Church, D., DeJong, P., Wilson, R. K., Pääbo, S., Rocchi, M., & Eichler, E. E. (2005). A genome-wide comparison of recent chimpanzee and human

segmental duplications. *Nature*, *437*(7055), 88–93. https://doi.org/10.1038/nature04000

Chiang, C., Layer, R. M., Faust, G. G., Lindberg, M. R., Rose, D. B., Garrison, E. P., Marth, G. T., Quinlan, A. R., & Hall, I. M. (2015). SpeedSeq: ultra-fast personal genome analysis and interpretation. *Nature Methods*, *12*(10), 966–968. https://doi.org/10.1038/nmeth.3505

Chiang, C., Scott, A. J., Davis, J. R., Tsang, E. K., Li, X., Kim, Y., Hadzic, T., Damani, F. N., Ganel, L., GTEx Consortium, Montgomery, S. B., Battle, A., Conrad, D. F., & Hall, I. M. (2017). The impact of structural variation on human gene expression. *Nature Genetics*, *49*(5), 692–699. https://doi.org/10.1038/ng.3834

Chimpanzee Sequencing and Analysis Consortium. (2005). Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature*, *437*(7055), 69–87. https://doi.org/10.1038/nature04072

Church, D. M., Schneider, V. A., Steinberg, K. M., Schatz, M. C., Quinlan, A. R., Chin, C.-S., Kitts, P. A., Aken, B., Marth, G. T., Hoffman, M. M., Herrero, J., Mendoza, M. L. Z., Durbin, R., & Flicek, P. (2015). Extending reference assembly models. *Genome Biology*, *16*, 13. https://doi.org/10.1186/s13059-015-0587-3

Cleary, J. G., Braithwaite, R., Gaastra, K., Hilbush, B. S., Inglis, S., Irvine, S. A., Jackson, A., Littin, R., Rathod, M., Ware, D., Zook, J. M., Trigg, L., & De La Vega, F. M. (2015). Comparing Variant Call Files for Performance Benchmarking of Next-Generation Sequencing Variant Calling Pipelines. In *bioRxiv* (p. 023754). https://doi.org/10.1101/023754

Collins, R. L., Brand, H., Karczewski, K. J., Zhao, X., Alföldi, J., Francioli, L. C., Khera, A. V., Lowther, C., Gauthier, L. D., Wang, H., Watts, N. A., Solomonson, M., O'Donnell-Luria, A., Baumann, A., Munshi, R., Walker, M., Whelan, C. W., Huang, Y., Brookings, T., … Talkowski, M. E. (2020). A structural variation reference for medical and population genetics. *Nature*, *581*(7809), 444–451. https://doi.org/10.1038/s41586-020-2287-8

Conrad, D. F., Pinto, D., Redon, R., Feuk, L., Gokcumen, O., Zhang, Y., Aerts, J., Andrews, T. D., Barnes, C., Campbell, P., Fitzgerald, T., Hu, M., Ihm, C. H., Kristiansson, K., Macarthur, D. G., Macdonald, J. R., Onyiah, I., Pang, A. W. C., Robson, S., … Hurles, M. E. (2010). Origins and functional impact of copy number variation in the human genome. *Nature*, *464*(7289), 704–712. https://doi.org/10.1038/nature08516

Damert, A. (2022). SVA Retrotransposons and a Low Copy Repeat in Humans and Great Apes: A Mobile Connection. *Molecular Biology and Evolution*, *39*(5). https://doi.org/10.1093/molbev/msac103

Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., Whitwham, A., Keane, T., McCarthy,

S. A., Davies, R. M., & Li, H. (2021). Twelve years of SAMtools and BCFtools. *GigaScience*, *10*(2). https://doi.org/10.1093/gigascience/giab008

de Manuel, M., Kuhlwilm, M., Frandsen, P., Sousa, V. C., Desai, T., Prado-Martinez, J., Hernandez-Rodriguez, J., Dupanloup, I., Lao, O., Hallast, P., Schmidt, J. M., Heredia-Genestar, J. M., Benazzo, A., Barbujani, G., Peter, B. M., Kuderna, L. F. K., Casals, F., Angedakin, S., Arandjelovic, M., … Marques-Bonet, T. (2016). Chimpanzee genomic diversity reveals ancient admixture with bonobos. *Science*, *354*(6311), 477–481. https://doi.org/10.1126/science.aag2602

Dennis, M. Y., & Eichler, E. E. (2016). Human adaptation and evolution by segmental duplication. *Current Opinion in Genetics & Development*, *41*, 44–52. https://doi.org/10.1016/j.gde.2016.08.001

Dennis, M. Y., Harshman, L., Nelson, B. J., Penn, O., Cantsilieris, S., Huddleston, J., Antonacci, F., Penewit, K., Denman, L., Raja, A., Baker, C., Mark, K., Malig, M., Janke, N., Espinoza, C., Stessman, H. A. F., Nuttle, X., Hoekzema, K., Lindsay-Graves, T. A., … Eichler, E. E. (2017). The evolution and population diversity of human-specific segmental duplications. *Nature Ecology & Evolution*, *1*(3), 69. https://doi.org/10.1038/s41559-016-0069

Dennis, M. Y., Nuttle, X., Sudmant, P. H., Antonacci, F., Graves, T. A., Nefedov, M., Rosenfeld, J. A., Sajjadian, S., Malig, M., Kotkiewicz, H., Curry, C. J., Shafer, S., Shaffer, L. G., de Jong, P. J., Wilson, R. K., & Eichler, E. E. (2012). Evolution of human-specific neural SRGAP2 genes by incomplete segmental duplication. *Cell*, *149*(4), 912–922. https://doi.org/10.1016/j.cell.2012.03.033

Derrien, T., Estellé, J., Marco Sola, S., Knowles, D. G., Raineri, E., Guigó, R., & Ribeca, P. (2012). Fast computation and applications of genome mappability. *PloS One*, *7*(1), e30377. https://doi.org/10.1371/journal.pone.0030377

Di Tommaso, P., Chatzou, M., Floden, E. W., Barja, P. P., Palumbo, E., & Notredame, C. (2017). Nextflow enables reproducible computational workflows. *Nature Biotechnology*, *35*(4), 316–319. https://doi.org/10.1038/nbt.3820

Dixon, J. R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J. S., & Ren, B. (2012). Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, *485*(7398), 376–380. https://doi.org/10.1038/nature11082

Dougherty, M. L., Nuttle, X., Penn, O., Nelson, B. J., Huddleston, J., Baker, C., Harshman, L., Duyzend, M. H.,

Ventura, M., Antonacci, F., Sandstrom, R., Dennis, M. Y., & Eichler, E. E. (2017). The birth of a human-specific neural gene by incomplete duplication and gene fusion. *Genome Biology*, *18*(1), 49. https://doi.org/10.1186/s13059-017-1163-9

Driessen, V. (2010, January 5). A successful Git branching model. *Nvie.com*. https://nvie.com/posts/a-successful-git-branching-model/

Dumont, B. L. (2015). Interlocus gene conversion explains at least 2.7% of single nucleotide variants in human segmental duplications. *BMC Genomics*, *16*, 456. https://doi.org/10.1186/s12864-015-1681-3

Durand, N. C., Shamim, M. S., Machol, I., Rao, S. S. P., Huntley, M. H., Lander, E. S., & Aiden, E. L. (2016). Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. *Cell Systems*, *3*(1), 95–98. https://doi.org/10.1016/j.cels.2016.07.002

Ebbert, M. T. W., Jensen, T. D., Jansen-West, K., Sens, J. P., Reddy, J. S., Ridge, P. G., Kauwe, J. S. K., Belzil, V., Pregent, L., Carrasquillo, M. M., Keene, D., Larson, E., Crane, P., Asmann, Y. W., Ertekin-Taner, N., Younkin, S. G., Ross, O. A., Rademakers, R., Petrucelli, L., & Fryer, J. D. (2019). Systematic analysis of dark and camouflaged genes reveals disease-relevant genes hiding in plain sight. *Genome Biology*, *20*(1), 97. https://doi.org/10.1186/s13059-019-1707-2

Ebert, P., Audano, P. A., Zhu, Q., Rodriguez-Martin, B., Porubsky, D., Bonder, M. J., Sulovari, A., Ebler, J., Zhou, W., Serra Mari, R., Yilmaz, F., Zhao, X., Hsieh, P., Lee, J., Kumar, S., Lin, J., Rausch, T., Chen, Y., Ren, J., … Eichler, E. E. (2021). Haplotype-resolved diverse human genomes and integrated analysis of structural variation. *Science*, *372*(6537). https://doi.org/10.1126/science.abf7117

Ebrahim, A., Lerman, J. A., Palsson, B. O., & Hyduke, D. R. (2013). COBRApy: COnstraints-Based Reconstruction and Analysis for Python. *BMC Systems Biology*, *7*, 74. https://doi.org/10.1186/1752-0509-7-74

Eichler, E. E. (2001). Segmental duplications: what's missing, misassigned, and misassembled--and should we care? *Genome Research*, *11*(5), 653–656. https://doi.org/10.1101/gr.188901

Eizenga, J. M., Novak, A. M., Sibbesen, J. A., Heumos, S., Ghaffaari, A., Hickey, G., Chang, X., Seaman, J. D., Rounthwaite, R., Ebler, J., Rautiainen, M., Garg, S., Paten, B., Marschall, T., Sirén, J., & Garrison, E. (2020). Pangenome Graphs. *Annual Review of Genomics and Human Genetics*, *21*, 139–162. https://doi.org/10.1146/annurev-genom-120219-080406

ENCODE Project Consortium, Moore, J. E., Purcaro, M. J., Pratt, H. E., Epstein, C. B., Shoresh, N., Adrian, J., Kawli,

T., Davis, C. A., Dobin, A., Kaul, R., Halow, J., Van Nostrand, E. L., Freese, P., Gorkin, D. U., Shen, Y., He, Y., Mackiewicz, M., Pauli-Behn, F., … Weng, Z. (2020). Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature*, *583*(7818), 699–710. https://doi.org/10.1038/s41586-020-2493-4

Eres, I. E., Luo, K., Hsiao, C. J., Blake, L. E., & Gilad, Y. (2019). Reorganization of 3D genome structure may contribute to gene regulatory evolution in primates. *PLoS Genetics*, *15*(7), e1008278. https://doi.org/10.1371/journal.pgen.1008278

Feuk, L., MacDonald, J. R., Tang, T., Carson, A. R., Li, M., Rao, G., Khaja, R., & Scherer, S. W. (2005). Discovery of human inversion polymorphisms by comparative analysis of human and chimpanzee DNA sequence assemblies. *PLoS Genetics*, *1*(4), e56. https://doi.org/10.1371/journal.pgen.0010056

Fiddes, I. T., Lodewijk, G. A., Mooring, M., Bosworth, C. M., Ewing, A. D., Mantalas, G. L., Novak, A. M., van den Bout, A., Bishara, A., Rosenkrantz, J. L., Lorig-Roach, R., Field, A. R., Haeussler, M., Russo, L., Bhaduri, A., Nowakowski, T. J., Pollen, A. A., Dougherty, M. L., Nuttle, X., … Haussler, D. (2018). Human-Specific NOTCH2NL Genes Affect Notch Signaling and Cortical Neurogenesis. *Cell*, *173*(6), 1356-1369.e22. https://doi.org/10.1016/j.cell.2018.03.051

Fiddes, I. T., Pollen, A. A., Davis, J. M., & Sikela, J. M. (2019). Paired involvement of human-specific Olduvai domains and NOTCH2NL genes in human brain evolution. *Human Genetics*, *138*(7), 715–721. https://doi.org/10.1007/s00439-019-02018-4

Fietz, S. A., Lachmann, R., Brandl, H., Kircher, M., Samusik, N., Schröder, R., Lakshmanaperumal, N., Henry, I., Vogt, J., Riehn, A., Distler, W., Nitsch, R., Enard, W., Pääbo, S., & Huttner, W. B. (2012). Transcriptomes of germinal zones of human and mouse fetal neocortex suggest a role of extracellular matrix in progenitor self-renewal. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(29), 11836–11841. https://doi.org/10.1073/pnas.1209647109

Florio, M., Albert, M., Taverna, E., Namba, T., Brandl, H., Lewitus, E., Haffner, C., Sykes, A., Wong, F. K., Peters, J., Guhr, E., Klemroth, S., Prüfer, K., Kelso, J., Naumann, R., Nüsslein, I., Dahl, A., Lachmann, R., Pääbo, S., & Huttner, W. B. (2015). Human-specific gene *ARHGAP11B* promotes basal progenitor amplification and neocortex expansion. *Science*, *347*(6229), 1465–1470. https://doi.org/10.1126/science.aaa1975

Florio, M., Heide, M., Pinson, A., Brandl, H., Albert, M., Winkler, S., Wimberger, P., Huttner, W. B., & Hiller, M. (2018). Evolution and cell-type specificity of human-specific genes preferentially expressed in progenitors

of fetal neocortex. *ELife*, *7*. https://doi.org/10.7554/eLife.32332

Florio, M., Namba, T., Pääbo, S., Hiller, M., & Huttner, W. B. (2016). A single splice site mutation in human-specific ARHGAP11B causes basal progenitor amplification. *Science Advances*, *2*(12), e1601941. https://doi.org/10.1126/sciadv.1601941

Franke, M., Ibrahim, D. M., Andrey, G., Schwarzer, W., Heinrich, V., Schöpflin, R., Kraft, K., Kempfer, R., Jerković, I., Chan, W.-L., Spielmann, M., Timmermann, B., Wittler, L., Kurth, I., Cambiaso, P., Zuffardi, O., Houge, G., Lambie, L., Brancati, F., … Mundlos, S. (2016). Formation of new chromatin domains determines pathogenicity of genomic duplications. *Nature*, *538*(7624), 265–269. https://doi.org/10.1038/nature19800

Fudenberg, G., & Pollard, K. S. (2019). Chromatin features constrain structural variation across evolutionary timescales. *Proceedings of the National Academy of Sciences of the United States of America*, *116*(6), 2175–2180. https://doi.org/10.1073/pnas.1808631116

Gallego Romero, I., Pavlovic, B. J., Hernando-Herraez, I., Zhou, X., Ward, M. C., Banovich, N. E., Kagan, C. L., Burnett, J. E., Huang, C. H., Mitrano, A., Chavarria, C. I., Friedrich Ben-Nun, I., Li, Y., Sabatini, K., Leonardo, T. R., Parast, M., Marques-Bonet, T., Laurent, L. C., Loring, J. F., & Gilad, Y. (2015). A panel of induced pluripotent stem cells from chimpanzees: a resource for comparative functional genomics. *ELife*, *4*, e07103. https://doi.org/10.7554/eLife.07103

Gel, B., & Serra, E. (2017). karyoploteR: an R/Bioconductor package to plot customizable genomes displaying arbitrary data. In *Bioinformatics* (Vol. 33, Issue 19, pp. 3088–3090). https://doi.org/10.1093/bioinformatics/btx346

Ghavi-Helm, Y., Jankowski, A., Meiers, S., Viales, R. R., Korbel, J. O., & Furlong, E. E. M. (2019). Highly rearranged chromosomes reveal uncoupling between genome topology and gene expression. *Nature Genetics*, *51*(8), 1272–1282. https://doi.org/10.1038/s41588-019-0462-3

Giannuzzi, G., Schmidt, P. J., Porcu, E., Willemin, G., Munson, K. M., Nuttle, X., Earl, R., Chrast, J., Hoekzema, K., Risso, D., Männik, K., De Nittis, P., Baratz, E. D., 16p11.2 Consortium, Herault, Y., Gao, X., Philpott, C. C., Bernier, R. A., Kutalik, Z., … Reymond, A. (2019). The Human-Specific BOLA2 Duplication Modifies Iron Homeostasis and Anemia Predisposition in Chromosome 16p11.2 Autism Individuals. *American Journal of Human Genetics*, *105*(5), 947–958. https://doi.org/10.1016/j.ajhg.2019.09.023

Giner-Delgado, C., Villatoro, S., Lerga-Jaso, J., Gayà-Vidal, M., Oliva, M., Castellano, D., Pantano, L., Bitarello, B.

D., Izquierdo, D., Noguera, I., Olalde, I., Delprat, A., Blancher, A., Lalueza-Fox, C., Esko, T., O'Reilly, P. F., Andrés, A. M., Ferretti, L., Puig, M., & Cáceres, M. (2019). Evolutionary and functional impact of common polymorphic inversions in the human genome. *Nature Communications*, *10*(1), 4222. https://doi.org/10.1038/s41467-019-12173-x

Goidts, V., Szamalek, J. M., de Jong, P. J., Cooper, D. N., Chuzhanova, N., Hameister, H., & Kehrer-Sawatzki, H. (2005). Independent intrachromosomal recombination events underlie the pericentric inversions of chimpanzee and gorilla chromosomes homologous to human chromosome 16. *Genome Research*, *15*(9), 1232–1242. https://doi.org/10.1101/gr.3732505

Gokcumen, O., Tischler, V., Tica, J., Zhu, Q., Iskow, R. C., Lee, E., Fritz, M. H.-Y., Langdon, A., Stütz, A. M., Pavlidis, P., Benes, V., Mills, R. E., Park, P. J., Lee, C., & Korbel, J. O. (2013). Primate genome architecture influences structural variation mechanisms and functional consequences. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(39), 15764–15769. https://doi.org/10.1073/pnas.1305904110

Gómez-Robles, A. (2019). Dental evolutionary rates and its implications for the Neanderthal–modern human divergence. *Science Advances*, *5*(5), eaaw1268. https://doi.org/10.1126/sciadv.aaw1268

Goodwin, S., McPherson, J. D., & McCombie, W. R. (2016). Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews. Genetics*, *17*(6), 333–351. https://doi.org/10.1038/nrg.2016.49

Gordon, D., Huddleston, J., Chaisson, M. J. P., Hill, C. M., Kronenberg, Z. N., Munson, K. M., Malig, M., Raja, A., Fiddes, I., Hillier, L. W., Dunn, C., Baker, C., Armstrong, J., Diekhans, M., Paten, B., Shendure, J., Wilson, R. K., Haussler, D., Chin, C.-S., & Eichler, E. E. (2016). Long-read sequence assembly of the gorilla genome. *Science*, *352*(6281), aae0344. https://doi.org/10.1126/science.aae0344

Green, R. E., Krause, J., Briggs, A. W., Maricic, T., Stenzel, U., Kircher, M., Patterson, N., Li, H., Zhai, W., Fritz, M. H.-Y., Hansen, N. F., Durand, E. Y., Malaspinas, A.-S., Jensen, J. D., Marques-Bonet, T., Alkan, C., Prüfer, K., Meyer, M., Burbano, H. A., … Pääbo, S. (2010). A draft sequence of the Neandertal genome. *Science*, *328*(5979), 710–722. https://doi.org/10.1126/science.1188021

Grüning, B. A., Lampa, S., Vaudel, M., & Blankenberg, D. (2019). Software engineering for scientific big data analysis. *GigaScience*, *8*(5). https://doi.org/10.1093/gigascience/giz054

GTEx Consortium. (2020). The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science*,

369(6509), 1318–1330. https://doi.org/10.1126/science.aaz1776

Guerrier, S., Coutinho-Budd, J., Sassa, T., Gresset, A., Jordan, N. V., Chen, K., Jin, W.-L., Frost, A., & Polleux, F. (2009). The F-BAR domain of srGAP2 induces membrane protrusions required for neuronal migration and morphogenesis. *Cell*, *138*(5), 990–1004. https://doi.org/10.1016/j.cell.2009.06.047

Guipponi, M., Tapparel, C., Jousson, O., Scamuffa, N., Mas, C., Rossier, C., Hutter, P., Meda, P., Lyle, R., Reymond, A., & Antonarakis, S. E. (2001). The murine orthologue of the Golgi-localized TPTE protein provides clues to the evolutionary history of the human TPTE gene family. *Human Genetics*, *109*(6), 569–575. https://doi.org/10.1007/s004390100607

Guipponi, M., Yaspo, M. L., Riesselman, L., Chen, H., De Sario, A., Roizès, G., & Antonarakis, S. E. (2000). Genomic structure of a copy of the human TPTE gene which encompasses 87 kb on the short arm of chromosome 21. *Human Genetics*, *107*(2), 127–131. https://doi.org/10.1007/s004390000343

Gulko, B., Hubisz, M. J., Gronau, I., & Siepel, A. (2015). A method for calculating probabilities of fitness consequences for point mutations across the human genome. *Nature Genetics*, *47*(3), 276–283. https://doi.org/10.1038/ng.3196

Gürünlüoğlu, K., Koç, A., Durmuş, K., Gözükara Bağ, H., Ceran, C., Gürünlüoğlu, S., Yıldız, T., Gül, M., & Demircan, M. (2020). Whole exome sequencing analysis for mutations in isolated type III biliary atresia patients. *Clinical and Experimental Hepatology*, *6*(4), 347–353. https://doi.org/10.5114/ceh.2020.102156

Halldorsson, B. V., Eggertsson, H. P., Moore, K. H. S., Hauswedell, H., Eiriksson, O., Ulfarsson, M. O., Palsson, G., Hardarson, M. T., Oddsson, A., Jensson, B. O., Kristmundsdottir, S., Sigurpalsdottir, B. D., Stefansson, O. A., Beyter, D., Holley, G., Tragante, V., Gylfason, A., Olason, P. I., Zink, F., … Stefansson, K. (2022). The sequences of 150,119 genomes in the UK Biobank. *Nature*, 1–9. https://doi.org/10.1038/s41586-022-04965-x

Handsaker, R. E., Van Doren, V., Berman, J. R., Genovese, G., Kashin, S., Boettger, L. M., & McCarroll, S. A. (2015). Large multiallelic copy number variations in humans. *Nature Genetics*, *47*(3), 296–303. https://doi.org/10.1038/ng.3200

Hao, Y., Hao, S., Andersen-Nissen, E., Mauck, W. M., 3rd, Zheng, S., Butler, A., Lee, M. J., Wilk, A. J., Darby, C., Zager, M., Hoffman, P., Stoeckius, M., Papalexi, E., Mimitou, E. P., Jain, J., Srivastava, A., Stuart, T., Fleming, L. M., Yeung, B., … Satija, R. (2021). Integrated analysis of multimodal single-cell data. *Cell*,

184(13), 3573-3587.e29. https://doi.org/10.1016/j.cell.2021.04.048

Hardwick, R. J., Ménard, A., Sironi, M., Milet, J., Garcia, A., Sese, C., Yang, F., Fu, B., Courtin, D., & Hollox, E. J. (2014). Haptoglobin (HP) and Haptoglobin-related protein (HPR) copy number variation, natural selection, and trypanosomiasis. *Human Genetics*, *133*(1), 69–83. https://doi.org/10.1007/s00439-013-1352-x

Hartasánchez, D. A., Brasó-Vives, M., Heredia-Genestar, J. M., Pybus, M., & Navarro, A. (2018). Effect of Collapsed Duplications on Diversity Estimates: What to Expect. *Genome Biology and Evolution*, *10*(11), 2899–2905. https://doi.org/10.1093/gbe/evy223

Hastie, A. R., Lam, E. T., Pang, A. W. C., Zhang, X., Andrews, W., Lee, J., Liang, T. Y., Wang, J., Zhou, X., Zhu, Z., Anantharaman, T., Džakula, Ž., Bocklandt, S., Surti, U., Saghbini, M., Austin, M. D., Borodkin, M., Erik Holmlin, R., & Cao, H. (2017). Rapid Automated Large Structural Variation Detection in a Diploid Genome by NanoChannel Based Next-Generation Mapping. In *bioRxiv* (p. 102764). https://doi.org/10.1101/102764

Hauer, P. (2018, July 31). *Code Review Guidelines for Humans*. https://phauer.com/2018/code-review-guidelines/

Havrilla, J. M., Pedersen, B. S., Layer, R. M., & Quinlan, A. R. (2019). A map of constrained coding regions in the human genome. *Nature Genetics*, *51*(1), 88–95. https://doi.org/10.1038/s41588-018-0294-6

Heft, I. E., Mostovoy, Y., Levy-Sakin, M., Ma, W., Stevens, A. J., Pastor, S., McCaffrey, J., Boffelli, D., Martin, D. I., Xiao, M., Kennedy, M. A., Kwok, P.-Y., & Sikela, J. M. (2020). The Driver of Extreme Human-Specific Olduvai Repeat Expansion Remains Highly Active in the Human Genome. *Genetics*, *214*(1), 179–191. https://doi.org/10.1534/genetics.119.302782

Hehir-Kwa, J. Y., Marschall, T., Kloosterman, W. P., Francioli, L. C., Baaijens, J. A., Dijkstra, L. J., Abdellaoui, A., Koval, V., Thung, D. T., Wardenaar, R., Renkens, I., Coe, B. P., Deelen, P., de Ligt, J., Lameijer, E.-W., van Dijk, F., Hormozdiari, F., Genome of the Netherlands Consortium, Uitterlinden, A. G., … Guryev, V. (2016). A high-quality human reference panel reveals the complexity and distribution of genomic structural variants. *Nature Communications*, *7*, 12989. https://doi.org/10.1038/ncomms12989

Himmelstein, D. S., Rubinetti, V., Slochower, D. R., Hu, D., Malladi, V. S., Greene, C. S., & Gitter, A. (2019). Open collaborative writing with Manubot. *PLoS Computational Biology*, *15*(6), e1007128. https://doi.org/10.1371/journal.pcbi.1007128

Hinds, D. A., Kloek, A. P., Jen, M., Chen, X., & Frazer, K. A. (2006). Common deletions and SNPs are in linkage disequilibrium in the human genome. *Nature Genetics*, *38*(1), 82–85. https://doi.org/10.1038/ng1695

Hinrichs, A. S., Karolchik, D., Baertsch, R., Barber, G. P., Bejerano, G., Clawson, H., Diekhans, M., Furey, T. S., Harte, R. A., Hsu, F., Hillman-Jackson, J., Kuhn, R. M., Pedersen, J. S., Pohl, A., Raney, B. J., Rosenbloom, K. R., Siepel, A., Smith, K. E., Sugnet, C. W., … Kent, W. J. (2006). The UCSC Genome Browser Database: update 2006. *Nucleic Acids Research*, *34*(Database issue), D590-8. https://doi.org/10.1093/nar/gkj144

Ho, S. S., Urban, A. E., & Mills, R. E. (2019). Structural variation in the sequencing era. *Nature Reviews. Genetics*. https://doi.org/10.1038/s41576-019-0180-9

Hollox, E. J., Armour, J. A. L., & Barber, J. C. K. (2003). Extensive Normal Copy Number Variation of a β-Defensin Antimicrobial-Gene Cluster. *American Journal of Human Genetics*, *73*(3), 591–600. https://doi.org/10.1086/378157

Hollox, Edward J., Zuccherato, L. W., & Tucci, S. (2022). Genome structural variation in human evolution. *Trends in Genetics: TIG*, *38*(1), 45–58. https://doi.org/10.1016/j.tig.2021.06.015

Hoyt, S. J., Storer, J. M., Hartley, G. A., Grady, P. G. S., Gershman, A., de Lima, L. G., Limouse, C., Halabian, R., Wojenski, L., Rodriguez, M., Altemose, N., Rhie, A., Core, L. J., Gerton, J. L., Makalowski, W., Olson, D., Rosen, J., Smit, A. F. A., Straight, A. F., … O'Neill, R. J. (2022). From telomere to telomere: The transcriptional and epigenetic state of human repeat elements. *Science*, *376*(6588), eabk3112. https://doi.org/10.1126/science.abk3112

Hsieh, P., Dang, V., Vollger, M. R., Mao, Y., Huang, T.-H., Dishuck, P. C., Baker, C., Cantsilieris, S., Lewis, A. P., Munson, K. M., Sorensen, M., Welch, A. E., Underwood, J. G., & Eichler, E. E. (2021). Evidence for opposing selective forces operating on human-specific duplicated TCAF genes in Neanderthals and humans. *Nature Communications*, *12*(1), 1–14. https://doi.org/10.1038/s41467-021-25435-4

Hsieh, P., Vollger, M. R., Dang, V., Porubsky, D., Baker, C., Cantsilieris, S., Hoekzema, K., Lewis, A. P., Munson, K. M., Sorensen, M., Kronenberg, Z. N., Murali, S., Nelson, B. J., Chiatante, G., Maggiolini, F. A. M., Blanché, H., Underwood, J. G., Antonacci, F., Deleuze, J.-F., & Eichler, E. E. (2019). Adaptive archaic introgression of copy number variants and the discovery of previously unknown human genes. *Science*, *366*(6463). https://doi.org/10.1126/science.aax2083

Huang, D. W., Sherman, B. T., & Lempicki, R. A. (2009a). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protocols*, *4*(1), 44–57. https://doi.org/10.1038/nprot.2008.211

Huang, D. W., Sherman, B. T., & Lempicki, R. A. (2009b). Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Research*, *37*(1), 1–13. https://doi.org/10.1093/nar/gkn923

Huddleston, J., Ranade, S., Malig, M., Antonacci, F., Chaisson, M., Hon, L., Sudmant, P. H., Graves, T. A., Alkan, C., Dennis, M. Y., Wilson, R. K., Turner, S. W., Korlach, J., & Eichler, E. E. (2014). Reconstructing complex regions of genomes using long-read sequencing technology. *Genome Research*, *24*(4), 688–696. https://doi.org/10.1101/gr.168450.113

Hudson, R. R., Kreitman, M., & Aguadé, M. (1987). A test of neutral molecular evolution based on nucleotide data. *Genetics*, *116*(1), 153–159. https://www.ncbi.nlm.nih.gov/pubmed/3110004

Huerta, M., Seto, B., & Liu, Y. (2000). *Nih working definition of bioinformatics and computational biology*. https://www.kennedykrieger.org/sites/default/files/library/documents/research/center-labs-cores/bioinformatics/bioinformatics-def.pdf

Huynh, L., & Hormozdiari, F. (2019). TAD fusion score: discovery and ranking the contribution of deletions to genome structure. *Genome Biology*, *20*(1), 60. https://doi.org/10.1186/s13059-019-1666-7

Inoue, K., & Lupski, J. R. (2002). Molecular mechanisms for genomic disorders. *Annual Review of Genomics and Human Genetics*, *3*, 199–242. https://doi.org/10.1146/annurev.genom.3.032802.120023

International Human Genome Sequencing Consortium. (2004). Finishing the euchromatic sequence of the human genome. *Nature*, *431*(7011), 931–945. https://doi.org/10.1038/nature03001

Iossifov, I., O'Roak, B. J., Sanders, S. J., Ronemus, M., Krumm, N., Levy, D., Stessman, H. A., Witherspoon, K. T., Vives, L., Patterson, K. E., Smith, J. D., Paeper, B., Nickerson, D. A., Dea, J., Dong, S., Gonzalez, L. E., Mandell, J. D., Mane, S. M., Murtha, M. T., … Wigler, M. (2014). The contribution of de novo coding mutations to autism spectrum disorder. *Nature*, *515*(7526), 216–221. https://doi.org/10.1038/nature13908

Iskow, R. C., Gokcumen, O., Abyzov, A., Malukiewicz, J., Zhu, Q., Sukumar, A. T., Pai, A. A., Mills, R. E., Habegger, L., Cusanovich, D. A., Rubel, M. A., Perry, G. H., Gerstein, M., Stone, A. C., Gilad, Y., & Lee, C. (2012). Regulatory element copy number differences shape primate expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(31), 12656–12661. https://doi.org/10.1073/pnas.1205199109

Jaganathan, K., Kyriazopoulou Panagiotopoulou, S., McRae, J. F., Darbandi, S. F., Knowles, D., Li, Y. I., Kosmicki,

J. A., Arbelaez, J., Cui, W., Schwartz, G. B., Chow, E. D., Kanterakis, E., Gao, H., Kia, A., Batzoglou, S., Sanders, S. J., & Farh, K. K.-H. (2019). Predicting Splicing from Primary Sequence with Deep Learning. *Cell*, *176*(3), 535-548.e24. https://doi.org/10.1016/j.cell.2018.12.015

Jain, C., Rhie, A., Zhang, H., Chu, C., Walenz, B. P., Koren, S., & Phillippy, A. M. (2020). Weighted minimizer sampling improves long read mapping. *Bioinformatics* , *36*(Suppl_1), i111–i118. https://doi.org/10.1093/bioinformatics/btaa435

Jain, M., Koren, S., Miga, K. H., Quick, J., Rand, A. C., Sasani, T. A., Tyson, J. R., Beggs, A. D., Dilthey, A. T., Fiddes, I. T., Malla, S., Marriott, H., Nieto, T., O'Grady, J., Olsen, H. E., Pedersen, B. S., Rhie, A., Richardson, H., Quinlan, A. R., … Loose, M. (2018). Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nature Biotechnology*, *36*(4), 338–345. https://doi.org/10.1038/nbt.4060

Jakobsson, M., Scholz, S. W., Scheet, P., Gibbs, J. R., VanLiere, J. M., Fung, H.-C., Szpiech, Z. A., Degnan, J. H., Wang, K., Guerreiro, R., Bras, J. M., Schymick, J. C., Hernandez, D. G., Traynor, B. J., Simon-Sanchez, J., Matarin, M., Britton, A., van de Leemput, J., Rafferty, I., … Singleton, A. B. (2008). Genotype, haplotype and copy-number variation in worldwide human populations. *Nature*, *451*(7181), 998–1003. https://doi.org/10.1038/nature06742

Jakubosky, D., Smith, E. N., D'Antonio, M., Jan Bonder, M., Young Greenwald, W. W., D'Antonio-Chronowska, A., Matsui, H., Bonder, M. J., Cai, N., Carcamo-Orive, I., D'Antonio, M., Frazer, K. A., Young Greenwald, W. W., Jakubosky, D., Knowles, J. W., Matsui, H., McCarthy, D. J., Mirauta, B. A., Montgomery, S. B., … i2QTL Consortium. (2020). Discovery and quality analysis of a comprehensive set of structural variants and short tandem repeats. *Nature Communications*, *11*(1), 2928. https://doi.org/10.1038/s41467-020-16481-5

Jegadesan, N. K., & Branzei, D. (2021). DDX11 loss causes replication stress and pharmacologically exploitable DNA repair defects. *Proceedings of the National Academy of Sciences of the United States of America*, *118*(17). https://doi.org/10.1073/pnas.2024258118

Jiang, Z., Tang, H., Ventura, M., Cardone, M. F., Marques-Bonet, T., She, X., Pevzner, P. A., & Eichler, E. E. (2007). Ancestral reconstruction of segmental duplications reveals punctuated cores of human genome evolution. *Nature Genetics*, *39*(11), 1361–1368. https://doi.org/10.1038/ng.2007.9

Johansen, N., & Quon, G. (2019). scAlign: a tool for alignment, integration, and rare cell identification from scRNA-seq data. *Genome Biology*, *20*(1), 166. https://doi.org/10.1186/s13059-019-1766-4

Johnson, M. E., Viggiano, L., Bailey, J. A., Abdul-Rauf, M., Goodwin, G., Rocchi, M., & Eichler, E. E. (2001). Positive selection of a gene family during the emergence of humans and African apes. *Nature*, *413*(6855), 514–519. https://doi.org/10.1038/35097067

Johnson, Matthew E., National Institute of Health Intramural Sequencing Center Comparative Sequencing Program, Cheng, Z., Morrison, V. A., Scherer, S., Ventura, M., Gibbs, R. A., Green, E. D., & Eichler, E. E. (2006). Recurrent duplication-driven transposition of DNA during hominoid evolution. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(47), 17626–17631. https://doi.org/10.1073/pnas.0605426103

Ju, X.-C., Hou, Q.-Q., Sheng, A.-L., Wu, K.-Y., Zhou, Y., Jin, Y., Wen, T., Yang, Z., Wang, X., & Luo, Z.-G. (2016). The hominoid-specific gene TBC1D3 promotes generation of basal neural progenitors and induces cortical folding in mice. *ELife*, *5*. https://doi.org/10.7554/eLife.18197

Kaggle. (2021). *2021 Kaggle Machine Learning & Data Science Survey*. https://www.kaggle.com/competitions/kaggle-survey-2021

Karczewski, K. J., Francioli, L. C., Tiao, G., Cummings, B. B., Alföldi, J., Wang, Q., Collins, R. L., Laricchia, K. M., Ganna, A., Birnbaum, D. P., Gauthier, L. D., Brand, H., Solomonson, M., Watts, N. A., Rhodes, D., Singer-Berk, M., England, E. M., Seaby, E. G., Kosmicki, J. A., … MacArthur, D. G. (2019). Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes. In *bioRxiv* (p. 531210). https://doi.org/10.1101/531210

Karczewski, K. J., Francioli, L. C., Tiao, G., Cummings, B. B., Alföldi, J., Wang, Q., Collins, R. L., Laricchia, K. M., Ganna, A., Birnbaum, D. P., Gauthier, L. D., Brand, H., Solomonson, M., Watts, N. A., Rhodes, D., Singer-Berk, M., England, E. M., Seaby, E. G., Kosmicki, J. A., … MacArthur, D. G. (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*, *581*(7809), 434–443. https://doi.org/10.1038/s41586-020-2308-7

Kazazian, H. H., & Moran, J. V. (2017). Mobile DNA in Health and Disease. *The New England Journal of Medicine*, *377*(4). https://doi.org/10.1056/NEJMra1510092

Kehrer-Sawatzki, H., Sandig, C. A., Goidts, V., & Hameister, H. (2005). Breakpoint analysis of the pericentric inversion between chimpanzee chromosome 10 and the homologous chromosome 12 in humans. *Cytogenetic and Genome Research*, *108*(1–3), 91–97. https://doi.org/10.1159/000080806

Kehrer-Sawatzki, Hildegard, Sandig, C., Chuzhanova, N., Goidts, V., Szamalek, J. M., Tänzer, S., Müller, S., Platzer, M., Cooper, D. N., & Hameister, H. (2005). Breakpoint analysis of the pericentric inversion distinguishing human chromosome 4 from the homologous chromosome in the chimpanzee (Pan troglodytes). *Human Mutation*, *25*(1), 45–55. https://doi.org/10.1002/humu.20116

Kehrer-Sawatzki, Hildegard, Szamalek, J. M., Tänzer, S., Platzer, M., & Hameister, H. (2005). Molecular characterization of the pericentric inversion of chimpanzee chromosome 11 homologous to human chromosome 9. *Genomics*, *85*(5), 542–550. https://doi.org/10.1016/j.ygeno.2005.01.012

Khalilipour, N., Baranova, A., Jebelli, A., Heravi-Moussavi, A., Bruskin, S., & Abbaszadegan, M. R. (2018). Familial Esophageal Squamous Cell Carcinoma with damaging rare/germline mutations in KCNJ12/KCNJ18 and GPRIN2 genes. *Cancer Genetics*, *221*, 46–52. https://doi.org/10.1016/j.cancergen.2017.11.011

Khan, Z., Ford, M. J., Cusanovich, D. A., Mitrano, A., Pritchard, J. K., & Gilad, Y. (2013). Primate transcript and protein expression levels evolve under compensatory selection pressures. *Science*, *342*(6162), 1100–1104. https://doi.org/10.1126/science.1242379

Kidd, J. M., Graves, T., Newman, T. L., Fulton, R., Hayden, H. S., Malig, M., Kallicki, J., Kaul, R., Wilson, R. K., & Eichler, E. E. (2010). A human genome structural variation sequencing resource reveals insights into mutational mechanisms. *Cell*, *143*(5), 837–847. https://doi.org/10.1016/j.cell.2010.10.027

Kidd, J. M., Newman, T. L., Tuzun, E., Kaul, R., & Eichler, E. E. (2007). Population stratification of a common APOBEC gene deletion polymorphism. *PLoS Genetics*, *3*(4), e63. https://doi.org/10.1371/journal.pgen.0030063

Kim, P. M., Lam, H. Y. K., Urban, A. E., Korbel, J. O., Affourtit, J., Grubert, F., Chen, X., Weissman, S., Snyder, M., & Gerstein, M. B. (2008). Analysis of copy number variants and segmental duplications in the human genome: Evidence for a change in the process of formation in recent evolutionary history. *Genome Research*, *18*(12), 1865–1874. https://doi.org/10.1101/gr.081422.108

Kircher, M., Witten, D. M., Jain, P., O'Roak, B. J., Cooper, G. M., & Shendure, J. (2014). A general framework for estimating the relative pathogenicity of human genetic variants. *Nature Genetics*, *46*(3), 310–315. https://doi.org/10.1038/ng.2892

Kirsche, M., Prabhu, G., Sherman, R., Ni, B., Aganezov, S., & Schatz, M. C. (2021). Jasmine: Population-scale structural variant comparison and analysis. *BioRxiv*. https://doi.org/10.1101/2021.05.27.445886

Kloosterman, W. P., Francioli, L. C., Hormozdiari, F., Marschall, T., Hehir-Kwa, J. Y., Abdellaoui, A., Lameijer, E.-W., Moed, M. H., Koval, V., Renkens, I., van Roosmalen, M. J., Arp, P., Karssen, L. C., Coe, B. P., Handsaker, R. E., Suchiman, E. D., Cuppen, E., Thung, D. T., McVey, M., … Guryev, V. (2015). Characteristics of de novo structural changes in the human genome. *Genome Research*, *25*(6), 792–801. https://doi.org/10.1101/gr.185041.114

Korbel, J. O., Urban, A. E., Affourtit, J. P., Godwin, B., Grubert, F., Simons, J. F., Kim, P. M., Palejev, D., Carriero, N. J., Du, L., Taillon, B. E., Chen, Z., Tanzer, A., Saunders, A. C. E., Chi, J., Yang, F., Carter, N. P., Hurles, M. E., Weissman, S. M., … Snyder, M. (2007). Paired-end mapping reveals extensive structural variation in the human genome. *Science*, *318*(5849), 420–426. https://doi.org/10.1126/science.1149504

Kronenberg, Z. N., Fiddes, I. T., Gordon, D., Murali, S., Cantsilieris, S., Meyerson, O. S., Underwood, J. G., Nelson, B. J., Chaisson, M. J. P., Dougherty, M. L., Munson, K. M., Hastie, A. R., Diekhans, M., Hormozdiari, F., Lorusso, N., Hoekzema, K., Qiu, R., Clark, K., Raja, A., … Eichler, E. E. (2018). High-resolution comparative analysis of great ape genomes. *Science*, *360*(6393). https://doi.org/10.1126/science.aar6343

Krusche, P., Trigg, L., Boutros, P. C., Mason, C. E., De La Vega, F. M., Moore, B. L., Gonzalez-Porta, M., Eberle, M. A., Tezak, Z., Lababidi, S., Truty, R., Asimenos, G., Funke, B., Fleharty, M., Chapman, B. A., Salit, M., Zook, J. M., & Global Alliance for Genomics and Health Benchmarking Team. (2019). Best practices for benchmarking germline small-variant calls in human genomes. *Nature Biotechnology*, *37*(5), 555–560. https://doi.org/10.1038/s41587-019-0054-x

Kuderna, L. F. K., Tomlinson, C., Hillier, L. W., Tran, A., Fiddes, I. T., Armstrong, J., Laayouni, H., Gordon, D., Huddleston, J., Garcia Perez, R., Povolotskaya, I., Serres Armero, A., Gómez Garrido, J., Ho, D., Ribeca, P., Alioto, T., Green, R. E., Paten, B., Navarro, A., … Marques-Bonet, T. (2017). A 3-way hybrid approach to generate a new high-quality chimpanzee reference genome (Pan_tro_3.0). *GigaScience*, *6*(11), 1–6. https://doi.org/10.1093/gigascience/gix098

Lalrohlui, F., Zohmingthanga, J., Hruaii, V., Vanlallawma, A., & Senthil Kumar, N. (2021). Whole exome sequencing identifies the novel putative gene variants related with type 2 diabetes in Mizo population, northeast India. *Gene*, *769*, 145229. https://doi.org/10.1016/j.gene.2020.145229

Lam, H. Y. K., Mu, X. J., Stütz, A. M., Tanzer, A., Cayting, P. D., Snyder, M., Kim, P. M., Korbel, J. O., & Gerstein, M. B. (2010). Nucleotide-resolution analysis of structural variants using BreakSeq and a breakpoint library.

Nature Biotechnology, *28*(1), 47–55. https://doi.org/10.1038/nbt.1600

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., … International Human Genome Sequencing Consortium. (2001). Initial sequencing and analysis of the human genome. *Nature*, *409*(6822), 860–921. https://doi.org/10.1038/35057062

Landrum, M. J., Lee, J. M., Benson, M., Brown, G. R., Chao, C., Chitipiralla, S., Gu, B., Hart, J., Hoffman, D., Jang, W., Karapetyan, K., Katz, K., Liu, C., Maddipatla, Z., Malheiro, A., McDaniel, K., Ovetsky, M., Riley, G., Zhou, G., … Maglott, D. R. (2018). ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Research*, *46*(D1), D1062–D1067. https://doi.org/10.1093/nar/gkx1153

Langergraber, K. E., Prüfer, K., Rowney, C., Boesch, C., Crockford, C., Fawcett, K., Inoue, E., Inoue-Muruyama, M., Mitani, J. C., Muller, M. N., Robbins, M. M., Schubert, G., Stoinski, T. S., Viola, B., Watts, D., Wittig, R. M., Wrangham, R. W., Zuberbühler, K., Pääbo, S., & Vigilant, L. (2012). Generation times in wild chimpanzees and gorillas suggest earlier divergence times in great ape and human evolution. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(39), 15716–15721. https://doi.org/10.1073/pnas.1211740109

Larson, J. L., Silver, A. J., Chan, D., Borroto, C., Spurrier, B., & Silver, L. M. (2015). Validation of a high resolution NGS method for detecting spinal muscular atrophy carriers among phase 3 participants in the 1000 Genomes Project. *BMC Medical Genetics*, *16*, 100. https://doi.org/10.1186/s12881-015-0246-2

Law, C. W., Chen, Y., Shi, W., & Smyth, G. K. (2014). voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biology*, *15*(2), R29. https://doi.org/10.1186/gb-2014-15-2-r29

Lazar, N. H., Nevonen, K. A., O'Connell, B., McCann, C., O'Neill, R. J., Green, R. E., Meyer, T. J., Okhovat, M., & Carbone, L. (2018). Epigenetic maintenance of topological domains in the highly rearranged gibbon genome. *Genome Research*, *28*(7), 983–997. https://doi.org/10.1101/gr.233874.117

Leblond, C. S., Le, T.-L., Malesys, S., Cliquet, F., Tabet, A.-C., Delorme, R., Rolland, T., & Bourgeron, T. (2021). Operative list of genes associated with autism and neurodevelopmental disorders based on database review. *Molecular and Cellular Neurosciences*, *113*, 103623. https://doi.org/10.1016/j.mcn.2021.103623

Lee, G., Bacon, S., Bush, I., Fortunato, L., Gavaghan, D., Lestang, T., Morton, C., Robinson, M., Rocca-Serra, P., Sansone, S.-A., & Webb, H. (2021). Barely sufficient practices in scientific computing. *Patterns (New York,*

N.Y.), *2*(2), 100206. https://doi.org/10.1016/j.patter.2021.100206

Lee, H., & Schatz, M. C. (2012). Genomic dark matter: the reliability of short read mapping illustrated by the genome mappability score. *Bioinformatics* , *28*(16), 2097–2105. https://doi.org/10.1093/bioinformatics/bts330

Leffler, E. M., Band, G., Busby, G. B. J., Kivinen, K., Le, Q. S., Clarke, G. M., Bojang, K. A., Conway, D. J., Jallow, M., Sisay-Joof, F., Bougouma, E. C., Mangano, V. D., Modiano, D., Sirima, S. B., Achidi, E., Apinjoh, T. O., Marsh, K., Ndila, C. M., Peshu, N., … Malaria Genomic Epidemiology Network. (2017). Resistance to malaria through structural variation of red blood cell invasion receptors. *Science*, *356*(6343). https://doi.org/10.1126/science.aam6393

Lek, M., Karczewski, K. J., Minikel, E. V., Samocha, K. E., Banks, E., Fennell, T., O'Donnell-Luria, A. H., Ware, J. S., Hill, A. J., Cummings, B. B., Tukiainen, T., Birnbaum, D. P., Kosmicki, J. A., Duncan, L. E., Estrada, K., Zhao, F., Zou, J., Pierce-Hoffman, E., Berghout, J., … Exome Aggregation Consortium. (2016). Analysis of protein-coding genetic variation in 60,706 humans. *Nature*, *536*(7616), 285–291. https://doi.org/10.1038/nature19057

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. In *arXiv [q-bio.GN]*. arXiv. http://arxiv.org/abs/1303.3997

Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* , *34*(18), 3094–3100. https://doi.org/10.1093/bioinformatics/bty191

Li, H., Bloom, J. M., Farjoun, Y., Fleharty, M., Gauthier, L., Neale, B., & MacArthur, D. (2018). A synthetic-diploid benchmark for accurate variant-calling evaluation. *Nature Methods*, *15*(8), 595–597. https://doi.org/10.1038/s41592-018-0054-7

Li, J.-P., Yang, C.-Y., Chuang, H.-C., Lan, J.-L., Chen, D.-Y., Chen, Y.-M., Wang, X., Chen, A. J., Belmont, J. W., & Tan, T.-H. (2014). The phosphatase JKAP/DUSP22 inhibits T-cell receptor signalling and autoimmunity by inactivating Lck. *Nature Communications*, *5*, 3618. https://doi.org/10.1038/ncomms4618

Lin, Y.-L., & Gokcumen, O. (2019). Fine-Scale Characterization of Genomic Structural Variation in the Human Genome Reveals Adaptive and Biomedically Relevant Hotspots. *Genome Biology and Evolution*, *11*(4), 1136–1151. https://doi.org/10.1093/gbe/evz058

Lin, Y.-L., Pavlidis, P., Karakoc, E., Ajay, J., & Gokcumen, O. (2015). The evolution and functional impact of human deletion variants shared with archaic hominin genomes. *Molecular Biology and Evolution*, *32*(4), 1008–1019.

https://doi.org/10.1093/molbev/msu405

Liu, G. E., Ventura, M., Cellamare, A., Chen, L., Cheng, Z., Zhu, B., Li, C., Song, J., & Eichler, E. E. (2009). Analysis of recent segmental duplications in the bovine genome. *BMC Genomics*, *10*, 571. https://doi.org/10.1186/1471-2164-10-571

Locke, D. P., Segraves, R., Carbone, L., Archidiacono, N., Albertson, D. G., Pinkel, D., & Eichler, E. E. (2003). Large-scale variation among human and great ape genomes determined by array comparative genomic hybridization. *Genome Research*, *13*(3), 347–357. https://doi.org/10.1101/gr.1003303

Locke, Devin P., Sharp, A. J., McCarroll, S. A., McGrath, S. D., Newman, T. L., Cheng, Z., Schwartz, S., Albertson, D. G., Pinkel, D., Altshuler, D. M., & Eichler, E. E. (2006). Linkage disequilibrium and heritability of copy-number polymorphisms within duplicated regions of the human genome. *American Journal of Human Genetics*, *79*(2), 275–290. https://doi.org/10.1086/505653

Loman, & Watson. (n.d.). So you want to be a computational biologist? *Nature Biotechnology*. https://idp.nature.com/authorize/casa?redirect_uri=https://www.nature.com/articles/nbt.2740&casa_token=44cj4-BY0W8AAAAA:HW2cnK51yGA49JtE8PAYPQif8ryRxPKx2GKU2N82VL_B73w4NI-mgDOtbSi7oTVxfgh3Lem2yZPRlX1sxg

Lorente-Galdos, B., Bleyhl, J., Santpere, G., Vives, L., Ramírez, O., Hernandez, J., Anglada, R., Cooper, G. M., Navarro, A., Eichler, E. E., & Marques-Bonet, T. (2013). Accelerated exon evolution within primate segmental duplications. *Genome Biology*, *14*(1), R9. https://doi.org/10.1186/gb-2013-14-1-r9

Lupiáñez, D. G., Kraft, K., Heinrich, V., Krawitz, P., Brancati, F., Klopocki, E., Horn, D., Kayserili, H., Opitz, J. M., Laxova, R., Santos-Simarro, F., Gilbert-Dussardier, B., Wittler, L., Borschiwer, M., Haas, S. A., Osterwalder, M., Franke, M., Timmermann, B., Hecht, J., … Mundlos, S. (2015). Disruptions of Topological Chromatin Domains Cause Pathogenic Rewiring of Gene-Enhancer Interactions. In *Cell* (Vol. 161, Issue 5, pp. 1012–1025). https://doi.org/10.1016/j.cell.2015.04.004

Luscombe, N. M., Greenbaum, D., & Gerstein, M. (2001). What is bioinformatics? A proposed definition and overview of the field. *Methods of Information in Medicine*, *40*(04), 346–358. https://doi.org/10.1055/s-0038-1634431

Lynch, M., & Conery, J. S. (2000). The evolutionary fate and consequences of duplicate genes [Review of *The evolutionary fate and consequences of duplicate genes*]. *Science*, *290*(5494), 1151–1155.

https://doi.org/10.1126/science.290.5494.1151

MacArthur, D. G., Balasubramanian, S., Frankish, A., Huang, N., Morris, J., Walter, K., Jostins, L., Habegger, L., Pickrell, J. K., Montgomery, S. B., Albers, C. A., Zhang, Z. D., Conrad, D. F., Lunter, G., Zheng, H., Ayub, Q., DePristo, M. A., Banks, E., Hu, M., … Tyler-Smith, C. (2012). A systematic survey of loss-of-function variants in human protein-coding genes. *Science*, *335*(6070), 823–828. https://doi.org/10.1126/science.1215040

Mahmoud, M., Gobet, N., Cruz-Dávalos, D. I., Mounier, N., Dessimoz, C., & Sedlazeck, F. J. (2019). Structural variant calling: the long and the short of it. *Genome Biology*, *20*(1), 246. https://doi.org/10.1186/s13059-019-1828-7

Mallick, S., Li, H., Lipson, M., Mathieson, I., Gymrek, M., Racimo, F., Zhao, M., Chennagiri, N., Nordenfelt, S., Tandon, A., Skoglund, P., Lazaridis, I., Sankararaman, S., Fu, Q., Rohland, N., Renaud, G., Erlich, Y., Willems, T., Gallo, C., … Reich, D. (2016). The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature*, *538*(7624), 201–206. https://doi.org/10.1038/nature18964

Mantere, T., Kersten, S., & Hoischen, A. (2019). Long-Read Sequencing Emerging in Medical Genetics. *Frontiers in Genetics*, *10*, 426. https://doi.org/10.3389/fgene.2019.00426

Mao, Y., Catacchio, C. R., Hillier, L. W., Porubsky, D., Li, R., Sulovari, A., Fernandes, J. D., Montinaro, F., Gordon, D. S., Storer, J. M., Haukness, M., Fiddes, I. T., Murali, S. C., Dishuck, P. C., Hsieh, P., Harvey, W. T., Audano, P. A., Mercuri, L., Piccolo, I., … Eichler, E. E. (2021). A high-quality bonobo genome refines the analysis of hominid evolution. *Nature*. https://doi.org/10.1038/s41586-021-03519-x

Marques-Bonet, T., & Eichler, E. E. (2009). The evolution of human segmental duplications and the core duplicon hypothesis. *Cold Spring Harbor Symposia on Quantitative Biology*, *74*, 355–362. https://doi.org/10.1101/sqb.2009.74.011

Marques-Bonet, Tomas, Kidd, J. M., Ventura, M., Graves, T. A., Cheng, Z., Hillier, L. W., Jiang, Z., Baker, C., Malfavon-Borja, R., Fulton, L. A., Alkan, C., Aksay, G., Girirajan, S., Siswara, P., Chen, L., Cardone, M. F., Navarro, A., Mardis, E. R., Wilson, R. K., & Eichler, E. E. (2009). A burst of segmental duplications in the genome of the African great ape ancestor. *Nature*, *457*(7231), 877–881. https://doi.org/10.1038/nature07744

McCarroll, S. A., Hadnott, T. N., Perry, G. H., Sabeti, P. C., Zody, M. C., Barrett, J. C., Dallaire, S., Gabriel, S. B., Lee, C., Daly, M. J., Altshuler, D. M., & International HapMap Consortium. (2006). Common deletion

polymorphisms in the human genome. *Nature Genetics*, *38*(1), 86–92. https://doi.org/10.1038/ng1696

McCartney, A. M., Shafin, K., Alonge, M., Bzikadze, A. V., Formenti, G., Fungtammasan, A., Howe, K., Jain, C., Koren, S., Logsdon, G. A., Miga, K. H., Mikheenko, A., Paten, B., Shumate, A., Soto, D. C., Sović, I., Wood, J. M. D., Zook, J. M., Phillippy, A. M., & Rhie, A. (2022). Chasing perfection: validation and polishing strategies for telomere-to-telomere genome assemblies. *Nature Methods*. https://doi.org/10.1038/s41592-022-01440-3

McClintock, B. (1931). *Cytological observations of deficiencies involving known genes, translocations and an inversion in Zea mays*. https://mospace.umsystem.edu/xmlui/bitstream/handle/10355/52974/age000163.pdf?sequence=1

McLaren, W., Gil, L., Hunt, S. E., Riat, H. S., Ritchie, G. R. S., Thormann, A., Flicek, P., & Cunningham, F. (2016). The Ensembl Variant Effect Predictor. *Genome Biology*, *17*(1), 122. https://doi.org/10.1186/s13059-016-0974-4

McLean, C. Y., Reno, P. L., Pollen, A. A., Bassan, A. I., Capellini, T. D., Guenther, C., Indjeian, V. B., Lim, X., Menke, D. B., Schaar, B. T., Wenger, A. M., Bejerano, G., & Kingsley, D. M. (2011). Human-specific loss of regulatory DNA and the evolution of human-specific traits. *Nature*, *471*(7337), 216–219. https://doi.org/10.1038/nature09774

McVicker, G., van de Geijn, B., Degner, J. F., Cain, C. E., Banovich, N. E., Raj, A., Lewellen, N., Myrthil, M., Gilad, Y., & Pritchard, J. K. (2013). Identification of genetic variants that affect histone modifications in human cells. *Science*, *342*(6159), 747–749. https://doi.org/10.1126/science.1242429

Meyer, M., Kircher, M., Gansauge, M.-T., Li, H., Racimo, F., Mallick, S., Schraiber, J. G., Jay, F., Prüfer, K., de Filippo, C., Sudmant, P. H., Alkan, C., Fu, Q., Do, R., Rohland, N., Tandon, A., Siebauer, M., Green, R. E., Bryc, K., … Pääbo, S. (2012). A high-coverage genome sequence from an archaic Denisovan individual. *Science*, *338*(6104), 222–226. https://doi.org/10.1126/science.1224344

Miga, K. H., Newton, Y., Jain, M., Altemose, N., Willard, H. F., & Kent, W. J. (2014). Centromere reference models for human chromosomes X and Y satellite arrays. *Genome Research*, *24*(4), 697–707. https://doi.org/10.1101/gr.159624.113

Miga, K. H., & Wang, T. (2021). The Need for a Human Pangenome Reference Sequence. *Annual Review of Genomics and Human Genetics*, *22*, 81–102. https://doi.org/10.1146/annurev-genom-120120-081921

Miller, C. A., Walker, J. R., Jensen, T. L., Hooper, W. F., Fulton, R. S., Painter, J. S., Sekeres, M. A., Ley, T. J., Spencer, D. H., Goll, J. B., & Walter, M. J. (2021). Failure to detect mutations in U2AF1 due to changes in the GRCh38 reference sequence. *BioRxiv*. https://doi.org/10.1101/2021.05.07.442430

Mills, R. E., Walter, K., Stewart, C., Handsaker, R. E., Chen, K., Alkan, C., Abyzov, A., Yoon, S. C., Ye, K., Cheetham, R. K., Chinwalla, A., Conrad, D. F., Fu, Y., Grubert, F., Hajirasouliha, I., Hormozdiari, F., Iakoucheva, L. M., Iqbal, Z., Kang, S., … 1000 Genomes Project. (2011). Mapping copy number variation by population-scale genome sequencing. *Nature*, *470*(7332), 59–65. https://doi.org/10.1038/nature09708

Mölder, F., Jablonski, K. P., Letcher, B., Hall, M. B., Tomkins-Tinch, C. H., Sochat, V., Forster, J., Lee, S., Twardziok, S. O., Kanitz, A., Wilm, A., Holtgrewe, M., Rahmann, S., Nahnsen, S., & Köster, J. (2021). Sustainable data analysis with Snakemake. *F1000Research*, *10*, 33. https://doi.org/10.12688/f1000research.29032.2

Monajemi, H., Fontijn, R. D., Pannekoek, H., & Horrevoets, A. J. G. (2002). The apolipoprotein L gene cluster has emerged recently in evolution and is expressed in human vascular tissue. *Genomics*, *79*(4), 539–546. https://doi.org/10.1006/geno.2002.6729

Moreno-Igoa, M., Hernández-Charro, B., Bengoa-Alonso, A., Pérez-Juana-del-Casal, A., Romero-Ibarra, C., Nieva-Echebarria, B., & Ramos-Arroyo, M. A. (2015). KANSL1 gene disruption associated with the full clinical spectrum of 17q21.31 microdeletion syndrome. *BMC Medical Genetics*, *16*, 68. https://doi.org/10.1186/s12881-015-0211-0

Munchel, S., Hoang, Y., Zhao, Y., Cottrell, J., Klotzle, B., Godwin, A. K., Koestler, D., Beyerlein, P., Fan, J.-B., Bibikova, M., & Chien, J. (2015). Targeted or whole genome sequencing of formalin fixed tissue samples: potential applications in cancer genomics. In *Oncotarget* (Vol. 6, Issue 28, pp. 25943–25961). https://doi.org/10.18632/oncotarget.4671

Namba, T., Dóczi, J., Pinson, A., Xing, L., Kalebic, N., Wilsch-Bräuninger, M., Long, K. R., Vaid, S., Lauer, J., Bogdanova, A., Borgonovo, B., Shevchenko, A., Keller, P., Drechsel, D., Kurzchalia, T., Wimberger, P., Chinopoulos, C., & Huttner, W. B. (2020). Human-Specific ARHGAP11B Acts in Mitochondria to Expand Neocortical Progenitors by Glutaminolysis. *Neuron*, *105*(5), 867-881.e9. https://doi.org/10.1016/j.neuron.2019.11.027

Navarro Gonzalez, J., Zweig, A. S., Speir, M. L., Schmelter, D., Rosenbloom, K. R., Raney, B. J., Powell, C. C., Nassar, L. R., Maulding, N. D., Lee, C. M., Lee, B. T., Hinrichs, A. S., Fyfe, A. C., Fernandes, J. D.,

Diekhans, M., Clawson, H., Casper, J., Benet-Pagès, A., Barber, G. P., … Kent, W. J. (2021). The UCSC Genome Browser database: 2021 update. *Nucleic Acids Research*, *49*(D1), D1046–D1057. https://doi.org/10.1093/nar/gkaa1070

Newman, T. L., Tuzun, E., Morrison, V. A., Hayden, K. E., Ventura, M., McGrath, S. D., Rocchi, M., & Eichler, E. E. (2005). A genome-wide survey of structural variation between human and chimpanzee. *Genome Research*, *15*(10), 1344–1356. https://doi.org/10.1101/gr.4338005

Nicholas, T. J., Cheng, Z., Ventura, M., Mealey, K., Eichler, E. E., & Akey, J. M. (2009). The genomic architecture of segmental duplications and associated copy number variants in dogs. *Genome Research*, *19*(3), 491–499. https://doi.org/10.1101/gr.084715.108

Nickerson, E., & Nelson, D. L. (1998). Molecular definition of pericentric inversion breakpoints occurring during the evolution of humans and chimpanzees. *Genomics*, *50*(3), 368–372. https://doi.org/10.1006/geno.1998.5332

Noble, W. S. (2009). A quick guide to organizing computational biology projects. *PLoS Computational Biology*, *5*(7), e1000424. https://doi.org/10.1371/journal.pcbi.1000424

Nozawa, M., Kawahara, Y., & Nei, M. (2007). Genomic drift and copy number variation of sensory receptor genes in humans. *Proceedings of the National Academy of Sciences of the United States of America*, *104*(51), 20421–20426. https://doi.org/10.1073/pnas.0709956104

Numanagic, I., Gökkaya, A. S., Zhang, L., Berger, B., Alkan, C., & Hach, F. (2018). Fast characterization of segmental duplications in genome assemblies. *Bioinformatics* , *34*(17), i706–i714. https://doi.org/10.1093/bioinformatics/bty586

Nurk, S., Koren, S., Rhie, A., Rautiainen, M., Bzikadze, A. V., Mikheenko, A., Vollger, M. R., Altemose, N., Uralsky, L., Gershman, A., Aganezov, S., Hoyt, S. J., Diekhans, M., Logsdon, G. A., Alonge, M., Antonarakis, S. E., Borchers, M., Bouffard, G. G., Brooks, S. Y., … Phillippy, A. M. (2022). The complete sequence of a human genome. *Science*, *376*(6588), 44–53. https://doi.org/10.1126/science.abj6987

O'Bleness, M., Searles, V. B., Dickens, C. M., Astling, D., Albracht, D., Mak, A. C., Lai, Y. Y., Lin, C., Chu, C., Graves, T., Kwok, P. Y., Wilson, R. K., & Sikela, J. M. (2014). Finished sequence and assembly of the DUF1220-rich 1q21 region using a haploid human genome. *BMC Genomics*, *15*, 387. https://doi.org/10.1186/1471-2164-15-387

O'Bleness, Majesta, Searles, V. B., Varki, A., Gagneux, P., & Sikela, J. M. (2012). Evolution of genetic and genomic

features unique to the human lineage. *Nature Reviews. Genetics*, *13*(12), 853–866. https://doi.org/10.1038/nrg3336

Ohno, S. (1970). *Evolution by gene duplication*. Springer-Verlag. https://play.google.com/store/books/details?id=sxUDAAAAMAAJ

Olson, M. V. (1999). When less is more: gene loss as an engine of evolutionary change. *American Journal of Human Genetics*, *64*(1), 18–23. https://doi.org/10.1086/302219

OMIM Entry - # 617768 - KLEEFSTRA SYNDROME 2; KLEFS2. (n.d.). Retrieved October 20, 2021, from https://www.omim.org/entry/617768

Pääbo, S. (2014). The Human Condition—A Molecular Approach. In *Cell* (Vol. 157, Issue 1, pp. 216–226). https://doi.org/10.1016/j.cell.2013.12.036

Pajic, P., Lin, Y.-L., Xu, D., & Gokcumen, O. (2016). The psoriasis-associated deletion of late cornified envelope genes LCE3B and LCE3C has been maintained under balancing selection since Human Denisovan divergence. *BMC Evolutionary Biology*, *16*(1), 265. https://doi.org/10.1186/s12862-016-0842-6

Pajic, P., Pavlidis, P., Dean, K., Neznanova, L., Romano, R.-A., Garneau, D., Daugherity, E., Globig, A., Ruhl, S., & Gokcumen, O. (2019). Independent amylase gene copy number bursts correlate with dietary preferences in mammals. *ELife*, *8*. https://doi.org/10.7554/eLife.44628

Parsons, J. D. (1995). Miropeats: graphical DNA sequence comparisons. *Computer Applications in the Biosciences: CABIOS*, *11*(6), 615–619. https://www.ncbi.nlm.nih.gov/pubmed/8808577

Patro, R., Duggal, G., Love, M. I., Irizarry, R. A., & Kingsford, C. (2017). Salmon provides fast and bias-aware quantification of transcript expression. *Nature Methods*, *14*(4), 417–419. https://doi.org/10.1038/nmeth.4197

Patterson, N., Price, A. L., & Reich, D. (2006). Population structure and eigenanalysis. *PLoS Genetics*, *2*(12), e190. https://doi.org/10.1371/journal.pgen.0020190

Patterson, N., Richter, D. J., Gnerre, S., Lander, E. S., & Reich, D. (2006). Genetic evidence for complex speciation of humans and chimpanzees. *Nature*, *441*(7097), 1103–1108. https://doi.org/10.1038/nature04789

Pavlovic, B. J., Blake, L. E., Roux, J., Chavarria, C., & Gilad, Y. (2018). A Comparative Assessment of Human and Chimpanzee iPSC-derived Cardiomyocytes with Primary Heart Tissues. *Scientific Reports*, *8*(1), 15312. https://doi.org/10.1038/s41598-018-33478-9

Perez-Riverol, Y., Gatto, L., Wang, R., Sachsenberg, T., Uszkoreit, J., Leprevost, F. da V., Fufezan, C., Ternent, T.,

Eglen, S. J., Katz, D. S., Pollard, T. J., Konovalov, A., Flight, R. M., Blin, K., & Vizcaíno, J. A. (2016). Ten Simple Rules for Taking Advantage of Git and GitHub. *PLoS Computational Biology*, *12*(7), e1004947. https://doi.org/10.1371/journal.pcbi.1004947

Perkel, J. M. (2020). Why scientists are turning to Rust. *Nature*, *588*(7836), 185–186. https://doi.org/10.1038/d41586-020-03382-2

Perry, G. H., Dominy, N. J., Claw, K. G., Lee, A. S., Fiegler, H., Redon, R., Werner, J., Villanea, F. A., Mountain, J. L., Misra, R., Carter, N. P., Lee, C., & Stone, A. C. (2007). Diet and the evolution of human amylase gene copy number variation. *Nature Genetics*, *39*(10), 1256–1260. https://doi.org/10.1038/ng2123

Pickrell, J. K., Marioni, J. C., Pai, A. A., Degner, J. F., Engelhardt, B. E., Nkadori, E., Veyrieras, J.-B., Stephens, M., Gilad, Y., & Pritchard, J. K. (2010). Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature*, *464*(7289), 768–772. https://doi.org/10.1038/nature08872

Pinto, G., Wiese, I., & Dias, L. F. (2018). How do scientists develop scientific software? An external replication. *2018 IEEE 25th International Conference on Software Analysis, Evolution and Reengineering (SANER)*, 582–591. https://doi.org/10.1109/SANER.2018.8330263

Pollen, A. A., Bhaduri, A., Andrews, M. G., Nowakowski, T. J., Meyerson, O. S., Mostajo-Radji, M. A., Di Lullo, E., Alvarado, B., Bedolli, M., Dougherty, M. L., Fiddes, I. T., Kronenberg, Z. N., Shuga, J., Leyrat, A. A., West, J. A., Bershteyn, M., Lowe, C. B., Pavlovic, B. J., Salama, S. R., … Kriegstein, A. R. (2019). Establishing Cerebral Organoids as Models of Human-Specific Brain Evolution. *Cell*, *176*(4), 743-756.e17. https://doi.org/10.1016/j.cell.2019.01.017

Poplin, R., Ruano-Rubio, V., DePristo, M. A., Fennell, T. J., Carneiro, M. O., Van der Auwera, G. A., Kling, D. E., Gauthier, L. D., Levy-Moonshine, A., Roazen, D., Shakir, K., Thibault, J., Chandran, S., Whelan, C., Lek, M., Gabriel, S., Daly, M. J., Neale, B., MacArthur, D. G., & Banks, E. (2018). Scaling accurate genetic variant discovery to tens of thousands of samples. *BioRxiv*. https://doi.org/10.1101/201178

Prado-Martinez, J., Sudmant, P. H., Kidd, J. M., Li, H., Kelley, J. L., Lorente-Galdos, B., Veeramah, K. R., Woerner, A. E., O'Connor, T. D., Santpere, G., Cagan, A., Theunert, C., Casals, F., Laayouni, H., Munch, K., Hobolth, A., Halager, A. E., Malig, M., Hernandez-Rodriguez, J., … Marques-Bonet, T. (2013). Great ape genetic diversity and population history. *Nature*, *499*(7459), 471–475. https://doi.org/10.1038/nature12228

Prescott, S. L., Srinivasan, R., Marchetto, M. C., Grishina, I., Narvaiza, I., Selleri, L., Gage, F. H., Swigut, T., &

Wysocka, J. (2015). Enhancer divergence and cis-regulatory evolution in the human and chimp neural crest. *Cell*, *163*(1), 68–83. https://doi.org/10.1016/j.cell.2015.08.036

Preston-Werner, T. (n.d.). *Semantic Versioning 2.0.0*. Retrieved September 23, 2022, from https://semver.org/

Pritt, J., Chen, N.-C., & Langmead, B. (2018). FORGe: prioritizing variants for graph genomes. *Genome Biology*, *19*(1), 220. https://doi.org/10.1186/s13059-018-1595-x

Procida, D. (2017). Diátaxis: A systematic framework for technical documentation authoring. *Diátaxis*. https://diataxis.fr/

Prüfer, K., de Filippo, C., Grote, S., Mafessoni, F., Korlević, P., Hajdinjak, M., Vernot, B., Skov, L., Hsieh, P., Peyrégne, S., Reher, D., Hopfe, C., Nagel, S., Maricic, T., Fu, Q., Theunert, C., Rogers, R., Skoglund, P., Chintalapati, M., … Pääbo, S. (2017). A high-coverage Neandertal genome from Vindija Cave in Croatia. *Science*, *358*(6363), 655–658. https://doi.org/10.1126/science.aao1887

Prüfer, K., Racimo, F., Patterson, N., Jay, F., Sankararaman, S., Sawyer, S., Heinze, A., Renaud, G., Sudmant, P. H., de Filippo, C., Li, H., Mallick, S., Dannemann, M., Fu, Q., Kircher, M., Kuhlwilm, M., Lachmann, M., Meyer, M., Ongyerth, M., … Pääbo, S. (2014). The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature*, *505*(7481), 43–49. https://doi.org/10.1038/nature12886

Quinlan, A. R., & Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* , *26*(6), 841–842. https://doi.org/10.1093/bioinformatics/btq033

Ramani, V., Cusanovich, D. A., Hause, R. J., Ma, W., Qiu, R., Deng, X., Blau, C. A., Disteche, C. M., Noble, W. S., Shendure, J., & Duan, Z. (2016). Mapping 3D genome architecture through in situ DNase Hi-C. *Nature Protocols*, *11*(11), 2104–2121. https://doi.org/10.1038/nprot.2016.126

Rao, S. S. P., Huang, S.-C., Glenn St Hilaire, B., Engreitz, J. M., Perez, E. M., Kieffer-Kwon, K.-R., Sanborn, A. L., Johnstone, S. E., Bascom, G. D., Bochkov, I. D., Huang, X., Shamim, M. S., Shin, J., Turner, D., Ye, Z., Omer, A. D., Robinson, J. T., Schlick, T., Bernstein, B. E., … Aiden, E. L. (2017). Cohesin Loss Eliminates All Loop Domains. *Cell*, *171*(2), 305-320.e24. https://doi.org/10.1016/j.cell.2017.09.026

Rao, S. S. P., Huntley, M. H., Durand, N. C., Stamenova, E. K., Bochkov, I. D., Robinson, J. T., Sanborn, A. L., Machol, I., Omer, A. D., Lander, E. S., & Aiden, E. L. (2014). A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*, *159*(7), 1665–1680. https://doi.org/10.1016/j.cell.2014.11.021

Redon, R., Ishikawa, S., Fitch, K. R., Feuk, L., Perry, G. H., Andrews, T. D., Fiegler, H., Shapero, M. H., Carson, A. R., Chen, W., Cho, E. K., Dallaire, S., Freeman, J. L., González, J. R., Gratacòs, M., Huang, J., Kalaitzopoulos, D., Komura, D., MacDonald, J. R., … Hurles, M. E. (2006). Global variation in copy number in the human genome. *Nature*, *444*(7118), 444–454. https://doi.org/10.1038/nature05329

Rees, J. S., Castellano, S., & Andrés, A. M. (2020). The Genomics of Human Local Adaptation. *Trends in Genetics: TIG*, *36*(6), 415–428. https://doi.org/10.1016/j.tig.2020.03.006

Reiter, T., Brooks, P. T., Irber, L., Joslin, S. E. K., Reid, C. M., Scott, C., Brown, C. T., & Pierce-Ward, N. T. (2021). Streamlining data-intensive biology with workflow systems. *GigaScience*, *10*(1). https://doi.org/10.1093/gigascience/giaa140

Rentzsch, P., Schubach, M., Shendure, J., & Kircher, M. (2021). CADD-Splice-improving genome-wide variant effect prediction using deep learning-derived splice scores. *Genome Medicine*, *13*(1), 31. https://doi.org/10.1186/s13073-021-00835-9

Rhie, A., McCarthy, S. A., Fedrigo, O., Damas, J., Formenti, G., Koren, S., Uliano-Silva, M., Chow, W., Fungtammasan, A., Kim, J., Lee, C., Ko, B. J., Chaisson, M., Gedman, G. L., Cantin, L. J., Thibaud-Nissen, F., Haggerty, L., Bista, I., Smith, M., … Jarvis, E. D. (2021). Towards complete and error-free genome assemblies of all vertebrate species. *Nature*, *592*(7856), 737–746. https://doi.org/10.1038/s41586-021-03451-0

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., & Smyth, G. K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, *43*(7), e47. https://doi.org/10.1093/nar/gkv007

Robinson, J. T., Thorvaldsdottir, H., Winckler, W., Guttman, M., Lander, E. S., Getz, G., & Mesirov, J. P. (2011). Integrative genomics viewer. *Nature Biotechnology*, *29*(1), 24–26. https://doi.org/10.1038/nbt.1754

Rochette, C. F., Gilbert, N., & Simard, L. R. (2001). SMN gene duplication and the emergence of the SMN2 gene occurred in distinct hominids: SMN2 is unique to Homo sapiens. *Human Genetics*, *108*(3), 255–266. https://doi.org/10.1007/s004390100473

Rodin, S. N., Parkhomchuk, D. V., & Riggs, A. D. (2005). Epigenetic changes and repositioning determine the evolutionary fate of duplicated genes. *Biochemistry. Biokhimiia*, *70*(5), 559–567. https://doi.org/10.1007/s10541-005-0149-5

Rogers, J., & Gibbs, R. A. (2014). Comparative primate genomics: emerging patterns of genome content and dynamics. *Nature Reviews. Genetics*, *15*(5), 347–359. https://doi.org/10.1038/nrg3707

Rule, A., Birmingham, A., Zuniga, C., Altintas, I., Huang, S.-C., Knight, R., Moshiri, N., Nguyen, M. H., Rosenthal, S. B., Pérez, F., & Rose, P. W. (2018). Ten Simple Rules for Reproducible Research in Jupyter Notebooks. In *arXiv [cs.OH]*. arXiv. http://arxiv.org/abs/1810.08055

Ryan, D. P., da Silva, M. R. D., Soong, T. W., Fontaine, B., Donaldson, M. R., Kung, A. W. C., Jongjaroenprasert, W., Liang, M. C., Khoo, D. H. C., Cheah, J. S., Ho, S. C., Bernstein, H. S., Maciel, R. M. B., Brown, R. H., Jr, & Ptácek, L. J. (2010). Mutations in potassium channel Kir2.6 cause susceptibility to thyrotoxic hypokalemic periodic paralysis. *Cell*, *140*(1), 88–98. https://doi.org/10.1016/j.cell.2009.12.024

Saitou, M., Satta, Y., & Gokcumen, O. (2018). Complex Haplotypes of GSTM1 Gene Deletions Harbor Signatures of a Selective Sweep in East Asian Populations. *G3* , *8*(9), 2953–2966. https://doi.org/10.1534/g3.118.200462

Saitou, M., Satta, Y., Gokcumen, O., & Ishida, T. (2018). Complex evolution of the GSTM gene family involves sharing of GSTM1 deletion polymorphism in humans and chimpanzees. *BMC Genomics*, *19*(1), 293. https://doi.org/10.1186/s12864-018-4676-z

Saitou, Marie, & Gokcumen, O. (2019a). Resolving the Insertion Sites of Polymorphic Duplications Reveals a HERC2 Haplotype under Selection. *Genome Biology and Evolution*, *11*(6), 1679–1690. https://doi.org/10.1093/gbe/evz107

Saitou, Marie, & Gokcumen, O. (2019b). An Evolutionary Perspective on the Impact of Genomic Copy Number Variation on Human Health. *Journal of Molecular Evolution*. https://doi.org/10.1007/s00239-019-09911-6

Saitou, Marie, Masuda, N., & Gokcumen, O. (2021). Similarity-based analysis of allele frequency distribution among multiple populations identifies adaptive genomic structural variants. *Molecular Biology and Evolution*. https://doi.org/10.1093/molbev/msab313

Samocha, K. E., Robinson, E. B., Sanders, S. J., Stevens, C., Sabo, A., McGrath, L. M., Kosmicki, J. A., Rehnström, K., Mallick, S., Kirby, A., Wall, D. P., MacArthur, D. G., Gabriel, S. B., DePristo, M., Purcell, S. M., Palotie, A., Boerwinkle, E., Buxbaum, J. D., Cook, E. H., Jr, … Daly, M. J. (2014). A framework for the interpretation of de novo mutation in human disease. *Nature Genetics*, *46*(9), 944–950. https://doi.org/10.1038/ng.3050

Santos, D., Mahtab, M., Boavida, A., & Pisani, F. M. (2021). Role of the DDX11 DNA Helicase in Warsaw Breakage Syndrome Etiology. *International Journal of Molecular Sciences*, *22*(5).

https://doi.org/10.3390/ijms22052308

Schatz, M. C., Philippakis, A. A., Afgan, E., Banks, E., Carey, V. J., Carroll, R. J., Culotti, A., Ellrott, K., Goecks, J., Grossman, R. L., Hall, I., Hansen, K. D., Lawson, J., Leek, J. T., Luria, A. O., Mosher, S., Morgan, M., Nekrutenko, A., O'Connor, B. D., … AnVIL Team. (2021). Inverting the model of genomics data sharing with the NHGRI Genomic Data Science Analysis, Visualization, and Informatics Lab-space (AnVIL). *BioRxiv*. https://doi.org/10.1101/2021.04.22.436044

Schatz, M. C., Philippakis, A. A., Afgan, E., Banks, E., Carey, V. J., Carroll, R. J., Culotti, A., Ellrott, K., Goecks, J., Grossman, R. L., Hall, I. M., Hansen, K. D., Lawson, J., Leek, J. T., Luria, A. O., Mosher, S., Morgan, M., Nekrutenko, A., O'Connor, B. D., … Wuichet, K. (2022). Inverting the model of genomics data sharing with the NHGRI Genomic Data Science Analysis, Visualization, and Informatics Lab-space. *Cell Genomics*, *2*(1). https://doi.org/10.1016/j.xgen.2021.100085

Schmidt, J. M., de Manuel, M., Marques-Bonet, T., Castellano, S., & Andrés, A. M. (2019). The impact of genetic adaptation on chimpanzee subspecies differentiation. *PLoS Genetics*, *15*(11), e1008485. https://doi.org/10.1371/journal.pgen.1008485

Schneider, V. A., Graves-Lindsay, T., Howe, K., Bouk, N., Chen, H.-C., Kitts, P. A., Murphy, T. D., Pruitt, K. D., Thibaud-Nissen, F., Albracht, D., Fulton, R. S., Kremitzki, M., Magrini, V., Markovic, C., McGrath, S., Steinberg, K. M., Auger, K., Chow, W., Collins, J., … Church, D. M. (2017). Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. *Genome Research*, *27*(5), 849–864. https://doi.org/10.1101/gr.213611.116

Scott, A. J., Chiang, C., & Hall, I. M. (2021). Structural variants are a major source of gene expression differences in humans and often affect multiple nearby genes. In *Cold Spring Harbor Laboratory* (p. 2021.03.06.434233). https://doi.org/10.1101/2021.03.06.434233

Sedlazeck, F. J., Lee, H., Darby, C. A., & Schatz, M. C. (2018). Piercing the dark matter: bioinformatics of long-range sequencing and mapping. *Nature Reviews. Genetics*, *19*(6), 329–346. https://doi.org/10.1038/s41576-018-0003-4

Sedlazeck, F. J., Rescheneder, P., Smolka, M., Fang, H., Nattestad, M., von Haeseler, A., & Schatz, M. C. (2018). Accurate detection of complex structural variations using single-molecule sequencing. *Nature Methods*, *15*(6), 461–468. https://doi.org/10.1038/s41592-018-0001-7

Seo, J.-S., Rhie, A., Kim, J., Lee, S., Sohn, M.-H., Kim, C.-U., Hastie, A., Cao, H., Yun, J.-Y., Kim, J., Kuk, J., Park, G. H., Kim, J., Ryu, H., Kim, J., Roh, M., Baek, J., Hunkapiller, M. W., Korlach, J., … Kim, C. (2016). De novo assembly and phasing of a Korean human genome. *Nature*, *538*(7624), 243–247. https://doi.org/10.1038/nature20098

Shafin, K., Pesout, T., Chang, P.-C., Nattestad, M., Kolesnikov, A., Goel, S., Baid, G., Eizenga, J. M., Miga, K. H., Carnevali, P., Jain, M., Carroll, A., & Paten, B. (2021). Haplotype-aware variant calling enables high accuracy in nanopore long-reads using deep neural networks. *BioRxiv*. https://doi.org/10.1101/2021.03.04.433952

Shafin, K., Pesout, T., Lorig-Roach, R., Haukness, M., Olsen, H. E., Bosworth, C., Armstrong, J., Tigyi, K., Maurer, N., Koren, S., Sedlazeck, F. J., Marschall, T., Mayes, S., Costa, V., Zook, J. M., Liu, K. J., Kilburn, D., Sorensen, M., Munson, K. M., … Paten, B. (2020). Nanopore sequencing and the Shasta toolkit enable efficient de novo assembly of eleven human genomes. *Nature Biotechnology*. https://doi.org/10.1038/s41587-020-0503-6

Sharp, A. J., Locke, D. P., McGrath, S. D., Cheng, Z., Bailey, J. A., Vallente, R. U., Pertz, L. M., Clark, R. A., Schwartz, S., Segraves, R., Oseroff, V. V., Albertson, D. G., Pinkel, D., & Eichler, E. E. (2005). Segmental duplications and copy-number variation in the human genome. *American Journal of Human Genetics*, *77*(1), 78–88. https://doi.org/10.1086/431652

She, X., Cheng, Z., Zöllner, S., Church, D. M., & Eichler, E. E. (2008). Mouse segmental duplication and copy number variation. *Nature Genetics*, *40*(7), 909–914. https://doi.org/10.1038/ng.172

Shebanits, K., Andersson-Assarsson, J. C., Larsson, I., Carlsson, L. M. S., Feuk, L., & Larhammar, D. (2018). Copy number of pancreatic polypeptide receptor gene NPY4R correlates with body mass index and waist circumference. *PloS One*, *13*(4), e0194668. https://doi.org/10.1371/journal.pone.0194668

Shen, F., & Kidd, J. M. (2020). Rapid, Paralog-Sensitive CNV Analysis of 2457 Human Genomes Using QuicK-mer2. *Genes*, *11*(2), 141. https://doi.org/10.3390/genes11020141

Sheng, J.-J., & Jin, J.-P. (2016). TNNI1, TNNI2 and TNNI3: Evolution, regulation, and protein structure-function relationships. *Gene*, *576*(1 Pt 3), 385–394. https://doi.org/10.1016/j.gene.2015.10.052

Sherman, R. M., Forman, J., Antonescu, V., Puiu, D., Daya, M., Rafaels, N., Boorgula, M. P., Chavan, S., Vergara, C., Ortega, V. E., Levin, A. M., Eng, C., Yazdanbakhsh, M., Wilson, J. G., Marrugo, J., Lange, L. A.,

Williams, L. K., Watson, H., Ware, L. B., … Salzberg, S. L. (2019). Assembly of a pan-genome from deep sequencing of 910 humans of African descent. *Nature Genetics*, *51*(1), 30–35. https://doi.org/10.1038/s41588-018-0273-y

Sherry, S. T., Ward, M., & Sirotkin, K. (1999). dbSNP-database for single nucleotide polymorphisms and other classes of minor genetic variation. *Genome Research*, *9*(8), 677–679. https://www.ncbi.nlm.nih.gov/pubmed/10447503

Shew, C. J., Carmona-Mora, P., Soto, D. C., Mastoras, M., Roberts, E., Rosas, J., Jagannathan, D., Kaya, G., O'Geene, H., & Dennis, M. Y. (2021). Diverse molecular mechanisms contribute to differential expression of human duplicated genes. *Molecular Biology and Evolution*. https://doi.org/10.1093/molbev/msab131

Shi, L., Guo, Y., Dong, C., Huddleston, J., Yang, H., Han, X., Fu, A., Li, Q., Li, N., Gong, S., Lintner, K. E., Ding, Q., Wang, Z., Hu, J., Wang, D., Wang, F., Wang, L., Lyon, G. J., Guan, Y., … Wang, K. (2016). Long-read sequencing and de novo assembly of a Chinese genome. *Nature Communications*, *7*, 12065. https://doi.org/10.1038/ncomms12065

Shimada, M. K., Kim, C.-G., Kitano, T., Ferrell, R. E., Kohara, Y., & Saitou, N. (2005). Nucleotide sequence comparison of a chromosome rearrangement on human chromosome 12 and the corresponding ape chromosomes. *Cytogenetic and Genome Research*, *108*(1–3), 83–90. https://doi.org/10.1159/000080805

Shin, H., Shi, Y., Dai, C., Tjong, H., Gong, K., Alber, F., & Zhou, X. J. (2016). TopDom: an efficient and deterministic method for identifying topological domains in genomes. *Nucleic Acids Research*, *44*(7), e70. https://doi.org/10.1093/nar/gkv1505

Sirén, J., Monlong, J., Chang, X., Novak, A. M., Eizenga, J. M., Markello, C., Sibbesen, J. A., Hickey, G., Chang, P.-C., Carroll, A., Gupta, N., Gabriel, S., Blackwell, T. W., Ratan, A., Taylor, K. D., Rich, S. S., Rotter, J. I., Haussler, D., Garrison, E., & Paten, B. (2021). Pangenomics enables genotyping of known structural variants in 5202 diverse genomes. *Science*, *374*(6574), abg8871. https://doi.org/10.1126/science.abg8871

Soneson, C., Love, M. I., & Robinson, M. D. (2015). Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. *F1000Research*, *4*, 1521. https://doi.org/10.12688/f1000research.7563.2

Soto, D. C., Shew, C., Mastoras, M., Schmidt, J. M., Sahasrabudhe, R., Kaya, G., Andrés, A. M., & Dennis, M. Y. (2020). Identification of Structural Variation in Chimpanzees Using Optical Mapping and Nanopore Sequencing. *Genes*, *11*(3), 276. https://doi.org/10.3390/genes11030276

Sousa, A. M. M., Meyer, K. A., Santpere, G., Gulden, F. O., & Sestan, N. (2017). Evolution of the Human Nervous System Function, Structure, and Development. *Cell*, *170*(2), 226–247. https://doi.org/10.1016/j.cell.2017.06.036

Spielmann, M., Lupiáñez, D. G., & Mundlos, S. (2018). Structural variation in the 3D genome. *Nature Reviews. Genetics*, *19*(7), 453–467. https://doi.org/10.1038/s41576-018-0007-0

Stack Overflow. (2021). *Stack Overflow Developer Survey 2021*. https://insights.stackoverflow.com/survey/2021

Stankiewicz, P., & Lupski, J. R. (2002). Genome architecture, rearrangements and genomic disorders. *Trends in Genetics: TIG*, *18*(2), 74–82. https://www.ncbi.nlm.nih.gov/pubmed/11818139

Stankiewicz, P., & Lupski, J. R. (2010). Structural variation in the human genome and its role in disease. *Annual Review of Medicine*, *61*, 437–455. https://doi.org/10.1146/annurev-med-100708-204735

Stefansson, H., Helgason, A., Thorleifsson, G., Steinthorsdottir, V., Masson, G., Barnard, J., Baker, A., Jonasdottir, A., Ingason, A., Gudnadottir, V. G., Desnica, N., Hicks, A., Gylfason, A., Gudbjartsson, D. F., Jonsdottir, G. M., Sainz, J., Agnarsson, K., Birgisdottir, B., Ghosh, S., … Stefansson, K. (2005). A common inversion under selection in Europeans. *Nature Genetics*, *37*(2), 129–137. https://doi.org/10.1038/ng1508

Steinberg, K. M., Antonacci, F., Sudmant, P. H., Kidd, J. M., Campbell, C. D., Vives, L., Malig, M., Scheinfeldt, L., Beggs, W., Ibrahim, M., Lema, G., Nyambo, T. B., Omar, S. A., Bodo, J.-M., Froment, A., Donnelly, M. P., Kidd, K. K., Tishkoff, S. A., & Eichler, E. E. (2012). Structural diversity and African origin of the 17q21.31 inversion polymorphism. *Nature Genetics*, *44*(8), 872–880. https://doi.org/10.1038/ng.2335

Steinberg, K. M., Schneider, V. A., Graves-Lindsay, T. A., Fulton, R. S., Agarwala, R., Huddleston, J., Shiryev, S. A., Morgulis, A., Surti, U., Warren, W. C., Church, D. M., Eichler, E. E., & Wilson, R. K. (2014). Single haplotype assembly of the human genome from a hydatidiform mole. *Genome Research*, *24*(12), 2066–2076. https://doi.org/10.1101/gr.180893.114

Stephens, Z. D., Lee, S. Y., Faghri, F., Campbell, R. H., Zhai, C., Efron, M. J., Iyer, R., Schatz, M. C., Sinha, S., & Robinson, G. E. (2015). Big Data: Astronomical or Genomical? *PLoS Biology*, *13*(7), e1002195. https://doi.org/10.1371/journal.pbio.1002195

Sturtevant, A. H. (1913). The linear arrangement of six sex-linked factors in Drosophila, as shown by their mode of association. In *Journal of Experimental Zoology* (Vol. 14, Issue 1, pp. 43–59). https://doi.org/10.1002/jez.1400140104

Sudmant, P. H., Huddleston, J., Catacchio, C. R., Malig, M., Hillier, L. W., Baker, C., Mohajeri, K., Kondova, I., Bontrop, R. E., Persengiev, S., Antonacci, F., Ventura, M., Prado-Martinez, J., Great Ape Genome Project, Marques-Bonet, T., & Eichler, E. E. (2013). Evolution and diversity of copy number variation in the great ape lineage. *Genome Research*, *23*(9), 1373–1382. https://doi.org/10.1101/gr.158543.113

Sudmant, P. H., Kitzman, J. O., Antonacci, F., Alkan, C., Malig, M., Tsalenko, A., Sampas, N., Bruhn, L., Shendure, J., 1000 Genomes Project, & Eichler, E. E. (2010). Diversity of human copy number variation and multicopy genes. *Science*, *330*(6004), 641–646. https://doi.org/10.1126/science.1197005

Sudmant, P. H., Mallick, S., Nelson, B. J., Hormozdiari, F., Krumm, N., Huddleston, J., Coe, B. P., Baker, C., Nordenfelt, S., Bamshad, M., Jorde, L. B., Posukh, O. L., Sahakyan, H., Watkins, W. S., Yepiskoposyan, L., Abdullah, M. S., Bravi, C. M., Capelli, C., Hervig, T., … Eichler, E. E. (2015). Global diversity, population stratification, and selection of human copy-number variation. *Science*, *349*(6253), aab3761. https://doi.org/10.1126/science.aab3761

Sudmant, P. H., Rausch, T., Gardner, E. J., Handsaker, R. E., Abyzov, A., Huddleston, J., Zhang, Y., Ye, K., Jun, G., Fritz, M. H.-Y., Konkel, M. K., Malhotra, A., Stütz, A. M., Shi, X., Casale, F. P., Chen, J., Hormozdiari, F., Dayama, G., Chen, K., … Korbel, J. O. (2015). An integrated map of structural variation in 2,504 human genomes. *Nature*, *526*(7571), 75–81. https://doi.org/10.1038/nature15394

Sung, S. S., Brassington, A.-M. E., Krakowiak, P. A., Carey, J. C., Jorde, L. B., & Bamshad, M. (2003). Mutations in TNNT3 cause multiple congenital contractures: a second locus for distal arthrogryposis type 2B. *American Journal of Human Genetics*, *73*(1), 212–214. https://doi.org/10.1086/376418

Suzuki, I. K., Gacquer, D., Van Heurck, R., Kumar, D., Wojno, M., Bilheu, A., Herpoel, A., Lambert, N., Cheron, J., Polleux, F., Detours, V., & Vanderhaeghen, P. (2018). Human-Specific NOTCH2NL Genes Expand Cortical Neurogenesis through Delta/Notch Regulation. *Cell*, *173*(6), 1370-1384.e16. https://doi.org/10.1016/j.cell.2018.03.067

Svensson, V., Vento-Tormo, R., & Teichmann, S. A. (2018). Exponential scaling of single-cell RNA-seq in the past decade. *Nature Protocols*, *13*(4), 599–604. https://doi.org/10.1038/nprot.2017.149

Szamalek, J. M., Goidts, V., Searle, J. B., Cooper, D. N., Hameister, H., & Kehrer-Sawatzki, H. (2006). The chimpanzee-specific pericentric inversions that distinguish humans and chimpanzees have identical breakpoints in Pan troglodytes and Pan paniscus. *Genomics*, *87*(1), 39–45.

https://doi.org/10.1016/j.ygeno.2005.09.003

Taillon-Miller, P., Bauer-Sardiña, I., Zakeri, H., Hillier, L., Mutch, D. G., & Kwok, P. Y. (1997). The homozygous complete hydatidiform mole: a unique resource for genome studies. *Genomics*, *46*(2), 307–310. https://doi.org/10.1006/geno.1997.5042

Tajima, F. (1989). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*, *123*(3), 585–595. https://doi.org/10.1093/genetics/123.3.585

Takahashi, S., Cui, Y.-H., Han, Y.-H., Fagerness, J. A., Galloway, B., Shen, Y.-C., Kojima, T., Uchiyama, M., Faraone, S. V., & Tsuang, M. T. (2008). Association of SNPs and haplotypes in APOL1, 2 and 4 with schizophrenia. In *Schizophrenia Research* (Vol. 104, Issues 1–3, pp. 153–164). https://doi.org/10.1016/j.schres.2008.05.028

Taliun, D., Harris, D. N., Kessler, M. D., Carlson, J., Szpiech, Z. A., Torres, R., Taliun, S. A. G., Corvelo, A., Gogarten, S. M., Kang, H. M., Pitsillides, A. N., LeFaive, J., Lee, S.-B., Tian, X., Browning, B. L., Das, S., Emde, A.-K., Clarke, W. E., Loesch, D. P., … Abecasis, G. R. (2021). Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature*, *590*(7845), 290–299. https://doi.org/10.1038/s41586-021-03205-y

Tan, Z., Shon, A. M., & Ober, C. (2005). Evidence of balancing selection at the HLA-G promoter region. *Human Molecular Genetics*, *14*(23), 3619–3628. https://doi.org/10.1093/hmg/ddi389

Tapparel, C., Reymond, A., Girardet, C., Guillou, L., Lyle, R., Lamon, C., Hutter, P., & Antonarakis, S. E. (2003). The TPTE gene family: cellular expression, subcellular localization and alternative splicing. *Gene*, *323*, 189–199. https://doi.org/10.1016/j.gene.2003.09.038

Than, Nándor Gábor, Balogh, A., Romero, R., Kárpáti, E., Erez, O., Szilágyi, A., Kovalszky, I., Sammar, M., Gizurarson, S., Matkó, J., Závodszky, P., Papp, Z., & Meiri, H. (2014). Placental Protein 13 (PP13) - A Placental Immunoregulatory Galectin Protecting Pregnancy. *Frontiers in Immunology*, *5*, 348. https://doi.org/10.3389/fimmu.2014.00348

Than, Nandor Gabor, Romero, R., Goodman, M., Weckle, A., Xing, J., Dong, Z., Xu, Y., Tarquini, F., Szilagyi, A., Gal, P., Hou, Z., Tarca, A. L., Kim, C. J., Kim, J.-S., Haidarian, S., Uddin, M., Bohn, H., Benirschke, K., Santolaya-Forgas, J., … Wildman, D. E. (2009). A primate subfamily of galectins expressed at the maternal-fetal interface that promote immune cell death. *Proceedings of the National Academy of Sciences of the*

United States of America, *106*(24), 9731–9736. https://doi.org/10.1073/pnas.0903568106

The 1000 Genomes Project Consortium. (2015). A global reference for human genetic variation. *Nature*, *526*(7571), 68–74. https://doi.org/10.1038/nature15393

Treangen, T. J., & Salzberg, S. L. (2011). Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nature Reviews. Genetics*, *13*(1), 36–46. https://doi.org/10.1038/nrg3117

Van der Auwera, G. A., Carneiro, M. O., Hartl, C., Poplin, R., Del Angel, G., Levy-Moonshine, A., Jordan, T., Shakir, K., Roazen, D., Thibault, J., Banks, E., Garimella, K. V., Altshuler, D., Gabriel, S., & DePristo, M. A. (2013). From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Current Protocols in Bioinformatics / Editoral Board, Andreas D. Baxevanis ... [et Al.]*, *43*, 11.10.1-33. https://doi.org/10.1002/0471250953.bi1110s43

Van der Auwera, G. A., & O'Connor, B. D. (2020). *Genomics in the Cloud: Using Docker, GATK, and WDL in Terra*. "O'Reilly Media, Inc." https://play.google.com/store/books/details?id=vsXaDwAAQBAJ

Varki, A. (2005). Comparing the human and chimpanzee genomes: Searching for needles in a haystack. In *Genome Research* (Vol. 15, Issue 12, pp. 1746–1758). https://doi.org/10.1101/gr.3737405

Varki, Ajit, & Altheide, T. K. (2005). Comparing the human and chimpanzee genomes: Searching for needles in a haystack. In *Genome Research* (Vol. 15, Issue 12, pp. 1746–1758). https://doi.org/10.1101/gr.3737405

Vollger, M. R., DeWitt, W. S., Dishuck, P. C., Harvey, W. T., Guitart, X., Goldberg, M. E., Rozanski, A. N., Lucas, J., Asri, M., The Human Pangenome Reference Consortium, Munson, K. M., Lewis, A. P., Hoekzema, K., Logsdon, G. A., Porubsky, D., Paten, B., Harris, K., Hsieh, P., & Eichler, E. E. (2022). Increased mutation rate and interlocus gene conversion within human segmental duplications. In *bioRxiv* (p. 2022.07.06.498021). https://doi.org/10.1101/2022.07.06.498021

Vollger, M. R., Dishuck, P. C., Sorensen, M., Welch, A. E., Dang, V., Dougherty, M. L., Graves-Lindsay, T. A., Wilson, R. K., Chaisson, M. J. P., & Eichler, E. E. (2019). Long-read sequence and assembly of segmental duplications. *Nature Methods*, *16*(1), 88–94. https://doi.org/10.1038/s41592-018-0236-3

Vollger, M. R., Guitart, X., Dishuck, P. C., Mercuri, L., Harvey, W. T., Gershman, A., Diekhans, M., Sulovari, A., Munson, K. M., Lewis, A. P., Hoekzema, K., Porubsky, D., Li, R., Nurk, S., Koren, S., Miga, K. H., Phillippy, A. M., Timp, W., Ventura, M., & Eichler, E. E. (2022). Segmental duplications and their variation in a complete human genome. *Science*, *376*(6588), eabj6965. https://doi.org/10.1126/science.abj6965

Vollger, M. R., Logsdon, G. A., Audano, P. A., Sulovari, A., Porubsky, D., Peluso, P., Wenger, A. M., Concepcion, G. T., Kronenberg, Z. N., Munson, K. M., Baker, C., Sanders, A. D., Spierings, D. C. J., Lansdorp, P. M., Surti, U., Hunkapiller, M. W., & Eichler, E. E. (2019). Improved assembly and variant detection of a haploid human genome using single-molecule, high-fidelity long reads. *Annals of Human Genetics*. https://doi.org/10.1111/ahg.12364

Voss, K., Gentry, J., & Van der Auwera, G. (2017). *Full-stack genomics pipelining with GATK4 + WDL + Cromwell*. F1000Research. https://doi.org/10.7490/f1000research.1114631.1

Wagner, J., Olson, N. D., Harris, L., McDaniel, J., Cheng, H., Fungtammasan, A., Hwang, Y.-C., Gupta, R., Wenger, A. M., Rowell, W. J., Khan, Z. M., Farek, J., Zhu, Y., Pisupati, A., Mahmoud, M., Xiao, C., Yoo, B., Sahraeian, S. M. E., Miller, D. E., … Sedlazeck, F. J. (2021). Towards a Comprehensive Variation Benchmark for Challenging Medically-Relevant Autosomal Genes. *BioRxiv*. https://doi.org/10.1101/2021.06.07.444885

Walker, S. M., Downes, C. P., & Leslie, N. R. (2001). TPIP: a novel phosphoinositide 3-phosphatase. *Biochemical Journal*, *360*(Pt 2), 277–283. https://doi.org/10.1042/0264-6021:3600277

Wang, T., Antonacci-Fulton, L., Howe, K., Lawson, H. A., Lucas, J. K., Phillippy, A. M., Popejoy, A. B., Asri, M., Carson, C., Chaisson, M. J. P., Chang, X., Cook-Deegan, R., Felsenfeld, A. L., Fulton, R. S., Garrison, E. P., Garrison, N. A., Graves-Lindsay, T. A., Ji, H., Kenny, E. E., … Human Pangenome Reference Consortium. (2022). The Human Pangenome Project: a global resource to map genomic diversity. *Nature*, *604*(7906), 437–446. https://doi.org/10.1038/s41586-022-04601-8

Warren, W. C., Harris, R. A., Haukness, M., Fiddes, I. T., Murali, S. C., Fernandes, J., Dishuck, P. C., Storer, J. M., Raveendran, M., Hillier, L. W., Porubsky, D., Mao, Y., Gordon, D., Vollger, M. R., Lewis, A. P., Munson, K. M., DeVogelaere, E., Armstrong, J., Diekhans, M., … Eichler, E. E. (2020). Sequence diversity analyses of an improved rhesus macaque genome enhance its biomedical utility. *Science*, *370*(6523). https://doi.org/10.1126/science.abc6617

Watson, C. T., Marques-Bonet, T., Sharp, A. J., & Mefford, H. C. (2014). The genetics of microdeletion and microduplication syndromes: an update. *Annual Review of Genomics and Human Genetics*, *15*, 215–244. https://doi.org/10.1146/annurev-genom-091212-153408

Wedenoja, S., Yoshihara, M., Teder, H., Sariola, H., Gissler, M., Katayama, S., Wedenoja, J., Häkkinen, I. M., Ezer,

S., Linder, N., Lundin, J., Skoog, T., Sahlin, E., Iwarsson, E., Pettersson, K., Kajantie, E., Mokkonen, M., Heinonen, S., Laivuori, H., … Kere, J. (2019). Balancing Selection at HLA-G Modulates Fetal Survival, Preeclampsia and Human Birth Sex Ratio. In *bioRxiv* (p. 851089). https://doi.org/10.1101/851089

Wenger, A. M., Peluso, P., Rowell, W. J., Chang, P.-C., Hall, R. J., Concepcion, G. T., Ebler, J., Fungtammasan, A., Kolesnikov, A., Olson, N. D., Töpfer, A., Alonge, M., Mahmoud, M., Qian, Y., Chin, C.-S., Phillippy, A. M., Schatz, M. C., Myers, G., DePristo, M. A., … Hunkapiller, M. W. (2019). Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nature Biotechnology*. https://doi.org/10.1038/s41587-019-0217-9

Wickham, H. (2014). Tidy Data. *Journal of Statistical Software*, *59*, 1–23. https://doi.org/10.18637/jss.v059.i10

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T., Miller, E., Bache, S., Müller, K., Ooms, J., Robinson, D., Seidel, D., Spinu, V., … Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, *4*(43), 1686. https://doi.org/10.21105/joss.01686

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., … Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, *3*, 160018. https://doi.org/10.1038/sdata.2016.18

Williams, T. N., Wambua, S., Uyoga, S., Macharia, A., Mwacharo, J. K., Newton, C. R. J. C., & Maitland, K. (2005). Both heterozygous and homozygous alpha+ thalassemias protect against severe and fatal Plasmodium falciparum malaria on the coast of Kenya. *Blood*, *106*(1), 368–371. https://doi.org/10.1182/blood-2005-01-0313

Wilson, G., Aruliah, D. A., Brown, C. T., Chue Hong, N. P., Davis, M., Guy, R. T., Haddock, S. H. D., Huff, K. D., Mitchell, I. M., Plumbley, M. D., Waugh, B., White, E. P., & Wilson, P. (2014). Best practices for scientific computing. *PLoS Biology*, *12*(1), e1001745. https://doi.org/10.1371/journal.pbio.1001745

Wilson, G., Bryan, J., Cranston, K., Kitzes, J., Nederbragt, L., & Teal, T. K. (2017). Good enough practices in scientific computing. *PLoS Computational Biology*, *13*(6), e1005510. https://doi.org/10.1371/journal.pcbi.1005510

Wilson, G. M., Flibotte, S., Missirlis, P. I., Marra, M. A., Jones, S., Thornton, K., Clark, A. G., & Holt, R. A. (2006). Identification by full-coverage array CGH of human DNA copy number increases relative to chimpanzee and gorilla. *Genome Research*, *16*(2), 173–181. https://doi.org/10.1101/gr.4456006

Wu, Y., Dowbenko, D., Pisabarro, M. T., Dillard-Telm, L., Koeppen, H., & Lasky, L. A. (2001). PTEN 2, a Golgi-associated testis-specific homologue of the PTEN tumor suppressor lipid phosphatase. *The Journal of Biological Chemistry*, *276*(24), 21745–21753. https://doi.org/10.1074/jbc.M101480200

Xue, Y., Sun, D., Daly, A., Yang, F., Zhou, X., Zhao, M., Huang, N., Zerjal, T., Lee, C., Carter, N. P., Hurles, M. E., & Tyler-Smith, C. (2008). Adaptive evolution of UGT2B17 copy-number variation. *American Journal of Human Genetics*, *83*(3), 337–346. https://doi.org/10.1016/j.ajhg.2008.08.004

Yamanaka, M., Kato, Y., Angata, T., & Narimatsu, H. (2009). Deletion polymorphism of SIGLEC14 and its functional implications. *Glycobiology*, *19*(8), 841–846. https://doi.org/10.1093/glycob/cwp052

Yan, S. M., Sherman, R. M., Taylor, D. J., Nair, D. R., Bortvin, A. N., Schatz, M. C., & McCoy, R. C. (2021). Local adaptation and archaic introgression shape global diversity at human structural variant loci. *ELife*, *10*. https://doi.org/10.7554/eLife.67615

Yandell, M., Huff, C., Hu, H., Singleton, M., Moore, B., Xing, J., Jorde, L. B., & Reese, M. G. (2011). A probabilistic disease-gene finder for personal genomes. *Genome Research*, *21*(9), 1529–1542. https://doi.org/10.1101/gr.123158.111

Yohn, C. T., Jiang, Z., McGrath, S. D., Hayden, K. E., Khaitovich, P., Johnson, M. E., Eichler, M. Y., McPherson, J. D., Zhao, S., Pääbo, S., & Eichler, E. E. (2005). Lineage-Specific Expansions of Retroviral Insertions within the Genomes of African Great Apes but Not Humans and Orangutans. In *PLoS Biology* (Vol. 3, Issue 4, p. e110). https://doi.org/10.1371/journal.pbio.0030110

Young, J. M., Endicott, R. M., Parghi, S. S., Walker, M., Kidd, J. M., & Trask, B. J. (2008). Extensive copy-number variation of the human olfactory receptor gene family. *American Journal of Human Genetics*, *83*(2), 228–242. https://doi.org/10.1016/j.ajhg.2008.07.005

Yurkovich, J. T., Yurkovich, B. J., Dräger, A., Palsson, B. O., & King, Z. A. (2017). A Padawan Programmer's Guide to Developing Software Libraries. *Cell Systems*, *5*(5), 431–437. https://doi.org/10.1016/j.cels.2017.08.003

Zerbino, D. R., Frankish, A., & Flicek, P. (2020). Progress, Challenges, and Surprises in Annotating the Human Genome. *Annual Review of Genomics and Human Genetics*, *21*, 55–79. https://doi.org/10.1146/annurev-

genom-121119-083418

Zhang, F., Carvalho, C. M. B., & Lupski, J. R. (2009). Complex human chromosomal and genomic rearrangements. *Trends in Genetics: TIG*, *25*(7), 298–307. https://doi.org/10.1016/j.tig.2009.05.005

Zhao, J., Ma, J., Deng, Y., Kelly, J. A., Kim, K., Bang, S.-Y., Lee, H.-S., Li, Q.-Z., Wakeland, E. K., Qiu, R., Liu, M., Guo, J., Li, Z., Tan, W., Rasmussen, A., Lessard, C. J., Sivils, K. L., Hahn, B. H., Grossman, J. M., … Tsao, B. P. (2017). A missense variant in NCF1 is associated with susceptibility to multiple autoimmune diseases. *Nature Genetics*, *49*(3), 433–437. https://doi.org/10.1038/ng.3782

Zheng-Bradley, X., Streeter, I., Fairley, S., Richardson, D., Clarke, L., Flicek, P., & 1000 Genomes Project Consortium. (2017). Alignment of 1000 Genomes Project reads to reference assembly GRCh38. *GigaScience*, *6*(7), 1–8. https://doi.org/10.1093/gigascience/gix038

Zhou, X., Cain, C. E., Myrthil, M., Lewellen, N., Michelini, K., Davenport, E. R., Stephens, M., Pritchard, J. K., & Gilad, Y. (2014). Epigenetic modifications are associated with inter-species gene expression variation in primates. *Genome Biology*, *15*(12), 547. https://doi.org/10.1186/s13059-014-0547-3

Zook, J. M., Hansen, N. F., Olson, N. D., Chapman, L., Mullikin, J. C., Xiao, C., Sherry, S., Koren, S., Phillippy, A. M., Boutros, P. C., Sahraeian, S. M. E., Huang, V., Rouette, A., Alexander, N., Mason, C. E., Hajirasouliha, I., Ricketts, C., Lee, J., Tearle, R., … Salit, M. (2020). A robust benchmark for detection of germline large deletions and insertions. *Nature Biotechnology*, *38*(11), 1347–1355. https://doi.org/10.1038/s41587-020-0538-8

Zufferey, M., Tavernari, D., Oricchio, E., & Ciriello, G. (2018). Comparison of computational methods for the identification of topologically associating domains. *Genome Biology*, *19*(1), 217. https://doi.org/10.1186/s13059-018-1596-9

Supplemental Figures

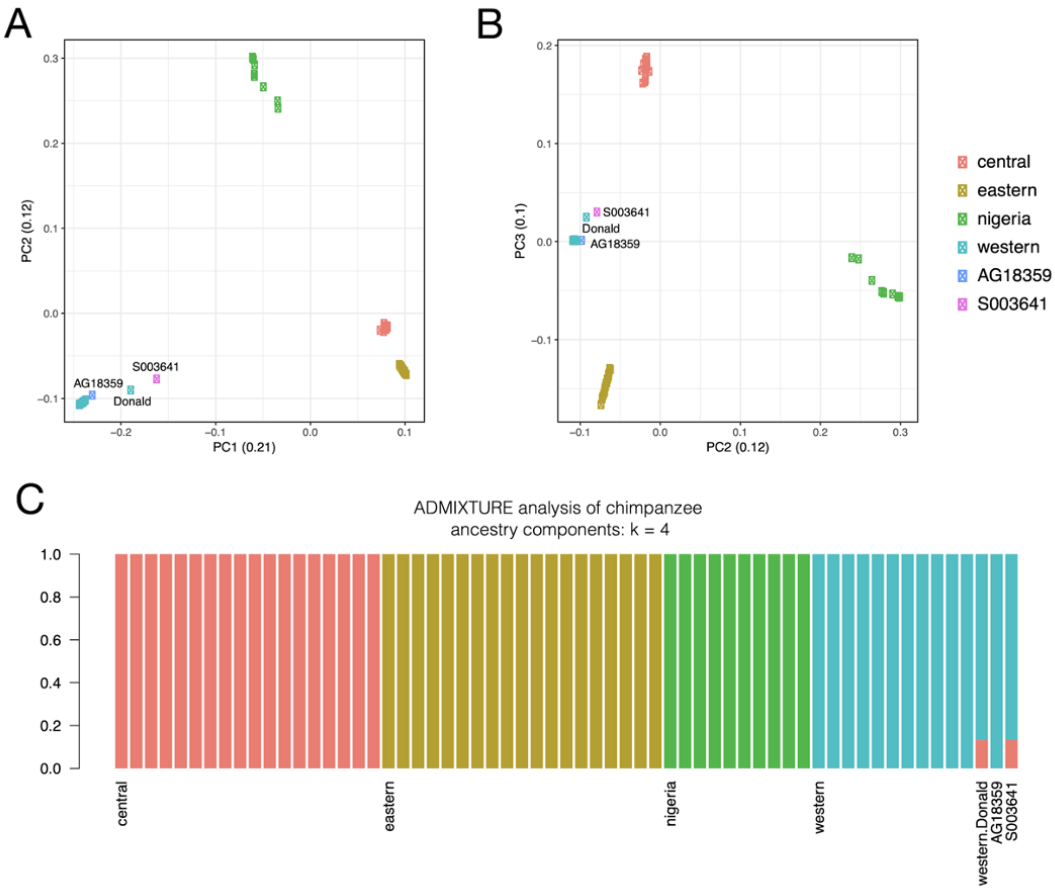


Figure S2.1. Chimpanzee subspecies identification. (A) and (B) PC analysis of chimpanzee genetic diversity. Both of the newly sequenced cell lines were projected onto PCs inferred from the 59 chimpanzees presented in de Manuel et al. (de Manuel et al., 2016). Both cell lines show closest affinity to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

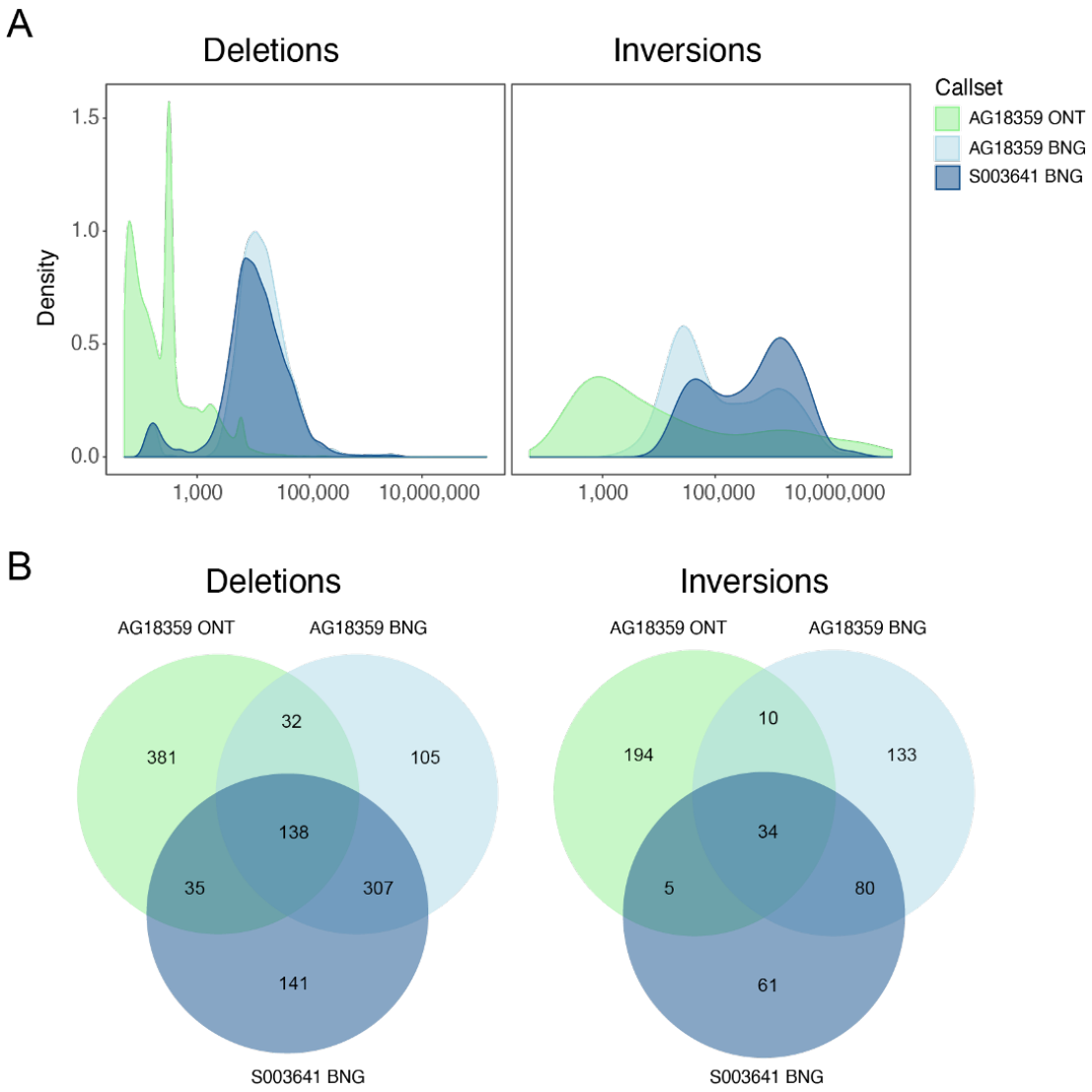


Figure S2.2. Description of SV discovery set. (A) Length distribution (x-axis in bp) of raw SV calls discovered by ONT (green) and BNG (light blue) from AG18359, and BNG from S003641 (dark blue). (B) Venn diagram comparing large (≥10 kbp) deletions (left) and inversions (right) discovered for each individual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

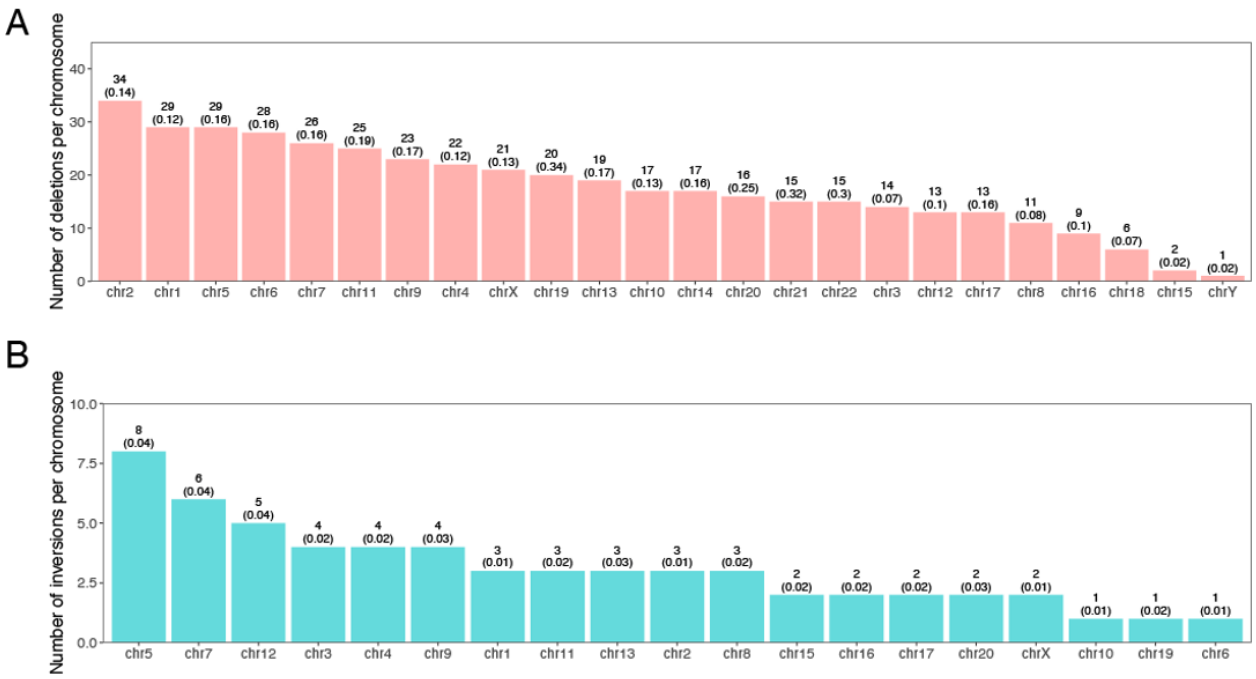


Figure S2.3. Histogram of identified SV events per chromosome. The number of high-confidence SV events discovered is depicted for (A) deletions and (B) inversions. The normalized number of events per Mbp for each chromosome is displayed in parentheses.

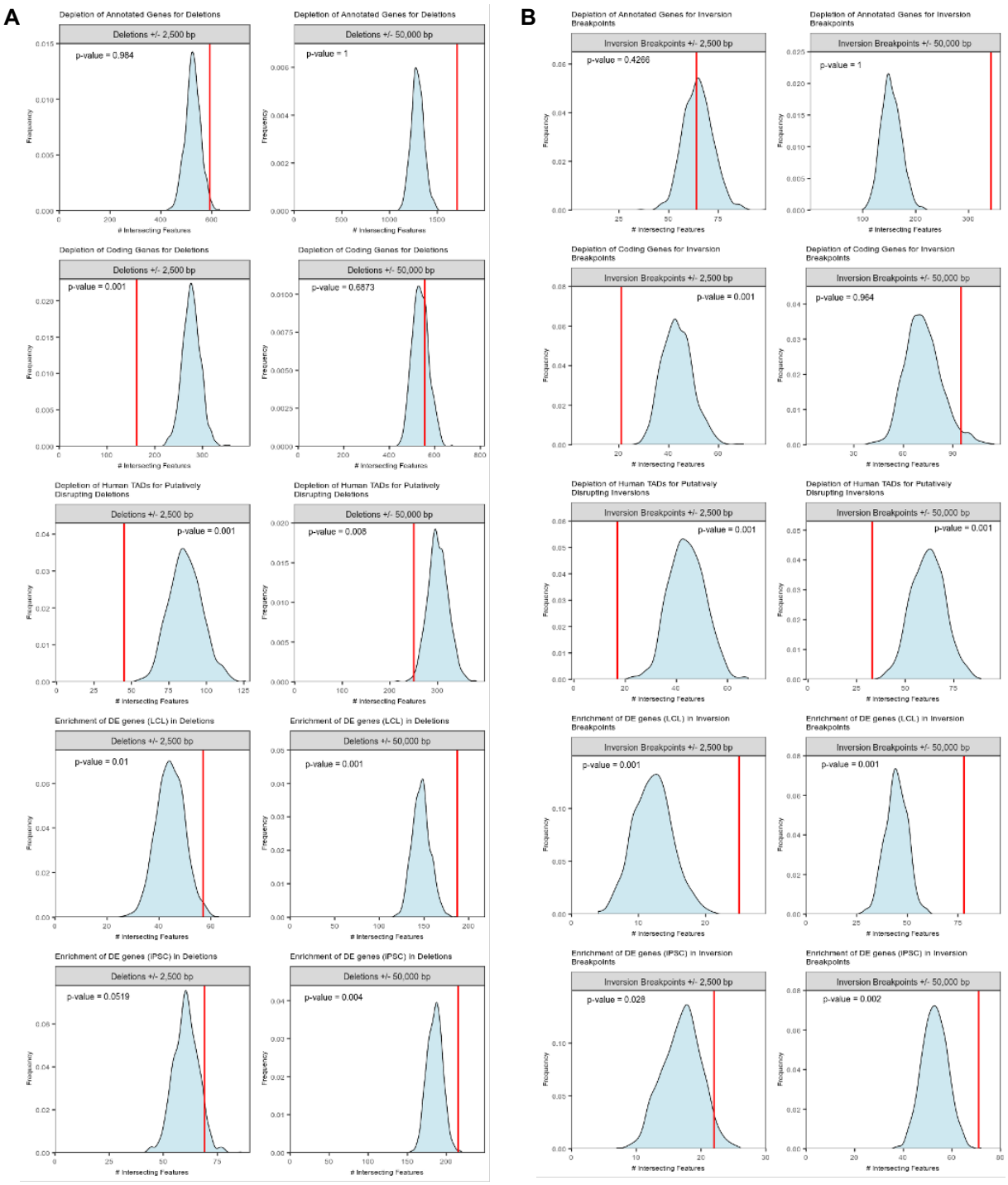


Figure S2.4. Enrichment/depletion of SV breakpoints for genomic features of interest as determined by permutation testing. Each plot compares the observed count of intersecting features (red vertical line) to a distribution of counts generated from 1000 permuted sets of coordinates (for testing depletion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

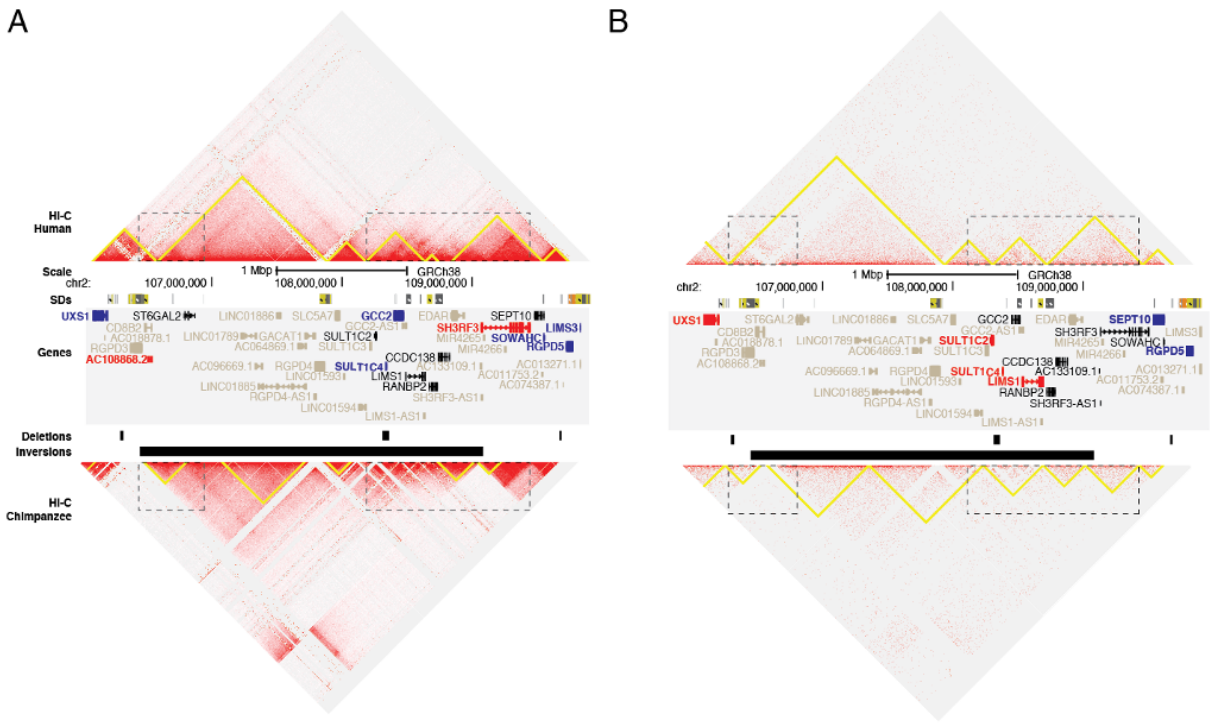


Figure S2.5. Genome organization of human chromosome 2q12.2-q13. The Hi-C genomic landscape of human (top) and chimpanzee (bottom) are depicted for iPSCs (A) and LCLs (B) using Juicebox at chr2:106,095,001-109,905,000 (GRCh38). Predicted TADs (yellow triangles) were compared between species, noting differences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

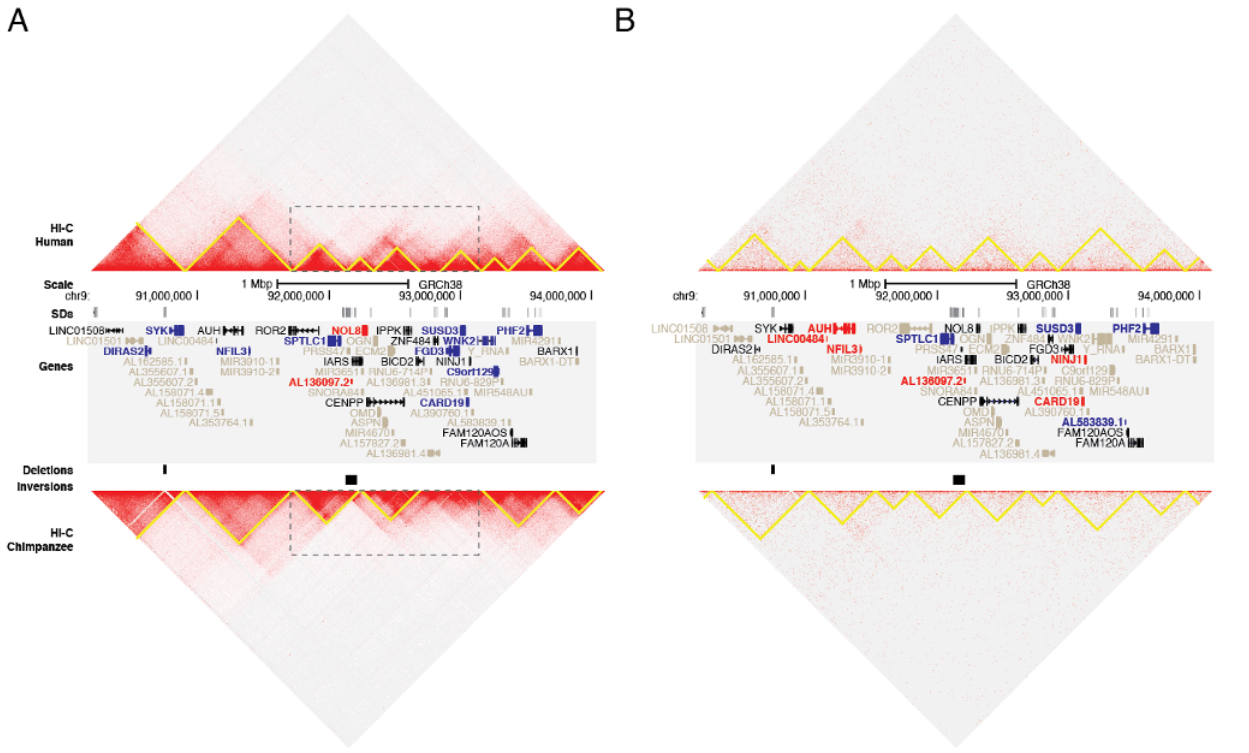


Figure S2.6. Genome organization of human chromosome 9q22.2-q22.32. The Hi-C genomic landscape of human (top) and chimpanzee (bottom) are depicted for iPSCs (A) and LCLs (B) using Juicebox at chr9:90,200,001-94,010,000 (GRCh38). Predicted TADs (yellow triangles) were compared between species, noting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

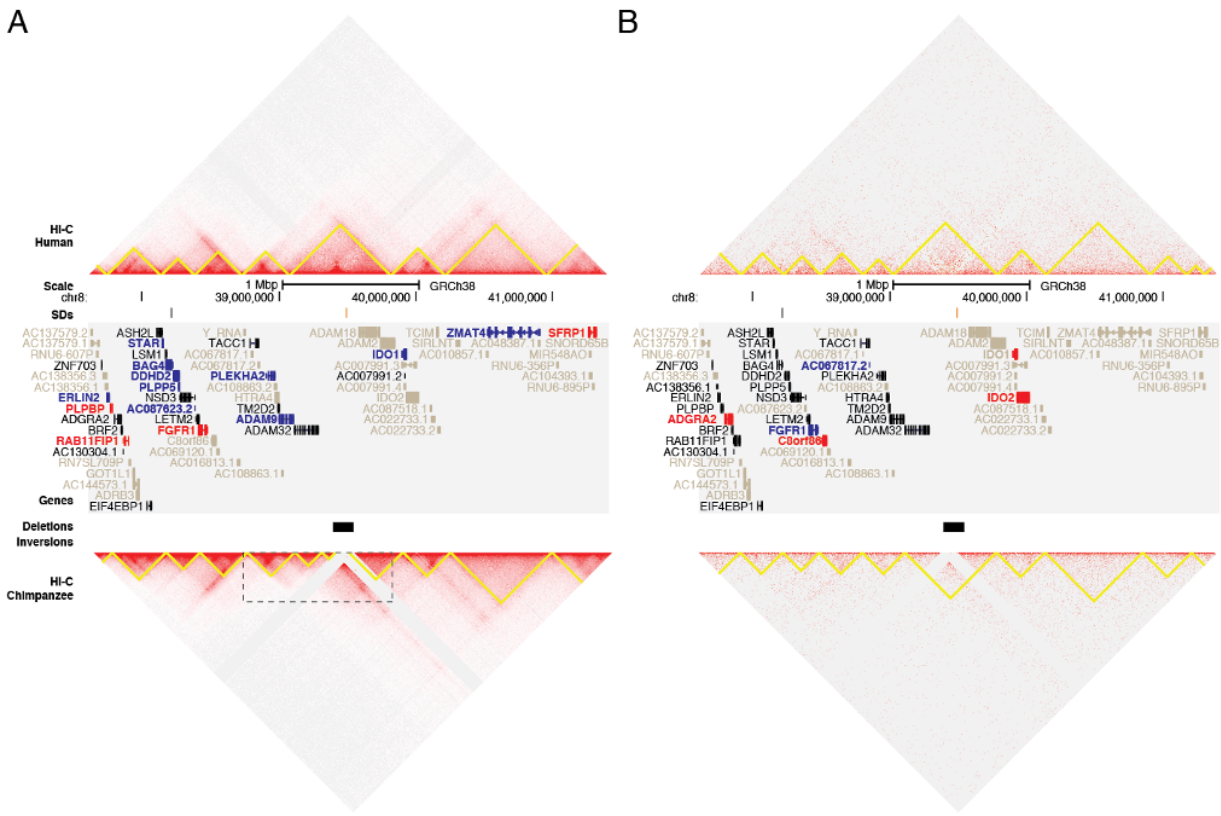


Figure S2.7. Genome organization of human chromosome 8p11.23-p11.21. The Hi-C genomic landscape of human (top) and chimpanzee (bottom) are depicted for iPSCs (A) and LCLs (B) using Juicebox at chr8:37,620,001-41,430,000 (GRCh38). Predicted TADs (yellow triangles) were compared between species, noting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

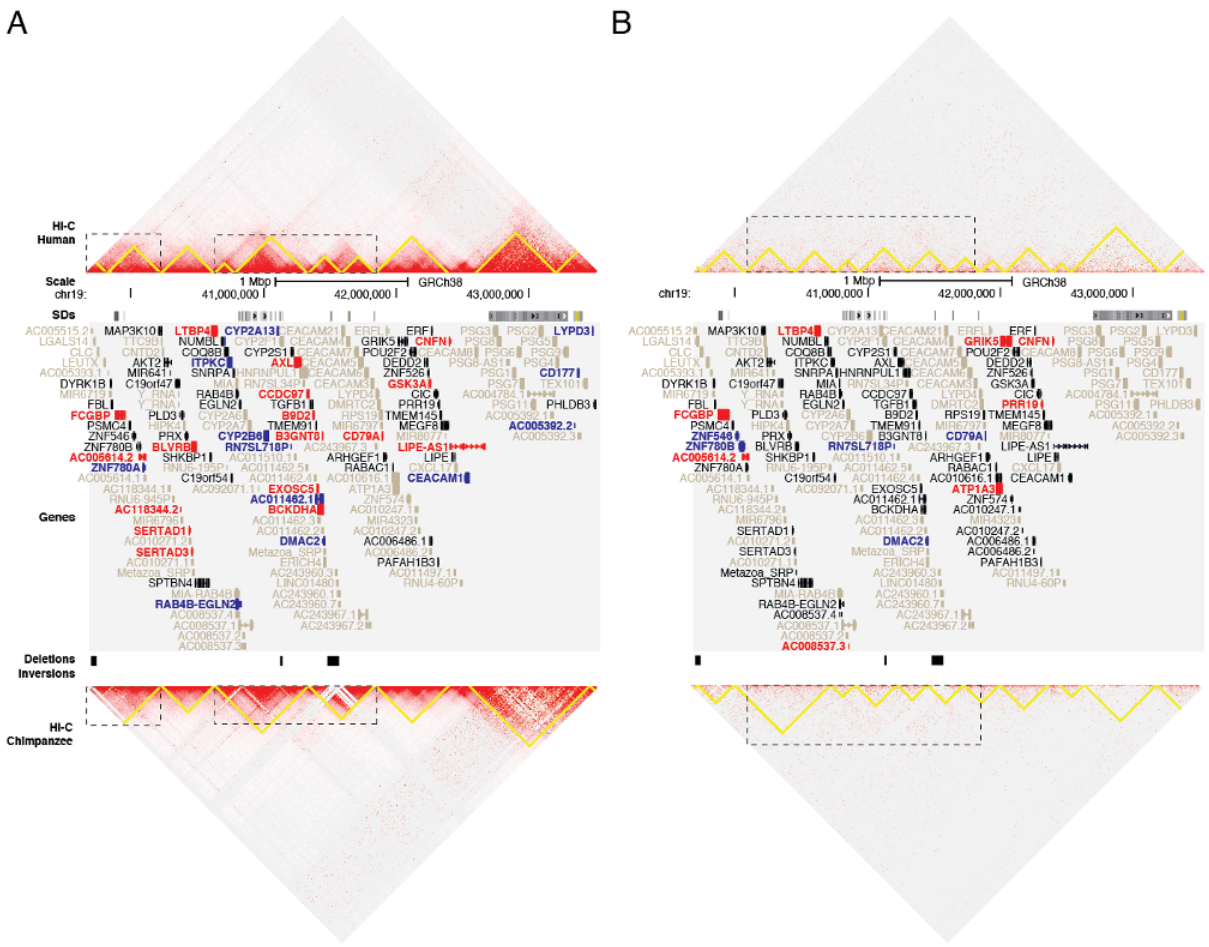


Figure S2.8. Genome organization of human chromosome 19q13.2-q13.31. The Hi-C genomic landscape of human (top) and chimpanzee (bottom) are depicted for iPSCs (A) and LCLs (B) using Juicebox at chr19:39,685,001-43,495,000 (GRCh38). Predicted TADs (yellow triangles) were compared between species, noting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

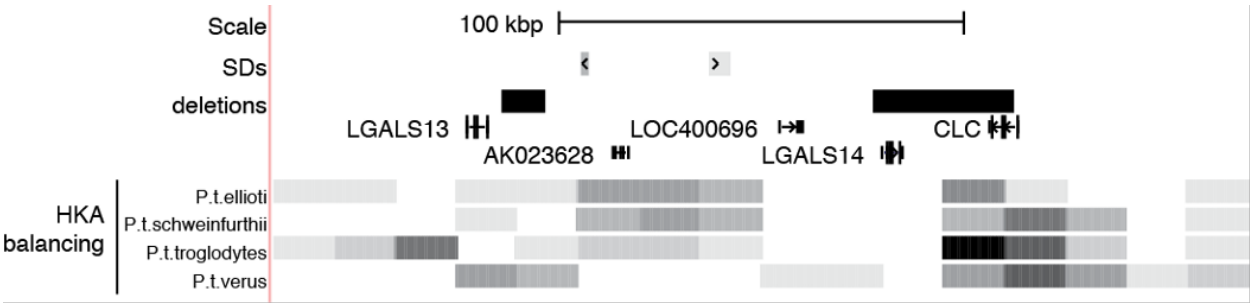


Figure S2.9. Chimpanzee-specific deletions of the galectin family of genes. Pictured is a UCSC Genome Browser snapshot of human chromosome 19p13.2. The locations of SDs (colored bars), deletions (black bars), and genes are indicated. For each subspecies of chimpanzee [*Pan troglodytes* (*P.t.*)], the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

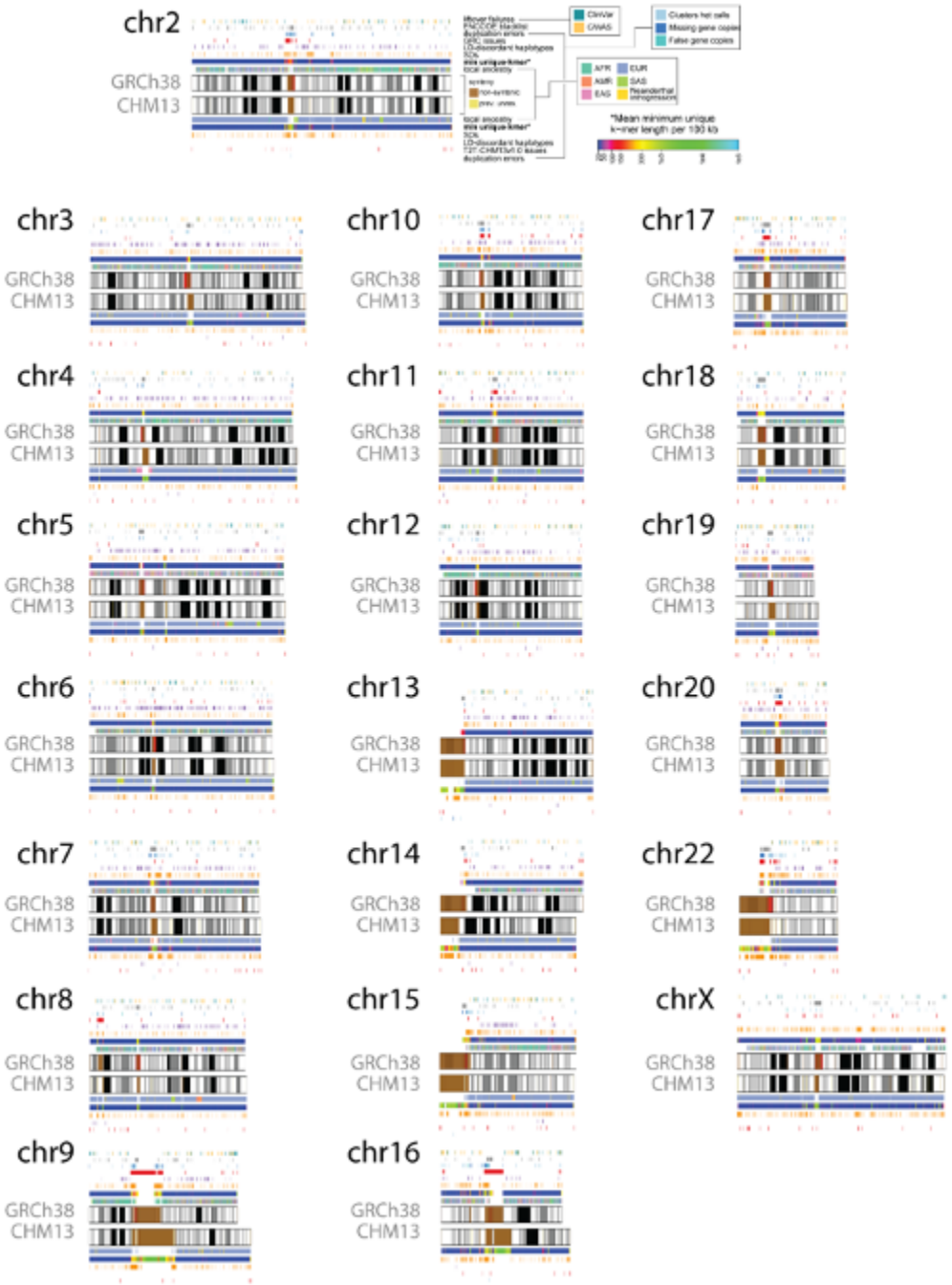


Figure S3.1. Annotation of all chromosomes. Overview of annotations available for GRCh38 and CHM13 all chromosomes (chromosomes 1 and 21 shown in **Figure 3.1A**) with colors indicated in legend. Cytobands are pictured as gray bands with red bands representing centromeric regions within ideograms.

Figure S3.2. Local ancestry of all chromosomes. Overview of local ancestry as separated tracks for all chromosomes for GRCh38 and T2T-CHM13 with ancestries indicated in the legend (AFR: African, AMR: Admixed American, EAS: East Asian, EUR: European, SAS: South Asian). Cytobands are pictured as gray bands with red bands representing centromeric regions within ideograms. Brown regions within chromosomes indicate non-syntenic regions between GRCh38 and CHM13, and yellow regions indicate previously unresolved sequences.

Figure S3.3. Duplication errors of all chromosomes. Overview of duplications errors as separated tracks in GRCh38 and T2T-CHM13 with categories indicated in the legend. Cytobands are pictured as gray bands with red bands representing centromeric regions within ideograms. Brown regions within chromosomes indicate non-syntenic regions between GRCh38 and CHM13, and yellow regions indicate previously unresolved sequences.

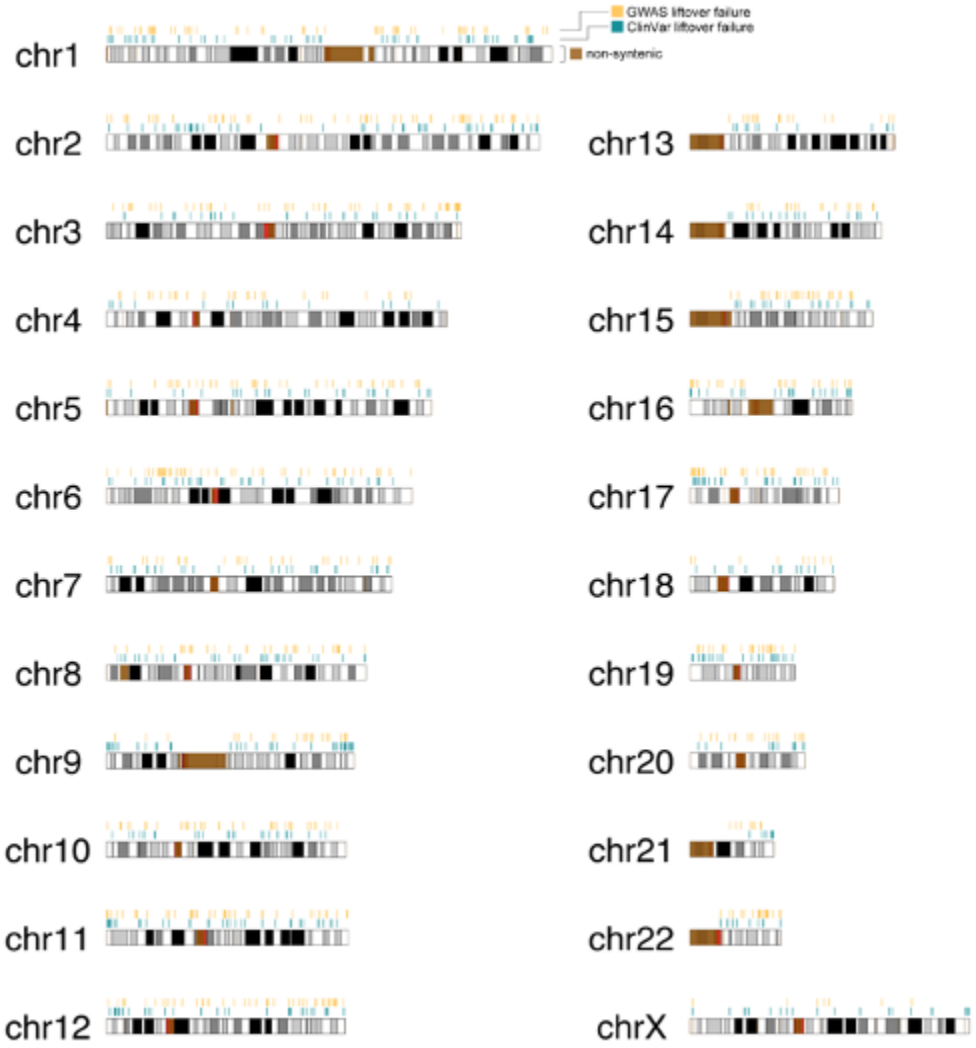


Figure S3.4. Liftover failures of all chromosomes. Overview of liftover failures for all chromosomes in GRCh38 for GWAS (top) and ClinVar (bottom). Cytobands are pictured as gray bands with red bands representing centromeric regions within ideograms. Brown regions within chromosomes indicate non-syntenic regions between GRCh38 and T2T-CHM13.

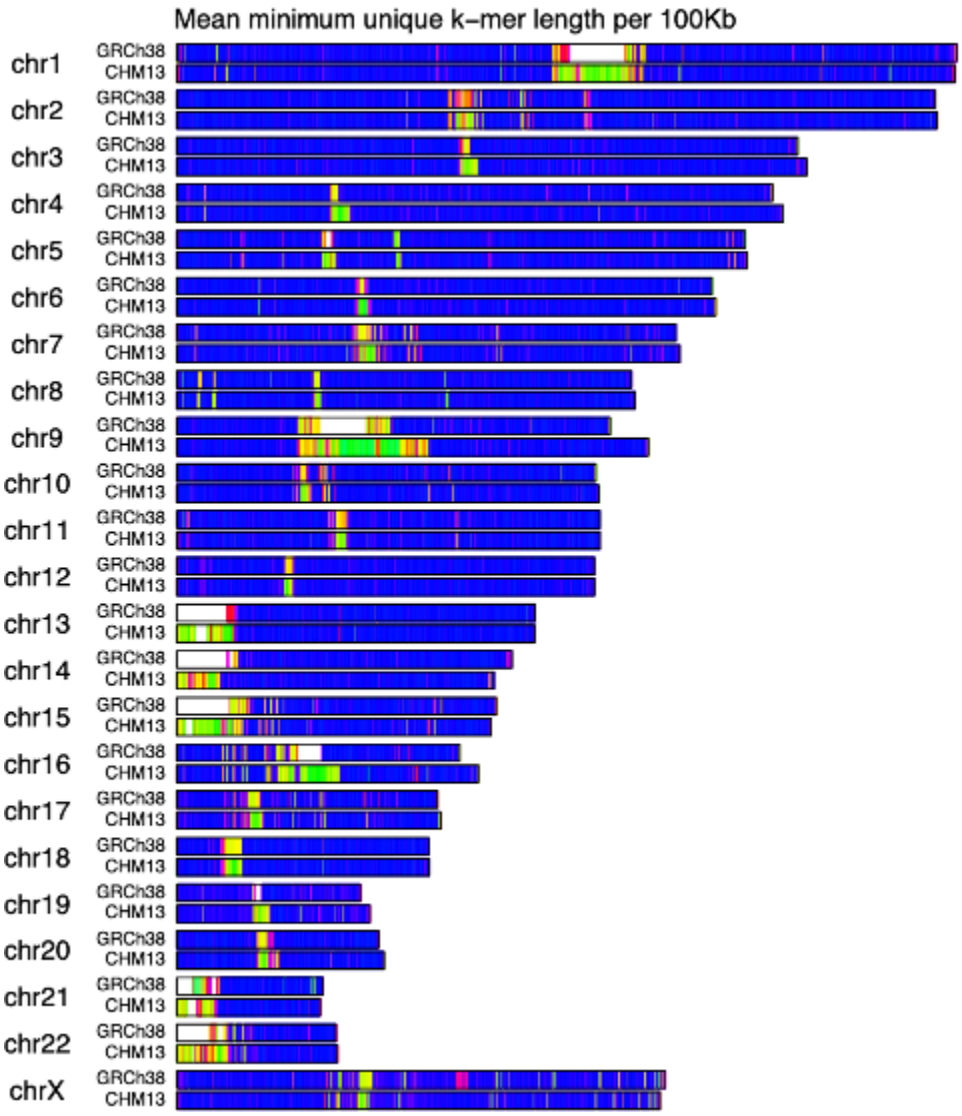


Figure S3.5. Minimum unique k-mer chromosome score maps. Minimum unique *k*-mer scores (left-anchored) were averaged in 100 Kb bins and plotted along the length of each chromosome using the color gradient displayed on the right. Regions denoted in white indicate that no valid score exists, either because the k-mer sequence contains at least one N or the length of the k-mer would cause it to overlap a chromosome end.

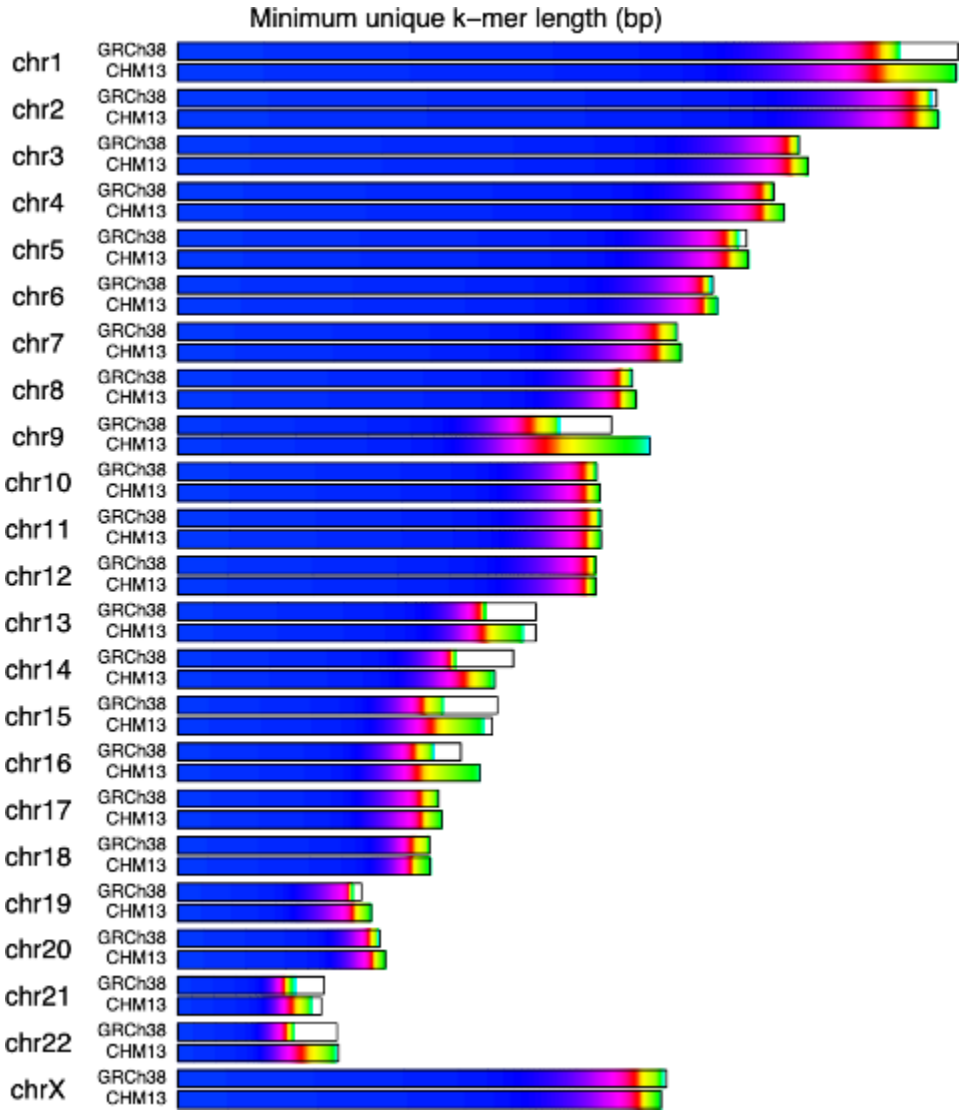


Figure S3.6. Minimum unique k-mer chromosome histograms. Minimum unique k-mer scores (left-anchored) were calculated for each base position, sorted on a per-chromosome basis, and plotted using the color gradient displayed on the right. A black outline represents the reported length of each chromosome. Regions of white indicate invalid scores due to either overlapping sequence containing at least one N or overlapping the end of a chromosome.

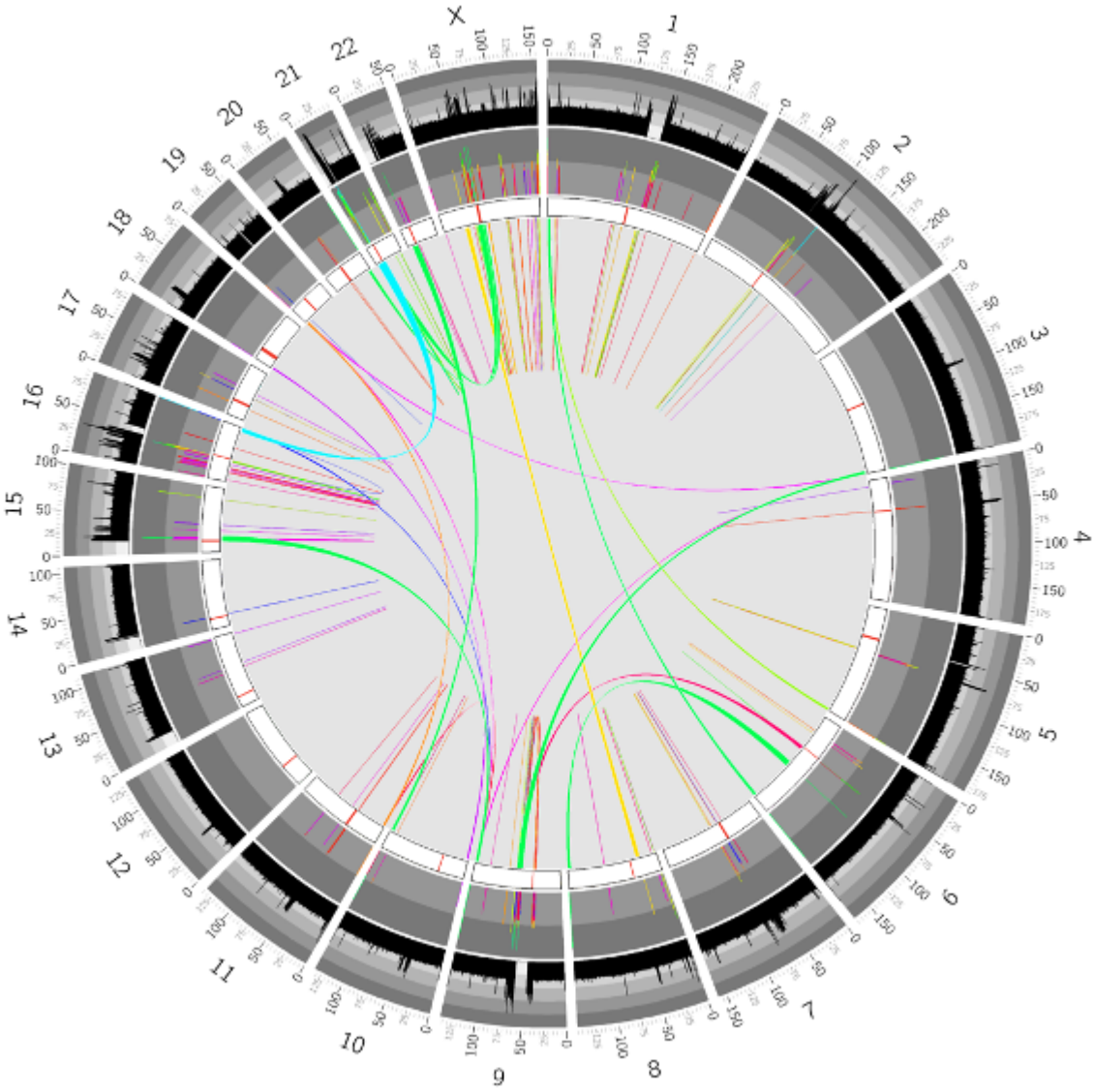


Figure S3.7. Large exact match sequence pairs in GRCh38. For all minimum non-unique sequences greater than 5 Kb (minimum unique k-mer length plus one), the pair of positions corresponding to each sequence were determined. Sequence sizes are denoted by both color and a colored barplot (middle ring) that ranges from 1 Kb to 100 Kb in $\log_{10}$ scale. The inner ring denotes the relative length of each chromosome with the annotated centromeric region indicated in read. The outer ring shows the minimum unique k-mer score, binned at 100 Kb intervals and is shown with a range of 1 to 100 Kb in $\log_{10}$ scale.

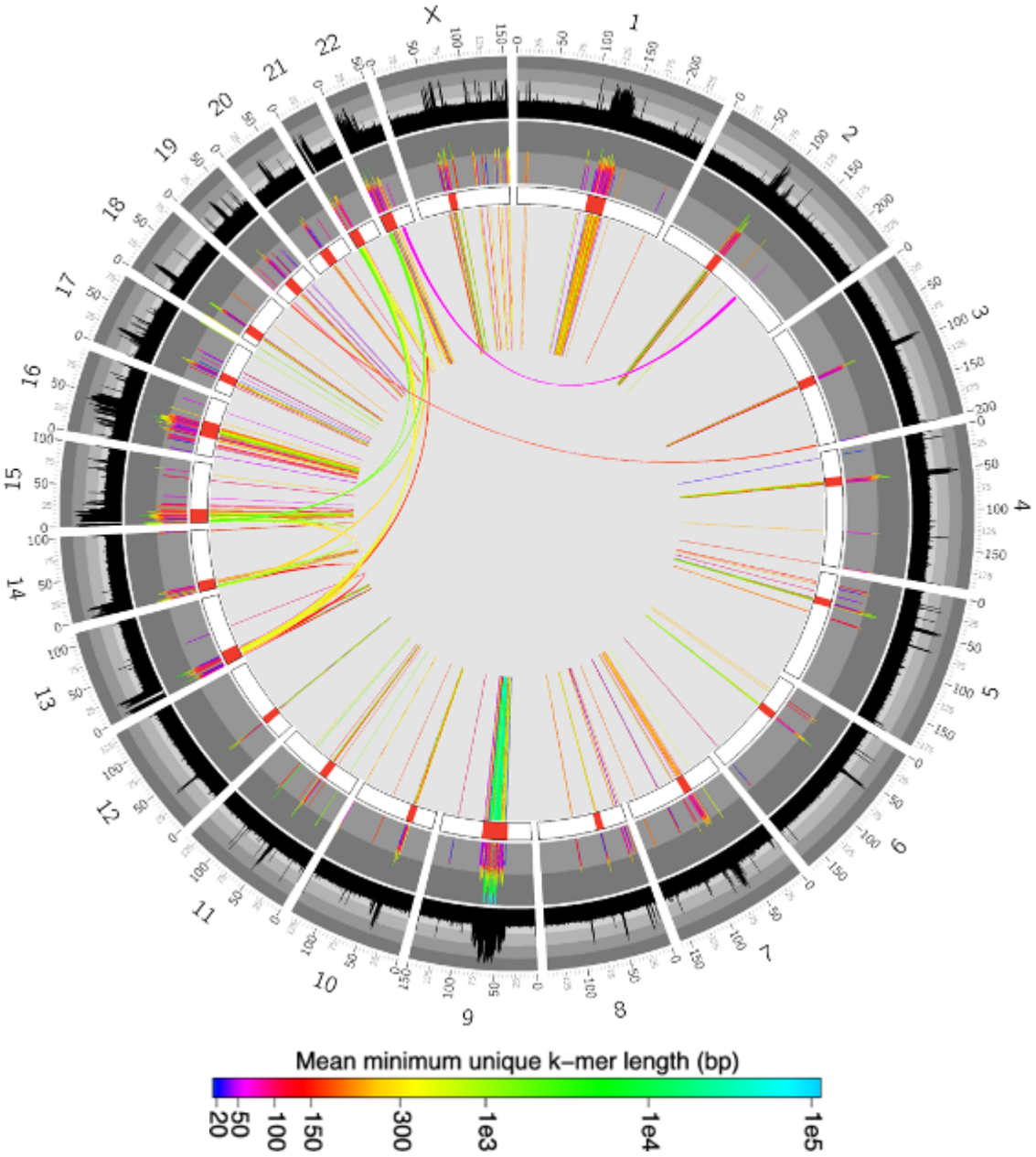


Figure S3.8. Large exact match sequence pairs in CHM13. For all minimum non-unique sequences greater than 5 Kb (minimum unique k-mer length plus one), the pair of positions corresponding to each sequence were determined. Sequence sizes are denoted by both color and a colored barplot (middle ring) that ranges from 1 Kb to 100 Kb in $\log_{10}$ scale. The inner ring denotes the relative length of each chromosome with the annotated centromeric region indicated in read. The outer ring shows the minimum unique k-mer score, binned at 100 Kb intervals and is shown with a range of 1 to 100 Kb in $\log_{10}$ scale.

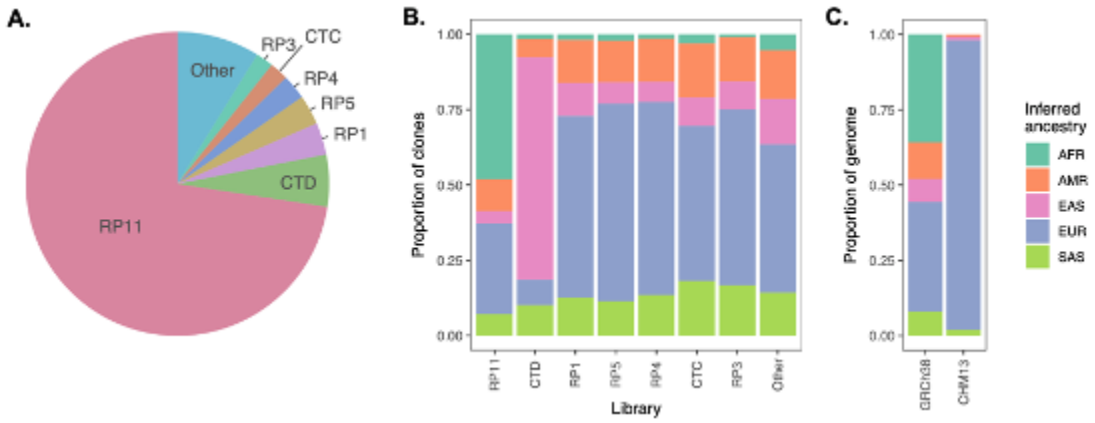


Figure S3.9. Local ancestry analysis of GRCh38 and CHM13. A. Proportion of the GRCh38 reference genome composed of clones from various libraries (RP11, CTD, etc.), each derived from DNA obtained from a distinct diploid donor, each indicated with a distinct color. B. Inferred local ancestry proportions for BAC libraries derived from different donor individuals that contributed to GRCh38. C. Total inferred local ancestry proportions for the GRCh38 and CHM13 reference genome. For panels B and C, ancestry is indicated with colors corresponding to 1KGP superpopulations (AFR: African, AMR: Admixed American, EAS: East Asian, EUR: European, SAS: South Asian).

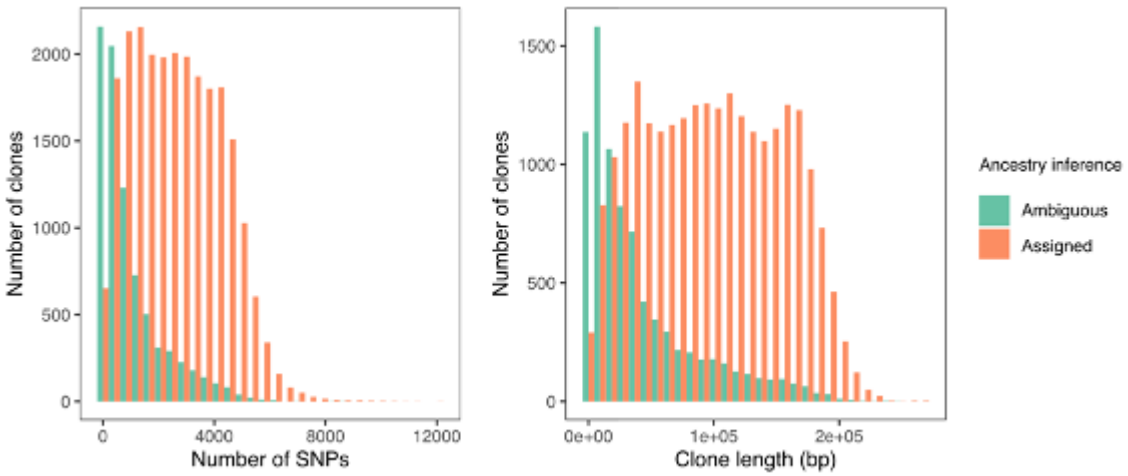


Figure S3.10. Ambiguity in ancestry inference for short GRCh38 clones with few markers. Number of SNPs (left panel) and length (right panel) of each GRCh38 clone for which ancestry was or was not inferred based on majority vote of nearest neighbor haplotypes in the phased 1KGP reference panel.

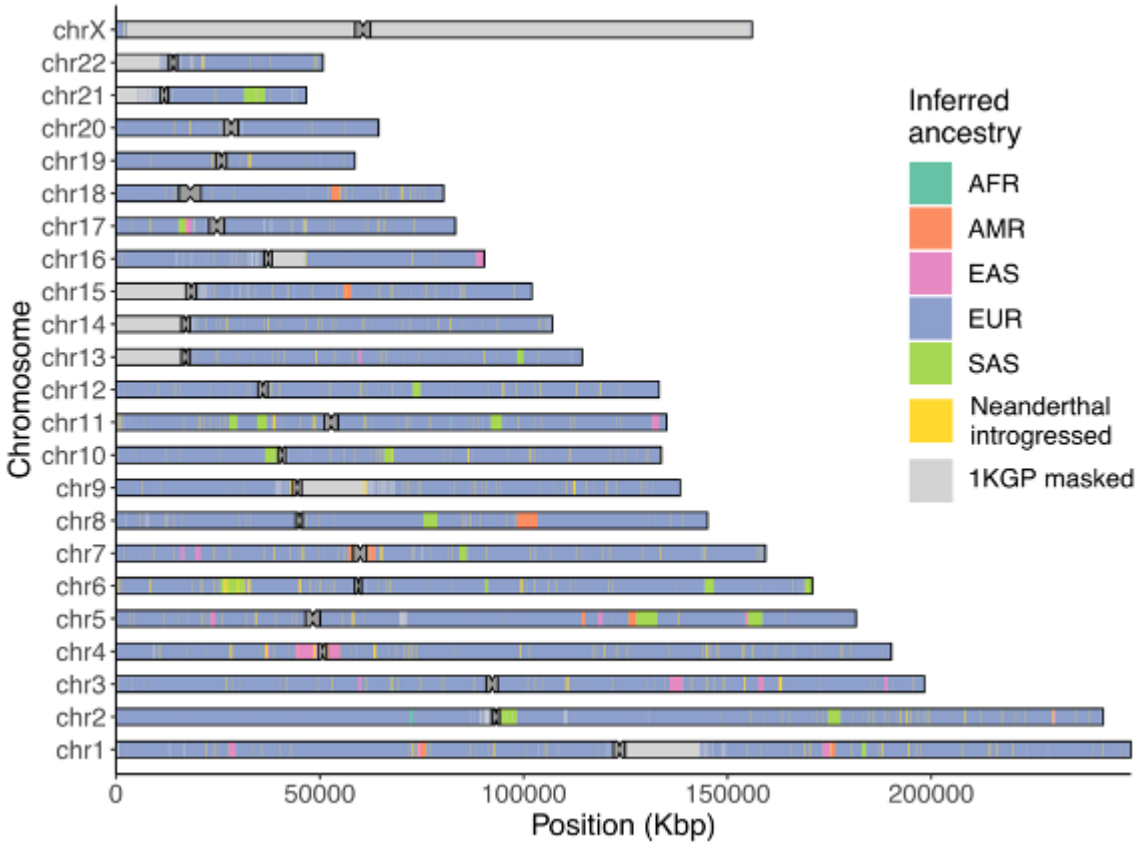


Figure S3.11. Local ancestry analysis of CHM13. Ideogram depicting RFMix-inferred local ancestry tracts for CHM13. Ancestry is divided by 1KGP superpopulation (AFR: African, AMR: Admixed American, EAS: East Asian, EUR: European, SAS: South Asian). Neanderthal-introgressed haplotypes, inferred with IBDmix, and regions masked by the 1KGP are superimposed.

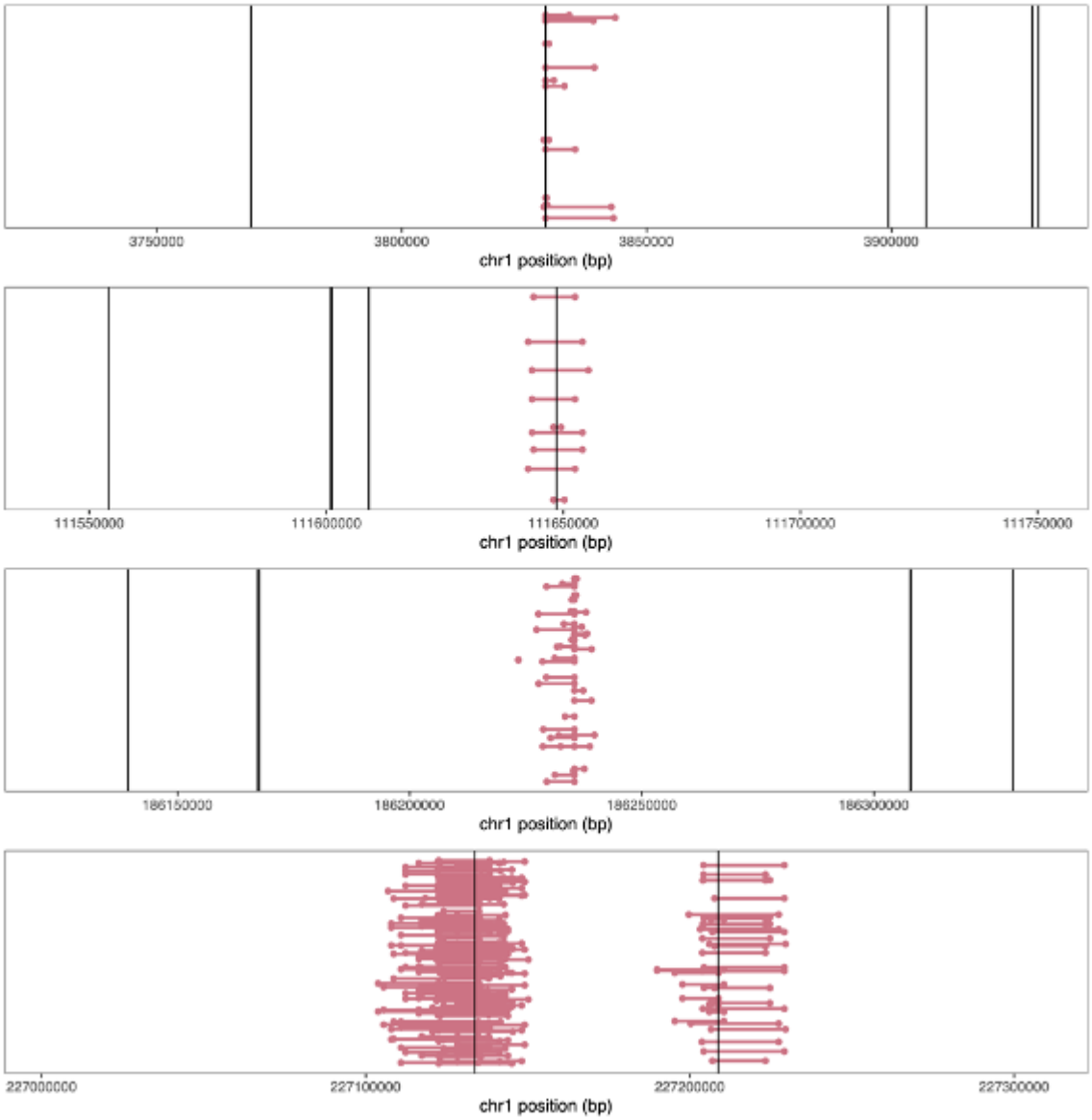


Figure S3.12. LD-discordant SNP pairs frequently span clone boundaries. Depiction of four representative "islands" of LD-discordant SNP pairs, where for common SNPs in perfect LD (R^2 = 1), GRCh38 possesses a combination of alleles that is never observed among the 1KGP samples. Linked SNP pairs are represented as dots connected by lines. Clone boundaries are represented as vertical lines. In all but one case (third row), SNP pairs straddle the annotated boundary of BAC clones.

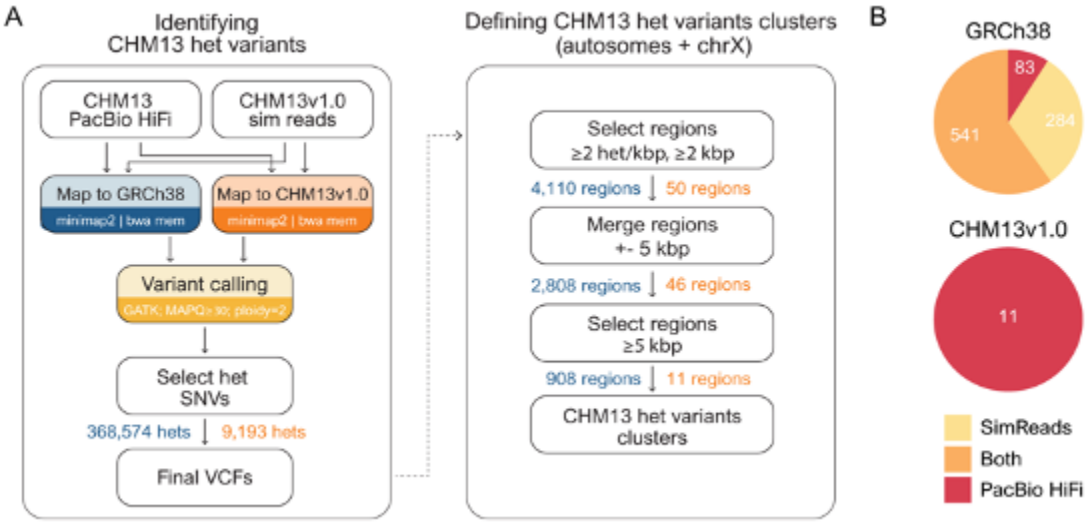


Figure S3.13. Strategy used to detect CHM13 heterozygous variant clusters. (A) The workflow to identify CHM13 heterozygous false positive (FP) heterozygous (het) regions in GRCh38 (blue) and CHM13v1.0 + Y chromosome (orange) from Illumina simulated reads (from T2T-CHM13v1.0 + Y chromosome) and PacBio (PB) HiFi reads generated from the CHM13 cell line. Variant calling (in yellow) was performed the same between both references. Numbers of features (excluding those associated with chrY) are indicated after each step in colored text using the same scheme (blue: GRCh38, orange: CHM13v1.0). (B) The sequencing-platform source of the FP het regions are shown in a Venn diagram for GRCh38 and CHM13v1.0 (SimReads: Simulated Illumina reads, PacBio HiFi reads, or Both).

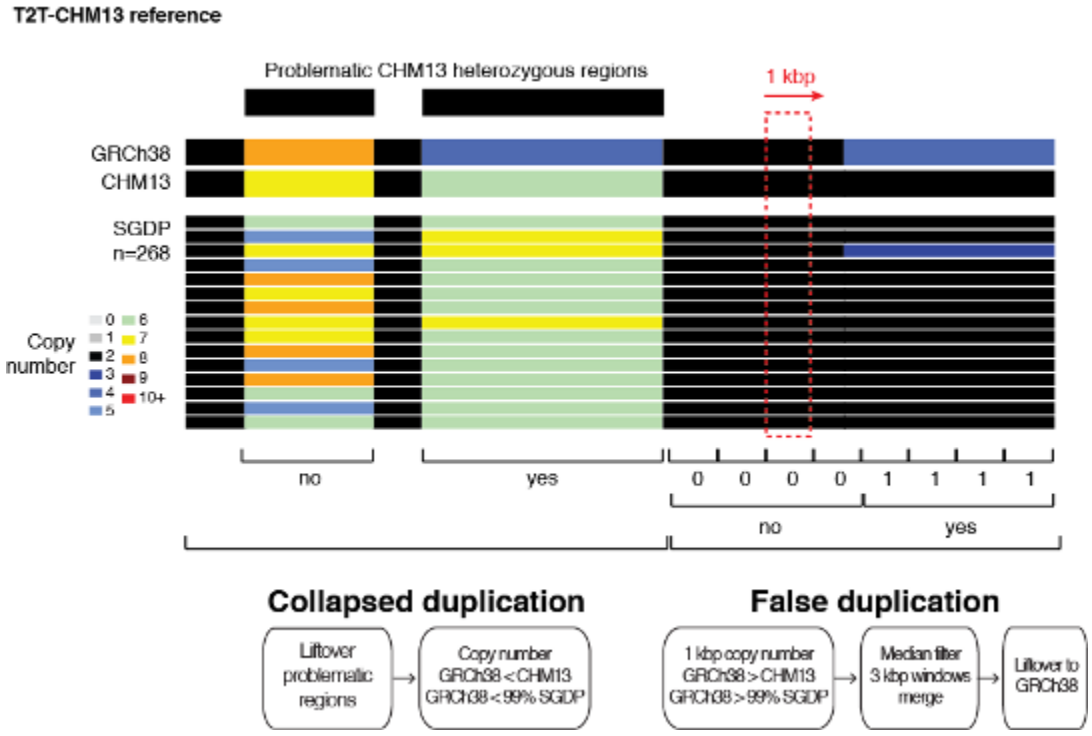


Figure S3.14. Strategy used to detect collapsed and false duplications in GRCh38. WSSD read-depth copy-number estimates were obtained for 'k-merized' versions of GRCh38 and T2T-CHM13v1.0 references, and Illumina reads from 268 SGDP individuals in the CHM13 reference. To identify putative collapsed duplications, the median copy-number of k-merized GRCh38 was compared to population and k-merized CHM13v1.0 copy numbers for each CHM13 problematic heterozygous region identified in either GRCh38 (lifted coordinates) or CHM13v1.0. To identify false duplications, k-merized GRCh38 copy-number estimates (depicted as colored regions as described by the legend) were compared to population and k-merized CHM13v1.0 copy-numbers using 1-kbp windows genome-wide.

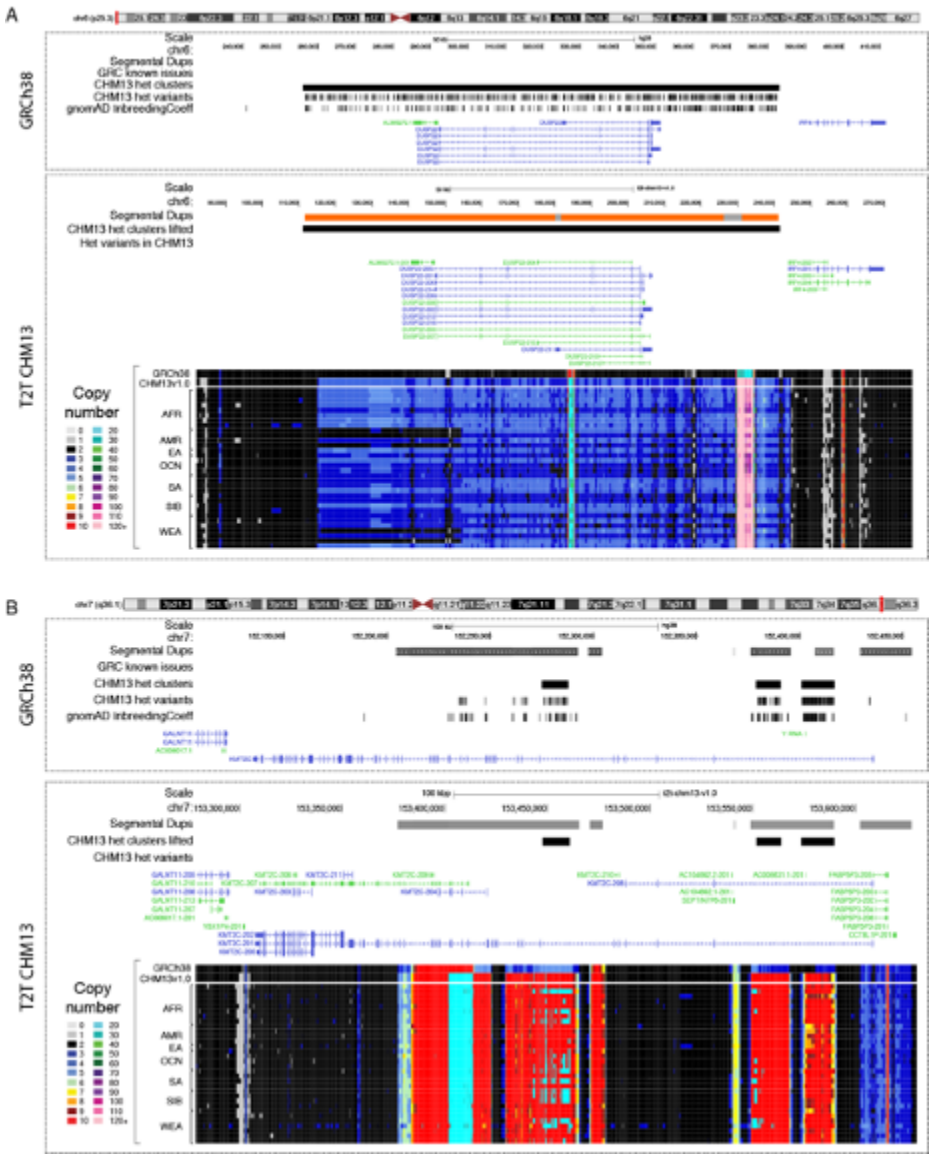


Figure S3.15. Examples of collapsed duplications impacting genes in GRCh38. UCSC Genome Browser snapshots of collapsed duplication in GRCh38 corrected in CHM13 impacting (A) DUSP22 and (B) KMT2C. For each loci, the position of the gene is depicted on a chromosomal ideogram. Segmental duplications (dups) are shown as bars colored by sequence similarity to paralog (gray: 90–98%; orange: >99%) and directionality versus paralog indicated as arrows (only for GRCh38). For both examples, CHM13 heterozygous variant clusters, CHM13 het variants, and gnomAD variants with the InbreedingCoeff flag are displayed for each reference. Gene models are depicted below variants, including coding (blue) and non-coding (green) isoforms. Read-depth copy-number estimates (with colors depicted in associated legends), displayed only for CHM13, are shown at the bottom for 'k-merized' versions of GRCh38 and T2T-CHM13v1.0 references, and Illumina reads from a diverse subset (n=34) of SGDP individuals.

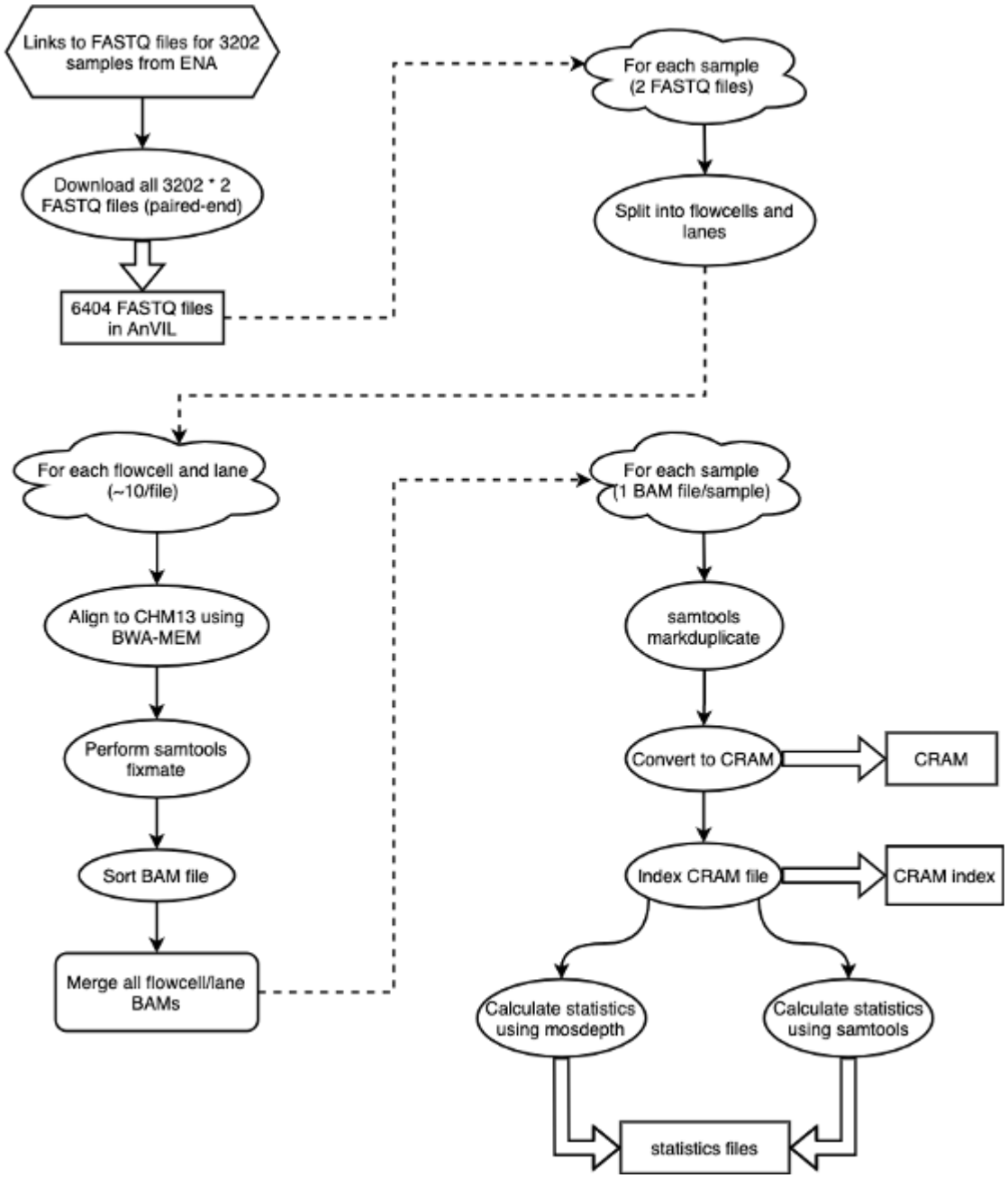


Figure S3.16. 1KGP Alignment Pipeline. Overview of the workflow for aligning 1KGP samples to CHM13, as adapted from the NYGC's pipeline for variant calling 1KGP samples on GRCh38 data. Hexagons represent input files, ellipses represent analysis steps, rectangles represent output files, clouds represent large cloud analyses, and rounded rectangles represent intermediary files. Dotted arrows precede cloud computations, whereas hollow arrows precede output files.

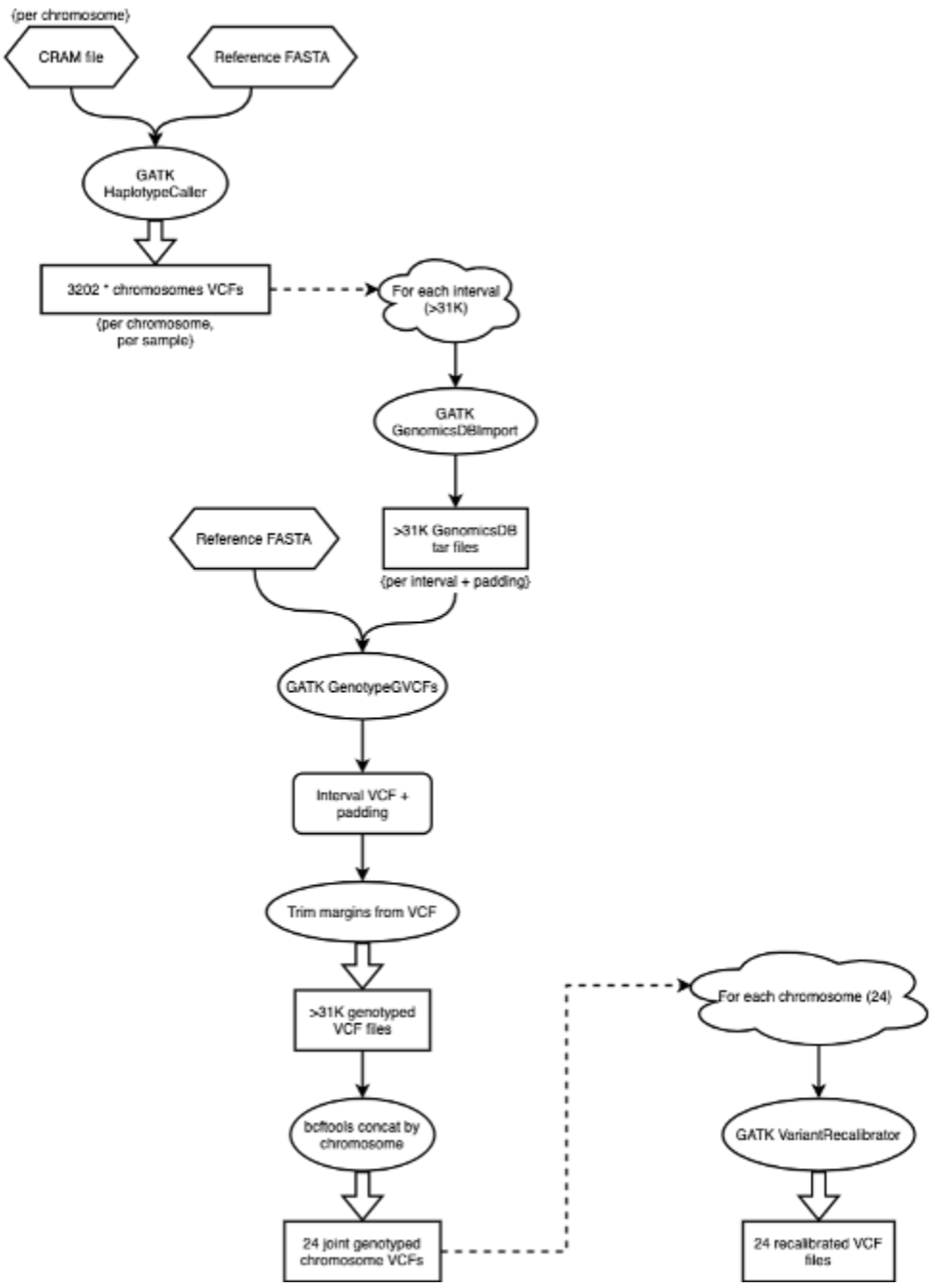


Figure S3.17. 1KGP Variant Calling Pipeline. Overview for the workflow for performing variant calling and joint genotyping on 1KGP samples after alignment to CHM13, as adapted from the NYGC's pipeline for variant calling 1KGP samples on GRCh38 data. Hexagons represent input files, ellipses represent analysis steps, rectangles represent output files, clouds represent large cloud analyses, and rounded rectangles represent intermediary files. Dotted arrows precede cloud computations, whereas hollow arrows precede final output files.

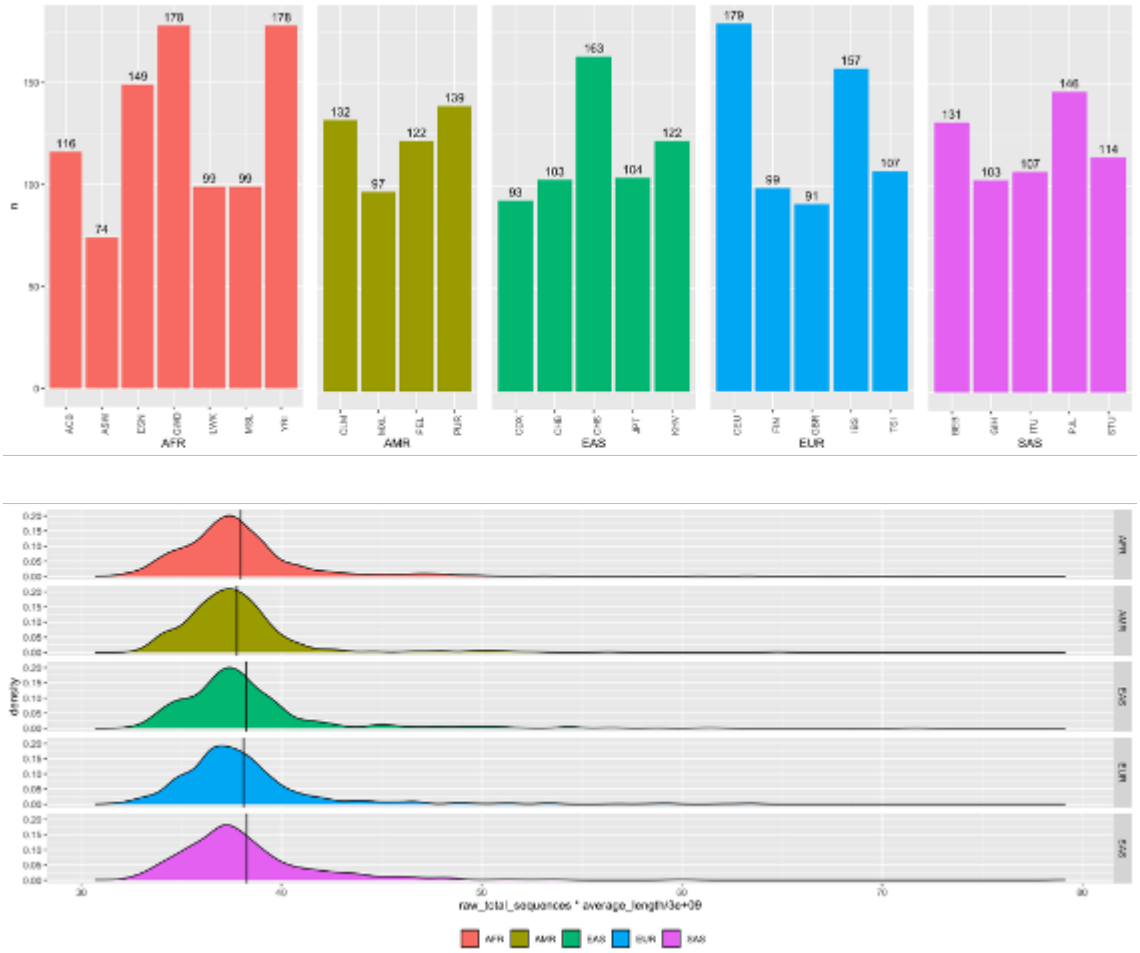


Figure S3.18. 1KGP Sample overview. (top) Total number of samples per population, including children in trios, grouped by superpopulation. (bottom) Violin plot of the total amount of raw sequencing coverage available per sample, assuming a 3.0Gbp genome size. Black vertical lines indicate the mean coverage per superpopulation (AFR: African, AMR: Admixed American, EAS: East Asian, EUR: European, SAS: South Asian).

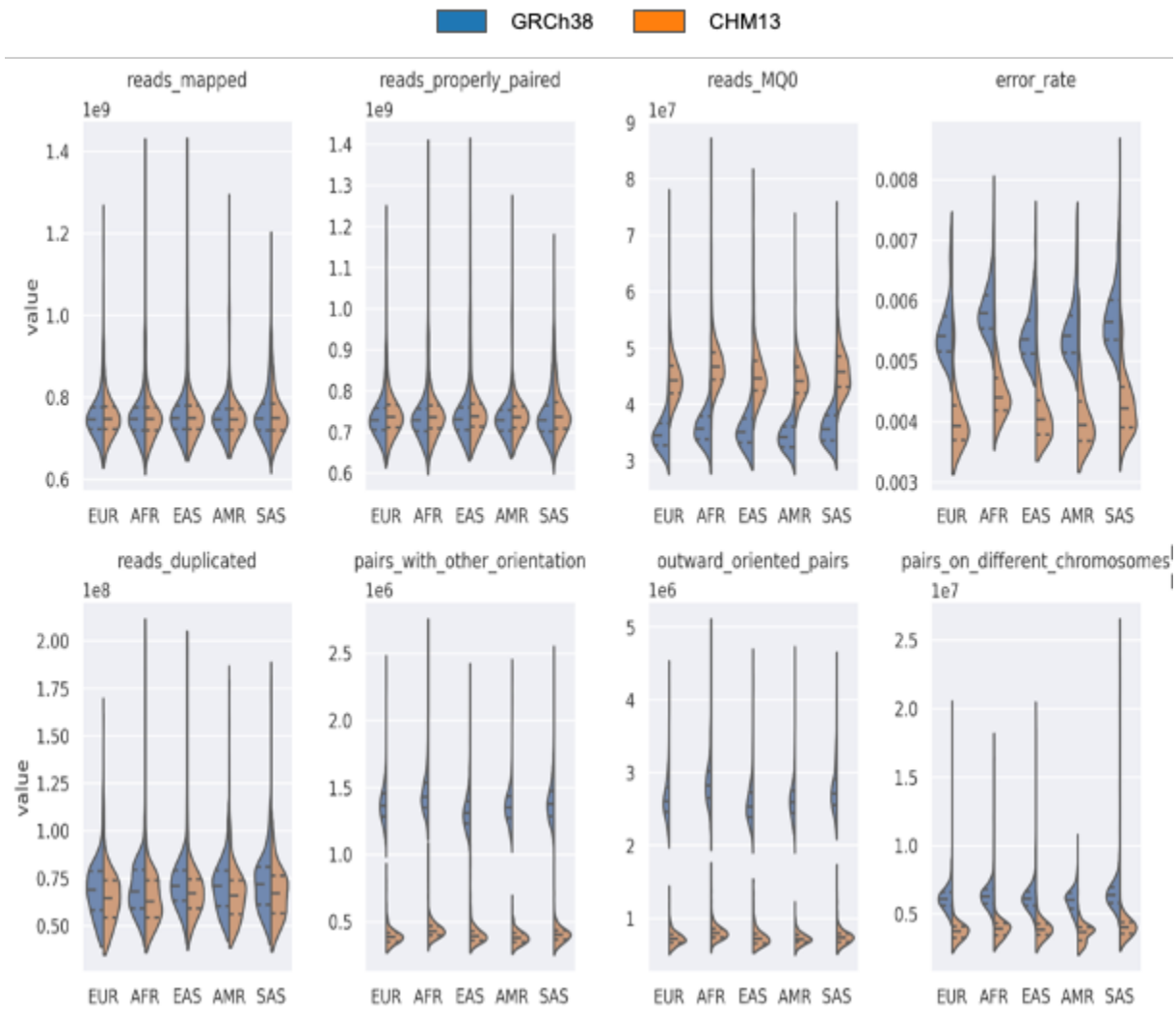


Figure S3.19. Short-read mapping statistics generated using samtools stats with GRCh38 and CHM13 as alignment target references. Results are stratified by superpopulation codes as per 1KGP dataset. Distribution quartile values are shown as dashed lines inside violin plots. Results for GRCh38 are shown in blue, CHM13 in orange for each superpopulation (AFR: African, AMR: Admixed American, EAS: East Asian, EUR: European, SAS: South Asian).

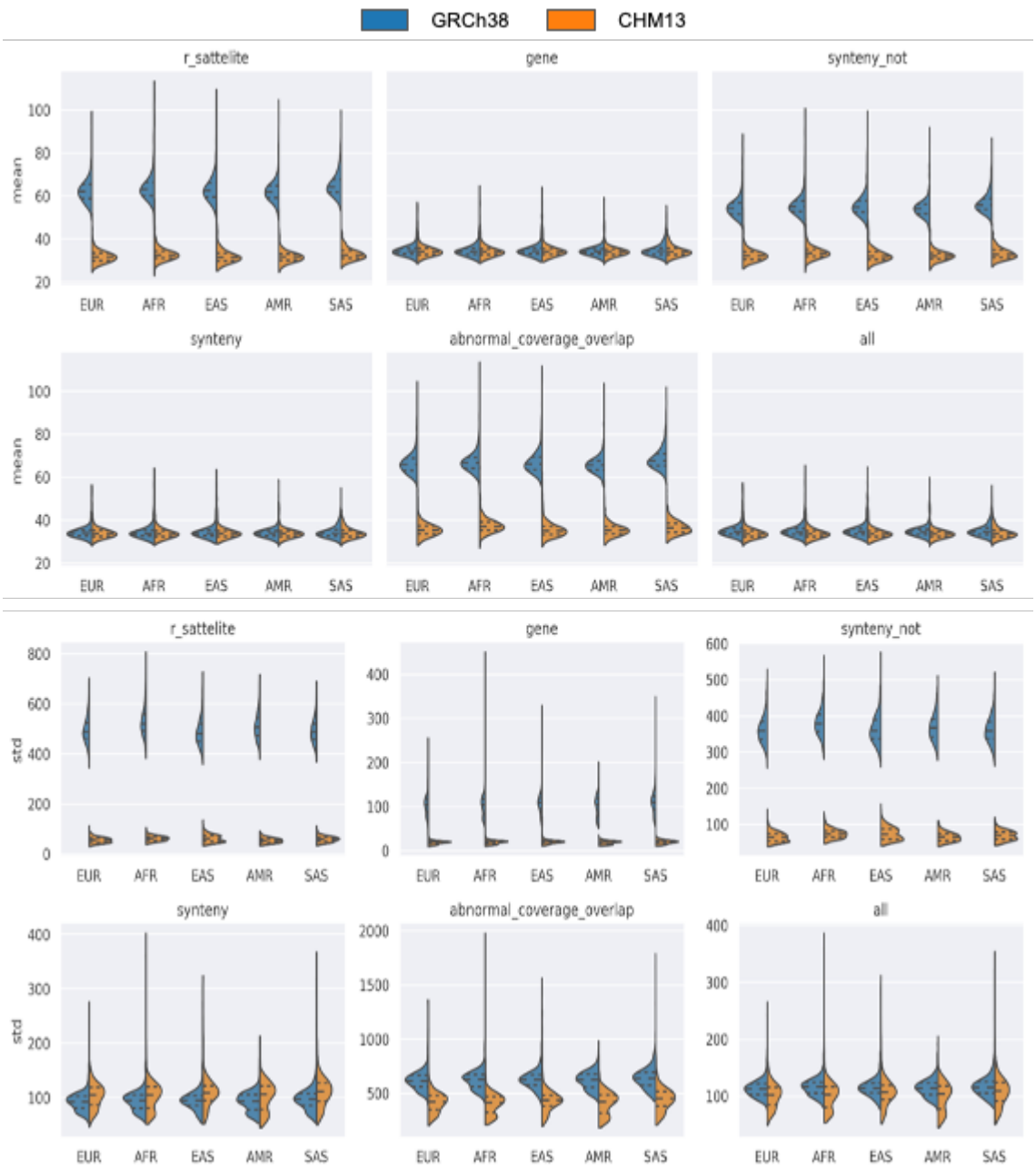


Figure S3.20. Short-read coverage statistics. Distributions of per-sample (A) mean and (B) standard deviation values for read depth context-stratified 500bp-windowed intervals with alignment target references GRCh38 (blue) and CHM13 (orange) are grouped based on 1KGP superpopulation sample annotations (AFR: African, AMR: Admixed American, EAS: East Asian, EUR: European, SAS: South Asian). Distribution quartiles are shown as dashed lines.

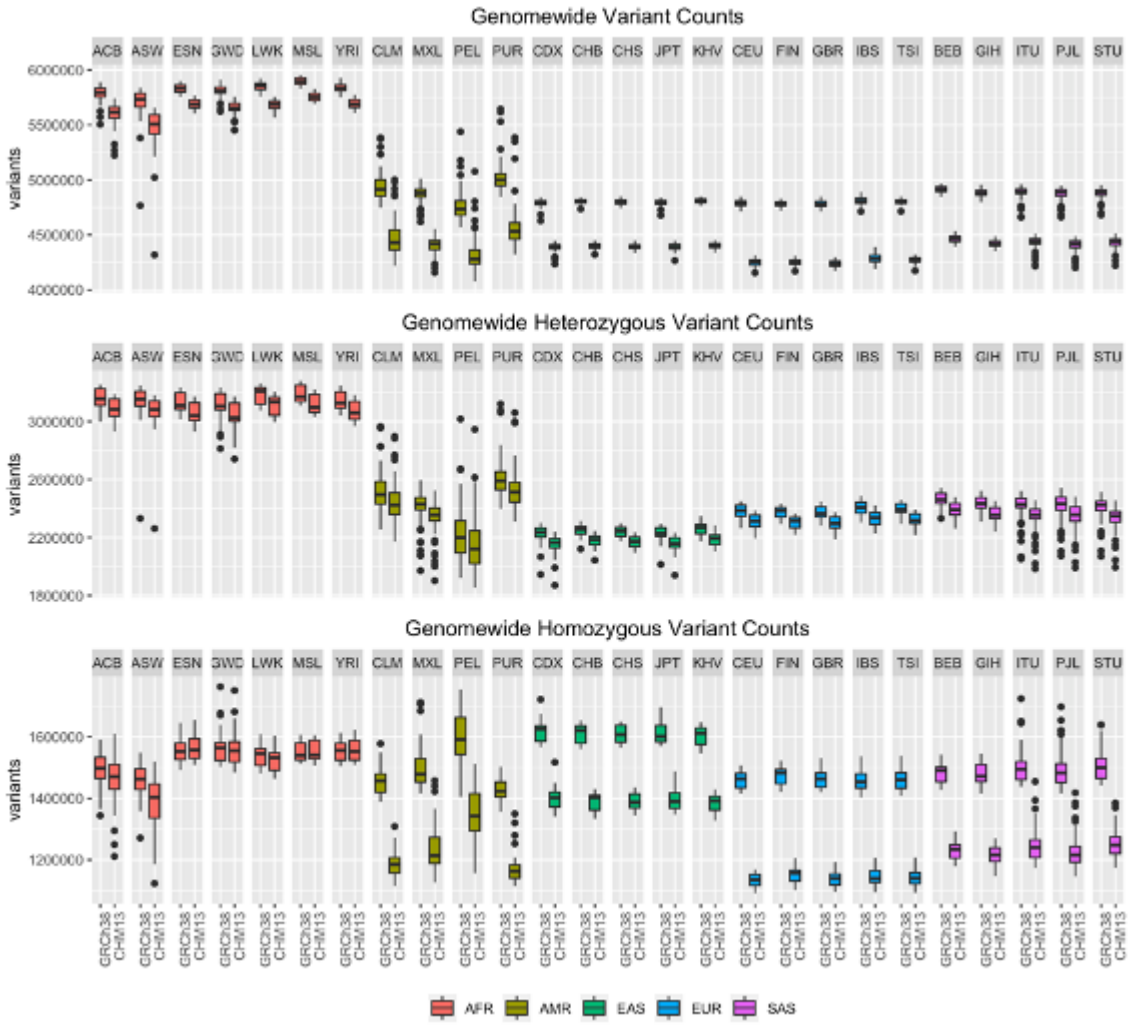
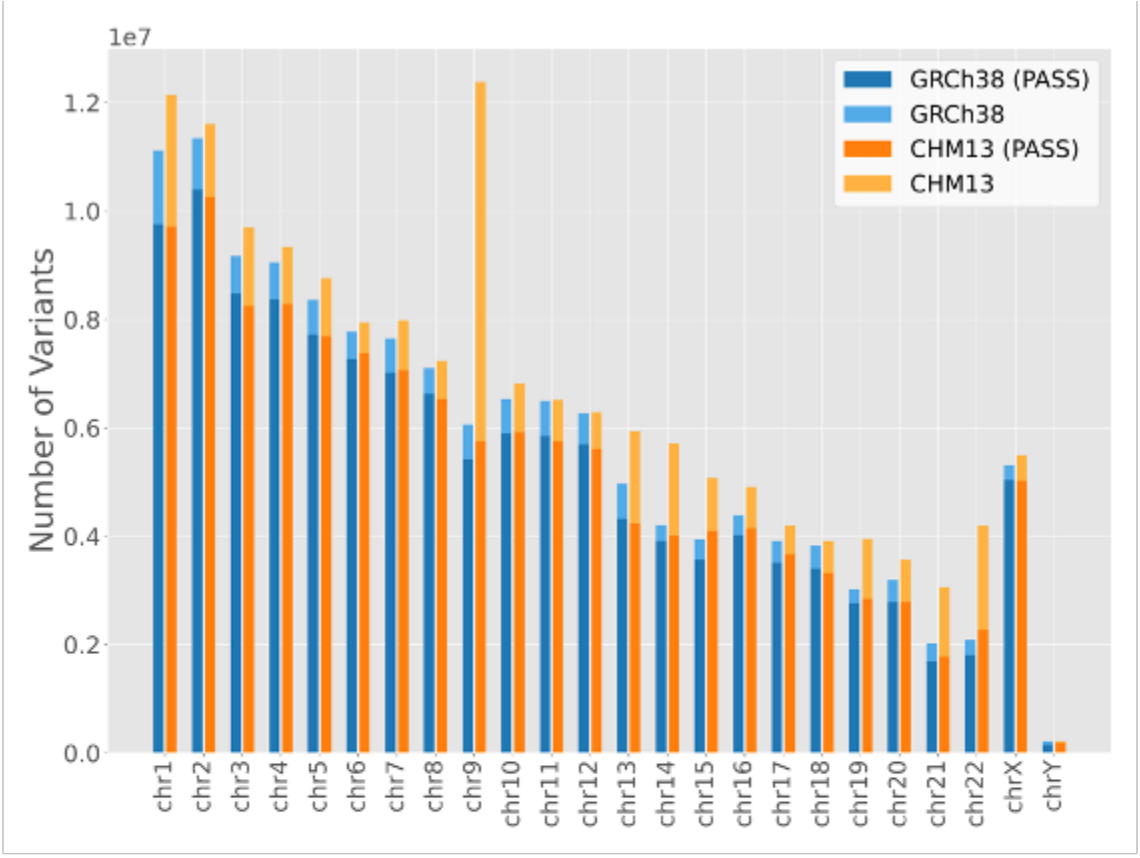


Figure S3.21. Population-specific variant counts. Population-specific boxplots of the number of (top) all variants, (middle) heterozygous variants, and (bottom) homozygous variants per sample, as computed in Figure 2B. Colors highlight superpopulations (AFR: African, AMR: Admixed American, EAS: East Asian, EUR: European, SAS: South Asian) for each population (CHB:Han Chinese, JPT:Japanese, CHS:Southern Han Chinese, CDX:Dai Chinese, KHV:Kinh Vietnamese, CHD:Denver Chinese, CEU:CEPH, TSI:Tuscan, GBR:British, FIN:Finnish, IBS:Spanish, YRI:Yoruba, LWK:Luhya, GWD:Gambian, MSL:Mende, ESN:Esan, ASW:African-American SW, ACB:African-Caribbean, MXL:Mexican-American, PUR:Puerto Rican, CLM:Colombian, PEL:Peruvian, GIH:Gujarati, PJL:Punjabi, BEB:Bengali, STU:Sri Lankan, ITU:Indian).



	GRCh38 (SNPs / indels)	CHM13 (SNPs / indels)
All	138,044,724 (119,092,057 / 18,952,667)	156,983,441 (134,902,604 / 22,080,837)
Filtered by PASS	125,484,020 (111,048,944 / 14,435,076)	126,591,489 (110,429,582 / 16,161,907)

Figure S3.22. Effect of PASS filtering. The per-chromosome (top) and genome-wide (bottom) number of variants in the 1KGP samples (allele count > 0) with respect to GRCh38 and CHM13, before and after filtering by GATK's "PASS" annotation.

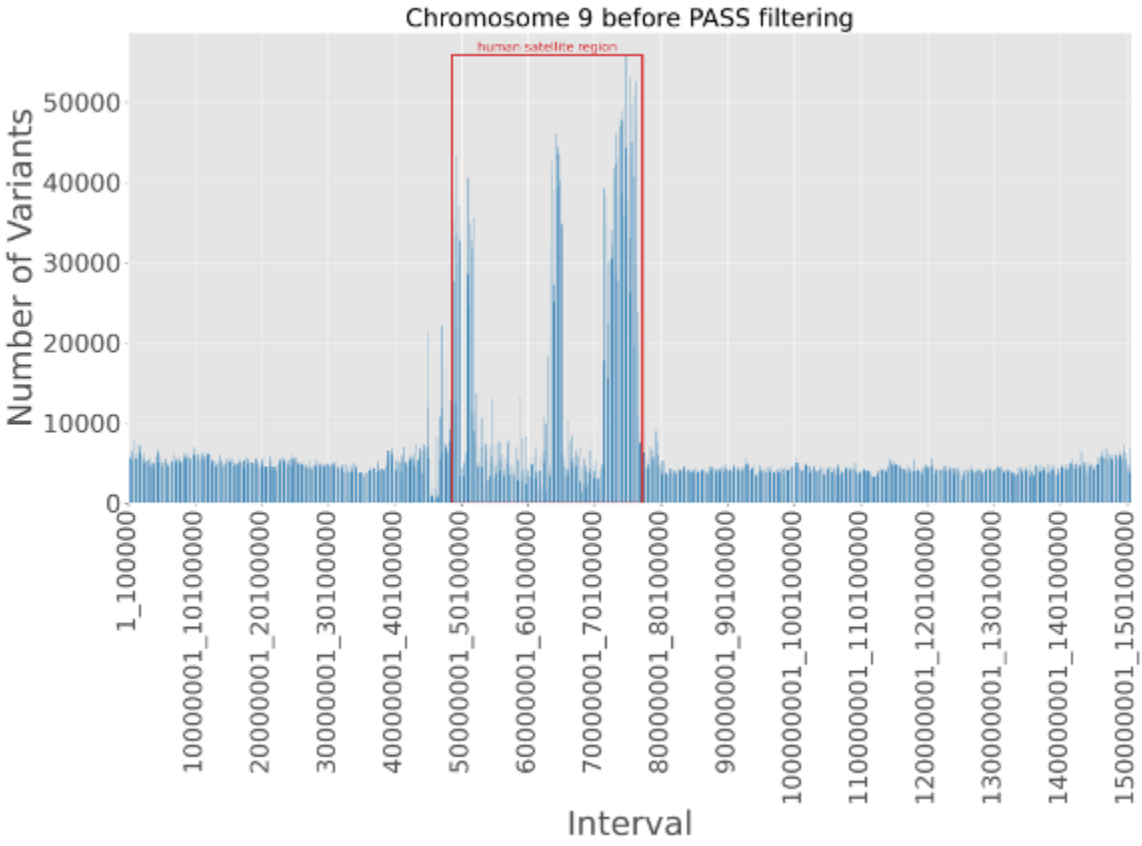


Figure S3.23. Complex regions affected by PASS filtering. The number of variants in chromosome 9 when aligned to CHM13 before filtering by the "PASS" annotation, with a complex human satellite region annotated in red. Most of these variants are filtered out with the "PASS" annotation.

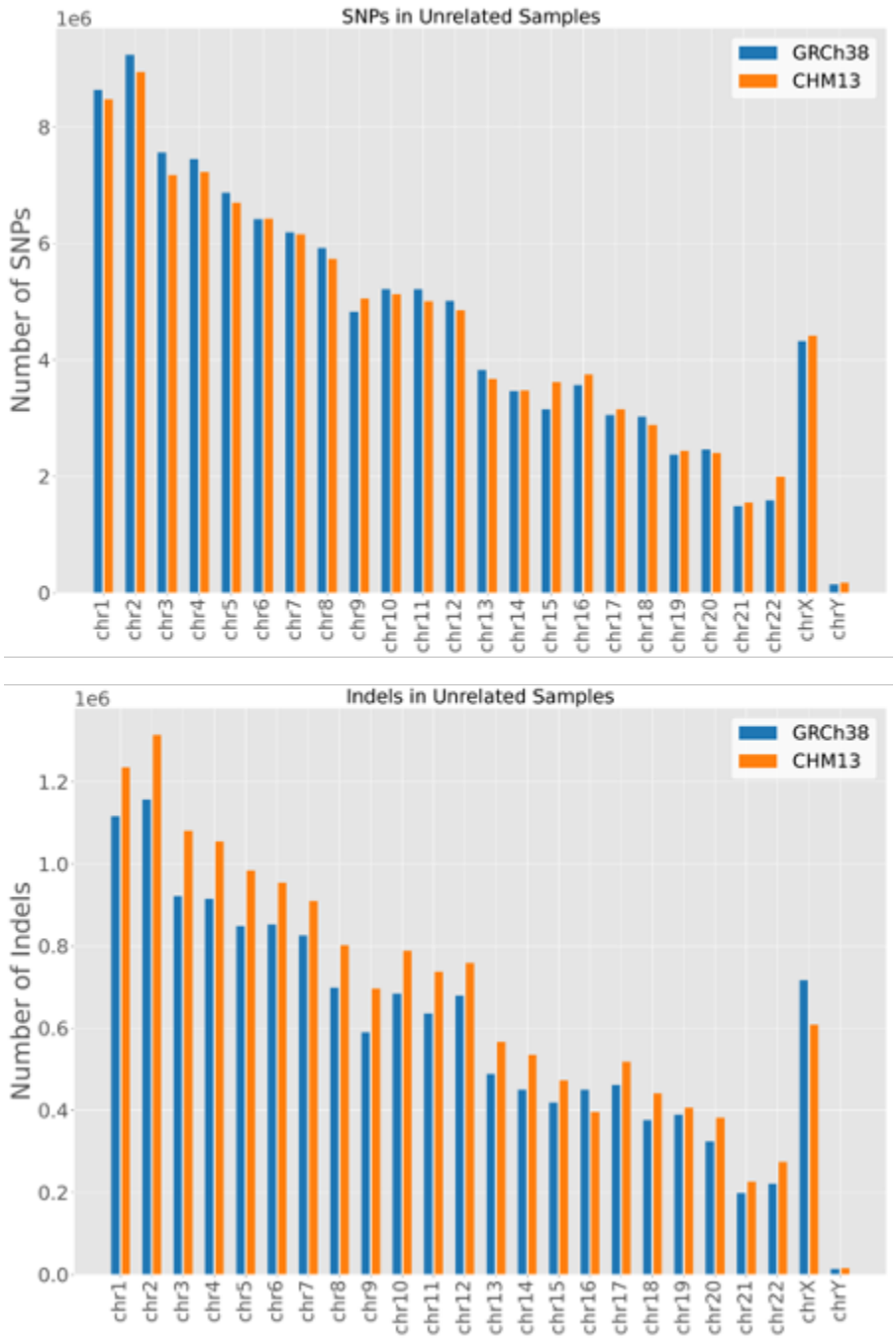


Figure S3.24. PASS-filtered SNPs and indels in unrelated samples. The number of PASS-filtered SNPs (top) and indels (bottom) across all 1KGP samples (allele count > 0) when aligned to GRCh38 (blue) and CHM13 (orange).

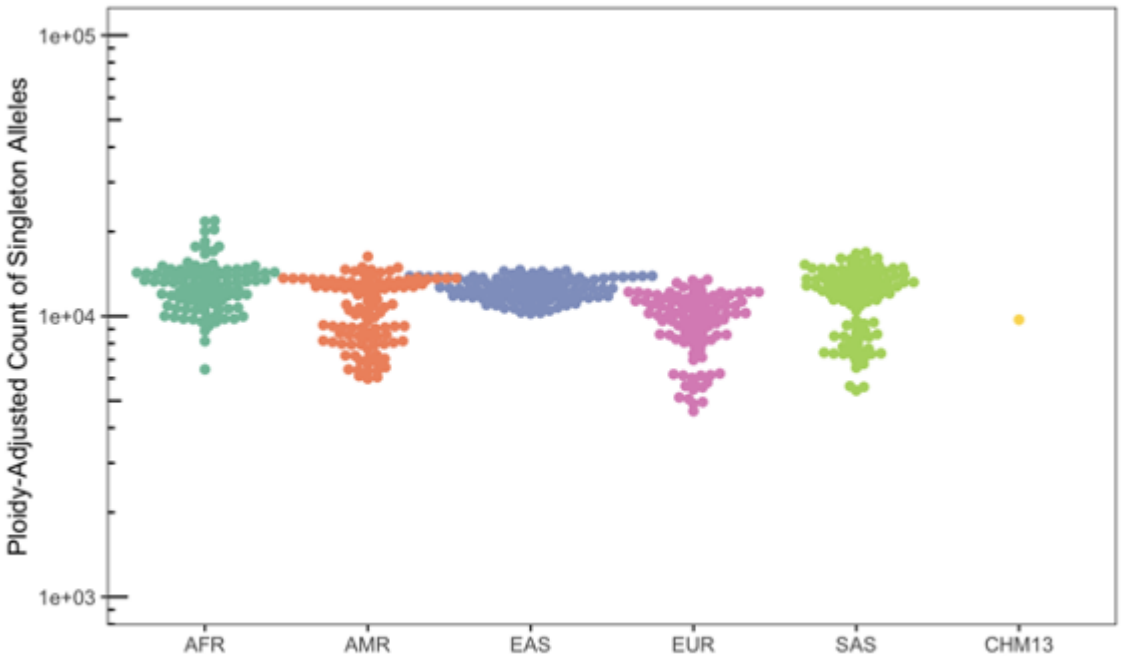


Figure S3.25. Counts of singleton alleles in 1KGP samples and CHM13. Each point represents the count of singleton alleles from one of the 1KGP individuals along with CHM13. For visual clarity, 100 random samples were selected from each 1KGP superpopulation (AFR: African, AMR: Admixed American, EAS: East Asian, EUR: European, SAS: South Asian) and the y-axis was log-scaled using the same color scheme described in the local ancestry legend of **Figure 3.1A**. The singleton count for each 1KGP sample was divided by two in order to adjust for the fact that these samples are diploid while CHM13 is effectively haploid.

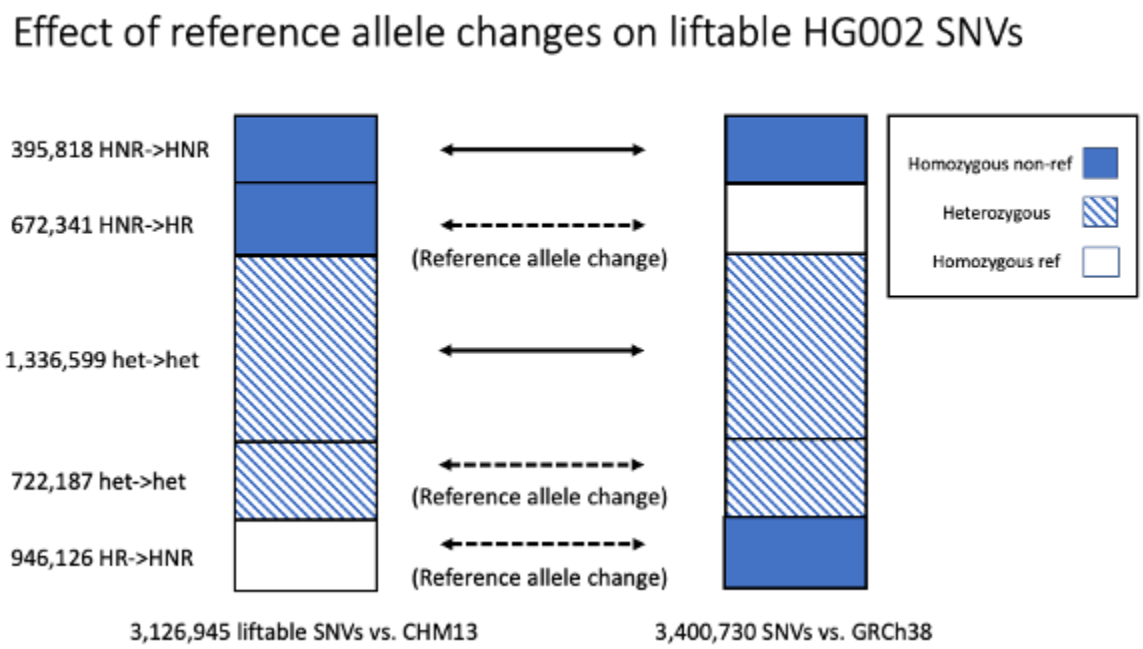


Figure S3.26. Effect of reference allele changes on SNV visibility in HG002. Many SNVs in the sample HG002 are not visible as variants with respect to either T2T-CHM13 or GRCh38 due to changes in the reference base such a homozygous variant call (HNR) changes to homozygous reference (HR). Many heterozygous SNVs (hets) also change due to reference base changes but are visible as variants on both references. Reported SNV counts are only among those with positions that lift successfully between references.

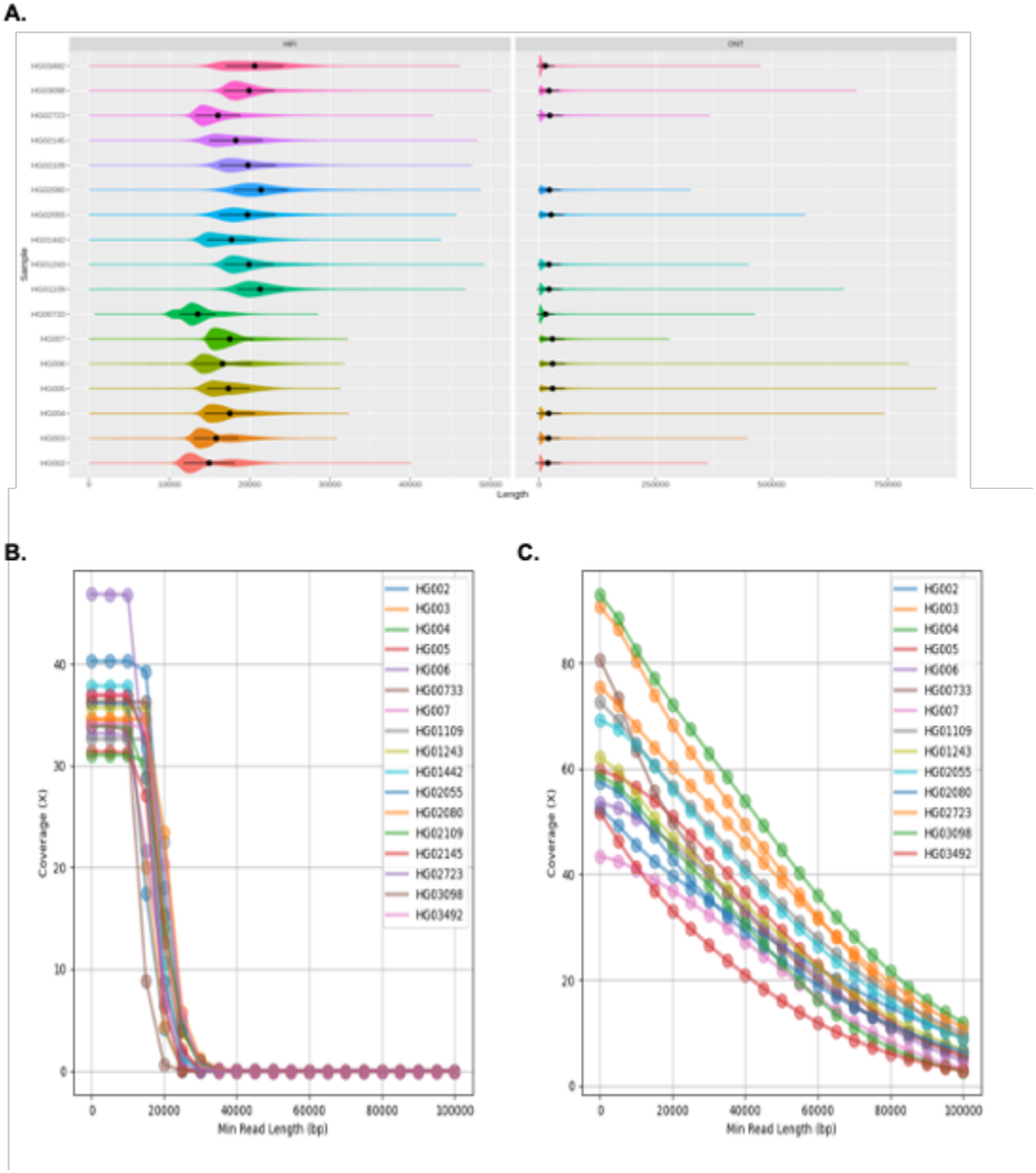


Figure S3.27. Read length distributions. (A) The distribution of HiFi read lengths in 17 samples and ONT read lengths in 14 of those samples. Points and error bars represent mean and standard deviation lengths in each sample. The mean of the per-sample mean lengths of HiFi reads is 18,130.2 bp. The mean of the per-sample mean lengths of ONT reads is 21,912.9 bp. (B) HiFi coverage in each sample as a function of minimum read length. (C) ONT coverage in each sample as a function of minimum read length.

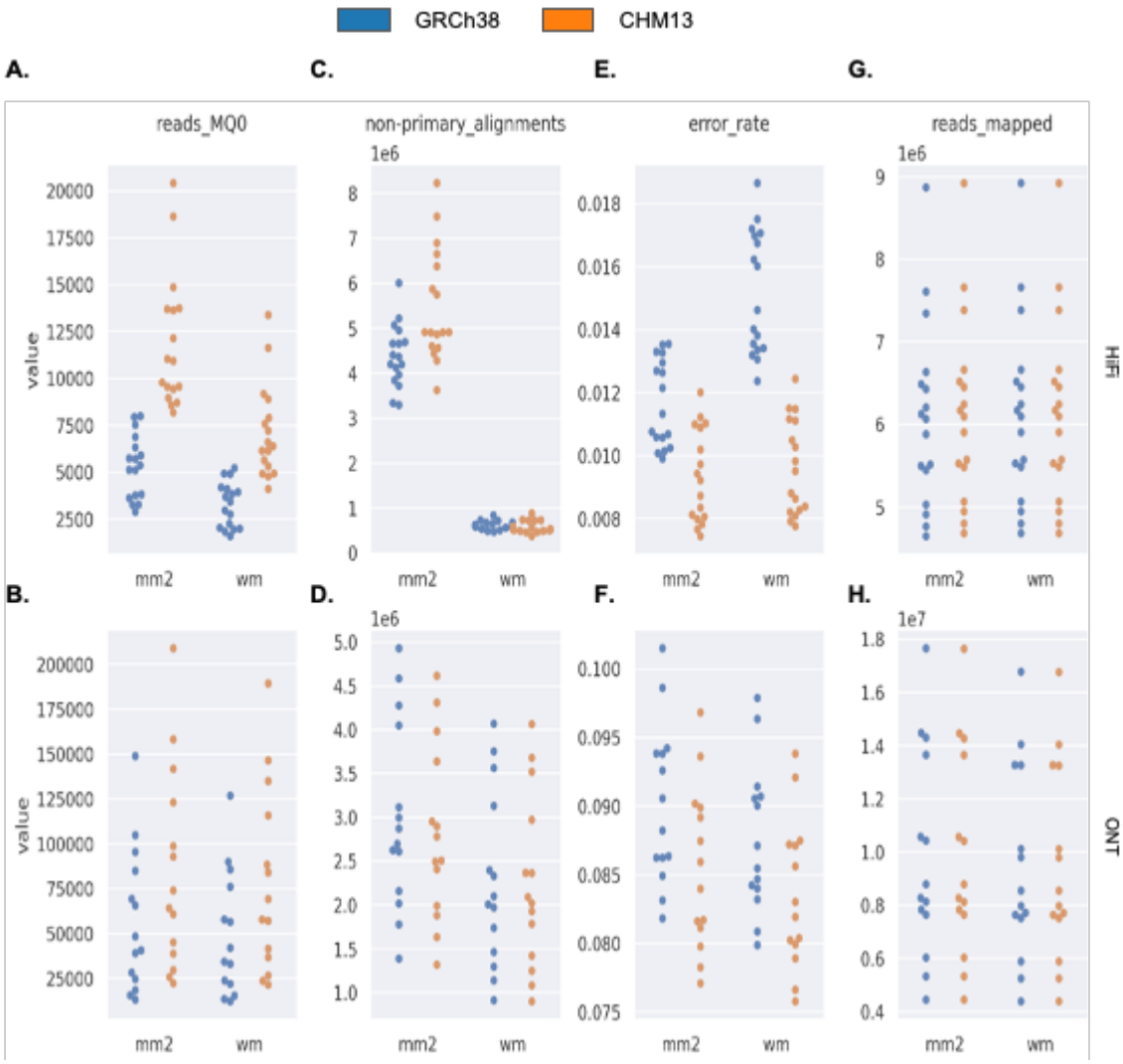


Figure S3.28. Long-Read mapping statistics generated with samtools stats.
A) The number of HiFi reads with 0 mapping quality across 17 samples in each reference.
B) The number of ONT reads with 0 mapping quality across 14 samples in each reference.
C) The number of HiFi reads with non-primary alignments across 17 samples in each reference.
D) The number of ONT reads with non-primary alignments across 14 samples in each reference.
E) The average error rate of HiFi reads across 17 samples in each reference.
F) The average error rate of ONT reads across 14 samples in each reference.
G) The number of HiFi reads mapped across 17 samples in each reference.
H) The number of ONT reads mapped across 14 samples in each reference.
mm2: minimap2; wm: winnowmap

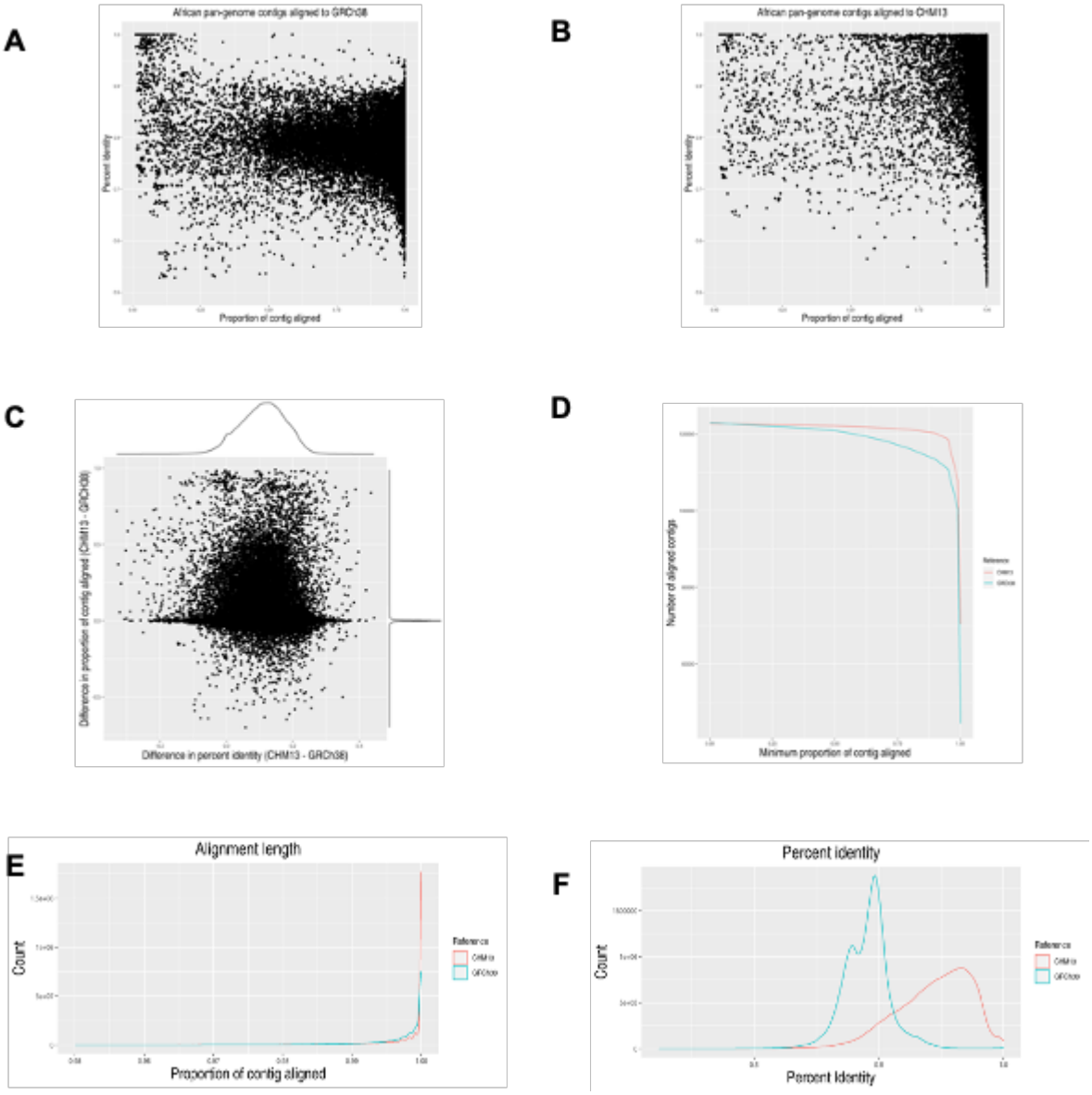


Figure S3.29. The results of aligning 124,240 contigs which were assembled from Illumina reads from 910 individuals of African descent that failed to align to GRCh38. (A) The proportion of each contig spanned by its longest alignment to GRCh38 and percent identity of those alignments. (B) The proportion of each contig spanned by its longest alignment to CHM13 and percent identity of those alignments. (C) The difference in alignment length and percent identity when aligning to CHM13 vs. GRCh38. (D) The number of contigs which meet different alignment length thresholds for each reference. (E) The distribution of proportion of contig length aligned to each reference. (F) The distribution of percent identity in alignments of the contigs to each reference.

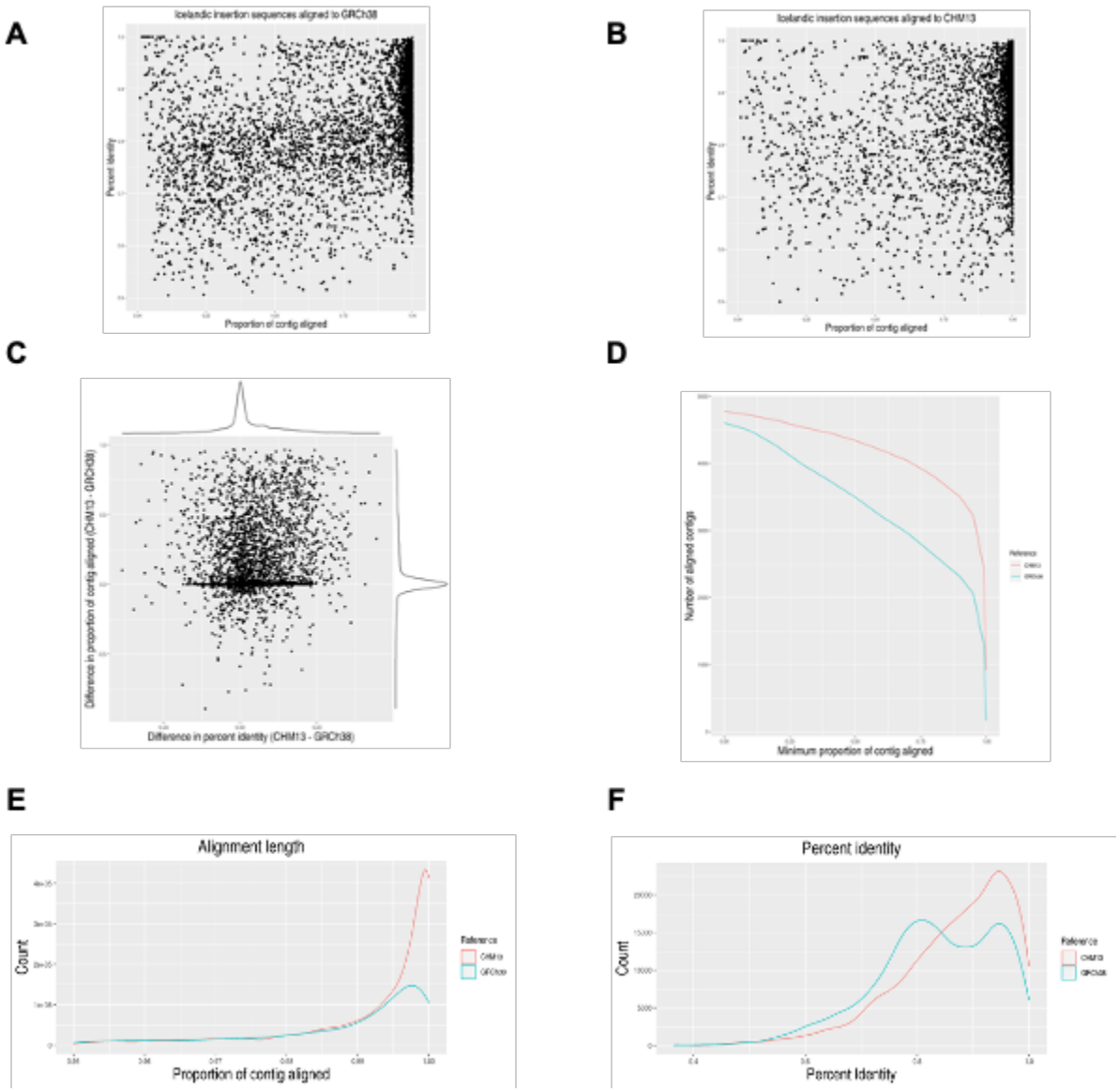


Figure S3.30. The results of aligning 4,953 insertion sequences >= 1kbp ("contigs") which were obtained by calling SVs in 3,622 Icelanders with respect to GRCh38. (A) The proportion of each contig spanned by its longest alignment to GRCh38 and percent identity of those alignments. (B) The proportion of each contig spanned by its longest alignment to CHM13 and percent identity of those alignments. (C) The difference in alignment length and percent identity when aligning to CHM13 vs. GRCh38. (D) The number of contigs which meet different alignment length thresholds for each reference. (E) The distribution of proportion of contig length aligned to each reference. (F) The distribution of percent identity in alignments of the contigs to each reference.

Figure S3.31. Long-read abnormal coverage statistics. The mean and standard deviation of coverage among 500bp bins in CHM13 and GRCh38 when using different combinations of aligners and sequencing technologies. The overall counts are displayed, as well as bins which overlap satellite repeats, genes, non-syntenic regions with respect to the other reference, syntenic regions, and bins with abnormal levels of coverage.
mm2: minimap2; wm: winnowmap.

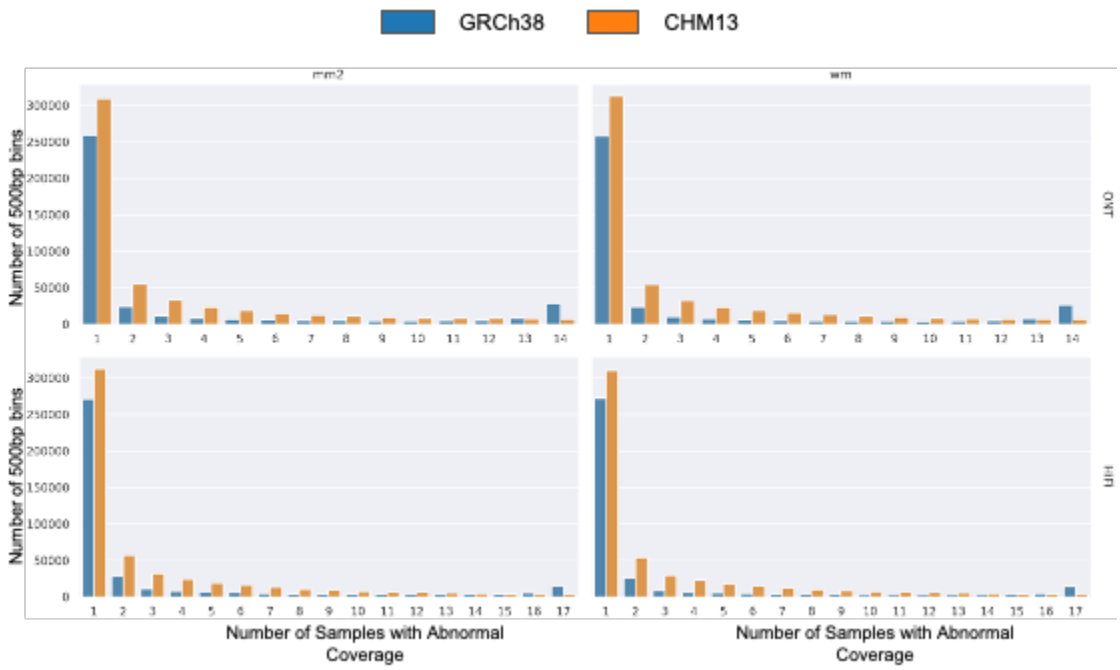


Figure S3.32. Long-read abnormal coverage frequency. A histogram of the number of 500bp bins with different frequencies of abnormal coverage among the samples studied. Here abnormal coverage is defined as outside the range [Median - 1.5 (Median - Q1), Median + 1.5 (Q3 - Median)] among all bins in the same reference.
mm2: minimap2; wm: winnowmap.

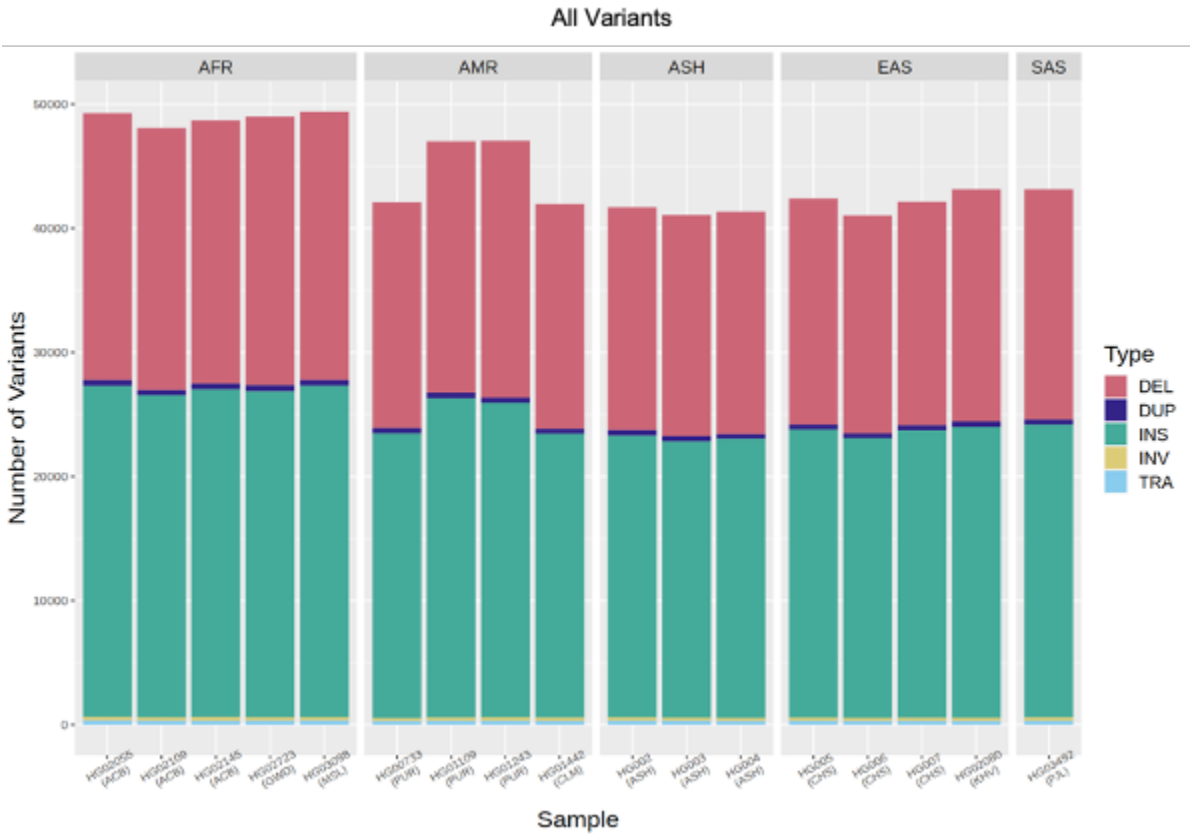


Figure S3.33. Per-sample SV counts. The counts of HiFi-derived SV calls in each of the 17 samples studied (DEL:Deletion, DUP:Duplication, INS:Insertion, INV:Inversion, TRA:Translocation). Counts are post-merging, so include variants which were called with high-confidence in that sample, as well as variants which were called with low-confidence in that sample but which were merged with high-confidence calls in other samples.

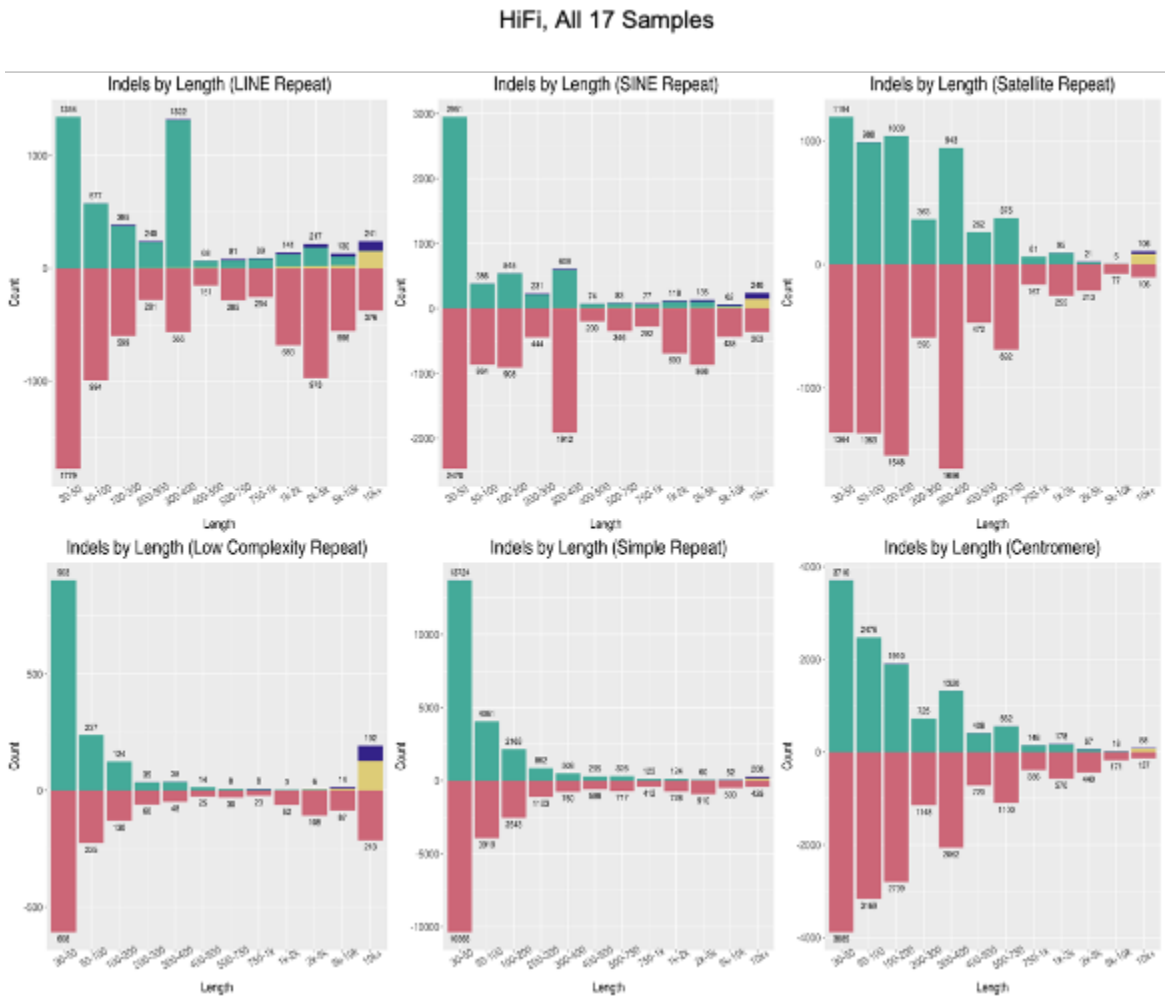


Figure S3.35. CHM13 variants in repeats. The lengths and types of SVs in CHM13 overlapping various repeat classes called from HiFi data in our cohort of 17 samples. The annotations are (top left) LINE Repeats, (top middle) SINE Repeats, (top right) Satellite Repeats, (bottom left) Low complex Repeats, (bottom middle) Simple repeats, and (bottom right) Centromeres. Colors are the same as used in **Figure S3.34**.

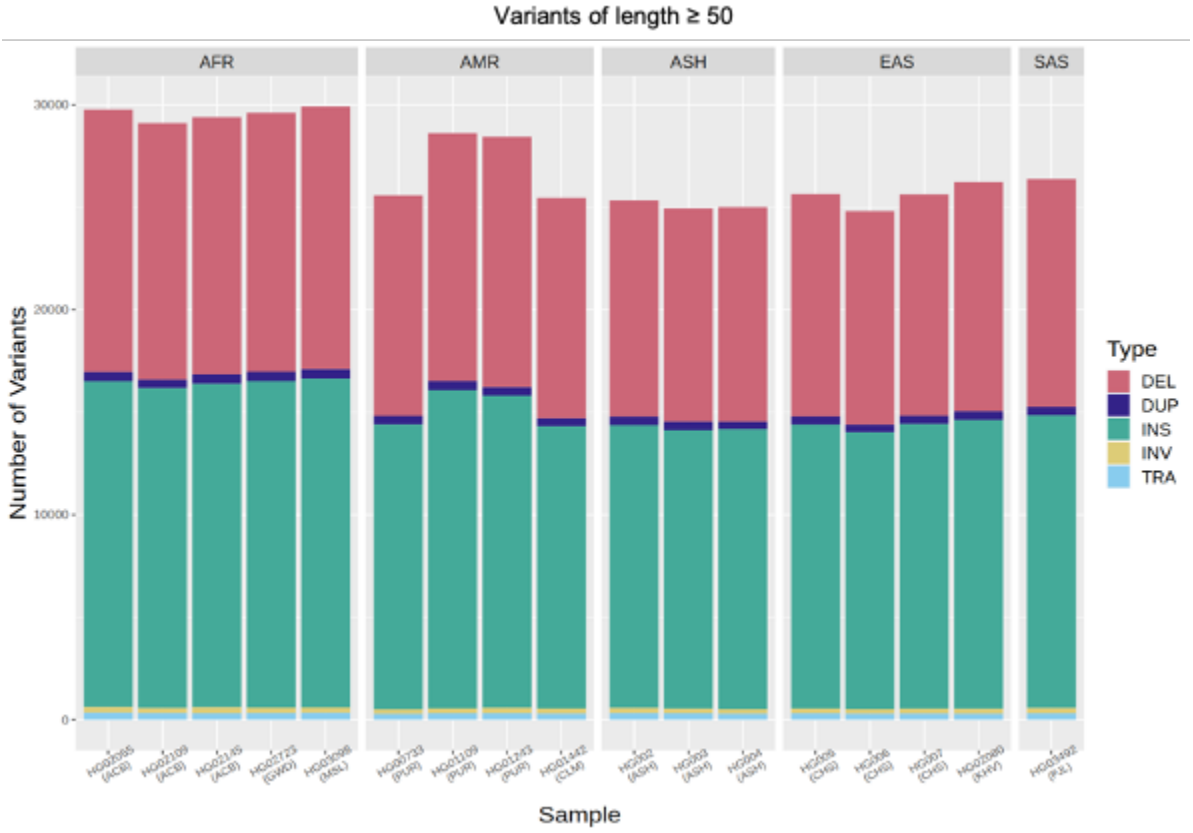


Figure S3.34. Per-sample variant counts (length ≥ 50). The counts of HiFi-derived SV calls with length at least 50, plus translocations, in each of the 17 samples studied (DEL:Deletion, DUP:Duplication, INS:Insertion, INV:Inversion, TRA:Translocation). Counts are post-merging, so include variants which were called with high-confidence in that sample, as well as variants which were called with low-confidence in that sample, but which were merged with high-confidence calls in other samples.

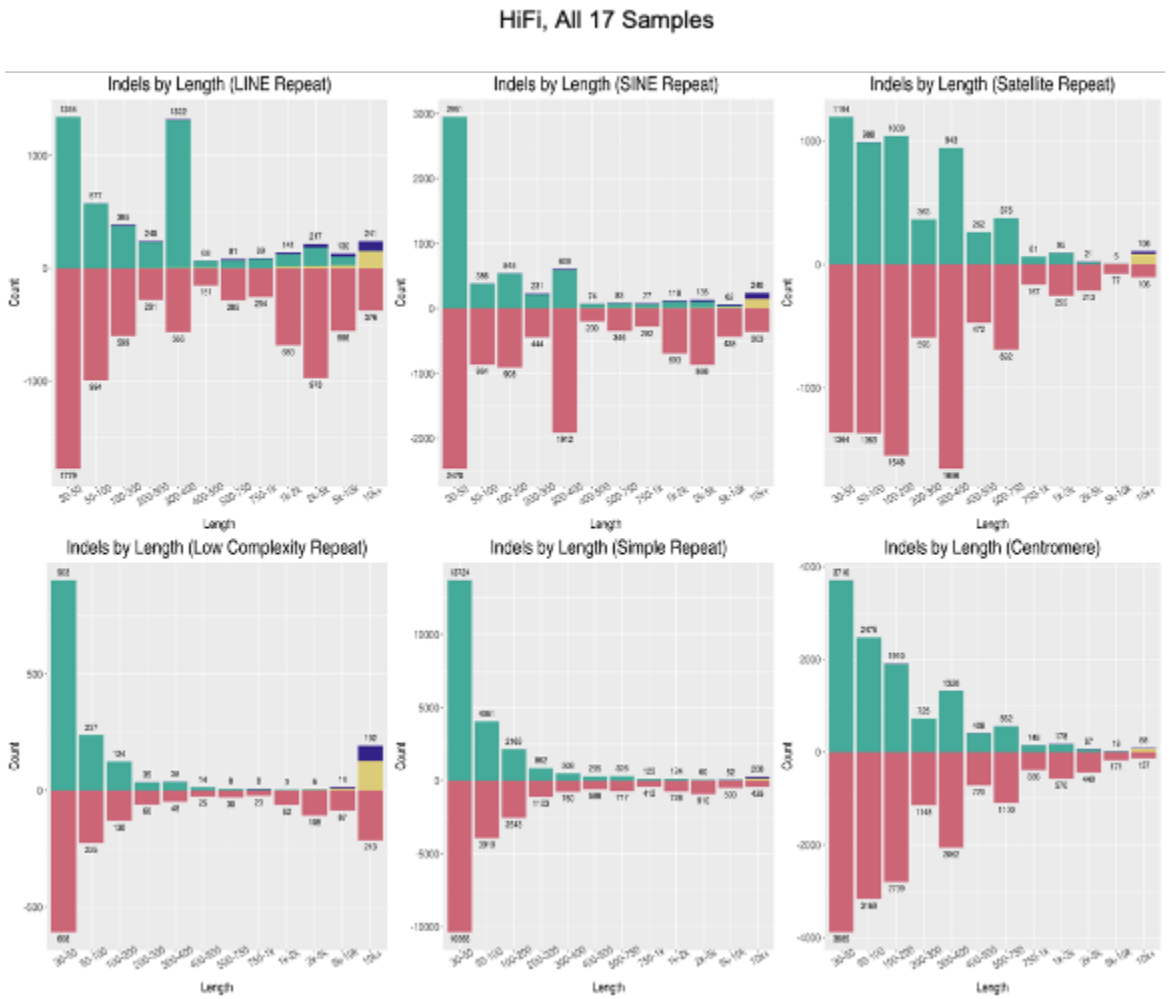


Figure S3.35. CHM13 variants in repeats. The lengths and types of SVs in CHM13 overlapping various repeat classes called from HiFi data in our cohort of 17 samples. The annotations are (top left) LINE Repeats, (top middle) SINE Repeats, (top right) Satellite Repeats, (bottom left) Low complex Repeats, (bottom middle) Simple repeats, and (bottom right) Centromeres. Colors are the same as used in **Figure S3.34**.

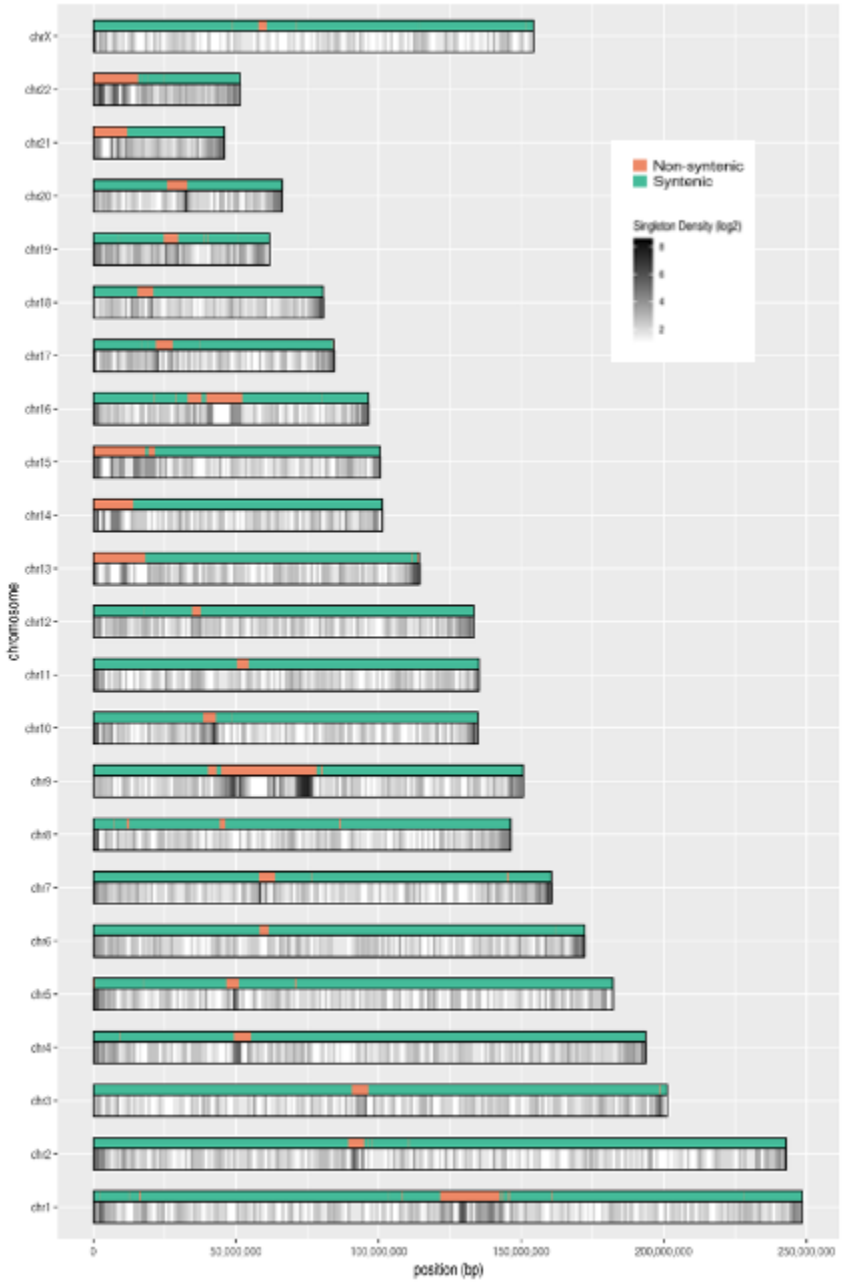


Figure S3.36. Singleton SV density across CHM13. The density of singletons SVs (i.e., present in only a single sample), across 1 Mbp bins of CHM13, represented as a grayscale heatmap across log2(number of variants). Color bands at the top highlight syntenic and non-syntenic regions of CHM13 compared to GRCh38.

SV Counts Overview

	HG002 Assembly						
	SV hg38	SV Chm13	SV unique to chm13 overlapping non-synteny	SV unique to chm13 overlapping non-synteny and at seg dups but not at centromere	SV found in both ref (hg38, chm13)	SV unique to hg38	SV unique to chm13
Deletion	1199	1379	147	37	501	698	878
Insertion	2771	1431	129	26	588	2182	843
Duplication	38	35	19	6	14	23	21
Inversion breakpoints	22	20	11	0	6	16	14

(SV counts after confidence filtering and clustering)

Figure S3.37. SV counts from Bionano. Comparison of SV calls >500 bp in HG002 from Bionano on GRCh38 (hg38) vs. T2T-CHM13, after filtering for quality and clustering equivalent calls. Balance of insertions and deletions is substantially improved. We also found a number of SVs uniquely called on T2T-CHM13 in non-syntenic regions, and in non-syntenic regions that are in segmental duplications but not in centromere-satellite regions.

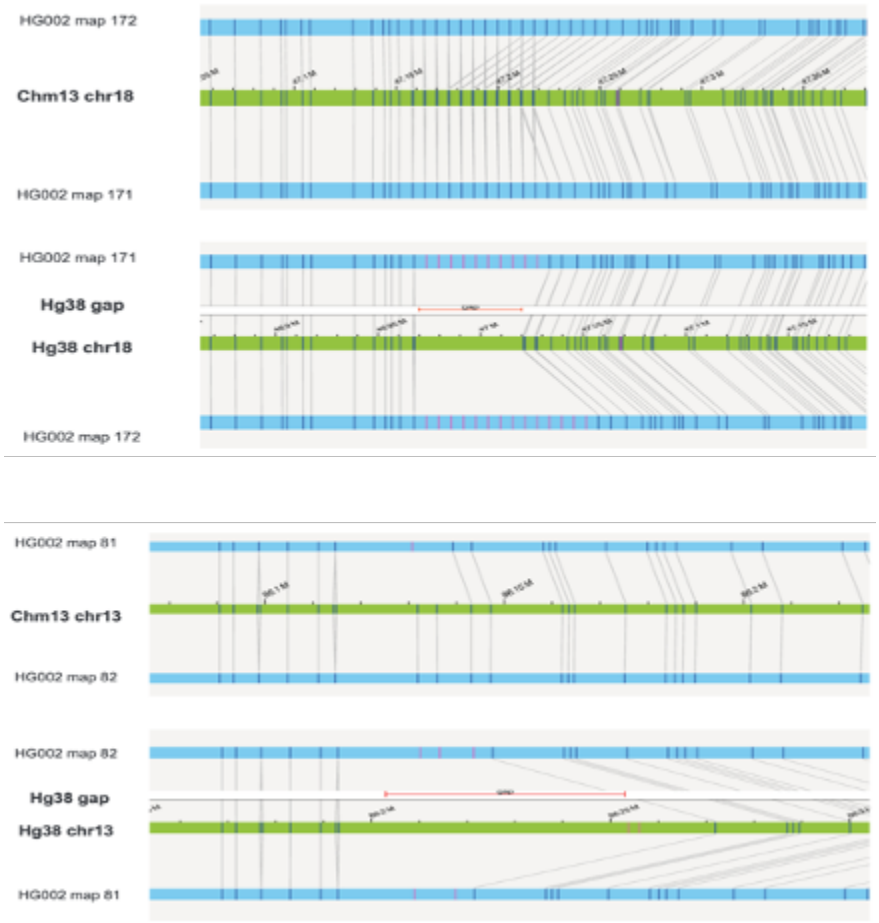


Figure S3.38. Improved Bionano alignment in CHM13 by closing gaps in GRCh38. Two examples showing improved resolution of SV calls that were gaps on chr18 and chr13 in GRCh38 but are fully sequenced in CHM13. Lines connect sequence markers in HG002 maternal and paternal assemblies to in silico digested references GRCh38 (hg38) or T2T-CHM13 (CHM13). Gaps in GRCh38 are identified by red bars.

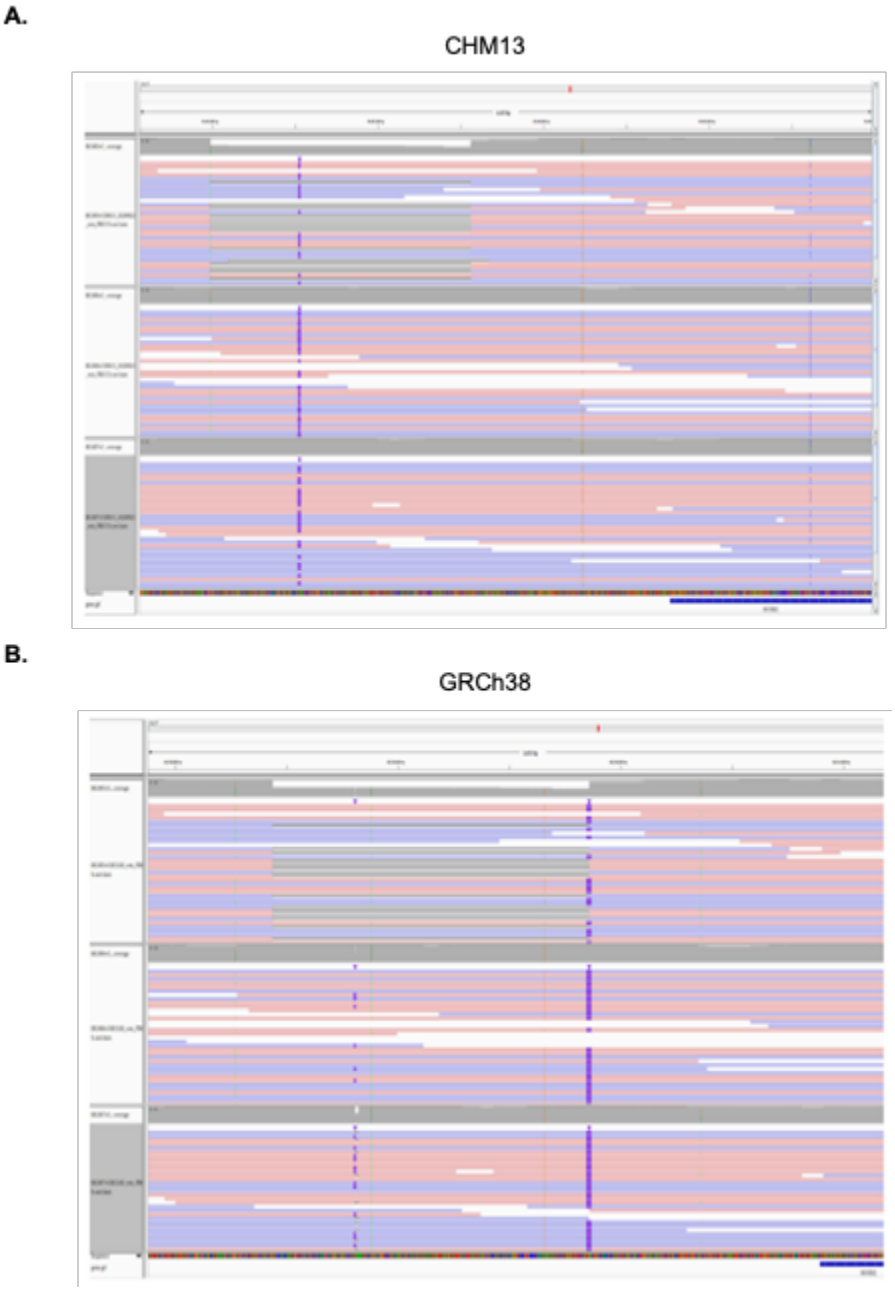


Figure S3.39. Potential de novo SV in HG005. IGV screenshots showing a putative de novo 1,571 bp deletion in HG005 at chr17:49,401,990 in CHM13. a.) The alignments of the reads of HG005 (child), HG006 (parent, 46XY), and HG007 (parent, 46XX) to CHM13 near the SV call, indicating the SV's presence in HG005 and absence in the parents. b.) The alignments of the same reads to GRCh38.

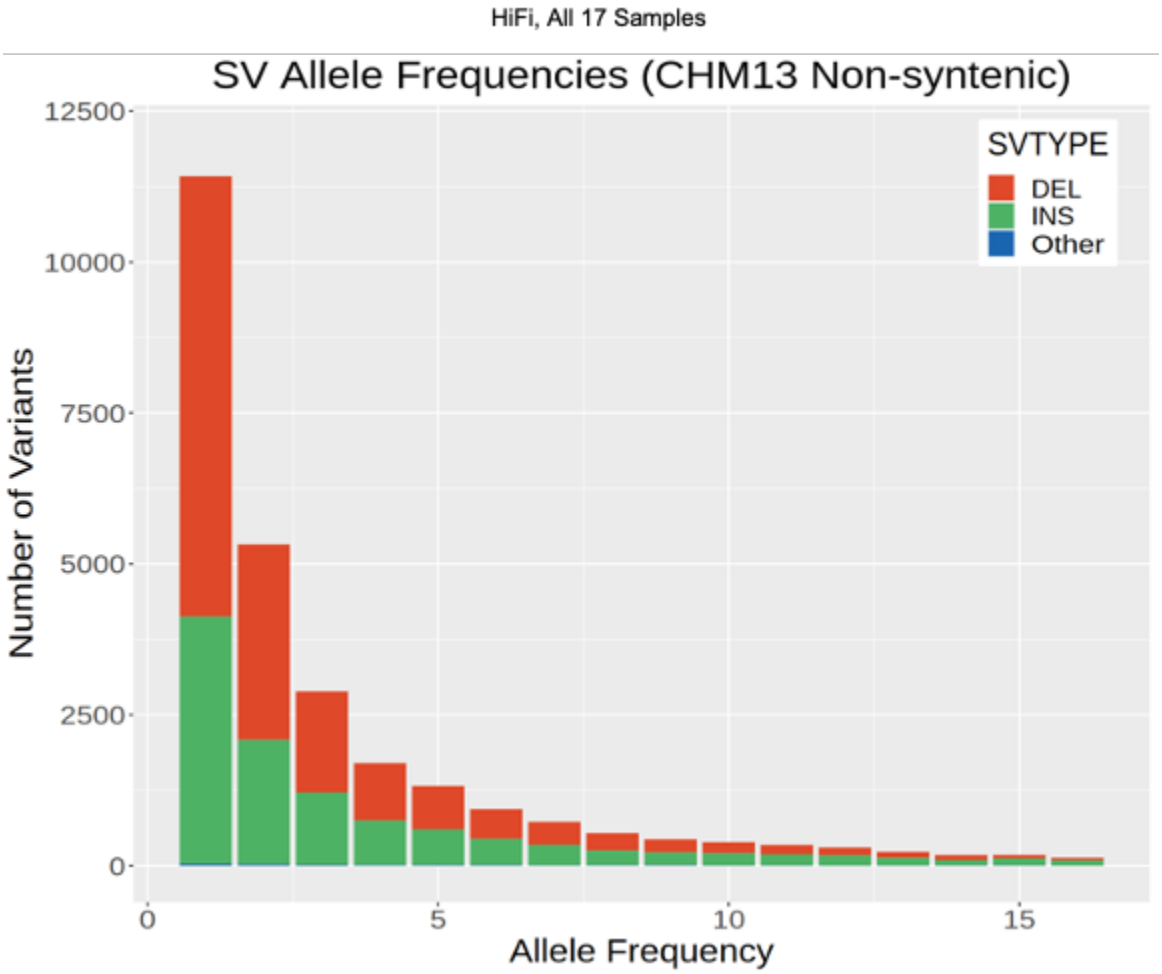


Figure S3.40. Allele frequencies of SVs in non-syntenic regions. The allele frequency of structural variants overlapping non-syntenic regions in CHM13, called from HiFi data in our cohort of 17 samples (DEL:Deletion, INS:Insertion, Other: Duplication + Inversion + Translocation).

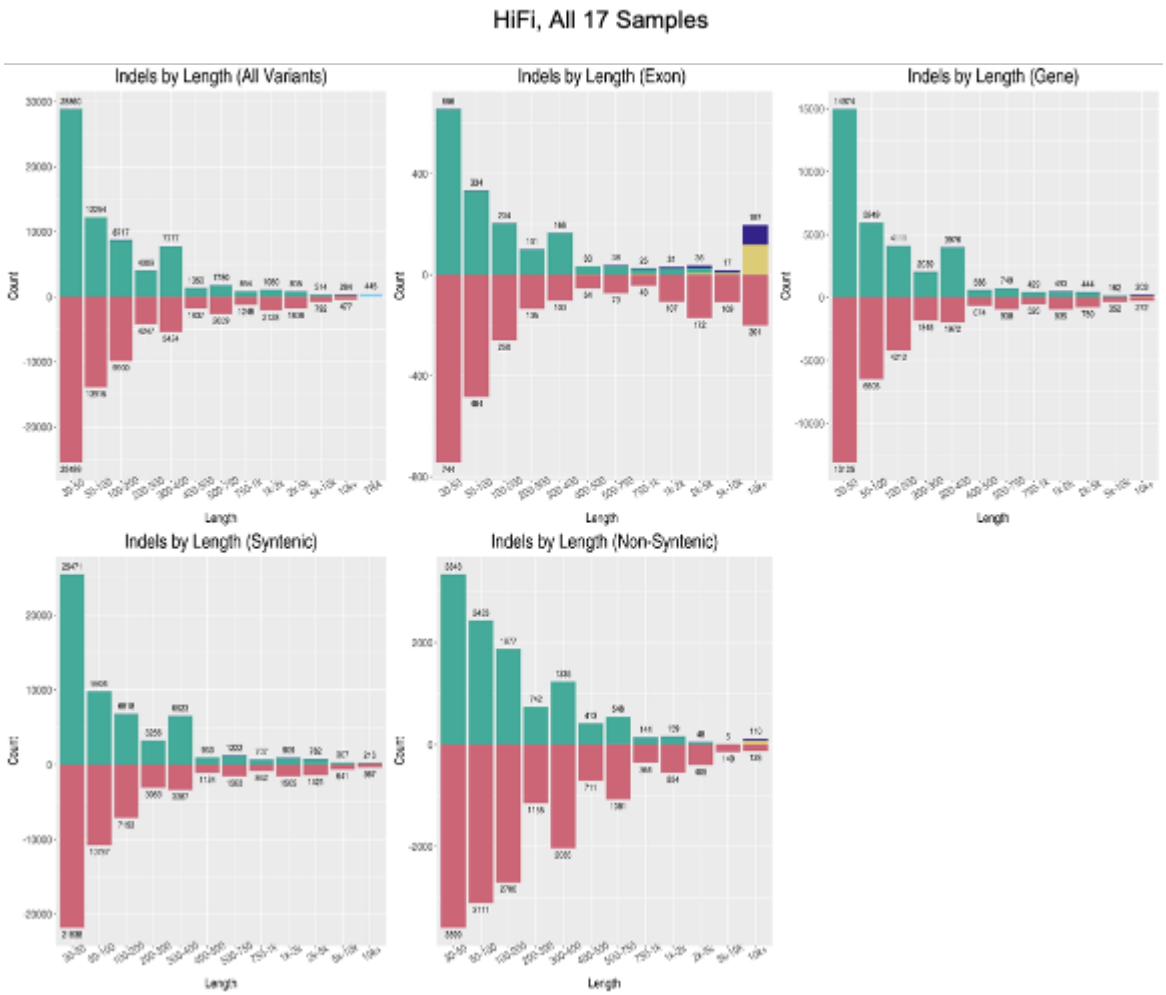


Figure S3.41. CHM13 variants across genomic contexts. The lengths and types of SVs in CHM13 overlapping genes, exons, regions syntenic to GRCh38, and regions non-syntenic to GRCh38 called from HiFi data in our cohort of 17 samples. Colors indicate the type of SV, using the same colors as in **Figure S3.34**.

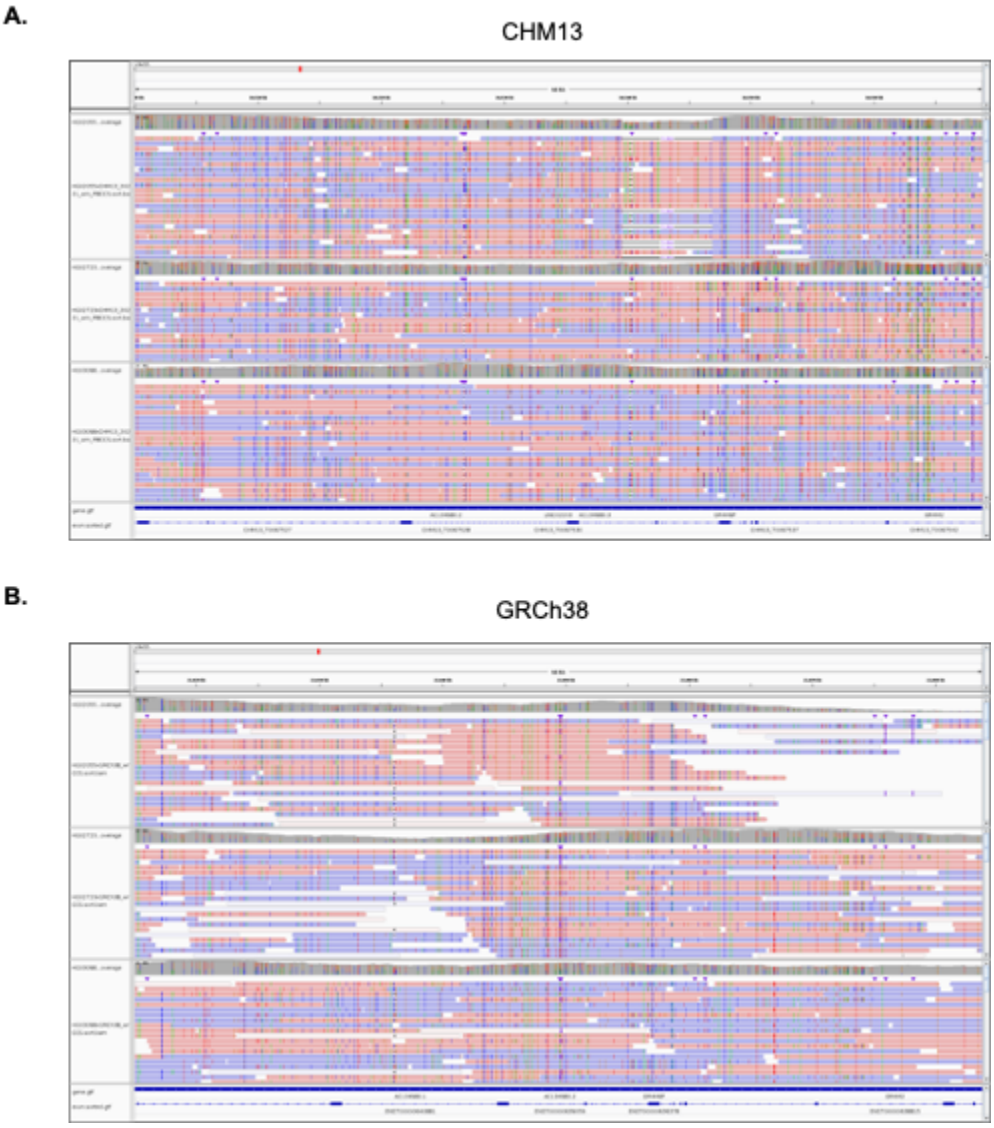


Figure S3.42. Alignments to *AC134980.2* among samples of African ancestry. IGV screenshots of alignments of three samples of African ancestry - HG02055, HG02723, and HG03098 respectively - to the region around an exon of AC134980.2 (shown in **Figure 3.4G** and **Figure 3.4H**) to a.) CHM13 and b.) GRCh38.

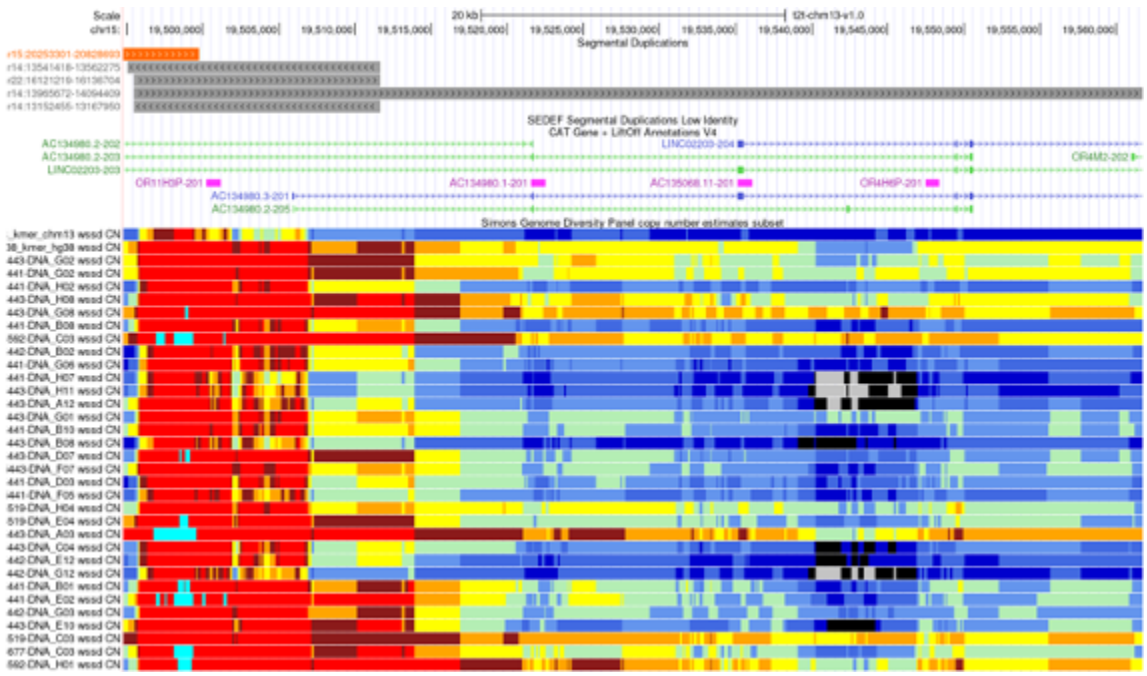


Figure S3.43. Copy number variation in *AC134980.2*. Copy number variation of the region surrounding an exon of *AC134980.2* (shown in **Figure 3.3G** and **Figure 3.3H**) among CHM13 and GRCh38, as well as samples from the Simons Diversity Panel. Color indicates copy number state using the same color palette as **Figure 3.5**. Different CAT gene colors are based on annotation type, including coding (blue), non-coding (green), and pseudogene (pink) isoforms.

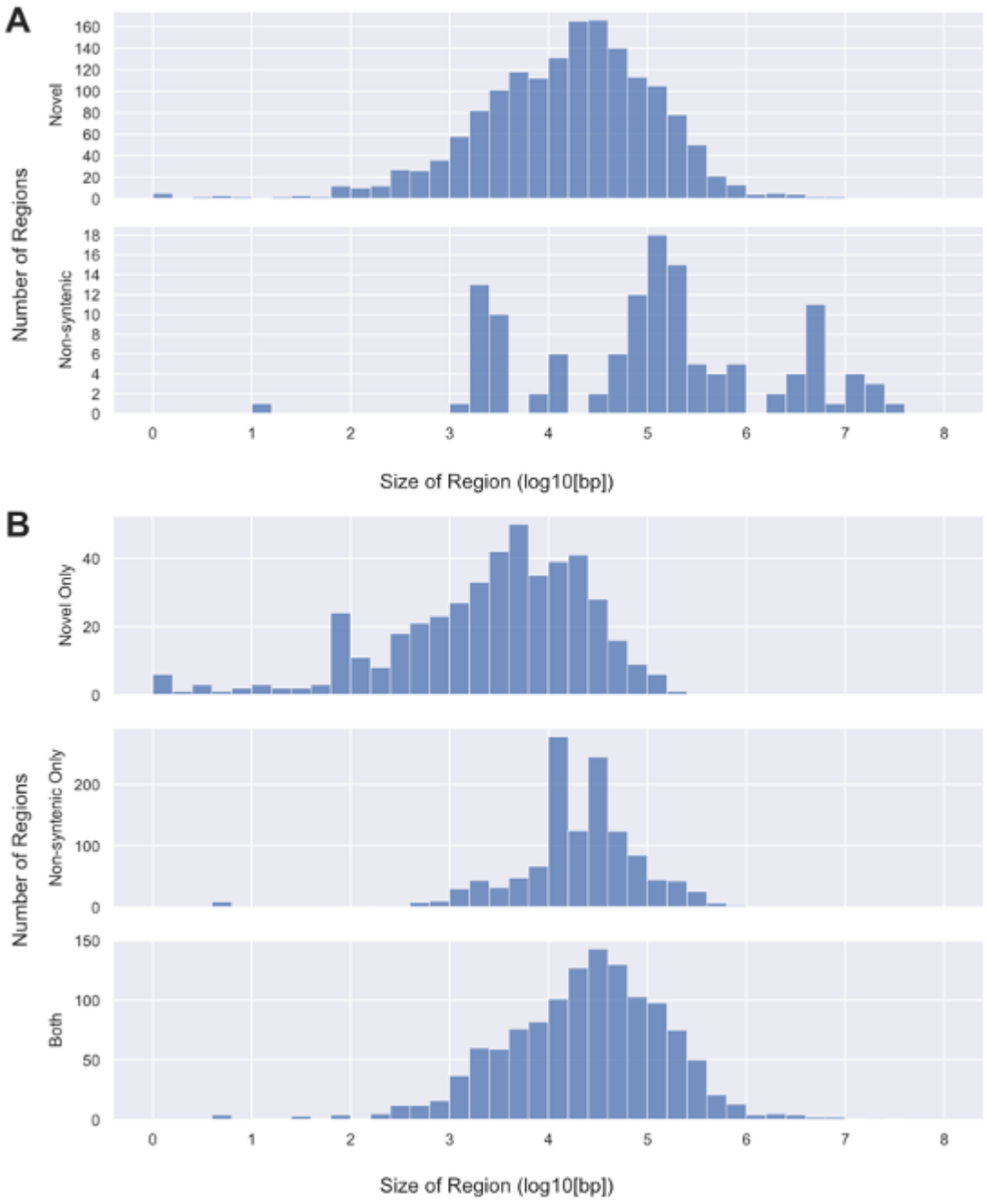


Figure S3.44. Distribution of region size for non-syntenic and previously unresolved regions. (A) Distribution of region sizes for previously unresolved (top) and non-syntenic (bottom) regions as explored in Figure 4A-C. Sizes are shown in units of $\log_{10}$(bp). (B) Distribution of region sizes for non-overlapping previously unresolved (top) and non-syntenic (middle) regions, and regions annotated as both previously unresolved and non-syntenic (bottom) as explored in Figure 4D. Sizes are shown in units of $\log_{10}$(bp).

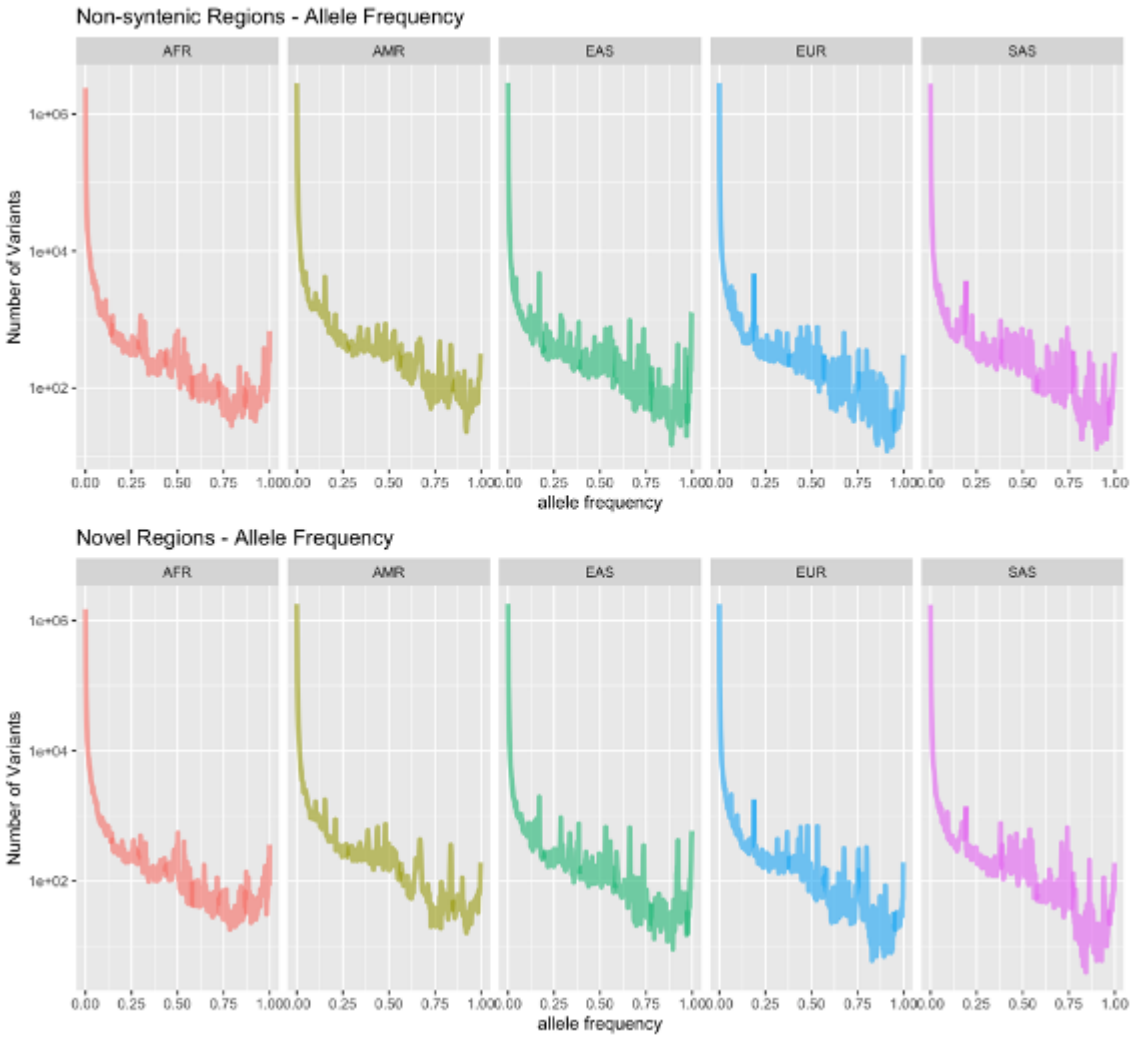


Figure S3.45. Variant allele frequencies within non-syntenic and previously unresolved regions. SNV allele frequencies within non-syntenic (top) and previously unresolved (bottom) regions across the 1KGP samples for each of the five superpopulations (AFR: African, AMR: Admixed American, EAS: East Asian, EUR: European, SAS: South Asian). Note, the color scheme here is different than depicted in the **Figure 3.1A** local ancestry legend. Values are computed using unrelated (founder) samples as in **Figure 3.2E** although only variants within the autosomes are counted here.

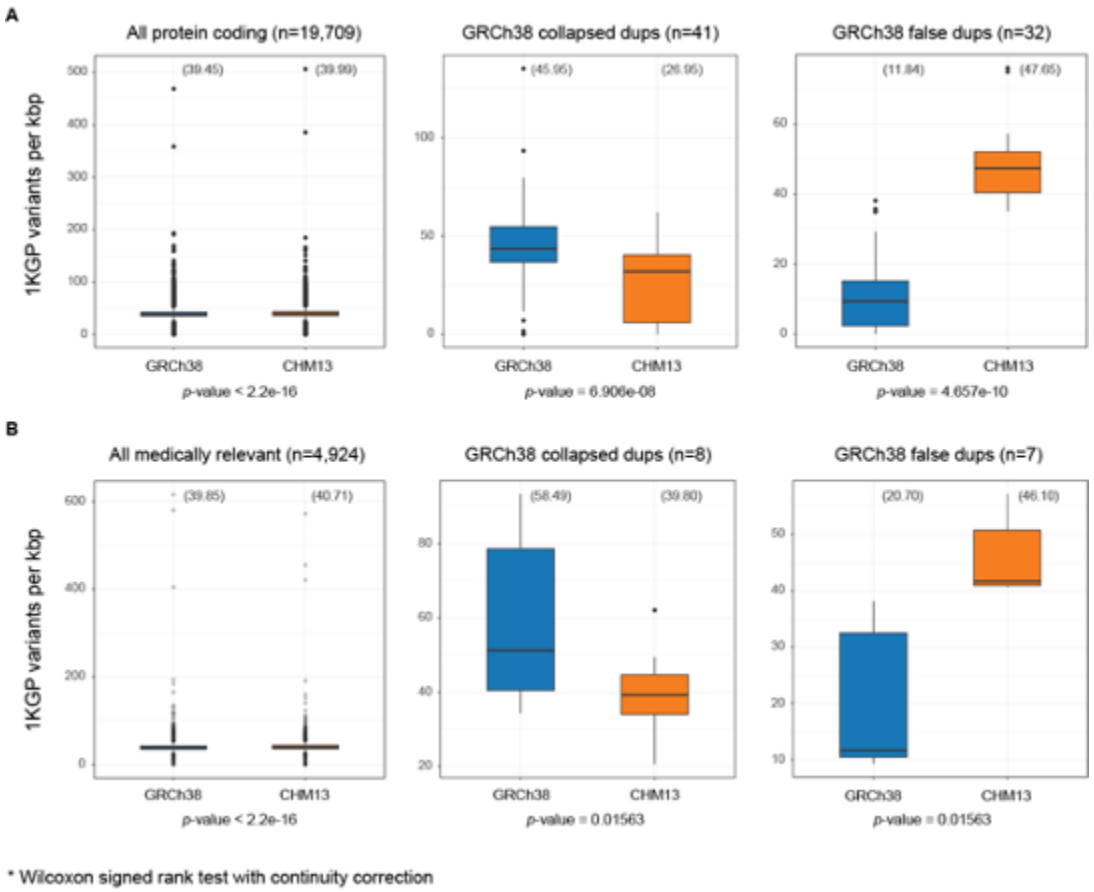


Figure S3.46. Comparison of 1 KGP variant densities within protein-coding genes between GRCh38 and T2T-CHM13. Variant densities are depicted as box plots across protein-coding genes in GRCh38 (blue) with successful lift over to T2T-CHM13 (orange) considering (A) all genes and (B) medically-relevant genes. Within these gene sets, variant densities were also determined for genes falling within GRCh38 collapsed duplications (dups) and false dups. Mean variant densities per reference genome is displayed in parentheses. *p*-values were calculated using a Wilcoxon signed-rank test with continuity correction.

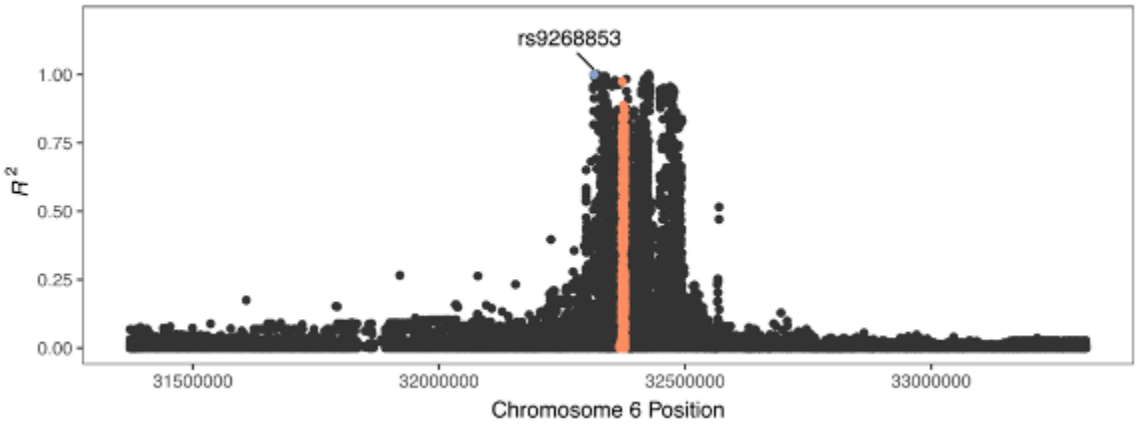


Figure S3.47. An example of pairwise LD between a known GWAS hit and variants in a non-syntenic region of T2T-CHM13. SNP rs9268853 (blue) is associated with fulminant type 1 diabetes in East Asian populations and segregates in strong LD ($R^2 > 0.8$) with 20 variants (orange) that were hidden to previous studies due to an insertion-deletion polymorphism that distinguishes GRCh38 from CHM13. While this example does not reflect an error or omission in GRCh38, it highlights the potential phenotypic and clinical relevance of such previously unresolved SNPs.

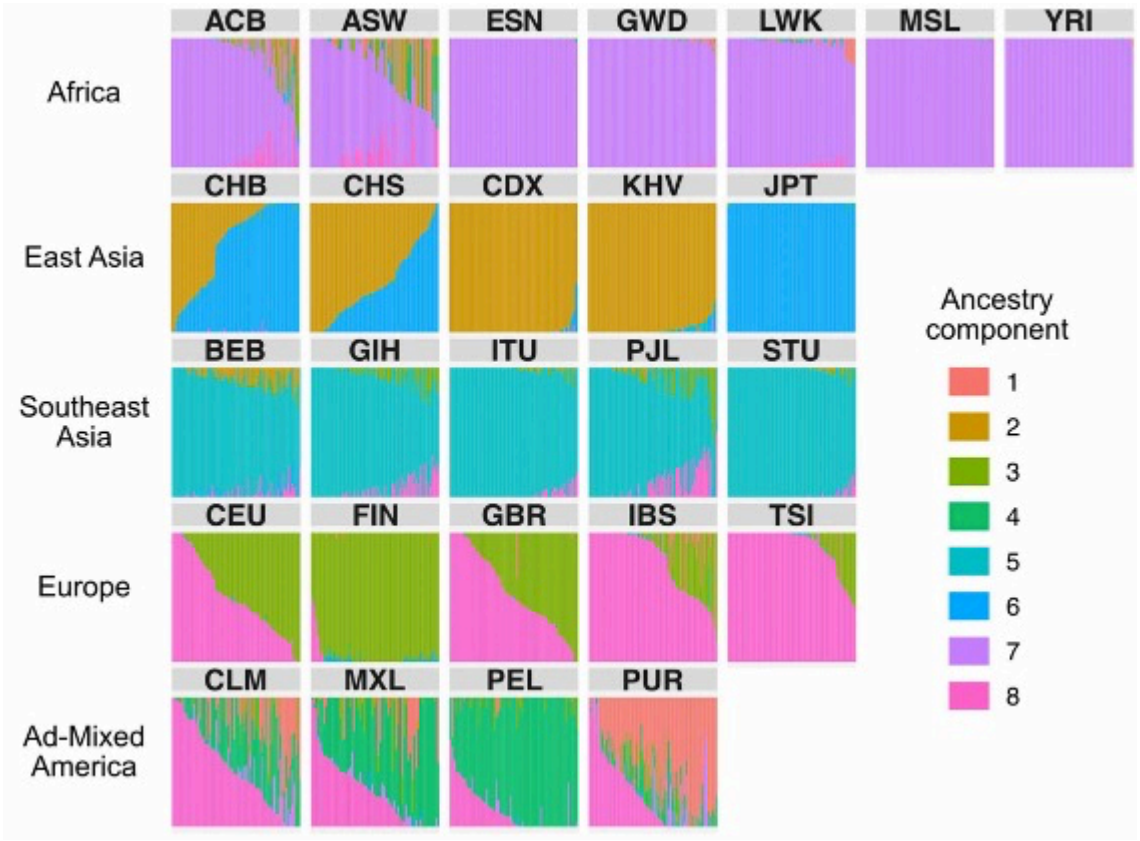


Figure S3.48. Inferred genome-wide ancestry of 1KGP individuals, used to identify frequency-differentiated SNPs. Admixture proportions (k = 8) for all samples in the 1KGP dataset, inferred by Ohana. Vertical bars represent individual genomes and are grouped by population. Ohana models each individual as a combination of k ancestry components and then searches for SNPs with evidence of frequency differentiation on these component lineages (CHB:Han Chinese, JPT:Japanese, CHS:Southern Han Chinese, CDX:Dai Chinese, KHV:Kinh Vietnamese, CHD:Denver Chinese, CEU:CEPH, TSI:Tuscan, GBR:British, FIN:Finnish, IBS:Spanish, YRI:Yoruba, LWK:Luhya, GWD:Gambian, MSL:Mende, ESN:Esan, ASW:African-American SW, ACB:African-Caribbean, MXL:Mexican-American, PUR:Puerto Rican, CLM:Colombian, PEL:Peruvian, GIH:Gujarati, PJL:Punjabi, BEB:Bengali, STU:Sri Lankan, ITU:Indian).

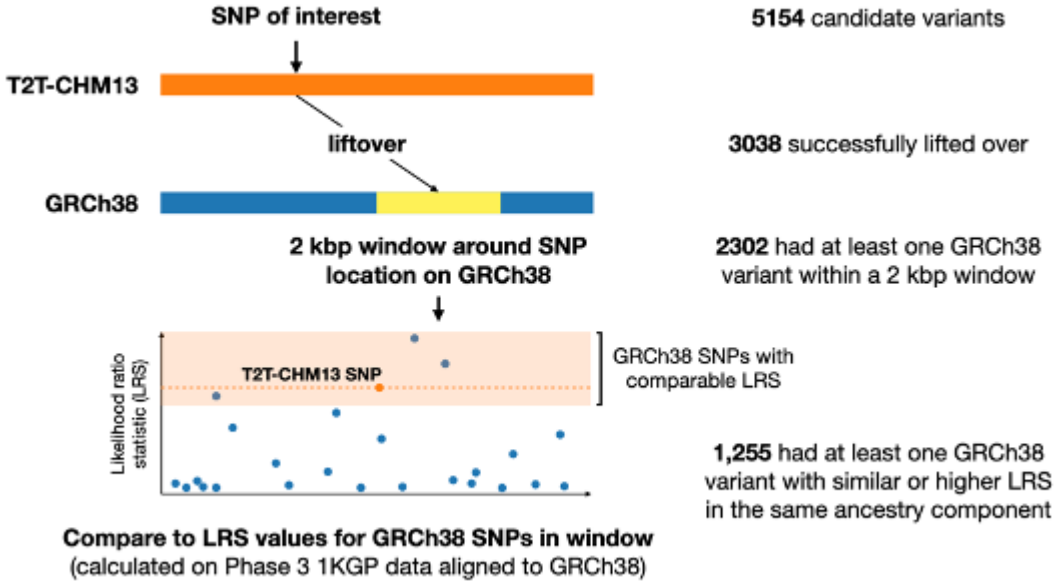


Figure S3.49. Comparing frequency-differentiated variants between T2T-CHM13 and GRCh38. Schematic for comparing likelihood ratio statistic (LRS) values between variants called on T2T-CHM13 and variants called from the 1KGP Phase 3 data aligned to GRCh38. The 5,154 most highly frequency-differentiated variants across ancestry components were lifted over from T2T-CHM13 to GRCh38. We selected all GRCh38 variants within a 2 kbp window of the lifted over position. We considered the LRS score of a GRCh38 variant comparable to that of the T2T-CHM13 SNP if it was within 10 of the T2T-CHM13 LRS value or greater.

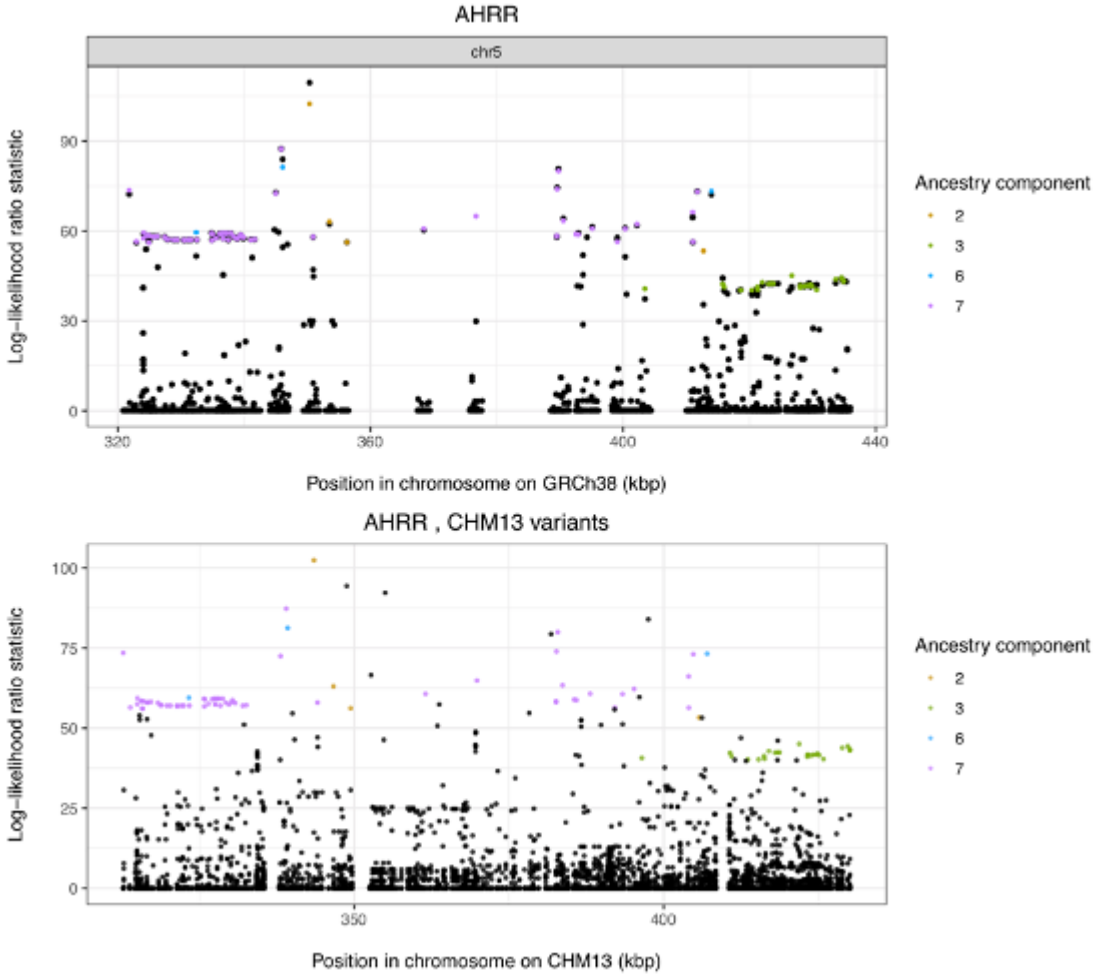


Figure S3.50. Example of a locus, overlapping the AHRR gene, with similar frequency differentiation results in T2T-CHM13 and GRCh38. Likelihood ratio statistics (LRS) for variants overlapping the AHRR gene on GRCh38 (top plot) and T2T-CHM13 (bottom plot). T2T-CHM13 variants with outlier LRS values, indicating strong allele frequency differences between ancestry components, are colored by ancestry. In the top plot, colored points indicate T2T-CHM13 variants lifted over to GRCh38 and black points indicate variants called on GRCh38. In the bottom plot, black points indicate non-outlier T2T-CHM13 variants.

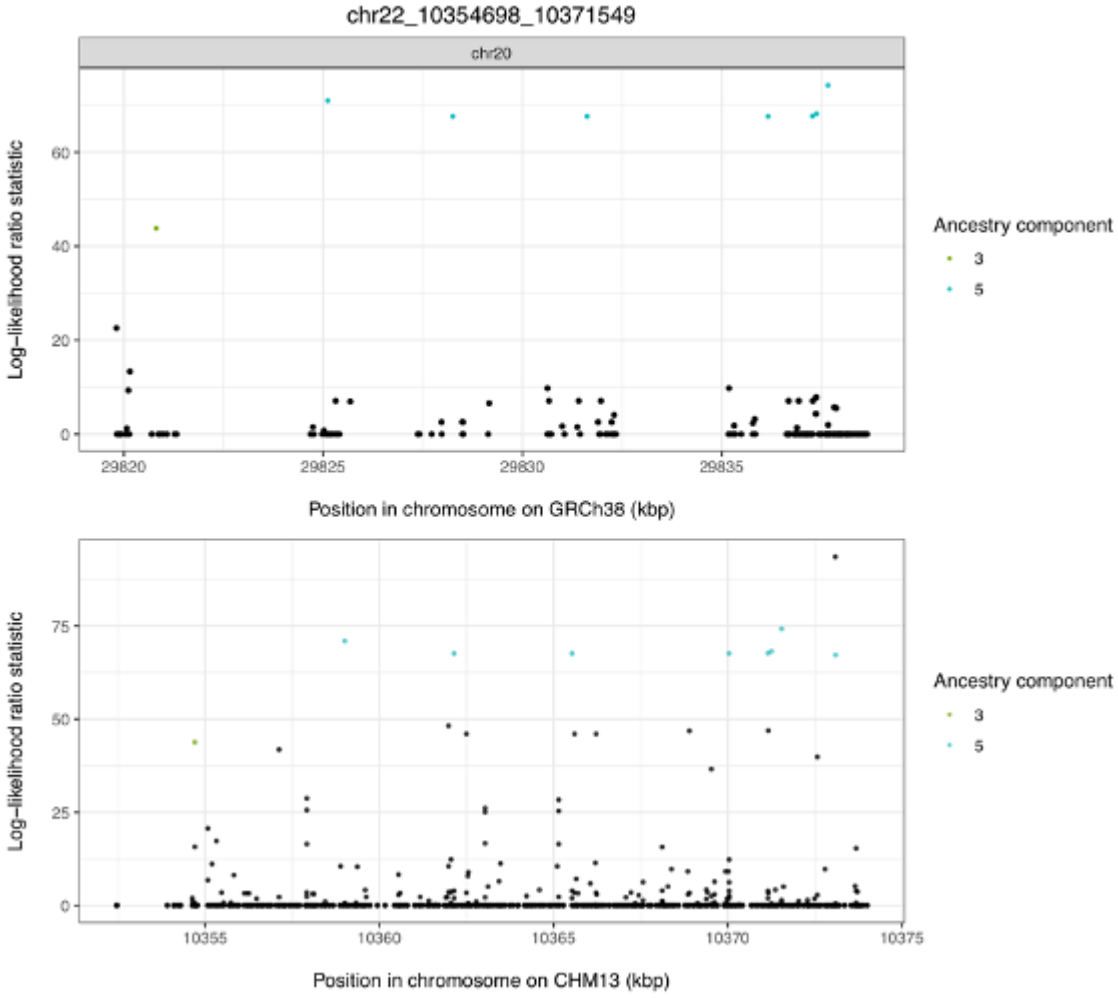


Figure S3.51. Example of a locus on chr22 with improved resolution in the T2T-CHM13 assembly. Likelihood ratio statistics (LRS) for variants in a region on chr22 of T2T-CHM13 (bottom plot) and GRCh38 (top plot), where it lifts over to chr20. T2T-CHM13 variants with outlier LRS values, indicating strong allele frequency differences between ancestry components, are colored by ancestry. In the top plot, colored points indicate T2T-CHM13 variants lifted over to GRCh38 and black points indicate variants called on GRCh38. In the bottom plot, black points indicate non-outlier T2T-CHM13 variants.

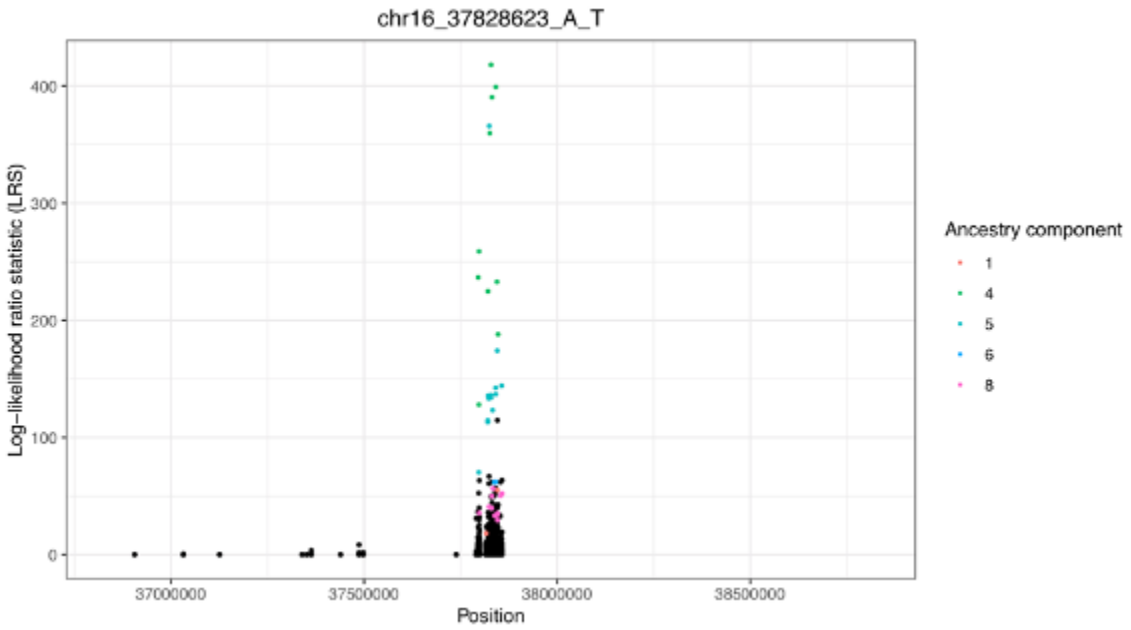


Figure S3.52. A frequency-differentiated, previously unresolved locus at chr16:37828623. Likelihood ratio statistics (LRS) for variants in a 2 Mbp window around chr16:37828623, a variant that reaches high allele frequencies in the Peruvian in Lima, Peru (PEL) population (ancestry component 4). T2T-CHM13 variants with outlier LRS values, indicating strong allele frequency differences between ancestry components, are colored by ancestry. Black points indicate non-outlier T2T-CHM13 variants.

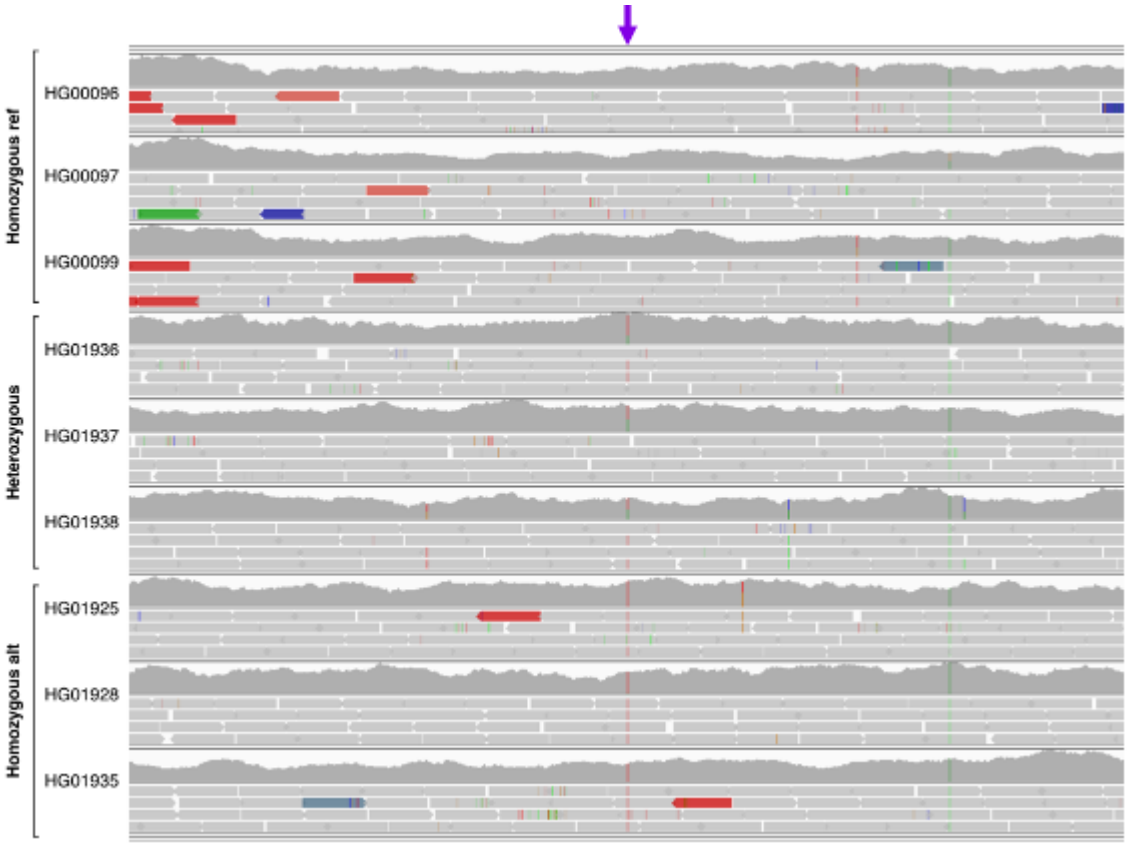


Figure S3.53. Alignments to the region around chr16:37828623. Alignments to the region around an A -> T SNV (arrow) at chr16:37828623 on T2T-CHM13, for three homozygous reference (HG00096, HG00097, HG00099; European ancestry), heterozygous (HG01936, HG01937, HG01938; Peruvian in Lima, Peru [PEL] ancestry), and homozygous alternate (HG01925, HG01928, HG01935; PEL ancestry) individuals.

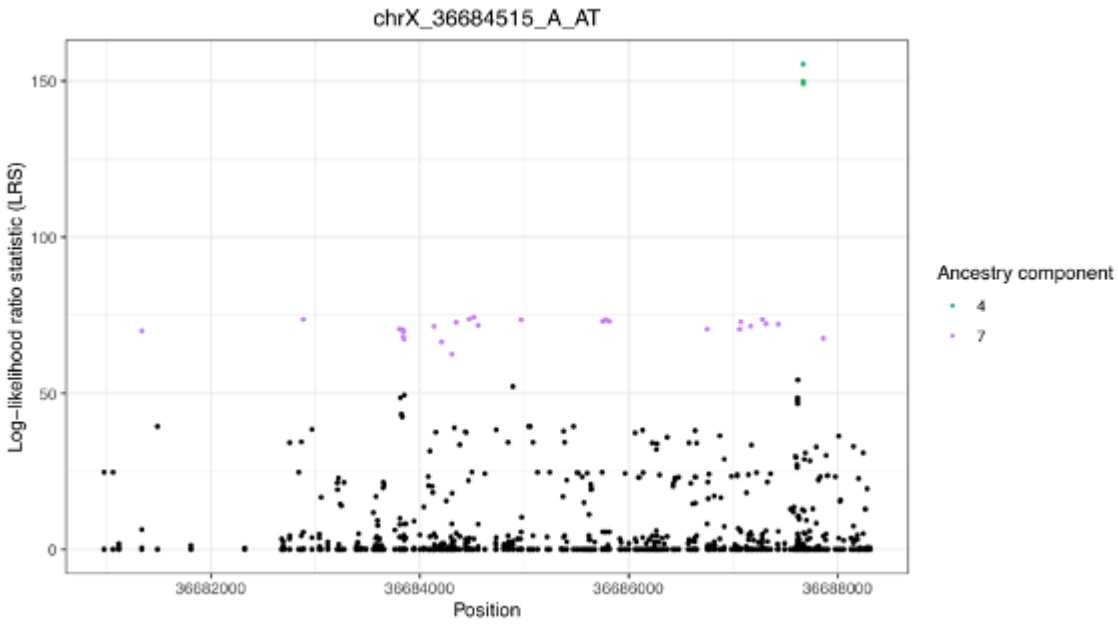


Figure S3.54. A frequency-differentiated, previously unresolved locus at chrX:36684515. Likelihood ratio statistics (LRS) for variants in a 2 Mbp window around chrX:36684515, a variant that reaches high allele frequencies in African populations (ancestry component 7). T2T-CHM13 variants with outlier LRS values, indicating strong allele frequency differences between ancestry components, are colored by ancestry. Black points indicate non-outlier T2T-CHM13 variants.

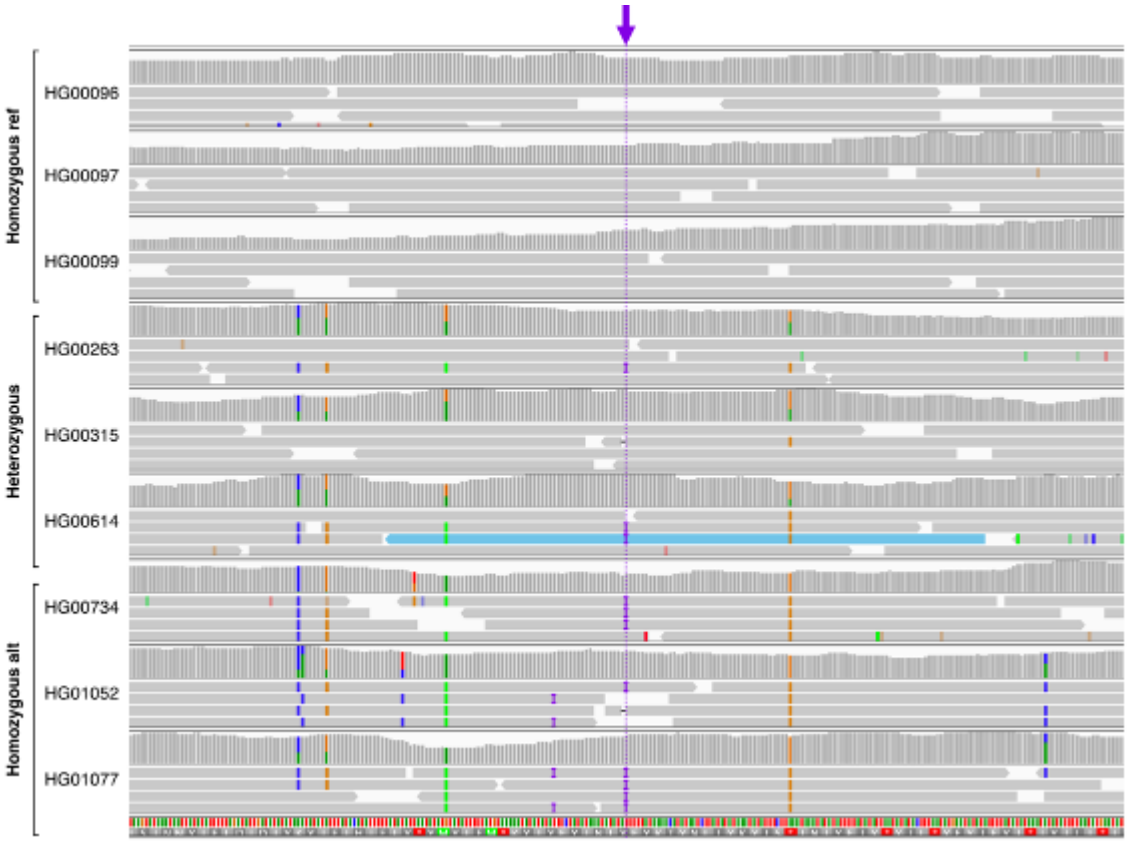


Figure S3.55. Alignments to the region around chrX:36684515. Alignments to the region around a 1bp insertion (arrow, dashed line) at chrX:36684515 on T2T-CHM13, for three homozygous reference (HG00096, HG00097, HG00099; European ancestry), heterozygous (HG00263, HG00315, HG00614; European and East Asian ancestry), and homozygous alternate (HG00734, HG01052, HG01077; African ancestry) individuals.

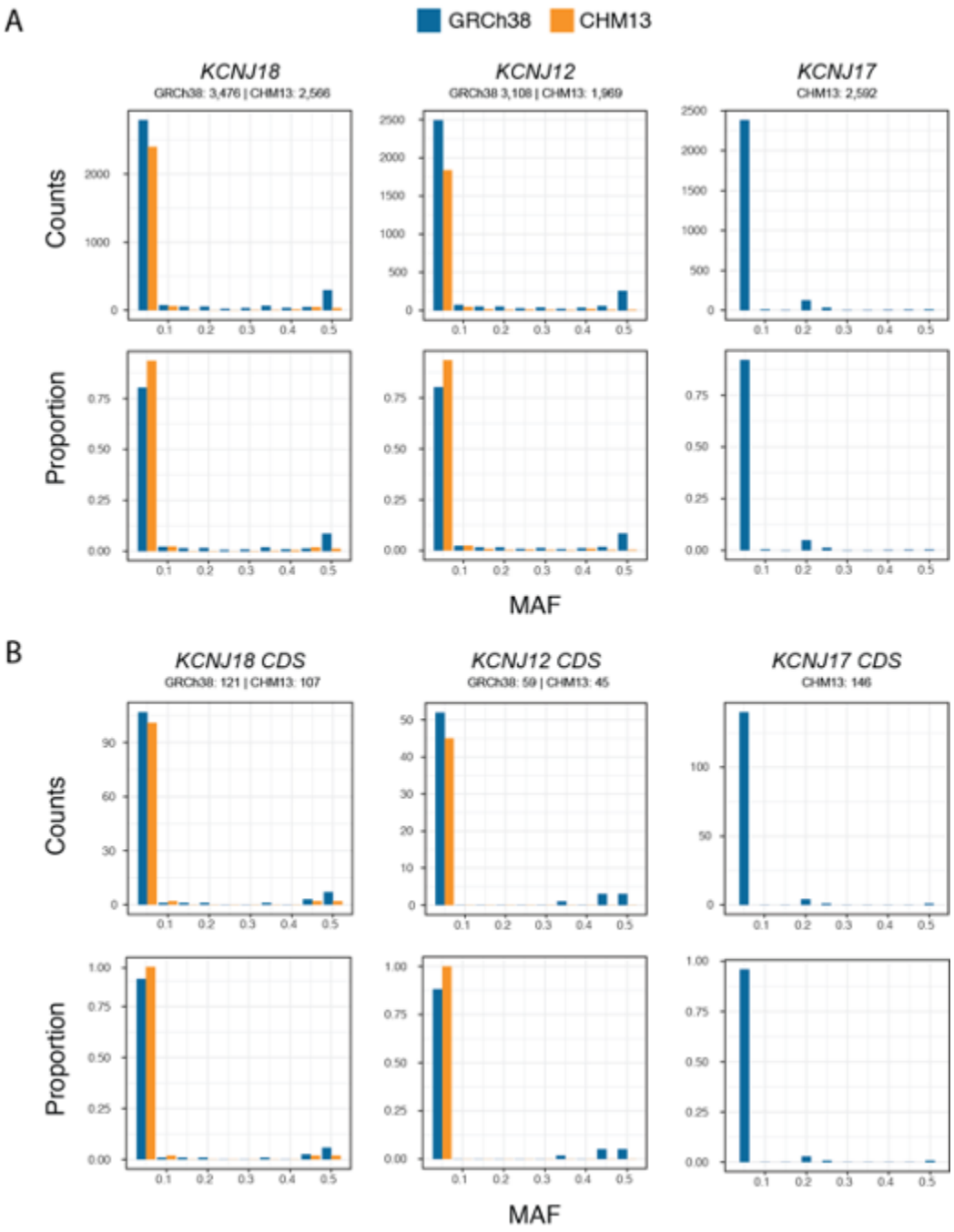


Figure S3.56. Folded site-frequency spectrum for *KCNJ18* paralogs in GRCh38 versus T2T-CHM13. Total counts and proportions are depicted for varied minor-allele frequencies (MAF) of variants discovered within entire SDs (A) and CDS (B) in GRCh38 (blue) and T2T-CHM13 (orange) from 1KGP datasets. Total numbers of bi-allelic SNVs detected at each locus is indicated beneath each gene header per reference genome.

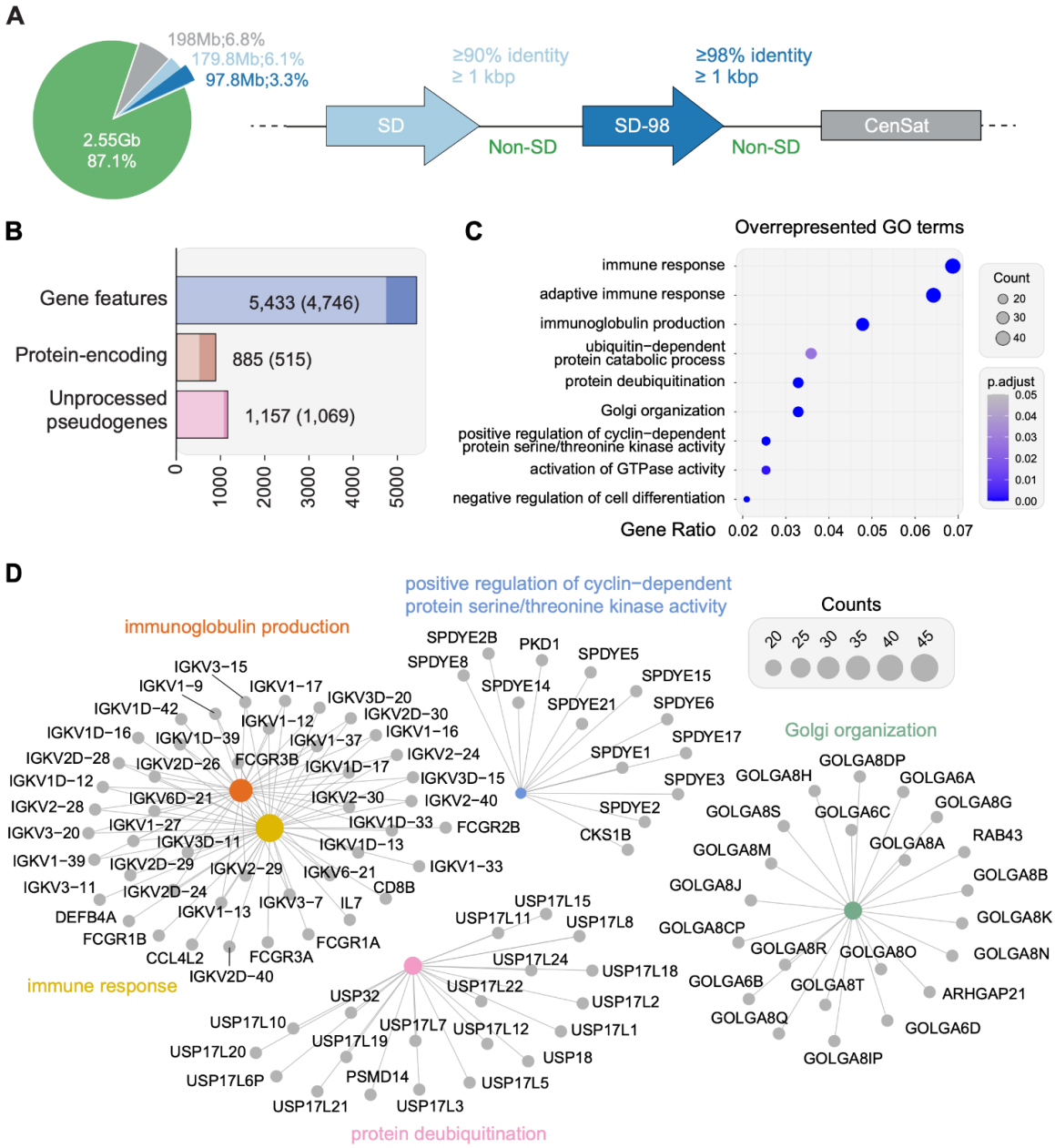


Figure S4.1. Nearly-identical human gene duplications in a complete human genome. (A) Size of autosomal segmental duplications (SD), SD with over 98% sequence identity (SD-98), CenSat (centromeric satellites excluding pericentromeric SD), and unique regions, which exclude SD and CenSat (Non-SD). (B) Gene features overlapping autosomal SD-98 regions. Numbers in parenthesis indicate gene features fully-contained in a SD-98 region. (C) Gene ontology (GO) enriched terms (FDR>0.1). (D) Gene concept network (cnetplot) of enriched GO terms of genes overlapping SD-98.

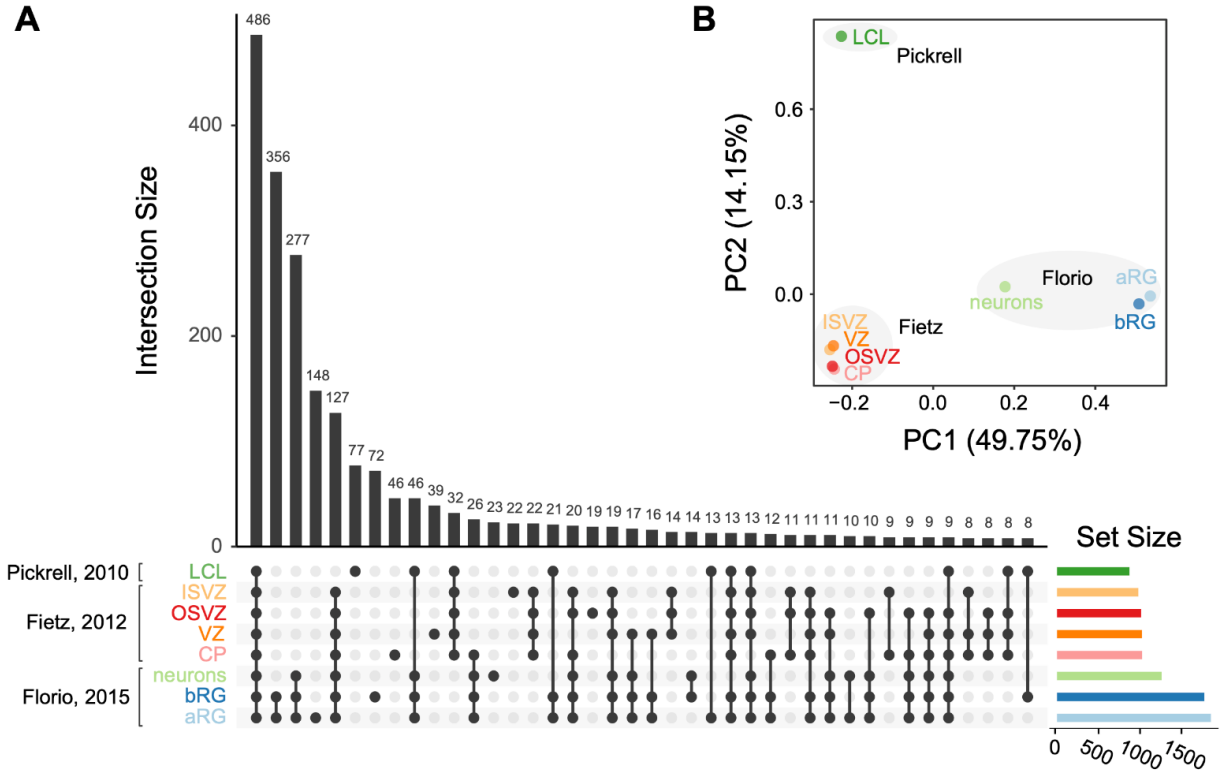


Figure S4.2. Expression patterns of SD-98 genes in LCL and human fetal brain samples remapped to T2T-CHM13 reference. (A) Expressed (TPM≥1) SD-98 genes shared between lymphoblastoid cell line (LCL), fetal brain tissues (inner subventricular zone [iSVZ], outer subventricular zone [OSVZ], ventricular zone [VZ], cortical plate [CP]), and fetal brain cell populations (apical radial glia [aRG]), basal radial glia [bRG], and neurons). (B) Principal component analysis of expressed and non-expressed SD-98 genes in LCL, fetal brain tissues, and fetal brain cell populations.

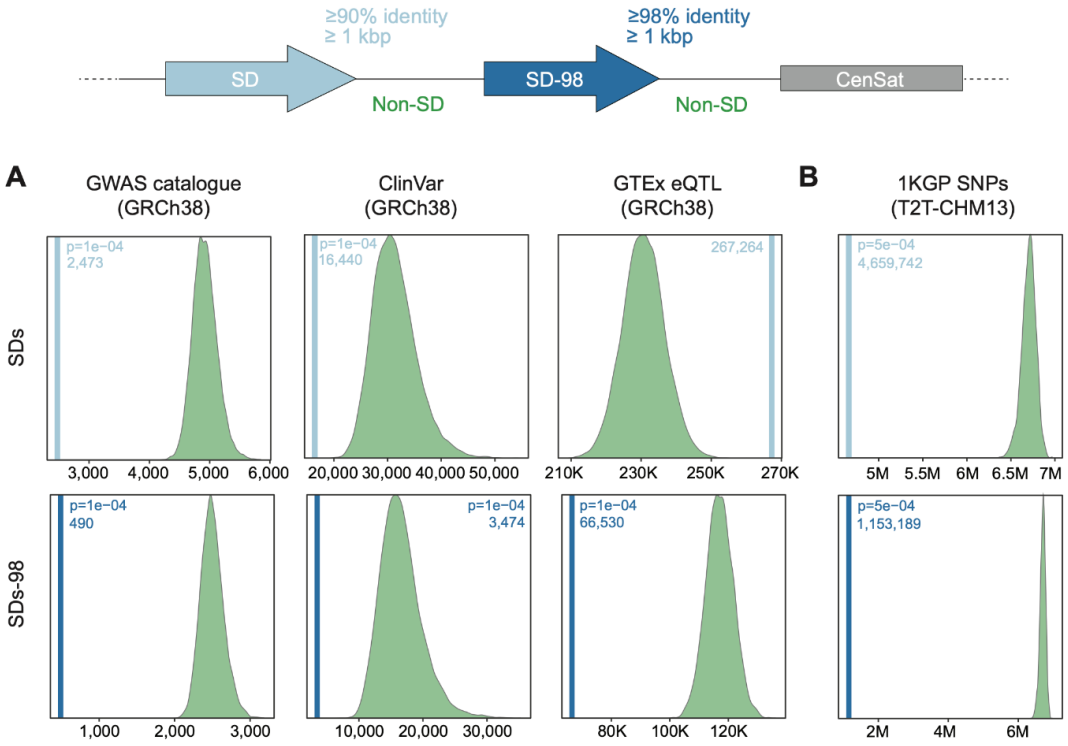


Figure S4.3. Variant depletion in SD and SD-98. (A) Depletion of variants in GWAS catalog, ClinVar, and GTEx eQTL databases and (B) depletion of biallelic SNVs in 1KGP (n=3,202) in T2T-CHM13 (v1.0). Vertical lines represent observed values and distributions correspond to the variants observed in 10,000 random permutations across the genome (excluding gaps and centromeric satellites), with empirical *p*-values within each plot.

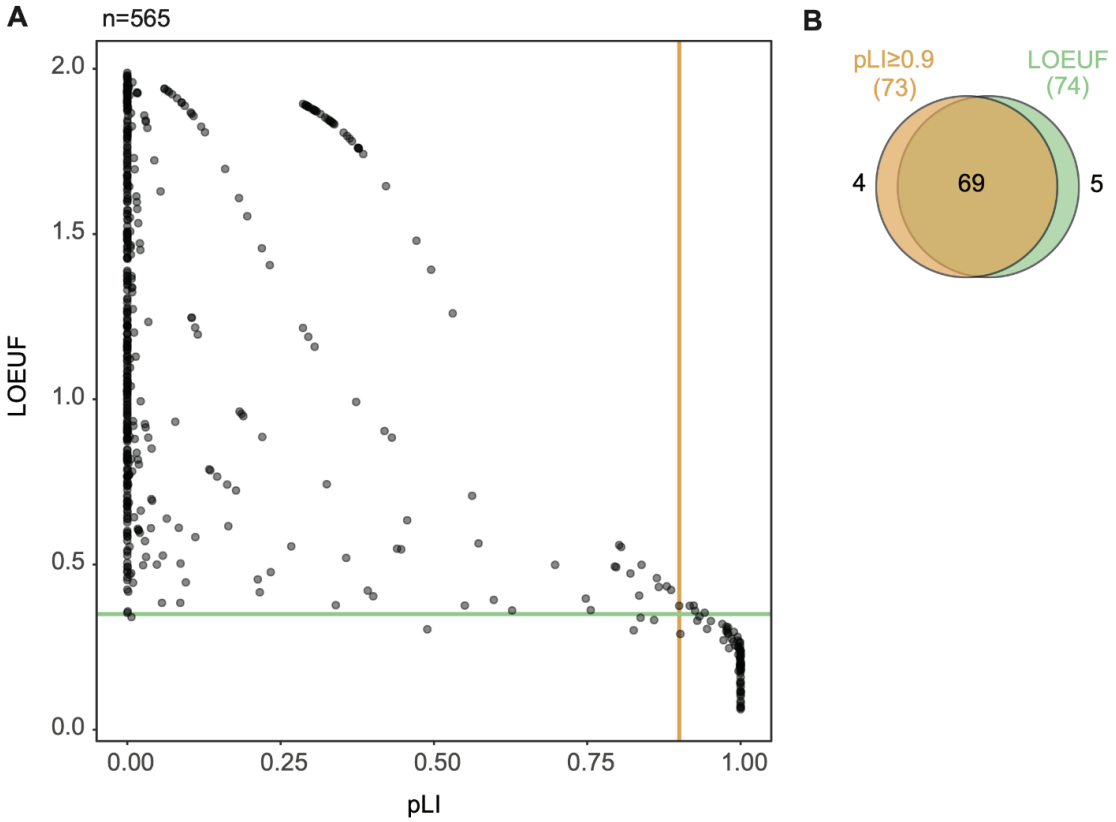


Figure S4.4. pLI and LOEUF scores of SD-98 genes. (A) Relationship between pLI and LOEUF values for 565 SD-98 genes with available scores. (B) Overlap loss of function intolerant genes as measured by pLI (≥0.9) and LOEUF (≤0.35) scores.

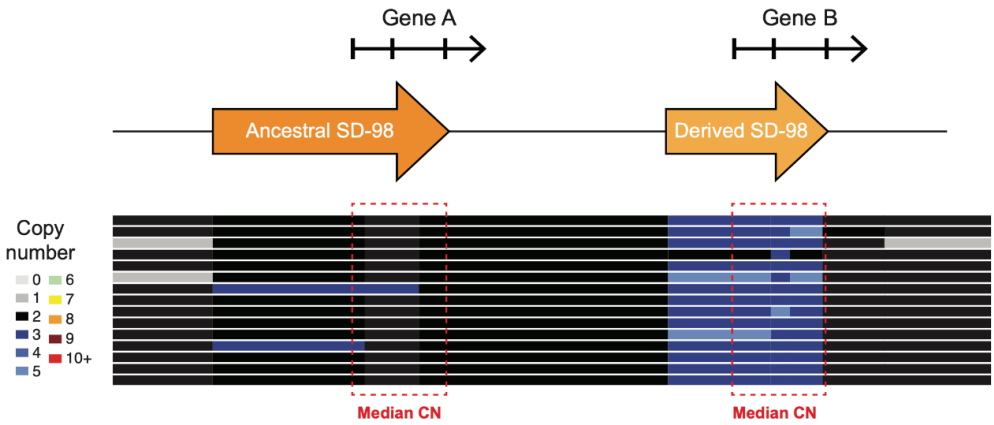


Figure S4.5. Diagram of copy-number genotyping approach. Heatmap represents CN estimates from QuicK-mer2 with values shown in legend. Each row represents a different individual. Red dashed lines indicate CN-genotyped SD-98 regions (including both ancestral and derived paralogs) used for CN analyses.

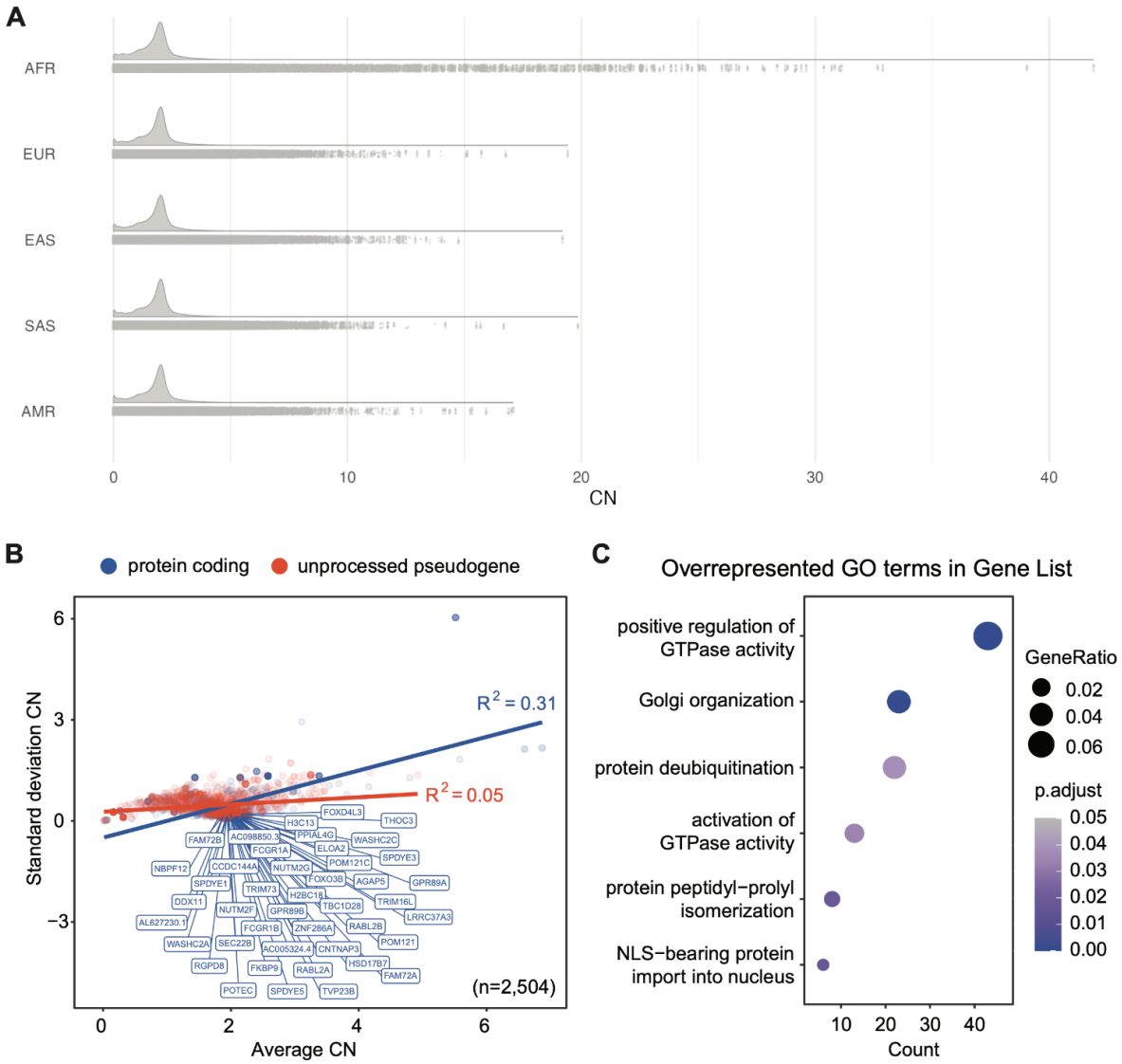


Figure S4.6. Analysis of CN estimates of SD-98 genes. (A) Copy-number distribution of SD-98 regions genotyped by superpopulation. (B) Relationship between CN mean average and standard deviation per genotyped SD-98 region. Colors indicated protein-encoding genes or unprocessed pseudogenes as classified by ENSEMBL. Linear regression and R-squared coefficient are indicated per each gene biotype. (C) Gene ontology terms overrepresented in CN fixed SD-98 genes (CN = 2 in ≥98% of individuals).

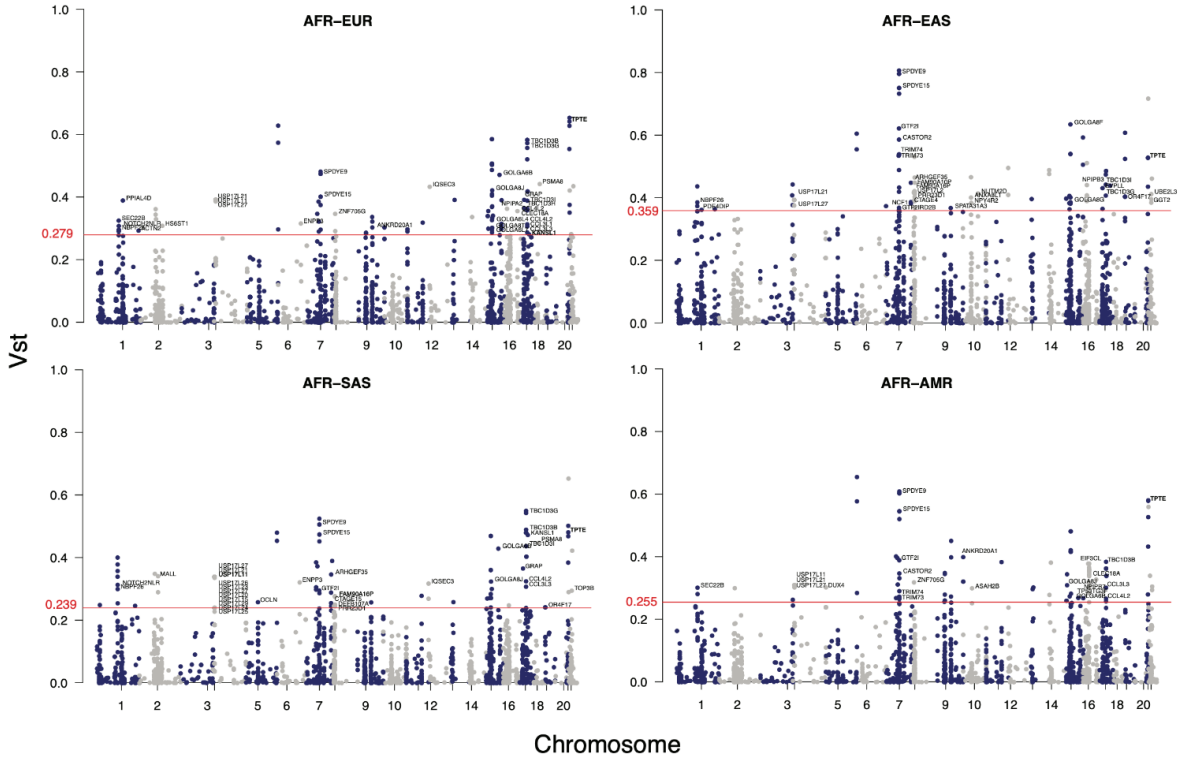


Figure S4.7. Distribution of V_{ST} values. Pairwise comparisons between Africans (AFR) and Europeans (EUR), East Asians (EAS), and Americans (AMR) are shown. Red line indicates the 95th percentile per pairwise comparison. Gene name labels correspond to protein-encoding genes above the 95th percentile threshold.

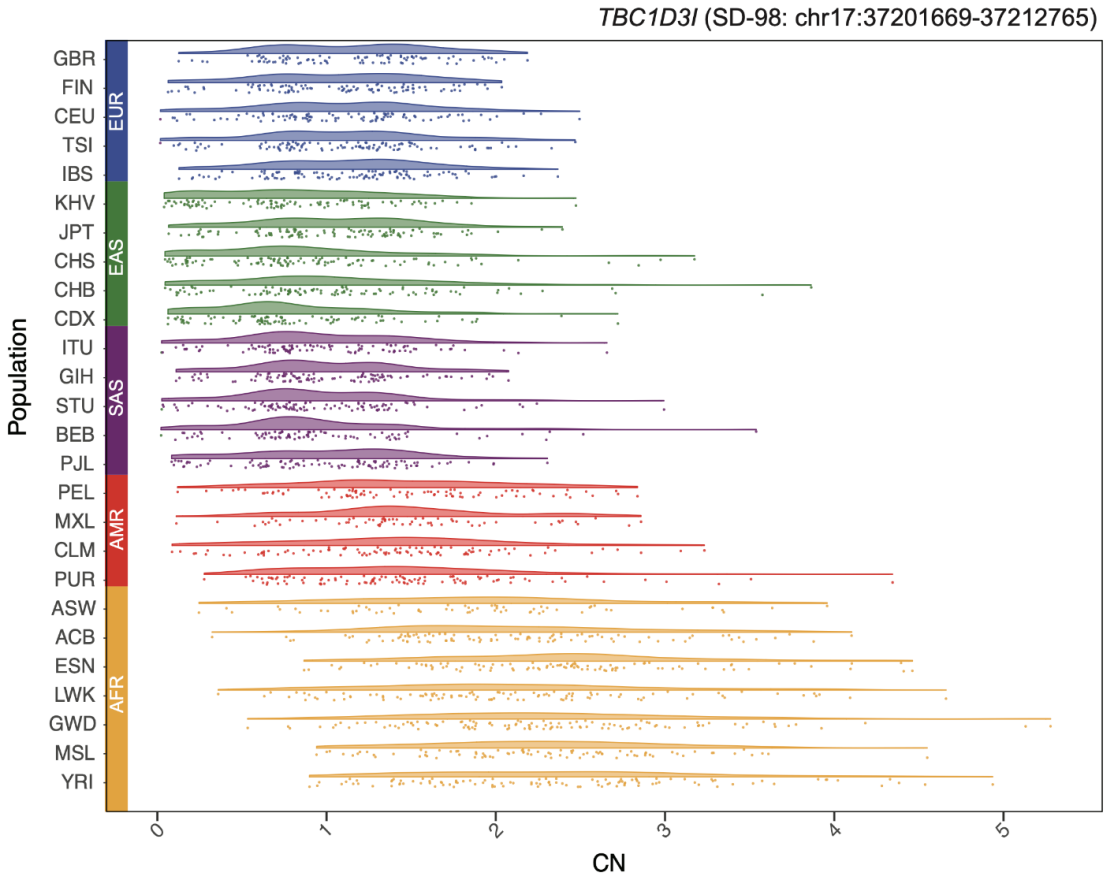


Figure S4.8. Copy-number dotplot of *TBC1D3I* gene. Copy-number estimates correspond to genotyped SD-98 region overlapping *TBC1D3I* gene. 1KGP populations are colored by superpopulations as indicated in figure legend.

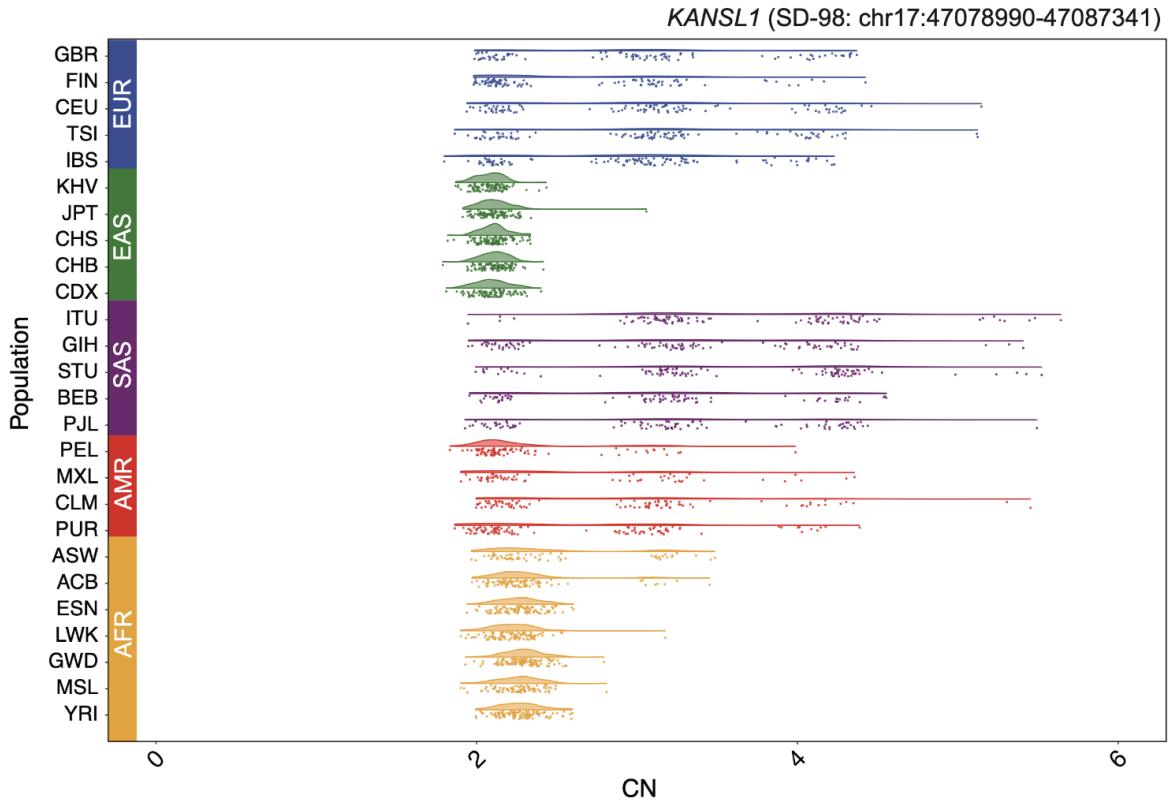


Figure S4.9. Copy-number dotplot of *KANSL1* gene. Copy-number estimates correspond to genotyped SD-98 region overlapping *KANSL1* gene. 1KGP populations are colored by superpopulations as indicated in figure legend.

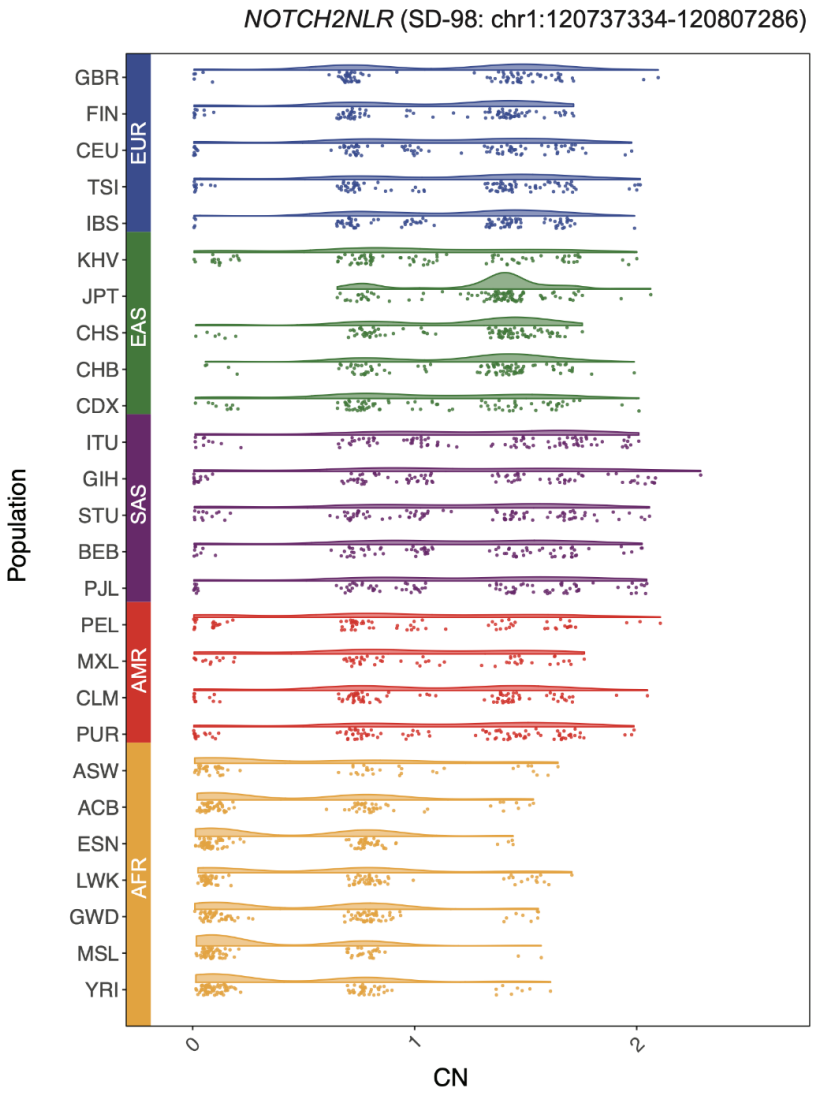


Figure S4.10. Copy-number dotplot of *NOTCH2NLR* gene. Copy-number estimates correspond to genotyped SD-98 region overlapping *NOTCH2NLR* gene. 1KGP populations are colored by superpopulations as indicated in figure legend.

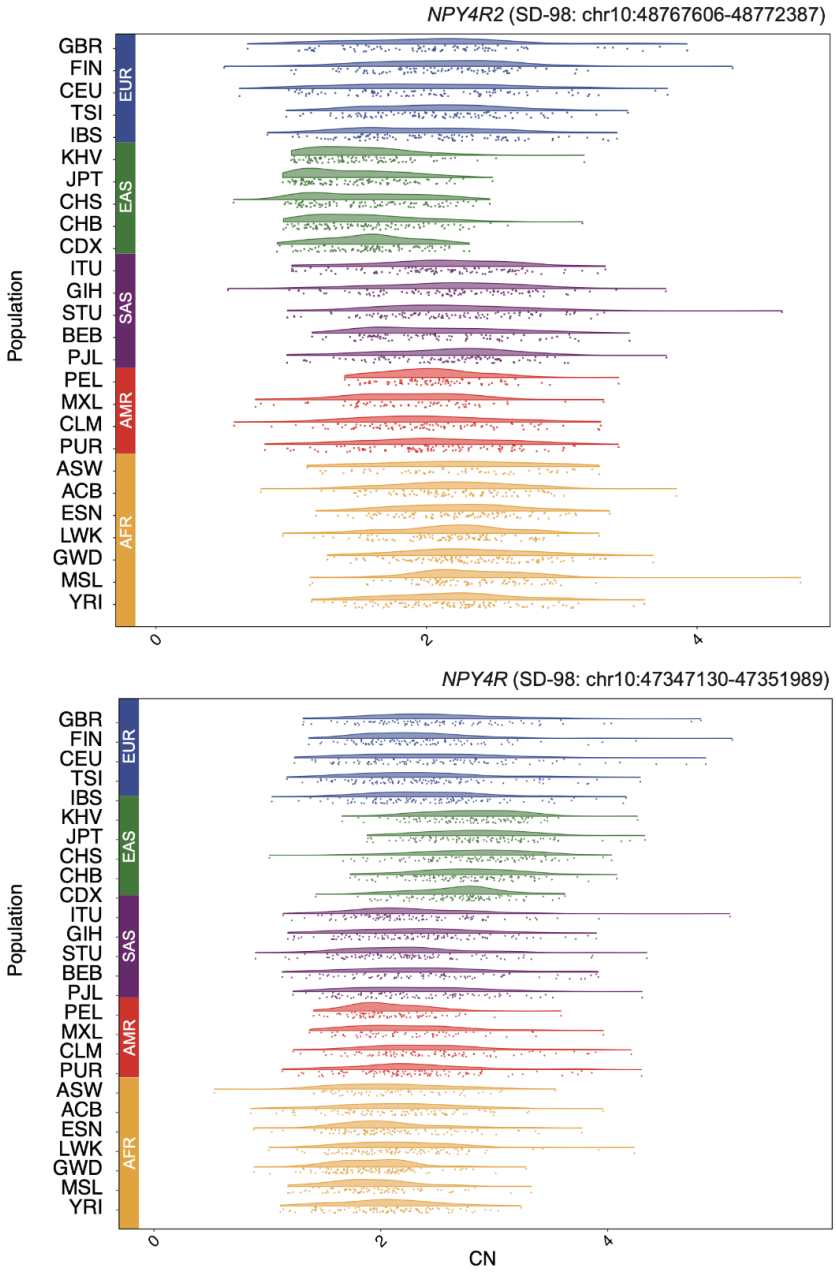


Figure S.4.11. Copy-number dotplot of *NPY4R and NPY4R2* genes. Copy-number estimates correspond to genotyped SD-98 region overlapping *NPY4R* and *NPY4R2* genes. 1KGP populations are colored by superpopulations as indicated in figure legend.

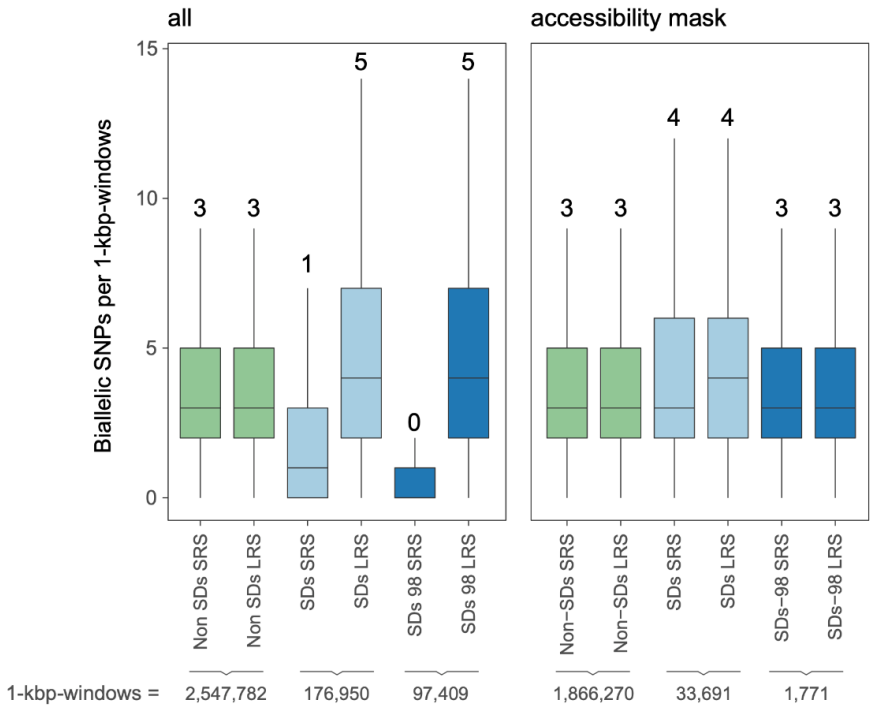


Figure S4.12. Comparison of short- and long-read SNV density in Non-SDs, SDs, and SD-98. Biallelic SNVs discovered with Illumina short-read sequencing (SRS) and PacBio HiFi long-read sequencing (LRS) in the same eight individuals were used.

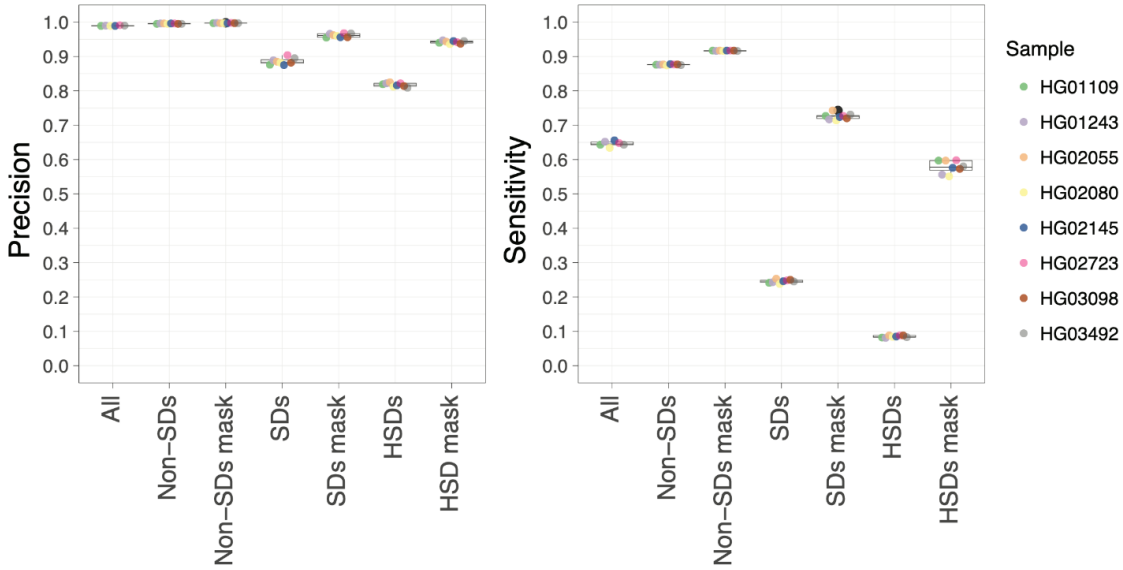


Figure S4.13. Short- and long-read variants concordance. Variants discovered with Illumina short-reads were benchmarked against PacBio HiFi variants in terms of precision sensitivity for eight individuals indicated in the legend sequenced with both platforms.

Supplemental Tables

Supplementary tables for chapters two, three and four are included as a zipped file.