Assembly of segmental duplications and their variation in humans

Mitchell R. Vollger

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2021

Reading Committee:

Evan E. Eichler, Chair

Sreeram Kannan

Phil Green

Program Authorized to Offer Degree:

Genome Sciences

University of Washington

Abstract

Assembly of segmental duplications and their variation in humans

Mitchell R. Vollger

Chair of the Supervisory Committee:
Evan E. Eichler
Genome Sciences

Despite their importance in disease and evolution, highly identical segmental duplications (SDs) have been among the last regions of the human reference genome to be finished. In this work I develop new methods, apply new sequencing technologies, and characterize the complete extent of SDs in a human genome. To assemble previously unresolved SDs in long-read assemblies, I develop a new method called Segmental Duplication Assembler (SDA) that constructs graphs using paralogous sequence variants. I demonstrate the utility of this method by applying it to three human genomes where I recover 33-79 Mbp of new sequence. Furthermore, the sequence produced by this method is highly accurate (>99.9%) so it can be applied to other complex genomes to resolve the last gene-rich gaps at the base-pair resolution. Using a new highly accurate long-read sequencing technology, I compared the accuracy, continuity, and gene

annotation of human genome assemblies generated from high-fidelity (HiFi) and continuous long-read (CLR) datasets. I find that the HiFi sequence data assemble an additional 10% of duplicated regions while incurring just one tenth of the sequencing cost and computational resources of the CLR assembly. Finally, using HiFi sequencing in combination with ultra-long sequencing from Oxford Nanopore, I contribute to a telomere-to-telomere (T2T) assembly of a human genome. I then use this genome to gain the first comprehensive view of human SD organization where I find that nearly one third of the additional sequence in T2T CHM13 corresponds to SDs, increasing the genome-wide estimate from 5.4% to 7.0% (219 Mbp). I identify novel regions of genomic instability, locate methylation differences between SD clusters, and predict 182 new gene models. I find that 91% of the new T2T SD sequence (68.3 Mbp) better represents human copy number based on analysis of 266 human genomes. Comparing long-read assemblies from other human (n=12) and nonhuman primate (n=6) genomes, I use T2T CHM13 to systematically reconstruct the evolution and structural diversity of biomedically relevant and duplicate genes important in the expansion of the human frontal cortex. The analysis reveals unprecedented patterns of structural heterozygosity and massive evolutionary differences between man and his closest living relatives.

TABLE OF CONTENTS

List of Figures vii

List of Tables viii

Chapter 1. Introduction 1

1.1 The impact of segmental duplications in human genetics 2

1.2 The first maps of segmental duplications in the human genome 3

1.3 The assembly problem 4

1.3.1 Assembly by hierarchical and whole-genome shotgun sequencing 4

1.3.2 Advances in sequencing technologies and assembly 6

1.4 Goals of this dissertation 8

1.5 Topics in this dissertation 9

1.5.1 Assembling highly identical segmental duplications from low accuracy long reads 9

1.5.2 Improvements in assembly from new sequencing technologies 10

1.5.3 The complete scope of segmental duplications in a human genome 10

Chapter 2. Long-read sequence and assembly of segmental duplications 12

2.1 Abstract 12

2.2 Introduction 13

2.3 Results 14

2.3.1 The problem: Unresolved SDs 14

2.3.2 The approach: Segmental Duplication Assembler (SDA) 15

2.3.3 Resolving SDs using SDA 17

2.3.4 Characterization of diverged duplications 19
2.4 Discussion 21
2.5 Methods 23
2.5.1 Human genome assemblies. 23
2.5.2 SD characterization. 23
2.5.3 Assembly collapse and PSV definition. 24
2.5.4 PSV graph construction. 25
2.5.5 Correlation clustering. 25
2.5.6 PSV read partition and assembly. 26
2.5.7 BAC clone insert sequencing. 26
2.5.8 Data availability 28
2.5.9 Code availability 29
2.6 Figures 30
2.7 Tables 35
2.8 Acknowledgements 35
2.9 Author contributions 35
Chapter 3. Improved assembly and variant detection of a haploid human genome using single-molecule, high-fidelity long reads 37
3.1 Abstract 37
3.2 Introduction 38
3.3 Results 39
3.3.1 Whole-genome assembly with HiFi versus CLR reads 39
3.3.2 Segmental duplication analyses 42

3.3.3 Tandem repeat resolution 44
3.3.4 Structural variant analyses 46
3.3.5 Gene open reading frame annotations 46
3.4 Discussion 48
3.5 Methods 51
3.5.1 Cell lines 51
3.5.2 CCS library preparation 51
3.5.3 Strand-seq library preparation 52
3.5.4 BAC clone insert sequencing 52
3.5.5 Genome assembly 53
3.5.6 QV calculations 53
3.5.7 SD analyses 54
3.5.8 Pericentromeric analyses 54
3.5.9 Tandem repeat analyses 55
3.5.10 SV analyses 55
3.5.11 Gene annotation 56
3.5.12 RepeatMasker analysis of unmappable sequences 57
3.5.13 Data Access 57
3.6 Figures 58
3.7 Tables 62
3.8 Acknowledgments 66
3.9 Author Contributions 66
Chapter 4. Segmental duplications and their variation in a complete human genome 67

4.1 Abstract 67
4.2 Introduction 68
4.3 Results 69
4.3.1 SD content and organization. 69
4.3.2 Validation and heteromorphic variation. 71
4.3.3 Single-nucleotide and copy number variation within SDs. 73
4.3.4 Structural variation and massive evolutionary changes in the human lineage. 75
4.3.5 New gene models and variable duplicate genes 77
4.3.6 SD methylation and transcription. 78
4.4 Discussion 80
4.5 Methods 84
4.5.1 Estimating the number of rDNA copies in the assembly 84
4.5.2 Estimating amount of missing rDNA sequence in the assembly. 84
4.5.3 Repeat Masking. 85
4.5.4 SD Annotation. 86
4.5.5 Defining syntenic regions between T2T CHM13 and GRCh38. 86
4.5.6 Calculating the number of SD alignments in 5 Mbp windows. 87
4.5.7 WSSD detection and genotyping. 87
4.5.8 Gene annotations with Liftoff. 88
4.5.9 Counting the number of high identity SD genes 89
4.5.10 Cell culture. 89
4.5.11 FISH characterization and validation 89
4.5.12 ONT validation. 90

4.5.13 TBC1D3 phylogenetic tree construction. 90
4.5.14 Defining structurally variable haplotypes. 91
4.5.15 Variation graphs for SD loci. 91
4.5.16 Methylation analysis. 92
4.5.17 Custom ideogram and homology visualizations. 92
4.5.18 Data Availability 93
4.6 Figures 94
4.7 Tables 100
4.8 Acknowledgments 100
4.9 Author contributions 100
Chapter 5. Discussion and future directions 101
5.1 Impacts of the presented work 101
5.1.1 The resolution and consequences of collapsed SDs in long-read assemblies 101
5.1.2 Advances in sequencing technology and routine human genome assembly 102
5.1.3 Sequence-resolved variation in SDs for nearly complete human genomes 102
5.2 Looking forward 103
5.2.1 Diploid telomere to telomere assemblies 103
5.2.2 Encoding evolutionary history and variation in SDs 105
5.2.3 Functional annotation of duplications 109
5.3 Addressing the aims of this dissertation 110
5.4 A final remark 111
Bibliography 112

Appendix A. Supplement for chapter 2 133

Appendix B. Supplement for chapter 3..... 165

Appendix C. Supplement for chapter 4..... 180

LIST OF FIGURES

Figure 1.1. Ambiguity in assembly due to repetitive genomic sequences. 5
Figure 1.2. Collapse of paralogous sequences from multiple segmental duplications. 7
Figure 2.1. Flowchart of Segmental Duplication Assembler (SDA) method. 30
Figure 2.2. SDA results of the CHM1 human genome assembly. 31
Figure 2.3. Sequence and assembly of *SRGAP2* loci in the CHM13 human genome. 32
Figure 2.4. Correspondence between SDA sequence-diverged contigs and BACs. 33
Figure 2.5. Gene discovery in previously unresolved duplications. 34
Figure 3.1. Comparison between the CHM13 HiFi and CLR genome assemblies. 58
Figure 3.2. Segmental duplication resolution in the HiFi and CLR genome assemblies. 59
Figure 3.3. Tandem repeat resolution in the HiFi and CLR genome assemblies. 60
Figure 4.1. Segmental duplication (SD) content of the T2T CHM13 genome. 94
Figure 4.2. Validation of novel SDs in T2T CHM13 and heteromorphic variation. 95
Figure 4.3. Segmental duplication sequence and copy-number variation. 96
Figure 4.4. Human-specific expansion of *TBC1D3* compared to nonhuman primates. ... 97
Figure 4.5. Genic insights in new SD regions of T2T CHM13. 98
Figure 4.6. SD methylation and gene transcription. 99
Figure 5.1. The progression of human genome assembly (credit G. A. Logsdon) 104
Figure 5.2. Variation in *CYP2D6* across human haplotypes. 106

LIST OF TABLES

Table 2.1. SDA assembly statistics. 35
Table 3.2. Statistics of the HiFi and CLR genome assemblies 62
Table 3.3. Summary of SV calls in the HiFi and CLR assemblies 63
Table 3.4. Summary of indels in the HiFi and CLR assemblies 64
Table 3.5. Summary of disrupted RefSeq gene models in the HiFi and CLR assemblies 65
Table 4.6. Segmental duplication summary statistics. 100
Table 5.7. Summary of HPRC samples sequenced with HiFi and ultra-long ONT 108

ACKNOWLEDGEMENTS

Graduate school has simply been the happiest time of my life. Despite the moments that challenged my resolve and made me question the choice to pursue a PhD, there is no doubt in my mind that I am ending this journey as a better version of myself. Unsurprisingly, my personal and academic success can be contributed to all those who have helped me, and I sincerely thank all of you.

I am deeply indebted to my family—Julia, Kyra, Leanne, Jitka, and Helmuth—for the support you have always given me, which has allowed me to freely chase my dreams without hesitation or fear of failure. I miss you all constantly and I cannot wait for our next weekly phone call. I am forever grateful to my elementary school teacher, Shane Harmon: whether it was giving me a love for the ocean or taking it upon yourself to come to my home and teach me when I was sick from chemotherapy, you gave me a passion for life that I can never repay. To my undergraduate mentor, Alison Gammie: it is not is possible to overstate how important your role has been in my development as a scientist. Your patience and enthusiasm as a mentor helped me realize that my struggles as a student did not disqualify me from becoming a scientist.

My accomplishments during graduate school would mean very little without someone to share them with, so I sincerely thank my cohort in Genome Sciences for always being there to celebrate the little victories, taking photos in the general's hat after finishing our general exams, or going out to drinks after publishing a paper. To paraphrase a very knowledgeable hobbit: "Alas, four and some years is far too short a time to live among such excellent and admirable doctoral

candidates. I don't know half of your research half as well as I should like, and I like less than half of your journal club presentations half as well as they deserve." But I have been extremely fortunate to have spent so much time with such excellent people.

I am grateful to my thesis committee Phil Green, Sreeram Kannan, William Noble, and Daniela Witten for their time, patience, and advice throughout my PhD, but mostly for the opportunity to share my work with an engaged and excited audience. I am grateful to Tonia Brown for her careful inspection and editing of all my posters, manuscripts, grant applications, and now thesis. I thank all my labmates for being excellent colleagues to work with—my success is in no small way possible because of our day-to-day collaborations and deliberations. In particular, I would like to thank Mark Chaisson, who left me with a fantastic project and a solid foundation on which to start my PhD. I owe a great debt to my fellow graduate students Philip Dishuck and Michelle Noyes for constant companionship and frequent beer hours.

Finally, I would like to thank my esteemed advisor Evan E. Eichler for your willingness to accept my shortcomings and celebrate my strengths. Time and time again time you believed in me and my work long before I believed in myself, and your support has given me the confidence to do things I never thought myself capable of. I cannot adequately express my gratitude to you for allowing me to learn about and contribute to this incredible field of science.

April Lo, you haven been my partner in all things and I simply cannot imagine doing any of this without you. I love you and thank you for everything that helped me get to this day.

Chapter 1. INTRODUCTION

The consequence of gene duplication and whole-genome duplication as forces of evolutionary structural change and adaptation have been appreciated even before the discovery of the genetic code (Bridges 1936; Muller 1936). In his seminal work, Susumu Ohno proposed that the force of purifying selection is strong enough that while adaptation can come about through neutral evolution, the emergence of novel function via this mechanism would be exceedingly rare. He instead proposed that the redundancy in mechanism as a result of gene duplication would alleviate the pressures of purifying selection and allow for novel function to come about (Ohno 1970). This process, now commonly referred to as neofunctionalization (Force et al. 1999), is one of three accepted fates for gene duplications. The most common fate is pseudogenization where through the accumulation of random mutations the gene loses its function (Lynch and Conery 2000). The final option available to duplications is subfunctionalization where gene copies assume different aspects of the original gene creating modularization of the functional components (Stoltzfus 1999; Force et al. 1999).

The impact of gene duplications in primates is of particular importance because while the primate lineage has seen a marked slowdown in the rates of chromosomal changes, single-nucleotide polymorphisms, and retrotransposon activity (Waterston et al. 2002; Wu and Li 1985; Steiper et al. 2004; Yunis and Prakash 1982; King and Wilson 1975), there has been a significant increase in duplication since the African great ape ancestor (Marques-Bonet et al. 2009). Even between human and chimpanzee, duplications account for twofold more base-pair differences than single-nucleotide changes (Cheng et al. 2005).

1.1 The impact of segmental duplications in human genetics

My research interests lie in a particular type of duplication referred to as segmental duplications (SDs), which are limited to the largest (>1 kbp) and most highly identical (>90%) sequences. The importance of these sequences in human genetics is disproportionate to the space they occupy in the genome, contributing more than expected to normal genetic variation, evolution, and disease.

SDs make up half of all normal copy number variation despite being just one twentieth of the human genome (Sudmant et al. 2010, 2015b), and this variation is often genic since SDs contain 6% of the exons in the human genome (Bailey et al. 2002). Furthermore, a genomic analysis of a diversity panel of 236 human genomes identified over 1,000 copy number variants (CNVs) stratified between human populations (Sudmant et al. 2015a). These loci are candidates for regions of the genome that may be associated with positive selection and adaptation.

One of example of this is found in a duplication shared between the modern-day Melanesian population and Denisovans. This is not an entirely unexpected event since an estimated 4-6% of the Melanesian genome is introgressed from Neanderthal and Denisovan (Reich et al. 2010). However, Hsieh et al. were able to find signals of a selective sweep in the flanking sequence near this CNV, suggesting a possible case of adaptive introgression. Using long-read technologies, this study sequenced resolved a nearly 400 kbp duplication and confirmed the introgression from Denisovan into the Melanesian population. Interestingly, the introgressed copy is located directly on the distal flank of a microdeletion event associated with the second most common cause of autism (Nuttle et al. 2016; Weiss et al. 2008) creating more homology between the two duplication blocks that mediate this microdeletion. Intuition says that this event would not be selected for; however, within the new 400 kbp of sequence there is an expressed novel gene, *NPIPB16,* with

signals of positive selection associated with amino acid substitutions (Hsieh et al. 2019). Previous work has shown that the gene family NPIP has been the subject of positive selection since the ape divergence and is still evolving adaptively (Johnson et al. 2001). This is not a one-off case of adaptation though copy number variation in segmental duplications, as multiple human-specific gene duplications have been shown to be important in human neurocognition and evolution (Fiddes et al. 2018; Suzuki et al. 2018; Ju et al. 2016; Dennis et al. 2012; Heide et al. 2020).

SDs also contribute to human disease though the homology between inter and, more commonly, intrachromosomal SDs, which mediate non-allelic homologous recombination events (Bailey et al. 2002; Kidd et al. 2010). These events can delete or duplicate megabases of sequence, resulting in abnormal gene dosages and giving rise to genomic disorders (Stankiewicz and Lupski 2002a; Sharp et al. 2006).

1.2 The first maps of segmental duplications in the human genome

In principle, the first draft assemblies of the human genome in 2001 should have allowed for *a priori* identification of all human segmental duplications (Venter et al. 2001; Lander et al. 2001; Bailey et al. 2001). However, estimates of SD content in the human genome based on these assemblies were inaccurate for two reasons: 1) The assembly from the public effort had false duplications stemming from the incorporation of multiple alleles into the assembly and misjoins between BACs without unique overlapping sequence. In fact, it was shown that ~80% of the intra-assembly alignments of over 98% identity were due to the incorporation of multiple alleles into the assembly (Bailey et al. 2001; Eichler 2001). 2) The private effort, and to a lesser extent the public effort, failed to assemble SDs because of the high sequence identity and the limitations in read technologies. It was not until 2002 when Bailey et al. used a combination of whole-genome sequencing (WGS) from the private effort and clone sequences from the public assembly that the

first accurate map of human SDs was laid out (Bailey et al. 2002). This initial annotation contained 8,595 regions representing 130.5 megabases of sequence. Furthermore, using the new SD map, the authors identified 169 regions likely to undergo unequal crossing over due to large highly identical duplications in close proximity. At the time of publishing, 24 of these regions were already associated with genomic disorders (Stankiewicz and Lupski 2002b).

However, this story was incomplete, and in fact remains incomplete 20 years after the first draft assemblies of the human genome. Certain segmental duplications, centromeres, and other highly duplicated heterochromatic sequences have all persisted as gaps in the human reference due to their size, complexity, and identity.

1.3 The assembly problem

1.3.1 *Assembly by hierarchical and whole-genome shotgun sequencing*

SDs are exceptionally difficult to resolve because they create ambiguity in assembly graphs, which then cannot be confidently traversed to recreate the underlying genome (Figure 1.1). The Human Genome Project tackled this issue in part by performing hierarchical assembly in which human sequence was made into bacterial artificial chromosomes (BACs), which could be independently assembled into ~175 kbp contigs before being combined to make larger stretches of human sequence; however, even this very laborious and expensive effort struggled in regions of tandem duplications and with certain sequences that were toxic to the *e. coli* (Lander et al. 2001).

The private effort lead by Celera took a different tact and sequenced the genome using whole-genome shotgun sequencing to generate 500-700 bp fragments of sequence at random from the human genome. For their initial assembly they were able to generate ~5-fold sequence coverage of the human genome, and after fragmenting BAC sequences from the public project into

faux sequencing reads they were able to obtain ~8-fold coverage of the human genome (Venter et al. 2001). These reads where then assembled using the Celera Assembler, an improved version of a previously developed algorithm for the assembly of the *Drosophila* genome (Myers et al. 2000). However, this effort was only able to construct "ancient" SDs, leaving the more recent and highly identical SDs missing from the assembly.

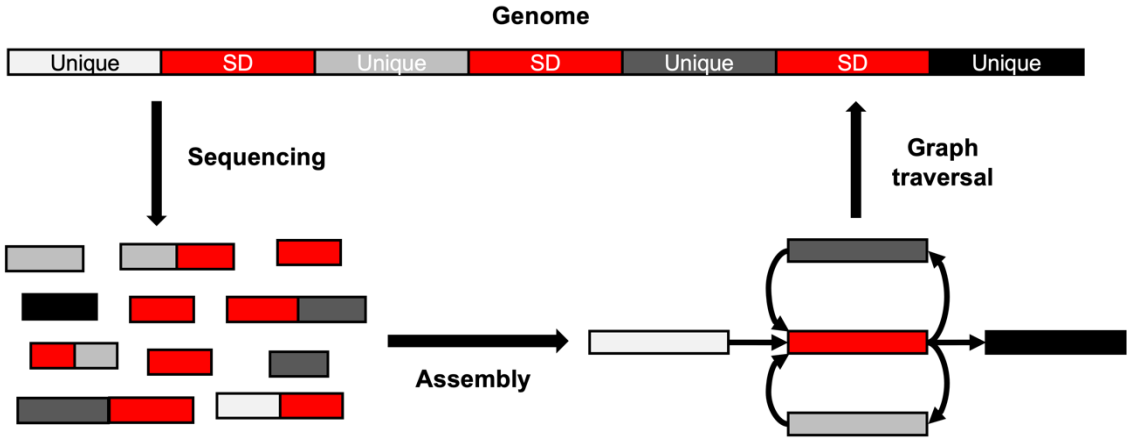


Figure 1.1. Ambiguity in assembly due to repetitive genomic sequences.
Genome assembly with WGS is accomplished by sequencing the genome such that each base is covered by many sequence reads and then overlapping the reads to make an assembly graph. This graph can then be traversed to recreate the genome; however, repetitive sequences add cycles to the graph making the traversal ambiguous.

Although the private effort was unable to resolve SDs, it did champion whole-genome shotgun sequencing, which has since become the preferred sequencing method due to its scalability and reduced cost. Even the public effort to assemble the mouse genome took advantage of WGS to construct contigs, which were only then scaffolded together using BACs and additional technologies (Waterston et al. 2002).

While there were fruitful areas of research in genome assembly from a theoretical perspective (Pevzner et al. 2001, 2004; Myers 2005; Chaisson et al. 2009), when it came to highly identical SDs in primates there was little progress in assembling them in whole-genome data in

the 2000s; although, there were many successful targeted efforts through the use of BACs and paired-end libraries (Skaletsky et al. 2003; Schmutz et al. 2004; Antonacci et al. 2010; Tuzun et al. 2005). This was due to the fundamental rule of repeats: without reads longer than the largest perfect repeat, it is impossible to accurately reconstruct the genome (Myers 1995).

1.3.2 *Advances in sequencing technologies and assembly*

The role of WGS in assembly was redefined with the advent of long-read sequencing technologies. Pacific Biosciences (PacBio) introduced their first commercial machine in 2010, boasting single-molecule sequencing with reads >10 kbp. This was made possible through zero-mode waveguides, which can isolate the signal from a fluorescing polymerase operating on a growing DNA strand (Eid et al. 2009). Not long after in 2014, Oxford Nanopore Technologies (ONT) released their first commercial sequencing machine, the MinION. This platform utilized the changes that occur in electrical current as nucleic acids pass through a protein nanopore to perform sequencing (Howorka et al. 2001).

These advances in sequencing technology necessitated the development of algorithms to utilize them and in 2013 the principles of overlap-layout-consensus (OLC) developed by Eugene Myers were incorporated into an assembly platform called HGAP for the long reads from the PacBio RS (Chin et al. 2013). This work was at first limited to small microbial genomes; however, it was not long before improvements in heuristics for all-by-all read alignment (Berlin et al. 2015) and new assembly methods based on string-graphs (Myers 2005; Chin et al. 2016) and the Celera assembler (Koren et al. 2017) allowed for structural variation detection and assembly of large mammalian genomes (Berlin et al. 2015; Chaisson et al. 2015a; Gordon et al. 2016; Chin et al. 2016; Koren et al. 2017). These assemblies were >10-fold more contiguous than all previous mammalian assemblies, save the human and mouse references (Chaisson et al. 2015b). Cell lines

and BAC sequences from complete hydatidiform moles (CHM) were a key resource (Eichler et al. 2002) in long-read assembly efforts particularly within duplications. CHM cell lines contain two copies of their paternal haplotype making them essentially haploid for the purposes of sequence and assembly. This has provided an essential benchmark for assembly since there is no ambiguity between paralogous and allelic variation in complex regions.

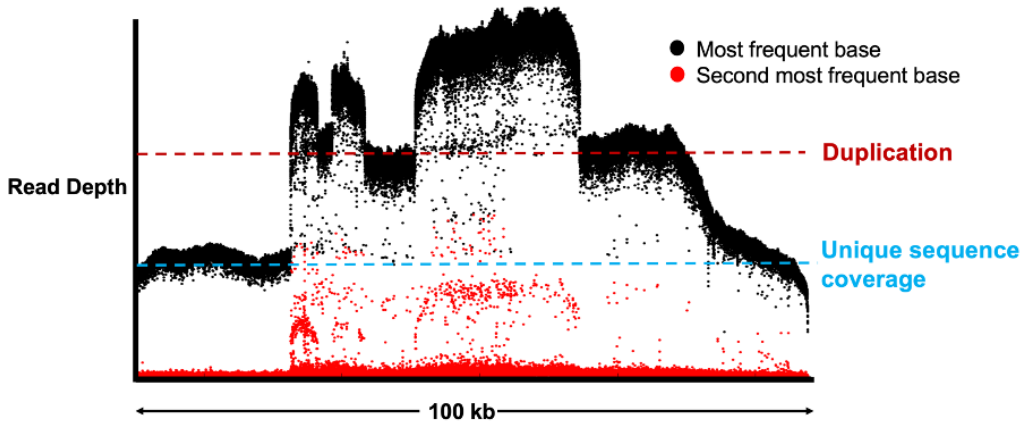


Figure 1.2. Collapse of paralogous sequences from multiple segmental duplications. Collapsed duplications are a result of under assembly of a duplication where there are more copies in the genome than in the assembly. Collapses are characterized by an apparent increase in read coverage and sequence variation; however, both of these effects can be attributed to the alignment of reads from paralogous duplications.

While read lengths were beginning to exceed the length of perfect repeats (at most ~50 kbp) in the human genome, the poor sequence accuracy of PacBio and ONT (~85%) made it impossible to distinguish duplications with less than ~3% divergence. Failure to assemble all the copies of duplicated loci in long-read assemblies often results in collapses (Figure 1.2) where the sequence from multiple paralogous loci are only represented in one collapsed version of the locus (Phillippy et al. 2008). Collapse is the fate of the largest and most highly identical SDs in genome assembly, and the underrepresentation of these sequences is particularly detrimental to the study

of human-specific gene duplication events [e.g., *NOTCH2NL*, *SRGAP2C*, *GRPIN2*, *TCAF1/2*, *ARHGAP11B* (Dennis et al. 2017; Chin et al. 2016)].

1.4 Goals of this dissertation

The goal of this dissertation is to develop methods using long-read sequencing to characterize the structure and organization of segmental duplications in the human genome. There have been many previous attempts to characterize structural variation. These attempts can be broken into three primary categories:

1) **Microarray approaches** such as single-nucleotide polymorphism microarray and array CGH (comparative genomic hybridization) could identify differences in copy number across large regions. However, the spacing between ascertained variants made the methods low resolution and they did not report the orientation, location, or sequence content of structural variants. (Fredman et al. 2004; Locke et al. 2003; McCarroll et al. 2006; Wellcome Trust Case Control Consortium et al. 2010; McCarroll et al. 2008)
2) **Targeted BAC assemblies** where individual BACs were selected and assembled with Sanger or long-read sequencing were very slow and expensive, which made it infeasible for whole-genome ascertainment and impossible for large numbers of samples. Additionally, assembly of tandem duplications is not assisted through the use of BACs and there are some human sequences that are unstable in or toxic to *e. coli*. (Schmutz et al. 2004; Eichler et al. 2002; Nuttle et al. 2016; Dennis et al. 2012; Mohajeri et al. 2016; Watson et al. 2013; Steinberg et al. 2014)

3) **Next-generation sequencing approaches** utilized read-depth or paired-end information in order to infer structural variation. These methods were limited in that they relied on a reference sequence, often could not identify copy number neutral events, and did not resolve the underlying sequence structure. (Alkan et al. 2009; Sudmant et al. 2015b, 2015a, 2010, 2013; Medvedev et al. 2010; Korbel et al. 2007; Mills et al. 2011)

This research will address these limitations by leveraging both novel computational methods and long-read sequencing technologies to close the last gaps in the human genome and understand the nature of normal structural variation in the human species.

1.5 Topics in this dissertation

1.5.1 *Assembling highly identical segmental duplications from low accuracy long reads*

Chapter 2 focuses on the development of a computational method for polyploid phasing of long sequence reads to resolve collapsed regions of segmental duplications within genome assemblies. My assembly tool Segmental Duplication Assembler (SDA) is able to construct graphs from the paralogous variation within collapses to ultimately resolve the underlying duplications. Paralog-specific variants (PSVs) identified in the collapsed sequence reads are used as nodes within the graph and these PSVs are connected via long-read sequencing data to form the edges of the graph. This formulation creates natural cliques that harbor the variation and reads that are particular to one copy of the duplication. By formalizing these cliques using a heuristic for the correlation clustering problem, we are able to reassemble collapsed SDs by phasing the reads into groups reflective of each paralog. To demonstrate the utility of the method, we applied it to single-

molecule, real-time sequence data from three human genomes and recovered 33-79 Mbp of duplications the equivalent of a small chromosome's worth of sequence.

1.5.2 *Improvements in assembly from new sequencing technologies*

PacBio high-fidelity sequencing, or HiFi sequencing, is a new technology that allows for single-molecule reads that are both long (>10 kbp) and highly accurate (>99%). This is accomplished by circularizing the DNA in the PacBio library preparation, which allows the polymerase to sequence the same molecule many times. The resulting "subreads" can then be computationally combined to create a highly accurate consensus read that is derived from a single DNA molecule (Wenger et al. 2019; Eid et al. 2009).

Chapter 3 focuses on this significant advance in sequencing technology and its impact on genome assembly and particularly the assembly of SDs. I compared the accuracy, continuity, and gene annotation of genome assemblies generated from either HiFi or continuous long-read (CLR) datasets from the same CHM. My findings show that HiFi assembled an additional 10% of duplicated regions and more accurately represented the structure of tandem repeats as well as the SD-rich pericentromeric regions. Additionally, the HiFi genome assembly was generated in significantly less time with fewer computational resources than the CLR assembly. Although the HiFi assembly improved continuity and accuracy in many complex regions of the genome, it still fell short in the assembly of centromeric DNA and the largest SD regions.

1.5.3 *The complete scope of segmental duplications in a human genome*

In addition to HiFi sequencing, the academic community developed protocols that allowed for ultra-long ONT sequencing with read lengths routinely exceeding 100 kbp and occasionally reaching >1 Mbp (Jain et al. 2018b). To leverage the incredible length of these reads, Adam

Phillippy and Karen Miga led the Telomere-to-Telomere (T2T) consortium in an effort to construct the first complete assembly of a human chromosome: chromosome X. However, poor sequence accuracy often obscured the variation that existed between the most complex duplications resulting in collapses where reads would map randomly to whichever copy had slightly higher quality. A collaborative effort with the Eichler lab resolved these issues and generated a highly accurate and complete assembly of the human X chromosome (Miga et al. 2020). This collaboration continued to be fruitful, leading to the development of HiCanu, an assembler purpose built for HiFi data and capable of resolving centromeres in an automated fashion (Nurk et al. 2020), and the completion of the first human autosome: chromosome 8 (Logsdon et al. 2020b).

Chapter 4 focuses on my work in the latest effort from the T2T consortium to finish all human chromosomes in the CHM13 genome, with the exception of five gaps that correspond to the rDNA clusters on each of the acrocentric short arms (Nurk et al., unpublished). Building on this assembly, we present the first comprehensive view of human SD organization in a complete assembly. From this assembly, we can increase the estimated fraction of the human genome composed of SDs from 5.4% to 7.0%. In addition, we identify novel regions of genomic instability, locate methylation differences between SD clusters, and predict 149 novel gene models. Copy number analysis of 266 human genomes shows that 91% of the new T2T SD sequence (68.3 Mbp) better represents human variation. We find that 63% (35.11/55.7 Mbp) of acrocentric chromosomes consist of SDs distinct from rDNA and satellite sequences. Comparing long-read assemblies from other human (n=11) and nonhuman (n=6) primate genomes, we use the T2T assembly to systematically reconstruct the evolution and structural haplotype diversity of biomedically relevant (*LPA*, *SMN*) and duplicated genes (*TBC1D3*, *SRGAP2C*, *ARHGAP11B*) important in the expansion of the human frontal cortex.

Chapter 2. LONG-READ SEQUENCE AND ASSEMBLY OF SEGMENTAL DUPLICATIONS

Chapter 2 is adapted with minimal modification from:

> Mitchell R. Vollger, Philip C. Dishuck, Melanie Sorensen, Annemarie E. Welch, Vy Dang, Max L. Dougherty, Tina A. Graves-Lindsay, Richard K. Wilson, Mark J. P. Chaisson, and Evan E. Eichler. 2019. "Long-Read Sequence and Assembly of Segmental Duplications." Nature Methods 16 (1): 88–94. http://dx.doi.org/10.1038/s41592-018-0236-3.

2.1 Abstract

We developed a computational method based on polyploid phasing of long sequence reads to resolve collapsed regions of segmental duplications within genome assemblies. Segmental Duplication Assembler (SDA), constructs graphs where paralogous sequence variants define the nodes and long-read sequences provide attraction and repulsion edges allowing us to partition and assemble long reads corresponding to distinct paralogs. We apply it to single-molecule, real-time sequence data from three human genomes and recover 33-79 Mbp of duplications where approximately half of the loci are diverged (<99.8%) when compared to the reference genome. We show that the corresponding sequence is highly accurate (>99.9%) and that the diverged sequence corresponds to copy number variable paralogs that are absent from the human reference. Our method can be applied to other complex genomes to resolve the last gene-rich gaps, improve duplicate gene annotation, and better understand copy number variant genetic diversity at the base-pair level.

2.2 Introduction

Advances in sequencing technologies and the development of novel computational assembly algorithms are central to the complete characterization of complex genomes. Recent developments in long-read sequencing technology have dramatically improved the contiguity and speed at which *de novo* assemblies of complex genomes can be generated (Alkan et al. 2011a, 2011b; Seo et al. 2016; Shi et al. 2016; Bickhart et al. 2017; Gordon et al. 2016; Koren et al. 2017; Chin et al. 2016). Individual labs, for example, can now accurately assemble >90% of the human euchromatin in less than 1,000 contigs within a few months(Huddleston et al. 2017; Kronenberg et al. 2018). Despite these recent advances, significant portions of the genome remain unresolved. This is especially true for larger, highly identical repetitive regions, including heterochromatin and gene-rich regions associated with segmental duplications (SDs), which are larger than majority of long reads (Kelley and Salzberg 2010; Pop 2004; Pevzner et al. 2001; Myers 2005; Pevzner et al. 2004).

SDs in most mammalian genomes are organized into complex regions typically >100 kbp in length and, by definition, are present at multiple locations. They contribute to dosage imbalance associated with disease (Stankiewicz and Lupski 2002b; Sharp et al. 2006) and are ten times more likely to contribute to normal copy number variation (Sudmant et al. 2015a). They are also a reservoir for gene innovations associated with species adaptations (Chen et al. 2015; Dennis and Eichler 2016; Abegglen et al. 2015). The size, copy number, and sequence identity of SDs means that they are usually the last regions of the genome to be sequenced and assembled often using large-insert BAC (bacterial artificial chromosomes) (Abegglen et al. 2015; Lander et al. 2001). More than half the gaps that remain in FALCON-based genome assemblies of single-molecule, real-time (SMRT) sequence data correspond to regions of SD. We estimate that the architecture of

only 29.2% of SD bases are resolved in an assembly of CHM1 (Figure S1, Supplementary Table S1, Methods) with most disease-associated regions unresolved (Table S2) (Emanuel and Shaikh 2001; Stankiewicz and Lupski 2002b). Similarly, an assembly of NA12878 using longer Oxford Nanopore Technologies (ONT) ultra-long reads (Jain et al. 2018b) shows moderate improvement (32.9% resolved) but leaves most SDs unresolved (Figure S1, Table S1).

Here, we develop and apply the Segmental Duplication Assembler (SDA) method that takes advantage of paralogous sequence variants (PSVs) and correlation clustering (Chaisson et al. 2017) to uniquely assemble different paralogs of SDs that were previously collapsed in long-read human genome assemblies. We apply it to actual SMRT and ONT long-read datasets to resolve SDs in recent assemblies and generate >30 Mbp of highly accurate, novel human genome sequence data. This method is computationally tractable and a generalizable solution for resolving collapsed repeat content in *de novo* assemblies of other mammalian genomes.

2.3 Results

2.3.1 ***The problem: Unresolved SDs***

While ONT and PacBio sequencing platforms generate long sequence reads, they also typically suffer from high error rates between 10-15% (Sedlazeck et al. 2015, 2018). The predominant long-read assembly methods for whole-genome shotgun sequence assembly (WGSA) are based on read correction and overlapping corrected reads to construct larger sequence contigs, e.g., Canu and FALCON (Koren et al. 2017; Chin et al. 2016). The high error rate of long-read sequencing platforms is particularly problematic for distinguishing paralogous and allelic sequence because the duplications are highly identical (>95%) and well within the range of error from long-read sequencing. This leads to sequence reads being recruited and merged from both paralogs and alleles during the assembly process creating collapses (Figure 2.1) where the

assembled sequence and corrected sequence contig are in error. To quantify the effect of collapse and misassembly, we compared several recent assemblies generated using both ONT and SMRT sequence data (Figure S1, Supplementary Note). Requiring a 50 kbp extension into unique sequence, we estimate that only 49.0-51.3 Mbp of SDs are fully resolved (Figure S2) leaving 71% (~125/175 Mbp) of SDs associated with gaps. We note that even without requiring an extension into unique sequence, 59.5-69.8% of SDs remain unresolved (Figure S1). We estimate that ~50 Mbp of the duplications correspond to regions where the assembly algorithm has collapsed highly identical duplications into the same contig. Analysis of an ONT assembly generated with ultra-long reads (2.5-fold coverage of reads over 100 kbp) showed a modest 8% improvement in SD assembly; however, most of the SDs still remained unresolved (Figure S1). As expected, the largest (>10 kbp) and most identical duplications (>95% identity) are particularly enriched in unresolved SDs (Figure S2a) and frequently correspond to annotated human genes (Figure S2b).

2.3.2 ***The approach: Segmental Duplication Assembler (SDA)***

Previously, we presented a computational algorithm (Chaisson et al. 2017) that could, in principle, assemble multi-copy duplications *de novo* using polyploid phasing (Das and Vikalo 2015; Aguiar and Istrail 2013; Berger et al. 2014; Puljiz and Vikalo 2016; Bonizzoni et al. 2016) and demonstrated its efficacy based on simulated datasets. Here, we develop SDA and apply it to WGSA collapsed duplications generated within existing human genome datasets. We specifically develop SDA to deal with different long-read datasets (Supplementary Note) and the generation of high-quality sequence contigs. Our method (Figure 2.1) identifies high-confidence PSVs *ab initio* and groups them using correlation clustering with defined attraction and repulsion edges into PSV graphs. We then assemble the partitioned reads independently, distinguishing the paralogous copies. Empirically we observe that we are able to assemble large duplications with less than 0.5%

sequence divergence (Supplementary Note). As a measure of reproducibility, we apply this method to four human genomes and validate the results and accuracy based on targeted BAC sequencing and analyses of specific duplicated loci.

We begin by identifying all collapsed duplications within each assembly based on an excess of sequencing read depth (Bailey et al. 2002; Kelley and Salzberg 2010) (Methods). Within the CHM1 assembly (Huddleston et al. 2017), for example, we identify 283 regions of collapse averaging 43 kbp in length (Table 2.1). When the 12.2 Mbp of collapsed CHM1 duplications are mapped back to the reference, they span 52.3 Mbp of sequence—93% (48.6 Mbp) are annotated as SDs and 88% of which (45.9 Mbp) overlap with regions of unresolved SDs in CHM1. Next, we define PSVs corresponding to each collapsed segment. We define candidate PSVs by classifying the second most frequent base at every position within the collapsed alignment and requiring sequence coverages consistent with a single-copy locus in order to distinguish PSVs from allelic variants (Methods). We next apply correlation clustering to filter false positive variants arising from sequencing error and uniquely assign each remaining PSV to the paralog from which it originates. For each collapsed region, we construct a graph where the PSVs define the nodes and the sequence reads define the edges. Attraction edges are formed when a read contains two or more PSVs connecting two or more nodes. Similarly, repulsion edges are formed when PSVs are mutually exclusive across all the sequence reads.

With this formulation of the problem, it is possible to address the correlation clustering objective, which is to minimize the number of repulsion edges within clusters and minimize the number of attraction edges between clusters. Correlation clustering offers a distinct advantage over many other clustering algorithms because it does not require the number of clusters as a starting input. It is therefore *ab initio* and defined entirely by the underlying sequence data. However,

correlation clustering is an NP (nondeterministic polynomial) complete problem; thus, we developed a heuristic to approximate the solution modeling after previous work (Ailon et al. 2005). The heuristic randomly assigns PSVs to clusters and then iteratively increases the size of the cluster by following positive edges that decrease the score of the entire graph (Methods).

2.3.3 *Resolving SDs using SDA*

We applied correlation clustering to each of the 283 collapsed regions in the CHM1 WGSA and generated a total of 668 distinct groupings. We created separate assemblies corresponding to each PSV graph partition using Canu followed by Quiver error correction. We successfully generated 590 assemblies where a single contig was produced corresponding to 33.1 Mbp of assembled sequence (Table 2.1, Figure 2.2) with an average sequence contig length of 60.7 kbp. The median assembly length was 53.0 kbp (mean 60.7 kbp), and the maximum sized assembly was 255.5 kbp. In general, the length of the assembly correlates ($r = 0.67$, Pearson's correlation) with the size of the collapse (Figure S3). Of the 668 PSV graphs, 59 failed to generate an assembly and 19 assembled into multiple contigs. An inspection of those clusters that failed to assemble showed that the majority did so due to an insufficient number of reads while clusters with multiple contigs were the result of either incomplete PSV separation among multiple contigs or variable sequence coverage.

In order to assess the accuracy and contiguity of the assembled SDs, we mapped each sequence contig back to the human reference genome (GRCh38). Of these assemblies, 48.5% (286/590) mapped to the human reference with at least 99.8% sequence identity over >90% of the contig length and accounted for ~18 Mbp of sequence. Interestingly, a similar fraction of assembled contigs (51.5% (304/590) (corresponding to 15.5 Mbp) showed greater sequence divergence ranging from 96% to 99.8% sequence identity (Figure 2.2). We consider the contigs

that “match” at high identity to GRCh38 to be correctly assembled and classify those with lower sequence identity than expected based on allelic variation (<99.8%) (1000 Genomes Project Consortium et al. 2015) to be “diverged.” Since >0.2% divergence lies outside the typical range of human allelic variation, such diverged sequence may represent different copies of the duplication not yet represented in the human genome. We examined, in detail, a few human-specific gene families (e.g., *SRGAP2* and *NOTCH2NL*) associated with neuroadaptation (Fiddes et al. 2018; Florio et al. 2018; Dennis et al. 2012; Nuttle et al. 2013; Dennis et al. 2017; Dennis and Eichler 2016) that have been the target of detailed BAC-based sequence assemblies (Table S3, Figure 2.3 and S4). Our analysis shows that we have successfully resolved the collapsed assemblies recreating the sequence and gene models present in the reference. This includes the identification and characterization of paralog-specific structural variation with most sequence assemblies matching ~99.8%-99.9% to their respective paralogs. Among these gene families, we estimate that 91%-93% of all PSVs have been correctly assigned.

We repeated this analysis for three additional long-read human genome assemblies, including a second haploid genome (CHM13) (Huddleston et al. 2017), a diploid genome of African descent (YRI19240) (Steinberg 2016), and a diploid genome assembled with ONT (NA12878) (Jain et al. 2018b) (Table 2.1, Supplementary Note, Figures S5-S7). The proportion of matched and diverged sequence assemblies as well as resolved SD regions was very similar among the PacBio genomes. For example, 83% (1,772/2,136) of clusters resolved into single contig assemblies for the African diploid genome assembly. In contrast, an analysis of a human genome assembly (NA12878) generated with ultra-long ONT reads showed more failed SD assemblies, although we note that the coverage of this genome was significantly less than that of the PacBio genome assemblies (Figure S7). Combining both the “matched” and “diverged” sequences, we

estimate that the SDA method adds an additional 72.6 and 78.6 Mbp of sequence corresponding to duplicated regions of the CHM13 and NA19240 human genomes, respectively.

2.3.4 ***Characterization of diverged duplications***

We focused on the diverged duplications and considered two possibilities: the sequence could represent misassembled sequence or, alternatively, may represent additional copies not yet present in the human reference genome. The latter may be expected given that SD regions are 10-fold more likely to be copy number polymorphic (Sudmant et al. 2015a) than unique regions of the genome. If diverged sequences resulted from the sequence and assembly of additional copies, we would expect a significant increase in the copy number difference for diverged sequences when compared to duplicated sequences that matched the human reference genome (>99.8% sequence identity). Indeed, a comparison of the copy number difference for these two categories clearly showed that diverged copies were more likely ($p = 2.0 \times 10\text{-}5$) to have a higher copy number in CHM1 (Figure 2.2) than duplicated sequences that matched the reference genome assembly.

As a more direct test, we sequenced and assembled 1,253 large-insert BAC clones (Table S4) corresponding to regions of SD from a genomic library (CHORI-17) derived from CHM1 (Chaisson et al. 2015a) (Methods). Restricting our analysis to the 304 diverged sequences assembled by SDA from CHM1, we identify 105 diverged duplications that match the CHORI-17 clones. Each of these 105 sequences aligned to a clone over at least 90% of its length and at >99.8% sequence identity (mean sequence identity of 99.97%) (Figure 2.4, Table S5). If we assume that our method targeted all SDs evenly across the whole genome, then we would expect approximately 37.4% of the bases across our diverged sequences to validate. We observe that 105 of our diverged sequences, or 36.3% of the bases, validate and show significantly better alignment to the CHM1 clone inserts when compared to GRCh38. We estimated the sequence accuracy for our assembled

duplications as 99.989 (quality value (QV) = 38.4) considering only single-base-pair mismatches and 99.857% (QV = 28.4) if indels and mismatches are counted. We note that many of the 105 validated assemblies contain sequences associated with gene families and, thus, have the potential to recover missing genic sequence not yet annotated. For example, we assembled a paralog of *NBPF1* that is 1.2% diverged from the human reference but maps with >99.99% sequence identity to a CHM1 clone (Figure 2.4, Table S6). Similarly, Sudmant and colleagues (Sudmant et al. 2015b) identified an additional duplication in 16p12.1 that exists in most individuals but was absent from the reference. Using SDA, we recovered the proposed duplication (Nuttle 2016) (Figure S8) with only one mismatched base pair across a 95 kbp alignment to the BAC-generated contig.

We analyzed more systematically the utility of these orphan SDA contigs to generate more accurate gene models for 37 human-specific segmental duplication (HSD) gene families. We selected 213,450 bulk single-molecule sequencing RNA reads (Iso-Seq) from fetal and adult human brain enriched for HSDs (Dougherty et al. 2018). We aligned Iso-Seq data and compared their mapping between SDA contigs versus previous collapsed contigs in the CHM13 assembly. Transcripts showed improved mapping to the SDA contigs for 11 gene families to varying degrees (Figure S9). We identified six gene families (Figure 2.5) where transcripts mapped better to the SDA assemblies than the human reference genome. A subset of transcripts from the *GPRIN2* (G-coupled protein inducer of neurite outgrowth) gene family are most striking with a 1.5% improvement. We aligned the second SDA *GPRIN2* contig that appears to be missing from the reference and found that it spans a gap in GRCh38 flanked by SDs (Figure 2.5). Moreover, a previous analysis of Illumina whole-genome shotgun (WGS) sequence shows that *GPRIN2* is polymorphic with copy number ranges from 3-7 copies with most humans carrying four in contrast to other apes which carry only one (diploid copy number = 2). Our analysis shows that both copies,

GPRIN2A and *GPRIN2B*, are transcribed and encode similar open reading frames, although GPRIN2B has a 3-amino-acid insertion as well as several amino acid differences when compared to the ancestral GPRIN2A (Figure S10). Interestingly, these PSVs have been erroneously classified as single-nucleotide variants (SNVs; with near 50% "allele" frequency in dbSNP) because the reference is missing this second copy (Table S7). Thus, the SDA contig not only improves gene annotation but also improves interpretation of human genetic variation.

2.4 Discussion

There are three strengths to SDA. First, our approach does not require PSVs to be predefined and, as such, can be applied to any genome assembly where long-read data of sufficient depth has been generated. A similar concept was recently applied to partition viral quasispecies (Artyomenko et al. 2017). Second, our validation results suggest that the paralog-specific assemblies are highly accurate (99.86%-99.99%). Importantly, the approach allows missing paralogs to be sequenced especially within regions of extensive copy number variation. This is particularly exciting because it allows previously uncharacterized forms of human genetic variation to be sequence-resolved for the first time. Finally, our analysis of the human genome suggests that the majority of collapsed duplications are at least partially resolved (Figure 2.2). Since unassembled SDs typically represent ~70-90 Mbp of sequence per genome, recovery of 33-79 Mbp is the equivalent of recovering an entire chromosome's worth of DNA for which accurate gene models can be constructed (Table 2.1, and Table S8). The method we have developed can be effectively applied to any genome for which long-read WGSA data exist providing access to the duplicated regions and the genes therein.

Notwithstanding these advances, limitations remain. The majority of the sequence contigs we generated with SDA are small (~54 kbp) and are not yet commensurate with the average contig

lengths generated by long-read sequence and assembly of unique regions of the genome. Only a small fraction (22%) of SDA contigs transition into unique sequence such that overlaps can be unambiguously assigned into the main genome assembly (Figure S11). Our new duplicated sequence contigs are not yet fully integrated into the genome and many of the resolved duplications remain "orphan" contigs in the absence of additional long-range mapping data. Directly integrating our SDA tool into popular long-read assemblers, to create long-range linkage information, may not be advisable even if it were possible. Optimizing parameters for SD assembly would likely come with costs for the remaining 95% of the genome. There are distinct advantages to performing bulk WGSA followed by a second-tier analysis to focus on the collapsed regions of the assembly. This is because overlap stringency should differ for high-identity duplications, and because PSVs provide important information for determining overlaps in these more difficult-to-assemble regions.

While we have shifted the accessible portions of SDs to larger (>50 kbp) and more identical regions (~99%), not all regions can be resolved using this approach. Duplications that are virtually identical cannot be distinguished and will require even longer read data, such as the ultra-long reads (>100 kbp) possible using ONT (Jain et al. 2018b). While we have developed and benchmarked SDA primarily with PacBio sequence data, we have also applied it to long-read sequence data from other platforms such as ONT (Supplementary Note). Our initial analysis of the ultra-long-read genome assembly of NA12878 (Jain et al. 2018b), for example, showed a slight improvement of 8% in SD assembly (Figure S1). However, most of the high-identity SDs remained unresolved with a similar number of collapsed duplications (n = 365) when compared to PacBio genome assemblies. Application of SDA to the ONT dataset resulted in far fewer resolved assemblies (Figure S7) with an overall lower accuracy of the assembled sequence contigs. An

important difference, however, is sequence coverage. The NA19240 PacBio assembly was sequenced at 73-fold sequence coverage versus the 35-fold ONT genome assembly. We note that while ultra-long ONT sequence reads were less successful in resolving SDs, they were useful as orthogonal data to validate PacBio SDA contigs (Supplementary Note). If long reads in excess of 200 kbp can be routinely generated with sufficient coverage to correct sequence error, it is possible that most SDs could be resolved by WGSA. The rapid advance of long-read sequencing technology may make the routine generation of ultra-long reads from low quantities of DNA a reality in the near future. Such advances would open up the possibility that other highly repetitive regions, such as centromeres and acrocentric DNA, could be routinely sequenced and assembled for the first time.

2.5 Methods

2.5.1 ***Human genome assemblies.***

We analyzed three human genome assemblies derived from haploid (CHM1 and CHM13) (Huddleston et al. 2017) and diploid source material (NA19240) (Steinberg 2016) of African descent. FALCON genome assemblies were previously generated from at least 61-fold SMRT sequence using P6C4 chemistry generated on the PacBio RS II sequencing platform. We also analyzed one recent human genome assembly (NA12878) generated with ultra-long ONT sequence reads (Jain et al. 2018b).

2.5.2 ***SD characterization.***

We mapped each human *de novo* assembly to the human reference genome GRCh38 using MashMap 2.0 (default settings) (Jain et al. 2018a) and defined SD regions based on intersection with annotated SDs in GRCh38. Sequence contigs overlapping SDs were defined as resolved if

the contig completely contained the SD sequence and extended at least 50 kbp on either side into unique sequence. We compared the number of resolved and unresolved contigs (**Figure 1a**) for each assembly as a function of SD block length and maximum percent identity. Scripts are available at https://github.com/mvollger/segDupPlots, as well as a more detailed description of the analysis in the README.

2.5.3 ***Assembly collapse and PSV definition.***

Within each assembly, we identified collapsed SDs by mapping SMRT or ONT sequencing reads back to each genome using BLASR (Chaisson and Tesler 2012) (version rc46) or minimap2 (Li 2018) (version 2.11) for ONT. Using unique regions, we computed the read coverage and standard deviation across 100 bp windows using the following BLASR settings (blasr $READS $ASM -sa $ASMSA \ -sdpTupleSize 13 -sdpMaxAnchorsPerPosition 10 -maxMatch 25 \ -minMapQV 30 -bestn 2 -advanceExactMatches 15 \ -clipping subread –sam). We excluded regions with >75% common repeat elements (RepeatMasker version 2004/03/06 –e wublast) and regions in the bottom or top two percentiles. We defined collapsed regions as those with a mean sequence coverage >3 standard deviations beyond the mean coverage and that were at least 9,000 bp in length (as smaller regions were routinely sequence and assembled). We examined all regions of collapse for the presence of SNVs and cataloged the second most common base at each position within the collapsed region using a more sensitive BLASR settings (blasr {input.basreads} {input.ref} \ -sam -preserveReadTitle -clipping subread \ -bestn 1 \ -mismatch 3 -insertion 9 -deletion 9 -minAlignLength 500). We defined these SNVs as potential PSVs if the sequence coverage was consistent with the read depth of unique regions. Three thresholds were applied to determine if an SNV was also a PSV. First, the total depth at the given position had to be at least the mean coverage plus three standard deviations. Second, the frequency of the second most

frequent base had to be less than the mean coverage. Finally, the frequency of the second most frequent base had to be greater than the mean coverage minus three standard deviations or half the mean coverage, whichever was greater. This process favors the selection of PSVs over allelic variants (Figure S4). We developed a Snakemake pipeline for this analysis ProcessCollapsedAssembly.py, which can be found at https://github.com/mvollger/SDA.

2.5.4 ***PSV graph construction.***

We constructed graphs for collapsed regions where each PSV corresponds to a node and sequence reads represent edges. Attraction edges are created when two PSV nodes have a substantial number of sequencing reads that contain both PSVs. Among reads containing both PSVs, we test whether each PSV is more likely to be real or a sequencing error using the ratio of two binomial tests. If at each PSV the log base 10 ratio of the two binomial tests was at least 1.5 (i.e., ~31 times more likely to be real than error), then an attraction edge was formed. Repulsion edges were created between any PSVs where less than 10% of the mean coverage of sequencing reads carried both PSVs.

2.5.5 ***Correlation clustering.***

We initially added all nodes to an unclustered set from which a node was randomly selected and then expanded upon by iteratively searching neighbors of this node that reduce the overall score of the PSV graph (i.e., minimize the objective function). As nodes that meet this criterion are added to the cluster, they are removed from the unclustered set. This process was repeated until there were no unclustered nodes as previously described (Chaisson et al. 2017). Next, all pairwise clusters are examined to see if they would improve the score of the graph if combined into a single cluster. Clusters are combined starting with the pairwise cluster that most improves the score of

the correlation clustering objective. Clusters of three or fewer nodes are removed. The correlation clustering heuristic is run independently 15 times each with different random initializations and the clustering that best minimizes the correlation clustering objective is used to construct the final PSV clusters. It can be the case that in the construction of the PSV graph the PSVs are already clustered appropriately as unconnected components in the graph. In this case the application of correlation clustering is unnecessary to phase PSVs.

2.5.6 ***PSV read partition and assembly.***

In order to partition SMRT or ONT sequencing reads according to the PSV clusters defined by correlation clustering, we apply WhatsHap (Patterson et al. 2015) (version 0.16) using the following parameters (whatshap haplotag $INPUT_VCF $INPUT_BAM -o $OUTPUT_BAM). Phasing was run on the entire set of reads for each PSV cluster, i.e., if there were five PSV clusters, WhatsHap was run five times to create five partitions of reads. After partitioning the reads into different paralogs, we independently assemble each correlation cluster with Canu version 1.5, followed by error correction (Quiver v 1.1.0) using the same set of reads. Specialized parameters are applied such that Canu can execute on such short contigs (https://github.com/mvollger/SDA/blob/master/SDA.2.snakemake.py).

2.5.7 ***BAC clone insert sequencing.***

BAC clones from CHORI-17 (CH17) clone libraries (http://bacpac.chori.org) were hybridized with probes targeting complex or highly duplicated regions of the human genome reference (GRCh38) (n = 727) or based on previously sequenced clones (n = 526) (Chaisson et al. 2015a). DNA from positive clones was isolated by a modified alkaline lysis miniprep procedure as follows: cell pellet was resuspended in 200 µL Qiagen buffer P1 with RNase and lysed with

200 µL of 0.2M NaOH/1%SDS solution for five minutes. Lysis was neutralized with 280 µL 3M NaOAc, pH 4.8. Neutralized lysate was incubated on ice for up to 20 minutes, collected by centrifugation for 30 min at 4000 rpm, concentrated by standard isopropanol and then ethanol precipitation, and resuspended in 25 µL 10 mM Tris-Cl pH 8.5. We prepared barcoded libraries from clone DNA using Illumina-compatible Nextera DNA sample prep kits (Epicentre, Cat. No. GA09115) as described previously (Steinberg 2012) and paired-end sequenced (125 bp reads) on an Illumina HiSeq 2500. Reads were then mapped to the reference genome (GRCh38) to identify singly unique nucleotide k-mers (SUNKs), defined as 30-mers that identify a region of the genome and can be used in conjunction with short-read sequencing data to genotype highly identical paralogs (Sudmant et al. 2010). This SUNK mapping was used to select a subset of positive clones for PacBio sequencing. BAC DNA from selected clones was isolated using a High Pure Plasmid Isolation Kit from Roche Applied Science per manufacturer instructions using 6 mL LB media with Chloramphenicol selective marker. We pooled non-overlapping BACs at equal molar amounts before library preparation. Approximately 1 µg of DNA per BAC was pooled and sheared using a Covaris® g-TUBE®. Libraries were processed using the PacBio SMRTbell Template Prep kit following the protocol 'Procedure and Checklist -20 kb Template Preparation Using BluePippin™ Size-Selection System'. Libraries were size-selected on the Sage PippinHT with a start value of 10,000-12,000 and an end value of 50000. DNA/Polymerase Binding Kit (P6-C4 chemistry) was used to bind DNA template to DNA polymerase and the MagBead kit was used to capture DNA polymerase/template complexes for loading. Libraries were sequenced on the PacBio RS II platform. We performed *de novo* assembly of pooled BAC inserts using Canu v1.5 (Koren et al. 2017). Reads were masked for vector sequence (pBACGK1.1) and assembled with Canu followed by consensus sequence calling with Quiver. Canu is specifically designed for

assembly with long error-prone reads, while Quiver is a multi-read consensus algorithm that uses the raw pulse and base call information generated during SMRT sequencing for error correction. PacBio assemblies were reviewed for misassembly by visualizing read depth of PacBio reads in Parasight (http://eichlerlab.gs.washington.edu/jeff/parasight/index.html) using coverage summaries generated during the resequencing protocol.

2.5.8 ***Data availability***

SMRT WGS for CHM1, CHM13, and NA12940 from this study are available at the NCBI Sequence Read Archive (SRA; https://www.ncbi.nlm.nih.gov/sra) under accession numbers SRP044331 for CHM1; SRX818607, SRX825542, and SRX825575-SRX825579 for CHM13; and SRX1093000, SRX1093555, SRX1093654, SRX1094289, SRX1094374, SRX1094388, and SRX1096798 for NA19240. ONT WGS data are available at https://github.com/nanopore-wgs-consortium/NA12878/blob/master/Genome.md. *De novo* assemblies of CHM1, CHM13, NA12940, and NA12878 from this study are available at the NCBI Assembly database (Assembly; https://www.ncbi.nlm.nih.gov/assembly/) under accession numbers GCA_001297185.1, GCA_000983455.2, GCA_001524155.4, and GCA_900232925.1, respectively. Assembled CHORI-17 BACs are available at the NCBI Clone DB (Clone; https://www.ncbi.nlm.nih.gov/clone/) under the accession numbers listed in Table S4. Information about length, PSVs, and mapping location in GRCh38 can be found for all the SDA contigs generated in Table S8. Additional data that support the findings of this study are available from the corresponding author upon request.

2.5.9 ***Code availability***

Code for analyzing the resolved and unresolved SDs in a *de novo* assembly can be found at https://github.com/mvollger/segDupPlots. Code for processing *de novo* assemblies with snakemake (Köster and Rahmann 2018) to find collapses and running SDA can be found at https://github.com/mvollger/SDA.

2.6 Figures

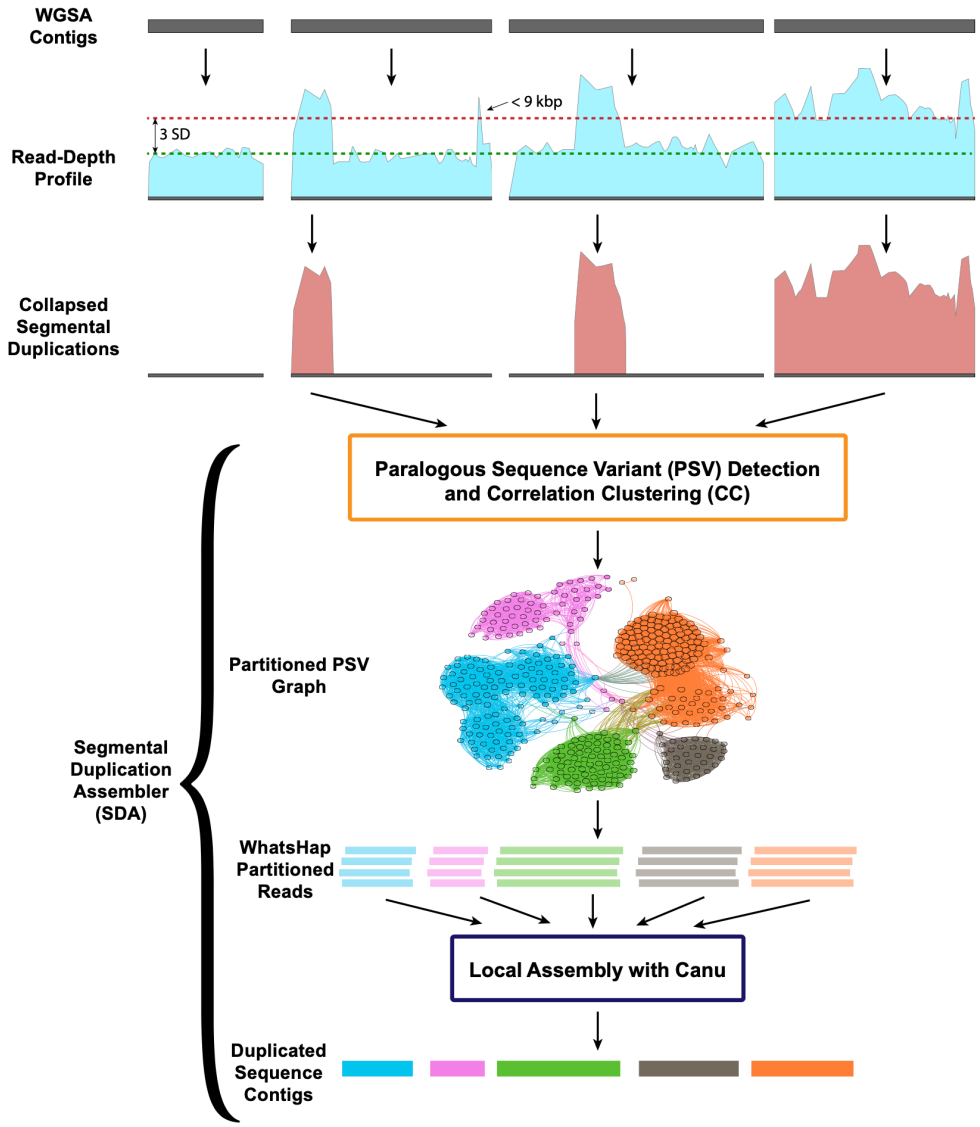


Figure 2.1. Flowchart of Segmental Duplication Assembler (SDA) method.

Regions of collapsed SDs are defined by assessing whole-genome shotgun (WGS) sequence read-depth profiles using BLASR across sequence contigs generated from a *de novo* WGSA. Regions (>9 kbp in length) with elevated sequence coverage (three standard deviations plus the mean) and not entirely composed of common repeats are considered collapsed SDs. Sequence reads corresponding to the collapsed SDs are recovered and examined for variants at each position along the collapse. Single-base-pair substitutions that appear at the same threshold as unique sequencing depth are identified and flagged as paralog-specific variants (PSVs) effectively partitioning reads into PSV clusters (WhatsHap). Sequence reads assigned to each PSV cluster are independently assembled using Canu and error-corrected using Quiver.

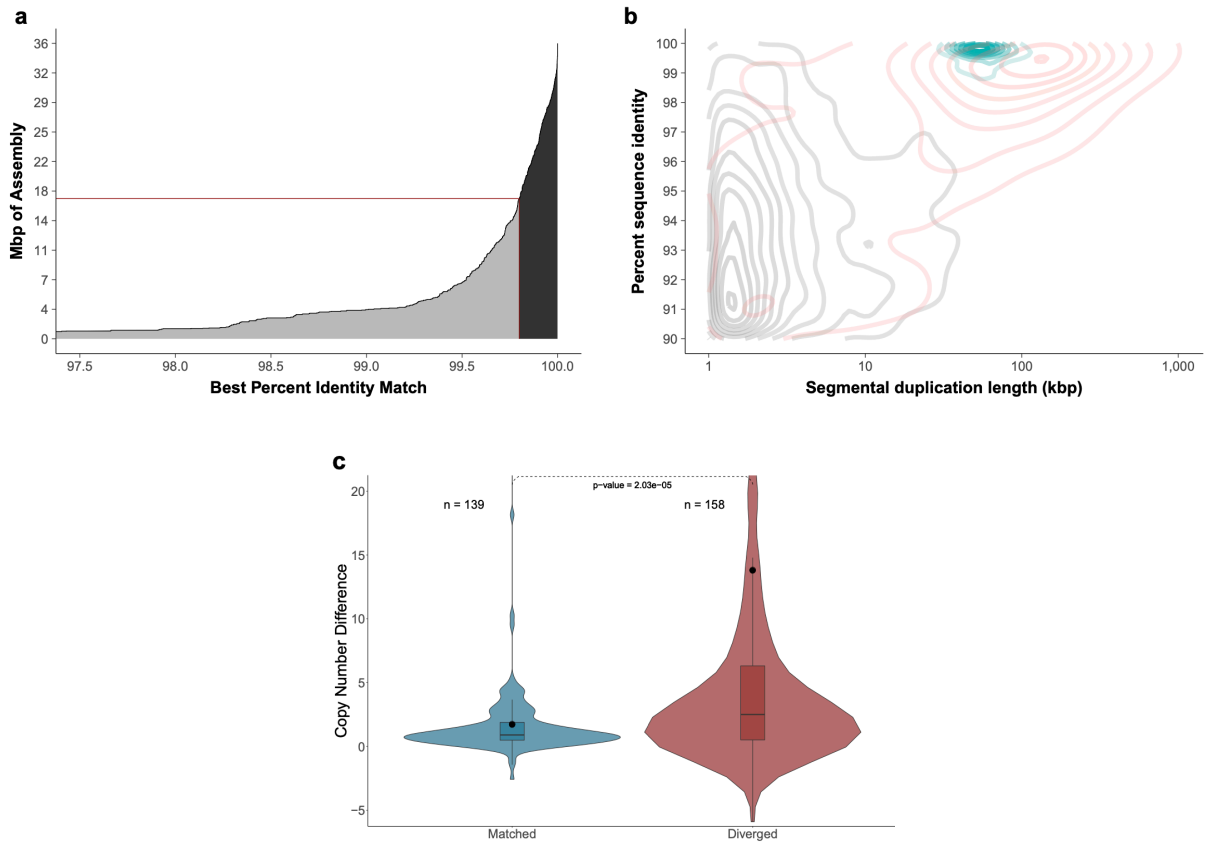


Figure 2.2. SDA results of the CHM1 human genome assembly.

a) A cumulative distribution of the SDA assemblies and their percent identity to their best match in the reference (<99.8% identity, gray; >99.8% identity, black). The number of assembly Mbp is calculated independently of a mapping to the reference, unlike in Table 2.1. b) Density plot of SDs plotted by length and percent identity. Black represents duplications resolved in the CHM1 assembly, red shows unresolved duplications in the CHM1 assembly, and blue represents paralogs assembled using SDA. Resolved SDA sequences are "content" resolved and not ordered within the genome, whereas SDs in the assembly must extend into unique sequence on both sides to be considered resolved. c) Copy number difference (CND) between CHM1 and the reference genome (CHM1 copy number – reference genome copy number) comparing n=139 SD regions that match (>99.8%) versus n=158 diverged SD regions (<99.8% identity). The mean CND of the matched sequence is 1.75, and the mean CND of the diverged sequence is 13.82 (black dot) (two-sided Mann-Whitney test; p=2.03*10-5). The boxes indicate the range between the first and third quartiles, with the bold line specifying the median. The whiskers show the minimum and maximum within 1.5 times the interquartile range extending from the first and third quartiles. Copy number was estimated in CHM1 examining k-mer frequency found in Illumina WGS reads; methods are described in Sudmant et al. 2015. A similar approach was used for estimating copy number in the reference except we generated simulated reads using the reference and then estimated copy number in the same fashion using the simulated reads.

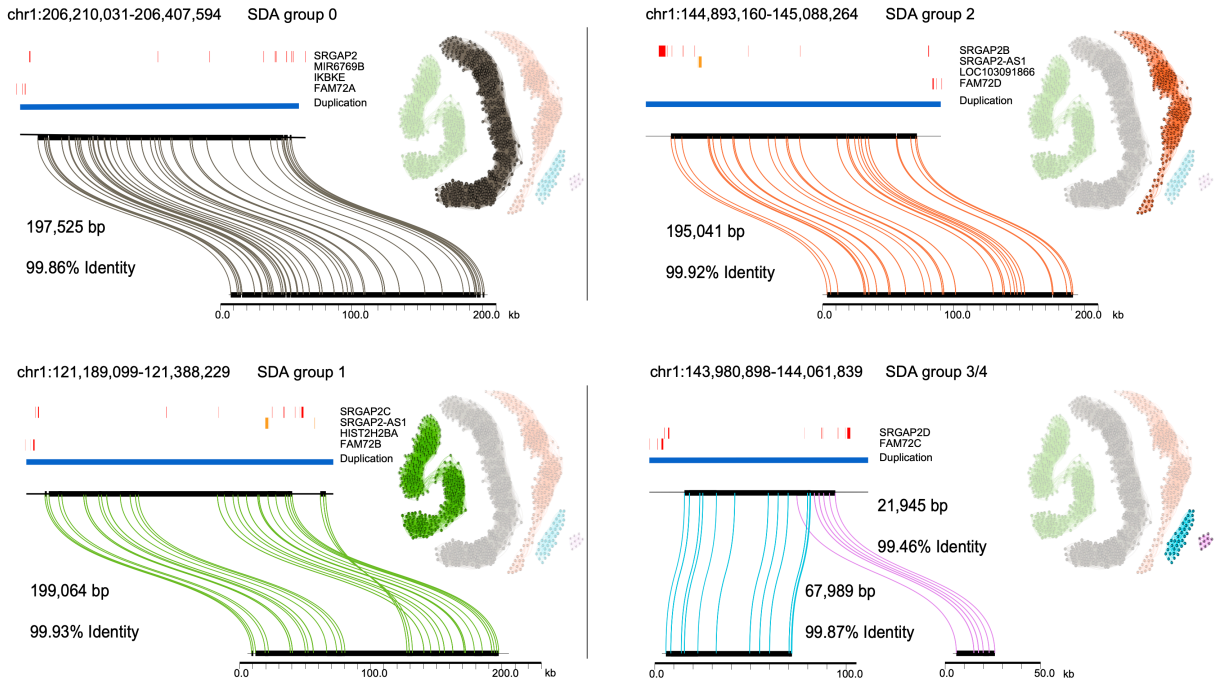


Figure 2.3. Sequence and assembly of *SRGAP2* loci in the CHM13 human genome.

SDA sequence contigs from CHM13 aligned to the GRCh38 loci for *SRGAP2*(*A/B/C/D*) using Miropeats (Parsons 1995). The length and percent identity of each alignment is shown. Similarly, in CHM1 we found that, on average, our sequence is 99.91% identical over all four loci and >99.999% identical if only mismatched bases are counted as errors as opposed to including indels. Adjacent to each alignment is the PSV graph with the relevant PSVs highlighted. Each node represents a PSV and loci are colored and numbered to reflect the grouping determined by correlation clustering. An edge is added between two nodes (PSVs) when a sequencing read contains both PSVs. The opacity of each node scales from 25% to 100% to reflect the position of the PSV along the collapse: 25% opacity reflects the first position along the collapse and 100% reflects the final position. For a more detailed view of the opacity of the nodes, see Figure S12. Clusters 3 and 4 in the PSV graph represent the fourth paralog (*SRGAP2D*), which carries a large deletion in the middle relative to the other paralogs.

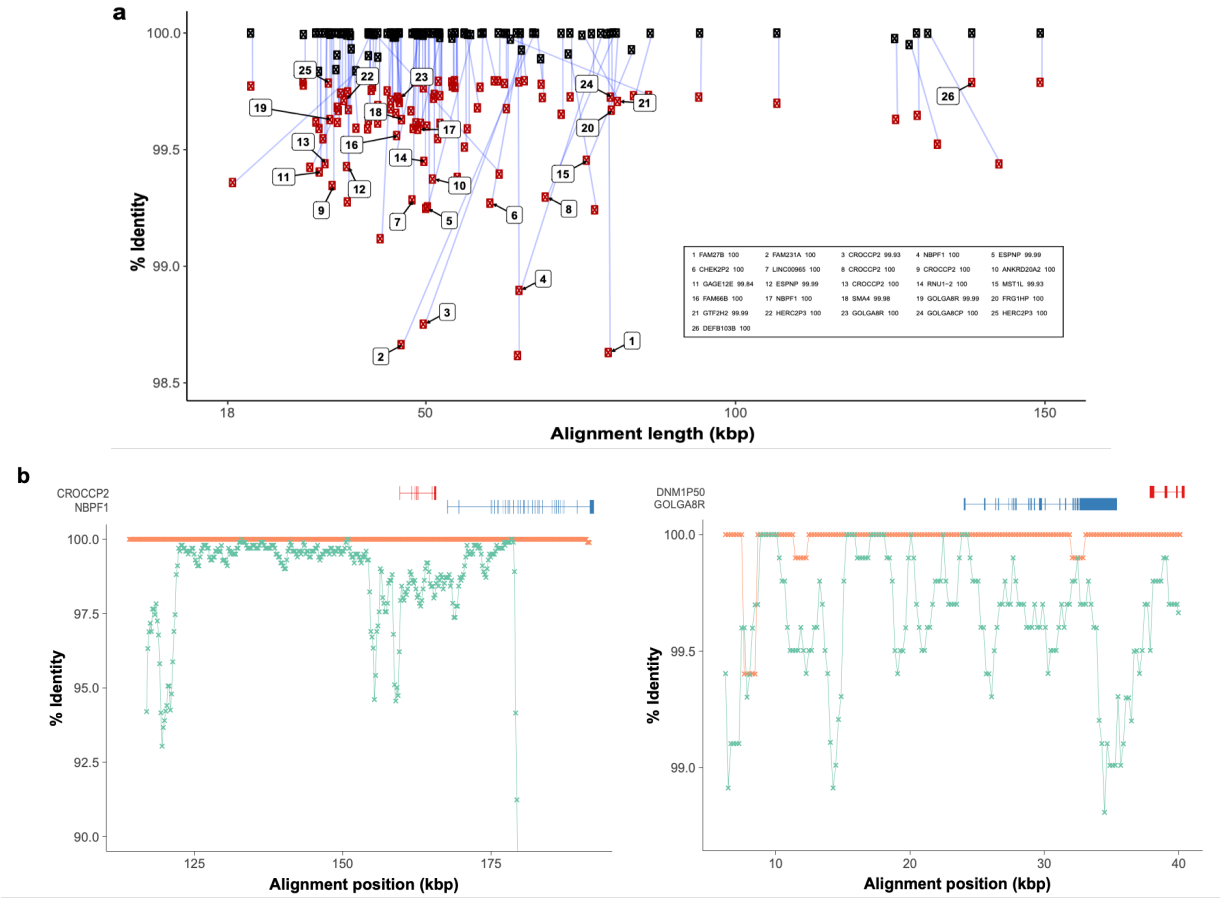


Figure 2.4. Correspondence between SDA sequence-diverged contigs and BACs.

a) Alignment length and percent identity sequence match for n=105 diverged SDA contigs compared to BAC clones (black) sequenced from the same source individual (CHM1) and the human reference genome (GRCh38) (red). (See Tables S5 and S6 for more details.) b) Two examples of genes corresponding to diverged duplications are shown where the SDA sequence is aligned to both the reference genome (blue) and the CHM1 BACs (orange). BLASR alignments are computed in 1000 bp windows sliding 500 bp (steps).

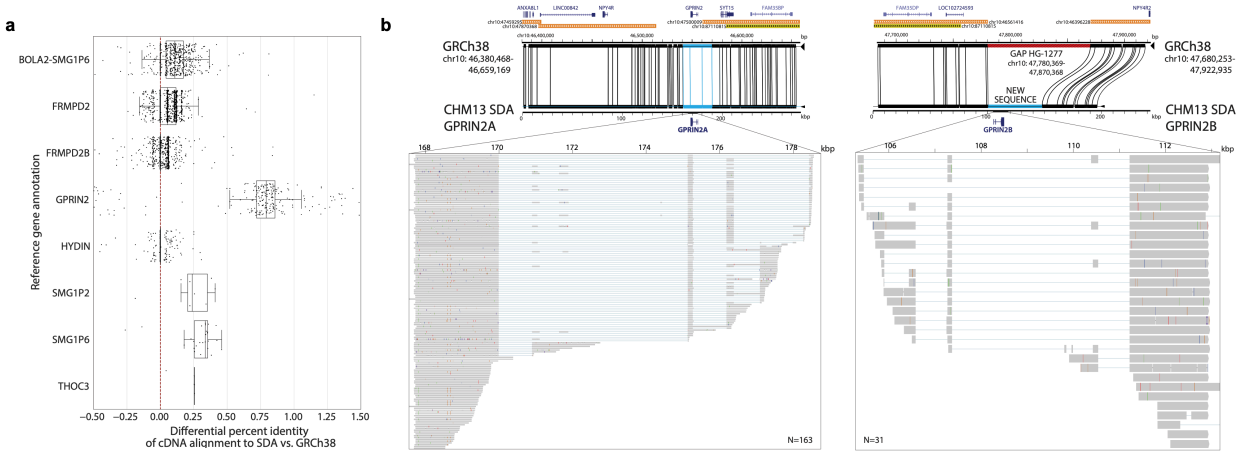


Figure 2.5. Gene discovery in previously unresolved duplications.

a) The percent identity differential of the mapping of full-length Iso-Seq transcripts (n=4,718) from human-specific duplications (HSDs) to both GRCh38 and SDA results on CHM13. The red dotted line represents equal mapping between the two, whereas points to the right represent an improved mapping with the SDA contigs. The boxes indicate the range between the first and third quartiles, with the bold line specifying the median. The whiskers show the minimum and maximum within 1.5 times the interquartile range extending from the first and third quartiles. b) *GPRIN2* SDA contigs compared (Miropeats) to the human reference assembly (GRCh38) with gene and SD annotation. The SDA contigs close a gap (red) in GRCh38, which contains a duplicate copy of *GPRIN2A* denoted here as *GPRIN2B* (Figure S10, Table S7).

2.7 Tables

Table 2.1. SDA assembly statistics.

	De novo Assembly					Segmental Duplication Assembler (SDA)				
Sample	**Assembly Accession**	**Contig N50 (Mbp)**	**Sequence Coverage**	**Read N50 (kbp)**	**Unresolved SDs (Mbp)**	**Collapses (count / Mbp)**	**Matched (count / Mbp)**	**Diverged (count / Mbp)**	**Multiple Assemblies (count / Mbp)**	**Failed**
CHM1[9]	GCA_001297185.1	26.9	61	20.5	124.1	283 / 52.3	286 / 17.98	304 / 15.51	19 / 1	59
CHM13[9]	GCA_002884485.1	29.3	67	18.2	126.5	527 / 86.6	685 / 39.1	755 / 35.0	69 / 3.1	339
NA19240[42]	GCA_001524155.4	29.1	61	17.5	124.1	489 / 82.4	789 / 38.8	983 / 40.9	107 / 5.8	257
NA12878[27]	GCA_900232925.1	7.7	35	12.5	117.7	365 / 52.5	38 / 0.066	792 / 22.1	8 / 0.21	1062

Genome summary statistics for four human genomes sequenced (SMRT/ONT) and assembled (FALCON/Canu) with long-read data.

Collapses from the assemblies were subjected to SDA and the number and Mbp of "matched" and "diverged" contig assemblies to the human reference genome (GRCh38) are shown.

2.8 Acknowledgements

The authors thank S. Cantsilieris and D. Gordon for technical assistance, J. Underwood for recommendations regarding the analysis of HSDs and Iso-Seq data, and T. Brown for help in editing this manuscript. This work was supported, in part, by grants from the U.S. National Institutes of Health (NIH grants HG002385 to E.E.E.; HG007635 to R.K.W. and E.E.E.; and HG003079 to R.K.W.). M.R.V. was supported by a National Library of Medicine (NLM) Big Data Training Grant for Genomics and Neuroscience (5T32LM012419-04). P.C.D. was supported by a National Human Genome Research Institute (NHGRI) Training Grant (5T32HG000035-23). E.E.E. is an investigator of the Howard Hughes Medical Institute.

2.9 Author contributions

SDA method development: M.R.V., M.J.P.C., E.E.E.; PacBio genome sequence generation: R.K.W., T.A.G-L.; BAC clone-insert sequencing and analysis; M.S., A.E.W., M.R.V.,

V.D.; Iso-Seq analysis: P.C.D., M.R.V., M.L.D.; organization of supplementary material: M.R.V.; manuscript writing: M.R.V., E.E.E., M.J.P.C.; display items: M.R.V., P.C.D.

Chapter 3. IMPROVED ASSEMBLY AND VARIANT DETECTION OF A HAPLOID HUMAN GENOME USING SINGLE-MOLECULE, HIGH-FIDELITY LONG READS

Chapter 3 is adapted with minimal modification from:

> Mitchell R. Vollger*, Glennis A. Logsdon*, Peter A. Audano, Arvis Sulovari, David Porubsky, Paul Peluso, Aaron M. Wenger, et al. 2020. "Improved Assembly and Variant Detection of a Haploid Human Genome Using Single-molecule, High-fidelity Long Reads." *Annals of Human Genetics* 84 (2): 125–40. https://doi.org/10.1111/ahg.12364.

First authorship is shared between MRV and GAL.

3.1 Abstract

The sequence and assembly of human genomes using long-read sequencing technologies has revolutionized our understanding of structural variation and genome organization. We compared the accuracy, continuity, and gene annotation of genome assemblies generated from either high-fidelity (HiFi) or continuous long-read (CLR) datasets from the same complete hydatidiform mole human genome. We find that the HiFi sequence data assemble an additional 10% of duplicated regions and more accurately represent the structure of tandem repeats, as validated with orthogonal analyses. As a result, an additional 5 Mbp of pericentromeric sequences are recovered in the HiFi assembly, resulting in a 2.5-fold increase in the NG50 within 1 Mbp of the centromere (HiFi 480.6 kbp, CLR 191.5 kbp). Additionally, the HiFi genome assembly was generated in significantly less time with fewer computational resources than the CLR assembly. Although the HiFi assembly has significantly improved continuity and accuracy in many complex regions of the genome, it still falls short of the assembly of centromeric DNA and the largest regions of segmental duplication using existing assemblers. Despite these shortcomings, our

results suggest that HiFi may be the most effective stand-alone technology for *de novo* assembly of human genomes.

3.2 Introduction

Recent advances in long-read sequencing technologies, including Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (ONT), have revolutionized the assembly of highly contiguous mammalian genomes (Bickhart et al. 2017; Chaisson et al. 2015a; Gordon et al. 2016; Huddleston et al. 2017; Jain et al. 2018b; Kronenberg et al. 2018; Low et al. 2019; Steinberg et al. 2016). For example, individual laboratories can now accurately assemble >90% of mammalian euchromatin in less than 1,000 contigs within a few months. However, the generation of high-quality datasets is costly and requires computational resources unavailable to most researchers. Long-read *de novo* assemblies of human samples typically require 20,000–50,000 CPU hours (Chin et al. 2016; Koren et al. 2017) and terabytes of data storage.

The accessibility of *de novo* assembly using single-molecule, real-time (SMRT) sequencing data has significantly improved with the recent introduction of high-fidelity (HiFi) sequence data from PacBio and the development of the SMRT Cell 8M. With 28-fold sequence coverage of the Genome in a Bottle Ashkenazim sample HG002, Wenger and colleagues demonstrated that it is possible to create a *de novo* assembly comparable to previous long-read assemblies with half the data and one-tenth the compute power (Wenger et al. 2019). While compute time and throughput have improved, there is little comparison of the HiFi assembly quality of HG002 to a previous continuous long-read (CLR) HG002 genome assembly and limited assessment of the more difficult regions of the genome.

Here, we generate 24-fold sequence coverage and produce a *de novo* assembly of a complete hydatidiform mole human genome (CHM13) with HiFi data. We directly compare it to

a previous assembly of CHM13 produced with CLR data (Kronenberg et al., 2018). The accurate assembly of the CHM13 genome is valuable for several reasons. First, due to its single-haplotype nature, it allows for better resolution of highly duplicated sequences, including segmental duplications (SDs) and tandem repeats. This 5-8% portion of the genome represents some of the most challenging regions to resolve. Second, its monoallelic nature permits the detection and unambiguous resolution of structural variants (SVs) that are crucial in disease and evolution. Finally, it allows for complete and absolute deduction of the sequence accuracy of a genome assembly [i.e., quality value (QV)] because there is only one haplotype for comparison. As a result, large-insert BAC clone sequences from the same source material can be expected to align at nearly 100% sequence identity and therefore be used to reliably compute the accuracy of different sequencing platforms and assembly approaches.

3.3 Results

3.3.1 ***Whole-genome assembly with HiFi versus CLR reads***

To assess the utility of PacBio's HiFi technology (Wenger et al. 2019) for *de novo* assembly, we set out to compare assemblies of the CHM13 genome using either HiFi (generated on the Sequel II platform) or CLR (generated on the RS II platform) data. To do this, we generated 24-fold HiFi circular consensus sequence (CCS) data from four SMRT Cells 8M. Each SMRT Cell produced, on average, 19.1 Gbp of QV > 20 sequence data (range 14-25 Gbp) with an average consensus read length of 10.9 kbp (Figure S1A). The long-read sequence data were of high quality, with an estimated 54.6% of the quality-filtered CCS reads having a QV > 30 (Figure S1B,C). The generation of HiFi data using the CCS algorithm took on average 12,500 CPU hours for each SMRT Cell 8M.

Using Canu (Koren et al. 2017) (Materials and Methods), we generated a *de novo* assembly (assembly FTP) with the HiFi CCS data (hereafter termed "HiFi assembly") and compared it to a previous FALCON assembly of CHM13 (accession GCA_002884485.1;(Kronenberg et al. 2018) generated with 77-fold CLR data (hereafter termed "CLR assembly") (Figure 3.1). The HiFi assembly required only 2,800 CPU hours, whereas the CLR assembly required more than 50,000 CPU hours. This reduction in runtime is because the correction step common to both FALCON and Canu can be skipped with adequate input read quality (Table S1). It might be expected that the shorter read length of the HiFi data (N50 10.9 vs. 17.5 kbp; Figure S1A) might lead to a less continuous assembly; however, we observed that the HiFi assembly had an N50 of 25.5 Mbp, which is comparable to the N50 of the CLR assembly (29.3 Mbp; Table 3.2, Figure 3.1). We confirmed that these results were not driven by the different assembly algorithms, but rather by the different data types, by generating additional assemblies that controlled for input coverage and assembly algorithm (Table S1, Supplemental Note).

To determine assembly base-pair accuracy, we sequenced and assembled the inserts of 31 randomly selected BACs from a genomic library produced from the CHM13 cell line (VMRC59; Materials and Methods). We estimated assembly accuracy by aligning these sequence inserts to the HiFi and CLR assemblies. We found that, before any polishing, the consensus accuracy of the HiFi assembly was much higher than the CLR assembly (median QV 40.4 vs. 27.5; Table 3.2, Figure S2). Next, we polished the CLR assembly using 77-fold coverage of CLR reads with Quiver and the HiFi assembly using 355-fold coverage of CCS subreads with Arrow. In this experiment, once again, the HiFi assembly was superior to the CLR assembly with respect to accuracy (median QV 43.3 vs. 40.7; Table 3.2, Figure S2).

While the initial assembly of the HiFi data was relatively rapid (2,800 CPU hours), subsequent polishing with Arrow required an additional 7,200 CPU hours. We were curious if we could reduce the polishing time by not incorporating subread information and using only the HiFi data. To do this, we applied Racon (Vaser et al. 2017) to polish our assembly with only the HiFi CCS reads. This Racon-based polishing step finished in only 135 CPU hours (100 for alignment and 35 for polishing) and offered improved accuracy over Arrow (median QV 45.0 vs. 43.3; Table 3.2, Figure S2). After a second round of Racon polishing, there was only one single-nucleotide difference between the HiFi assembly and the BACs excluding indels. Using Illumina WGS data as a third orthogonal platform, we determined that this difference is likely not a sequence error but rather a *bona fide* mutational change that represents a divergence between the propagated VMRC59 BAC and the CHM13 cell line (Figure S3). With the exception of remaining single-base-pair indels, this finding suggests that the QVs reported here should be considered lower bounds due to subsequent propagation errors in BAC DNA (Supplemental Note).

To evaluate the global contiguity of the respective assemblies, we generated and applied 2.8-fold sequencing data from strand-specific sequencing (Strand-seq) of the CHM13 cell line. Strand-seq is able to preserve structural contiguity of individual homologs by tracking the read directionality and, therefore, can be used for detection of misassembled contigs in *de novo* assemblies (Falconer et al. 2012; Sanders et al. 2016). Using this analysis, we detected six misassembled contigs that contain seven breakpoints in the HiFi assembly (Table S2, Figure S4). In contrast, we detected a slightly lower number of misassembled contigs (5) and breakpoints (5) in the CLR assembly (Table S2). However, given the number of assembled contigs, these results demonstrate that both assemblies are highly accurate, with <0.5% misassembly.

3.3.2 *Segmental duplication analyses*

SDs are often recalcitrant to genome assembly due to their high (>90%) sequence identity, length (>1 kbp), and complex modular organization. Therefore, the accuracy and completeness of SDs is a particularly useful metric for assembly quality since these most often correspond to the last gaps in the euchromatic portions of long-read assemblies (Chaisson et al. 2015a). We performed a number of analyses to assess the SD resolution in the HiFi and CLR assemblies (Figure 3.2). First, we compared the percentage of SDs resolved in both genome assemblies, as well as the human reference genome and several recently published long-read assemblies (Materials and Methods; (Vollger et al. 2019). Requiring that SDs are anchored contiguously with unique flanking sequence, we found that, on average, 42% of SDs are resolved in the CHM13 HiFi assembly compared to 32% in the CLR assembly (Figure 3.2). Although the majority of human SDs remain unassembled, this is the highest fraction of resolved SDs for any of the published assemblies analyzed thus far (Huddleston et al. 2017; Jain et al. 2018b; Seo et al. 2016; Shi et al. 2016), with an average 12% increase over even the ultra-long ONT assembly of NA12878 (Figure 3.2). Additionally, the number of bases with significantly elevated coverage (mean + three standard deviations) (Vollger et al. 2019) in the HiFi assembly was reduced by 15% as compared to the CLR assembly (27.3 vs. 32.1 Mbp). This indicates that the HiFi assembly has fewer collapsed sequences compared to the CLR assembly, with multiple SDs now represented by a single contig.

Next, we specifically focused on the pericentromeric regions of the genome where megabases of interchromosomal duplications have accumulated during the course of great ape evolution (She et al. 2004a, 2006). We first assessed the contiguity and coverage within the 1 Mbp regions flanking each centromere by calculating a pericentromere-specific NG50. We found that

the HiFi assembly had an NG50 of 480.6 kbp, whereas the CLR assembly had a NG50 of only 191.5 kbp (Figure 3.2). Next, we assessed contiguity within the pericentromeric regions by counting the number of contigs within the 1 Mbp region flanking the centromeres for each assembly (Figure S5A). Assemblies with fewer contigs have increased contiguity and improved assembly; therefore, we expected that the HiFi assembly would have fewer contigs within many of these regions. Indeed, we found that the HiFi assembly had reduced or the same number of contigs at 52.2% (24/46) of the 1 Mbp pericentromeric regions when compared to the CLR assembly [30.4% (14/46) of the pericentromeric regions had fewer contigs, and 21.7% (10/46) had the same number of contigs in both assemblies]. The remaining pericentromeric regions were split between having no contig representation (8.7%; 4/46) and an increased number of contigs (39.1%; 18/46) in the HiFi assembly relative to the CLR assembly. We hypothesized that the increased number of contigs in these regions in the HiFi assembly may be indicative of fragmented sequences not found in the CLR assembly (Figure S5B). When we tested this hypothesis by summing up the total contig coverage in the 1 Mbp windows flanking the centromeres, we found that, indeed, the HiFi assembly had recovered an additional 5.03 Mbp of pericentromeric sequence missing from the CLR assembly (Figure S5C).

To assess the sequence accuracy and contiguity within SD regions, we compared HiFi and CLR assemblies to 310 sequenced and assembled large-insert BAC clones of CHM13 origin. Once again, we found that the HiFi assembly is more accurate (median QV 33.5, n = 139) than the CLR assembly (median QV 31.3, n = 102) against BACs that align along at least 95% of their length (Figure 3.2). We suspect the increased QV is due to the inability of the correction step in FALCON to correctly resolve paralog-specific reads into different groups. Although the HiFi assembly has a higher QV, it should be noted that both assemblies are far less accurate for SDs than unique

regions of the genome. Additionally, we find that the HiFi-assembled contigs are more continuous within the sampled SD regions: in 253 of the 310 (82%) BACs, the alignment length to the HiFi assembly is greater than or equal to the alignment length to the CLR assembly (Figure 3.2).

A significant fraction of high-identity duplications remain collapsed and unassembled in both the CLR and HiFi assemblies. However, we recently developed a method, Segmental Duplication Assembler (SDA), that can resolve collapsed duplications by taking advantage of long reads that share multiple paralog-specific variants (PSVs) and then grouping them using correlation clustering (Vollger et al. 2019). The algorithm depends on the length of the underlying reads, and since HiFi reads are substantially shorter (N50 10.9 vs. 17.5 kbp), we were concerned that SDA would be limited. To test the ability of HiFi and CLR to resolve collapses, we selected five problematic gene-rich regions of biomedical and biological importance and directly compared the potential of correlation clustering to partition and assemble such regions (Table S3; these regions contained the genes *OPN1LW*, *NOTCH2NL*, *SRGAP2*, *FCGR2/3*, *KANSL1*). Of the five regions: two were resolved more accurately by the CLR reads (*OPN1LW, KANSL1*), one was equivalent between HiFi and CLR (*SRGAP2*), and two were better resolved by the HiFi reads (*NOTCH2NL*, *FCGR2/3*). These results are encouraging since SDA was optimized to handle CLR data (Vollger et al. 2019), and we believe future improvements to SDA that take advantage of the high-quality single-nucleotide variants embedded within the HiFi data will resolve even more collapsed regions of genomes.

3.3.3 ***Tandem repeat resolution***

Since tandem repeat sequences are often difficult to resolve for both length and content, we assessed whether short tandem repeats (STRs) and variable number of tandem repeats (VNTRs) were correctly assembled in the HiFi and CLR assemblies (Figure 3.3). We identified 3,074

tandem repeats that were ≥1 kbp, on average, across the six Human Genome Structural Variation Consortium (HGSVC) haplotype-resolved assemblies (Chaisson et al. 2019). For each locus, we compared the length of the region in the HiFi and CLR assemblies against an orthogonal set of ultra-long ONT reads generated from CHM13 (Materials and Methods). A total of 2,969 (96.6%) and 2,936 (95.5%) of the tandem repeats assembled with HiFi and CLR reads, respectively. Both HiFi and CLR assemblies had a high length concordance with ONT reads (Pearson's correlation coefficients $\rho = 0.816$ and $\rho = 0.809$, respectively) over tandem repeats that were resolved in at most a single contig by each assembly and spanned by more than one ONT read (n = 2,898). When we compared loci within each assembly to the mean length of the region in ultra-long ONT reads (with at least one spanning read) (Figure 3.3), we found that the HiFi contigs had a lower root-mean-square (RMS) error of 0.886 kbp, while the CLR contigs had an RMS error of 0.952 kbp.

Further restricting the analysis to VNTRs present in HiFi but completely absent from the CLR assembly (n = 87), 53% (n = 46) of the loci agreed in length with the ONT reads. Inversely, restricting the analysis to VNTRs present in CLR but completely absent from the HiFi assembly (n = 54), 59% (n = 32) of the loci agreed in length with the ONT reads. The N50 of the 46 validated HiFi-only tandem repeats was 4,968 bp, while the N50 for the 32 validated CLR-only tandem repeats was 3,306 bp. Additionally, the largest VNTRs resolved by HiFi and CLR assemblies were 19,397 bp and 14,250 bp, respectively. This pattern suggests that HiFi reads accurately assemble large tandem repeats that may be inaccessible to CLR. Several of these loci were genic, such as the 439 copy 15-mer in the intron of *RTEL1* (Figure 3.3) and the expansion of a 35-mer in the intron of *ZNF717* from 15 (CLR) to 89 (HiFi) tandem repeat copies (Figure 3.3). Overall, the HiFi assembly more accurately represented the content and sequence length of the tandem repeats,

particularly in previously unrepresented or collapsed regions of the CLR assembly, based on orthogonal validation experiments.

3.3.4 *Structural variant analyses*

Since errors in an assembly will lead to false-positive variant calls, we assessed the utility of assembled HiFi data as a variant discovery tool and used it as a metric to evaluate assembly quality. For each assembly, we called insertions and deletions against GRCh38 from contig alignments and filtered for consensus regions (loci where the assembly had one mapped contig, Materials and Methods). We generated a callset for each assembly before and after polishing using a variety of tools, including Racon, Quiver, Arrow, Pilon, and a FreeBayes-based indel correction pipeline (PacBio; Chin et al. 2013; Kronenberg et al. 2018; Walker et al. 2014). We found that SV (indels ≥50 bp) calls were largely consistent among assemblies (Table 3.3). Although HiFi read quality is substantially higher, polishing was required to reduce the number of false positive indel calls (Table 3.4). Overall, we found that the number of insertions and deletions was comparable between polished HiFi and CLR assemblies. When we compare SVs to published CHM13 calls, we see very strong concordance, with 89.5% of insertions and 86.8% of deletions called in both (Figure S6).

3.3.5 *Gene open reading frame annotations*

Long-read sequencing platforms exhibit high indel error rates due to missed and erroneous incorporations during real-time sequencing. As a result, predicted open reading frames are often disrupted, leading to potential problems in gene annotation (Watson and Warr 2019) unless additional error correction steps are employed (Kronenberg et al. 2018). We compared the SV and indel callsets to human RefSeq annotations and identified likely gene-disruptive events (Materials

and Methods). In the unpolished HiFi assembly, we found 16,158 SVs and indels putatively disrupting 4,151 of 18,045 RefSeq genes within the assembly consensus regions (23%), which reduced to 134 after polishing with two rounds of Racon (0.74%) (Table 3.5). Before polishing, these predicted gene-disruptive SVs and indels were overwhelmingly single-base-pair errors (98%; 15,822 of 16,158), which were greatly reduced after polishing (56%; 93 of 165). As expected, the CLR assembly had more likely disrupted genes before polishing (64%; 11,593 of 17,991 genes in its consensus region), but this declined to 209 after polishing (1.2%). We found fewer predicted disrupted genes outside of repetitive events in the HiFi assembly (53 in HiFi vs. 58 in CLR), and this trend increases inside SDs where short reads may not polish as effectively (39 in HiFi vs. 101 in CLR). It is worth noting that 2,412 protein-coding genes (13%) have exons in SDs, and this difference between the HiFi and CLR assemblies represents 2.7% of these duplicated protein-coding genes.

Since true biological variation and reference errors will contribute to gene-disrupted events, we expect many of these to be biological and not necessarily assembly artifacts. When we intersect the disrupted genes from the polished HiFi and CLR assemblies, we find that the HiFi genes are largely a subset of the CLR genes, but the converse is not true (Figure S7). To provide additional support for these events, we intersected gene-disrupting variants with CHM13 calls from SMRT-SV (Audano et al. 2019) and a FreeBayes callset from Illumina CHM13 whole-genome sequence reads (ERR1341795) (Materials and Methods). We applied this to both the polished HiFi assembly (2 times with Racon) and the fully polished CLR assembly. In the HiFi assembly, 13% (17 of 135) of the disrupted genes had no orthogonal support with the majority corresponding to duplicated genes (14 genes). We conclude that the events in these 17 genes are likely false positives; however, only three of these remaining unsupported gene-disrupting indels

mapped to unique sequence. In the CLR assembly, 44% (93 of 209) of the gene-disruptive events had no orthogonal support with the majority (80 genes) mapping to SDs. These experiments suggest that there are approximately 120 genes in CHM13 altered by *bona fide* frame-shifting indels and SVs when compared to GRCh38 and RefSeq annotations.

3.4 Discussion

The generation and assembly of HiFi and CLR long-read sequence data from the same haploid source material allows us to directly compare the accuracy and contiguity of these technologies without the added complication of disentangling haplotypes needed to resolve SV alleles. We conclude that there are three key strengths of the HiFi technology over CLR technology. First, the time to generate the *de novo* assembly is reduced 10-fold, and it will likely be reduced further as HiFi assemblers are developed and optimized. This not only makes *de novo* assembly of human genomes accessible to a larger number of research groups, but it also paves the ways for larger cohorts of individuals to be sequenced and assembled. Although assembly time is drastically reduced, the background compute time required to generate HiFi data by the CCS algorithm remains substantial (~50,000 CPU hours in total).

Second, our analyses confirm that, both in terms of quality and continuity, the HiFi assembly is generally superior or at least comparable to the CLR assembly despite the shorter read lengths and effectively reduced genome coverage (Wenger et al. 2019). One significant advance is that HiFi assembly can be polished without reverting to the underlying subreads, which saves approximately 1 terabyte of subread data and 7,000 hours of additional compute time. Polishing remains an absolute requirement to reduce indel errors and obtain a high-quality final assembly. Human CLR datasets ultimately require orthogonal Illumina data, and our results show that the

HiFi sequencing platform alone achieves a greater level of accuracy for annotated protein-coding genes.

Finally, we demonstrate that, in some of the most difficult regions of the genome (i.e., SDs, pericentromeric regions, and tandem repeats), the HiFi assembly shows improved continuity and representation, but relatively modest accuracy improvements. Highly accurate HiFi data allows for the assembly of an additional 10% of duplicated sequences and better recovers the structure of tandem repeats such that they more exactly reflect the genomic length of VNTRs and STRs as confirmed by orthogonal analyses. We note, however, that the accuracy of the duplicated and tandem repeat regions is still lower than that of unique regions of the genome. Follow-up procedures such as SDA, which are designed to target and further resolve collapsed regions, show mixed results especially among the most highly identical human duplications. Our analyses suggest that this is a limitation of the shorter read lengths of HiFi (N50 of 10.9 vs. 17.5 kbp), which reduces the power needed to phase PSVs and assign collapsed reads to their respective duplicated loci. Nevertheless, we believe the results are encouraging since methods such as SDA were optimized to handle CLR data (Vollger et al. 2019). Future improvements to SDA that take advantage of the high-quality single-nucleotide variants embedded within the HiFi data in duplicated regions will resolve even more collapsed regions of assembled genomes.

Because of these three strengths, we conclude that HiFi technology is currently the best choice for *de novo* genome assembly when speed, quality, and resolution of repetitive sequences are priorities. Additionally, there is currently no other single technology available that can accurately recreate genes models and confidently call diverse types of genetic variation, from large SVs down to single-nucleotide variants (Wenger et al. 2019).

Next steps involve benchmarking and optimization of performance within diploid genome assemblies. Much of the recent advances in improving the contiguity of genome assemblies from telomere to telomere (Miga et al. 2020) have been based on the same haploid source material analyzed here. It is clear that current HiFi genome assemblies are not as contiguous as those generated with high-coverage, ultra-long ONT data, or with combinations of PacBio and ONT data. While the haploid source material has been extremely useful for benchmarking, the ultimate challenge is the accurate assembly of human diploid genomes where both chromosomal haplotypes are fully resolved. Incorporation of linking-read technologies, such as Strand-seq, Hi-C, and 10x Genomics, or trio-binning approaches have been shown to significantly improve phasing and SV sequence and assembly (Chaisson et al. 2019; Koren et al. 2017, 2018; Kronenberg et al. 2018, 2019). It is likely that such approaches could be combined with HiFi datasets to enhance telomere-to-telomere phasing and improve the accuracy of more complex repeats. Alternatively, the use of ultra-long-read datasets coupled with HiFi sequencing on the same samples will likely enhance both the phasing and accuracy of diploid genome assemblies. A useful standard for diploid genome assembly will be to repeat these analyses for two haploid source genomes in order to model the effect and accuracy of *in silico* diploid genomes as we (Huddleston et al. 2017) and others (Li et al. 2018) have shown.

Notwithstanding these advances, significant challenges remain for complete genome assembly, including large SDs, centromeric satellites, and acrocentric regions. For example, although the CHM13 HiFi assembly we generated is highly contiguous (N50 25.5), an analysis of the unmappable reads shows an abundance of repetitive DNA (70.4%; Figure S8). Of these sequences, 49.5% consist of various classes of satellite repeats, which populate centromeres and the acrocentric portions of human chromosomes. Given the accuracy of these unmapped sequence

reads, they will be quite valuable in obtaining the first overview of the sequence content and composition of these more complex heterochromatic regions. Obtaining even longer HiFi reads than used in this assembly (i.e., >11 kbp average used here) will be necessary to accurately anchor and sequence-resolve these repeat regions in future genome assemblies. Coupled with advances from other long-read technologies, such as ONT, it is clear that highly accurate telomere-to-telomere assemblies of diploid genomes will soon be achievable.

3.5 Methods

3.5.1 *Cell lines*

Cells from a complete human hydatidiform mole, CHM13 (46X,X), were immortalized with human telomerase reverse transcriptase (hTERT) and cultured in complete AmnioMAX C-100 Basal Medium (ThermoFisher Scientific, Carlsbad, CA) supplemented with 15% AmnioMAX supplement (ThermoFisher Scientific, Carlsbad, CA) and 1% penicillin and streptomycin. Cells were maintained at 37°C in a humidified incubator with 5% CO_2.

3.5.2 *CCS library preparation*

High-molecular-weight DNA was isolated from cultured CHM13 cells using a modified Qiagen Gentra Puregene Cell Kit protocol (Huddleston et al. 2014). A HiFi library with an average insert length of ~11 kbp was generated according to the protocol in (Wenger et al. 2019) and sequenced on four SMRT Cells 8M on a Sequel II instrument using Sequel II Sequencing Chemistry 1.0, 12-hour pre-extension, and 30-hour movies. Raw data was processed using the CCS algorithm (version 3.4.1, parameters: --minPasses 3 -- minPredictedAccuracy 0.99 -- maxLength 21000) to yield 75.7 Gbp in 6.9 million reads with an average read length of 10.9 kbp and estimated median QV of 32.85. Sequence data is available via NCBI SRA

(https://www.ncbi.nlm.nih.gov/sra/SRX5633451). Average run time for the CCS algorithm was ~12,500 CPU core hours per SMRT Cell (~50,000 total).

3.5.3 ***Strand-seq library preparation***

Cultured CHM13 cells were pulsed with BrdU and used for preparation of single-cell Strand-seq libraries as previously described (Sanders et al. 2016, 2017).

3.5.4 ***BAC clone insert sequencing***

BAC clones from the VMRC59 clone library were hybridized with probes targeting complex or highly duplicated regions of GRCh38 (n = 310), or selected from random regions of the genome not intersecting with SD (n = 31). DNA from positive clones was isolated, screened for genome location, and prepared for long-insert PacBio sequencing as previously described (Vollger et al. 2019). Libraries were sequenced on the PacBio RS II and Sequel platforms with the P6-C4 or Sequel 2.1/Sequel 3.0 chemistries, respectively. We performed *de novo* assembly of pooled BAC inserts using Canu v1.5 (Koren et al. 2017). After assembly, we removed vector sequence (pCCBAC1), restitched the insert, and then polished with Quiver or Arrow. Canu is specifically designed for assembly with long error-prone reads, whereas Quiver/Arrow is a multi-read consensus algorithm that uses the raw pulse and base call information generated during SMRT sequencing for error correction. We reviewed PacBio assemblies for misassembly by visualizing the read depth of PacBio reads in Parasight (http://eichlerlab.gs.washington.edu/jeff/parasight/index.html), using coverage summaries generated during the resequencing protocol.

3.5.5 *Genome assembly*

Canu v1.7.1 was applied with the following parameters to generate the HiFi *de novo* assembly:

```
genomeSize=3.1g correctedErrorRate=0.015
ovlMerThreshold=75 batOptions="-eg 0.01 -eM 0.01 -dg
6 -db 6 -dr 1 -ca 50 -cp 5" -pacbio-corrected
```

Assemblies were mapped to GRCh38 with minimap2 (Li 2018) version 2.15 using the following parameters:

```
--secondary=no -a --eqx -Y -x asm20 -m 10000 -z
10000,50 -r 50000 --end-bonus=100 -O 5,56 -E 4,1 -B
5. These alignments were used for downstream SV
calling and ideogram visualizations.
```

Error correction with Quiver, Arrow, Pilon, and indel correction was done as previously described (Chin et al. 2013; Kronenberg et al. 2018; Vaser et al. 2017; Walker et al. 2014). Error correction with Racon was executed with the following steps:

```
minimap2 -ax map-pb --eqx -m 5000 -t {threads} --
secondary=no {ref} {fastq} | samtools view -F 1796 -
> {sam}
racon {fastq} {sam} {ref} -u -t {threads} >
{output.fasta}
```

3.5.6 *QV calculations*

QV calculations were made by alignments to 31 sequenced and assembled BACs falling within unique regions of the genome (>10 kbp away from the closest SD) where at least 95% of the BAC sequence was aligned. The following formula was used to calculate the QV, and gaps of size N were counted as N errors: $QV = -10\log_{10}[1 - (\text{percent identity}/100)]$. QV calculations within SDs were done in the same manner but against 310 BACs that overlap with SD regions.

3.5.7 *SD analyses*

SDs were defined as resolved or unresolved based on their alignments to GRCh38 using the minimap2 parameters described above. Alignments that extended a minimum number of base pairs beyond the annotated SDs were considered to be resolved. This minimum extension varied from -10,000 to 50,000 bp and the average difference between assemblies was used to define the percent difference reported.

The number of collapsed bases was determined by aligning the CLR reads to both the CLR and the HiFi assemblies. Regions were defined as collapsed if they met the following conditions: coverage greater than the mean coverage plus three standard deviations, 15 kbp of consecutive increased coverage or more, and <80% repeat content as defined by RepeatMasker.

3.5.8 *Pericentromeric analyses*

The number of contigs within each pericentromeric region was calculated by first aligning the contigs from the HiFi or CLR assemblies to GRCh38 using the minimap2 parameters described above. Alignments were limited to be within 1 Mbp on either side of the centromere decoys, and then unique contig names were counted.

The representation within the pericentromeric regions was calculated using BEDTools (Quinlan and Hall 2010) to collapse all filtered contigs within the pericentromeric region for the HiFi and CLR assemblies. The resulting size of the collapsed contigs within the CLR assembly was subtracted from the size calculated in the corresponding region in the HiFi assembly.

The pericentromere-specific NG50 statistic was calculated using a G of 46 Mbp (accounting for the 1 Mbp size of each pericentromeric region on the 23 chromosomes).

3.5.9 *Tandem repeat analyses*

Tandem Repeats Finder (Benson 1999) was run on the six haplotype-resolved assemblies (Chaisson et al. 2019) as well as the CLR CHM13 assembly using the following parameters: 2 7 7 80 10 50 2000 -h -d -ngs. After identifying all tandem repeats not represented or collapsed in the CLR assembly relative to the six human haplotypes, we obtained a final set of 3,074 large tandem repeats, all of which were anchored in GRCh38. Second, we retrieved sequence from each of these loci using the two assemblies and our orthogonal CHM13 ONT data source. For each region in both assemblies and aligned ultra-long ONT reads, we extracted the sequence that mapped from the start of the region to the end using the alignment CIGAR strings as a guide. Since multiple sequences may map to a region, we recorded the number of alignments and computed the average length of the region for each dataset. Concordance with ONT reads was defined by allowing ≤5% variation in the average ONT read length. For our in-depth sequence analysis of the two VNTR loci, we used repeat homology plots, which were constructed using a pairwise alignment between the motif and assembled sequence in every tiling window of the same length as the repeat unit length (i.e., 15 bp and 53 bp, respectively, for the two VNTRs; Figure 3.3). At any given window, the repeat unit (i.e., the motif) was circularized in 1 bp increments, and the maximal sequence identity was reported at each tiling window. The dotplots were generated using Gepard (Krumsiek et al. 2007)

3.5.10 *SV analyses*

For assembly in each polishing stage, contigs mapped to GRCh38 were used to create a consensus region, which included all loci with exactly one aligned contig. Next, we called indels and SVs from the alignments using a previously validated method (Chaisson et al. 2015a) implemented in PrintGaps.py distributed in the SMRT-SV v2 pipeline

(https://github.com/EichlerLab/smrtsv2). We then filtered for variants within the assembly's consensus region. We further filtered out variants in pericentromeric loci where callsets are difficult to reproduce (Audano et al. 2019). This process was repeated for each assembly in each polishing stage.

For gene annotations, SVs were intersected with a callset from SMRT-SV and FreeBayes. For the SMRT-SV indels, we retrieved the CHM13 contigs, called SVs and indels from them using the same PrintGaps.py method. SMRT-SV generates a BED file linking regions of GRCh38 to the best contig for variant calling, and we used this BED to filter the SV and indel calls from the overlapping assembly contigs. We then intersected HiFi and CLR variants with either SMRT-SV or FreeBayes SVs and indels using custom code that requires either a variant length match by 50% and maximum distance between events is no more than 50 bp or 50% reciprocal overlap. Matching by size and distance reduces overlap bias for short indels while matching by reciprocal overlap allows larger SVs to intersect even when they are shifted, which is common for calling insertions associated with tandem duplications or repetitive sequence.

3.5.11 ***Gene annotation***

With custom code using the SV and indel callset, the number of bases in coding regions of RefSeq annotations (retrieved 2019-04-24 from UCSC RefSeq track on GRCh38) were quantified. Briefly, if an insertion was located in a coding region, its entire length was taken as the number of coding bases it affects. For deletions, the number of bases falling inside the coding region were quantified. From these results, we obtained a set of genes where at least one variant inserts or deletes a number of bases that is not a multiple of three within any isoform of the gene. For this analysis, we excluded RefSeq noncoding RNA annotations.

We intersected RefSeq exons with tandem repeats (UCSC hg38 “Simple Repeats” track) and SDs (UCSC hg38 “Segmental Dups” track) to annotate them as either containing or absent of SDs or tandem repeats. For each assembly, we calculated results using only RefSeq genes that are fully contained within its consensus region.

3.5.12 ***RepeatMasker analysis of unmappable sequences***

All HiFi sequence reads were mapped to the *de novo* assemblies using the following minimap2 parameters: -x asm20 -m 4000 --secondary=no --paf-no-hit. Reads that did not map to the *de novo* assemblies were subjected to RepeatMasker analysis (Smit, Hubley, & Green, 1996) to determine their repeat content.

3.5.13 ***Data Access***

HiFi assemblies with varying levels of polishing are available here: https://eichlerlab.gs.washington.edu/help/mvollger/papers/chm13_hifi/rebasecalled/HiFi_Asms/. CLR assemblies with varying levels of polishing are available here: https://eichlerlab.gs.washington.edu/help/mvollger/papers/chm13_hifi/rebasecalled/CLR_Asms/. HiFi sequence data (SRX5633451), CLR sequence data (SRX818607, SRX825542, and SRX825575-SRX825579), and assembled BACs from the VMRC59 clone library are available via NCBI SRA.

3.6 Figures

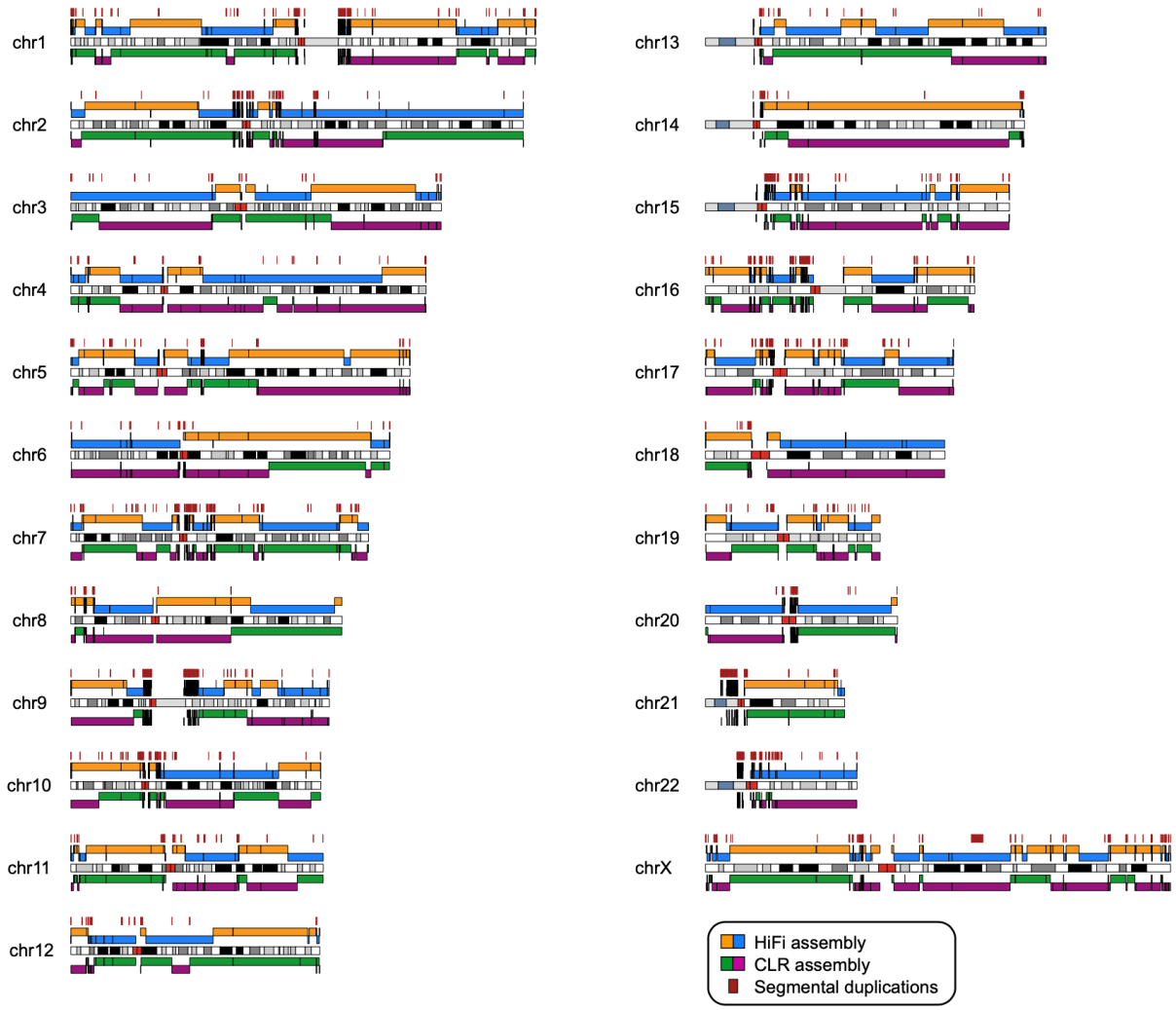


Figure 3.1. Comparison between the CHM13 HiFi and CLR genome assemblies.

Shown are alignments of the HiFi assembly (blue and orange) and the CLR assembly (green and purple) to GRCh38, as well as segmental duplication (SD) blocks greater than 25 kbp in length (dark red) projected onto a karyotype (chromosome banding is indicated in white, black, and gray, with centromeres in bright red and acrocentric regions in blue-gray; CHM13 has a 46X,X karyotype). The alignments are colored by contig name such that when the contig name changes, so does the alignment color. Black bars within a solid color block represent a break in the alignment within the same contig name, which are likely to be locations of structural variants between CHM13 and GRCh38. The large majority of contig alignments over 100 kbp in length end within 50 kbp of an SD [158/166 (95%) in HiFi and 177/182 (97%) in CLR].

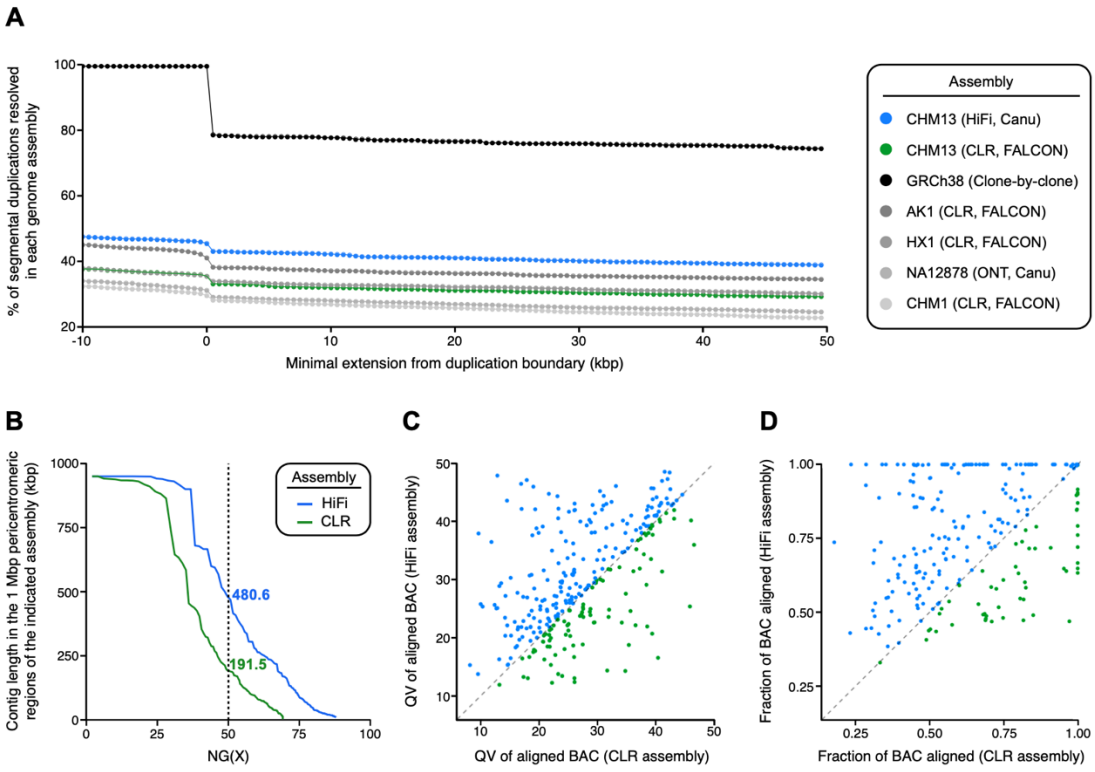


Figure 3.2. Segmental duplication resolution in the HiFi and CLR genome assemblies.

A) Shown is the percent of resolved SDs as defined in GRCh38 across the indicated de novo assemblies. To be considered resolved, the alignment of the de novo assembly must extend X number of base pairs beyond the annotated duplication block on either side. GRCh38 is not 100% resolved after a minimum extension of zero base pairs because many SDs in GRCh38 are flanked by gaps. B) Shown is the NG(X) of the HiFi and CLR assemblies in the 1 Mbp regions flanking the centromeres. NG(X) is defined as the sequence length of the shortest contig at X% of the total pericentromeric region length, which is 46 Mbp (1 Mbp for each pericentromere). The HiFi assembly has an NG50 2.5-fold greater than the CLR assembly in these regions. C) Plot of the quality value (QV) score for each of 310 BACs aligning to SDs within the HiFi and CLR assemblies. Data points above the dashed line have a higher QV score, and, therefore, better sequence identity, in the HiFi assembly relative to the CLR assembly. The accuracy of the HiFi assembly within SDs (median QV 33.5) is increased compared to the CLR assembly (median QV 31.3). D) Plot of the fraction of each of 310 BACs aligning to the HiFi and CLR assemblies. Data points above the dashed line have a higher alignment length in the HiFi assembly relative to the CLR assembly. In 253 of the 310 (82%) BACs, the alignment length to the HiFi assembly is greater than or equal to the alignment length in the CLR assembly.

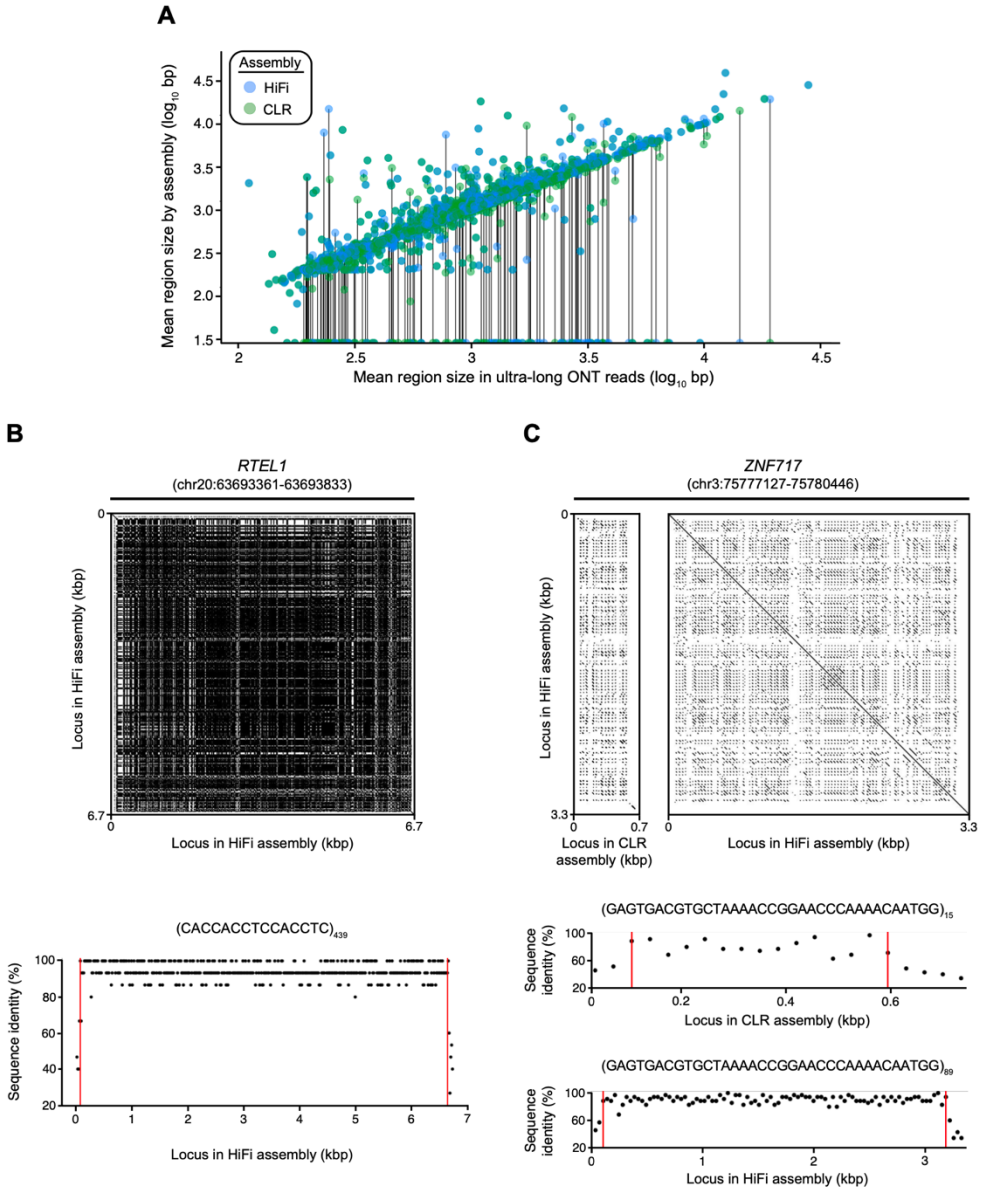


Figure 3.3. Tandem repeat resolution in the HiFi and CLR genome assemblies.

A) Plot of the length of tandem repeat loci in the HiFi and CLR assemblies vs. the mean size of these loci in ultra-long CHM13 ONT reads. Discordancy between HiFi and CLR assemblies map off the diagonal, with dropouts clustering as points along on the horizontal axis. For this plot, we include only regions with more than one spanning ONT read and no more than one spanning contig in either assembly (n = 2,898 regions). B) Dot plot of a 6.7 kbp VNTR in the intron of RTEL1 (chr20:63693361-63693833) (top panel), which was resolved in the HiFi assembly only. The CLR assembly contained a gap over this region. The overall structure and length of this VNTR was supported by the ONT reads mapping to this location, which averaged 5,956 +/- 1799 bp (n = 5 ultra-long ONT reads), placing the HiFi sequence length at <1 standard deviation away from the average ONT read. The motif homology plot (bottom panel) indicates that the content of the RTEL1 VNTR is relatively pure, with an average sequence identity to the 15-mer repeat unit of 94.49%

across the 439 copies. C) Dot plot of the zinc finger protein gene ZNF717 (GRCh38 coordinates: chr3:75777127-75780446) (top panel), which was collapsed in the CLR assembly but fully represented in the HiFi assembly. The number of copies of this 35 bp repeat unit increased from 15 in the CLR assembly to 89 in the HiFi assembly. The large amount of variation between individual copies of this VNTR is shown in the region between the red lines in the motif homology plots (bottom panels). The level of purity within the VNTR increased from 80.38% sequence identity in the CLR assembly to 90.75% sequence identity in the HiFi assembly. The red vertical lines indicate the start and end position of the VNTR.

3.7 Tables

Table 3.2. Statistics of the HiFi and CLR genome assemblies

	Polisher	Total size (Gbp)	N50 (Mbp)	No. of contigs	Median QV	No. of CPU hours for assembly
HiFi						
CHM13 genome						
Canu assembly	None	3.03	25.51	5,296	40.41	~2,800
	Arrow	3.03	25.51	5,296	43.29	~10,000
	Racon	3.03	25.51	5,296	44.95	~2,950
	2x Racon	3.03	25.51	5,296	45.25	~3,100
	2x Racon+	3.03	25.51	5,296	45.25	~4,200
CLR						
CHM13 genome						
FALCON assembly	None	2.88	29.26	1,916	27.49	>50,000
	Quiver	2.88	29.26	1,916	40.73	>55,000
	Quiver+	2.88	29.26	1,916	42.70	>55,000
Assemblies available for comparison						
HiFi						
HG002 genome						
FALCON assembly[#]	None	2.89	29.07	2,541	Not reported*	~2,650
ONT						
NA12878 genome						
Canu assembly[†]	Nanopolish+	2.87	7.67	2,337	Not reported*	~151,000
CLR/ONT						
CHM13 genome						
Canu assembly[‡]	Multitechnology	2.93	71.70	590	42.2	Not reported

HiFi: HiFi assembly [24-fold sequencing depth]
CLR: CLR assembly [77-fold sequencing depth]
2x Racon: Two rounds of Racon
2x Racon+: Two rounds of Racon and one round of Pilon
Quiver+: Quiver, Pilon, and FreeBayes-based indel correction
Nanopolish+: One round of Nanopolish and one round of Pilon
Multitechnology: Two rounds of Racon, two rounds of Nanopolish, two rounds of Arrow, and one round of Long Ranger
Median QV: Median QV over 31 BACs
*The median QV was not reported using a BAC-based formula for these diploid genomes
[#]Wenger et al. 2019
[†]Jain et al. 2018
[‡]Miga et al. 2019

Table 3.3. Summary of SV calls in the HiFi and CLR assemblies

		Insertions			Deletions			All		
	Polishing	**No. of events**	**Mean length (bp)**	**Total length (bp)**	**No. of events**	**Mean length (bp)**	**Total length (bp)**	**No. of events**	**Mean length (bp)**	**Total length (bp)**
HiFi										
CHM13 genome										
Canu assembly	None	10,650	569	6,063,301	6,254	482	3,012,598	16,904	537	9,075,899
	Arrow	10,608	570	6,050,398	6,243	482	3,009,027	16,851	538	9,059,425
	Racon	10,632	569	6,044,348	6,273	478	3,000,394	16,905	535	9,044,742
	2x Racon	10,603	570	6,048,876	6,273	479	3,005,723	16,876	537	9,054,599
	2x Racon+	10,579	571	6,044,564	6,468	475	3,072,543	17,047	535	9,117,107
CLR										
CHM13 genome										
FALCON assembly	None	10,655	558	5,947,788	6,405	471	3,018,503	17,060	526	8,966,291
	Quiver	10,664	559	5,959,522	6,497	476	3,095,325	17,161	528	9,054,847
	Quiver+	10,627	560	5,950,702	6,992	469	3,275,883	17,619	524	9,226,585
Assemblies available for comparison										
HiFi										
HG002 genome										
FALCON assembly[#]	None	11,093	567	6,285,361	6,691	417	2,791,417	17,784	510	9,076,778
ONT										
NA12878 genome										
Canu assembly[†]	Nanopolish+	7,578	578	4,382,730	55,354	250	13,818,995	62,932	289	18,201,725
CLR/ONT										
CHM13 genome										
Canu assembly[‡]	Multitechnology	10,878	599	6,513,259	6,549	497	3,257,052	17,427	561	9,770,311

HiFi: HiFi assembly
CLR: CLR assembly
2x Racon: Two rounds of Racon
2x Racon+: Two rounds of Racon and one round of Pilon
Quiver+: Quiver, Pilon, and FreeBayes-based indel correction
Nanopolish+: One round of Nanopolish and one round of Pilon
Multitechnology: Two rounds of Racon, two rounds of Nanopolish, two rounds of Arrow, and one round of Long Ranger
SV: indel ≥ 50 bp
Excludes SVs mapping to pericentromeric regions (see **Materials and Methods**)
[#]Wenger et al. 2019
[†]Jain et al. 2018
[‡]Miga et al. 2019

Table 3.4. Summary of indels in the HiFi and CLR assemblies

		Insertions				Deletions			
	Polishing	**No. of events**	**% of 1 bp events**	**Mean length (bp)**	**Total length (bp)**	**No. of events**	**% of 1 bp events**	**Mean length (bp)**	**Total length (bp)**
HiFi									
CHM13 genome									
Canu assembly	None	1,014,192	80%	1.87	1,891,600	1,340,486	86%	1.69	2,269,778
	Arrow	339,429	50%	3.46	1,172,772	345,678	48%	3.66	1,264,044
	Racon	343,927	50%	3.44	1,182,117	342,106	48%	3.67	1,254,412
	2x Racon	343,733	50%	3.44	1,181,197	339,831	48%	3.68	1,251,757
	2x Racon+	343,132	50%	3.44	1,179,652	339,787	48%	3.69	1,252,463
CLR									
CHM13 genome									
FALCON assembly	None	943,936	75%	1.99	1,860,855	3,616,964	82%	1.46	5,271,942
	Quiver	350,924	50%	3.41	1,196,221	509,229	62%	2.87	1,460,402
	Quiver+	353,245	50%	3.39	1,198,153	392,657	52%	3.41	1,337,638
		Assemblies available for comparison							
HiFi									
HG002 genome									
FALCON assembly[#]	None	3,436,790	92%	1.30	4,460,278	3,956,047	94%	1.26	4,984,037
ONT									
NA12878 genome									
Canu assembly[†]	Nanopolish+	1,308,650	83%	1.52	1,983,468	5,929,008	86%	1.44	8,522,029
CLR/ONT									
CHM13 genome									
Canu assembly[‡]	Multitechnology	371,940	52%	3.29	1,225,435	376,524	51%	3.47	1,308,150

HiFi: HiFi assembly
CLR: CLR assembly
2x Racon: Two rounds of Racon
2x Racon+: Two rounds of Racon and one round of Pilon
Quiver+: Quiver, Pilon, and FreeBayes-based indel correction
Nanopolish+: One round of Nanopolish and one round of Pilon
Multitechnology: Two rounds of Racon, two rounds of Nanopolish, two rounds of Arrow, and one round of Long Ranger
SV: indel ≥ 50 bp
Excludes SVs mapping to pericentromeric regions (see **Materials and Methods**)
[#]Wenger et al. 2019
[†]Jain et al. 2018
[‡]Miga et al. 2019

Table 3.5. Summary of disrupted RefSeq gene models in the HiFi and CLR assemblies

	Polishing	No. of events in whole genome	No. of events in whole genome excluding TRs/SDs	No. of events in SDs only
HiFi				
CHM13 genome				
Canu assembly	None	4,151	2,481	360
	Arrow	138	55	39
	Racon	154	65	40
	2x Racon	135	54	39
	2x Racon+	134	53	39
CLR				
CHM13 genome				
FALCON assembly	None	11,593	6,686	1,249
	Quiver	653	261	159
	Quiver+	209	58	101
Assemblies available for comparison				
HiFi				
HG002 genome				
FALCON assembly[#]	None	14,369	8,526	1,462
ONT				
NA12878 genome				
Canu assembly[†]	Nanopolish+	14,384	8,413	1,595
CLR/ONT				
CHM13 genome				
Canu assembly[‡]	Multitechnology	183	63	70

HiFi: HiFi assembly
CLR: CLR assembly
2x Racon: Two rounds of Racon
2x Racon+: Two rounds of Racon and one round of Pilon
Quiver+: Quiver, Pilon, and FreeBayes-based indel correction
Nanopolish+: One round of Nanopolish and one round of Pilon
Multitechnology: Two rounds of Racon, two rounds of Nanopolish, two rounds of Arrow, and one round of Long Ranger
No. of events in whole genome: All RefSeq gene models within the assembly consensus regions were counted. Total gene count is 18,045 (HiFi assembly) and 17,991 (CLR assembly).
No. of events in whole genome excluding TRs/SDs: All RefSeq gene models within the assembly consensus were counted except for those with exons intersecting tandem repeats (TRs) or segmental duplications (SDs). Total gene count is 10,853 (HiFi assembly) and 10,850 (CLR assembly).
No. of events in SDs only: Only RefSeq gene models within SDs were counted. Total gene count is 2,005 (HiFi assembly) and 1,951 (CLR assembly).
[#]Wenger et al. 2019
[†]Jain et al. 2018
[‡]Miga et al. 2019

3.8 Acknowledgments

The authors thank T. Brown for assistance in editing this manuscript. This work was supported, in part, by grants from the U.S. National Institutes of Health (NIH grants HG002385 and HG010169 to E.E.E.) and an Advanced Grant from the European Research Council (P.M.L.). M.R.V. was supported by a National Library of Medicine (NLM) Big Data Training Grant for Genomics and Neuroscience (5T32LM012419-04). A.S. was supported by a National Human Genome Research Institute (NHGRI) Training Grant (5T32HG000035-23). E.E.E. is an investigator of the Howard Hughes Medical Institute.

3.9 Author Contributions

M.R.V., G.A.L., and E.E.E. wrote the manuscript; M.R.V., G.A.L., P.A.A., A.S., and D.P. produced the display items; M.R.V. performed the assembly and polishing with suggestions from Z.N.K. and A.M.W.; M.R.V. and A.M.W. performed the QV analysis; D.P. performed the Strand-seq analysis; A.D.S, D.C.J.S, and P.M.L generated the Strand-seq data; M.R.V. and G.A.L. performed the SD analyses; A.S. and P.A.A. performed the tandem repeat analysis; P.A.A. performed the SV and gene annotation analyses; G.A.L. performed the unassembled sequence analysis; M.R.V. and G.A.L. organized the supplementary material; P.P., G.T.C., K.M.M., C.B., and M.W.H. generated the PacBio genome sequence data; U.S. developed and supplied the homozygous CHM13hTERT cell line.

Chapter 4. SEGMENTAL DUPLICATIONS AND THEIR VARIATION IN A COMPLETE HUMAN GENOME

Chapter 4 is a manuscript that will soon be submitted:

Mitchell R. Vollger, Xavi Guitart, Philip C. Dishuck, Ludovica Mercuri, William T. Harvey, Ariel Gershman, Mark Diekhans, Arvis Sulovari, Katherine M. Munson, Kendra Hoekzema, Alexandra M. Lewis, David Porubsky, Ruiyang Li, Sergey Nurk, Sergey Koren, Karen H. Miga, Adam Phillippy, Winston Timp, Mario Ventura, Evan E. Eichler

4.1 Abstract

Despite their importance in disease and evolution, highly identical segmental duplications (SDs) have been among the last regions of the human reference genome (GRCh38) to be finished. Based on a complete telomere-to-telomere human genome (T2T CHM13), we present the first comprehensive view of human SD organization. Nearly one third of the additional sequence in T2T CHM13 corresponds to SDs, increasing the genome-wide estimate from 5.4 to 7.0% (218 Mbp). We identify novel regions of genomic instability, locate methylation differences between SD clusters, and predict 182 protein coding gene models in the additional sequence. We find that 91% of the new T2T CHM13 SD sequence (68.3 Mbp) better represents human copy number based on analysis of 266 human genomes. We find that 63% (35.11/55.7 Mbp) of acrocentric chromosomes consist of SDs distinct from rDNA and satellite sequences. Acrocentric SDs are 1.75-fold longer (p=0.00034) than other SDs, are frequently shared with autosomal pericentromeric regions, and are heteromorphic among human chromosomes. Comparing long-read assemblies from other human (n=12) and nonhuman (n=6) primate genomes, we use the T2T CHM13 to systematically reconstruct the evolution and structural haplotype diversity of biomedically relevant (*LPA*, *SMN*) and duplicated (*TBC1D3*, *SRGAP2C*, *ARHGAP11B*) genes

important in the expansion of the human frontal cortex. The analysis reveals unprecedented patterns of structural heterozygosity and massive evolutionary differences between man and his closest living relatives.

4.2 Introduction

Genomic duplications have long been recognized as important sources of structural change and gene innovation (Ohno et al. 1968; Ohno 1970). In humans, for example, the most recent and highly identical sequences (>90%) referred to as segmental duplications (SDs) (Bailey et al. 2001) promote meiotic unequal crossover events contributing to recurrent rearrangements associated with ~5% of developmental delay and autism (Cooper et al. 2011). These same SDs are reservoirs for human-specific genes important in increasing synaptic density and the expansion of the frontal cortex since humans diverged from other ape lineages (Dennis et al. 2012; Florio et al. 2018; Ju et al. 2016; Fiddes et al. 2018). SDs are also enriched ~10-fold for normal copy number variation although most of this genetic diversity has yet to be fully characterized or associated with human phenotypes (Sudmant et al. 2010, 2015a). Their length (frequently >100 kbp), sequence identity, and extensive structural diversity among human haplotypes have significantly hampered our ability to characterize these regions at a genetic level because sequence reads have been insufficiently long and human haplotypes too structurally diverse to resolve duplicate copies or distinguish allelic variants. One of the first whole-genome sequence (WGS) assembly drafts from human based on Sanger sequence reads was almost completely devoid of SDs and their underlying genes (Venter et al. 2001; She et al. 2004b). Similarly, BAC-based approaches to assemble the human genome from different haplotypes led to many misjoins creating *de facto* gaps that took years to resolve (Lander et al. 2001). While combining WGS- and BAC-based data from the first human genomes provided a road map of the SD landscape (Bailey et al. 2002), more than 50% of

the reference gaps within the human reference genome have corresponded to regions of complex SDs. The development of genomic resources (Eichler et al. 2002; Fredman et al. 2004; Chaisson et al. 2015a), including BAC libraries and long-read sequence data from complete hydatidiform moles (which represents a single human haplotype), was motivated in large part by efforts to resolve the organization of these regions and concomitantly complete the human reference genome. Here we present the most complete view of SDs in a human genome and highlight their importance in completing our understanding of human genetic diversity, evolution, and disease.

4.3 Results

4.3.1 ***SD content and organization.***

We characterized the SD content of the T2T CHM13 assembly based on sequence read-depth and pairwise sequence alignments (>90% and >1 kbp) (Numanagic et al. 2018). Our analysis of the assembly identifies 208 Mbp of nonredundant segmentally duplicated sequence within chromosome-level scaffolds, compared to just 167 Mbp in the current reference (GRCh38) (Table 4.6, Figure 4.1). This raises the percent estimate of the human genome that is segmentally duplicated from 5.4% to 6.7%. Five gaps remain in the current T2T CHM13 assembly. Each corresponds to a cluster of tandemly repeated rDNA genes on each acrocentric chromosome where we confirm long-read sequence pileups representing the last unresolved SDs of the human genome. To estimate the amount of missing duplicated rDNA sequence, we applied digital droplet PCR (Bell et al. 2018). Assuming a canonical repeat length of 45 kbp for the rDNA molecule (Gonzalez and Sylvester 1995; Kim et al. 2018), we approximate that there are ~10 Mbp and 223 copies of unresolved rDNA sequence (Methods). Including this, the overall SD content of the human genome is 7.0% (see Table 4.6 for statistics breakdown by SD type).

One third (81.3 Mbp) of SDs are new or differ structurally when comparing the T2T CHM13 assembly to GRCh38. Most of these involve large, high-identity SDs. For example, there is a 70% increase (41,289/24,280) in the number of SD pairs and a doubling of the number of bases in pairwise alignments with greater than 95% identity (Figure 4.1c). Among these new or variable SDs, 13,258 (35.04 Mbp) map to the acrocentric short arms of chromosomes 13, 14, 15, 21, and 22 (Figure 4.1b, Table 4.6), which are assembled for the first time. These SDs do not correspond to rDNA duplications but are predominantly shared among acrocentric (n=5,332 alignments) and the pericentromeric regions of specific non-acrocentric chromosomes (n= 5,500 alignments corresponding to 14.9 Mbp of SDs (Table S1). In particular the pericentromeric regions of chromosomes 1, 3, 4, 7, 9, 16 and 20 show the most extensive SD homology with acrocentric portions (Figure 4.1b). We find that non-rDNA acrocentric SDs are 1.75-fold longer than all other SDs (N50: 74,704 vs. 42,842) and significantly longer than any other defined category of SD [intrachromosomal, interchromosomal, pericentromeric, and telomeric (Figure S1)].

We annotated all T2T CHM13 SDs using DupMasker (Jiang et al. 2008), which defines ancestral evolutionary units of duplication based on mammalian outgroups and a repeat graph (Jiang et al. 2007). Focusing on duplicons that carry genes or duplicated portions of genes, we identify 30 duplicons that show the greatest copy number change between the two assemblies and all differences favor a significant increase in copy number for the T2T CHM 13 assembly (Figure 1d, Table S2). We also compared the number of SDs more directly by aligning syntenic regions between GRCh38 and T2T CHM13 and counting the number of SD alignments in 5 Mbp windows (Methods). Of the 15 windows with the largest increase, nine mapped to the acrocentric short arms while six were in pericentromeric regions (Figure S1, Table S3). In particular, the intervals between the centromeric satellite and secondary constrictions (qh regions) on chromosomes 1, 9,

and 16 show a 4.5-fold increase in the number of SDs (5,254/1,141) and show the most dramatic differences in organization when compared to GRCh38. SDs in these regions are almost exclusively interchromosomal and depleted for intrachromosomal duplications (Figure S2-S3).

4.3.2 ***Validation and heteromorphic variation.***

Because the acrocentric short arms as well as the qh regions on chr1, chr9, and chr16 were either newly assembled or showed the most significant differences in terms of SD content, we focused first on validating their organization. We mapped available end-sequence data from a human fosmid genome library (Kidd et al. 2008) to the T2T CHM13 assembly and selected nine distinct clones as probes (Figure 4.2a) to confirm the patterns of high-identity (>95%) SDs. Of the 30 distinct duplication predictions based on T2T CHM13 SDs, all 30 were corroborated by FISH against chromosomal metaphases of the CHM13 cell line (Figure 4.2b, Table S4). Interestingly, FISH also revealed nine additional signals not originally predicted by our SD analysis. However, we were able to identify lower identity duplications confirming seven of these sites leading to an overall concordance of 95% (37/39) between FISH and the T2T CHM13 SD assembly content. We extended this analysis to five additional human cell lines of diploid origin because both pericentromeric and acrocentric portions of chromosomes have been shown to be cytogenetically heteromorphic ((Bhasin 2005; Hsu et al. 1987; Barber 1994). In total, we identified 61 distinct cytogenetic locations of which 28 (46%) were fixed while 33 (54%) were variable in their presence or absence on specific homologues (both acrocentric and pericentromeric regions of the human genome) (Figure S4). Of the 61 FISH signals all but six were observed in more than one of the six human cell lines indicating that such heteromorphic variation is common and prevalent. The characterization of fixed versus variable sequence-based probes anchored at specific positions

within the T2T CHM13 assembly will be valuable for future characterization of acrocentric variation including disease-associated Robertsonian translocations.

We assessed genome-wide copy number variation between the assembly and Illumina data generated from CHM13 and found that there is a Pearson correlation of 0.96 (Methods). The same analysis between two high-coverage Illumina datasets from CHM13 had a correlation of only 0.92, indicating that the remaining differences in copy number (CN) estimates are likely dominated by Illumina sequencing biases. Finally, we focused on assessing orientation differences between GRCh38 and T2T CHM13—most of which were mediated by SDs that frequently map to the breakpoints of inversion polymorphisms (Chaisson et al. 2019; Sanders et al. 2016; Kidd et al. 2008). We validated 65 inversions relative to GRCh38 based on Strand-seq analysis of the CHM13 assembly (Figure S5-S6**,** Methods). While 32 of these represent known human polymorphisms, 33 are novel compared to six human genome samples previously analyzed for structural variation (Chaisson et al. 2019). However, by analysis of Strand-seq data from one additional human haplotype (CHM1), we further confirmed 30 of these inversions suggesting that at least 95.3% (62/65) represent human inversion polymorphisms (Figure S5). Inversions associated with SDs (30) are significantly longer than those not associated with SDs (p value < 0.01, one-sided Wilcoxon rank-sum test) and are all polymorphic between humans (Figure S6). One striking example of duplication-rich inversion difference between GRCh38 and T2T CHM13 is a potential polymorphism mapping to chromosome 1q21. It is a complex event consisting of two inversions (262.3 kbp, 2.26 Mbp) originally predicted by Sanders and colleagues (Sanders et al. 2016) and one relocation (767.6 kbp) (Figure 4.2c). The large inversion (chr1:146,350,000-148,610,000) is flanked by the core duplicon *NPBF* and in combination with the other rearrangements changes the order of human-specific genes *NOTCH2NL*A, B and C, which have been implicated in the

expansion of the frontal cortex (Fiddes et al. 2018; Suzuki et al. 2018). As a final test, we sequence resolved this region in eight additional human haplotypes (Methods)—all of which support the T2T CHM13 configuration with one exception (CHM1), which was used to resolve this region in GRCh38 (Figure S7).

4.3.3 ***Single-nucleotide and copy number variation within SDs.***

The high quality and single haplotype nature of both the T2T CHM13 and GRCh38 reference genomes provides us an opportunity to compare the genome-wide pattern of single-nucleotide variation in regions that have been typically excluded from most previous analyses due to their repetitive nature. We aligned GRCh38 to T2T CHM13 in 5 kbp windows and retained only regions deemed to be "syntenic" based on an ambiguous 1-1 correspondence between both reference genomes (Methods). As expected, most unique regions of the genome (2,693 Mbp) could be compared while only 60% (124 Mbp) of the SDs within T2T CHM13 had a clear orthologous relationship between the two human references. As expected, the X chromosome and the region corresponding to the major histocompatibility complex (MHC) are the least and most diverged regions, respectively (Figure 4.3a), due to the deep coalescence of MHC and slower rate of evolution of the female X. Interestingly, SD sequences are significantly more diverged than unique sequences (p-value < 0.001, one-sided Mann-Whitney U test) (Table S5, Figure S8). This could be due to an increased number of mutational mechanisms that act on SDs (e.g., interlocus gene conversion), or a deeper average coalescence of duplicated sequences.

As part of this analysis, we also identified regions that structurally differ or are absent from GRCh38 when compared to the T2T CHM13 assembly. Based on 1 Mbp LASTZ alignments (Methods), we identified 126 non-syntenic regions for a total of 240 Mbp (N50 length of 12.7 Mbp; Figure S9). Of these, 33.9% (81.34/240 Mbp) overlap SD regions. Using sequence read-

depth (Methods) from 268 human genomes (Simons Genome Diversity Project or SGDP), we compared the copy number of both CHM13 and GRCh38 (Mallick et al. 2016) successfully genotyping 1,292 distinct copy number variable regions (74.85 Mbp). We find that CHM13 maps within 2 standard deviations of the median human CN from SGDP for 94% of bases (70.6 Mbp) in contrast to GRCh38 where 57% (42.8 Mbp) meet this metric (Figure S10). In particular, we find that human copy number is 9.0 times (59.26/6.55 Mbp) more likely to match the CHM13 copy number rather than GRCh38 (Figure 4.3b). Using the reference as a predictor of human copy number, we find that CHM13 is a much better predictor (AUC 0.91) than GRCh38 (AUC 0.77) and close to the best theoretical reference (AUC 0.96, Figure 4.3c). GRCh38 tends to underestimate normal human CN (by on average 9.2 copies or median of 3.0 copies).

We identify 119 protein-encoding genes (65 for GRCh38) where CHM13 copy number better represents the true human copy number state (Table S6). These include both biomedically important genes relevant to disease risk *(LPA*, *MUC3A*, *FCGR2*) (Clarke et al. 2009; Coassin et al. 2019; Kronenberg and Utermann 2013; Schmidt et al. 2016; Gum et al. 1990; Pratt et al. 2000; Kyo et al. 2001; Kyogoku et al. 2002; Willcocks et al. 2010) as well as gene families that have been implicated in the evolutionary adaptation of the human lineage (*TBC1D3, NPIP, NPBF*) (Ju et al. 2016; Paulding et al. 2003; Cantsilieris et al. 2020; Marques-Bonet and Eichler 2009) important loci (Figure 4.3d, Table S6) (Marques-Bonet and Eichler 2009). In T2T CHM13, for example, there are additional copies of *NPIP, NPBF,* and *GOLGA* that are absent from GRCh38—each of these has been described as core duplicons responsible for the expansion of interspersed duplications in the human genome (Jiang et al. 2007) as well as the emergence of human-specific gene families. Interestingly, African genomes tend to have overall a higher copy number status when compared to non-African genomes. In particular, *TBC1D3* shows ~7 fewer copies in non-

Africans when compared to Africans (p-value < 1e-12). These findings suggest that higher copy is likely ancestral (Table S7) and CHM13 better represents that diversity.

4.3.4 ***Structural variation and massive evolutionary changes in the human lineage.***

Because of advances in long-read genome assembly (Nurk et al. 2020; Cheng et al. 2021), we have the potential to sequence resolve complex structural variation associated with SDs at the haplotypic level. We generated or used existing HiFi sequence data from 12 human and 6 nonhuman primate genomes to understand both the structural diversity and evolution of specific SD regions. Comparing the chimpanzee and the T2T CHM13 assemblies, we specifically searched for gene-rich, large-scale genetic differences (>50 kbp in length) and selected 10 loci for a more detailed evolutionary analyses, including regions of particular biomedical importance and regions associated with adaptive changes and the expansion of the frontal cortex (Tables S8-S10; Figure S11). Of 10 targeted loci, eight confirmed the structural organization of T2T CHM13 when compared to GRCh38 where 5/10 were completely supported by another human haplotype. Overall, 73% of human haplotype assemblies were successfully reconstructed (extending at least 50 kbp beyond the duplicated portion) (Table S8). The level of sequence completion in diploid samples varied, however, depending on the complexity of the locus. For example, in the case of the 8.9 Mbp region corresponding to *NOTCH2NL* and *SRGAP2D/2D*, we recovered only 37.5% of human haplotypes (Table S8, Figure S7). Similarly, we resolved only 6 haplotypes (from a potential of 24 haplotypes) for the 3.4 Mbp region harboring the *SMN1 and SMN2* loci (Figure S12).

Among those loci that could be resolved, we find a high degree of structural heterozygosity (67%, Methods) with 249 kbp differing on average among human haplotypes compared to T2T CHM13 (Table S9). In some cases, the structural changes are simple, such as ~12 kbp insertion or

deletion of *CYPD26*, which contributes to differential drug metabolism activity as well other human disease susceptibilities (Bertilsson et al. 2002; Hammer and Sjöqvist 1967; Alexanderson et al. 1969; Skoda et al. 1988; Dahl et al. 1992; Gaedigk et al. 1991; Johansson et al. 1993) (Figure S13). In other cases, the patterns of structural variation are complex involving hundreds of kbp of inserted or deleted gene-rich sequence along with large-scale inversion events that alter gene order for specific human haplotypes (see *ARGHAP11A/B*; Figure S14 and *NOTCH2NLA/B;* Figure S7). The spinal muscular atrophy (SMA) locus containing *SMN1* and *SMN2—one of the most difficult regions to finish as part of chromosome 5* (Schmutz et al. 2004)—shows a unique structure for each of the six assembled haplotypes that we resolved (plus GRCh38). Some haplotypes not only show increases in *SMN2* copy number (Figure S12), a known genetic modifier associated with the delay of SMA (Butchbach 2016), but also potential functional differences in the organization and composition of *SMN2*. Since *SMN2* serves as a target for small molecule drug therapy improving splice-site efficiency compensating for the loss *SMN1* in SMA patients (Bebee et al. 2010), this level of sequence resolution is of practical utility for disease risk and treatment of patients.

Of particular interest is the *TBC1D3* gene family (Paulding et al. 2003) (Figure 4.4, Figures S15-S16) whose protein products modulate epidermal growth factor receptor signaling and trafficking (Wainszelbaum et al. 2008) and whose duplication in humans has been associated with expansion of the human prefrontal cortex as evidenced by mouse transgenic experiments (Ju et al. 2016). A comparison to chimpanzee (Figure 4.4a) shows two massive genomic expansions in the human lineage (323.0 and 124.4 kbp). Both the high sequence identity (99.6%) and sequence read-depth comparisons of *TBC1D3* copy number are consistent with expansion occurring in the human lineage after divergence from chimpanzee (Figure 4.4b). We extended this analysis to other nonhuman primates by generating HiFi assemblies for bonobo, gorilla, orangutan, and macaque.

We identified *TBC1D3* homologues in each species and constructed a maximum likelihood phylogeny based on intronic or noncoding sequence flanking the gene (Figure 4.4c). The analysis reveals recurrent and independent expansions of *TBC1D3* in the orangutan, gorilla, and macaque species at different time points during primate evolution with the most recent occurring 2 and 2.6 million years ago, near the emergence of the *Homo* genus (Wood 1992).

Complete sequencing of human *TBC1D3* haplotypes reveals remarkable structural diversity (Figure 4.4d) with *TBC1D3* copy number ranging from three to fourteen *TBC1D3* copies in expansion site #1, and two to nine copies in expansion site #2. In total, approximately one third of human expansion site #2 shows large-scale structural variation and we identify >1.8 Mbp of duplicated sequence and >650 kbp of inverted sequence across the 18 haplotypes (including GRCh38). We estimate the heterozygosity of this locus to be over 77.8% (14/18 haplotypes are structurally distinct) (Figure S16). *TBC1D3* expansion site #1 is similarly structurally heterozygous with 63.6% (14/22) of the haplotypes displaying unique structures corresponding to copy number differences in the *TBC1D3* gene family (Figure S15). Using orthogonal Oxford Nanopore ultra-long-read sequencing technology, we validate these complex patterns of structural variation in a subset of the samples investigated here (Methods, Figures S17-S18). The nature of human genetic variation at these loci complicates simple linear alignments and is better captured by a graph-based representation (Li et al. 2020), which identifies two *TBC1D3* genes as common among all human haplotypes examined thus far (*TBC1D3B* at site #1 and *TBC1D3(A)* at site #2).

4.3.5 ***New gene models and variable duplicate genes.***

We identified 182 candidate new or non-syntenic genes in the T2T genome assembly with open reading frames and multiple exons. Of these 91% (166) corresponded to SD gene families (Figure 4.5a). Many of these represent expanded tandem duplications (e.g., *GAGE* gene family

members on the X chromosome) or large interspersed duplications (e.g., beta-defensin locus) adding additional copies of nearly identical genes to the human genome (Figure 4.5a). We searched for evidence that these copy number polymorphic or structurally variant regions were transcribed by aligning long-read transcript sequencing data and searching for perfect matches (Methods). We constructed a database of 44.2 million full-length cDNA transcripts derived from 31 human tissue samples and compared them to both the GRCh38 and T2T CHM13 human genome references. For those 182 novel protein-coding genes where an unambiguous assignment could be made, 36% (65/182, >20 IsoSeq reads) were confirmed to be expressed with 23 showing the majority of reads mapping better to T2T CHM13 when compared to GRCh38 (Figure 4.5b). Overall across the entire genome, 12% of full-length transcripts align at least 0.2% better to CHM13, while 8% align better to GRCh38. These results are consistent with the notion that the T2T CHM13 is more complete but that both are capturing structurally variant haplotypes of genic potential. In addition to entirely new genes, we identify several gene models that are complete for the first time—many of which encode proteins with large tandem repeat domains (ZNF, LPA, Mucin, Figure 4.5c) etc. Among these is the complete gene structure of the Kringle IV domain of the lipoprotein A gene—one of the strongest genetic associations with cardiovascular disease, especially among African Americans (Clarke et al. 2009; Kronenberg 2016; Kronenberg and Utermann 2013; Coassin et al. 2019; Schmidt et al. 2016)— where we identify not only length variation but other forms of functional variation potentially relevant to disease risk (Figure 4.5d).

4.3.6 ***SD methylation and transcription.***

Since methylation is an important consideration in regulating gene transcription, we took advantage of the signal inherent to ultra-long-read ONT data (Loman et al. 2015; Quick et al. 2016; Simpson et al. 2017) to investigate CpG methylation status of SD genes within the CHM13

genome (Methods). SD blocks follow a clear bimodal distribution in methylation level; (Figure 4.6a, Figure S19); 452 SD blocks flanked (127.7 Mbp) by unique sequences are hypermethylated in contrast to 222 hypomethylated SD blocks (52.1 Mbp). Methylation status does not appear to be driven by genomic location, e.g., proximity to the centromeres, acrocentric short arms, or telomeres (Figure 4.6a). Using full-length transcript data from CHM13, we compared methylation and transcription status of duplicated genes (Methods). If we stratify genes by their number of full-length transcripts, we observe distinct methylation patterns for transcribed and non-transcribed SD genes (Figure 4.6b). For highly transcribed SDs and unique genes, the gene body and flanking sequence are generally hypermethylated with a dramatic dip near the transcription start site (TSS)/promoter. In contrast, non-transcribed genes show moderate to low methylation across the gene body and flanking sequence. Restricting the analysis to genes mapping within SDs, we find that transcriptionally silenced duplicate genes are more likely (10,000 permutations, $p=0.0018$) to map to hypomethylated regions of SD sequence (Figure 4.6a) when compared to transcribed duplicate genes. Additionally, in untranscribed SD genes we observe a statistically significant (one-sided Mann-Whitney) increase in TSS methylation (6.6% increase) when compared to unique genes where the TSS is more likely to be depleted for methylation (8.2% decrease).

One important consideration in this analysis is the presence of a CpG island within 1500 bp of the promoter (Saxonov et al. 2006). In our analysis of CHM13, for example, unexpressed unique genes have a low CpG count, consistent with a lack of CpG islands. If we repeat the same analysis on SD genes, we find that the unexpressed SD genes exist with and without CpG islands (Figure S20). In total, these observations suggest a process of epigenetic silencing for a subset of duplicate genes through general demethylation of the gene body but hypermethylation of promoter regions. Based on these observed signatures, we investigated whether it might be possible to

predict individually actively transcribed duplicate genes from these epigenetic features. We investigated a recently duplicated hominid (*NPIPA*) (Johnson et al. 2001) where sufficient paralogous sequence differences exist to unambiguously assign full-length transcripts to specific loci. While promoter / TSS signatures are less evident at the individual gene level, the gene body methylation signal appears diagnostic (Figure 4.5c). *NPIPA1* and *NPIPA9*, for example, are the most transcriptionally active and show demonstrably distinct methylation patterns providing an epigenetic mark to define the transcriptionally active loci associated with high-identity gene families. We show this trend also holds for other high-copy number gene families (Figure S21).

4.4 Discussion

This work represents a significant advance for three reasons. First, it provides the first comprehensive view of the organization of SDs in the human genome. The new reference adds a chromosome's worth of new SDs increasing the human genome average from 5 to 7% nearly doubling the number of SD pairwise relationships (41 vs. 24 thousand) and, as a result, predicts new regions of genomic instability. By every metric, T2T CHM13 is a better representation than GRCh38 of the structure of the human genome. This includes the first sequence-based organization of the short arms of chromosomes 13, 14, 15, 20 and 21 where we find that SDs account for more sequence (34.6 Mbp) than either heterochromatic satellite (26.7 Mbp) or rDNA (10 Mbp)**.** Acrocentric SDs are almost twice as large when compared to non-acrocentric regions likely due to non-allelic homologous exchange events occurring among the short arms which associate more frequently during the formation of the nucleolus (Arnheim et al. 1983). Interestingly, nearly half of the acrocentric SDs involve duplications with non-acrocentric pericentromeric regions of chromosomes 1, 3, 4, 7, 9, 16 and 20. While the underlying mechanism for these large-scale duplications is unknown, it is noteworthy that three of these have large secondary constriction sites

(chromosomes 1q, 9q and 16q) composed almost entirely of heterochromatic satellites (HSAT2 & 3) (Figure S2). These particular SD blocks, thus, are bracketed by large tracts of heterochromatic satellites and such configurations may make them particularly prone to double-strand breakage events (Luke et al. 1992). The organization of these pericentromeric regions is, thus, similar to acrocentric regions helping to explain such regions are enriched for interchromosomal duplications (Figure S3).

Second, the new reference along with resources from other human genomes provides a baseline for investigating more complex forms of human genetic variation at various levels. The completion of the reference sequence, for example, facilitates the design of sequence-anchored probes to systematically discover and characterize SD heteromorphic variation where chromosome organization differs among individuals (Figure 4.2). Such chromosomal heteromorphisms have been traditionally investigated cytogenetically and are thought to be clinically benign (Bhasin 2005; Hsu et al. 1987; Barber 1994); however, more recent work indicates that these large-scale variants associate with infertility by increasing sperm aneuploidy, decreasing rates of embryonic cleavage (IVF), and increasing miscarriage (Barber et al. 2006; Sahin et al. 2008); (Codina-Pascual et al. 2006; Caglayan et al. 2010; Madon et al. 2005; Minocherhomji et al. 2009; Hong et al. 2011; Kalantari et al. 2001). Distinguishing between fixed and heteromorphic acrocentric SDs will facilitate such research as well as the characterization of breakpoints associated with Robertstonian translocations—the most common form of human translocation (Wilch and Morton 2018).

At a finer-grained level, the new reference and the use of long reads from other human genomes provides access to other complex forms of variation involving duplicated gene families. Short-read copy number variation analyses and SNP microarray have long predicted that SDs are enriched 10-fold for copy number variation but the structural differences underlying these regions

as well as their functional consequences have remained elusive (Sudmant et al. 2015a; Locke et al. 2004). We reveal unprecedented levels of human genetic variation in genes important for neurodevelopment (*TBC1D3*) and human disease (*LPA*, *SMN*). Even between just two genomes (GRCh38 and CHM13) we find that 37% (81 Mbp) of SD bases are structurally variable and this predicts 184 copy number variable genes between two human haplotypes (Table S6). In some cases (*TBC1D3*), we find that nearly every human haplotype varies structurally carrying different complements and arrangements of the *TBC1D3* gene family. The potential ramifications of this dramatic expansion in humans versus chimpanzees and of such high structural heterozygosity among humans are intriguing given the genes purported role in expansion of the frontal cortex (Ju et al. 2016). Similarly, we were able to reconstruct the complete structure of the *LPA* gene model in multiple human genotypes. While this is only a single gene, variability in the tandemly repeated 5.2 kbp protein-encoding Kringle-IV domain underlies one of the most significant genetic risk factors for cardiovascular disease. Sequence resolution of the structural variation as well as underlying amino-acid differences allow us to predict novel risk alleles for disease (Figure 4.5). Sequence-resolved structural variation improves genotyping and tests of selection (Ebler et al. 2020; Ebert et al. 2020; Hsieh et al. 2019) providing a path forward for understanding the disease and evolutionary implications of these complex forms of genetic variation.

Third, and perhaps most importantly, the new reference coupled with other long-read datasets enables genome-wide functional characterization of recently duplicated genes. Both gene annotation and large-scale efforts to characterize the regulatory landscape of the human genome have typically excluded repetitive regions, including the 859 human genes mapping to high-identity SDs (GTEx Consortium et al. 2017; Dougherty et al. 2018). This is because the underlying short sequence read data limits RNA-seq or Chip-seq data from being assigned unambiguously to

specific duplicated genes. In this study, we generated long-read full-length transcript data (Iso-Seq) with long-read methylation data from ONT from the same genome source allowing us to investigate epigenetic and transcriptional data simultaneously against a fully assembled reference genome. The long-read data from the same haploid source facilitated the unique assignment of these functional readouts allowing us to correlate methylation and transcript abundance. Our initial analyses suggests that a large fraction of duplicate genes are in fact epigenetically silenced (characterized by hypermethylation of the promoter and hypomethylation of the gene body) and that this epigenetic mark may be used to predict actively transcribed loci even when genes are virtually identical (Figure 4.6). While more human genomes and diverse tissues will need to be interrogated to assess the significance of this observation, it is clear that phased genome assemblies (Ebert et al. 2021) with long-read functional readouts such as methylation (Simpson et al. 2017), transcription, or Fiber-seq (Stergachis et al. 2020; Abdulhay et al. 2020) provide a powerful approach to understanding the regulatory landscape of duplicated and copy number polymorphic genes in the human genome.

There are several remaining challenges. First, not all human haplotypes corresponding to specific duplicated regions could be fully sequence resolved by strictly applying long-read HiFi sequencing technology. There is compelling evidence that such unresolved regions correspond to some of the largest and most variable regions of the human genome (Ebert et al. 2021). For example, only 25% of *SMN1*/*SMN2* haplotypes were fully resolved and unresolved loci are predicted to carry some of the most complex SV patterns. Assembly methods that effectively combine the extreme read lengths of ONT and accuracy of HiFi will likely be critical to complete characterization of SD haplotypes and a diploid T2T assembly (Miga et al. 2020; Logsdon et al. 2020b). Indeed, combining the two long-read technologies was an important factor to the success

of T2T CHM13 (Nurk et al., unpublished). Another important challenge going forward will be how to accurately represent these more complex forms of human genetic variation, including functional annotation. Most agree that linear references will be insufficient for this purpose (Eizenga et al. 2020). While a sufficiently complex pangenome reference graph could overcome these limitations, practically it is unclear how this will be achieved or how it will be adopted by the genomics and clinical community. This highlights the importance of not only the construction of a pangenome reference but the necessary tools that will distinguish paralogous and orthologous sequences within duplications to allow for comparison between haplotypes with different SD architectures. The work currently underway by the Human Pangenome Reference Consortium (HPRC), Human Genome Structural Variation Consortium (HGSVC), and Telomere-to-Telomere (T2T) consortium will be key in developing these methods; however, the magnitude and complexity of variation in SDs still remains underestimated.

4.5 Methods

4.5.1 *Estimating the number of rDNA copies in the assembly.*

To estimate the CN in the assembly of the rDNA repeats we aligned KY962518.fasta (https://www.ncbi.nlm.nih.gov/nuccore/KY962518) to the whole-genome assembly using minimap2 with the following settings and counted the number of alignments:

```
minimap2 -ax asm20 -N 100 -p 0.5 --secondary=yes --
eqx -r 500000 chm13.draft_v1.0.fasta KY962518.fasta |
samtools view -c
```

4.5.2 *Estimating amount of missing rDNA sequence in the assembly.*

To estimate the amount of missing duplicated rDNA sequence, we applied digital droplet PCR (Bell et al. 2018). See the main paper for methods (Nurk et al., unpublished). As an orthogonal

test to estimate the total rDNA present in T2T CHM13, we aligned Illumina WGS data to copies of the 45S rDNA that were in the assembly. We then divided the coverage of the regions by the genome average to get a CN estimate, which was then multiplied by the length of the rDNA motif (45 kbp).

```
chr13 5774372 5780085 CN=13.9611
chr13 9830085 9841446 CN=24.5189
chr13 9874433 9887769 CN=15.6735
chr13 9919303 9932657 CN=14.5361
chr13 9968556 9981897 CN=13.9473
chr13 10014553 10020808 CN=10.2396
chr14 2781023 2794382 CN=16.8869
chr14 2825169 2838549 CN=14.3628
chr14 2869275 2875496 CN=6.76065
chr15 2510161 2523520 CN=17.6727
chr15 5286318 5292607 CN=8.41476
chr21 3112034 3125412 CN=16.3837
chr21 6296778 6310122 CN=17.8963
chr21 6345500 6349728 CN=7.72519
chr22 4797590 4809971 CN=12.3878
chr22 5709971 5711101 CN=11.1573
chr22 5743105 5749418 CN=11.7682
totalCN=234.293 totalMbp=10.5432
```

Finally, we also estimated the CN using a k-mer analysis, which predicts a total of 10.035 Mbp. See the methods of the main paper for details.

4.5.3 ***Repeat Masking.***

Common repeats were masked with RepeatMasker v4.1 (Smit et al. 1996) and TRF (Benson 1999). The full pipeline for these masking steps are provided for convenience at https://github.com/mrvollger/assembly_workflows/ under workflows/mask.smk. In brief RepeatMasker was run with the following settings:

```
RepeatMasker -s -xsmall -e ncbi -species human -dir
$(dirname {input.fasta}) -pa {threads} {input.fasta}
```

And TRF was run with:

```
trf {input.fasta} 2 7 7 80 10 50 15 -l 25 -h -ngs >
{output.dat}
```

4.5.4 ***SD Annotation.***

To annotate SDs we identified homologous segments used SEDEF [v1.1-31-g68de243, (Numanagic et al. 2018)] on a masked version the CHM13 v1.0 assembly that included chrY from GRCh38. SDs were filtered to contain at most 70% satellite sequence as determined by RepeatMasker. Additionally, SDs had to be at least 90% identical by % identical, 50% identical including indels, and at least 1 kbp of aligned sequence or else they were filtered into a set of smaller and lower identity duplications. Pericentromeric and telomeric SDs were defined as being within 500 kbp and 5 Mbp of the telomere and centromere, respectively. The full pipeline for these masking steps are provided for convenience at https://github.com/mrvollger/assembly_workflows/ under workflows/sedef.smk. The same workflow was applied to the chromosome-level scaffolds of GRCh38 for all SD comparisons made in the paper.

4.5.5 ***Defining syntenic regions between T2T CHM13 and GRCh38.***

The T2T CHM13 to GRCh38 synteny track was constructed using the Cactus HAL file (available at the following link: http://t2t.gi.ucsc.edu/chm13/hub/t2t-chm13-v1.0/cactus/t2t-chm13-v1.0.aln1.hal) with 1 mega-base resolution and a maximum anchor distance of 50 kbp. We used the tool halSynteny to construct syntenic blocks from the Cactus alignments the methods of which are described in detail in Krasheninnikova et al., 2020 (Krasheninnikova et al. 2020). This track is available at the following link: http://t2t.gi.ucsc.edu/chm13/hub/t2t-chm13-v1.0/synteny/synteny.1mb.bigPsl. To define the new and variable regions T2T CHM13, we inverted the 1 Mbp synteny track retaining all regions without an alignment to GRCh38.

4.5.6 ***Calculating the number of SD alignments in 5 Mbp windows.***

We first offset the coordinates of SDs in GRCh38 such that the largest gaps (acrocentric short arms, centromeres, and HSAT arrays) matched the length of the assembled sequence in CHM13 T2T. We then normalized the GRCh38 coordinates so that the length of the chromosomes in GRCh38 were equal to those in T2T CHM13. After this we took 5Mbp non-overlapping windows from T2T CHM13 and the normalized GRCh38 and calculated the difference in the number of SDs within each window (Table S3).

4.5.7 ***WSSD detection and genotyping.***

As an orthogonal method to estimate copy number of SDs, we applied the whole genome shotgun sequence detection (WSSD) pipeline which uses sequence read-depth as a proxy (Bailey et al. 2002). Short-read sequence data were processed into 36 bp non-overlapping fragments and mapped to a masked T2T CHM13 reference using mrsFAST (Hach et al. 2010) with a maximum of two substitution mismatches not allowing for indels. Masking was determined by Tandem Repeats Finder and RepeatMasker. Read-depth across the genome was corrected for GC-bias and copy number was determined using linear regression on read-depth versus known fixed copy number control regions. Finally, integer genotypes were estimated by using the predicted mean and variance of the Gaussian distributions underlying different copy numbers to create a series of models to represent the likely distribution of read-depths underlying a region of specific copy number.

For defining genotyping intervals, we applied the changepoint package in R (Killick et al. 2016) to identify regions where the CHM13 WSSD copy number estimate was consistent. Specifically, we used a log transformed continuous CN estimate from WSSD for sliding windows

across the assembly and then applied binary segmentation to identify regions where the CN remained the same. We used the following R command:

```
cpt.mean(Log_cn, method = "BinSeg", Q=Q)
```

Where Log_cn is a vector of log scaled CN estimates and Q is the number of independent 50 kbp windows within each chromosome. To validate the CN of assemblies we fragmented the assemblies in 36 bp windows with a 1 bp slide and used it as input read data for our WSSD CN pipeline. Then every CN estimate within an SD space was compared between the Illumina estimate and assembly estimate and a Pearson's correlation was calculated.

4.5.8 ***Gene annotations with Liftoff.***

Gene annotations on T2T CHM13 were made using Liftoff (Shumate and Salzberg 2020a) and then processed with gffread (Pertea and Pertea 2020) to filter for only transcripts with open reading frames. The full pipeline for gene annotation is provided for convenience at https://github.com/mrvollger/assembly_workflows/ under workflows/liftoff.smk. In brief Liftoff was called with the following command:

```
liftoff -dir {output.temp} -f <(echo "locus") -flank
0.1 -sc 0.85 -copies -p {threads} -g {input.gff} -o
{output.gff} -u {output.unmapped} {input.t} {input.r}
```

Using as input the gencode v34 annotation gff3 available at ftp://ftp.ebi.ac.uk/pub/databases/gencode/Gencode_human/release_34/gencode.v34.annotation.gff3.gz and GRCh38 fasta available at https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/001/405/GCF_000001405.39_GRCh38.p13/GRCh38_major_release_seqs_for_alignment_pipelines/GCA_000001405.15_GRCh38_no_alt_analysis_set.fna.gz.

4.5.9 ***Counting the number of high identity SD genes.***

We counted all protein-encoding genes with at least one exon mapping fully within a >95% identical SD and had the additional condition that at least 50% of the full-length gene maps to SD space without the identity limitation.

4.5.10 ***Cell culture.***

CHM13 and CHM1 cells were cultured in complete AmnioMax C-100 Basal Medium (Thermo Fisher Scientific, 17001082) supplemented with 15% AmnioMax C-100 Supplement (Thermo Fisher Scientific, 12556015) and 1% penicillin-streptomycin (Thermo Fisher Scientific, 15140122). GM24385, GM19240, HG00514 and HG00733 cells were cultured in RPMI 1640 with L-glutamine medium (Thermo Fisher Scientific, 11875093) supplemented with 15% FBS (Thermo Fisher Scientific, 16000-044) and 1% penicillin-streptomycin (Thermo Fisher Scientific, 15140122). All cells were cultured in a humidity-controlled environment at 37°C with 5% CO_2.

4.5.11 ***FISH characterization and validation.***

Fosmid probes for FISH experiments were selected by mapping fosmid end sequences (FES) from the ABC10 (NA19240 Yoruban) library (Kidd et al. 2008) to the T2T CHM13 reference using blast (Altschul et al. 1990). Human fosmids clones were used as probes in one- or two-color FISH experiments and hybridized on metaphases obtained from CHM13, CHM1, GM24385, GM19240, HG00514, and HG00733 lymphoblastoid cell lines. FISH experiments were essentially performed as previously described (Cardone et al. 2006). Slides were imaged on an inverted fluorescence microscope (Leica DMI6000) equipped with a charge-coupled device camera (Leica DFC365 FX). Mapping was performed following comparison to the conventional

classical cytogenetics G-banding (Standing Committee on Human Cytogenetic Nomenclature 1995).

4.5.12 ***ONT validation.***

To validate structural variant configurations predicted by HiFi sequence and assembly, we aligned ultra-long ONT data from two samples (HG002, HG00733) and assessed the uniformity of coverage over the *TBC1D3* assemblies for these four haplotypes. We find no obvious sign of collapsed duplications (read coverage abnormalities) or misjoins in the assemblies (every 25 kbp segment with 1 kbp slide is spanned by four or more reads) in the ultra-long ONT data (Figures S18-19).

4.5.13 ***TBC1D3 phylogenetic tree construction.***

Orthologous sequences for the two human *TBC1D3* expansion sites were identified in T2T CHM13 using minimap2 (Li 2018) and gene models were annotated using Liftoff (Shumate and Salzberg 2020b). *TBC1D3* transcripts with open reading frames were identified using gffread (Pertea and Pertea 2020). Exons were masked and removed using BEDTools maskfasta and getfasta functions (Quinlan and Hall 2010) in order to construct neutrally evolving phylogenetic trees. With exon-free paralogs of both CHM13 and nonhuman primates, a multiple sequence alignment (MSA) was generated using MAFFT (Katoh et al. 2002). To produce the most confident MSA, an iterative refinement algorithm described was used with the option for iterating 1000 times (Berger and Munson 1991; Gotoh 1993).

```
mafft --reorder --maxiterate 1000 --thread 16
{input.fasta} > {output.MSA.fasta}
```

The MSA was subsequently used to generate a maximum likelihood phylogeny, using RAxML (Stamatakis 2014). For this phylogeny, the rapid bootstrapping analysis was utilized to identify

the best ML tree, a gamma model was used to model rate heterogeneity, and macaque *TBC1D3* sequences were used as outgroup sequences.

```
raxmlHPC-PTHREADS -f a -p 12345 -x 12345 -s
{input.fasta} -m GTRGAMMA -# 100 -T 8 -n
{output.fasta.name} -o {outgroup.sequence.names}
```

4.5.14 ***Defining structurally variable haplotypes.***

To define the set of structurally distinct haplotypes for the evolutionary and biomedically important loci we performed an all against all pairwise alignment for each of the haplotypes using the following minimap2 command (Li 2018):

```
minimap2 -r 50000 -ax asm20 --eqx -Y
```

Sequences aligned to the same haplotype for at least 90% of their length at >99% identity without deletions or insertions of 50 kbp or more were grouped into a single structural haplotype and considered not structurally variable. Structurally variable haplotypes were then defined as the mutually exclusive groups where every haplotype in a given group did not align to the haplotype of any other group for >90% of its length at >99% identity.

4.5.15 ***Variation graphs for SD loci.***

We applied minigraph v0.14 (Li et al. 2020) to construct variation graphs using all structurally distinct haplotypes with the parameters:

```
minigraph -xggs -L 5000 -r 100000 -t {threads}
*.fasta
```

All haplotypes were aligned back the the graph to call variants:

```
minigraph -x asm -t {threads} {input.gfa}
{input.fasta}
```

4.5.16 ***Methylation analysis.***

Methylation analysis was performed using the same data and methods described by Gershman et al., unpublished. In brief CHM13 ultra long nanopore reads were aligned to the CHM13 reference with Winnowmap2 (Jain et al. 2020) with a kmer size of 15 and filtered for primary alignments for read lengths greater than 50 Kb. To measure CpG methylation in nanopore data we used Nanopolish (v0.13.2) (Simpson et al. 2017)filtered methylation calls using the nanopore_methylation_utilities tool (https://github.com/timplab/nanopore-methylation-utilities), which uses a log-likelihood ratio of 1.5 as a threshold for calling methylation. Methylation data was then loaded into R for all downstream analysis with GenomicRanges and dplyr.

4.5.17 ***Custom ideogram and homology visualizations***.

Linear ideograms were constructed using the karyoploteR package (Gel and Serra 2017) and circular ideograms were made using circlize (Gu et al. 2014). R code used to make these figures is shared for convenience at https://github.com/mrvollger/Vollger_2020_Figures; however, this is not a software package and therefore provided without installation instructions and will not run on other machines without modifications. Sequence homology plots were made with a modified version of Miropeats (Parsons 1995) that uses minimap2 to identify alignments. Code for the homology plots can be found in https://github.com/mrvollger/assembly_workflows under workflows/minimiro.smk. In brief sequences are aligned using the following minimap2 parameters:

```
minimap2 -x asm20 -r 200000 -s 100000 -N 1000 --
secondary=no --cs {input.ref} {input.query} >
{output.paf}
```

and then processed into a postscript file using scripts/minimiro.py and converted into a pdf.

4.5.18 ***Data Availability***

PacBio HiFi data has been deposited into NCBI SRA under the following accessions: SRX7897688, SRX7897687, SRX7897686, and SRX7897685 for CHM13; ERX3831682 for HG00733; SRR10382244, SRR10382245, SRR10382248 and SRR10382249 for HG002; PRJNA540705 for NA12878; PRJEB36100 for HG00733, NA19240, and HG00514; and PRJNA659034 for all NHP samples. PacBio HiFi data for the T2T diversity panel individuals can be found at: https://github.com/human-pangenomics/hpgp-data. The complete T2T CHM13 assembly and all CHM13 ONT data, including raw signal files (FAST5), base calls (FASTQ), and alignments (BAM/CRAM), are available at https://github.com/nanopore-wgs-consortium/chm13. The assembly can also be found on NCBI (GCA_009914755.2). Two human PacBio Iso-Seq datasets from fetal brain and testis are accessioned under NCBI BioProject PRJNA659539. SD annotations and Liftoff gene models can be found on the UCSC T2T CHM13 genome browser (http://genome.ucsc.edu/cgi-bin/hgTrackUi?hgsid=1027772427_04qK1w44AcW9pKwhyMjS4bxYTzvA&g=hub_2395475_sedefSegDups, http://genome.ucsc.edu/cgi-bin/hgTrackUi?hgsid=1027772427_04qK1w44AcW9pKwhyMjS4bxYTzvA&g=hub_2395475_liftOffGenes). The canonical rDNA unit used to estimate copy number can be found on the NCBI nucleotide repository (KY962518.1).

4.6 Figures

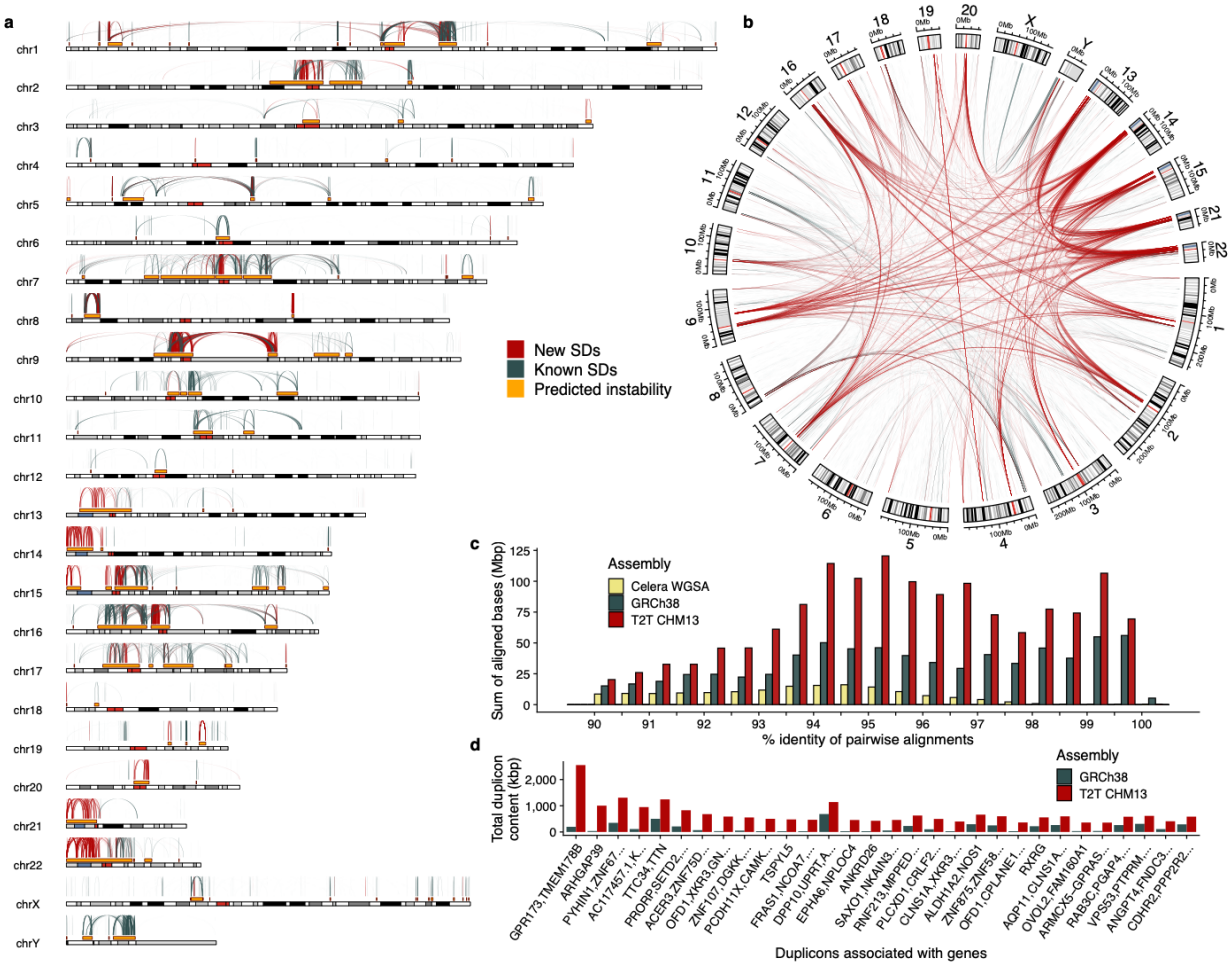


Figure 4.1. Segmental duplication (SD) content of the T2T CHM13 genome.

a) The pattern of novel or structurally variant intrachromosomal duplication in T2T CHM13 (red) compared to known duplications in GRCh38 (blue). These predict hotspots of genomic instability (gold) flanked by large (>10 kbp), high-identity (>95%) interspersed (>50 kbp) SDs. b) Circos plot highlighting novel interchromosomal SDs (red) shows the preponderance of new SDs mapping to pericentromeric and acrocentric regions. c) A histogram comparing SD content in different human reference genomes. The sum of bases in pairwise SD alignments stratified by their percent identity for the Celera (yellow, Sanger-based), GRCh38 (blue-gray, BAC-based), and T2T CHM13 (red, long read) assemblies. d) The 30 genic duplicons (ancestral repeat units) with the greatest copy number difference between GRCh38 and T2T CHM13 as determined by DupMasker. All of the 30 largest differences are present in T2T CHM13.

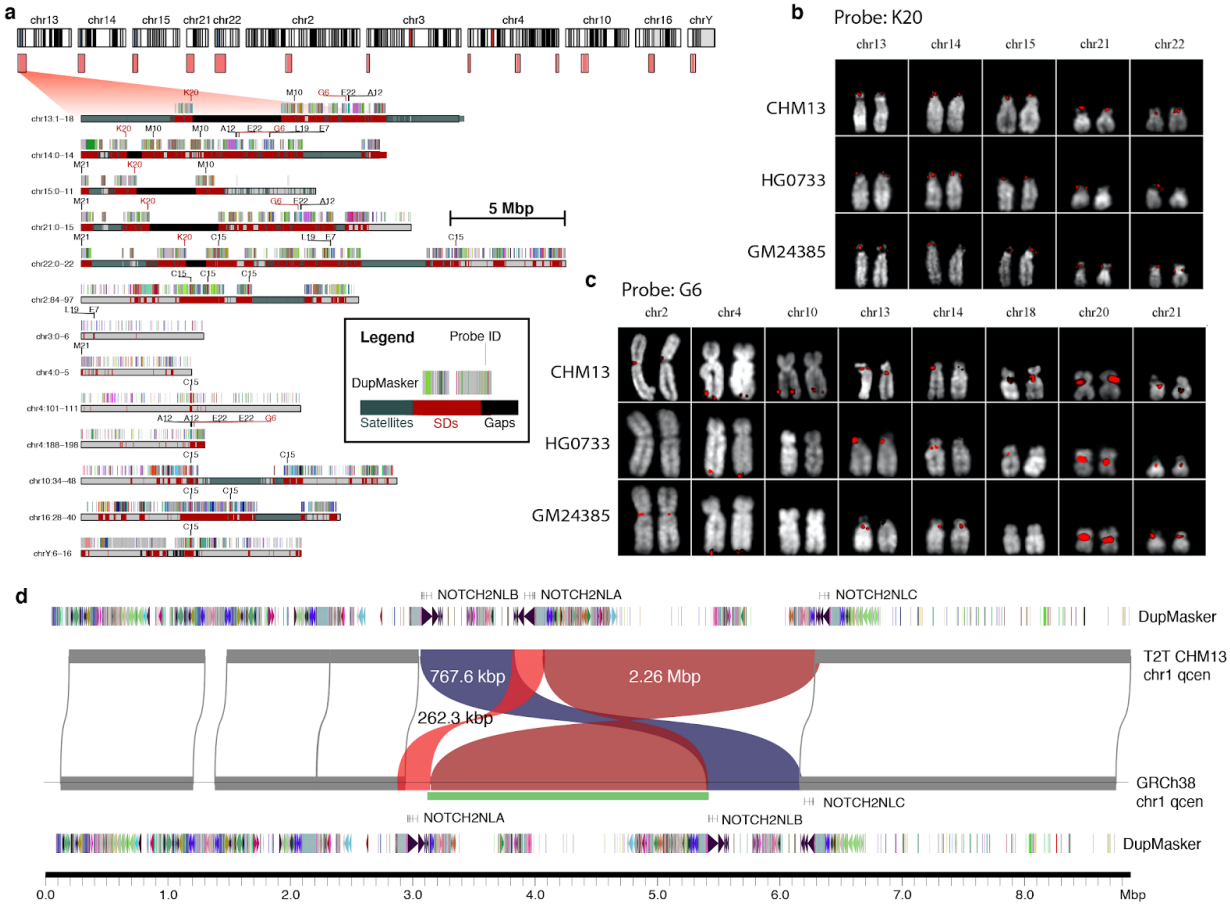


Figure 4.2. Validation of novel SDs in T2T CHM13 and heteromorphic variation.

a) Ideogram (top) shows large SD regions (light red) present in T2T but absent from the current reference human genome(GRCh38). An expanded view of the duplication (red) and satellite organization (blue-gray) are depicted below showing the location of fosmid FISH probes (eg. C15) an SD organization compared to ancestral duplicon segments (multi colored bars) (see inset). b,c) Show FISH signals (red) on extracted metaphase for two probes and three human cell lines. Probe K20 shows a fixed signal (except for one heterozygous signal), and G6 is heteromorphic among humans (see Table S4, Figure S4 for complete description for all nine probes). c) Inversion polymorphism (green bar) between T2T CHM13 and GRCh38 in the pericentromeric chromosome 1q region. Inversion (green bar) as confirmed by Strand-seq (Sanders et al, 2016) but sequencing shows a more complex structure including two inversions (red) and one reordered segment (blue) mapping near NOTCH2NL human-specific duplications.

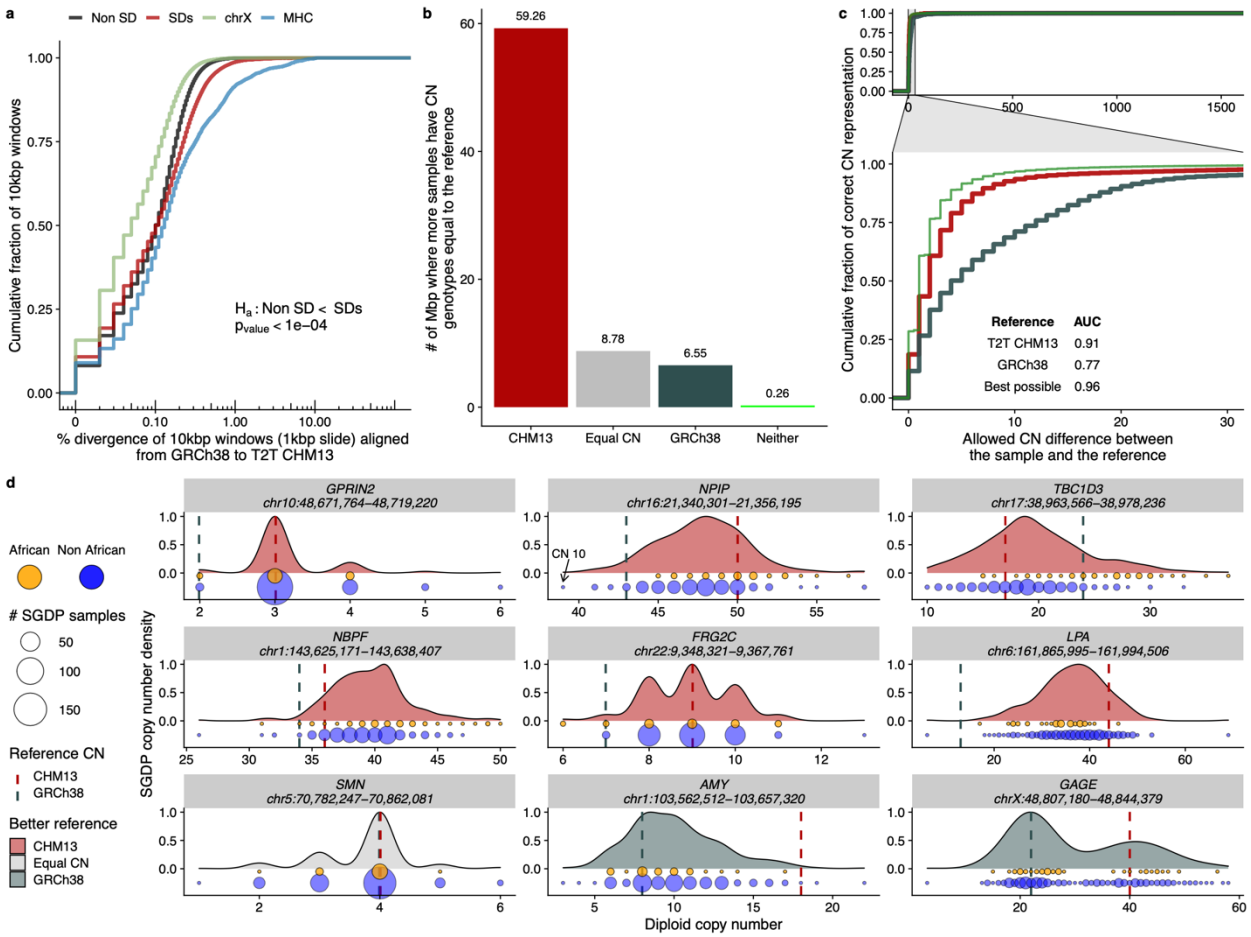


Figure 4.3. Segmental duplication sequence and copy-number variation.

a) SNV sequence divergence (% in 10 kbp windows) between GRCh38 and T2TCHM13 is compared for SD regions (bue), their flanks (yellow) and unique regions (black) of the genome. Syntenic SD regions show significantly more divergence when compared to unique but not as much as MHC regions. b) Copy number of SD regions that are new or structurally different in T2T CHM13 compared to GRCh38 based on 268 human genomes from the Simons Genome Diversity Project (SGDP). The histogram shows the number of Mbp where more samples support the CN of the given assembly [T2T CHM13 (red), GRCh38 (blue), neither (green), or both equally (Equal CN)]. c) Empirical cumulative distribution showing how many samples genotype correctly with either GRCh38 or T2T CHM13 as a function of the allowed difference between sample and reference CN. The inset shows the area under the curve (AUC) calculation for both references allowing a maximum CN difference of 30. The green curve shows the theoretical best possible reference made by taking the average CN of the SGDP samples at each site. d) Genic copy number variation. CNV of nine gene families is shown (based on SGDP) and distribution is colored according to which reference better reflects the median CN; GRCh38 generally underestimates copy number (vertical lines) and Africans (orange) tend to show higher copy number than non-Africans (blue); circle size indicates # of samples.

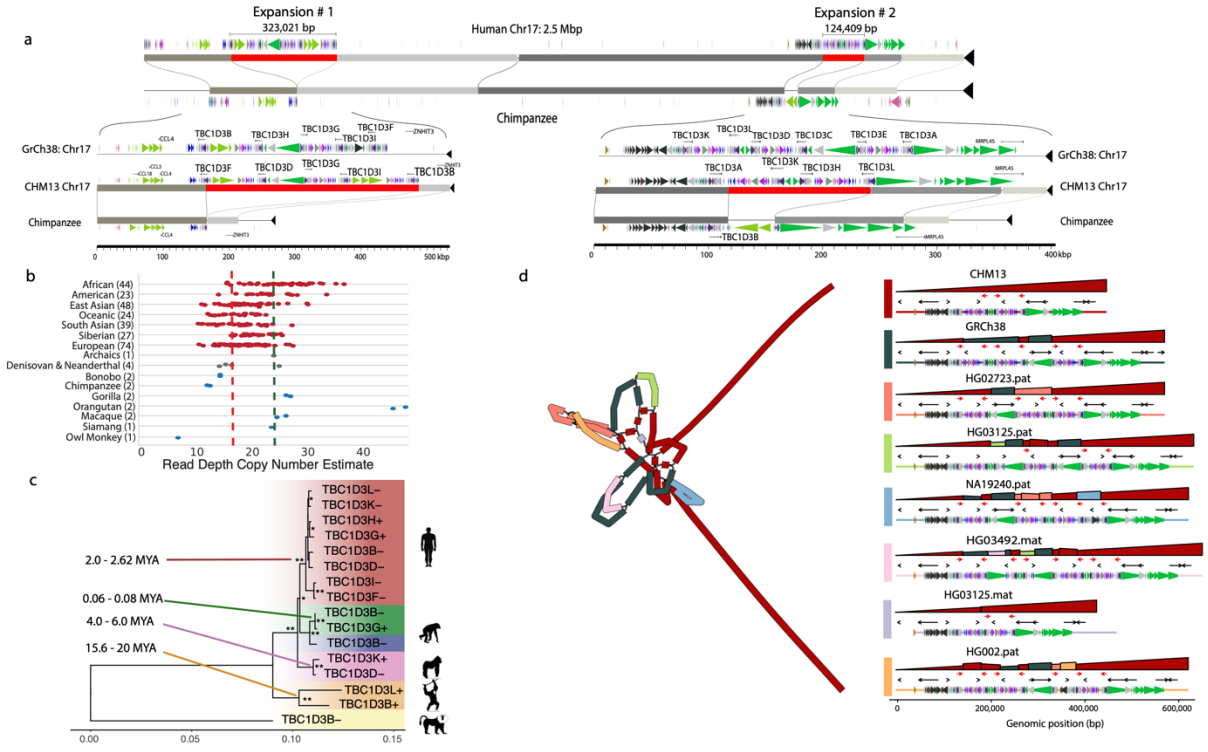


Figure 4.4. Human-specific expansion of *TBC1D3* compared to nonhuman primates.

a) Regions of homology between human T2T CHM13's Chromosome 17 (top) and a HiFi assembly of chimpanzee Clint PTR (bottom). Red blocks represent regions of human-specific expansion, including TBC1D3 duplications. Colored arrows above and below the homologous sequence represent unique ancestral units (duplicons) identified by DupMasker. Inset plots for both expansion sites are included below with the gene models identified with Liftoff. b) Copy number estimates from an Illumina read-depth analysis of SGDP, ancient hominids, and nonhuman primates for a TBC1D3 paralog. Copy number estimates are diploid and include pseudogenes (5) not included in the phylogeny or homology sequence, explaining the higher counts observed. c) Phylogeny of TBC1D3 copies at these two expansion sites as well as nonhuman primate copies. Single asterixis at nodes indicate bootstrap values greater than or equal to 70%, while double indicate 100%. The data illustrate a human-specific expansion, as well as several independent expansions in the macaque, gorilla, and orangutan. Using macaque sequence as an outgroup, we estimate the human-specific expansion to be 2.31 MYA. d) Variation in human haplotypes across the first TBC1D3 expansion site. On the left is a graph representation (rGFA) of the locus where colors indicate the source genome for the sequence, and on the right the path for each haplotype-resolved assembly through the graph. The "squashed dot plot" represents a vertically compressed dot plot comparing the haplotype-resolved sequence (horizontal) against the graph (vertical). Color represents the source haplotype for the vertical sequence. Structural variants can be identified from discontinuities in height (deletion), changes between colors (insertion), or changes in the direction of the polygon (inversion). Below is shown the gene of interest (red arrow) and other genic content in the region (black arrow). The final line is a duplicons track, showing the ancestral duplications sites (color) that make up the larger duplication block.

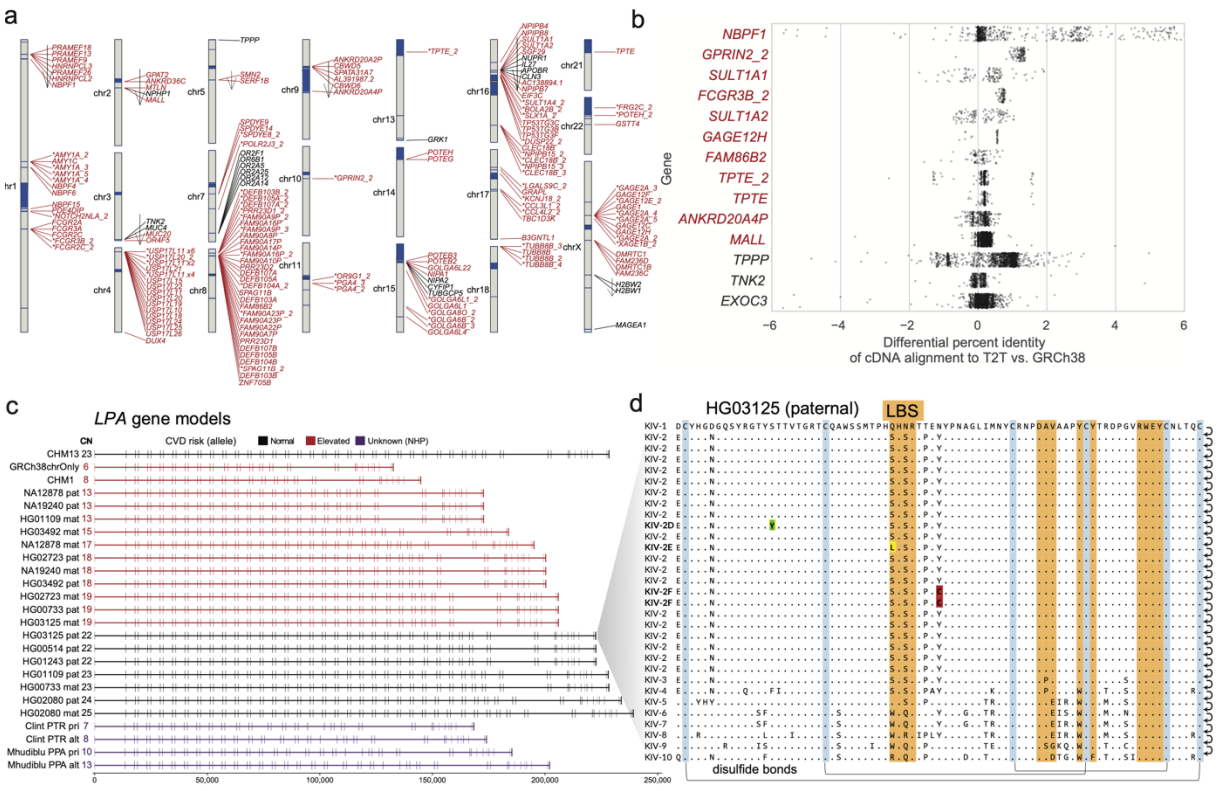


Figure 4.5. Genic insights in new SD regions of T2T CHM13.

a) Ideogram showing the new or non-syntenic gene models with ORFs, multiple exons, and at least 200 bp of CDS in the CHM13 T2T assembly as predicted by Liftoff. Genes colored red are in SD sequence and genes with asterisks are predicted to be an expansion in the gene family relative to GRCh38 (Methods). Arrows indicate inverted regions. b) Percent improvement in mapping of CHM13 Iso-Seq reads in segmentally duplicated (red) genes in the non-syntenic regions of the T2T CHM13 assembly. Reads with a positive value indicate a better alignment to T2T CHM13 than GRCh38. c) Gene models of LPA with ORF generated from haplotype-resolved HiFi assemblies. The double-exon repeat in these gene models encode for the Kringle IV subtype 2 domain of the LPA protein. Highlighted in red are haplotypes with low enough CN for Kringle IV subtype 2 that result in increased risk of CVD. d) Amino acid variation in the Kringle IV subtype 2 repeat in the paternal haplotype of HG01325. This haplotype contains a previously unknown set of amino acid substitutions: Ser42Leu in the active site, Ser24Tyr and Tyr49Cys.

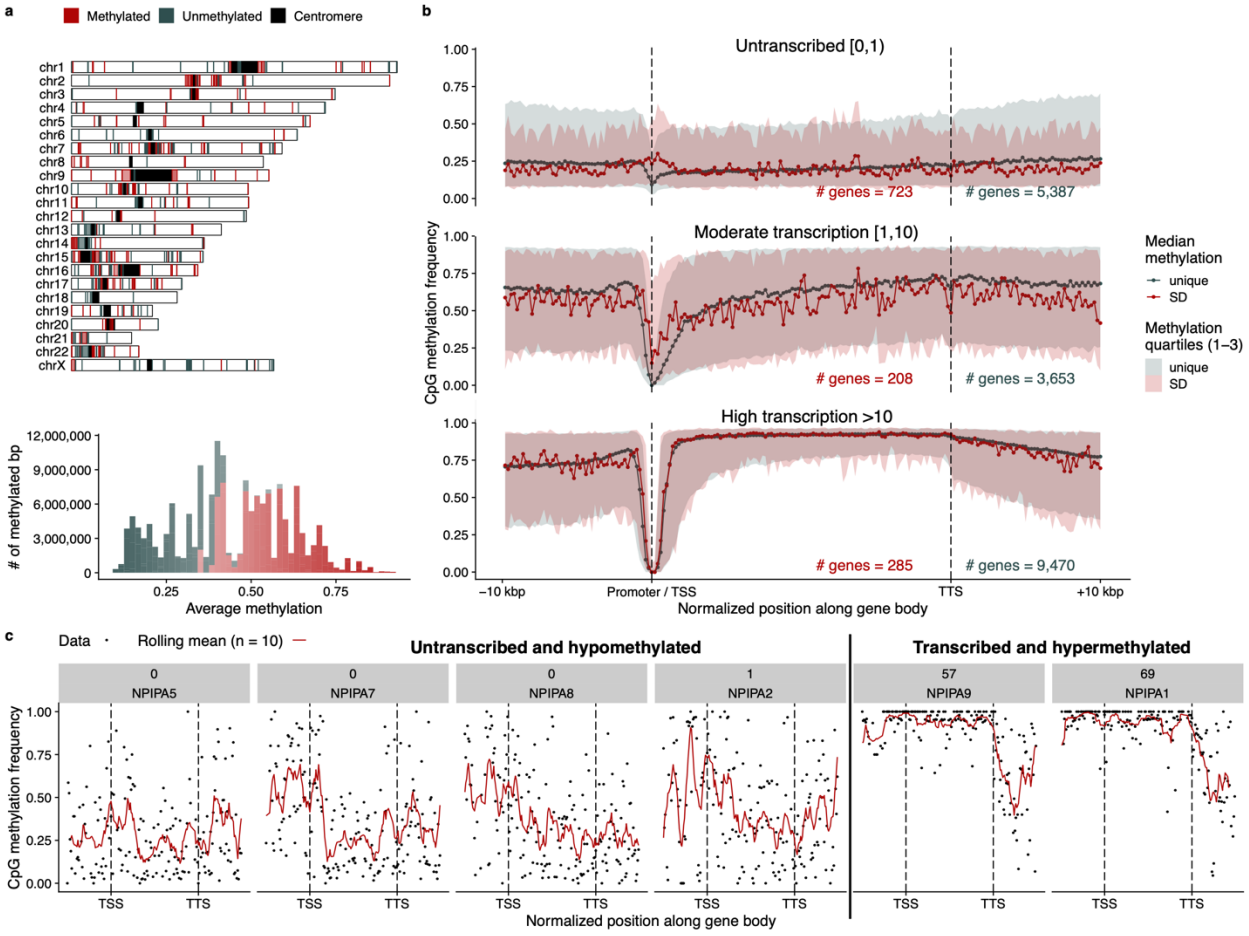


Figure 4.6. SD methylation and gene transcription.

a) Shows all SD blocks in the CHM13 genome with at least 50 kbp of unique flanking sequence and whether these sequences are hypermethylated (red) or hypomethylated (blue-gray). The histogram shows the distribution of average methylation across these regions. b) Median methylation signal of SD (red) and unique (blue-gray) genes stratified by their Iso-Seq expression levels in CHM13. The filled intervals represent the 25 and 75 quartiles of the observed data. c) Methylation signal across the *NPIPA* gene family in CHM13, showing increased methylation in transcriptionally active copies. Black points are individual methylation calls and the red line is a rolling mean across 10 methylation sites. The labels in gray show the number of CHM13 Iso-Seq transcripts and the gene name.

4.7 Tables

Table 4.6. Segmental duplication summary statistics.

Table 1. Summary statistics of segmental duplications in T2T CHM13 and GRCh38.

Assembly	*Gbp*	*% SD*	*SD (Mbp)*	*# SDs*	*inter (Mbp)*	*# inter*	*intra (Mbp)*	*# intra*	*acro (Mbp)*	*# acro*	*peri (Mbp)*	*# peri*	*telo (Mbp)*	*# telo*
T2T CHM13	3.11	6.67	207.56	41289	121.11	30484	142.96	10805	35.11	13264	88.61	24985	10.98	4998
GRCh38	3.11	5.37	167.30	24280	83.56	16348	120.71	7932	6.62	1407	53.94	10606	8.93	1529
Difference	0.00	1.29	40.27	17009	37.56	14136	22.25	2873	28.48	11857	34.66	14379	2.05	3469
New or structurally variable	0.24	33.88	81.34	25161	61.87	20579	54.93	4582	35.04	13258	54.04	19607	5.62	4005
T2T CHM13 with rDNA (estimate)	3.11	6.99	217.60	66042	131.15	49213	152.99	16829	45.14	38017	98.64	49738	10.98	4998

peri: within 5 Mbp of the centromere; telo: within 500 kbp of the telomere; acro: within the short arms of the acrocentric chromosomes

4.8 Acknowledgments

The authors thank T. Brown for help in editing this manuscript. This work was supported, in part, by grants from the U.S. National Institutes of Health (NIH grants 5R01HG002385 to E.E.E; 5U01HG010971 to E.E.E., I.M.H., D.H.H., and E.D.J.; and 1U01HG010973 to M.C., E.E.E., and M.T.). E.E.E. is an investigator of the Howard Hughes Medical Institute.

4.9 Author contributions

Identification of SDs in T2T CHM13 and analysis: M.R.V.; PacBio genome sequence generation: K.M.M, A.M.L., K.H.; FISH experiments and analysis; L.M., M.V., M.R.V., E.E.E; Iso-Seq analysis: P.C.D., M.R.V., R.L.; TBC1D3 analysis: X.G., M.R.V.; copy number analysis: M.R.V., W.T.H.; inversion analysis: D.P. M.R.V.; T2T CHM13 assembly generation: S.N., S.K., A.M.P.; refinement of SDs annotations near centromeres K.H.M., M.R.V.; UCSC browser: M.D., W.T.H., M.R.V.; methylation analysis: M.R.V, A.G., W.T., E.E.E.; organization of tables: M.R.V., P.C.D., X.G.; organization of supplementary material: M.R.V.; manuscript writing: M.R.V., E.E.E. X.G.; display items: M.R.V., X.G., P.C.D.

Chapter 5. DISCUSSION AND FUTURE DIRECTIONS

5.1 Impacts of the presented work

5.1.1 *The resolution and consequences of collapsed SDs in long-read assemblies*

There are several important conclusions from the work presented in the second chapter. First, our understanding of duplicated regions in primates has been inhibited by the presence of collapsed sequences in even the best reference assemblies (Kronenberg et al. 2018; Berlin et al. 2015; Chin et al. 2016; Koren et al. 2017; Seo et al. 2016; Shi et al. 2016; Steinberg et al. 2016). We estimate there is between 70-90 Mbp of collapsed sequence corresponding to regions of segmental duplication in human long-read assemblies. Furthermore, these collapses affect our interpretations of human variation in unexpected ways. The gene *GPRIN2*, which exists as a single copy in GRCh38, has 15 amino acid variants present in databases such as gnomAD and dbSNP that are in fact paralogous variants in a second missing copy of *GPRIN2*. Second, there exists sufficient paralogous variation to partially resolve most collapsed segmental duplications (SDs) using reads with lengths on the order of 10 kbp. Our tool, Segmental Duplication Assembler (SDA), is able to utilize this variation to resolve between 30-70 Mbp of previously collapsed sequence in human long-read assemblies. Finally, there are limitations of this work—the most important of which is phasing of human haplotypes. We know from previous studies that SDs contain half of all copy number variation (Sudmant et al. 2015a), but we are still unable to phase SDs into maternal and paternal haplotypes using SDA. This shortcoming likely leads to a significant amount of missed variation in our assemblies.

5.1.2 *Advances in sequencing technology and routine human genome assembly*

Chapter three focused on the introduction of a new sequence technology, PacBio HiFi, and its impacts on genome assembly. This work showed improvements in the accuracy, contiguity, and repeat resolution in human genome assembly when using the HiFi data type. More importantly, it demonstrated the feasibility of human genome assembly with resources available to most research groups. Between 2016 and 2018, Kronenberg et al. used long-read sequencing to generate a new high-quality reference genome for gorilla; this effort cost over $70,000 in sequencing reagents and took more than 50,000 CPU hours to assemble (Kronenberg et al. 2018). In this paper we assemble a human genome of comparable size and improved base-pair quality with less than $10,000 in sequencing costs and only 5,000 hours of CPU time. In the same year new assembly methods were able to assemble and phase a human genome in less than 100 CPU hours (Chin and Khalak 2019; Porubsky et al. 2020), less time than it takes to map a high-coverage short-read dataset to the human reference (Logsdon et al. 2020a). With more high-quality phased assemblies, it will be possible to access human variation by comparing assemblies thereby removing the reference bias and size constraints implicit in variant calling via read mapping.

5.1.3 *Sequence-resolved variation in SDs for nearly complete human genomes*

Chapter four uses a telomere-to-telomere (T2T) assembly of a complete hydatidiform mole (CHM) to characterize the SD landscape in a near-complete human genome. A remarkable finding of this work was that 81 Mbp of SD sequence was new or structurally variable compared to GRCh38. Of these 81 Mbp, half represent sequence assembled for the first time (35 Mbp from acrocentric short arms and ~5 Mbp of sequence from pericentromeric regions on chr1, chr9, and chr16), but the remaining 41 Mbp appears to be structurally different between the two assemblies. There are three possibilities that explain this observation: 1) One or both of the assemblies are

often wrong creating the appearance of large structural variants. 2) For 23% (41/179 Mbp) of SDs there is real variation between GRCh38 and T2T CHM13 indicating that SDs are highly polymorphic. 3) Or it could be some combination of the previous possibilities. Our results indicate the second option is by far the most probable since comparison of k-mers in the assembly and short-read data, estimates of copy number, Bionano validation, Strand-seq validation, and BAC assessments all indicate that the T2T assembly of CHM13 is of the highest quality. Even this estimate of variation may be a lower bound since it does not reflect any comparison from the acrocentric short arms given they are unassembled in GRCh38. However, we do show that of 38 acrocentric duplications tested with FISH, 16 were polymorphic across six samples. Moreover, in the 10 loci we targeted for assembly in 11 other humans, we saw frequent variation that matched or exceeded the variation between T2T CHM13 and GRCh38. Together these results indicate that SDs may be an even larger force in human variation than previously anticipated affecting 10-fold more bases than single-nucleotide variants (SNVs; ~4 million bp), (1000 Genomes Project Consortium et al. 2015).

5.2 Looking forward

5.2.1 ***Diploid telomere to telomere assemblies***

In the past year there have been two impactful developments in the assembly of human genomes. PacBio HiFi and ultra-long ONT have made possible not only the assembly of a complete human chromosome (Miga et al. 2020; Logsdon et al. 2020b) but, in fact, the entire human genome (Nurk et al., unpublished) (Figure 5.1). We now have the ability to fully phase human haplotypes through the use of orthogonal sequencing technologies, whether it be with parental short-read information (Koren et al. 2018; Cheng et al. 2021), Strand-seq (Porubsky et al. 2020), or Hi-C (Garg et al. 2020).

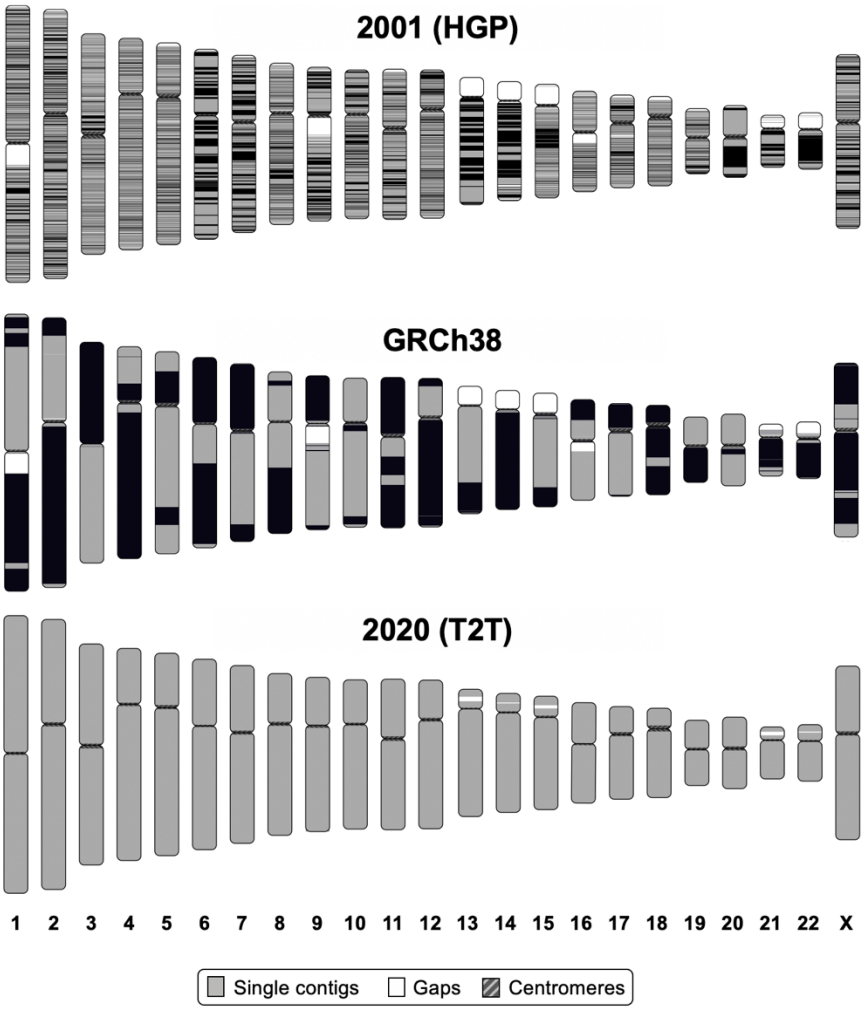


Figure 5.1. The progression of human genome assembly (credit G. A. Logsdon)

The clear next step in human assembly is at the junction of these two advances: orthogonal sequencing technologies will be combined with PacBio HiFi and ultra-long ONT to allow for fully phased telomere-to-telomere assemblies of diploid individuals. For select regions on chromosome 8 (e.g., the centromere), we have demonstrated that through careful inspection it is possible to phase and assemble the most complex regions of the genome using a combination HiFi and ultra-long ONT (Logsdon et al. 2020b). Automating and extending these methods will be essential to for fully phasing and assembling human chromosomes.

This goal is quite clear, but there are limitations in the current advances. As of now there is still no assembly for the rDNA arrays on the acrocentric chromosomes and no complete assembly of chromosome Y. The Y chromosome could be the hardest challenge yet due to sequences transposed from chromosome X with over 99% identity, X-degenerate sequences, large amplicon arrays with over 99.9% identity that include significant amounts (30%) of the euchromatin, and at least eight large (>1 Mbp) palindromic duplications containing testis genes (Skaletsky et al. 2003). In addition, the work presented here shows that some of the most complex duplications, like *SMN1* and *SMN2* on chromosome 5, are only assembled 25% of the time in human haplotypes.

There are also limitations in phasing that still need to be addressed. In addition to being very difficult to produce, Strand-seq requires mapping short-read data to identify informative SNVs, which limits its application to SDs (Porubsky et al. 2020; Sanders et al. 2017). Hi-C shares this dependency on short-read alignments but also has limitations in long-range phasing because mate pairs can only be a few megabases apart before interchromosomal interactions dominate the signal (Garg et al. 2020). And while parental short-read data avoids all of these limitations through the identification of k-mers that are unique to the maternal and paternal haplotypes, it requires parental sequencing information, which is often unavailable (Koren et al. 2018; Cheng et al. 2021).

Despite these limitations, with the many recent advances in sequencing technology and increased investment from the scientific community these will likely not be obstacles for long.

5.2.2 ***Encoding evolutionary history and variation in SDs***

In 2007 Jiang et al. used pairwise alignments of SDs to construct a repeat graph (modified A-Brujin graph) that could be used to define blocks of segmental duplications that shared an ancestral state (Jiang et al. 2007). These units of shared ancestral state were called ancestral

duplicons and they provided a vocabulary to describe the shared history within and between the mosaic SDs seen in primates. An updated version of this analysis using the new duplications in the T2T CHM13 assembly and better references for outgroup species (e.g., macaque) (Warren et al. 2020a) would likely reveal many duplicons missed in the initial 2007 analysis.

However, we are also in need of a new method that can define which duplications are orthologous and which are paralogous between multiple (>2) humans with potentially different duplication architectures. These methods are necessary so that we can identify the conserved elements of duplications that are likely to be functional (Dennis et al. 2012), which will then allow us to make associations between variants in conserved elements and phenotype. While Jiang's method revealed the shared state of SDs spanning tens of millions of years of evolution, we now need a new strategy that can operate on a much smaller time scale in order to compare the variation we see within humans.

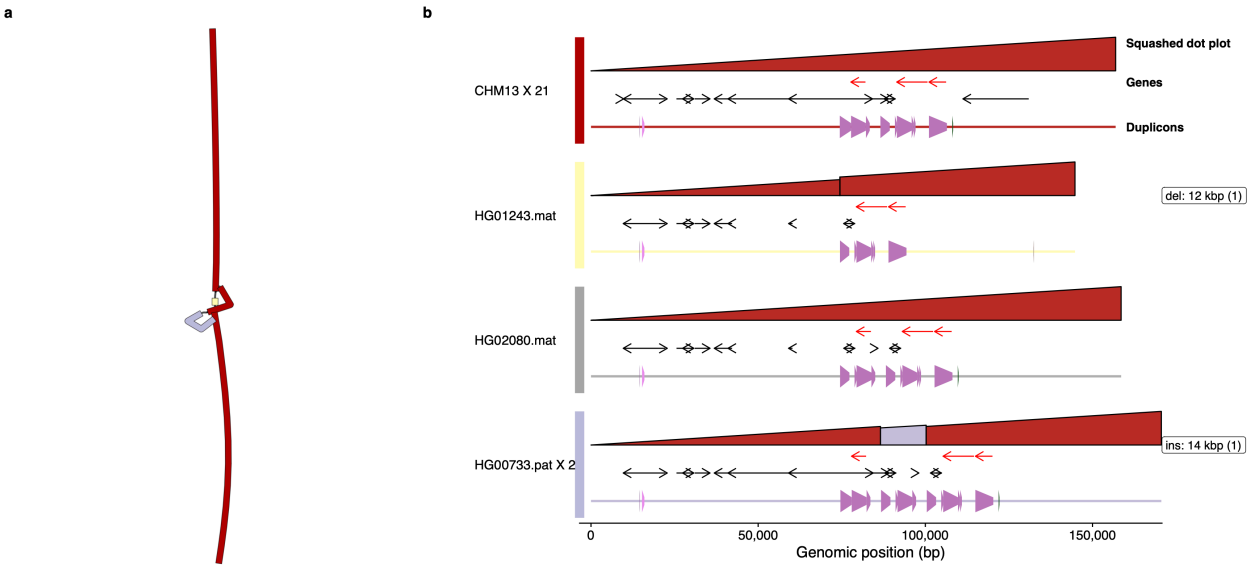


Figure 5.2. Variation in *CYP2D6* across human haplotypes.

a) Graph representation of the locus where colors indicate the source genome for the sequence. b) The path for each haplotype-resolved assembly through the graph. The "squashed dot plot" represents a vertically compressed dot plot comparing the haplotype-resolved sequence (horizontal) against the graph (vertical). Color represents the source

haplotype for the vertical sequence. Structural variants can be identified from discontinuities in height (deletion), changes between colors (insertion), or changes in the direction of the polygon (inversion). Below is shown the gene of interest (red arrow) and other genic content in the region (black arrow). The final line is a duplicon track showing the ancestral duplications (color).

There are some tools that currently exist that encode information from multiple humans into a variation graph (Garrison et al. 2018; Li et al. 2020) (Figure 5.2); however, to date these tools do not take into account the history of repetitive sequence when encoding the graph. There is a set of tools that uses phylogenetic trees to inform whole-genome alignments called Cactus (Armstrong et al. 2020; Paten et al. 2011b, 2011a), but its intraspecies pangenome construction relies on minigraph alignments (Li et al. 2020) that have failed to reliably infer the history of highly identical duplicated loci in our experience.

The need of a solution to this problem is imminent. The Human Pangenome Reference Consortium (HPRC), which was formed to create phased assemblies for 350 individuals to explore the full extent of human variation, has just released all the data from its first year of funding (https://humanpangenome.org/year-1-sequencing-data-release/). This data includes a multitude of sequencing types (PacBio HiFi, ultra-long ONT, Strand-seq, Bionano, Illumina Hi-C, and parental short-read data) for 45 human samples (Table 5.7). Furthermore, the HPRC has taken steps to choose a more representative of set of individuals to capture human variation, including 20 African individuals and only two Europeans. Given that the T2T consortium has achieved gapless assembly of a haploid human genome with these sequencing technologies, it is likely that in the near future the same feat will be accomplished in a diploid. If this advance is achieved, the data from these 45 individuals will provide 90 haplotypes with near-complete sequence resolution. Comparing the variation between 90 haplotypes with even the most conservative estimates of variation in SDs will be much too complex for current methods. Adding to this problem, gene

Table 5.7. Summary of HPRC samples sequenced with HiFi and ultra-long ONT

	Sample	Super Population	Population
1	HG00438	EAS	Southern Han Chinese
2	HG00514	EAS	Southern Han Chinese
3	HG00621	EAS	Southern Han Chinese
4	HG00673	EAS	Southern Han Chinese
5	HG00733	AMR	Puerto Ricans from Puerto Rico
6	HG00735	AMR	Puerto Ricans from Puerto Rico
7	HG00741	AMR	Puerto Ricans from Puerto Rico
8	HG01071	AMR	Puerto Ricans from Puerto Rico
9	HG01106	AMR	Puerto Ricans from Puerto Rico
10	HG01109	AMR	Puerto Ricans from Puerto Rico
11	HG01123	AMR	Colombians from Medellin, Colombia
12	HG01175	AMR	Puerto Ricans from Puerto Rico
13	HG01243	AMR	Puerto Ricans from Puerto Rico
14	HG01258	AMR	Colombians from Medellin, Colombia
15	HG01358	AMR	Colombians from Medellin, Colombia
16	HG01361	AMR	Colombians from Medellin, Colombia
17	HG01891	AFR	African Caribbeans in Barbados
18	HG01928	AMR	Peruvians from Lima, Peru
19	HG01952	AMR	Peruvians from Lima, Peru
20	HG01978	AMR	Peruvians from Lima, Peru
21	HG02080	EAS	Kinh in Ho Chi Minh City, Vietnam
22	HG02148	AMR	Peruvians from Lima, Peru
23	HG02257	AFR	African Caribbeans in Barbados
24	HG02486	AFR	African Caribbeans in Barbados
25	HG02559	AFR	African Caribbeans in Barbados
26	HG02572	AFR	Gambian in Western Divisions in the Gambia
27	HG02622	AFR	Gambian in Western Divisions in the Gambia
28	HG02630	AFR	Gambian in Western Divisions in the Gambia
29	HG02717	AFR	Gambian in Western Divisions in the Gambia
30	HG02723	AFR	Gambian in Western Divisions in the Gambia
31	HG02818	AFR	Gambian in Western Divisions in the Gambia
32	HG02886	AFR	Gambian in Western Divisions in the Gambia
33	HG03125	AFR	Esan in Nigeria
34	HG03453	AFR	Mende in Sierra Leone
35	HG03486	AFR	Mende in Sierra Leone
36	HG03492	SAS	Punjabi from Lahore, Pakistan
37	HG03516	AFR	Esan in Nigeria
38	HG03540	AFR	Gambian in Western Divisions in the Gambia
39	HG03579	AFR	Mende in Sierra Leone
40	NA12878	EUR	Utah Residents (CEPH) with Northern and Western European Ancestry
41	HG002	EUR	Ashkenazi
42	NA18906	AFR	Yoruba in Ibadan, Nigeria
43	NA19240	AFR	Yoruba in Ibadan, Nigeria
44	NA20129	AFR	Americans of African Ancestry in SW USA
45	NA21309	AFR	Maasai, Kenya

conversion will take otherwise orthologous loci and add paralogous variation making it even more difficult to distinguish.

As has often been the case for complex regions of the genome, I believe manual inspection and evaluation of specific loci will be essential in guiding the development of tools that can make genome-wide inferences about the relationships between segmental duplications across samples.

5.2.3 *Functional annotation of duplications*

Once we have the ability to comprehensively assemble and compare segmental duplications at the genetic level, we will have the foundation for functional analysis of these repetitive regions. Transcriptionally active SDs can be annotated by long-read cDNA and RNA from PacBio and ONT, respectively. And with complete assemblies generated from the same source materials as the transcript data, there will not be ambiguity in the alignments stemming from allelic versus paralogous variation that has plagued previous analyses in recent duplications (Dougherty et al. 2018, 2017; Vollger et al. 2019). Furthermore, developments in single-cell long-read transcriptional sequencing (Joglekar et al. 2020; Zheng et al. 2020) in combination with iPSC-derived neurons (Fiddes et al. 2018) may allow us to investigate human-specific duplications that have been implicated in human cognition and neurodevelopment at higher resolution (*ARHGAP11B*, *TBC1D3*, *NOTCH2NL*, *SRGAP2C*).

There has already been much progress annotating the transcriptome over SDs (Dougherty et al. 2018, 2017; Hsieh et al. 2019; Sahlin et al. 2018; Cantsilieris et al. 2020; Warren et al. 2020b; Kronenberg et al. 2018), and sequencing technologies are expanding to the next frontier: the epigenome. It is possible to automatically access methylation without modification of the sequencing protocol using ONT sequencing (Simpson et al. 2017). This technology has been used in part to identify the location of the functional kinetochore in centromere arrays (Miga et al. 2020; Logsdon et al. 2020b) and could be used to identify methylation differences associated with disease in tandem repeats (Sutcliffe et al. 1992; Eichler et al. 1994; T. and M 2010; Sakamoto et al. 1999; Saveliev et al. 2003; Tian et al. 2019).

In addition to methylation signatures, long-read technologies can also provide insight into the accessibility of DNA sequences. Two recent methods have shown that with the kinetic

information from PacBio sequencing it is possible to infer the location of histones on single molecules of DNA thereby revealing the accessibility of the sequence (Stergachis et al. 2020; Abdulhay et al. 2020). These methods differ from previous accessibility assays in two important ways: 1) They are based in long reads, so they can be utilized in repetitive regions. 2) The heterogeneity of accessibility can be quantified because the readout is per DNA molecule. Without per-molecule readout, a moderate peak in accessibility from a DNase footprinting assay could arise from all DNA molecules being partially accessible, or that half being totally inaccessible and half completely accessible. The single-molecule readout of the long-read methods can answer these questions unambiguously, which could be particularly important in haplotype-specific epigenetic signatures.

5.3 Addressing the aims of this dissertation

The goal of this thesis was to develop methods to uncover the full extent of human sequence variation in segmental duplications. By building a complete human genome and comparing its segmental duplication content to other samples, I have addressed many of my chief objectives. However, there are numerous outstanding questions important to the completion of this work: How do we assemble a human diploid from telomere to telomere? How do we compare complex structural variation across many human haplotypes? How do we functionally annotate segmental duplications and their variants in the human population?

In this discussion I have outlined some of the immediate steps I see as essential in addressing these questions, and progress in these areas will rely on improvements in sequencing technologies and computational approaches that utilize these improvements. Currently, I believe that methods for analysis are lagging behind our sequencing capabilities; however, problems that

seem sizable right now could become trivial given new technologies (e.g., phasing human haplotypes given high accuracy reads of ~50 kbp in length).

I predict that within the next five years we will have comprehensive representations of human variation and the ability to compare complex regions of the genome across many haplotypes. Furthermore, assembly of diploid samples from telomere to telomere will be as easy as incomplete assembly is today. I also predict that advances in our ability to compare complex variation will be accomplished primarily through the development of new algorithms and that diploid T2T assembly will be possible primarily through new or improved sequencing technologies.

5.4 A final remark

My thoughts on how a complete view of segmental duplications will impact human genetics unsurprisingly echo the thoughts of my advisor.

> In short, exceptional duplicated regions underlie exceptional biology. Consequently, I look forward with great anticipation to the unabridged version of the human genome.
>
> – Evan E. Eichler (Eichler 2001)

It is also with great anticipation that I look forward to utilizing multiple unabridged human genomes to explore exceptional duplications. I hope only that I do not have to wait quite as long as my advisor has.

BIBLIOGRAPHY

1000 Genomes Project Consortium, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, Marchini JL, McCarthy S, McVean GA, et al. 2015. A global reference for human genetic variation. *Nature* **526**: 68–74. http://dx.doi.org/10.1038/nature15393.

Abdulhay NJ, McNally CP, Hsieh LJ, Kasinathan S, Keith A, Estes LS, Karimzadeh M, Underwood JG, Goodarzi H, Narlikar GJ, et al. 2020. Massively multiplex single-molecule oligonucleosome footprinting. *Elife* **9**. http://dx.doi.org/10.7554/eLife.59404.

Abegglen LM, Caulin AF, Chan A, Lee K, Robinson R, Campbell MS, Kiso WK, Schmitt DL, Waddell PJ, Bhaskara S, et al. 2015. Potential mechanisms for cancer resistance in elephants and comparative cellular response to DNA damage in humans. *JAMA* **314**: 1850. http://dx.doi.org/10.1001/jama.2015.13134.

Aguiar D, Istrail S. 2013. Haplotype assembly in polyploid genomes and identical by descent shared tracts. *Bioinformatics* **29**: 352–360. http://dx.doi.org/10.1093/bioinformatics/btt213.

Ailon N, Charikar M, Newman A. 2005. Aggregating inconsistent information. In *Proceedings of the thirty-seventh annual ACM symposium on Theory of computing - STOC '05*, ACM Press, New York, New York, USA http://dx.doi.org/10.1145/1060590.1060692.

Alexanderson B, Evans DA, Sjöqvist F. 1969. Steady-state plasma levels of nortriptyline in twins: influence of genetic factors and drug therapy. *Br Med J* **4**: 764–768. http://dx.doi.org/10.1136/bmj.4.5686.764.

Alkan C, Coe BP, Eichler EE. 2011a. Genome structural variation discovery and genotyping. *Nat Rev Genet* **12**: 363–376. http://dx.doi.org/10.1038/nrg2958.

Alkan C, Kidd JM, Marques-Bonet T, Aksay G, Antonacci F, Hormozdiari F, Kitzman JO, Baker C, Malig M, Mutlu O, et al. 2009. Personalized copy number and segmental duplication maps using next-generation sequencing. *Nat Genet* **41**: 1061–1067. http://dx.doi.org/10.1038/ng.437.

Alkan C, Sajjadian S, Eichler EE. 2011b. Limitations of next-generation genome sequence assembly. *Nat Methods* **8**: 61–65. http://dx.doi.org/10.1038/nmeth.1527.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J Mol Biol* **215**: 403–410. http://www.sciencedirect.com/science/article/pii/S0022283605803602.

Antonacci F, Kidd JM, Marques-Bonet T, Teague B, Ventura M, Girirajan S, Alkan C, Campbell CD, Vives L, Malig M, et al. 2010. A large and complex structural polymorphism at 16p12.1 underlies microdeletion disease risk. *Nat Genet* **42**: 745–750. http://dx.doi.org/10.1038/ng.643.

Armstrong J, Hickey G, Diekhans M, Fiddes IT, Novak AM, Deran A, Fang Q, Xie D, Feng S, Stiller J, et al. 2020. Progressive Cactus is a multiple-genome aligner for the thousand-genome era. *Nature* **587**: 246–251. http://dx.doi.org/10.1038/s41586-020-2871-y.

Arnheim N, Nei M, Koehn RK. 1983. Evolution of genes and proteins. *Sinauer, Sunderland, MA* 38–61.

Artyomenko A, Wu NC, Mangul S, Eskin E, Sun R, Zelikovsky A. 2017. Long Single-Molecule Reads Can Resolve the Complexity of the Influenza Virus Composed of Rare, Closely Related Mutant Variants. *J Comput Biol* **24**: 558–570. http://online.liebertpub.com/doi/10.1089/cmb.2016.0146.

Audano PA, Sulovari A, Graves-Lindsay TA, Cantsilieris S, Sorensen M, Welch AE, Dougherty ML, Nelson BJ, Shah A, Dutcher SK, et al. 2019. Characterizing the Major Structural Variant Alleles of the Human Genome. *Cell* **176**: 663-675.e19. http://dx.doi.org/10.1016/j.cell.2018.12.019.

Bailey JA, Gu Z, Clark RA, Reinert K, Samonte RV, Schwartz S, Adams MD, Myers EW, Li PW, Eichler EE. 2002. Recent segmental duplications in the human genome. *Science* **297**: 1003–1007. http://dx.doi.org/10.1126/science.1072047.

Bailey JA, Yavor AM, Massa HF, Trask BJ, Eichler EE. 2001. Segmental duplications: organization and impact within the current human genome project assembly. *Genome Res* **11**: 1005–1017. http://dx.doi.org/10.1101/gr.gr-1871r.

Barber JC. 1994. Euchromatic heteromorphism or duplication without phenotypic effect? *Prenat Diagn* **14**: 323–324. http://dx.doi.org/10.1002/pd.1970140418.

Barber JCK, Zhang S, Friend N, Collins AL, Maloney VK, Hastings R, Farren B, Barnicoat A, Polityko AD, Rumyantseva NV, et al. 2006. Duplications of proximal 16q flanked by heterochromatin are not euchromatic variants and show no evidence of heterochromatic position effect. *Cytogenet Genome Res* **114**: 351–358. https://www.karger.com/DOI/10.1159/000094225.

Bebee TW, Thomas, Bebee W. 2010. Splicing regulation of the Survival Motor Neuron genes and implications for treatment of spinal muscular atrophy. *Frontiers in Bioscience* **15**: 1191. http://dx.doi.org/10.2741/3670.

Bell AD, Usher CL, McCarroll SA. 2018. Analyzing Copy Number Variation with Droplet Digital PCR. *Methods Mol Biol* **1768**: 143–160. http://dx.doi.org/10.1007/978-1-4939-7778-9_9.

Benson G. 1999. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res* **27**: 573–580. http://dx.doi.org/10.1093/nar/27.2.573.

Berger E, Yorukoglu D, Peng J, Berger B. 2014. HapTree: a novel Bayesian framework for single individual polyplotyping using NGS data. *PLoS Comput Biol* **10**: e1003502. http://dx.doi.org/10.1371/journal.pcbi.1003502.

Berger MP, Munson PJ. 1991. A novel randomized iterative strategy for aligning multiple protein sequences. *Comput Appl Biosci* **7**: 479–484. http://dx.doi.org/10.1093/bioinformatics/7.4.479.

Berlin K, Koren S, Chin C-S, Drake JP, Landolin JM, Phillippy AM. 2015. Assembling large genomes with single-molecule sequencing and locality-sensitive hashing. *Nat Biotechnol* **33**: 623–630. http://www.nature.com/doifinder/10.1038/nbt.3238.

Bertilsson L, Dahl M-L, Dalén P, Al-Shurbaji A. 2002. Molecular genetics of CYP2D6: clinical relevance with focus on psychotropic drugs. *Br J Clin Pharmacol* **53**: 111–122. http://dx.doi.org/10.1046/j.0306-5251.2001.01548.x.

Bhasin MK. 2005. Human population cytogenetics: A review. *Int J Hum Genet* **5**: 83–152. https://www.tandfonline.com/doi/full/10.1080/09723757.2005.11885918.

Bickhart DM, Rosen BD, Koren S, Sayre BL, Hastie AR, Chan S, Lee J, Lam ET, Liachko I, Sullivan ST, et al. 2017. Single-molecule sequencing and chromatin conformation capture enable de novo reference assembly of the domestic goat genome. *Nat Genet* **49**: 643–650. http://dx.doi.org/10.1038/ng.3802.

Bonizzoni P, Dondi R, Klau GW, Pirola Y, Pisanti N, Zaccaria S. 2016. On the minimum error correction problem for haplotype assembly in diploid and polyploid genomes. *J Comput Biol* **23**: 718–736. http://online.liebertpub.com/doi/10.1089/cmb.2015.0220.

Bridges CB. 1936. The bar "gene" a duplication. *Science* **83**: 210–211. http://dx.doi.org/10.1126/science.83.2148.210.

Butchbach MER. 2016. Copy Number Variations in the Survival Motor Neuron Genes: Implications for Spinal Muscular Atrophy and Other Neurodegenerative Diseases. *Front Mol Biosci* **3**: 7. http://dx.doi.org/10.3389/fmolb.2016.00007.

Caglayan AO, Ozyazgan I, Demiryilmaz F, Ozgun MT. 2010. Are heterochromatin polymorphisms associated with recurrent miscarriage? *J Obstet Gynaecol Res* **36**: 774–776. http://dx.doi.org/10.1111/j.1447-0756.2010.01207.x.

Cantsilieris S, Sunkin SM, Johnson ME, Anaclerio F, Huddleston J, Baker C, Dougherty ML, Underwood JG, Sulovari A, Hsieh P, et al. 2020. An evolutionary driver of interspersed segmental duplications in primates. *Genome Biol* **21**: 202. http://dx.doi.org/10.1186/s13059-020-02074-4.

Cardone MF, Alonso A, Pazienza M, Ventura M, Montemurro G, Carbone L, de Jong PJ, Stanyon R, D'Addabbo P, Archidiacono N, et al. 2006. Independent centromere formation in a capricious, gene-free domain of chromosome 13q21 in Old World monkeys and pigs. *Genome Biol* **7**: R91. http://dx.doi.org/10.1186/gb-2006-7-10-r91.

Chaisson MJ, Brinza D, Pevzner PA. 2009. De novo fragment assembly with short mate-paired reads: Does the read length matter? *Genome Res* **19**: 336–346. http://dx.doi.org/10.1101/gr.079053.108.

Chaisson MJ, Mukherjee S, Kannan S, Eichler EE. 2017. Resolving multicopy duplications de novo using polyploid phasing. *Res Comput Mol Biol* **10229**: 117–133. http://dx.doi.org/10.1007/978-3-319-56970-3_8.

Chaisson MJ, Tesler G. 2012. Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. *BMC Bioinformatics*.

Chaisson MJP, Huddleston J, Dennis MY, Sudmant PH, Malig M, Hormozdiari F, Antonacci F, Surti U, Sandstrom R, Boitano M, et al. 2015a. Resolving the complexity of the human genome using single-molecule sequencing. *Nature* **517**: 608–611. http://dx.doi.org/10.1038/nature13907.

Chaisson MJP, Sanders AD, Zhao X, Malhotra A, Porubsky D, Rausch T, Gardner EJ, Rodriguez OL, Guo L, Collins RL, et al. 2019. Multi-platform discovery of haplotype-resolved structural variation in human genomes. *Nat Commun* **10**: 1784. http://dx.doi.org/10.1038/s41467-018-08148-z.

Chaisson MJP, Wilson RK, Eichler EE. 2015b. Genetic variation and the de novo assembly of human genomes. *Nat Rev Genet* **16**: 627–640. http://dx.doi.org/10.1038/nrg3933.

Chen J, Huddleston J, Buckley RM, Malig M, Lawhon SD, Skow LC, Lee MO, Eichler EE, Andersson L, Womack JE. 2015. Bovine NK-lysin: Copy number variation and functional diversification. *Proc Natl Acad Sci U S A* 7223–7229. https://www.ncbi.nlm.nih.gov/pubmed/26668394.

Cheng H, Concepcion GT, Feng X, Zhang H, Li H. 2021. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat Methods* **18**: 170–175. http://dx.doi.org/10.1038/s41592-020-01056-5.

Cheng Z, Ventura M, She X, Khaitovich P, Graves T, Osoegawa K, Church D, DeJong P, Wilson RK, Pääbo S, et al. 2005. A genome-wide comparison of recent chimpanzee and human segmental duplications. *Nature* **437**: 88–93. http://dx.doi.org/10.1038/nature04000.

Chin CS, Alexander DH, Marks P, Klammer AA, Drake J, Heiner C, Clum A, Copeland A, Huddleston J, Eichler EE, et al. 2013. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nat Methods* **10**: 563–569. http://dx.doi.org/10.1038/nmeth.2474.

Chin C-S, Khalak A. 2019. Human Genome Assembly in 100 Minutes. *bioRxiv* 705616. https://www.biorxiv.org/content/10.1101/705616v1 (Accessed October 16, 2019).

Chin C-S, Peluso P, Sedlazeck FJ, Nattestad M, Concepcion GT, Clum A, Dunn C, O'Malley R, Figueroa-Balderas R, Morales-Cruz A, et al. 2016. Phased diploid genome assembly with Single Molecule Real-Time Sequencing. *Nat Methods* **13**: 056887. http://biorxiv.org/lookup/doi/10.1101/056887.

Clarke R, Peden JF, Hopewell JC, Kyriakou T, Goel A, Heath SC, Parish S, Barlera S, Franzosi MG, Rust S, et al. 2009. Genetic variants associated with Lp(a) lipoprotein level and coronary disease. *N Engl J Med* **361**: 2518–2528. http://dx.doi.org/10.1056/NEJMoa0902604.

Coassin S, Schönherr S, Weissensteiner H, Erhart G, Forer L, Losso JL, Lamina C, Haun M, Utermann G, Paulweber B, et al. 2019. A comprehensive map of single-base polymorphisms in the hypervariable LPA kringle IV type 2 copy number variation region. *J Lipid Res* **60**: 186–199. http://dx.doi.org/10.1194/jlr.M090381.

Codina-Pascual M, Navarro J, Oliver-Bonet M, Kraus J, Speicher MR, Arango O, Egozcue J, Benet J. 2006. Behaviour of human heterochromatic regions during the synapsis of homologous chromosomes. *Hum Reprod* **21**: 1490–1497. http://dx.doi.org/10.1093/humrep/del028.

Cooper GM, Coe BP, Girirajan S, Rosenfeld JA, Vu TH, Baker C, Williams C, Stalker H, Hamid R, Hannig V, et al. 2011. A copy number variation morbidity map of developmental delay. *Nat Genet* **43**: 838–846. http://dx.doi.org/10.1038/ng.909.

Dahl ML, Johansson I, Palmertz MP, Ingelman-Sundberg M, Sjöqvist F. 1992. Analysis of the CYP2D6 gene in relation to debrisoquin and desipramine hydroxylation in a Swedish population. *Clin Pharmacol Ther* **51**: 12–17. http://dx.doi.org/10.1038/clpt.1992.2.

Das S, Vikalo H. 2015. SDhaP: Haplotype assembly for diploids and polyploids via semi-definite programming. *BMC Genomics* **16**: 1–16. http://dx.doi.org/10.1186/s12864-015-1408-5.

Dennis MY, Eichler EE. 2016. Human adaptation and evolution by segmental duplication. *Current Opinion in Genetics and Development* **41**. http://dx.doi.org/10.1016/j.gde.2016.08.001.

Dennis MY, Harshman L, Nelson BJ, Penn O, Cantsilieris S, Huddleston J, Antonacci F, Penewit K, Denman L, Raja A, et al. 2017. The evolution and population diversity of human-specific segmental duplications. *Nat Ecol Evol* **1**: 69. http://dx.doi.org/10.1038/s41559-016-0069.

Dennis MY, Nuttle X, Sudmant PH, Antonacci F, Graves TA, Nefedov M, Rosenfeld JA, Sajjadian S, Malig M, Kotkiewicz H, et al. 2012. Evolution of human-specific neural SRGAP2 genes by incomplete segmental duplication. *Cell* **149**: 912–922. http://dx.doi.org/10.1016/j.cell.2012.03.033.

Dougherty ML, Nuttle X, Penn O, Nelson BJ, Huddleston J, Baker C, Harshman L, Duyzend MH, Ventura M, Antonacci F, et al. 2017. The birth of a human-specific neural gene by incomplete duplication and gene fusion. *Genome Biol* **18**: 1–16. http://dx.doi.org/10.1186/s13059-017-1163-9.

Dougherty ML, Underwood JG, Nelson BJ, Tseng E, Munson KM, Penn O, Nowakowski TJ, Pollen AA, Eichler EE. 2018. Transcriptional fates of human-specific segmental

duplications in brain. *Genome Res* **28**: 1566–1576. http://dx.doi.org/10.1101/gr.237610.118.

Ebert P, Audano PA, Zhu Q, Rodriguez-Martin B, Porubsky D, Bonder MJ, Sulovari A, Ebler J, Zhou W, Mari RS, et al. 2020. De novo assembly of 64 haplotype-resolved human genomes of diverse ancestry and integrated analysis of structural variation. *Cold Spring Harbor Laboratory* 2020.12.16.423102. https://www.biorxiv.org/content/10.1101/2020.12.16.423102v1.full (Accessed February 8, 2021).

Ebert P, Audano PA, Zhu Q, Rodriguez-Martin B, Porubsky D, Bonder MJ, Sulovari A, Ebler J, Zhou W, Mari RS, et al. 2021. Haplotype-resolved diverse human genomes and integrated analysis of structural variation. *Science*. https://science.sciencemag.org/content/early/2021/02/24/science.abf7117 (Accessed February 26, 2021).

Ebler J, Clarke WE, Rausch T, Audano PA, Houwaart T, Korbel J, Eichler EE, Zody MC, Dilthey AT, Marschall T. 2020. Pangenome-based genome inference. *Cold Spring Harbor Laboratory* 2020.11.11.378133. https://www.biorxiv.org/content/10.1101/2020.11.11.378133v1.abstract (Accessed February 8, 2021).

Eichler EE. 2001. Segmental Duplications: What's Missing, Misassigned, and Misassembled—and Should We Care? *Genome Res* **11**: 653–656. http://genome.cshlp.org/content/11/5/653.short.

Eichler EE, Holden JJ, Popovich BW, Reiss AL, Snow K, Thibodeau SN, Richards CS, Ward PA, Nelson DL. 1994. Length of uninterrupted CGG repeats determines instability in the FMR1 gene. *Nat Genet* **8**: 88–94. http://dx.doi.org/10.1038/ng0994-88.

Eichler EE, Surti U, Ophoff R. 2002. Proposal for Construction a Human Haploid BAC library from Hydatidiform Mole Source Material.

Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G, Peluso P, Rank D, Baybayan P, Bettman B, et al. 2009. Real-time DNA sequencing from single polymerase molecules. *Science* **323**: 133–138. http://dx.doi.org/10.1126/science.1162986.

Eizenga JM, Novak AM, Sibbesen JA, Heumos S, Ghaffaari A, Hickey G, Chang X, Seaman JD, Rounthwaite R, Ebler J, et al. 2020. Pangenome graphs. *Annu Rev Genomics Hum Genet* **21**: 139–162. http://dx.doi.org/10.1146/annurev-genom-120219-080406.

Emanuel BS, Shaikh TH. 2001. Segmental duplications: An "expanding" role in genomic instability and disease. *Nat Rev Genet* **2**: 791–800. http://dx.doi.org/10.1038/35093500.

Falconer E, Hills M, Naumann U, Poon SSS, Chavez EA, Sanders AD, Zhao Y, Hirst M, Lansdorp PM. 2012. DNA template strand sequencing of single-cells maps genomic rearrangements at high resolution. *Nat Methods* **9**: 1107–1112. http://dx.doi.org/10.1038/nmeth.2206.

Fiddes IT, Lodewijk GA, Mooring M, Bosworth CM, Ewing AD, Mantalas GL, Novak AM, van den Bout A, Bishara A, Rosenkrantz JL, et al. 2018. Human-specific NOTCH2NL genes affect notch signaling and cortical neurogenesis. *Cell* **173**: 1356-1369.e22. http://dx.doi.org/10.1016/j.cell.2018.03.051.

Florio M, Heide M, Pinson A, Brandl H, Albert M, Winkler S, Wimberger P, Huttner WB, Hiller M. 2018. Evolution and cell-type specificity of human-specific genes preferentially expressed in progenitors of fetal neocortex. *Elife* **7**. http://dx.doi.org/10.7554/elife.32332.

Force A, Lynch M, Pickett FB, Amores A, Yan YL, Postlethwait J. 1999. Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* **151**: 1531–1545. https://www.ncbi.nlm.nih.gov/pubmed/10101175.

Fredman D, White SJ, Potter S, Eichler EE, Den Dunnen JT, Brookes AJ. 2004. Complex SNP-related sequence variation in segmental genome duplications. *Nat Genet* **36**: 861–866. http://dx.doi.org/10.1038/ng1401.

Gaedigk A, Blum M, Gaedigk R, Eichelbaum M, Meyer UA. 1991. Deletion of the entire cytochrome P450 CYP2D6 gene as a cause of impaired drug metabolism in poor metabolizers of the debrisoquine/sparteine polymorphism. *Am J Hum Genet* **48**: 943–950. https://www.ncbi.nlm.nih.gov/pubmed/1673290.

Garg S, Fungtammasan A, Carroll A, Chou M, Schmitt A, Zhou X, Mac S, Peluso P, Hatas E, Ghurye J, et al. 2020. Chromosome-scale, haplotype-resolved assembly of human genomes. *Nat Biotechnol*. http://dx.doi.org/10.1038/s41587-020-0711-0.

Garrison E, Sirén J, Novak AM, Hickey G, Eizenga JM, Dawson ET, Jones W, Garg S, Markello C, Lin MF, et al. 2018. Variation graph toolkit improves read mapping by representing genetic variation in the reference. *Nat Biotechnol* **36**: 875–879. http://dx.doi.org/10.1038/nbt.4227.

Gel B, Serra E. 2017. karyoploteR: an R/Bioconductor package to plot customizable genomes displaying arbitrary data. *Bioinformatics* **33**: 3088–3090. http://dx.doi.org/10.1093/bioinformatics/btx346.

Gonzalez IL, Sylvester JE. 1995. Complete sequence of the 43-kb human ribosomal DNA repeat: analysis of the intergenic spacer. *Genomics* **27**: 320–328. http://dx.doi.org/10.1006/geno.1995.1049.

Gordon D, Huddleston J, Chaisson MJP, Hill CM, Kronenberg ZN, Munson KM, Malig M, Raja A, Fiddes I, Hillier LW, et al. 2016. Long-read sequence assembly of the gorilla genome. *Science* **352**: aae0344. http://dx.doi.org/10.1126/science.aae0344.

Gotoh O. 1993. Optimal alignment between groups of sequences and its application to multiple sequence alignment. *Comput Appl Biosci* **9**: 361–370. http://dx.doi.org/10.1093/bioinformatics/9.3.361.

GTEx Consortium, Laboratory, Data Analysis &Coordinating Center (LDACC)—Analysis Working Group, Statistical Methods groups—Analysis Working Group, Enhancing GTEx (eGTEx) groups, NIH Common Fund, NIH/NCI, NIH/NHGRI, NIH/NIMH, NIH/NIDA, Biospecimen Collection Source Site—NDRI, et al. 2017. Genetic effects on gene expression across human tissues. *Nature* **550**: 204–213. http://dx.doi.org/10.1038/nature24277.

Gu Z, Gu L, Eils R, Schlesner M, Brors B. 2014. circlize Implements and enhances circular visualization in R. *Bioinformatics* **30**: 2811–2812. http://dx.doi.org/10.1093/bioinformatics/btu393.

Gum JR, Hicks JW, Swallow DM, Lagace RL, Byrd JC, Lamport DT, Siddiki B, Kim YS. 1990. Molecular cloning of cDNAs derived from a novel human intestinal mucin gene. *Biochem Biophys Res Commun* **171**: 407–415. http://dx.doi.org/10.1016/0006-291x(90)91408-k.

Hach F, Hormozdiari F, Alkan C, Hormozdiari F, Birol I, Eichler EE, Sahinalp SC. 2010. MrsFAST: A cache-oblivious algorithm for short-read mapping. *Nat Methods* **7**: 576–577. http://dx.doi.org/10.1038/nmeth0810-576.

Hammer W, Sjöqvist F. 1967. Plasma levels of monomethylated tricyclic antidepressants during treatment with imipramine-like compounds. *Life Sciences* **6**: 1895–1903. http://dx.doi.org/10.1016/0024-3205(67)90218-4.

Heide M, Haffner C, Murayama A, Kurotaki Y, Shinohara H, Okano H, Sasaki E, Huttner WB. 2020. Human-specific ARHGAP11B increases size and folding of primate neocortex in the fetal marmoset. *Science* **369**: 546–550. http://dx.doi.org/10.1126/science.abb2401.

Hong Y, Zhou Y-W, Tao J, Wang S-X, Zhao X-M. 2011. Do polymorphic variants of chromosomes affect the outcome of in vitro fertilization and embryo transfer treatment? *Hum Reprod* **26**: 933–940. http://dx.doi.org/10.1093/humrep/deq333.

Howorka S, Cheley S, Bayley H. 2001. Sequence-specific detection of individual DNA strands using engineered nanopores. *Nat Biotechnol* **19**: 636–639. http://dx.doi.org/10.1038/90236.

Hsieh P, Vollger MR, Dang V, Porubsky D, Baker C, Cantsilieris S, Hoekzema K, Lewis AP, Munson KM, Sorensen M, et al. 2019. Adaptive archaic introgression of copy number variants and the discovery of previously unknown human genes. *Science* **366**. http://dx.doi.org/10.1126/science.aax2083.

Hsu LY, Benn PA, Tannenbaum HL, Perlis TE, Carlson AD. 1987. Chromosomal polymorphisms of 1, 9, 16, and Y in 4 major ethnic groups: a large prenatal study. *Am J Med Genet* **26**: 95–101. http://doi.wiley.com/10.1002/ajmg.1320260116.

Huddleston J, Chaisson MJP, Steinberg KM, Warren W, Hoekzema K, Gordon D, Graves-Lindsay TA, Munson KM, Kronenberg ZN, Vives L, et al. 2017. Discovery and

genotyping of structural variation from long-read haploid genome sequence data. *Genome Res* **27**: 677–685. http://dx.doi.org/10.1101/gr.214007.116.

Huddleston J, Ranade S, Malig M, Antonacci F, Chaisson M, Hon L, Sudmant PH, Graves TA, Alkan C, Dennis MY, et al. 2014. Reconstructing complex regions of genomes using long-read sequencing technology. *Genome Res* **24**: 688–696. http://dx.doi.org/10.1101/gr.168450.113.

Jain C, Dilthey A, Koren S, Aluru S, Phillippy AM. 2018a. A fast approximate algorithm for mapping long reads to large reference databases. *J Comput Biol* **25**: 766–779. http://dx.doi.org/10.1089/cmb.2018.0036.

Jain C, Rhie A, Zhang H, Chu C, Walenz BP, Koren S, Phillippy AM. 2020. Weighted minimizer sampling improves long read mapping. *Bioinformatics* **36**: i111–i118. http://dx.doi.org/10.1093/bioinformatics/btaa435.

Jain M, Koren S, Miga KH, Quick J, Rand AC, Sasani TA, Tyson JR, Beggs AD, Dilthey AT, Fiddes IT, et al. 2018b. Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nat Biotechnol* **36**. http://dx.doi.org/10.1038/nbt.4060.

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

Jiang Z, Tang H, Ventura M, Cardone MF, Marques-Bonet T, She X, Pevzner P a., Eichler EE. 2007. Ancestral reconstruction of segmental duplications reveals punctuated cores of human genome evolution. *Nat Genet* **39**: 1361–1368. http://dx.doi.org/10.1038/ng.2007.9.

Joglekar A, Prjibelski A, Mahfouz A, Collier P, Lin S, Schlusche AK, Marrocco J, Williams SR, Haase B, Hayes A, et al. 2020. Cell-type, single-cell, and spatial signatures of brain-region specific splicing in postnatal development. *Cold Spring Harbor Laboratory* 2020.08.27.268730. https://www.biorxiv.org/content/10.1101/2020.08.27.268730v1 (Accessed February 23, 2021).

Johansson I, Lundqvist E, Bertilsson L, Dahl ML, Sjöqvist F, Ingelman-Sundberg M. 1993. Inherited amplification of an active gene in the cytochrome P450 CYP2D locus as a cause of ultrarapid metabolism of debrisoquine. *Proc Natl Acad Sci U S A* **90**: 11825–11829. http://dx.doi.org/10.1073/pnas.90.24.11825.

Johnson ME, Viggiano L, Bailey JA, Abdul-Rauf M, Goodwin G, Rocchi M, Eichler EE. 2001. Positive selection of a gene family during the emergence of humans and African apes. *Nature* **413**: 514–519. http://dx.doi.org/10.1038/35097067.

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**. http://dx.doi.org/10.7554/eLife.18197.

Kalantari P, Sepehri H, Behjati F, Ashtiani ZO, Akbari MT. 2001. Chromosomal studies in infertile men. *Tsitol Genet* **35**: 50–54. https://www.ncbi.nlm.nih.gov/pubmed/11944328.

Katoh K, Misawa K, Kuma K-I, Miyata T. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res* **30**: 3059–3066. http://dx.doi.org/10.1093/nar/gkf436.

Kelley DR, Salzberg SL. 2010. Detection and correction of false segmental duplications caused by genome mis-assembly. *Genome Biol* **11**. http://dx.doi.org/10.1186/gb-2010-11-3-r28.

Kidd JM, Cooper GM, Donahue WF, Hayden HS, Sampas N, Graves T, Hansen N, Teague B, Alkan C, Antonacci F, et al. 2008. Mapping and sequencing of structural variation from eight human genomes. *Nature* **453**: 56–64. http://dx.doi.org/10.1038/nature06862.

Kidd JM, Graves T, Newman TL, Fulton R, Hayden HS, Malig M, Kallicki J, Kaul R, Wilson RK, Eichler EE. 2010. A human genome structural variation sequencing resource reveals insights into mutational mechanisms. *Cell* **143**: 837–847. http://dx.doi.org/10.1016/j.cell.2010.10.027.

Killick R, Haynes K, Eckley I, Fearnhead P, Lee J. 2016. Package 'changepoint.' *R package version 0 4 -2011 -http://cran rproject org/web/packages/changepoint/index html*. https://cran.r-project.org/web/packages/changepoint/changepoint.pdf.

Kim J-H, Dilthey AT, Nagaraja R, Lee H-S, Koren S, Dudekula D, Wood WH Iii, Piao Y, Ogurtsov AY, Utani K, et al. 2018. Variation in human chromosome 21 ribosomal RNA genes characterized by TAR cloning and long-read sequencing. *Nucleic Acids Res* **46**: 6712–6725. http://dx.doi.org/10.1093/nar/gky442.

King MC, Wilson AC. 1975. Evolution at two levels in humans and chimpanzees. *Science* **188**: 107–116. http://dx.doi.org/10.1126/science.1090005.

Korbel JO, Urban AE, Affourtit JP, Godwin B, Grubert F, Simons JF, Kim PM, Palejev D, Carriero NJ, Du L, et al. 2007. Paired-end mapping reveals extensive structural variation in the human genome. *Science* **318**: 420–426. http://dx.doi.org/10.1126/science.1149504.

Koren S, Rhie A, Walenz BP, Dilthey AT, Bickhart DM, Kingan SB, Hiendleder S, Williams JL, Smith TPL, Phillippy AM. 2018. De novo assembly of haplotype-resolved genomes with trio binning. *Nat Biotechnol* **36**: 1174–1182. http://dx.doi.org/10.1038/nbt.4277.

Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. 2017. Canu: scalable and accurate long-read assembly via adaptive k -mer weighting and repeat separation. *Genome Res* 1–33. http://genome.cshlp.org/lookup/doi/10.1101/gr.215087.116.

Köster J, Rahmann S. 2018. Snakemake-a scalable bioinformatics workflow engine. *Bioinformatics* **34**: 3600. http://dx.doi.org/10.1093/bioinformatics/bty350.

Krasheninnikova K, Diekhans M, Armstrong J, Dievskii A, Paten B, O'Brien S. 2020. halSynteny: a fast, easy-to-use conserved synteny block construction method for multiple whole-genome alignments. *Gigascience* **9**. http://dx.doi.org/10.1093/gigascience/giaa047.

Kronenberg F. 2016. Human Genetics and the Causal Role of Lipoprotein(a) for Various Diseases. *Cardiovasc Drugs Ther* **30**: 87–100. http://dx.doi.org/10.1007/s10557-016-6648-3.

Kronenberg F, Utermann G. 2013. Lipoprotein(a): resurrected by genetics. *J Intern Med* **273**: 6–30. http://dx.doi.org/10.1111/j.1365-2796.2012.02592.x.

Kronenberg ZN, Fiddes IT, Gordon D, Murali S, Cantsilieris S, Meyerson OS, Underwood JG, Nelson BJ, Chaisson MJPP, Dougherty ML, et al. 2018. High-resolution comparative analysis of great ape genomes. *Science* **6343**. http://dx.doi.org/10.1126/science.aar6343.

Kronenberg ZN, Rhie A, Koren S, Concepcion GT, Peluso P, Munson KM, Hiendleder S, Fedrigo O, Jarvis ED, Phillippy AM, et al. 2019. Extended haplotype phasing of de novo genome assemblies with FALCON-Phase. *bioRxiv* 327064. https://www.biorxiv.org/content/10.1101/327064v2.full (Accessed May 10, 2019).

Krumsiek J, Arnold R, Rattei T. 2007. Gepard: a rapid and sensitive tool for creating dotplots on genome scale. *Bioinformatics* **23**: 1026–1028. http://dx.doi.org/10.1093/bioinformatics/btm039.

Kyo K, Muto T, Nagawa H, Lathrop GM, Nakamura Y. 2001. Associations of distinct variants of the intestinal mucin gene MUC3A with ulcerative colitis and Crohn's disease. *J Hum Genet* **46**: 5–20. http://dx.doi.org/10.1007/s100380170118.

Kyogoku C, Dijstelbloem HM, Tsuchiya N, Hatta Y, Kato H, Yamaguchi A, Fukazawa T, Jansen MD, Hashimoto H, van de Winkel JGJ, et al. 2002. Fcgamma receptor gene polymorphisms in Japanese patients with systemic lupus erythematosus: contribution of FCGR2B to genetic susceptibility. *Arthritis Rheum* **46**: 1242–1254. http://dx.doi.org/10.1002/art.10257.

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, et al. 2001. Initial sequencing and analysis of the human genome. *Nature* **409**: 860–921. http://www.ncbi.nlm.nih.gov/pubmed/11237011\nhttp://www.nature.com/nature/journal/v409/n6822/pdf/409860a0.pdf.

Li H. 2018. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**: 3094–3100. http://dx.doi.org/10.1093/bioinformatics/bty191.

Li H, Bloom JM, Farjoun Y, Fleharty M, Gauthier L, Neale B, MacArthur D. 2018. A synthetic-diploid benchmark for accurate variant-calling evaluation. *Nat Methods* **15**: 595–597. http://dx.doi.org/10.1038/s41592-018-0054-7.

Li H, Feng X, Chu C. 2020. The design and construction of reference pangenome graphs with minigraph. *Genome Biol* **21**: 265. http://dx.doi.org/10.1186/s13059-020-02168-z.

Locke DP, Segraves R, Carbone L, Archidiacono N, Albertson DG, Pinkel D, Eichler EE. 2003. Large-scale variation among human and great ape genomes determined by array comparative genomic hybridization. *Genome Res* **13**: 347–357. http://dx.doi.org/10.1101/gr.1003303.

Locke DP, Segraves R, Nicholls RD, Schwartz S, Pinkel D, Albertson DG, Eichler EE. 2004. BAC microarray analysis of 15q11–q13 rearrangements and the impact of segmental duplications. *J Med Genet* **41**: 175–182. https://jmg.bmj.com/content/41/3/175.short (Accessed February 26, 2021).

Logsdon GA, Vollger MR, Eichler EE. 2020a. Long-read human genome sequencing and its applications. *Nat Rev Genet*. http://dx.doi.org/10.1038/s41576-020-0236-x.

Logsdon GA, Vollger MR, Hsieh P, Mao Y, Liskovykh MA, Koren S, Nurk S, Mercuri L, Dishuck PC, Rhie A, et al. 2020b. The structure, function, and evolution of a complete human chromosome 8. *Cold Spring Harbor Laboratory* 2020.09.08.285395. https://www.biorxiv.org/content/10.1101/2020.09.08.285395v1 (Accessed November 30, 2020).

Loman NJ, Quick J, Simpson JT. 2015. A complete bacterial genome assembled de novo using only nanopore sequencing data. *Nat Methods* **12**: 733–735. http://dx.doi.org/10.1038/nmeth.3444.

Low WY, Tearle R, Bickhart DM, Rosen BD, Kingan SB, Swale T, Thibaud-Nissen F, Murphy TD, Young R, Lefevre L, et al. 2019. Chromosome-level assembly of the water buffalo genome surpasses human and goat genomes in sequence contiguity. *Nat Commun* **10**: 260. http://dx.doi.org/10.1038/s41467-018-08260-0.

Luke S, Verma RS, Conte RA, Mathews T. 1992. Molecular characterization of the secondary constriction region (qh) of human chromosome 9 with pericentric inversion. *J Cell Sci* **103 (Pt 4)**: 919–923. https://www.ncbi.nlm.nih.gov/pubmed/1487504.

Lynch M, Conery JS. 2000. The evolutionary fate and consequences of duplicate genes. *Science* **290**: 1151–1155. http://dx.doi.org/10.1126/science.290.5494.1151.

Madon PF, Athalye AS, Parikh FR. 2005. Polymorphic variants on chromosomes probably play a significant role in infertility. *Reprod Biomed Online* **11**: 726–732. http://dx.doi.org/10.1016/s1472-6483(10)61691-4.

Mallick S, Li H, Lipson M, Mathieson I, Gymrek M, Racimo F, Zhao M, Chennagiri N, Nordenfelt S, Tandon A, et al. 2016. The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature* **538**: 201–206. http://dx.doi.org/10.1038/nature18964.

Marques-Bonet T, Eichler EE. 2009. The evolution of human segmental duplications and the core duplicon hypothesis. *Cold Spring Harb Symp Quant Biol* **74**: 355–362. http://dx.doi.org/10.1101/sqb.2009.74.011.

Marques-Bonet T, Kidd JM, Ventura M, Graves TA, Cheng Z, Hillier LW, Jiang Z, Baker C, Malfavon-Borja R, Fulton LA, et al. 2009. A burst of segmental duplications in the genome of the African great ape ancestor. *Nature* **457**: 877–881. http://dx.doi.org/10.1038/nature07744.

McCarroll SA, Hadnott TN, Perry GH, Sabeti PC, Zody MC, Barrett JC, Dallaire S, Gabriel SB, Lee C, Daly MJ, et al. 2006. Common deletion polymorphisms in the human genome. *Nat Genet* **38**: 86–92. http://dx.doi.org/10.1038/ng1696.

McCarroll SA, Kuruvilla FG, Korn JM, Cawley S, Nemesh J, Wysoker A, Shapero MH, de Bakker PIW, Maller JB, Kirby A, et al. 2008. Integrated detection and population-genetic analysis of SNPs and copy number variation. *Nat Genet* **40**: 1166–1174. http://dx.doi.org/10.1038/ng.238.

Medvedev P, Fiume M, Dzamba M, Smith T, Brudno M. 2010. Detecting copy number variation with mated short reads. *Genome Res* **20**: 1613–1622. http://dx.doi.org/10.1101/gr.106344.110.

Miga KH, Koren S, Rhie A, Vollger MR, Gershman A, Bzikadze A, Brooks S, Howe E, Porubsky D, Logsdon GA, et al. 2020. Telomere-to-telomere assembly of a complete human X chromosome. *Nature* **585**: 79–84. http://dx.doi.org/10.1038/s41586-020-2547-7.

Mills RE, 1000 Genomes Project, Walter K, Stewart C, Handsaker RE, Chen K, Alkan C, Abyzov A, Yoon SC, Ye K, et al. 2011. Mapping copy number variation by population-scale genome sequencing. *Nature* **470**: 59–65. http://dx.doi.org/10.1038/nature09708.

Minocherhomji S, Athalye AS, Madon PF, Kulkarni D, Uttamchandani SA, Parikh FR. 2009. A case-control study identifying chromosomal polymorphic variations as forms of epigenetic alterations associated with the infertility phenotype. *Fertil Steril* **92**: 88–95. http://dx.doi.org/10.1016/j.fertnstert.2008.05.071.

Mohajeri K, Cantsilieris S, Huddleston J, Nelson BJ, Coe BP, Campbell CD, Baker C, Harshman L, Munson KM, Kronenberg ZN, et al. 2016. Interchromosomal core duplicons drive both evolutionary instability and disease susceptibility of the Chromosome 8p23.1 region. *Genome Res* **26**: 1453–1467. http://dx.doi.org/10.1101/gr.211284.116.

Muller HJ. 1936. BAR DUPLICATION. *Science* **83**: 528–530. http://dx.doi.org/10.1126/science.83.2161.528-a.

Myers EW. 2005. The fragment assembly string graph. *Bioinformatics* **21**: 79–85. http://dx.doi.org/10.1093/bioinformatics/bti1114.

Myers EW. 1995. Toward Simplifying and Accurately Formulating Fragment Assembly. *J Comput Biol* **2**: 275–290. https://doi.org/10.1089/cmb.1995.2.275.

Myers EW, Sutton GG, Delcher AL, Dew IM, Fasulo DP, Flanigan MJ, Kravitz SA, Mobarry CM, Reinert KH, Remington KA, et al. 2000. A whole-genome assembly of Drosophila. *Science* **287**: 2196–2204. http://dx.doi.org/10.1126/science.287.5461.2196.

Numanagic I, Gökkaya AS, Zhang L, Berger B, Alkan C, Hach F. 2018. Fast characterization of segmental duplications in genome assemblies. *Bioinformatics* **34**: i706–i714. http://dx.doi.org/10.1093/bioinformatics/bty586.

Nurk S, Walenz BP, Rhie A, Vollger MR, Logsdon GA, Grothe R, Miga KH, Eichler EE, Phillippy AM, Koren S. 2020. HiCanu: accurate assembly of segmental duplications, satellites, and allelic variants from high-fidelity long reads. *Genome Res* **30**: 1291–1305. http://dx.doi.org/10.1101/gr.263566.120.

Nuttle X. 2016. Emergence of a Homo sapiens-specific gene family and chromosome 16p11. 2 CNV susceptibility. *Nature* **536**: 205–209.

Nuttle X, Giannuzzi G, Duyzend MH, Schraiber JG, Narvaiza I, Sudmant PH, Penn O, Chiatante G, Malig M, Huddleston J, et al. 2016. Emergence of a Homo sapiens-specific gene family and chromosome 16p11.2 CNV susceptibility. *Nature* **536**: 205–209. http://www.nature.com/doifinder/10.1038/nature19075.

Nuttle X, Huddleston J, O'Roak BJ, Antonacci F, Fichera M, Romano C, Shendure J, Eichler EE. 2013. Rapid and accurate large-scale genotyping of duplicated genes and discovery of interlocus gene conversions. *Nat Methods* **10**: 903–909. http://dx.doi.org/10.1038/nmeth.2572.

Ohno S. 1970. *Evolution by Gene Duplication*. Springer Science & Business Media https://play.google.com/store/books/details?id=5SjqCAAAQBAJ.

Ohno S, Wolf U, Atkin NB. 1968. Evolution from fish to mammals by gene duplication. *Hereditas* **59**: 169–187. http://dx.doi.org/10.1111/j.1601-5223.1968.tb02169.x.

PacBio. Quiver. https://github.com/PacificBiosciences/GenomicConsensus.

Parsons JD. 1995. Miropeats: graphical DNA sequence comparisons. *Comput Appl Biosci* 615–619.

Paten B, Diekhans M, Earl D, John JS, Ma J, Suh B, Haussler D. 2011a. Cactus graphs for genome comparisons. *J Comput Biol* **18**: 469–481. http://dx.doi.org/10.1089/cmb.2010.0252.

Paten B, Earl D, Nguyen N, Diekhans M, Zerbino D, Haussler D. 2011b. Cactus: Algorithms for genome multiple sequence alignment. *Genome Res* **21**: 1512–1528. http://dx.doi.org/10.1101/gr.123356.111.

Patterson M, Marschall T, Pisanti N, van Iersel L, Stougie L, Klau GW, Schönhuth A. 2015. WhatsHap: Weighted haplotype assembly for future-generation sequencing reads. *J Comput Biol* **22**: 498–509. http://online.liebertpub.com/doi/10.1089/cmb.2014.0157.

Paulding CA, Ruvolo M, Haber DA. 2003. The Tre2 (USP6) oncogene is a hominoid-specific gene. *Proc Natl Acad Sci U S A* **100**: 2507–2511. http://dx.doi.org/10.1073/pnas.0437015100.

Pertea G, Pertea M. 2020. GFF Utilities: GffRead and GffCompare. *F1000Res* **9**. http://dx.doi.org/10.12688/f1000research.23297.2.

Pevzner PA, Tang H, Tesler G. 2004. De novo repeat classification and fragment assembly. *Genome Res* **14**: 1786–1796. http://dx.doi.org/10.1101/gr.2395204.

Pevzner PA, Tang H, Waterman MS. 2001. An Eulerian path approach to DNA fragment assembly. *Proceedings of the National Academy of Sciences* **98**: 9748–9753. http://www.pnas.org/cgi/doi/10.1073/pnas.171285098.

Phillippy AM, Schatz MC, Pop M. 2008. Genome assembly forensics: finding the elusive mis-assembly. *Genome Biol* **9**: R55. http://dx.doi.org/10.1186/gb-2008-9-3-r55.

Pop M. 2004. Shotgun sequence assembly. In *Advances in Computers*, *Advances in computers*, pp. 193–248, Elsevier http://dx.doi.org/10.1016/s0065-2458(03)60006-9.

Porubsky D, Ebert P, Audano PA, Vollger MR, Harvey WT, Marijon P, Ebler J, Munson KM, Sorensen M, Sulovari A, et al. 2020. Fully phased human genome assembly without parental data using single-cell strand sequencing and long reads. *Nat Biotechnol*. http://dx.doi.org/10.1038/s41587-020-0719-5.

Pratt WS, Crawley S, Hicks J, Ho J, Nash M, Kim YS, Gum JR, Swallow DM. 2000. Multiple transcripts of MUC3: evidence for two genes, MUC3A and MUC3B. *Biochem Biophys Res Commun* **275**: 916–923. http://dx.doi.org/10.1006/bbrc.2000.3406.

Puljiz Z, Vikalo H. 2016. Decoding genetic variations: communications-inspired haplotype assembly. *IEEE/ACM Trans Comput Biol Bioinform* **13**: 518–530. http://dx.doi.org/10.1109/TCBB.2015.2462367.

Quick J, Loman NJ, Duraffour S, Simpson JT, Severi E, Cowley L, Bore JA, Koundouno R, Dudas G, Mikhail A, et al. 2016. Real-time, portable genome sequencing for Ebola surveillance. *Nature* **530**: 228–232. http://dx.doi.org/10.1038/nature16996.

Quinlan AR, Hall IM. 2010. BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**: 841–842. http://dx.doi.org/10.1093/bioinformatics/btq033.

Reich D, Green RE, Kircher M, Krause J, Patterson N, Durand EY, Viola B, Briggs AW, Stenzel U, Johnson PLF, et al. 2010. Genetic history of an archaic hominin group from Denisova Cave in Siberia. *Nature* **468**: 1053–1060. http://dx.doi.org/10.1038/nature09710.

Sahin FI, Yilmaz Z, Yuregir OO, Bulakbasi T, Ozer O, Zeyneloglu HB. 2008. Chromosome heteromorphisms: an impact on infertility. *J Assist Reprod Genet* **25**: 191–195. http://dx.doi.org/10.1007/s10815-008-9216-3.

Sahlin K, Tomaszkiewicz M, Makova KD, Medvedev P. 2018. Deciphering highly similar multigene family transcripts from Iso-Seq data with IsoCon. *Nat Commun* **9**: 4601. http://dx.doi.org/10.1038/s41467-018-06910-x.

Sakamoto N, Chastain PD, Parniewski P, Ohshima K, Pandolfo M, Griffith JD, Wells RD. 1999. Sticky DNA: self-association properties of long GAA.TTC repeats in R.R.Y triplex structures from Friedreich's ataxia. *Mol Cell* **3**: 465–475. http://dx.doi.org/10.1016/s1097-2765(00)80474-8.

Sanders AD, Falconer E, Hills M, Spierings DCJ, Lansdorp PM. 2017. Single-cell template strand sequencing by Strand-seq enables the characterization of individual homologs. *Nat Protoc* **12**: 1151–1176. http://dx.doi.org/10.1038/nprot.2017.029.

Sanders AD, Hills M, Porubský D, Guryev V, Falconer E, Lansdorp PM. 2016. Characterizing polymorphic inversions in human genomes by single-cell sequencing. *Genome Res* **26**: 1575–1587. http://dx.doi.org/10.1101/gr.201160.115.

Saveliev A, Everett C, Sharpe T, Webster Z, Festenstein R. 2003. DNA triplet repeats mediate heterochromatin-protein-1-sensitive variegated gene silencing. *Nature* **422**: 909–913. http://dx.doi.org/10.1038/nature01596.

Saxonov S, Berg P, Brutlag DL. 2006. A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. *Proc Natl Acad Sci U S A* **103**: 1412–1417. http://dx.doi.org/10.1073/pnas.0510310103.

Schmidt K, Noureen A, Kronenberg F, Utermann G. 2016. Structure, function, and genetics of lipoprotein (a). *J Lipid Res* **57**: 1339–1359. http://dx.doi.org/10.1194/jlr.R067314.

Schmutz J, Martin J, Terry A, Couronne O, Grimwood J, Lowry S, Gordon LA, Scott D, Xie G, Huang W, et al. 2004. The DNA sequence and comparative analysis of human chromosome 5. *Nature* **431**: 268–274. http://dx.doi.org/10.1038/nature02919.

Sedlazeck F, Kingham B, Prof John Todd, Centre Director & Dr David Buck, Head of High throughput Genomics, Schmutz J, Steven Wiley H. 2015. Home - PacBio - sequence with confidence. https://www.pacb.com/. (Accessed February 10, 2021).

Seo J-SS, Rhie A, Kim JJJJ, Lee S, Sohn M-HH, Kim CC-UU, Hastie A, Cao H, Yun J-YY, Kim JJJJ, et al. 2016. De novo assembly and phasing of a Korean human genome. *Nature* **538**: 243–247. http://dx.doi.org/10.1038/nature20098.

Sharp AJ, Hansen S, Selzer RR, Cheng Z, Regan R, Hurst JA, Stewart H, Price SM, Blair E, Hennekam RC, et al. 2006. Discovery of previously unidentified genomic disorders from the duplication architecture of the human genome. *Nat Genet* **38**: 1038–1042. http://dx.doi.org/10.1038/ng1862.

She X, Horvath JE, Jiang Z, Liu G, Furey TS, Christ L, Clark R, Graves T, Gulden CL, Alkan C, et al. 2004a. The structure and evolution of centromeric transition regions within the human genome. *Nature* **430**: 857–864. http://dx.doi.org/10.1038/nature02806.

She X, Jiang Z, Clark RA, Liu G, Cheng Z, Tuzun E, Church DM, Sutton G, Halpern AL, Eichler EE. 2004b. Shotgun sequence assembly and recent segmental duplications within the human genome. *Nature* **431**: 927–930. http://dx.doi.org/10.1038/nature03062.

She X, Liu G, Ventura M, Zhao S, Misceo D, Roberto R, Cardone MF, Rocchi M, NISC Comparative Sequencing Program, Green ED, et al. 2006. A preliminary comparative analysis of primate segmental duplications shows elevated substitution rates and a great-ape expansion of intrachromosomal duplications. *Genome Res* **16**: 576–583. http://dx.doi.org/10.1101/gr.4949406.

Shi L, Guo Y, Dong C, Huddleston J, Yang H, Han X, Fu A, Li Q, Li N, Gong S, et al. 2016. Long-read sequencing and de novo assembly of a Chinese genome. *Nat Commun* **7**: 12065. http://dx.doi.org/10.1038/ncomms12065.

Shumate A, Salzberg SL. 2020a. Liftoff: accurate mapping of gene annotations. *Bioinformatics*. http://dx.doi.org/10.1093/bioinformatics/btaa1016.

Shumate A, Salzberg SL. 2020b. Liftoff: an accurate gene annotation mapping tool. *Cold Spring Harbor Laboratory* 2020.06.24.169680. https://www.biorxiv.org/content/10.1101/2020.06.24.169680v1.abstract (Accessed November 12, 2020).

Simpson JT, Workman RE, Zuzarte PC, David M, Dursi LJ, Timp W. 2017. Detecting DNA cytosine methylation using nanopore sequencing. *Nat Methods* **14**: 407–410. http://dx.doi.org/10.1038/nmeth.4184.

Skaletsky H, Kuroda-Kawaguchi T, Minx PJ, Cordum HS, Hillier L, Brown LG, Repping S, Pyntikova T, Ali J, Bieri T, et al. 2003. The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. *Nature* **423**: 825–837. http://dx.doi.org/10.1038/nature01722.

Skoda RC, Gonzalez FJ, Demierre A, Meyer UA. 1988. Two mutant alleles of the human cytochrome P-450db1 gene (P450C2D1) associated with genetically deficient metabolism of debrisoquine and other drugs. *Proc Natl Acad Sci U S A* **85**: 5240–5243. http://dx.doi.org/10.1073/pnas.85.14.5240.

Smit AFA, Hubley R, Green P. 1996. RepeatMasker.

Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**: 1312–1313. http://dx.doi.org/10.1093/bioinformatics/btu033.

Standing Committee on Human Cytogenetic Nomenclature. 1995. *ISCN 1995: An International System for Human Cytogenetic Nomenclature (1995) : Recommendations of the*

International Standing Committee on Human Cytogenetic Nomenclature, Memphis, Tennessee, USA, October 9-13, 1994. Karger Medical and Scientific Publishers https://play.google.com/store/books/details?id=7Lc10M3qJqEC.

Stankiewicz P, Lupski JR. 2002a. Genome architecture, rearrangements and genomic disorders. *Trends Genet* **18**: 74–82. http://dx.doi.org/10.1016/S0168-9525(02)02592-1.

Stankiewicz P, Lupski JR. 2002b. Genome architecture, rearrangements and genomic disorders. *Trends Genet* **18**: 74–82. http://dx.doi.org/10.1016/s0168-9525(02)02592-1.

Steinberg KM. 2016. High-quality assembly of an individual of yoruban descent. *bioRxiv* **067447**. http://dx.doi.org/10.

Steinberg KM. 2012. Structural diversity and African origin of the 17q21. 31 inversion polymorphism. *Nat Genet* **44**: 872–880.

Steinberg KM, Graves-Lindsay T, Schneider VA, Chaisson MJP, Tomlinson C, Huddleston JL, Minx P, Kremitzki M, Albrecht D, Magrini V, et al. 2016. High-quality assembly of an individual of yoruban descent. *bioRxiv* 067447. http://biorxiv.org/lookup/doi/10.1101/067447.

Steinberg KM, Schneider VA, Graves-lindsay TA, Fulton RS, Agarwala R, Huddleston J, Shiryev SA, Morgulis A, Surti U, Warren WC, et al. 2014. Single haplotype assembly of the human genome from a hydatidiform mole. *Genome Res* 1–12. http://dx.doi.org/10.1101/gr.180893.114.24.

Steiper ME, Young NM, Sukarna TY. 2004. Genomic data support the hominoid slowdown and an Early Oligocene estimate for the hominoid-cercopithecoid divergence. *Proc Natl Acad Sci U S A* **101**: 17021–17026. http://dx.doi.org/10.1073/pnas.0407270101.

Stergachis AB, Debo BM, Haugen E, Churchman LS, Stamatoyannopoulos JA. 2020. Single-molecule regulatory architectures captured by chromatin fiber sequencing. *Science* **368**: 1449–1454. http://dx.doi.org/10.1126/science.aaz1646.

Stoltzfus A. 1999. On the possibility of constructive neutral evolution. *J Mol Evol* **49**: 169–181. http://dx.doi.org/10.1007/pl00006540.

Sudmant PH, Huddleston J, Catacchio CR, Malig M, Hillier LW, Baker C, Mohajeri K, Kondova I, Bontrop RE, Persengiev S, et al. 2013. Evolution and diversity of copy number variation in the great ape lineage. *Genome Res* **23**: 1373–1382. http://dx.doi.org/10.1101/gr.158543.113.

Sudmant PH, Kitzman JO, Antonacci F, Alkan C, Malig M, Tsalenko A, Sampas N, Bruhn L, Shendure J. 2010. Diversity of human copy number. *Science* **11184**: 2–7. http://dx.doi.org/papers2://publication/uuid/C37D7A2A-43D1-4164-A5DE-4447C58EBC0D.

Sudmant PH, Mallick S, Nelson BJ, Hormozdiari F, Krumm N, Huddleston J, Coe BP, Baker C, Nordenfelt S, Bamshad M, et al. 2015a. Global diversity, population stratification, and selection of human copy-number variation. *Science* **349**. http://dx.doi.org/10.1126/science.aab3761.

Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, Huddleston J, Zhang Y, Ye K, Jun G, Fritz MH-Y, et al. 2015b. An integrated map of structural variation in 2,504 human genomes. *Nature* **526**: 75–81. http://dx.doi.org/10.1038/nature15394.

Sutcliffe JS, Nelson DL, Zhang F, Pieretti M, Caskey CT, Saxe D, Warren ST. 1992. DNA methylation represses FMR-1 transcription in fragile X syndrome. *Hum Mol Genet* **1**: 397–400. http://dx.doi.org/10.1093/hmg/1.6.397.

Suzuki IK, Gacquer D, Van Heurck R, Kumar D, Wojno M, Bilheu A, Herpoel A, Lambert N, Cheron J, Polleux F, et al. 2018. Human-Specific NOTCH2NL Genes Expand Cortical Neurogenesis through Delta/Notch Regulation. *Cell* **173**: 1370-1384.e16. http://dx.doi.org/10.1016/j.cell.2018.03.067.

T., M B. 2010. Long intronic GAA repeats causing Friedreich ataxia impede transcription elongation.

Tian Y, Wang J-L, Huang W, Zeng S, Jiao B, Liu Z, Chen Z, Li Y, Wang Y, Min H-X, et al. 2019. Expansion of human-specific GGC repeat in neuronal intranuclear inclusion disease-related disorders. *Am J Hum Genet* **105**: 166–176. http://dx.doi.org/10.1016/j.ajhg.2019.05.013.

Tuzun E, Sharp AJ, Bailey JA, Kaul R, Morrison VA, Pertz LM, Haugen E, Hayden H, Albertson D, Pinkel D, et al. 2005. Fine-scale structural variation of the human genome. *Nat Genet* **37**: 727–732. http://dx.doi.org/10.1038/ng1562.

Vaser R, Sović I, Nagarajan N, Šikić M. 2017. Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Res* **27**: 737–746. http://dx.doi.org/10.1101/gr.214270.116.

Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, et al. 2001. The sequence of the human genome. *Science* **291**: 1304–1351. http://dx.doi.org/10.1126/science.1058040.

Vollger MR, Dishuck PC, Sorensen M, Welch AE, Dang V, Dougherty ML, Graves-Lindsay TA, Wilson RK, Chaisson MJP, Eichler EE. 2019. Long-read sequence and assembly of segmental duplications. *Nat Methods* **16**: 88–94. http://dx.doi.org/10.1038/s41592-018-0236-3.

Wainszelbaum MJ, Charron AJ, Kong C, Kirkpatrick DS, Srikanth P, Barbieri MA, Gygi SP, Stahl PD. 2008. The hominoid-specific oncogene TBC1D3 activates Ras and modulates epidermal growth factor receptor signaling and trafficking. *J Biol Chem* **283**: 13233–13242. http://dx.doi.org/10.1074/jbc.M800234200.

Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, Cuomo CA, Zeng Q, Wortman J, Young SK, et al. 2014. Pilon: An integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One* **9**. http://dx.doi.org/10.1371/journal.pone.0112963.

Warren WC, Alan Harris R, Haukness M, Fiddes IT, Murali SC, Fernandes J, Dishuck PC, Storer JM, Raveendran M, Hillier LW, et al. 2020a. Sequence diversity analyses of an improved rhesus macaque genome enhance its biomedical utility. *Science* **370**. https://science.sciencemag.org/content/370/6523/eabc6617 (Accessed February 24, 2021).

Warren WC, Harris RA, Haukness M, Fiddes IT, Murali SC, Fernandes J, Dishuck PC, Storer JM, Raveendran M, Hillier LW, et al. 2020b. Sequence diversity analyses of an improved rhesus macaque genome enhance its biomedical utility. *Science* **370**. http://dx.doi.org/10.1126/science.abc6617.

Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, Agarwal P, Agarwala R, Ainscough R, Alexandersson M, An P, et al. 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**: 520–562. ://WOS:000179611600053\nhttp://www.nature.com/nature/journal/v420/n6915/pdf/nature01262.pdf">http://www.nature.com/doifinder/10.1038/nature01262\n<Go to ISI>://WOS:000179611600053\nhttp://www.nature.com/nature/journal/v420/n6915/pdf/nature01262.pdf.

Watson CT, Steinberg KM, Huddleston J, Warren RL, Malig M, Schein J, Willsey AJ, Joy JB, Scott JK, Graves TA, et al. 2013. Complete haplotype sequence of the human immunoglobulin heavy-chain variable, diversity, and joining genes and characterization of allelic and copy-number variation. *Am J Hum Genet* **92**: 530–546. http://dx.doi.org/10.1016/j.ajhg.2013.03.004.

Watson M, Warr A. 2019. Errors in long-read assemblies can critically affect protein prediction. *Nat Biotechnol* **37**: 124–126. http://dx.doi.org/10.1038/s41587-018-0004-z.

Weiss LA, Shen Y, Korn JM, Arking DE, Miller DT, Fossdal R, Saemundsen E, Stefansson H, Ferreira MAR, Green T, et al. 2008. Association between microdeletion and microduplication at 16p11.2 and autism. *N Engl J Med* **358**: 667–675. http://dx.doi.org/10.1056/NEJMoa075974.

Wellcome Trust Case Control Consortium, Craddock N, Hurles ME, Cardin N, Pearson RD, Plagnol V, Robson S, Vukcevic D, Barnes C, Conrad DF, et al. 2010. Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls. *Nature* **464**: 713–720. http://dx.doi.org/10.1038/nature08979.

Wenger AM, Peluso P, Rowell WJ, Chang P-C, Hall RJ, Concepcion GT, Ebler J, Fungtammasan A, Kolesnikov A, Olson ND, et al. 2019. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nat Biotechnol*. http://dx.doi.org/10.1038/s41587-019-0217-9.

Wilch ES, Morton CC. 2018. Historical and Clinical Perspectives on Chromosomal Translocations. *Adv Exp Med Biol* **1044**: 1–14. http://dx.doi.org/10.1007/978-981-13-0593-1_1.

Willcocks LC, Carr EJ, Niederer HA, Rayner TF, Williams TN, Yang W, Scott JAG, Urban BC, Peshu N, Vyse TJ, et al. 2010. A defunctioning polymorphism in FCGR2B is associated with protection against malaria but susceptibility to systemic lupus erythematosus. *Proc Natl Acad Sci U S A* **107**: 7881–7885. https://www.pnas.org/content/107/17/7881 (Accessed February 25, 2021).

Wood B. 1992. Origin and evolution of the genus Homo. *Nature* **355**: 783–790. http://dx.doi.org/10.1038/355783a0.

Wu CI, Li WH. 1985. Evidence for higher rates of nucleotide substitution in rodents than in man. *Proc Natl Acad Sci U S A* **82**: 1741–1745. http://dx.doi.org/10.1073/pnas.82.6.1741.

Yunis JJ, Prakash O. 1982. The origin of man: a chromosomal pictorial legacy. *Science* **215**: 1525–1530. http://dx.doi.org/10.1126/science.7063861.

Zheng Y-F, Chen Z-C, Shi Z-X, Hu K-H, Zhong J-Y, Wang C-X, Shi W, Chen Y, Xie S-Q, Luo F, et al. 2020. HIT-scISOseq: High-throughput and High-accuracy Single-cell Full-length Isoform Sequencing for Corneal Epithelium. *Cold Spring Harbor Laboratory* 2020.07.27.222349. https://www.biorxiv.org/content/10.1101/2020.07.27.222349v1 (Accessed February 23, 2021).

2018. Oxford Nanopore Technologies. https://nanoporetech.com (Accessed March 2018).

Appendix A. Supplement for chapter 2

A.1 Percentage of resolved SDs across genomes/assemblers/technologies

Figure S1 and Table S1 show the fraction of "Resolved" segmental duplications (SDs). Our working definition of resolved is that for an SD to be resolved the assembly must continue into unique sequence on either side of the SD by at least some minimal extension. Figure S1 shows the fraction of resolved bases as the minimal extension is varied from 0 to 250 kbp. The basic steps of identifying resolved versus unresolved duplications are as follows:

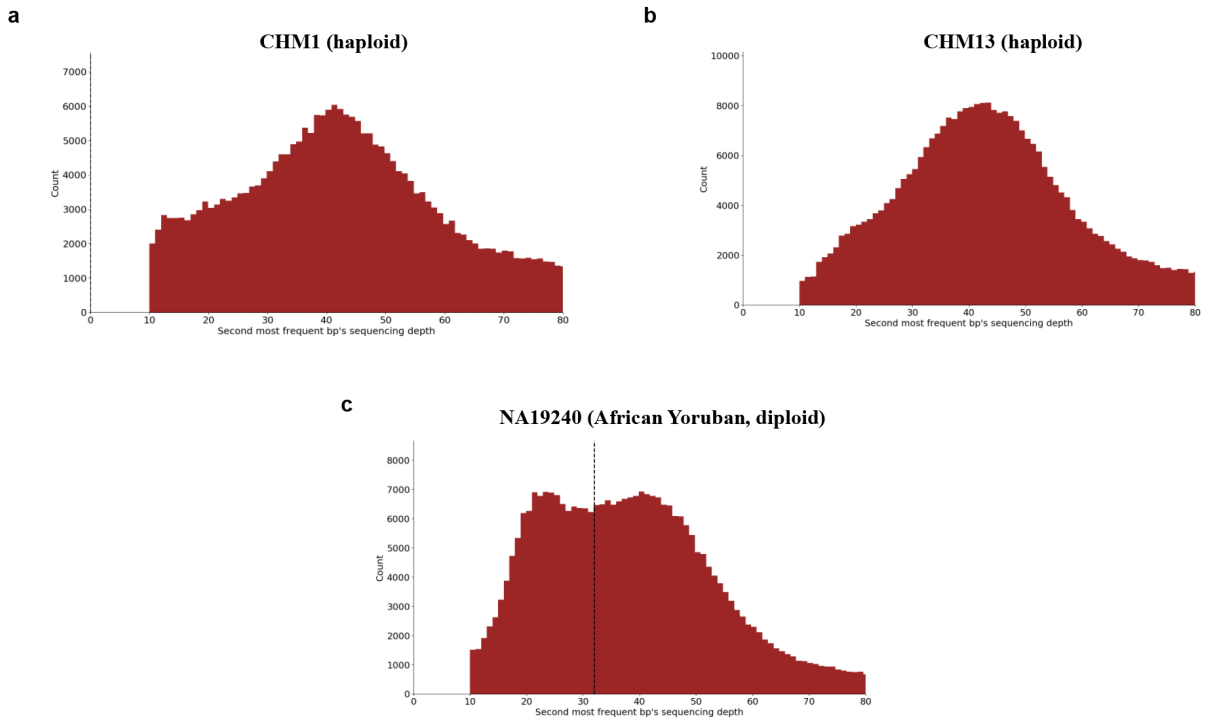
1) Map the *de novo* assembly to the human reference using MashMap 2.0 defaults.
2) Download the UCSC-annotated SD track and merge overlapping SDs by their maximum percent identity.
3) Intersect the *de novo* assembly track with the modified SD track.
4) Determine if and by how much the *de novo* assembly extends past SD blocks on either side.
5) Mark SDs as resolved or unresolved based on whether the *de novo* assembly extends at least X kbp into unique sequence on either side.
6) Plot the percentage of SD bases resolved as a function of the minimal extension into unique sequence past a duplication block.

Currently, there are two Oxford Nanopore Technologies (ONT) ultra-long assemblies of NA12878: one that is recently published (Jain et al. 2018b), and the other an updated assembly from the Phillippy lab. The ONT assemblies do outperform the PacBio assemblies; however, its improvement over the different PacBio assemblies is less than 10%. This still leaves the majority of SDs unresolved, motivating and highlighting the importance of our method. All the input

assemblies for this analysis were “contig” assemblies and not “scaffolded” assemblies. While there exist scaffolded assemblies for some of these genomes, we decided not to use them in order to make comparisons more consistent.

A.2 Application of SDA to NA19240 (diploid)

Using the two haploid (CHM1 and CHM13) and one diploid (NA19240) human genomes, we effectively modeled read depth corresponding to the second most common base pair (i.e., SNV or PSV). Because most paralogous variation is evolutionarily older than allelic variation, it is much more likely to be fixed and, as a result, true PSVs show a different sequence read depth than allelic variation (i.e., CHM1 shows a mode at ~40-fold read-depth, consistent with a fixed duplicate copy. In contrast, a diploid sample that harbors both allelic and paralogous variants shows a clear bimodal distribution. Thus, to avoid phasing allelic variation we set a minimum depth threshold at the mean coverage minus three standard deviations or half the mean coverage, whichever was greater. This is represented by the black dotted line and corresponds to the trough between the two peaks (~31-fold). This threshold enriches for true PSVs and prevents most alternate haplotypes from being assembled.



PSV read-depth distribution. Sequencing read-depth distribution shown for the second most common SNV across all collapsed regions of SDs in a) CHM1, b) CHM13, and c) NA19240 genome assemblies. a/b) For CHM1 and CHM13, we consider the distribution with a mode at a read depth of 42-fold to represent putative PSVs. There is a clear peak in SNV frequency around a sequencing read depth of ~42-fold (see Methods). c) In the case of NA19240, we observe a bimodal distribution and consider variants with a read depth ~45-fold to once again represent PSVs, while the second mode at read depth of ~23-fold represents possible allelic SNVs. Therefore, we also set a minimum PSV sequencing depth of 32X (black dashed line) for diploid genomes. SNVs with a read depth less than 10-fold sequence coverage are not displayed because they likely represent sequencing error and exist at a much higher frequency.

Please note that the recovered SDs would not be the same for an *in silico* diploid of CHM1 and CHM13 because only PSVs common to both CHM1 and CHM13 would be used for phasing. Thus, the resolution would be of paralogs and not alleles. Haplotype phasing of duplication regions remains an unaddressed layer of complexity and an area of future investigation. For the diploid genome NA19240, we focused our analysis on the discovery of PSVs occurring at the expected

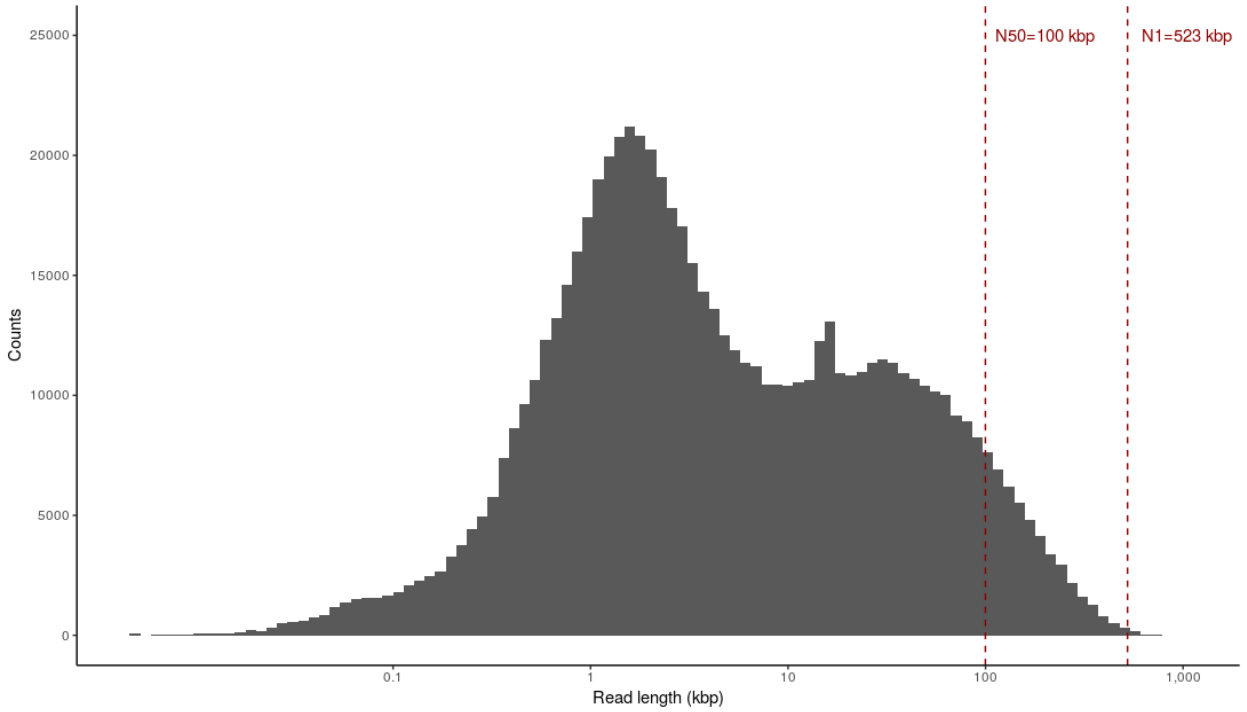
frequency of a duplicated copy and specifically excluded allelic variation by requiring sequence coverage consistent with a unique diploid region of the genome. However, given that paralogous variation can approximate or even become more identical than allelic variation[2], it is likely that SDA could be extended to distinguish and assemble haplotypes as well as paralogs. For example, many haplotypes of *HLA* share 90%-99% sequence identity[3], but *NOTCH2NL*, which we resolved using SDA, shares up to 99.7% sequence identity among the copies. It may be possible to integrate our SDA method with haplotype-aware assemblers such as FALCON-Unzip[4], which currently fail to resolve highly identical duplications within human genomes.

A.3 Application to ONT data

SDA is compatible with ONT data and we performed an analysis of collapsed SDs present in the ONT assembly of NA12878[1]. We identified 365 collapses, a similar number to that identified in the CHM1 PacBio assembly analyzed (283). We present the results compared to PacBio data for NA19240 (Figure S7). Overall, the accuracy of the ONT contigs is much lower. There are far more "failed" assemblies because of the lower sequencing coverage. PSVs are more difficult to identify since ONT has more mismatch errors than PacBio. While ONT data offers longer reads, the fundamental problem is its lower accuracy. The total assembly accuracy of the NA19240 assembly (assembled with PacBio) was 99.28%[5] before Illumina short-read polishing, whereas the assembly accuracy of NA12878 was only 95.20%[1].

We also note that the generation of ultra-long reads of >1 Mbp is not yet common. The longest read reported in Jain et al. 2018 was 882 kbp and reads >500 kbp represent ~1% of the data. Therefore, there is only ~0.05X coverage of 500 kbp reads and ~2.5X coverage of 100 kbp reads, which is not sufficient for proper read correction and assembly of SDs. Finally, the

generation of 1 Mbp length molecules is non-trivial and will be limited to a small fraction of samples where high-quality DNA can be prepared.



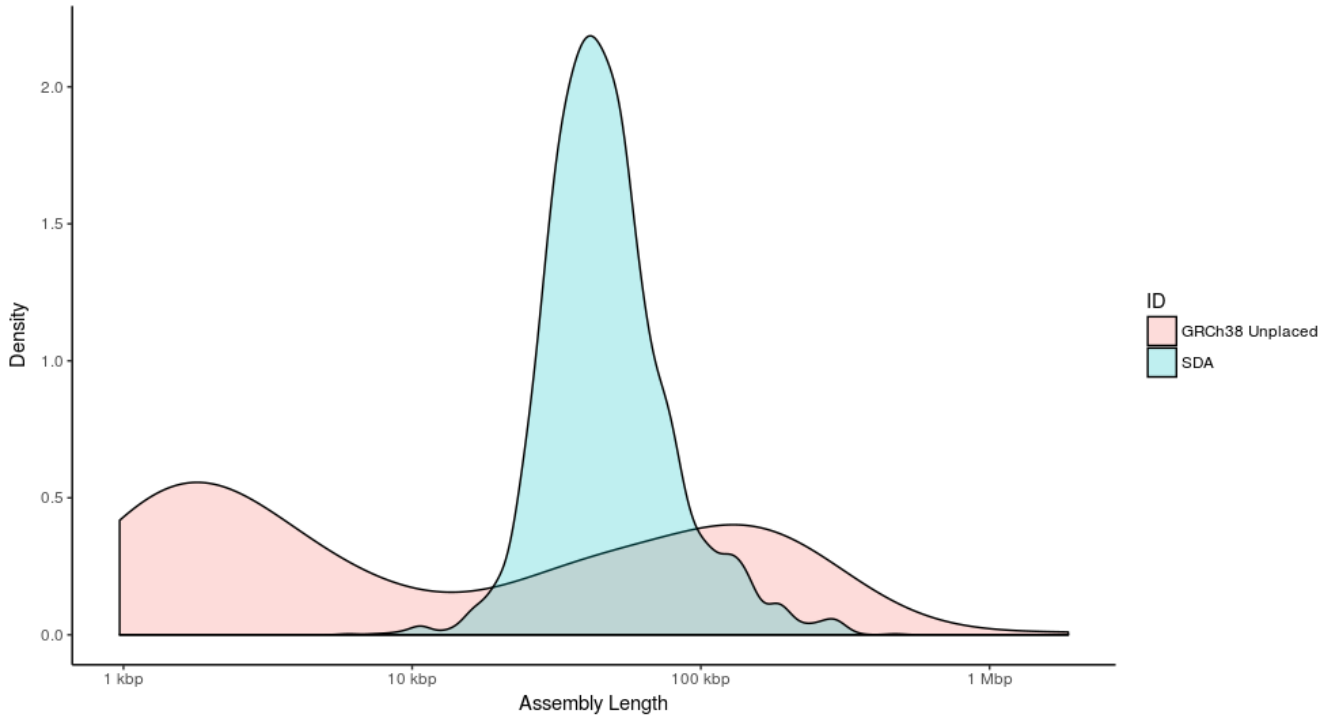
Distribution of ONT ultra-long reads from NA12878. Data available at https://github.com/nanopore-wgs-consortium/NA12878/blob/master/Genome.md. Only reads prepared using the ultra-long protocol are shown.

A.4 Integration of SDA contigs into the *de novo* assembly

The majority of our sequence contigs begin and end within SDs and do not transition into unique regions. This is due to the interspersed architecture SDs, which are frequently organized into very large blocks (>500 kbp in size) where structural variation and interlocus gene conversion occur. The latter creates pockets (often >50 kbp) with limited or no sequence divergence. As a result, the resolved sequences effectively represent islands of duplication with no transition into unique sequence. For example, of the 590 assembled sequences from our CHM1 assembly, we found that only 131 (22.2%) can be anchored to a unique sequence (Figure S11). Of these 131, only 28 overlap with unique sequence for at least 10 kbp. These contigs can be used to extend the

original FALCON assembly confidently. In total, there is 583 kbp of sequence from SDA contigs that overlap with unique sequence in the genome.

We note that even though our "orphan" assemblies are small, they are comparable in length to the unplaced contigs in GRCh38. More importantly, they are high quality and contiguous, making them useful for downstream genomic analyses. This is in sharp contrast to the small contigs typically generated by WGSA, which represent collapsed and fragmented mistakes of the assembly process of little biological utility. In the past, studies of duplication typically relied on generating similar high-quality sequence from BAC and fosmid clone inserts—a lengthy and costly prospect[6–9]. Here, we have generated the equivalent of 500-1,500 high-quality contigs of similar size per genome that otherwise would have been lost. The average size of these high-quality contigs (54 kbp) is sufficient for improved gene annotation.



Length distribution of orphaned contigs. Density plot of the assembly lengths of unplaced contigs in GRCh38 versus the contigs produced by SDA across CHM1, CHM13, and NA19240. The mean and median lengths for GRCh38 unplaced contigs are 67.8 kbp and

6.5 kbp, respectively, and the mean and median lengths for the SDA contigs are 54.3 kbp and 44.9 kbp, respectively.

In the event that others would like to use SDA within their whole-genome assembly, SDA creates a partitioned list of reads with assignments to SDA contigs that can be processed by other assemblers post hoc. Specifically, one would first run the assembler to produce contigs and resolve duplications with SDA. All reads processed by SDA can be processed as a data structure of tuples (read, duplication index). Assembly would be executed again. Given two reads that overlap, one could check if they are assigned a duplication index, and if so, whether they have the same duplication index.

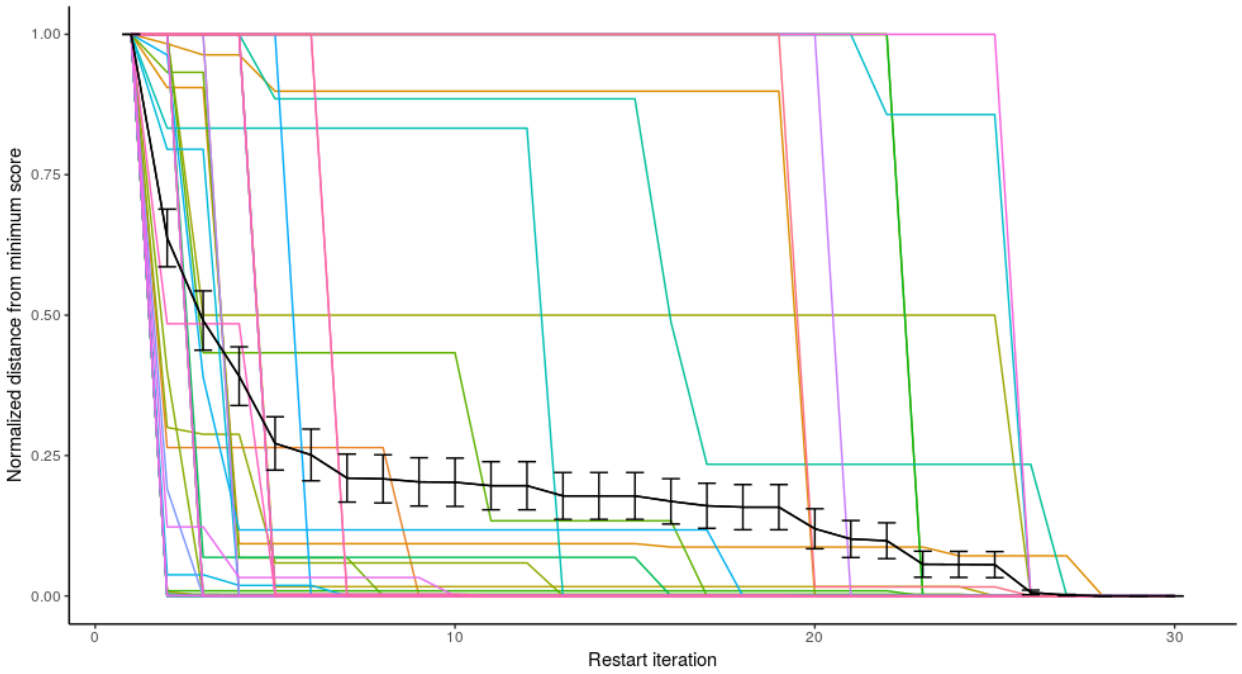
A.5 Improvements in SDA

The underlying PSV correlation clustering (CC) algorithm was presented as part of a RECOMB submission (Chaisson, 2017). This paper showed proof-of-principle of the algorithm and was based only on simulated data. Here, we develop the SDA method, including the computational infrastructure, apply it to real long-read WGS data, and perform a detailed analysis of the results. This required several developments and improvements. Specific improvements to the RECOMB algorithm include:

1. An optimization of the random sampling procedure to select the best sampled partition among many runs. The CC score was reduced per random sampling iteration for PSV graphs where multiple iterations had an impact.
2. The graph used by CC was modified to account for sparse stretches of PSVs that are more commonly seen in real data than the simulated data in the RECOMB paper. Previously, repulsion edges were made when two PSVs had overlapping reads, but there was no positive edge. However, this was problematic because PSVs that just missed the threshold to have an attraction edge would automatically become a repulsion edge even

though there was moderate evidence for an attraction edge. Here, we modified the definition of a repulsion edge to be two PSVs without any significant evidence for an attraction edge. This improved performance (fewer edges) and the results (less incorrect fracturing of paralogs).

3. During the process of cluster formation, all existing pairwise clusters are intermittently assessed to determine if the merging of any pair of them would improve the overall CC score. This development improved the performance of paralog separation, particularly for very long collapsed duplications.
4. In addition to these algorithmic developments, we modified SDA so that it can work with either ONT or PacBio long-read data as input and provided options such that different assemblers (e.g., Canu, miniasm and wtdbg) can be applied to resolve paralogs based on the partitioned reads. For a comparison of assembler performance, see Table S9.



Random restarts improve CC. This figure shows the normalized decrease in the CC score over n=76 different collapses in CHM13. Each line shows the minimum CC score observed at a given restart iteration. The black line shows the mean CC score with standard error bars.

A.6 Assembling collapsed regions using Canu

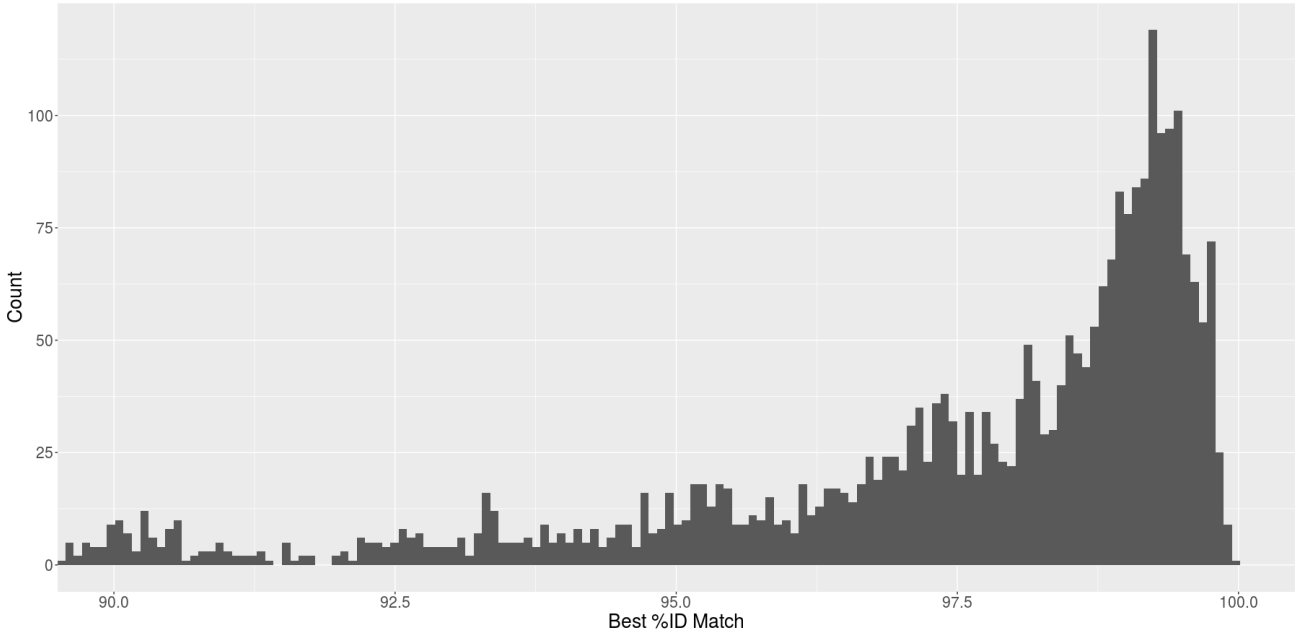
We assessed the effects of various parameter adjustments within Canu to see if SDs could be resolved without SDA. We specifically selected 10 regions of collapse and tried many parameter combinations using Canu and compared the results to our SDA. Results are summarized in Table S8. Assembling the individual collapses produced multiple paralogs in most cases; however, in all but one case, SDA was able to resolve more of the paralogs than any of the Canu assemblies, regardless of parameters. We found these two parameters essential to having any success in creating paralogs: corOutCoverage=300 and corMhapSensitivity=high. Setting corOutCoverage much higher than for the whole genomic coverage forces all reads to be corrected; similarly setting corMhapSensitivity to high ensures that the best overlaps are found. Both of these parameters are computationally impractical for whole-genome assemblies. Finally, we varied the corMaxEvidenceErate=[0.15, 0.25, 0.35, 0.45, 0.55] parameter to generate the ranges shown in Table S8. The corMaxEvidenceErate controls the maximum amount of error that can exist between two reads for them to be overlapped in the read correction step. Increasing this value generally increased the amount of assembled sequence but decreased the quality of the assembly.

A.7 Sequence divergence required for SDA

SDA is able to resolve large human-specific duplication events with less than 0.5% sequence divergence, see *NOTCH2NL* and S*RGAP2* (Table S3, Figures 3 and S4). However, there were events, such as the duplication surrounding *BOLA2*, with stretches of 50 kbp of identical sequence we were unable to resolve. Based on our results, we would argue that reads with 10-15%

error are sufficient to resolve duplications that are less than 0.5% diverged, as long as the reads have random errors and there is sufficient coverage (>60X).

To further examine the required sequence divergence between duplications to resolve them with SDA, we aligned and determined the percent identity between all the SDA contigs that we generated for CHM1. At 99.5% sequence identity, the distribution drops off precipitously suggesting a limit. Additionally, almost no sequences are 99.9% identical indicating that 0.1% probably reflects an upper bound of what SDA is able to resolve even in ideal cases.

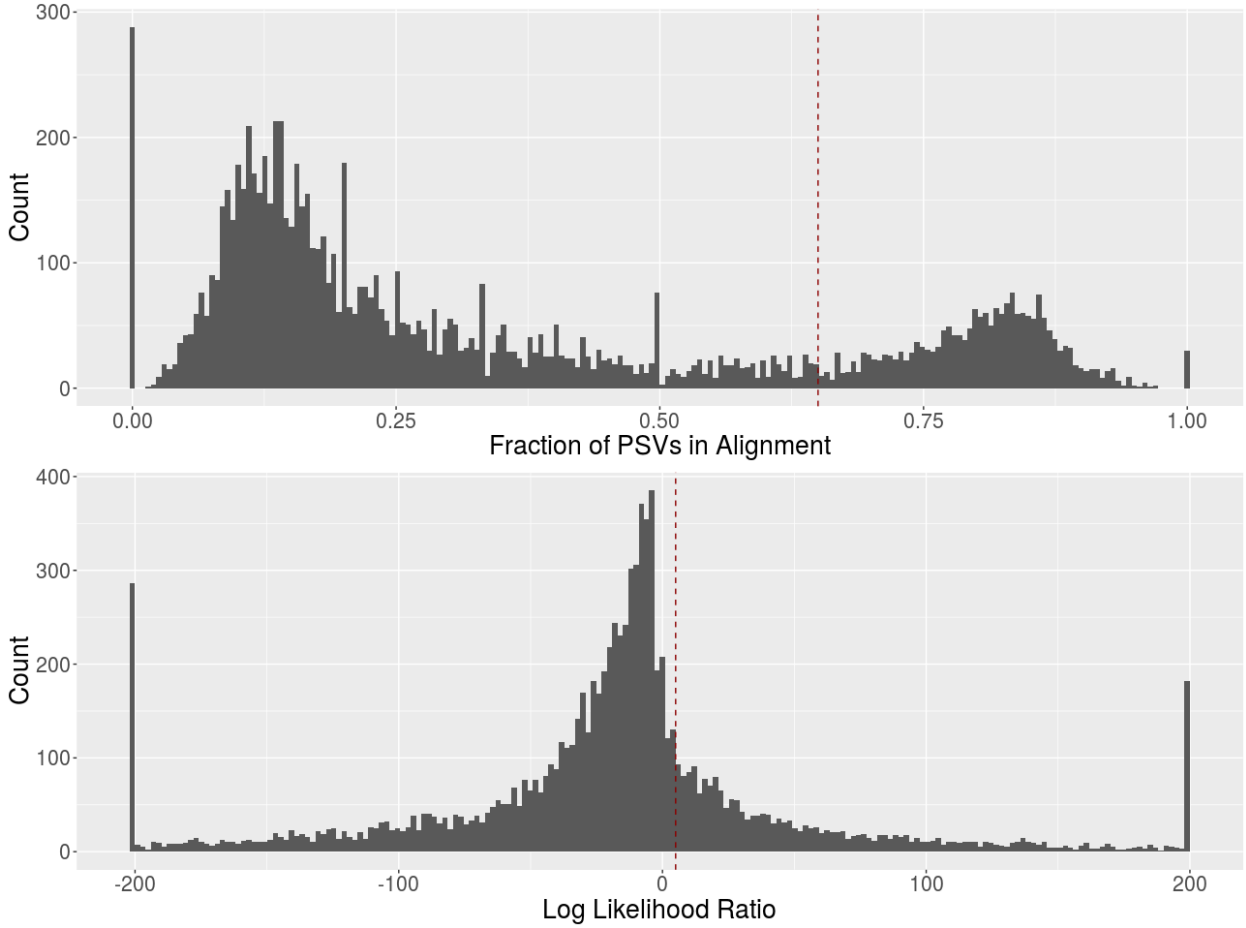


Percent identity between SDA contigs from CHM1. This figure shows the highest percent identity alignment between all pairs of SDA contigs from CHM1.

A.8 Ultra-long ONT as orthogonal support for SDA contigs

We investigated the ability of ultra-long ONT reads to provide orthogonal support for the existence of our SDA contigs. To overcome the high error rate of the individual ONT reads, we focused on identifying matches between the PSVs identified in our contigs and the ONT reads.

We required that at least 65% of the PSVs expected to be present in the overlap and that the log likelihood ratio between probability that the observed PSVs were real versus sequencing error to be greater than five**.** When we did this, we identified 1,932 ONT alignments between the ultra-long reads and our SDA contigs. On average, an ONT read maps 1.19 times to 1,184 SDA contigs providing orthogonal support for 641 (54%) of our SDA contigs.



PSV thresholds for determining correct ONT alignments. Each histogram shows the distribution of alignments of ONT reads to SDA contigs; the lines in red mark the thresholds used for filtering valid alignments. The first plot shows the distribution of the fraction of the expected number of PSVs in the alignment. The second plot shows the distribution of the log likelihood ratio between the probability that the observed PSVs are real or sequencing error.

A.9 Additional information on SDA results

Information about length, PSVs, and mapping location in GRCh38 can be found for all the SDA contigs generated in Table S10. When the collapsed sequences in CHM13 (24.3 Mbp) and NA19240 (22.6 Mbp) are mapped back to the reference, they represent 86.6 and 82.4 Mbp of sequence, respectively. Additionally, 73.1 (84.4%) and 64.4 (78.2%) Mbp of the mapped collapsed sequence overlaps with unresolved SDs. Approximately 52% (755/1,440) of CHM13 and 55% (973/1,772) of the African genome assemblies were diverged (<99.8% sequence identity) when compared to the reference genome. All of this is consistent with our results in CHM1 (Figure 2.2, S5, and S6).

A.10 BAC analysis with CHM1

In the main text we assert that we expect 37.4% of our BAC clones to validate. This is based on the alignment of 1,253 CHM1 BAC clones back to the reference genome where we found that they represent 65.7 Mbp, or 37.4% of the 175.5 Mbp of SDs annotated in GRCh38. Accession numbers for all BACs used to validate CHM1 SDA contigs can be found in Table S4.

A.11 Supplemental Figures

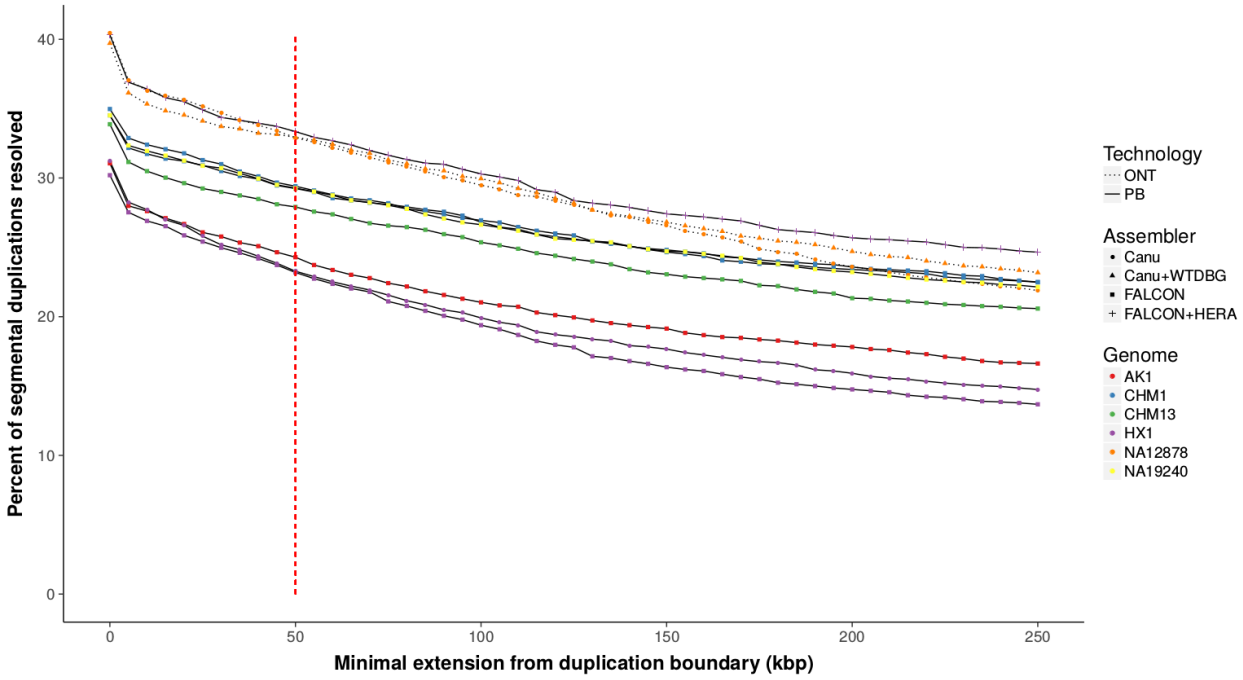


Figure S1. Proportion of resolved SDs in different PacBio (PB)/ONT genome assemblies. The figure shows the percent of SD bases that are resolved in human genome assemblies plotted as a function of the length of minimum extension of the alignment past the duplication. The number of resolved SD base pairs is relatively constant irrespective of the requirement of flanking unique base pairs. The dashed red line indicates the threshold chosen for our analysis used to generate the first panel in Figure S2 and the fraction of resolved SDs in Table S1.

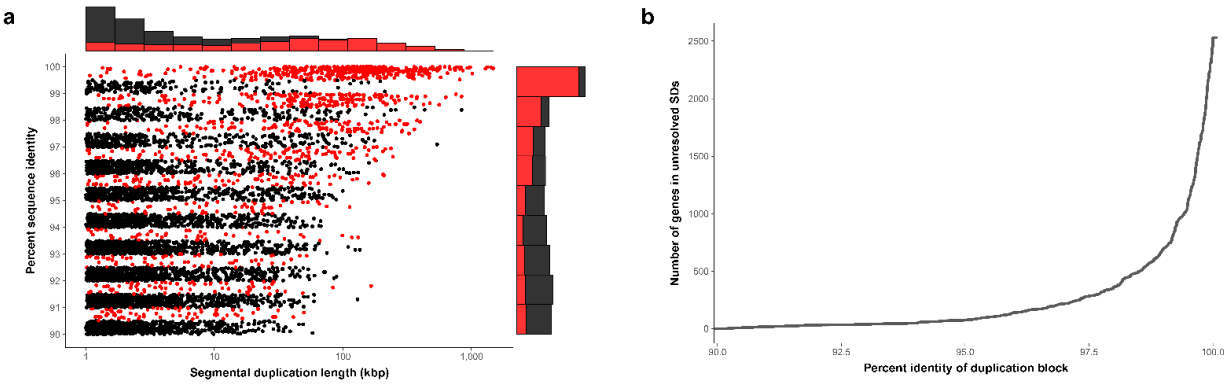


Figure S2. Resolution of SDs in SMRT genome assemblies. a) SDs (as a function of percent identity and length) in GRCh38 are marked as resolved (black) if present in the CHM1 assembly, or unresolved (red) if it only appears in the reference. The stacked marginal histograms show the relative number of resolved and unresolved SDs within each bin. Resolved duplications are defined as those mapping with high sequence identity, being completely contained, and extending at least 50 kbp into unique sequence on either side of the duplication block (Methods). See Figure S1 and Table S1 for the fraction of unresolved duplications across different genomes, assemblers, and technologies. Note that resolved and unresolved SDs are offset from one another along the y-axis to avoid overlapping. b) This plot shows the number of genes that exist within unresolved SDs blocks in the CHM1 assembly versus the maximum percent identity SD within that block.

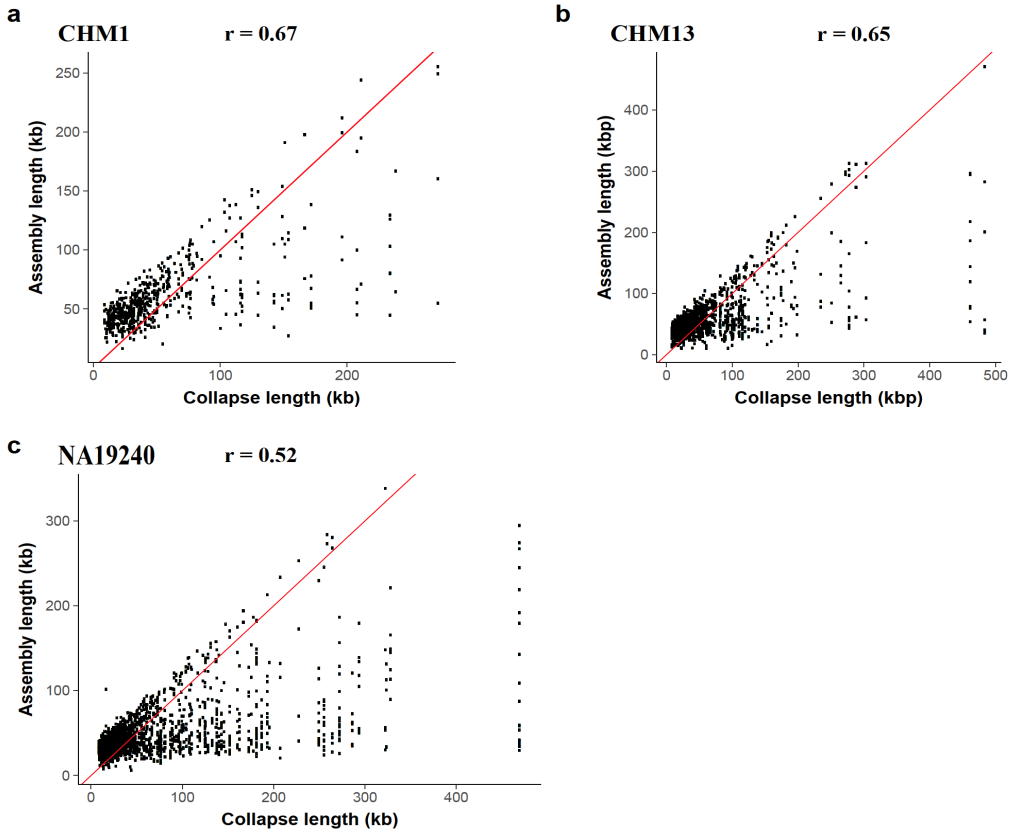


Figure S3. Length of collapsed SDs and SDA assemblies. Correlation of collapse length and SDA assembly length in a) CHM1 (n=590), b) CHM13 (n=1440), and c) NA19240 (n=1772) genome assemblies. In all three assemblies there is a strong correlation (Pearson's correlation) between the length of a collapsed SD and the length of the resulting SDA assembly. SDA is not restricted to assembling duplications less than the maximum read length (like other assemblers), but rather it is restricted by the size of the collapsed duplication.

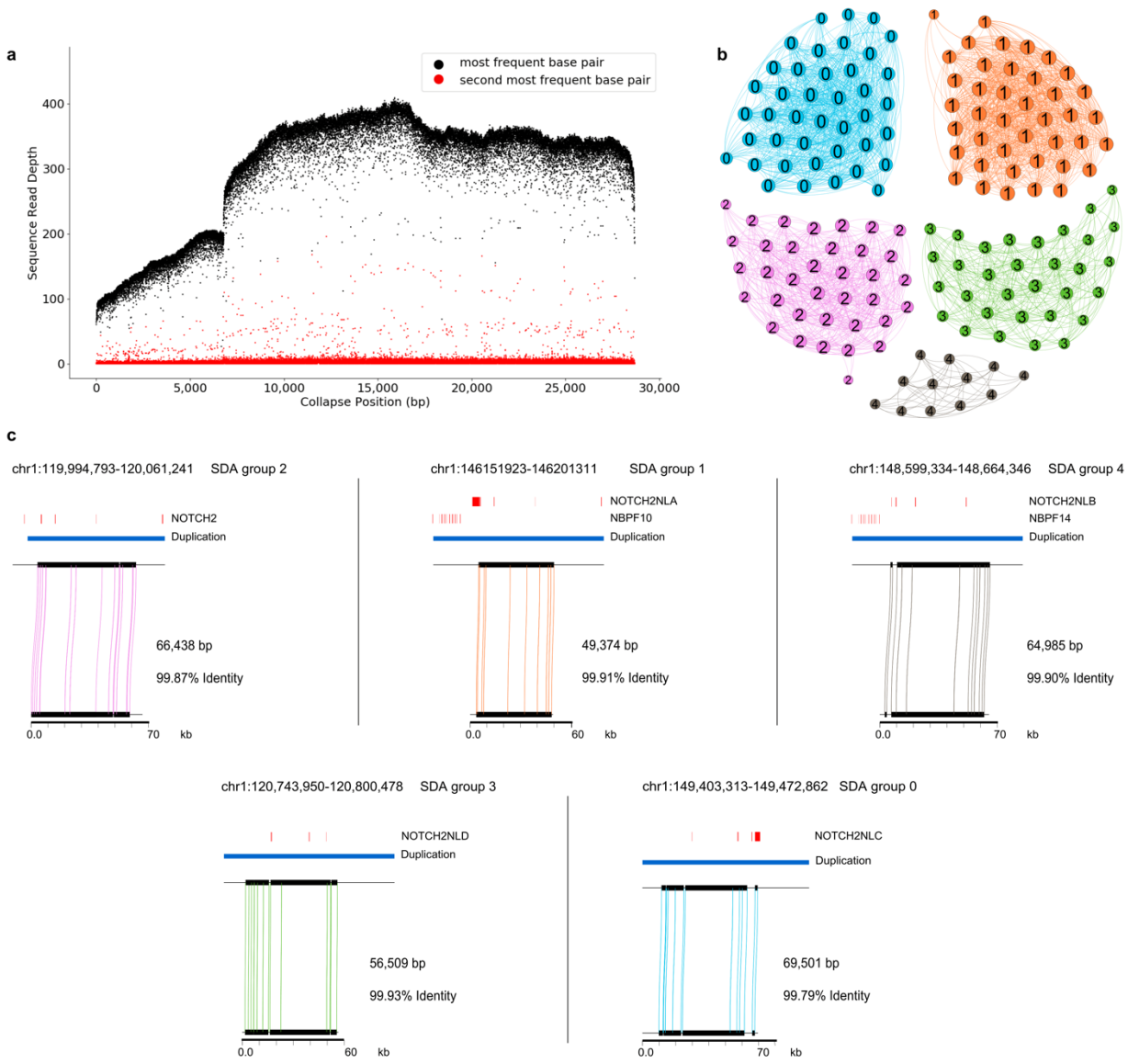


Figure S4. Sequence and assembly of *NOTCH2* loci in the CHM1 human genome. a) A collapsed representation of a portion of the *NOTCH2* loci is shown. Plotted is the read-depth profile over a collapsed representation of *NOTCH2*. Each black dot represents the coverage of the most frequent base pair at that position, while each red dot is the second most frequent. Secondary bases at low frequency represent sequencing error; however, those at high frequency represent PSV candidates. b) *NOTCH2* PSV graph resolves the collapse into five potential loci. c) The alignment of each SDA contig back to the loci for *NOTCH2* (./NLA/NLB/NLC/NLD) using Miropeats. Our assembled sequence is 99.88% identical over all five loci and >99.995% identical if only mismatched bases are counted as errors.

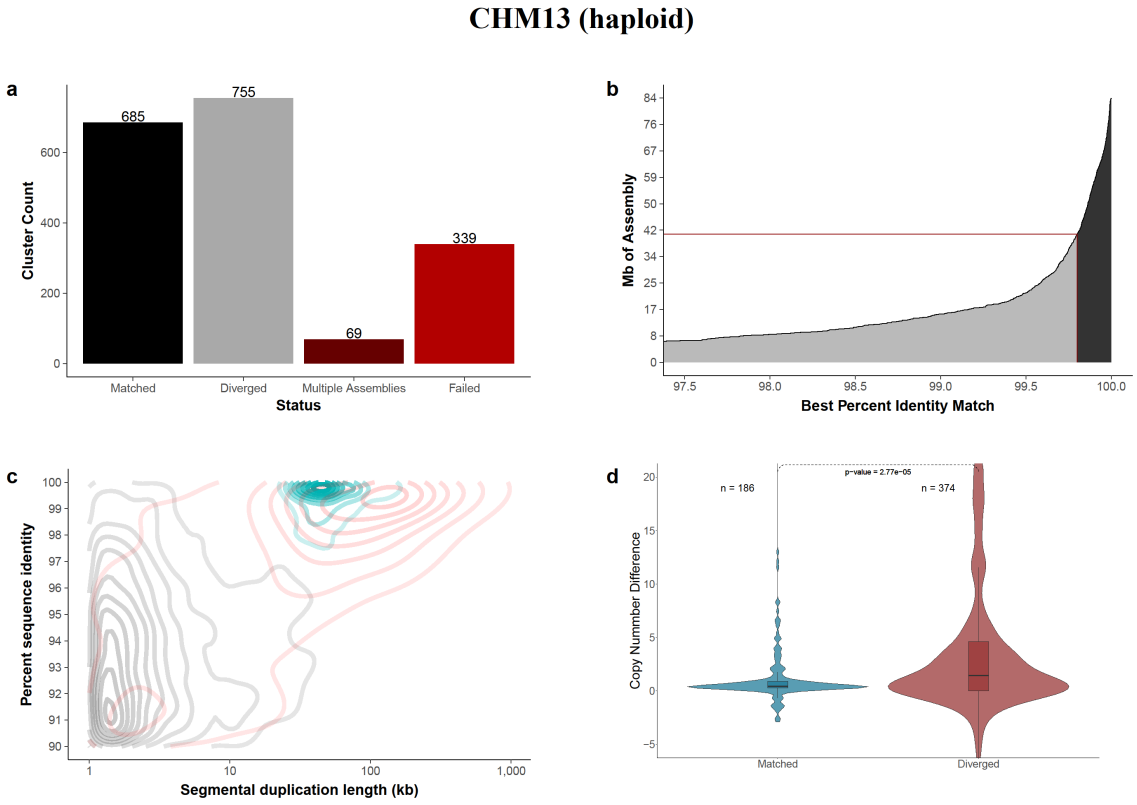


Figure S5. SDA results for the CHM13 assembly. a) SDA analysis of the CHM13 FALCON assembly generates 1,848 PSV clusters. b) Cumulative distribution of the assemblies and their percent identity to their best match in the reference. There are 40.4 Mbp of diverged assembly (gray) and 43.0 Mbp that map to the reference at high identity (black). c) A density plot of SDs plotted by length and percent identity. d) Copy number difference (CND) between CHM13 and the reference genome (CHM13 copy number - reference genome copy number) comparing n=186 SD regions that match (>99.8%) versus n=374 diverged SD regions (<99.8% identity). The mean CND of the matched sequence is 1.61 and the mean CND of the diverged sequence is 5.98, indicating that the diverged sequences are much more likely to represent additional duplicate copies that are unrepresented in the reference genome (GRCh38) (two-sided Mann-Whitney test; p=2.77*10-5). The boxes indicate the range between the first and third quartiles, with the bold line specifying the median. The whiskers show the minimum and maximum within 1.5 times the interquartile range extending from the first and third quartiles. (See Figure 2 for more details.)

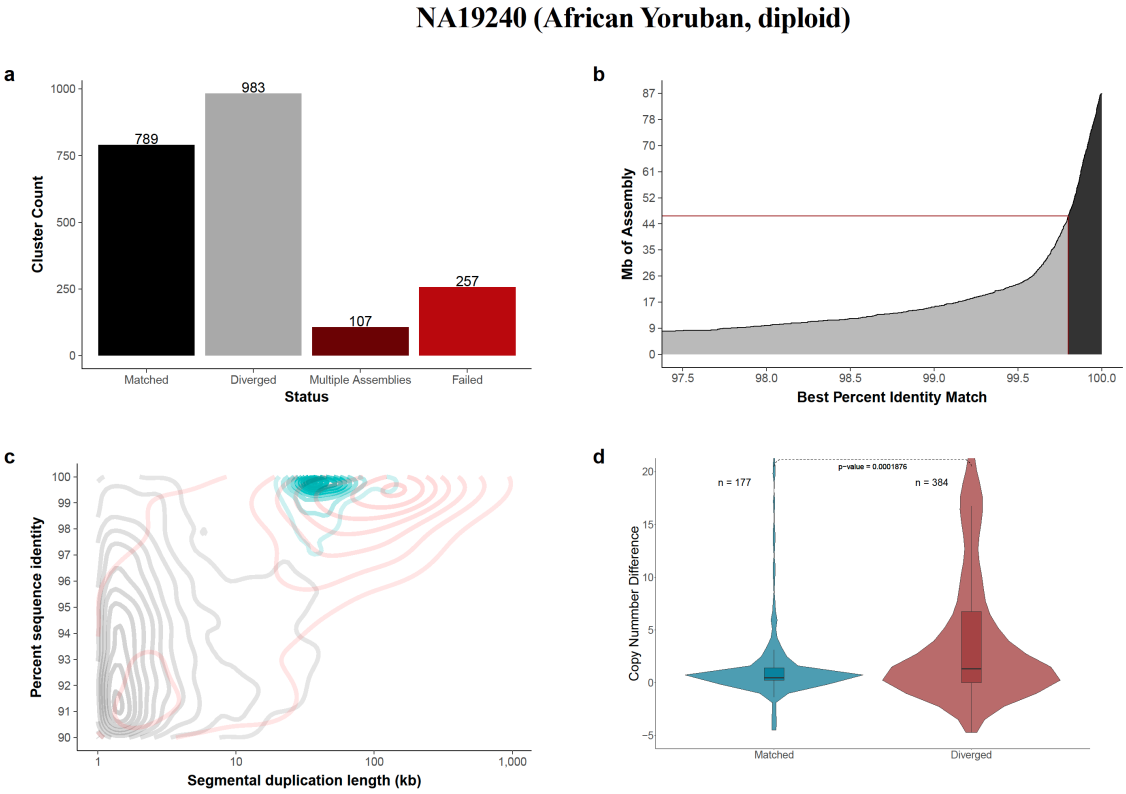


Figure S6. SDA results for the NA19240 (African Yoruban) assembly. a) SDA analysis of the NA19240 FALCON assembly generates 2,136 PSV clusters. b) Cumulative distribution of the assemblies and their percent identity to their best match in the reference. There are 46.1 Mbp of diverged assembly (gray) and 41.0 Mbp that maps to the reference at high identity (black). c) A density plot of SDs plotted by length and percent identity. d) CND between NA19240 and the reference genome (NA19240 copy number - reference genome copy number) comparing n=177 SD regions that match (>99.8%) versus n=384 diverged SD regions (<99.8% identity). The mean CND of the matched sequence is 4.11 and the mean CND of the diverged sequence is 10.87, indicating that the diverged sequences are much more likely to represent additional duplicate copies that are unrepresented in the reference genome (GRCh38) (two-sided Mann-Whitney test; p=1.88*10-4). The boxes indicate the range between the first and third quartiles, with the bold line specifying the median. The whiskers show the minimum and maximum within 1.5 times the interquartile range extending from the first and third quartiles. (See Figure 2 for more details.)

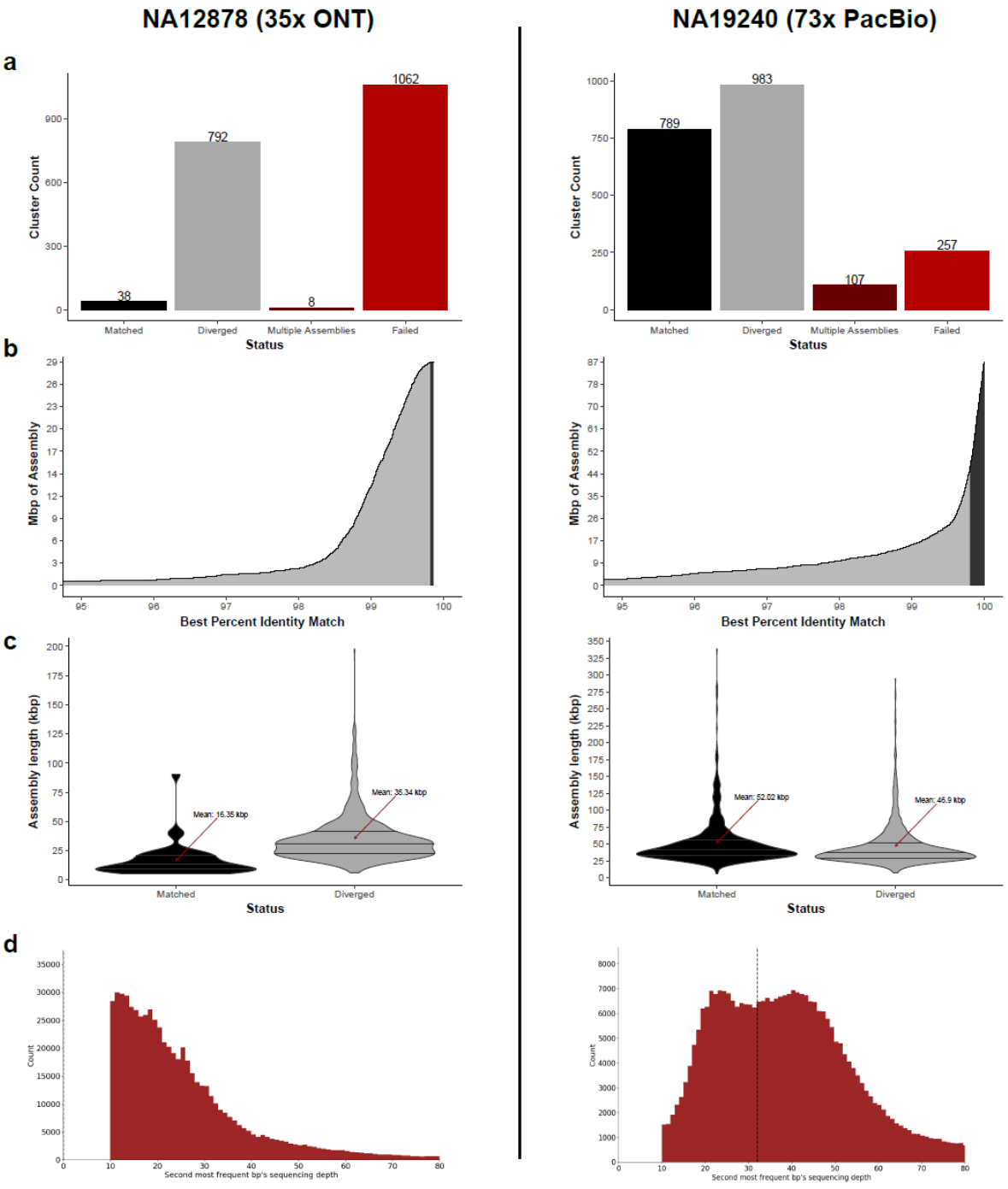


Figure S7. Comparison of SDA on ONT versus SMRT data. The left half of the figure shows the results of SDA applied to the ONT assembly of NA12878; on the right is the PacBio assembly of NA19240. a) SDA analysis of the NA12878 assembly generated 38 assemblies that mapped with >99.8% identity (matched) to GRCh38 and 792 mapped with <99.8% sequence identity (diverged). Failed clusters (n = 1,052) did not result in an assembly while multiple assemblies were PSV clusters with more than one contig produced by the Canu assembly. b) Cumulative distribution of the assemblies and their percent

identity to their best match in the reference. The number of assembly Mbp is calculated independently of a mapping to the reference. c) Length distribution of the matched and diverged assemblies (NA12878: matched n=38, diverged n=792; NA19240: matched n=789, diverged n=983). The lines on the violin plots indicate the first and third quartiles as well as the median. d) Sequencing read-depth distribution of the second most common SNV across all collapsed regions of SDs.

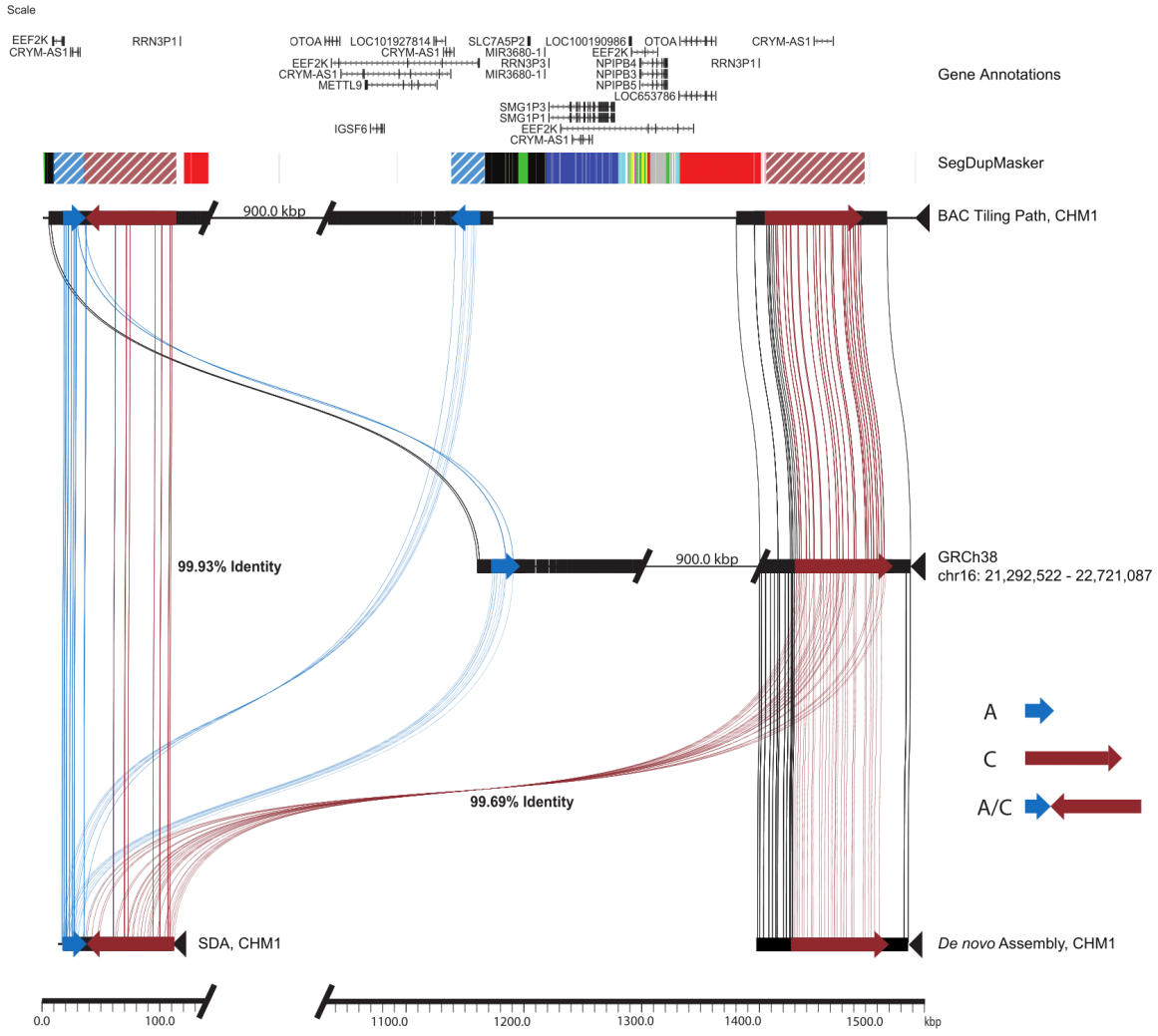


Figure S8. Sequence and assembly of a missing 16p12.1 duplication. The Miropeats alignments compare a BAC-based tiling path assembly of CHM1 (top line) to the human reference genome (GRCh38) (middle line) to a *de novo* assembly of CHM1 where SDA was applied (bottom line). The A/C duplication (red blue) proposed by Sudmant et al. that is present in most humans was correctly assembled using SDA and matches at high sequence identity (99.9%) to the BAC-based assembly structure.

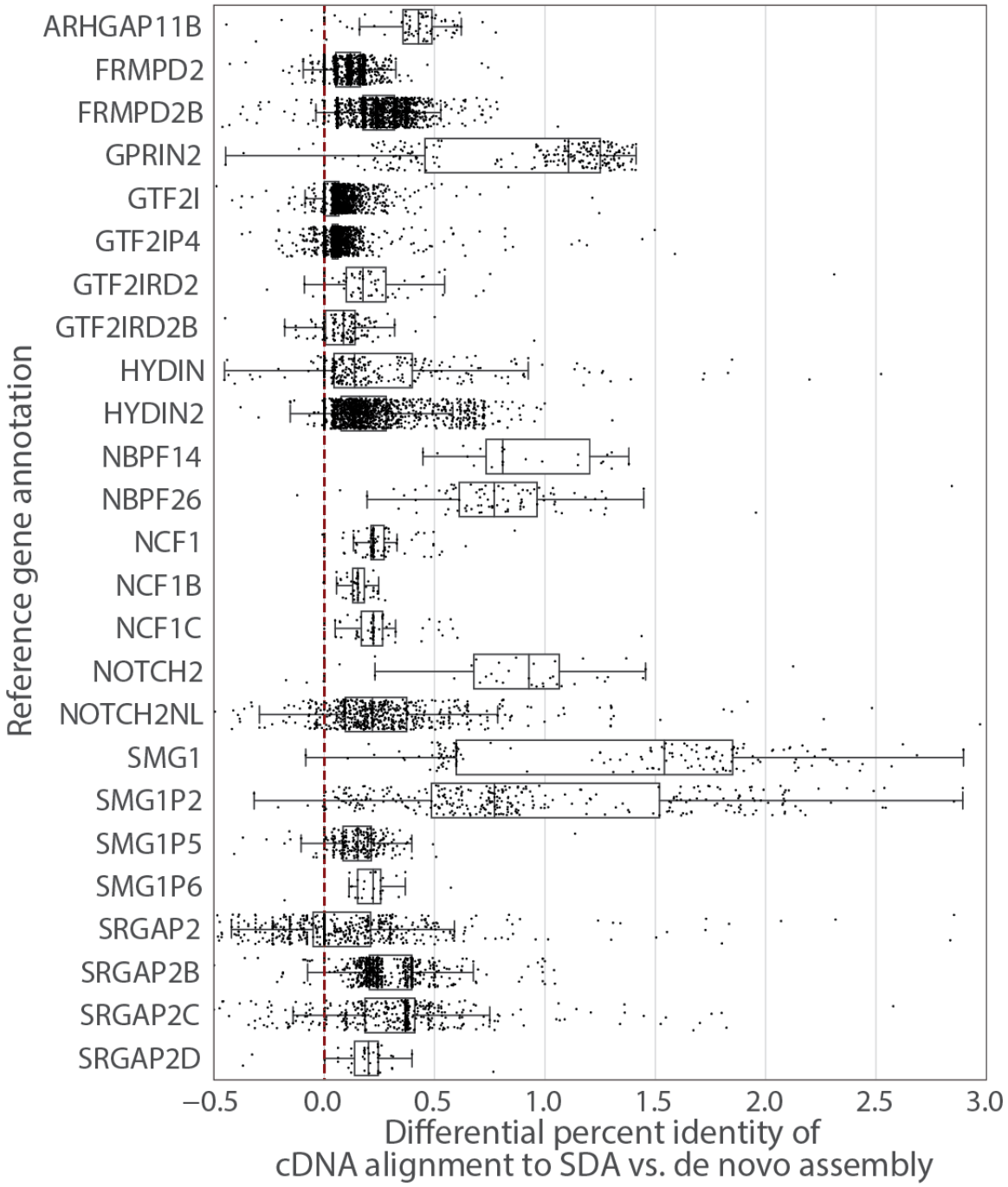


Figure S9. Mapping differential of transcripts between SDA and *de novo* CHM13. The percent identity differential of the mapping of full-length Iso-Seq transcripts (n=14,562) from human-specific segmental duplications (HSDs) to both the *de novo* assembly of CHM13 and the SDA results on CHM13 is shown. In total, 11 gene families showed significantly ($p < 0.001$, two-sided Wilcoxon signed-rank test) improved mapping to the SDA-resolved contigs. The boxes indicate the range between the first and third quartiles, with the bold line specifying the median. The whiskers show the minimum and maximum within 1.5 times the interquartile range extending from the first and third quartiles.

Figure S10. Multiple sequence alignment (MSA) between GRCh38 GPRIN2 and SDA GPRIN2A/B. Shown is the amino acid MSA between the copies of GPRIN2 resolved by SDA and the copy of GPRIN2 in GRCh38. Of the 15 differences in the MSA, 12 are annotated in dbSNP as variants in GPRIN2 when they are in fact differences between GPRIN2A and GPRIN2B. At p.Ser104Gly, p.Arg242Gly, and p.Val375Ala, the reference has the minor allele. Table S9 shows the allele frequencies for all variants seen in this alignment.

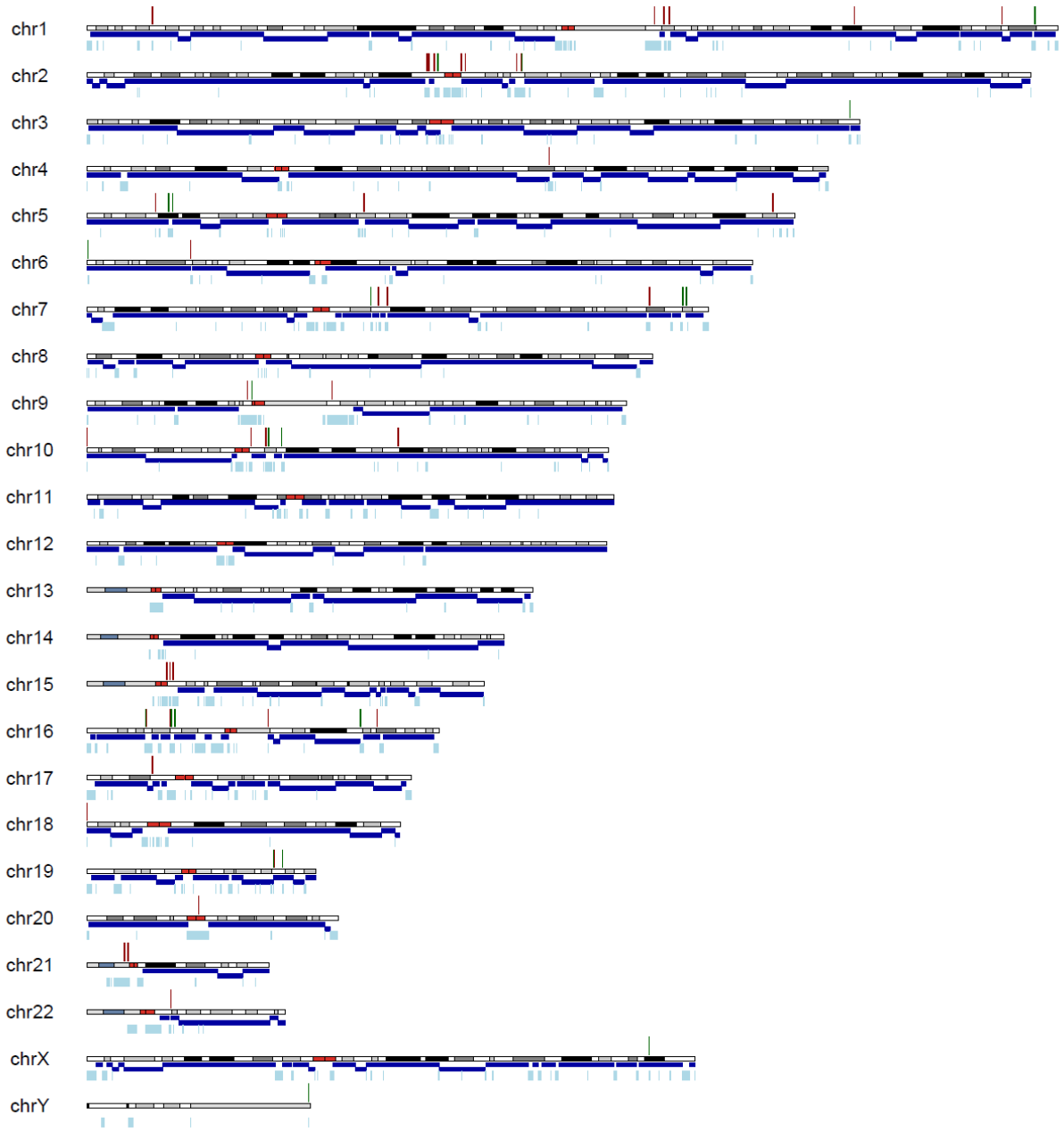


Figure S11. CHM1 SDA contigs that overlap with unique sequence. This ideogram shows where SDA contigs could extend the FALCON assembly. The bottom panel of each chromosome shows the FALCON assembly (contigs >1 Mbp (dark blue), contigs <1 Mbp (light blue)). The top panel shows where SDA contigs with unique overlaps map along the reference (contigs with >10 kbp of overlap (green), contig with <10 kbp (red)).

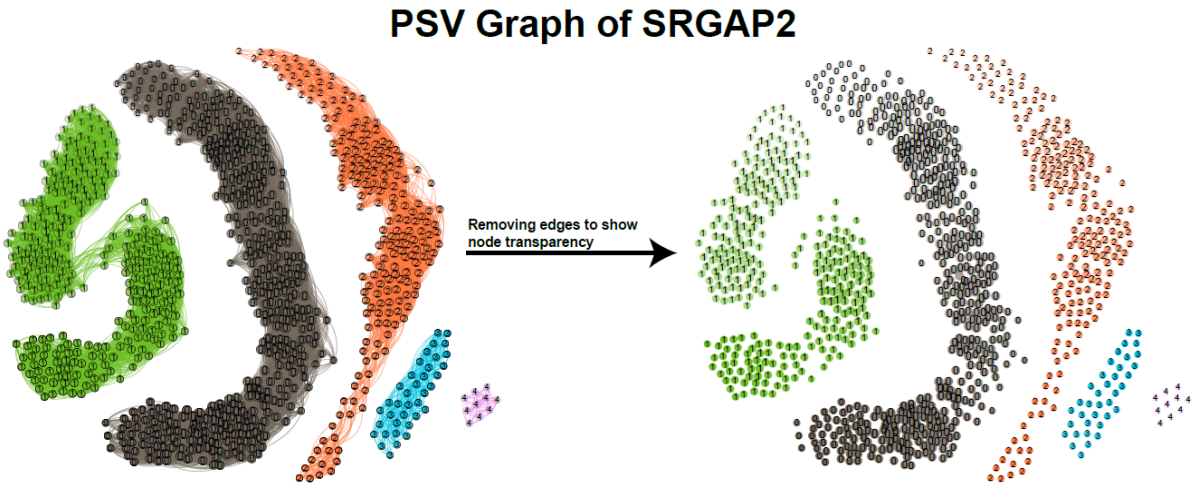


Figure S12. PSV graph without attraction edges. Reproduced above is the PSV graph shown in Figure 2.3 for *SRGAP2*. The left-hand side shows the attraction edges used in correlation clustering (CC). On the right-hand side, the edges are removed so that the transparency of the nodes is visible. The opacity of each node scales from 0.25 to 1, with 0.25 reflecting the start position on the contig and 1 representing the final position on the contig.

A.12 Supplemental Tables

Table S1. Fraction of resolved SDs in different *de novo* assemblies.

Assembly	**Mbp of aligned SD***	**% Resolved SD****	**Read Coverage**
AK1 GCA_001750385.2	113.6	24.3	101
CHM1 GCA_001297185.1	140.9	29.2	61
CHM1 Internally Assembled	116.1	29.4	61
CHM13 GCA_002884485.1	115.5	27.9	73.6
HX1 GCA_001708065.2	125.3	23.2	103
HX1 Canu	125.8	23.3	103
HX1 HERA	131.2	33.3	103
NA12878 Jain 2018	128.4	32.9	35.4 ONT (4.73 >50 kbp)
NA12878 Jain 2018 update	121.6	32.9	38.1 ONT (5.04 >50 kbp)
Yoruban GCA_001524155.4	123.7	29.3	73x

* Mbp of sequence aligned to the reference over annotated SDs.
** Percent of annotated SDs in the reference that are resolved in the *de novo* assembly.
For an SD to be resolved, the aligned contig must extend 50 kbp past the SD into unique space on both sides.

Table S2. Status of disease-mediating SDs in the FALCON CHM1 assembly.

Disease	Type of Rearrangement	Location	Coordinates	Mbp	OMIM	PMid	Resolved in CHM1
Charcot Marie tooth disease type 1A	Interstitial duplication	17p12	chr17:14,446,995-16,048,139	1.5	118220	11584295	No
Hereditary neuropathy with pressure palsies	Deletion	17p12	chr17:14,456,878-16,038,255	1.5	162500	11584295	No
SMS Smith Magenis syndrome	Deletion	17p11.2	chr17:15,112,335-20,380,493	5	182290	11584295	No
Potockl-Lupski syndrome	Interstitial duplication	17p11.2	chr17:16,100,000-22,700,000	5	610883	11584295	No
Neurofibromatosis type1 NF1	Deletion	17q11.2	chr17:30,293,798-32,178,804	1.5	162200	11584295	Yes
Prader-Willi syndrome	Deletion	15q11-15q13	chr15:22,833,353-26,969,005	4	176270	11584295	No
Angelman syndrome	Deletion	15q11-15q13	chr15:23,351,093-27,425,225	4	105830	11584295	Yes
Chromosome 15q11-q13 duplication syndrome	Supernumerary marker chromosome	15q11-15q14	chr15:20,083,333-24,416,666	4	608636	11584295	No
Williams Beuren syndrome	Deletion	7q11.23	chr7:74,529,630-76,070,370	1.6	194050	11584295	No
DiGeorge and velocardiofacial	Deletion	22q11.2	chr22:17,977,414-21,562,880	3	188400	11584295	No
Cat eye syndrome	Supernumerary marker chromosome	22q11.2	chr22:18,500,000-21,999,999	3	115470	11584295	No
X-linked ichthyosis	Deletion	xp22	chrX:6,329,207-8,172,686	1.9	308100	11584295	No
Hemophilia A	Inversion	Xq28	chrX:154,648,851-155,209,658	0.5	306700	11584295	Yes
Male infertility AZFa microdeletion	Deletion	yq11.2	chrY:12,344,706-13,146,789	0.8	415000	11818139	No
Male infertility AZFc microdeletion	Deletion	yq11.2	chrY:11,007,537-14,554,536	3.5	415000	11818139	No

A list of diseases and syndromes caused by large genomic rearrangements as described in Emanuel 2001 and Stankeiwicz 2002, and if they are contiguously assembled past the duplication boundaries in the CHM1 genome assembly.

Table S3. Sequence and assembly of *SRGAP2* and *NOTCH2NL* gene families.

Gene	SDA Group	Status	Percent Identity	GRCh38 Location	Length	Number of PSVs	# PSVs in GRCh38
SRGAP2	0	Resolved	99.96	chr1:206,210,031-206,407,594	197,525	451	407
SRGAP2C	1	Resolved	99.99	chr1:121,189,099-121,388,229	199,064	299	287
SRGAP2B	2	Resolved	99.99	chr1:144,893,160-145,088,264	195,041	203	188
SRGAP2D	3	Resolved	99.97	chr1:143,980,898-144,061,839	67,989	37	37
SRGAP2D	4	Resolved	99.89	chr1:143,980,898-144,061,839	21,945	10	0
NOTCH2NLC	0	Resolved	99.79	chr1:149,403,313-149,472,862	69,501	41	34
NOTCH2NLA	1	Resolved	99.91	chr1:146151923-146201311	49,374	41	41
NOTCH2	2	Resolved	99.87	chr1:119,994,793-120,061,241	66,438	36	35
NOTCH2NLD	3	Resolved	99.93	chr1:120,743,950-120,800,478	56,509	32	28
NOTCH2NLB	4	Resolved	99.9	chr1:148,599,334-148,664,346	64,985	12	12

The percent identity (GRCh38), contig length, and number of PSVs for four copies of *SRGAP2* and five copies of *NOTCH2NL* are shown.
Sequences were resolved by SDA and correlation clustering.

Table S4. CHORI-17 BAC clone sequences.

Too large to include see: 10.1038/s41592-018-0236-3

Table S5. BAC clone sequence analysis.

Too large to include see: 10.1038/s41592-018-0236-3

Table S6. Gene content analysis.

Too large to include see: 10.1038/s41592-018-0236-3

Table S7. Differences in the multiple sequence alignment of GPRIN2A/B and the reference copy of GPRIN2.

Position	Consequence	hg38_GPRIN	SDA_GPRIN2A	SDA_GPRIN2B	RSID*	Allele Frequency**	Het freq**	Allele Number**
5	p.Arg5His	R	R	H	rs3127817	0.50	1.00	193796
39	p.Leu39Val	L	V	L	rs4926045	0.48	0.96	232084
40	p.Arg40His	R	R	H	rs3127818	0.50	0.99	251024
47	p.Val47Met	V	V	M	rs3127819	0.50	0.99	251292
91	p.Trp91Arg	W	W	R	rs3127820	0.50	1.00	270506
100	p.Thr100Pro	T	P	T	rs7090312	0.48	0.97	260714
104	p.Ser104Gly	S	G	G	rs3127679	0.89	0.22	271564
202	p.Gly202Trp	G	W	G	rs1120465 8	0.48	0.96	255588
233	p.Ala233Ser	A	S	A	rs1120465 9	0.48	0.97	260010
239	p.Arg239Lys	R	K	R	rs7895979	0.48	0.97	257192
240	p.Met238_Glu240dup	-	-	MRE	rs1126204 25	0.50	1.00	309120
242	p.Arg242Gly	R	G	G	rs5540908 11	0.88	0.23	276863
348	p.Val348Leu	V	L	V	rs4926046	0.36	0.72	245486
375	p.Val375Ala	V	A	A	rs3127822	0.99	0.00	277212
400	p.Leu400Pro	L	L	P	rs3127823	0.50	1.00	275948

*Results from dbSNP.
**Results from gnomAD.

Table S8. Comparison of SDA and parameterized *de novo* assemblies of 10 individual collapses.

Region of Collapse	**SDA assembly size (kbp)**	**SDA %ID**	**De novo assembly sizes (kbp)***	**De novo % IDs***	**% increase in bases by using SDA***
SRGAP2	679.5	99.62	334.5 - 366.6	99.51 - 99.54	85.36 - 103.14
ROCK1	299.8	98.22	119.2 - 133.4	97.83 - 99.23	124.75 - 151.51
NPY4	493.2	99.25	224.7 - 276.8	99.1 - 99.14	78.16 - 119.45
NOTCH2	568	99.79	398.7 - 471.5	99.53 - 99.63	20.49 - 42.49
NAIP	345.7	99.47	234.1 - 323.9	99.01 - 99.48	6.75 - 47.7
HYDIN2	626.4	99.84	419.1 - 433.1	99.57 - 99.61	44.62 - 49.46
GTF2H2	122.1	99.49	62.6 - 74.9	99.36 - 99.51	63.11 - 95.12
FRMPD2	234.1	99.59	146.3 - 182.8	99.59 - 99.63	28.05 - 60.04
FCGR	128.1	99.08	97.6 - 166.3	99.21 - 99.27	-22.99 - 31.2

* Ranges reflect the minimum and maximum result from different *de novo* runs of Canu on the collapse.

Appendix B. Supplement for chapter 3

B.1 Polishing the assemblies.

Initially, the HiFi assembly was polished with Racon using the approximate alignments from minimap2 (i.e., the PAF output generated using the -x asm5 option) and fasta input of HiFi reads. However, we found that this polishing step only modestly increased the QV (by <0.01). When we polished the HiFi assembly with the exact alignments (i.e., the SAM output generated using the -ax map-pb option) and fastq input, we observed a large increase in the median QV (from 40.4 to 45.0). In addition, we observed that the QV achieved using these Racon parameters was greater than that achieved with Arrow (which used >1 TB of CCS subreads). Polishing a second time with Racon further increased the QV and significantly reduced the number of gene-disrupting indels. Adding Pilon polishing did not change the median QV but significantly reduced the total QV across all the BACs because it introduced a 660 bp insertion that appears to be an error relative to the AC275297.1 BAC. Additionally, Pilon polishing only reduced the number of indels genome wide by 645 out of 683,564 (0.094%) and resolved only one additional gene-disrupting event in unique sequence. It is, therefore, our suggestion to polish assemblies generated with HiFi data with two rounds of Racon using the parameters described above rather than with Arrow or Pilon.

B.2 BAC divergence.

In all of our polished assemblies (HiFi or CLR; Table 1), we noticed that the same two BACs (AC270121.1 and AC275290.1) had the lowest QV values of those assessed (Figure S2). We examined the alignments of these BACs to all the assemblies and to the HiFi reads and found that these BACs had contractions in tandem repeats relative to the CHM13 cell line. In AC270121.1, there was a 338 bp deletion of a (TCCCCC)n repeat, and in AC275290.1, there was

an 80 bp deletion in a (GGCTGAGG)n repeat. In addition, AC270136.1 showed a 62 bp expansion in a poly(T) tract, where the HiFi data supported the HiFi assembly, and AC270122.1 showed an 83 bp insertion, where both Illumina and HiFi data supported the HiFi assembly. Across all 31 BACs used for calculating QV, there was only one mismatched base (AC275285.1:148688-148688). This base appears to be correct in the HiFi assembly, as it was observed in both the HiFi and Illumina data (Figure S3). In combination, these results indicate that many of the BACs with QV < 40 are diverged in sequence when compared to the CHM13 genome due to a mutation that likely arose during BAC generation and/or clonal propagation and do not represent an error in the assemblies. For this reason, our QV values should be interpreted as a lower bound of the true QV.

B.3 Additional assemblies.

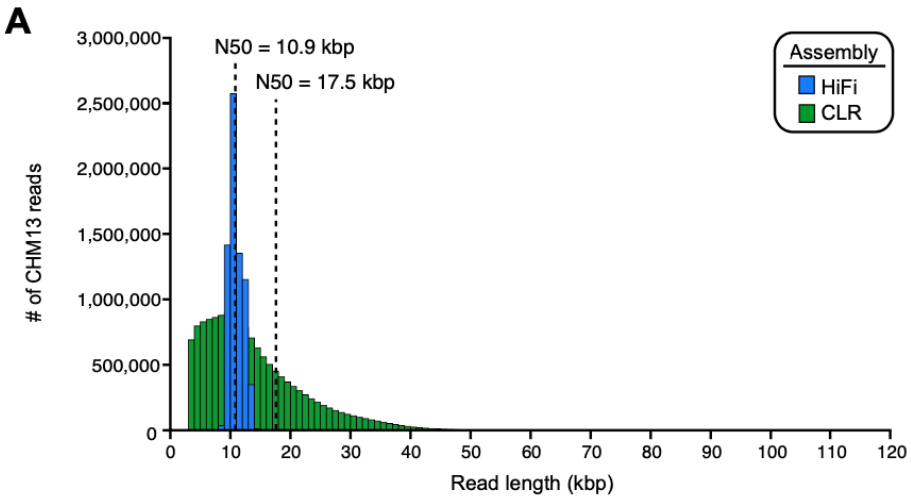
To determine whether the improvements in genome assembly quality observed in the HiFi assembly (Figures 1-3, Tables 1-4) are due to the assembler or the data type, we performed two control experiments. First, we generated a HiFi assembly using FALCON rather than Canu (hereafter termed “HiFi,FALCON”); second, we also produced a CLR assembly using Canu rather than FALCON [using a downsampled CLR dataset that has the equivalent coverage as the HiFi dataset (24-fold rather than 77-fold); hereafter termed “CLR,Canu”] (Table S1). Our results suggest that the improvements in the HiFi genome assembly quality are due to the data type and not the assembler.

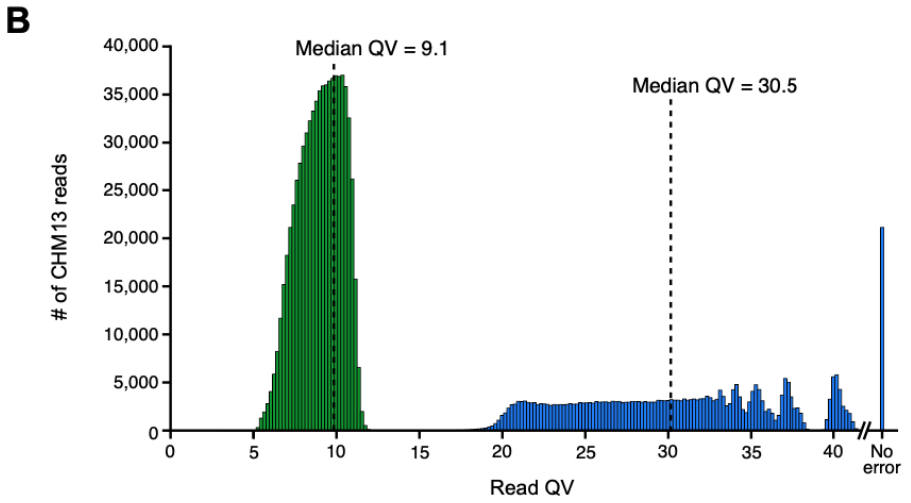
For the HiFi,FALCON assembly, we find that it is highly comparable to the HiFi,Canu assembly in size (3.00 Gbp vs. 3.03 Gbp), quality (median BAC QV of 44.45 vs. 45.25), and compute time (~4,400 CPU hours vs. ~2,800 CPU hours). The contiguity of the HiFi,FALCON assembly is improved compared to the HiFi,Canu assembly (N50 31.92 vs. 25.51) and slightly exceeds the CLR,FALCON assembly (N50 29.26). Additionally, error correction with Arrow on

the HiFi,FALCON assembly resulted in a higher quality assembly than when Quiver was applied to the CLR,FALCON assembly (QV 43.54 vs. 40.73), and two rounds of polishing with Racon performs better than polishing with Arrow on the HiFi,FALCON assembly (QV 44.45 vs. 43.54). Quiver polishing is not supported for sequencing data generated on the Sequel II, so it was not possible to generate a direct comparison on Quiver polishing. Compute time for the HiFi,FALCON assembly (~4,400 CPU hours) is slightly longer than the HiFi,Canu assembly (~2,800 CPU hours); however, the HiFi,FALCON assembly is still more than ten times faster than the CLR,FALCON assembly (>50,000 CPU hours). Overall, the HiFi,FALCON and HiFi,Canu assemblies are similar in terms of size (3.00 vs. 3.03 Gbp), median BAC QV (44.45 vs. 45.25), compute time (~4,400 vs. ~2,800 CPU hours), and contiguity (31.92 vs. 25.51).

For the CLR,Canu assembly, we find that it is not comparable to the HiFi,Canu assembly, and this is mainly due to the data type used to generate the assembly. When CLR coverage is downsampled to that of HiFi coverage (24-fold) and assembled with the same assembler as HiFi data (Canu), it produces a much smaller assembly (2.48 Gbp vs. 3.03 Gbp) with a much lower contiguity (N50 0.31 Mbp vs. 25.51) and many more contigs (21,267 vs. 5,296) while taking much longer compute time (~37,300 vs. ~2,800 CPU hours). Additionally, the CLR,Canu assembly has a median BAC QV that is much lower than the HiFi,Canu assembly (26.50 vs. 40.41). Overall, we find that the downsampled CLR,Canu assembly is much lower quality than the HiFi,Canu assembly, and this is largely due to the data type.

B.4 Supplemental Figures





C

	Read QV (% accuracy) including indels		Read QV (% accuracy) excluding indels	
	Median	**Mean**	**Median**	**Mean**
HiFi	30.5 (99.91%)	23.4 (99.54%)	No error (100.0%)	29.6 (99.89%)
CLR	9.1 (87.67%)	8.8 (86.78%)	16.1 (97.53%)	15.4 (97.10%)

Figure S1. Distribution of the CHM13 HiFi and CLR read lengths and quality values (QVs). A) Histogram of the CHM13 HiFi (blue) and CLR (green) read lengths. The read N50 of each dataset is shown. B) Histogram of the CHM13 HiFi (blue) and CLR (green) read QVs. QVs were estimated by aligning reads from each dataset to the curated, telomere-to-telomere assembly of the CHM13 X chromosome (Miga et al., 2019) and counting the differences in the alignments as errors in the reads. More than half (54.6%) of the HiFi

reads are QV 30 or greater. C) Table listing the CHM13 HiFi and CLR read QV and accuracy. Mean and median values were calculated by aligning reads from each dataset to the CHM13 telomere-to-telomere X chromosome assembly (Miga et al., 2019). Values were calculated with and without indels to demonstrate the high indel error rate in CLR data.

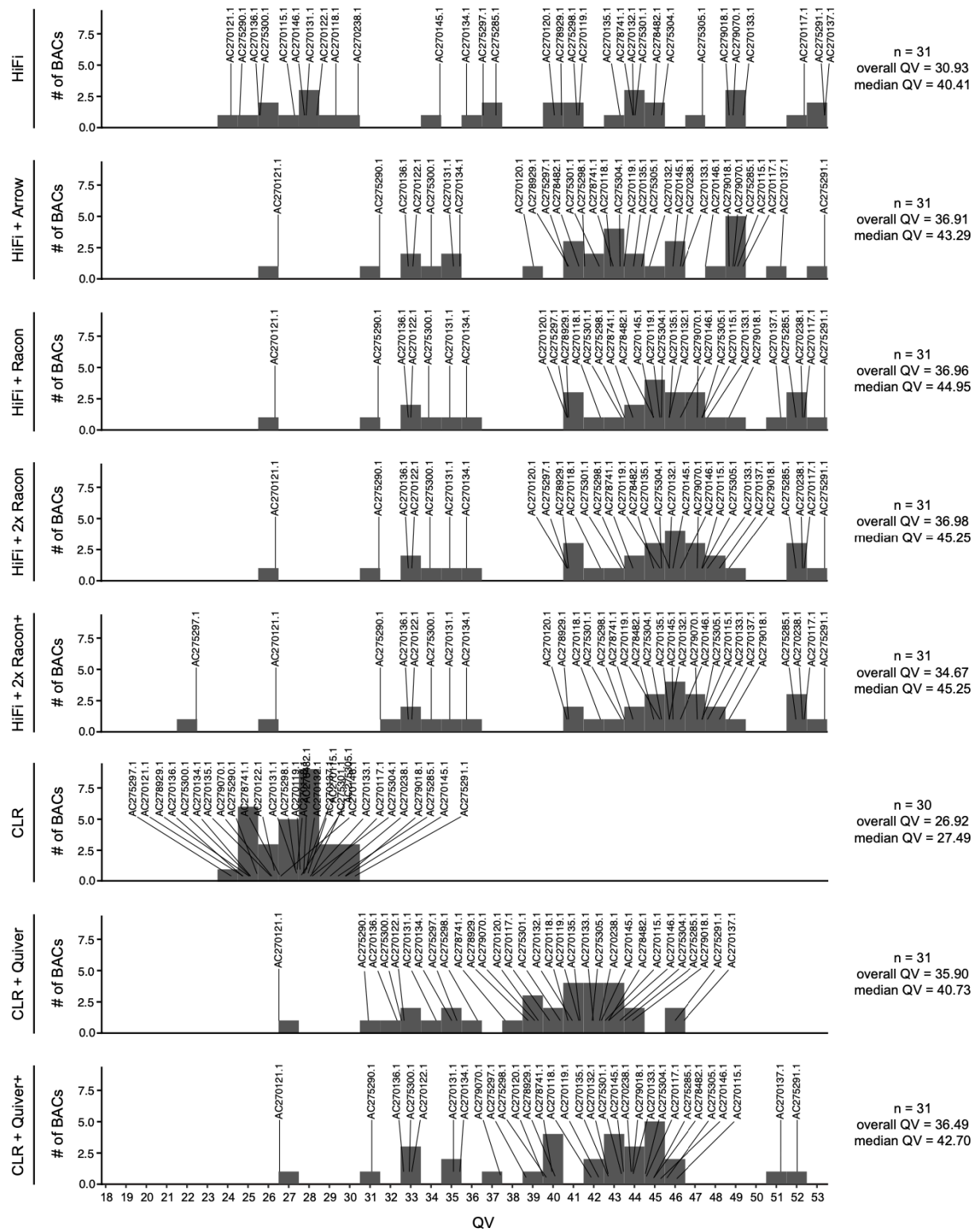


Figure S2. Assessment of QV score of each genome assembly with varying levels of polishing. The QV score histogram was derived from the alignment of 31 BACs to the indicated assembly. Each BAC clone accession name is indicated, and the overall and median QV scores for each genome assembly are shown.

Figure S3. Assessment of a base mismatch between a BAC clone and the HiFi assembly. Out of all 31 BACs assessed, there was a single mismatch observed between one BAC and the HiFi assembly (highlighted here in red and boxed with a dashed line). However, both the HiFi and Illumina data supported the base in the HiFi assembly ('C' rather than 'A'), indicating a divergence in sequence between BAC clone AC275285.1 and the CHM13 cell line (see **Supplemental Notes**).

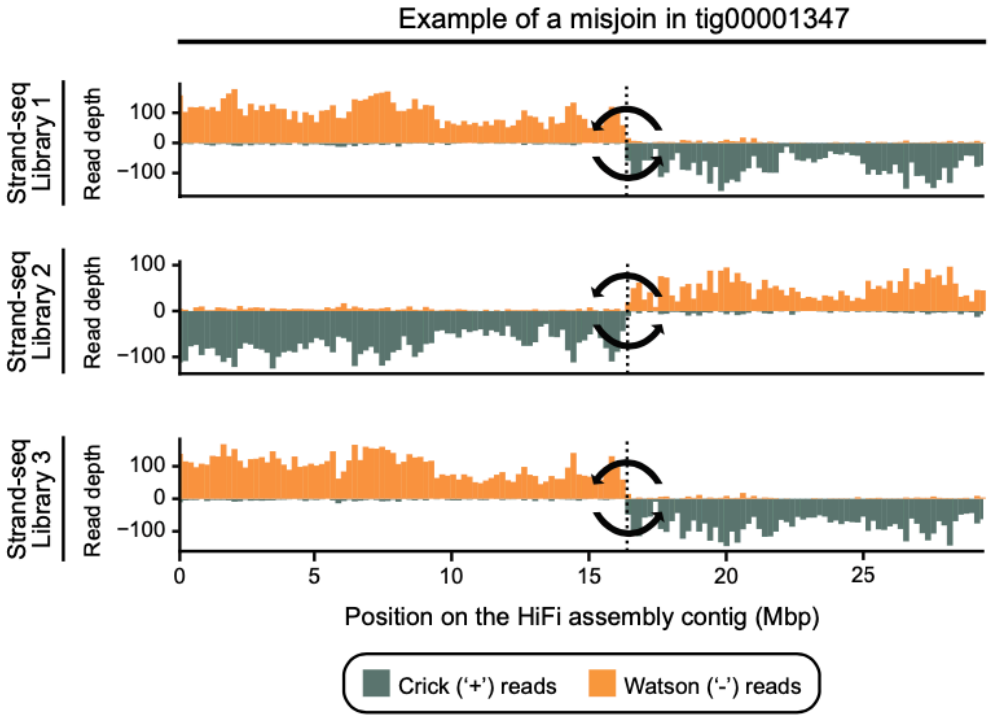


Figure S4. Example of a misjoined contig in the HiFi assembly. Shown is an example of a misjoined contig in the HiFi assembly**.** Reads mapping to the plus (Crick; teal) or minus (Watson; orange) strand of the reference genome are plotted as vertical bars along the contig. Each row shows one Strand-seq library. A recurrent change in read directionality in the middle of the contig suggests that left and right portions of this contig have flipped orientation with respect to each other and have likely been misjoined during the assembly process.

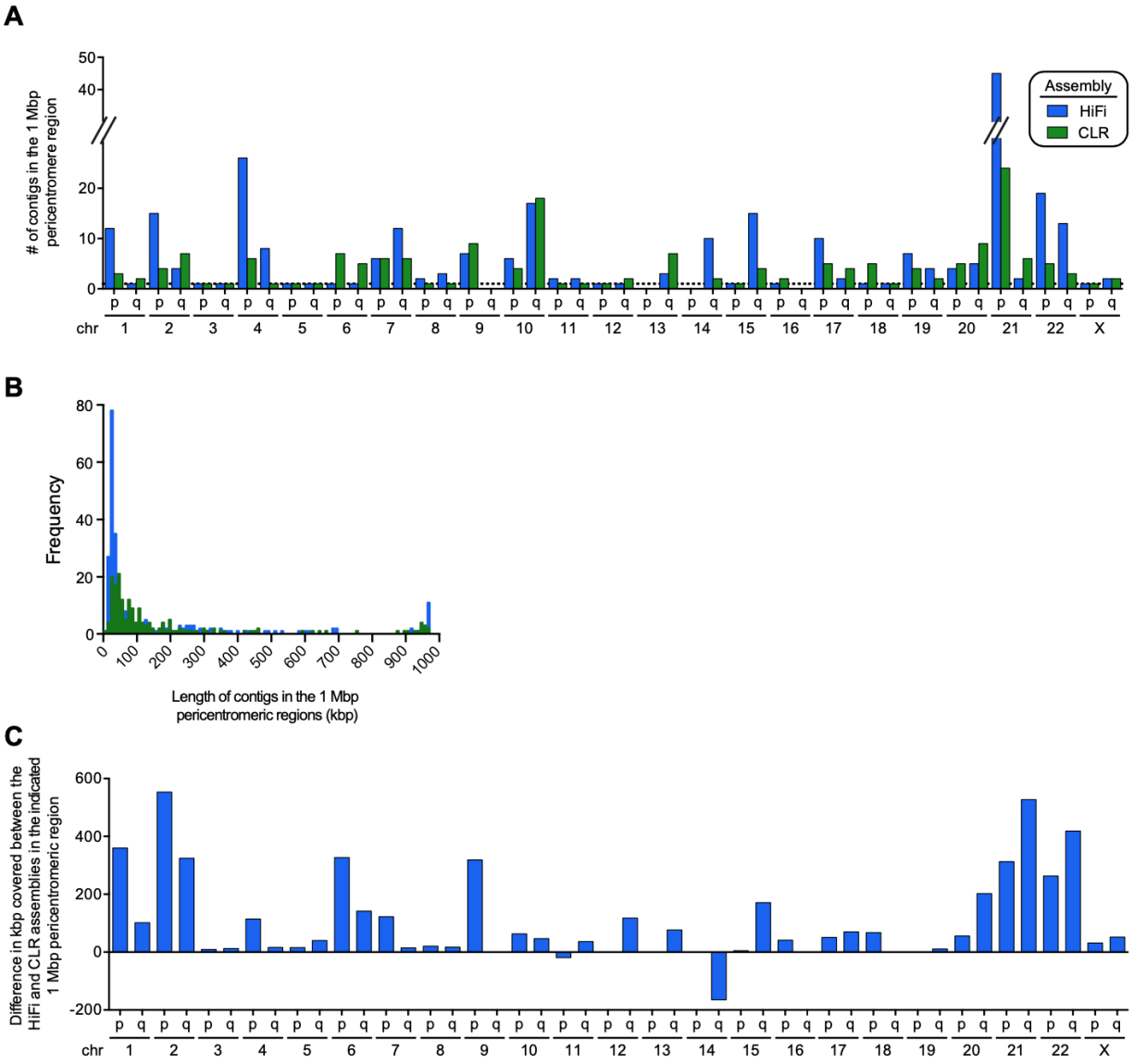


Figure S5. Assessment of continuity in the pericentromeric regions in the HiFi and CLR assemblies. A) Plot of the number of contigs in the 1 Mbp regions flanking each centromere in the HiFi and CLR assemblies. The majority of the pericentromeric regions in the HiFi assembly (52.2%) contained either a reduced number of contigs or the same number of contigs. The remaining pericentromeric regions either contained no contig (8.7%) or an increased number of contigs (39.1%) in the HiFi assembly relative to the CLR assembly. B) Histogram of the length of contigs in the 1 Mbp regions flanking the centromeres for each assembly. The HiFi assembly has more contigs than the CLR assembly overall, with an increase in the number of small contigs (<100 kbp) and large contigs (900-1000 kbp). The average contig length is 145.8 kbp in the HiFi assembly and 177.6 kbp in the CLR assembly. C) Plot of the difference in sequence coverage in the 1 Mbp regions flanking each centromere for the HiFi and CLR genome assemblies. Nearly all pericentromeric

regions contain additional sequences in the HiFi assembly relative to the CLR assembly. The HiFi assembly contains an additional 5.03 Mbp of pericentromeric sequence missing in the CLR assembly.

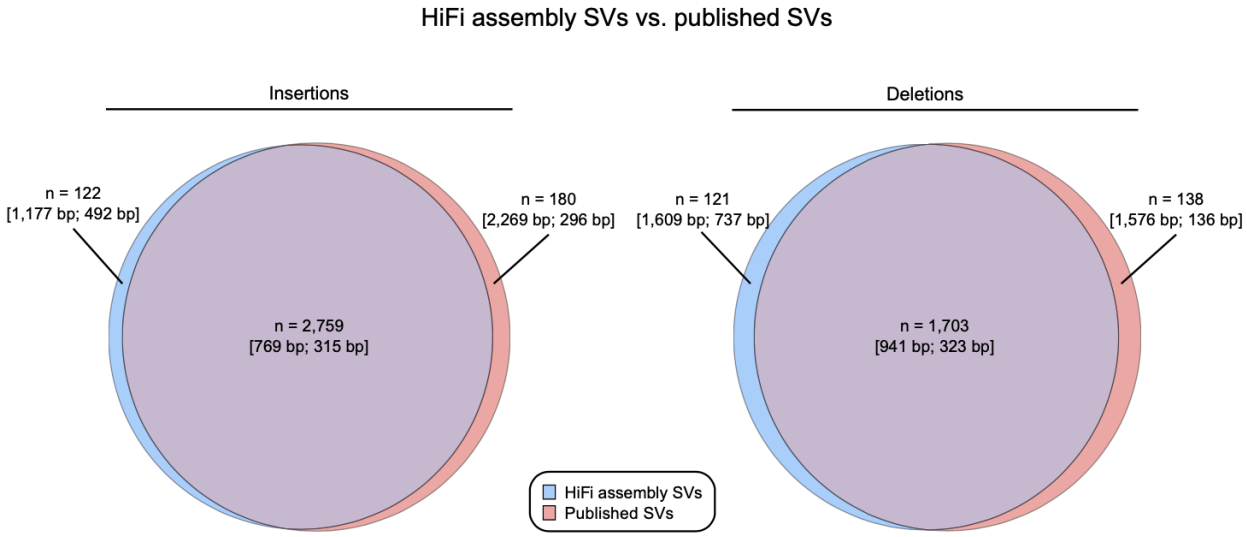


Figure S6. SVs discovered in the HiFi assembly are supported by published CHM13 calls. We intersected SVs with published CHM13 SVs excluding tandem repeat and segmental duplication (SD) loci, where variant comparisons are more challenging. Both insertions (left) and deletions (right) are strongly supported. Each Venn area is annotated with the total number of variants (n) along with the mean and median variant size, respectively, in brackets. For both insertions and deletions, the HiFi assembly calls more variants around 700 bp to 1 kbp, but the published variants have more calls in the 50-100 bp range as well as more larger calls (10+ kbp). These disagreements are reflected in the mean and medians, and they may be due to assembly errors or differences in mapping and assembly methods used in each study.

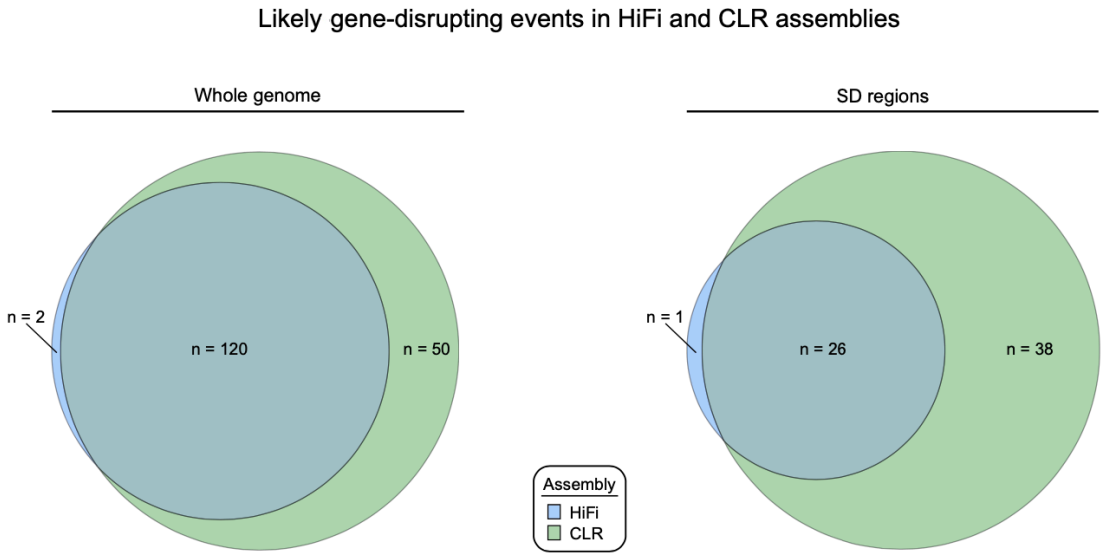


Figure S7. Disrupted genes in the HiFi assembly supported by the CLR assembly. For all loci where the polished HiFi and CLR assemblies had a single alignment to the reference (left), all but two genes disrupted by the HiFi assembly have support in the CLR assembly. We observe 50 genes disrupted in CLR without HiFi support (29% of CLR-disrupted genes). When we restrict our analysis to SDs (right), the percentage of genes disrupted in the CLR assembly without HiFi support increases to 59%.

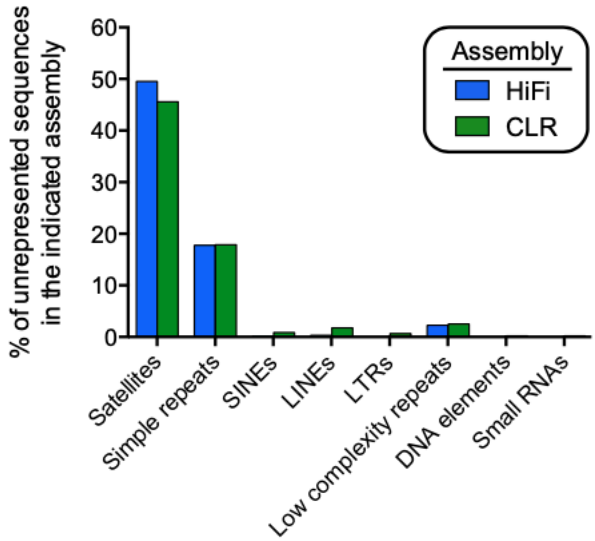


Figure S8. Repeat content of unassembled reads. Bar plot of the repeat composition of sequences not incorporated into the HiFi and CLR assemblies. Most of the unrepresented sequences consist of satellite repeats mapping to heterochromatin or pericentromeric DNA

(centromeres, acrocentric DNA and secondary constrictions of chromosomes). SINE, small interspersed nuclear element; LINE, long interspersed nuclear element; LTR, long terminal repeat.

B.5 Supplemental Tables

Table S1. Statistics of the HiFi and CLR genome assemblies produced by different assemblers

	Polishing	Total size (Gbp)	N50 (Mbp)	No. of contigs	Median QV	No. of CPU hours for assembly
HiFi (24-fold) *CHM13 genome Canu assembly*	None	3.03	25.51	5,296	40.41	~2,800
	Arrow	3.03	25.51	5,296	43.29	~10,000
	Racon	3.03	25.51	5,296	44.95	~2,950
	2x Racon	3.03	25.51	5,296	45.25	~3,100
	2x Racon+	3.03	25.51	5,296	45.25	~4,200
HiFi (24-fold) *CHM13 genome FALCON assembly*	None	3.00	31.91	2,115	27.65	~4,400
	Arrow	3.00	31.92	2,115	43.45	~12,400
	2x Racon	3.00	31.92	2,115	44.45	~4,700
CLR (77-fold) *CHM13 genome FALCON assembly*	None	2.88	29.26	1,916	27.49	>50,000
	Quiver	2.88	29.26	1,916	40.73	>55,000
	Quiver+	2.88	29.26	1,916	42.70	>55,000
CLR (24-fold) *CHM13 genome Canu assembly*	None	2.48	0.31	21,167	26.50	~37,300

HiFi: HiFi assembly
CLR: CLR assembly
ONT: Oxford Nanopore Technologies
2x Racon: Two rounds of Racon
2x Racon+: Two rounds of Racon and one round of Pilon
Quiver+: Quiver, Pilon, and FreeBayes-based indel correction
Median QV: Median QV over 31 BACs

Table S2. False joins identified by Strand-seq within *de novo* assembly contigs

	Contig	Start coordinate	End coordinate	No. of false joins
HiFi				
CHM13 genome				
Canu assembly	tig00000017	99,602,681	99,618,276	1
	tig00001347	16,388,858	16,426,941	1
	tig00003369	22,812,690	22,841,943	1
	tig00002385	3,056,424	3,204,103	1
	tig00002385	10,092,563	10,210,819	1
	tig00001433	2,544,909	2,612,082	1
	tig00004976	785,171	805,928	1
CLR				
CHM13 genome				
FALCON assembly	NTIA01000004.1	51,797,117	51,869,304	1
	NTIA01000039.1	16,068,231	16,072,573	1
	NTIA01000061.1	12,087,358	12,166,683	1
	NTIA01000067.1	4,639,556	4,640,945	1
	NTIA01000093.1	746,665	783,795	1

HiFi: HiFi assembly

CLR: CLR assembly

Table S3. Comparison of PSVs linked with SDA in HiFi and CLR assemblies

Locus	Expected no. of paralogs	Expected no. of phased bases (kbp)	No. of paralogs by SDA (HiFi, CLR)	No. of bases phased by SDA (HiFi, CLR; kbp)	Average length of SDA-phased block (HiFi, CLR; kbp)	No. of PSVs (HiFi, CLR)
OPN1LW	2	70	0, 1	0, 20	0, 20	0, 14
NOTCH2NL	5	500	9, 4	443, 345	49, 86	728, 482
SRGAP2	4	520	5, 5	493, 494	99, 99	1194, 821
FCGR2/3	3	220	3, 4	141, 140	47, 35	611, 139
KANSL1	2	280	4, 3	54, 150	13, 50	48, 92

Appendix C. Supplement for chapter 4

C.1 Supplemental Figures

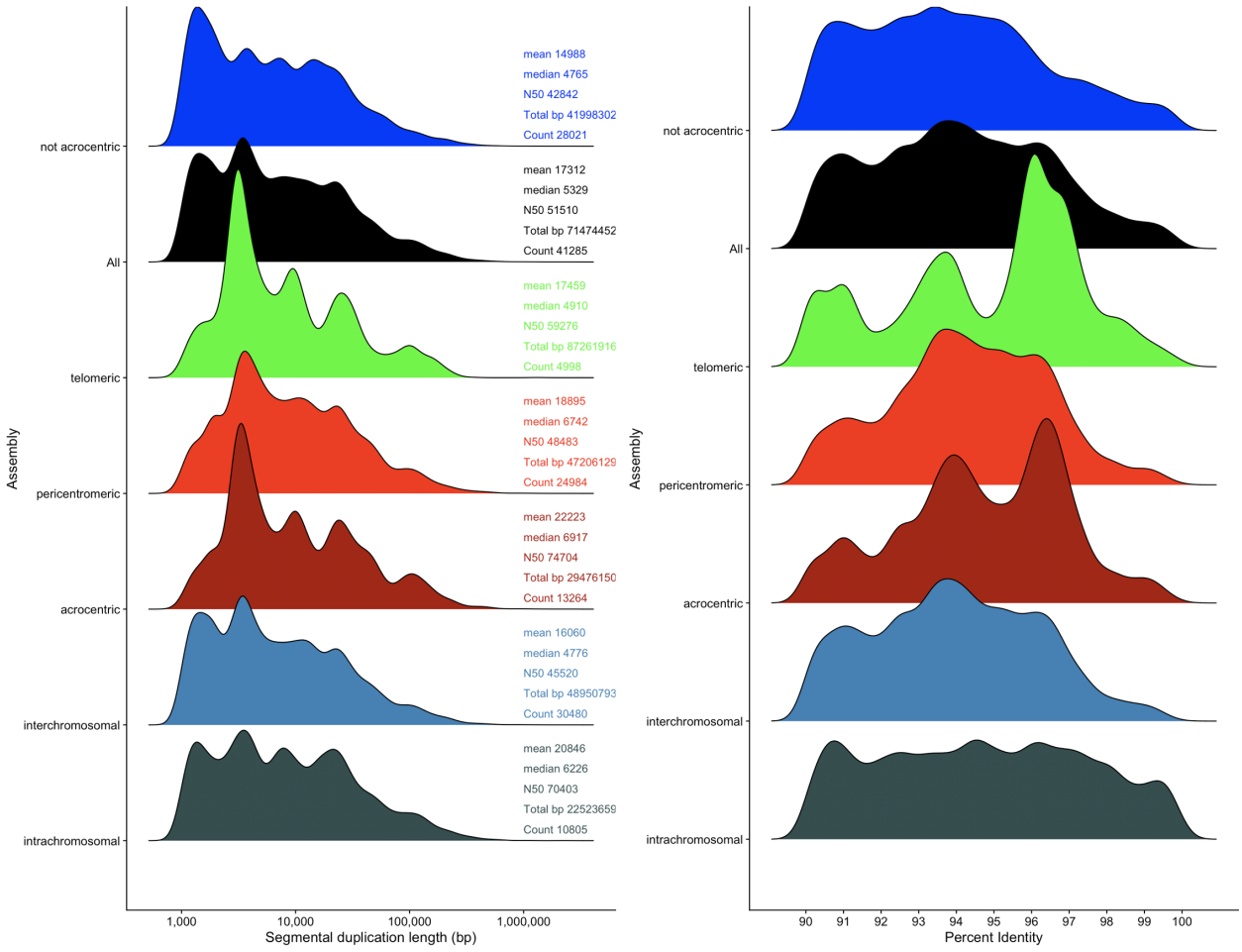


Figure S1. Comparison of SD length and identity in different regions of the genome. The length (left) and identity (right) of SDs across commonly delineated regions of the genome (colors).

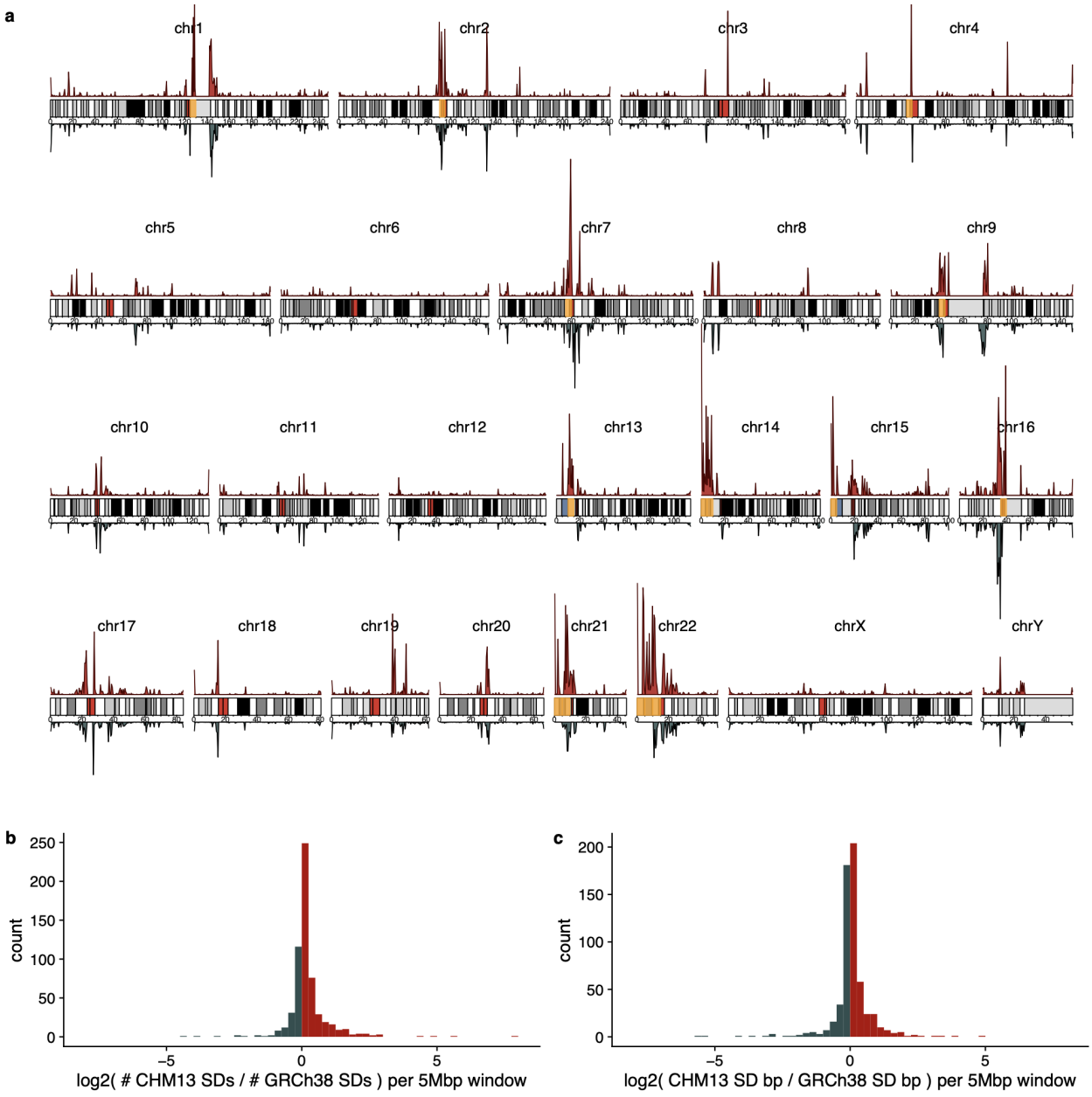


Figure S2. SD density comparison between CHM13 T2T and GRCh38. a) Density of SDs in T2T CHM13 (red) and GRCh38 (blue). In the ideogram highlighted in orange are the 15 regions with the largest increase in the number of SDs. b) Histogram showing the $\log_2$ fold change between the number of SDs in T2T CHM13 and GRCh38 per non-overlapping 5 Mbp window. c) Histogram showing the $\log_2$ fold change between the number of bp in SDs for CHM13 T2T and GRCh38 per non-overlapping 5 Mbp window.

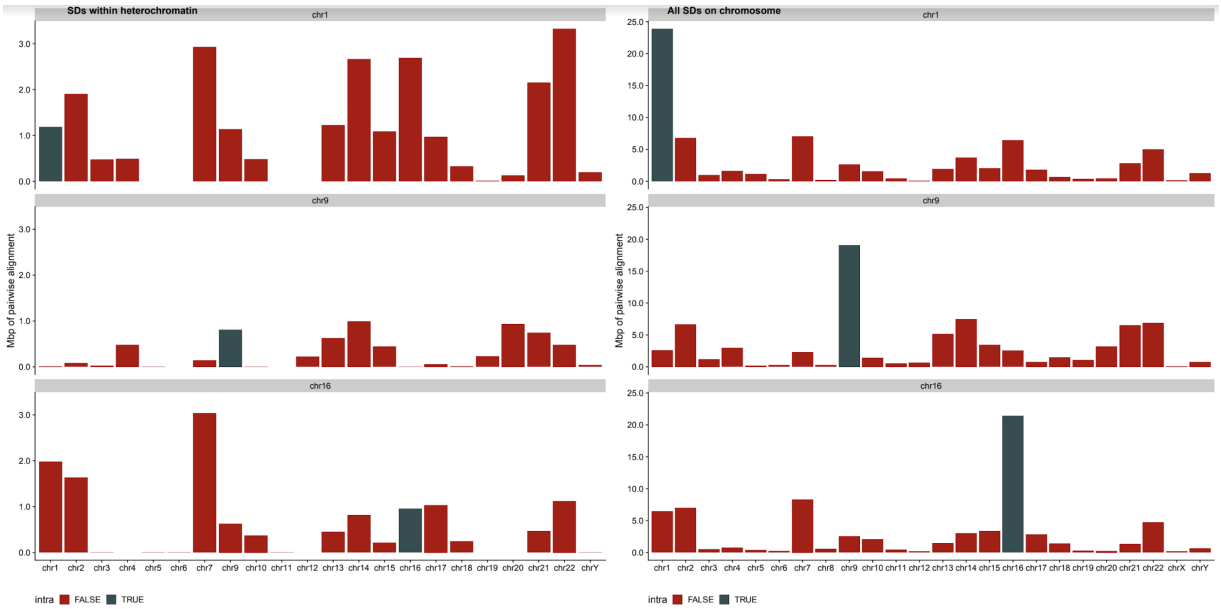


Figure S3. SDs within heterochromatin on chromosomes 1, 9, and 16. This figure shows where the SD clusters that separate the HSAT and centromere arrays on chromosomes 1, 9, and 16 align to (left) compared to the overall distribution of that chromosome (right). Blue are intrachromosomal SDs and red are interchromosomal.

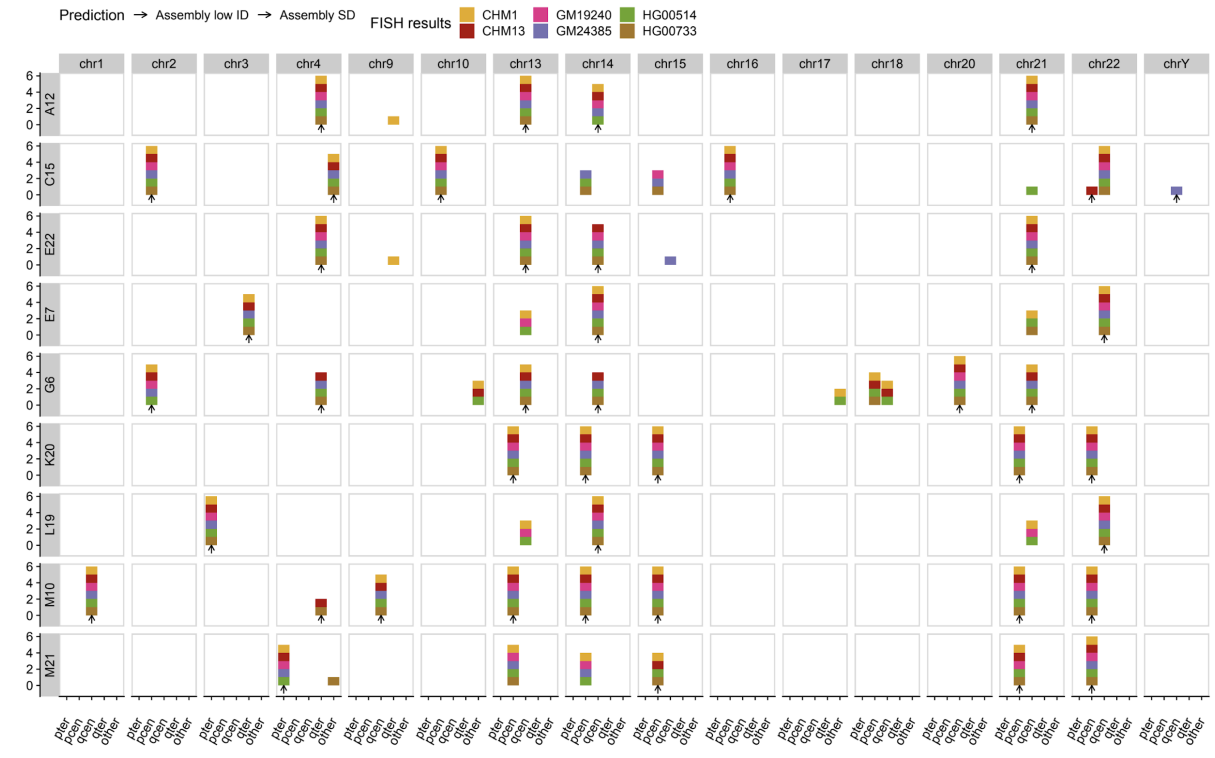


Figure S4. Support for new CHM13 duplications by FISH. This table shows the location and chromosomes of FISH signals (x) for each probe (y) across the different cell lines (color). Black shows the predicted locations of FISH signals from the T2T CHM13 assembly.

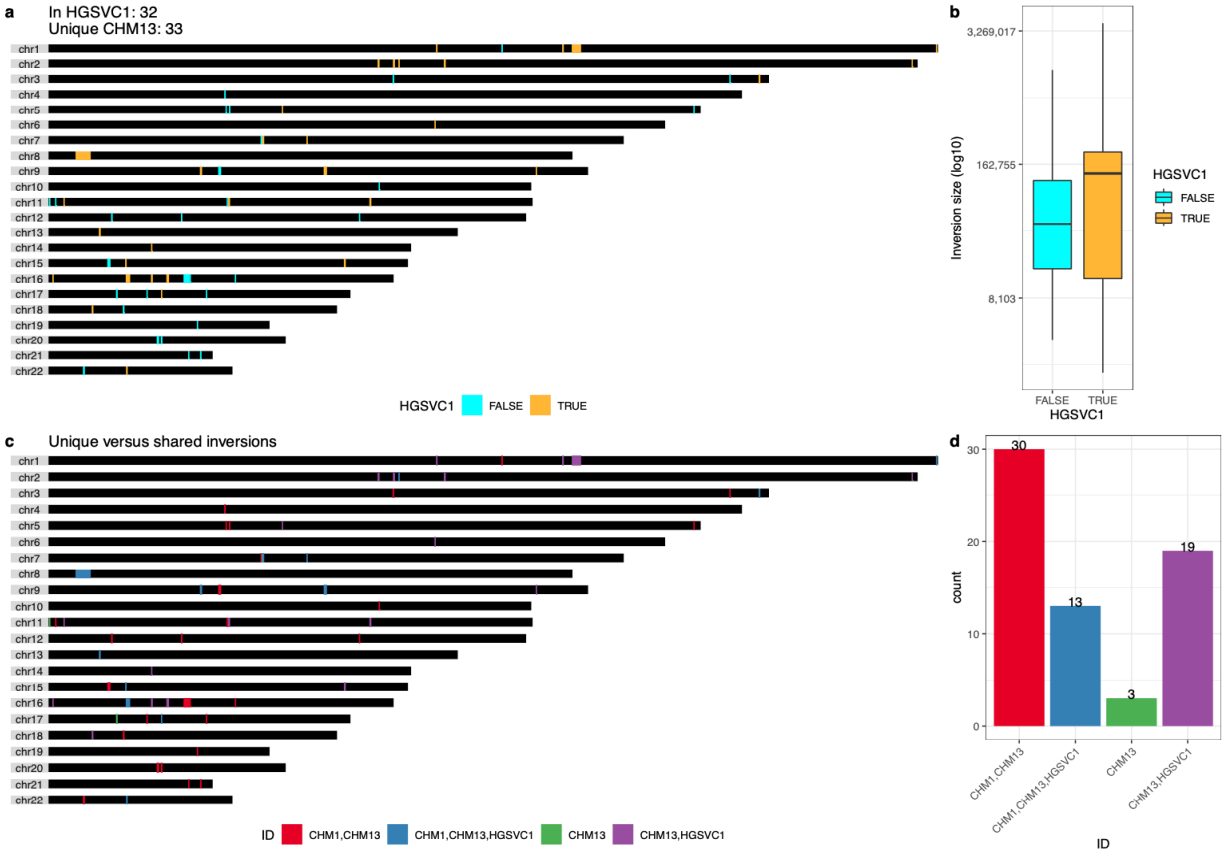


Figure S5. CHM13 inversions supported by Strand-seq. a) Locations of inversions in CHM13 relative to GRCh38 as identified with Strand-seq. The color indicates whether the inversion is shared (orange) with at least one sample from HGSVC1 (Chaisson et al. 2019) or unique to CHM13 (cyan). b) Size distribution of inversions in CHM13. c) Comparison of inversions shared between CHM1 and CHM13. d) Bar chart showing the counts of shared and unique inversions in CHM13.

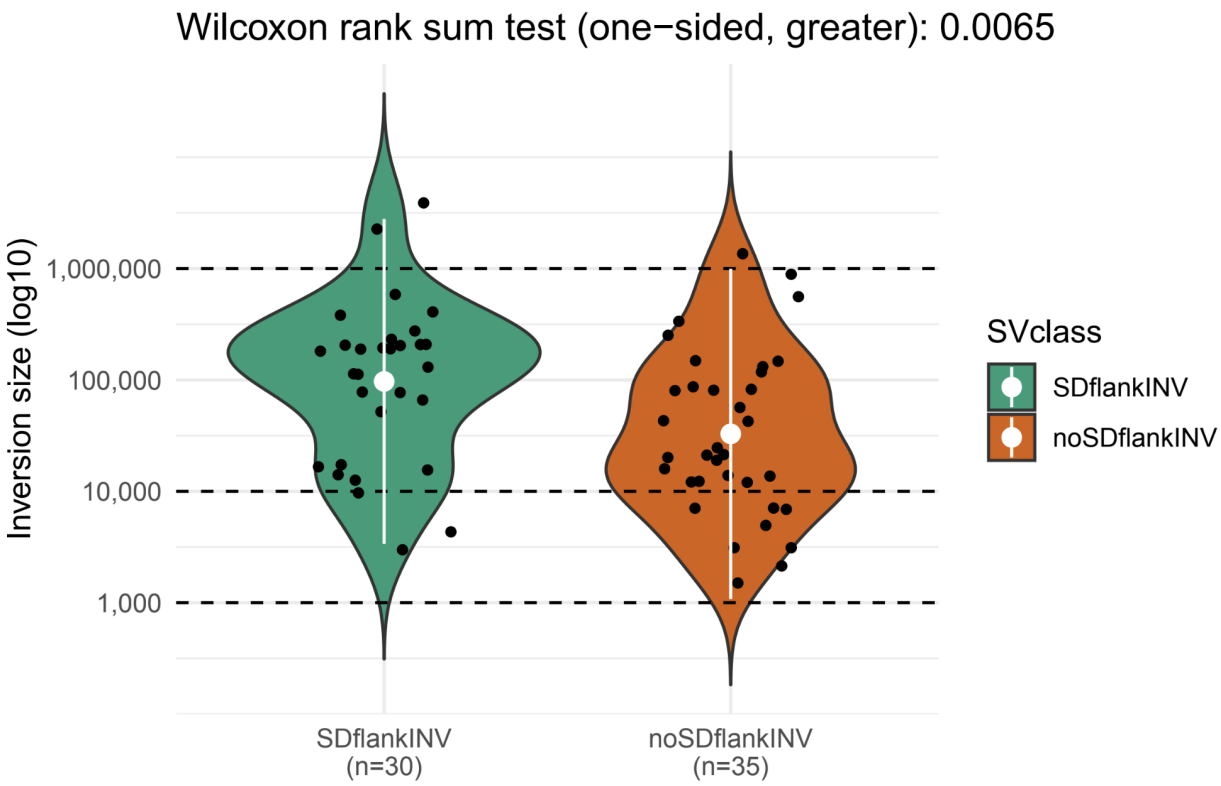


Figure S6. Length of CHM13 inversions. The length of inversions in CHM13 as predicted by Strand-seq stratified by the presence of flanking SDs (green) or lack thereof (orange). Inversions flanked by SDs are significantly longer than other inversions ($p = 0.0065$, one-sided Wilcoxon rank-sum test).

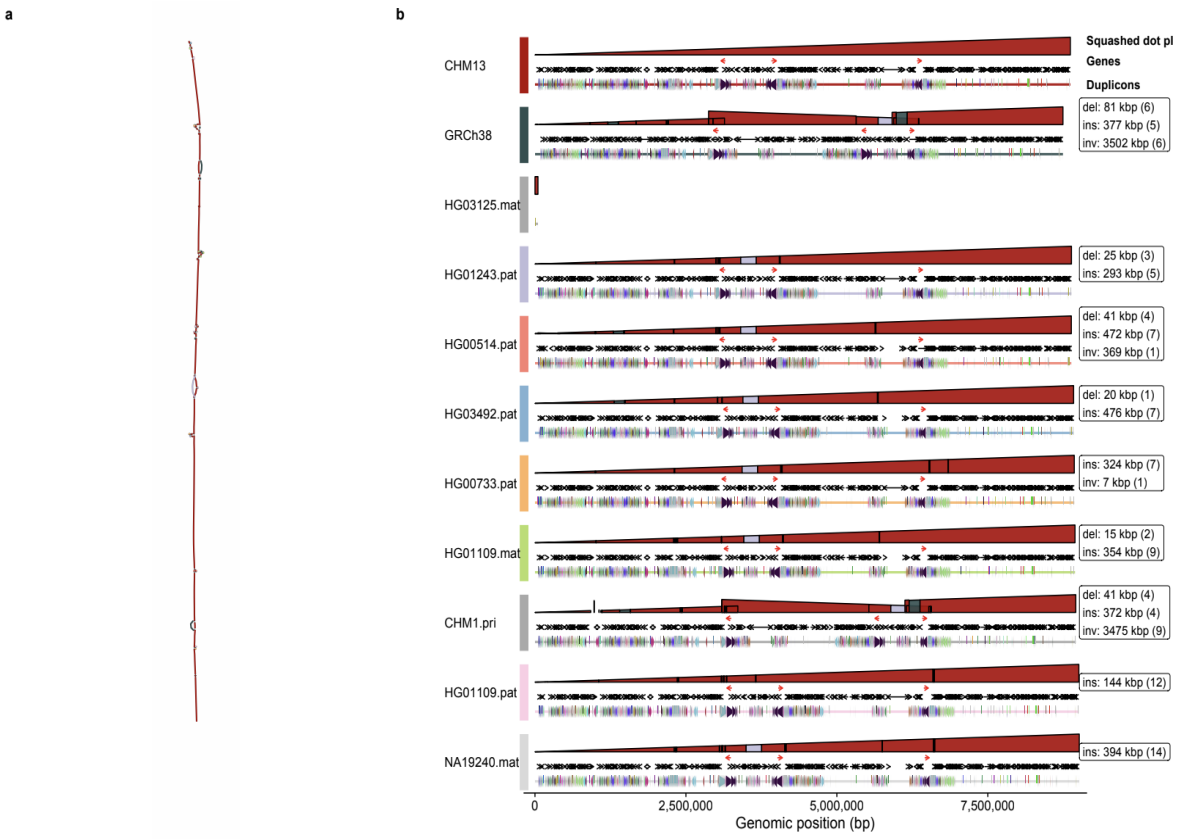


Figure S7. Pangenome graph of *NOTCH2NL* and *SRGAP2*. a) Graph representation (rGFA) of the locus where colors indicate the source genome for the sequence. b) The path for each haplotype-resolved assembly through the graph. The "squashed dot plot" represents a vertically compressed dot plot comparing the haplotype-resolved sequence (horizontal) against the graph (vertical). Color represents the source haplotype for the vertical sequence. Structural variants can be identified from discontinuities in height (deletion), changes between colors (insertion), or changes in the direction of the polygon (inversion). Below is shown the gene of interest (*NOTCH2NL*, red arrow) and other genic content in the region (black arrow). The final line is a duplicon track, showing the ancestral duplications (color) that make up the larger duplication block.

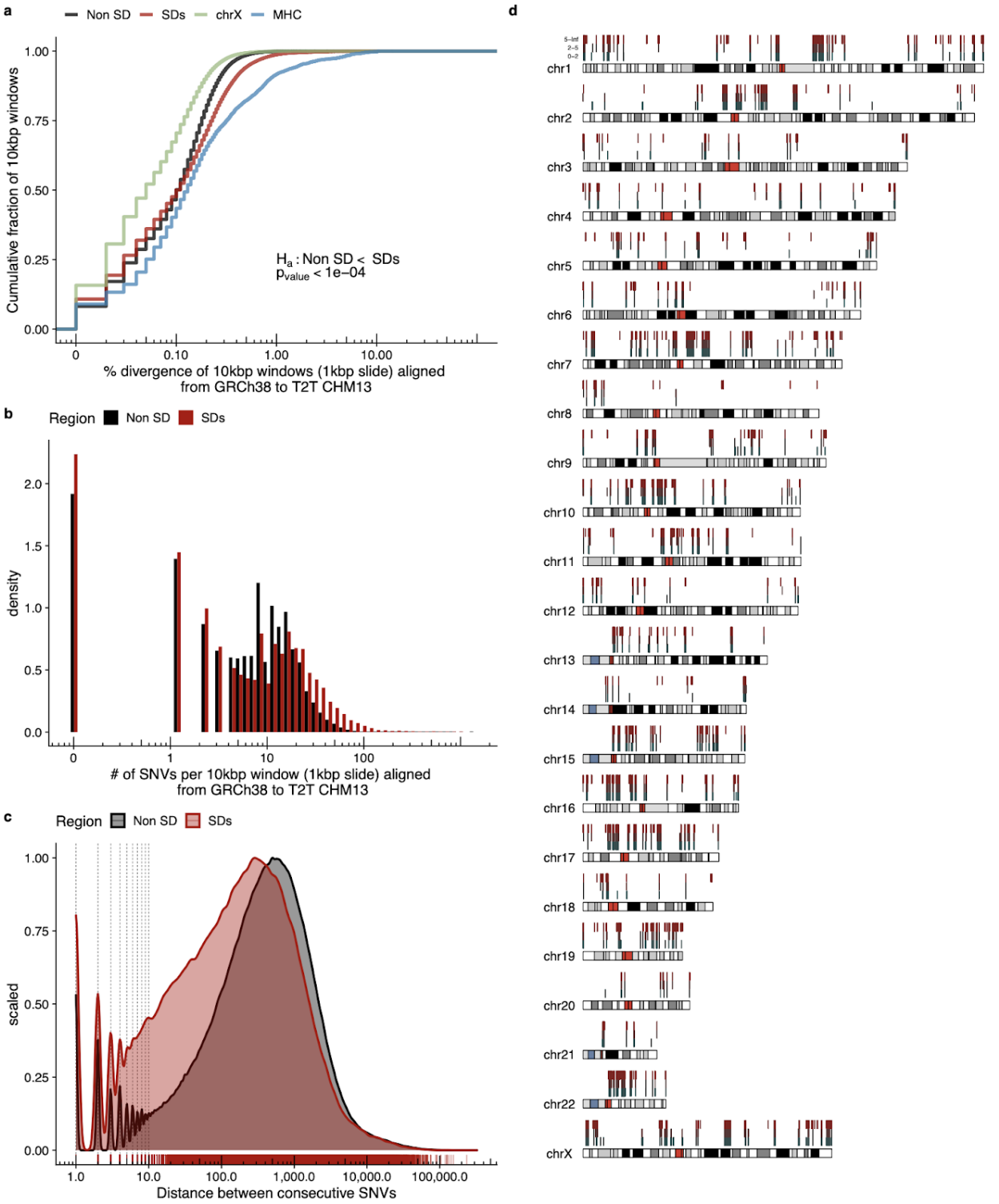


Figure S8. Single-nucleotide variation in SDs between T2T CHM13 and GRCh38. a) Divergence of 10 kbp windows with synteny between GRCh38 and CHM13 T2T. b) Distribution of the number of SNVs per 10 kbp windows aligned from GRCh38 to T2T CHM13 in unique and SD regions. c) Distribution of the distance between SNVs in the syntenic regions of GRCh38 and T2T CHM13. d) SD regions with synteny between T2T

CHM13 and GRCh38 and their average levels of single-nucleotide variation in 1 kbp windows. The bottom row has windows of SD with 0-2 SNVs per kbp, middle row 2-5 SNVs per kbp, and top row is greater than 5 SNVs per kbp.

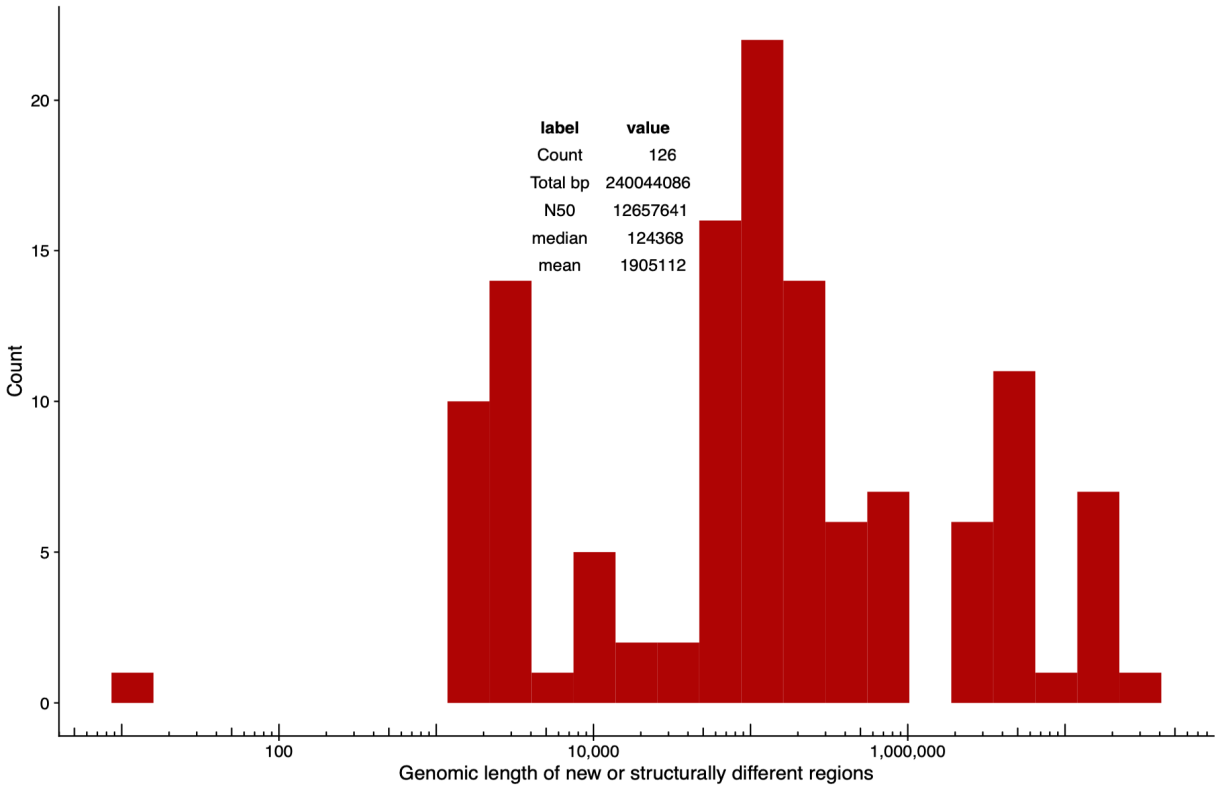


Figure S9. Size distribution of non-syntenic regions between GRCh38 and T2T CHM13. Histogram showing the size of non-syntenic regions (Methods) between GRCh38 and T2T CHM13, and a table of statistics on the lengths of the region.

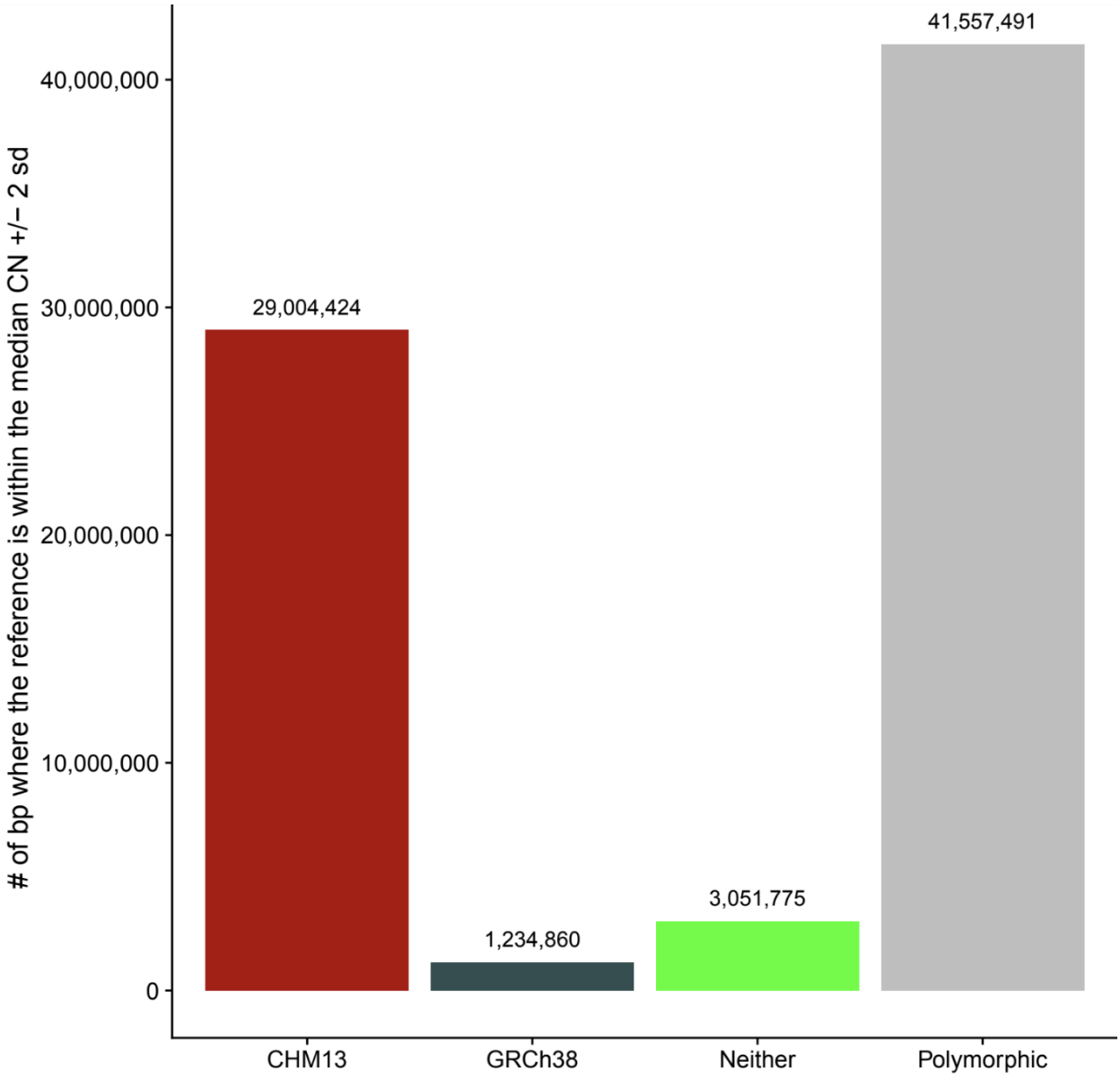


Figure S10. Non-syntenic regions where the reference copy number reflects SGDP. Copy number (CN) of SD regions that are new or structurally different in T2T CHM13 compared to GRCh38 and 268 individuals from the SGDP. The histogram shows the number of Mbp where the median sample CN from SGDP was within 2 standard deviations (sd) of the given assembly [T2T CHM13 (red), GRCh38 (blue), neither (green), or both (equal CN)].

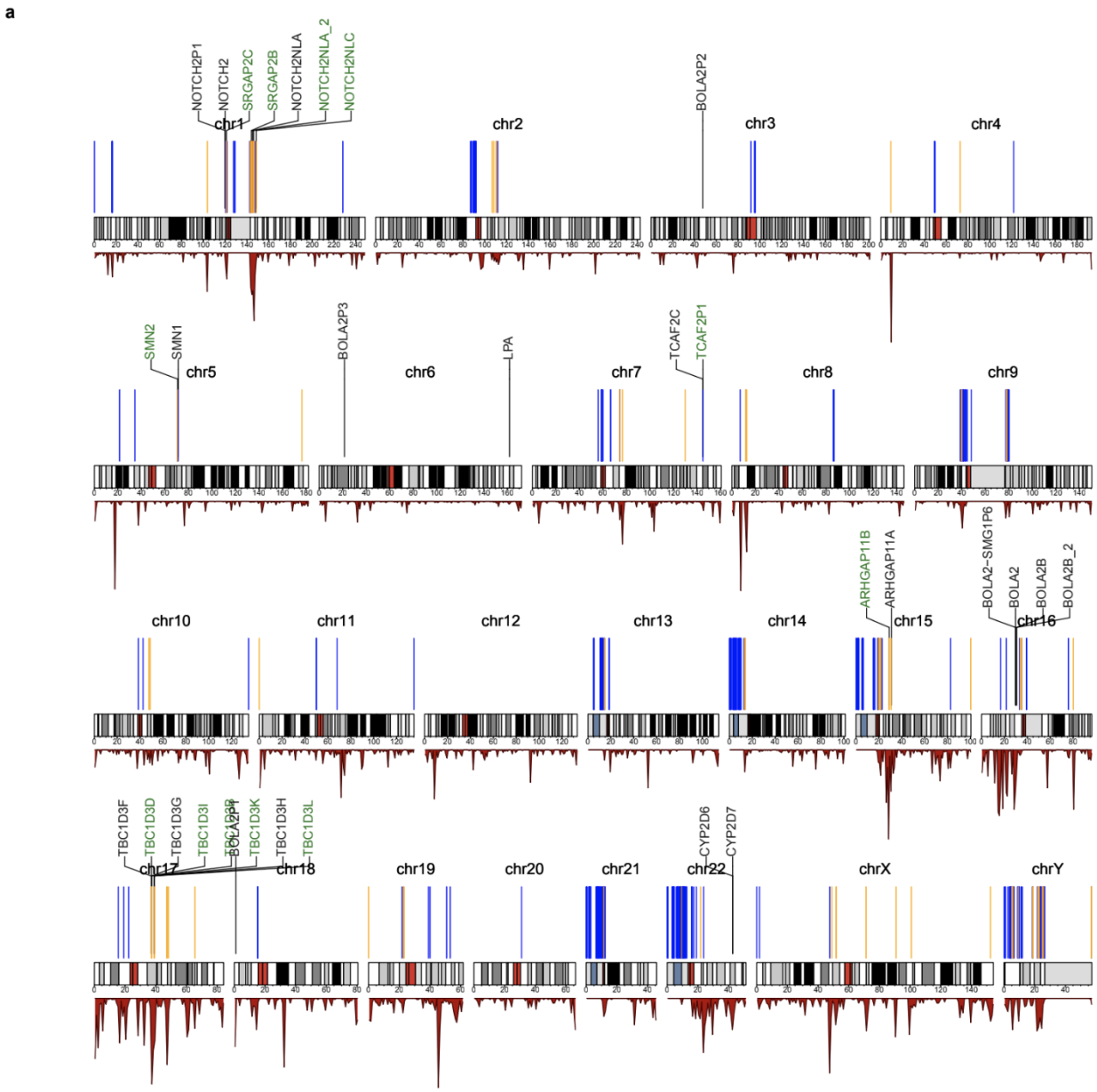


Figure S11. Genic SD expansions in T2T CHM13 relative to chimpanzee. The blue (no genes) and orange (containing genes) peaks in the ideogram show regions of expansion in CHM13 relative to the Clint_PTR assembly within SD space. The bottom panel shows the density of genic SDs in T2T CHM13. The genes highlighted as biomedicaly or evolutionarily important loci are labeled and colored green if they are part of a human expansion.

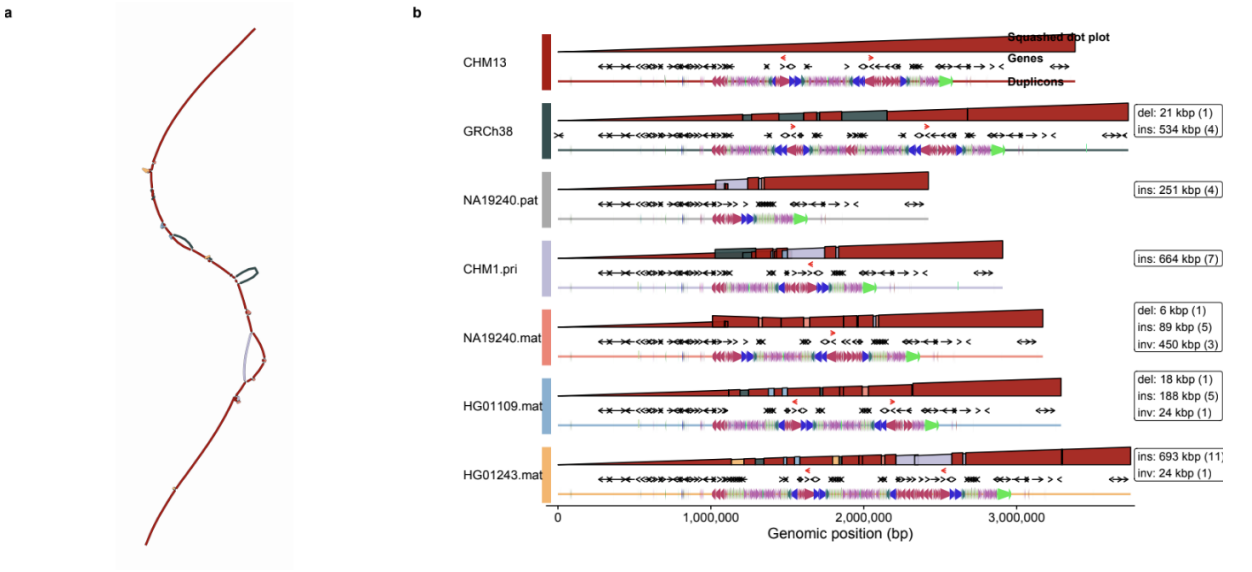


Figure S12. Pangenome graph of *SMN*. For a description of the elements within this figure, see Figure S7.

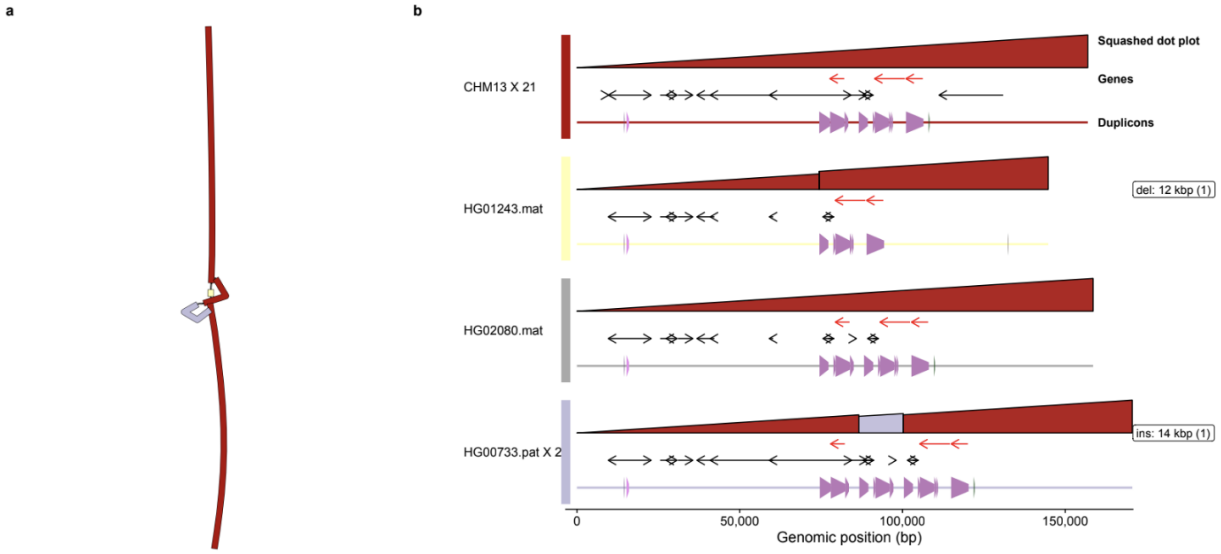


Figure S13. Pangenome graph of *CYP2D6*. For a description of the elements within this figure, see Figure S7.

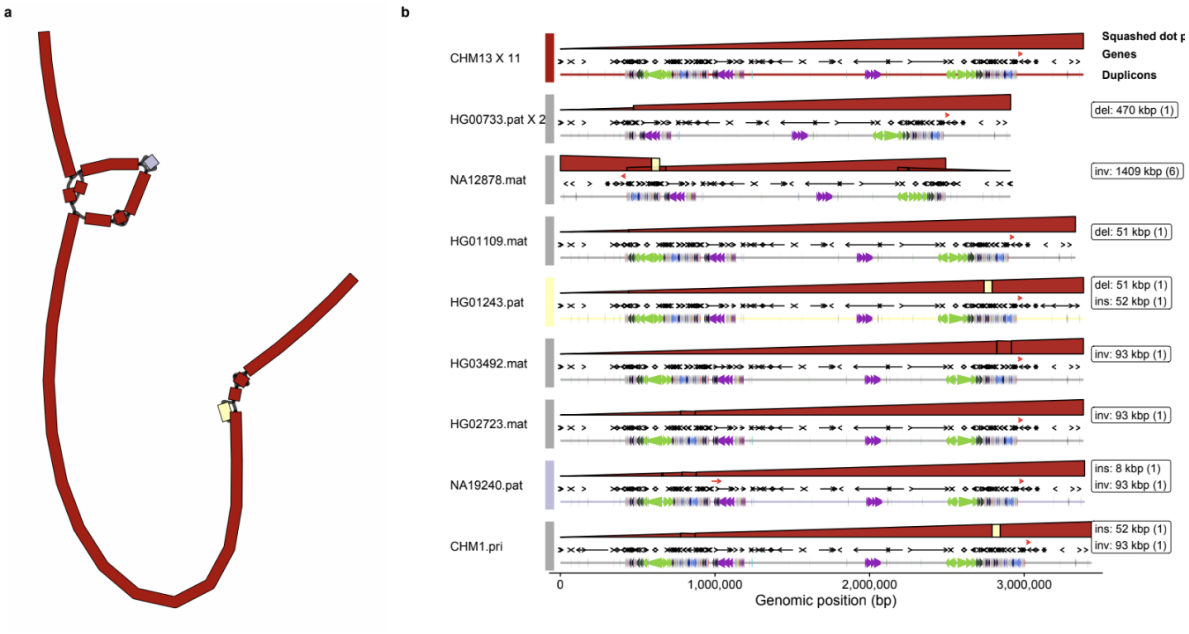


Figure S14. Pangenome graph of *ARHGAP11*. For a description of the elements within this figure, see Figure S7.

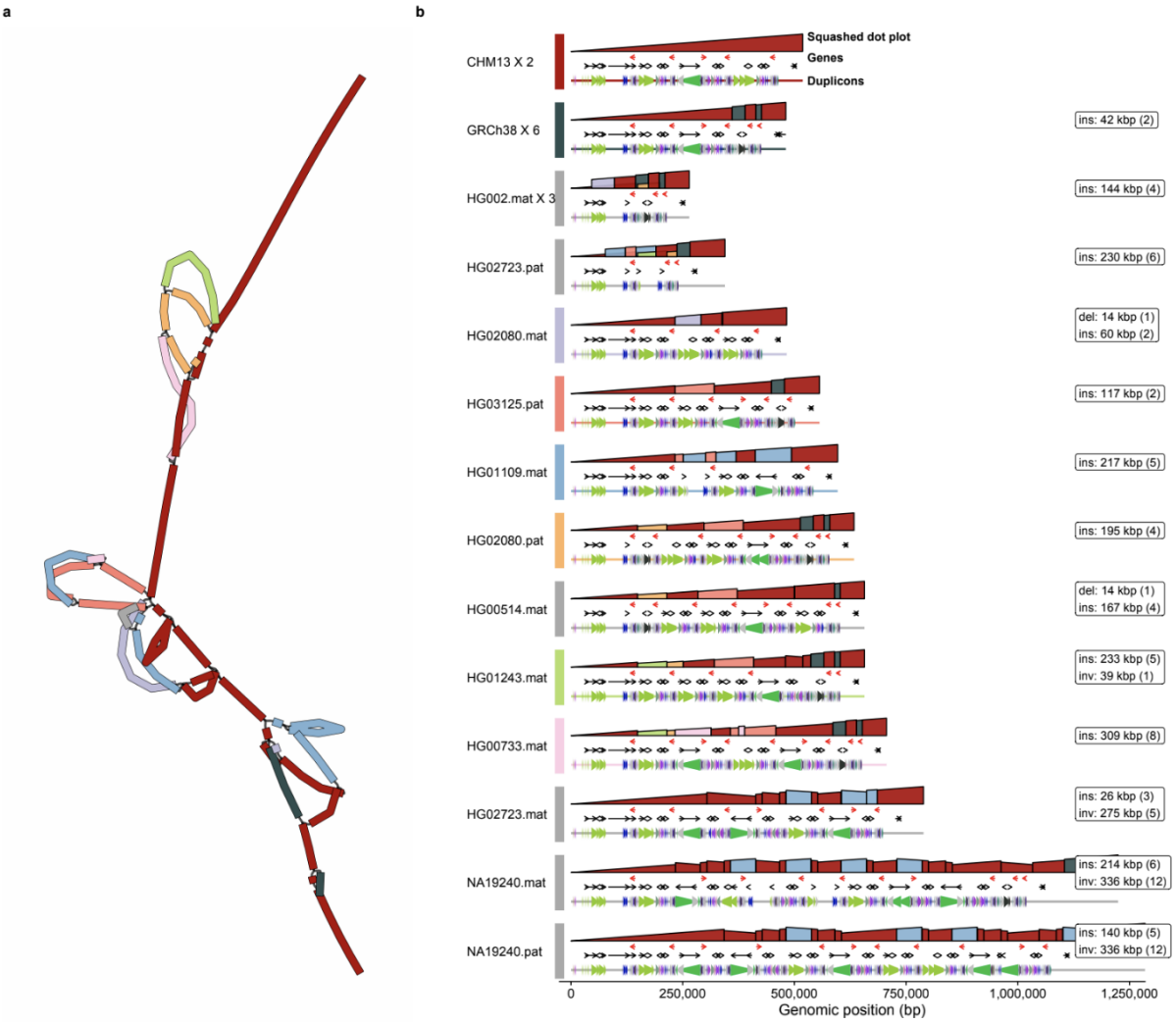


Figure S15. Pangenome graph of *TBC1D3* expansion site one. For a description of the elements within this figure, see Figure S7.

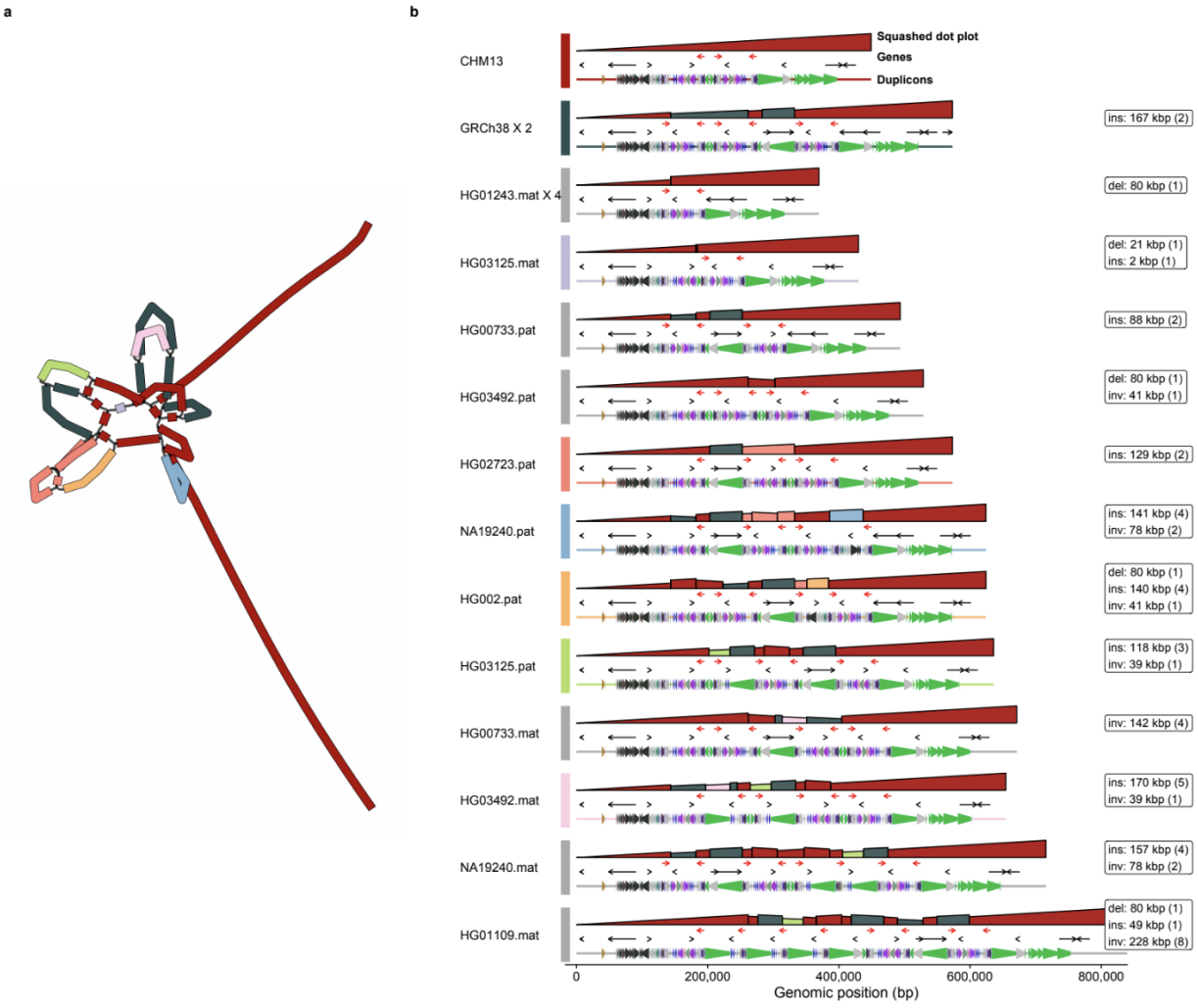


Figure S16. Pangenome graph of *TBC1D3* expansion site two. For a description of the elements within this figure, see Figure S7.

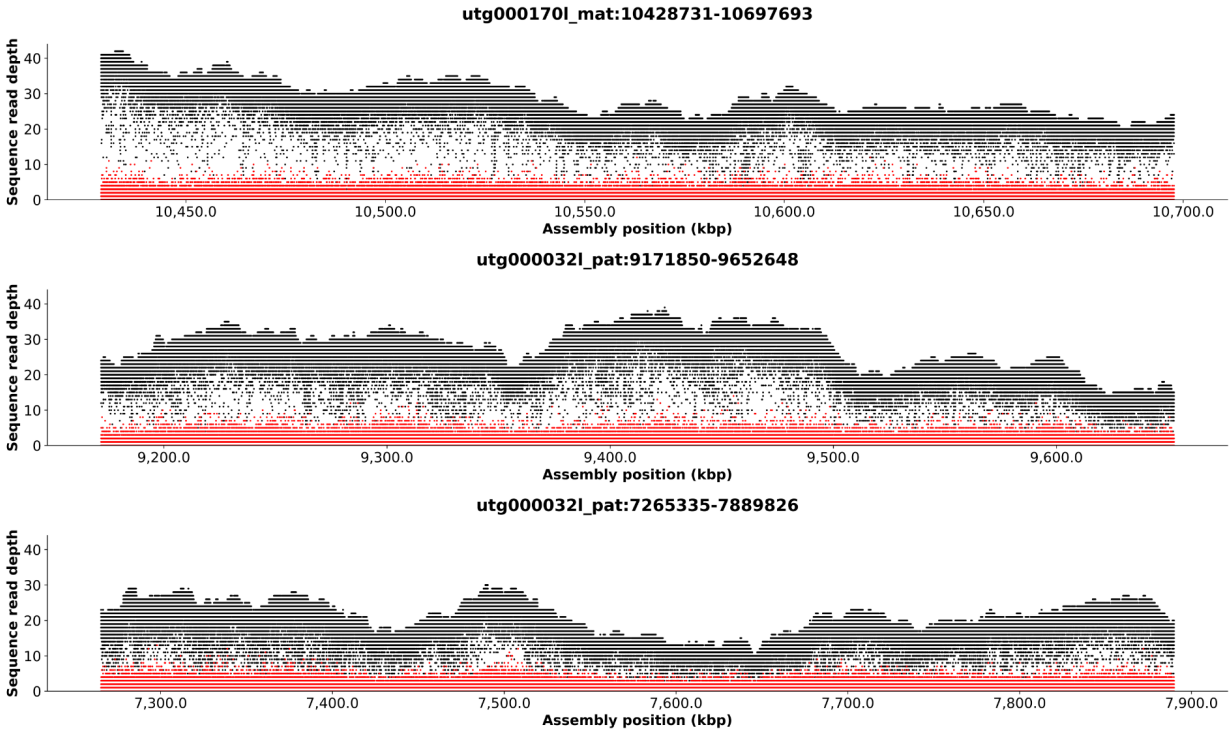


Figure S17. Validation of assembly using ONT coverage over *TBC1D3* for HG002. Ultra-long ONT coverage of HG002 across the maternal haplotype of *TBC1D3* expansion site one, and the coverage across the maternal and paternal haplotypes of *TBC1D3* expansion site two. Black dots show the coverage of the most frequent base at each genomic position and red dots show the coverage of the second most frequent base.

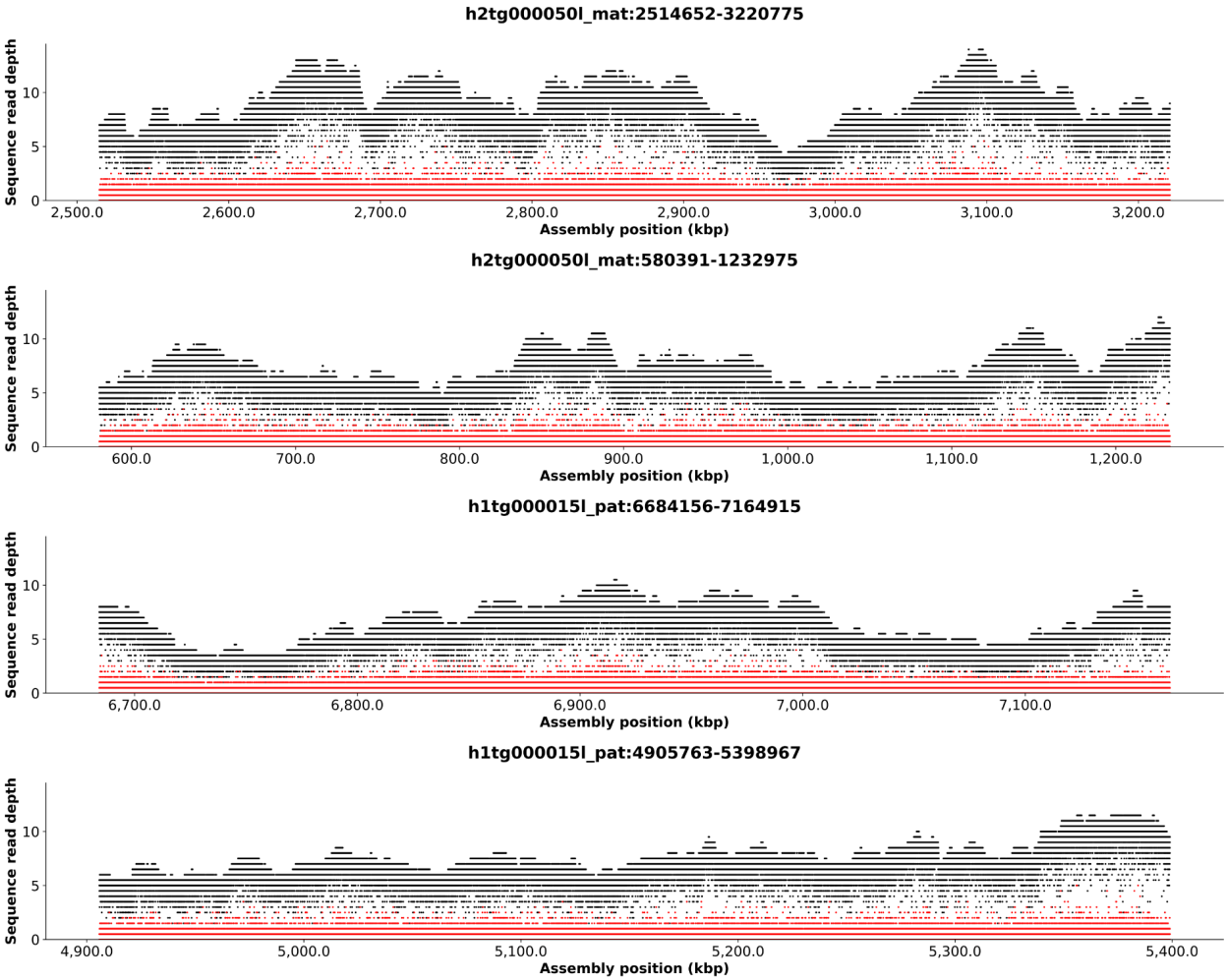


Figure S18. Validation of assembly using ONT coverage over *TBC1D3* for HG00733. Ultra-long ONT coverage of HG00733 across the maternal and paternal haplotypes of *TBC1D3* expansion site one, and the coverage across the maternal and paternal haplotypes of *TBC1D3* expansion site two. Black dots show the coverage of the most frequent base at each genomic position and red dots show the coverage of the second most frequent base.

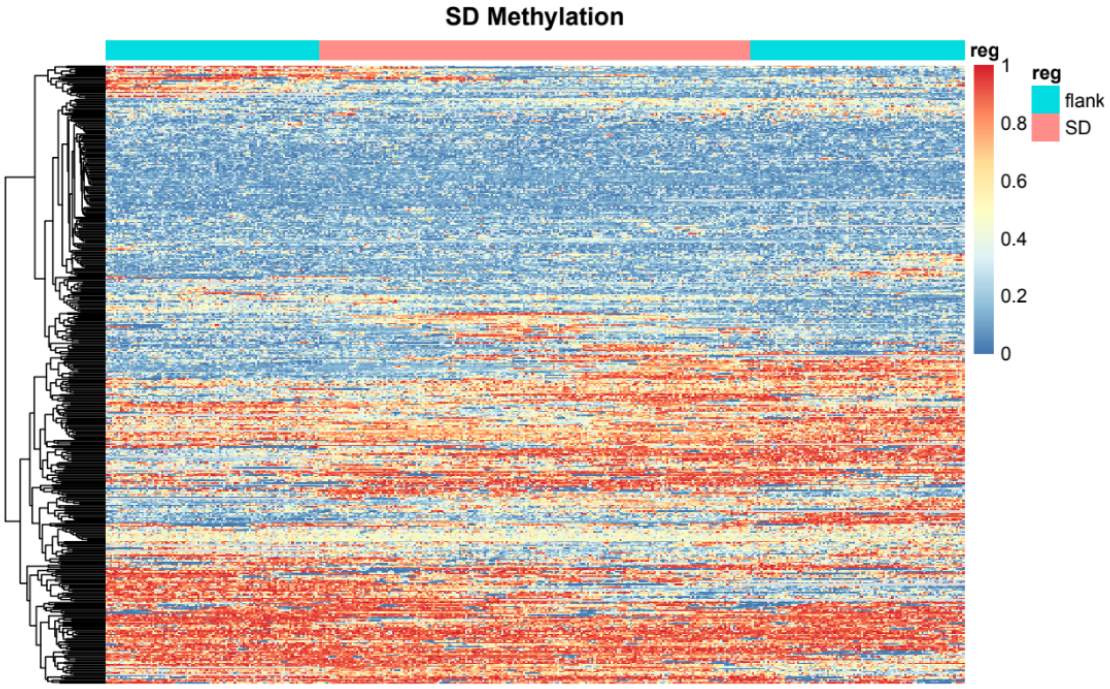


Figure S19. Clustering of methylation status in SD blocks. Heatmap of CpG methylation of all SD blocks with at least 50 kbp of flanking sequence clustered using the "pheatmap" package in R. The horizontal annotation shows in cyan the 50 kbp of unique flanking sequence and red the SD block.

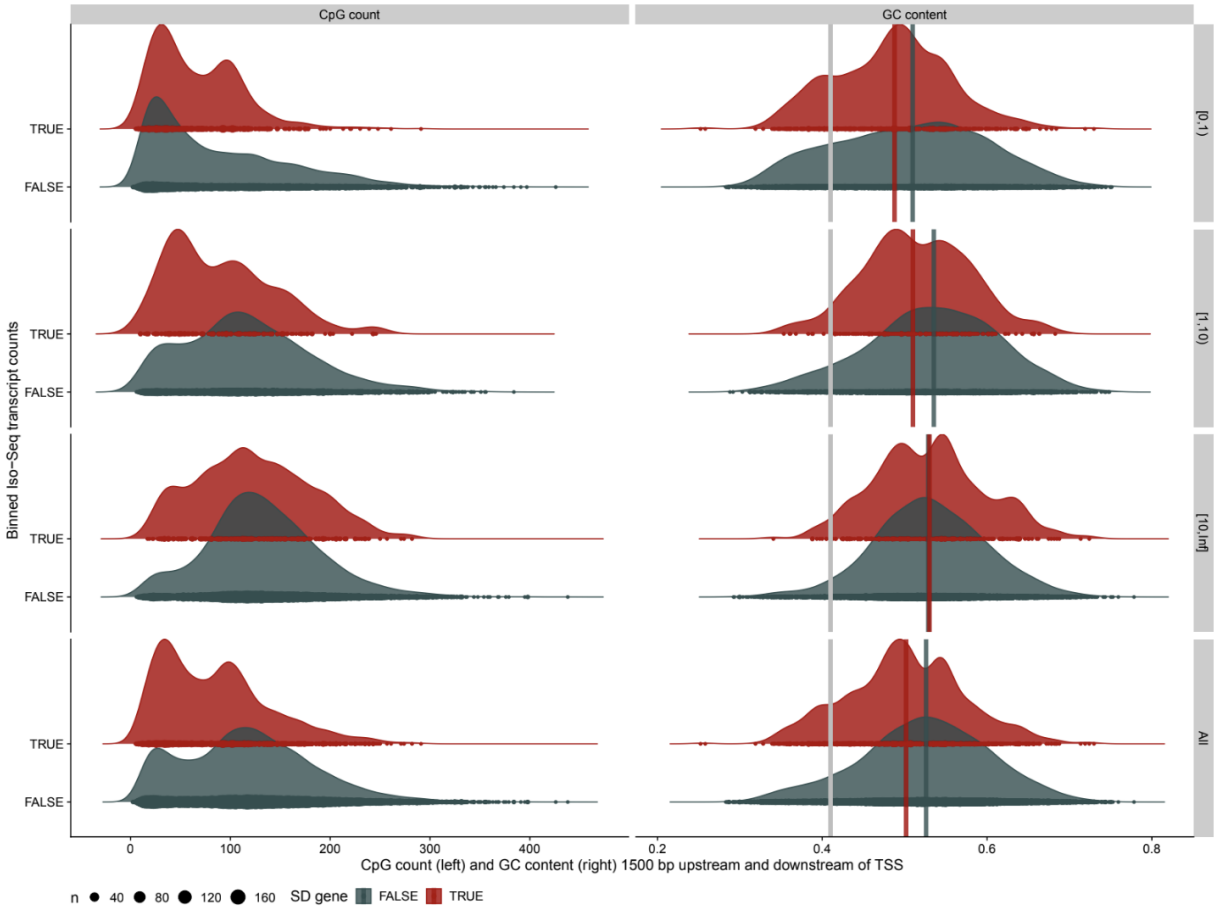


Figure S20. CpG count and GC content within 1,500 bp of TSS. Shown are the density of the number of CpGs within +/-1,500 bp of the TTS (left), and the density of bases that are G or C within +/-1,500 bp of the TTS (right), both stratified by the level of Iso-Seq transcription (vertical positioning) and SD content (color).

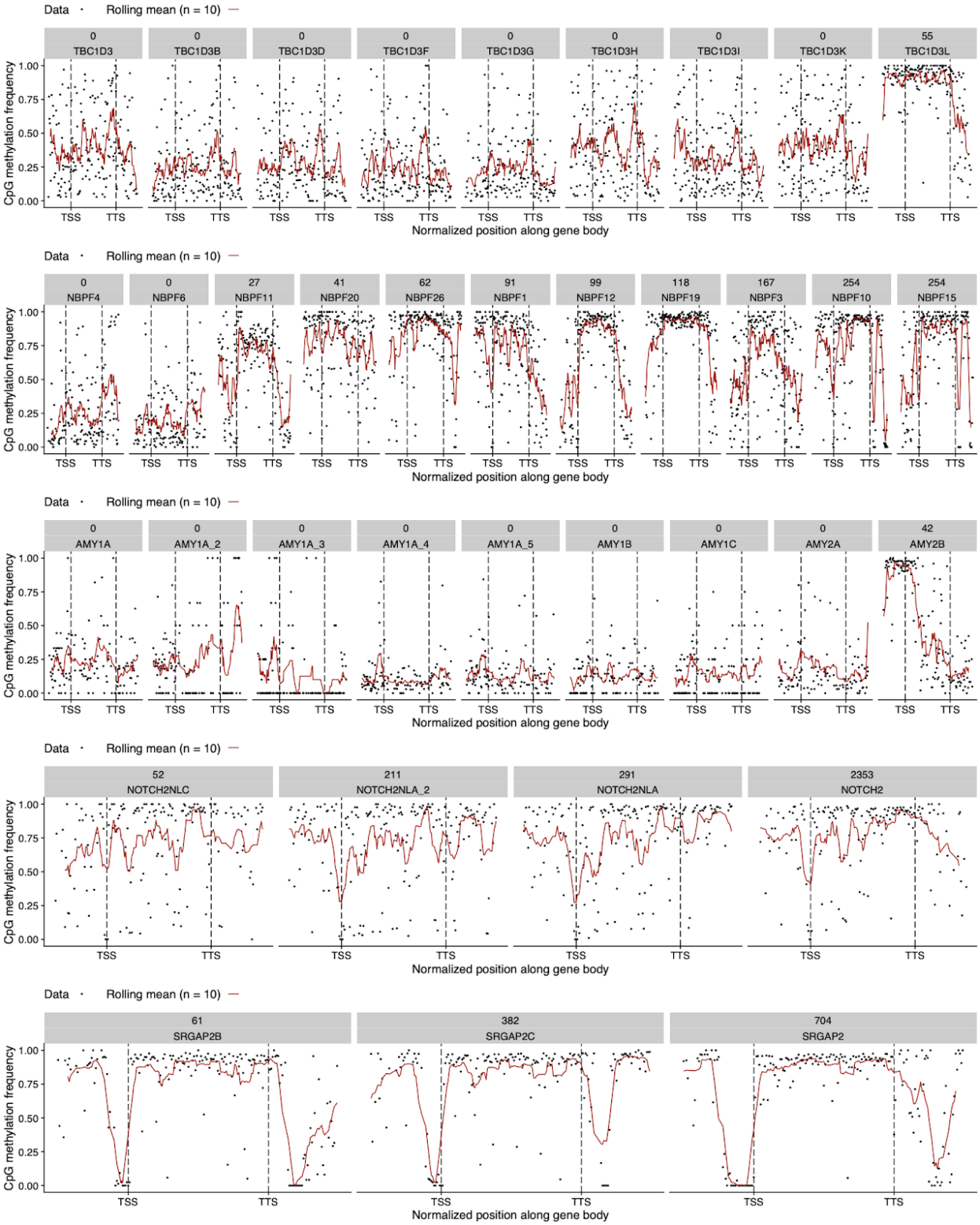


Figure S21. Methylation and transcription levels across multi-copy gene families

C.2 Supplemental Tables

Supplementary tables for chapter four were too large to include here so they are included as a zipped file.

VITA

Mitchell R. Vollger was born in 1992, in Carson City, NV, where he spent 11 years before moving to Elko, NV, and then Eureka, CA. After attending Academy of the Redwoods, Mitchell went to community college at College of the Redwoods and then on to study at Princeton University. In 2015 he graduated from Princeton with B.S.E. in computer science and a minor in quantitative and computational biology. During his time at Princeton, he worked in the lab of Alison Gammie improving the reference genome for the yeast strain W303. In 2016 Mitchell joined the Department of Genome Sciences at the University of Washington in pursuit of a PhD, ultimately joining the lab of Evan E. Eichler to study the sequence and assembly of segmental duplications in humans and nonhuman primates. Upon completion of his PhD, Mitchell plans to remain at Genome Sciences as a postdoctoral fellow while searching for a new fellowship position.