Improving Allele Calling in Highly Polymorphic and Repetitive Genomic Regions

by

Heather Marie Gibling

A thesis submitted in conformity with the requirements
for the degree of Doctor of Philosophy

Department of Molecular Genetics
University of Toronto

Improving Allele Calling in Highly Polymorphic and Repetitive Genomic Regions

Heather Marie Gibling

Doctor of Philosophy

Department of Molecular Genetics
University of Toronto

2023

Abstract

Many regions of the human genome are highly polymorphic, containing numerous variants not found in the current reference genome and making it difficult for variant calling software to accurately identify variants. The effects are particularly pronounced when analyzing structural variants and repetitive regions with short-read sequencing owing to unmapped reads. Long-read sequencing greatly improves structural variant identification, but there is an abundance of data already available generated from short-read technology. Adapting current variant calling approaches for problematic regions of the genome can provide novel insights about variation from existing datasets.

This thesis focuses on developing and assessing methods to improve variant calling for polymorphic repetitive genomic regions using the highly variable minisatellite-like zinc finger repeat array of the gene *PRDM9* as a proof-of-concept region. I developed short-read models that utilize *k*-mer information to distinguish between the numerous *PRDM9* alleles that differ by single nucleotide polymorphisms and copy number variants of zinc finger repeats. These models were able to call haploid and diploid *PRDM9* genotypes with average F1 scores of up to 0.99 and 0.73, respectively, using simulated sequencing reads under typical conditions. Using real Illumina data, the models were able to genotype 43% of 101 samples correctly and

ranked the correct genotype in the top 10 for 69% of the samples. I also developed long-read models that use realignment or consensus sequences to validate the short-read models and identify novel alleles. The models correctly genotyped 98% of 101 samples using PacBio reads and found five novel alleles. Finally, I assessed the use of graph-based reference genomes, which are flexible data structures that incorporate known variants to improve polymorphic representation. I constructed different *PRDM9* reference graphs and observed modest improvements in read alignment over mapping to GRCh38, with larger improvements for samples with more divergent *PRDM9* genotypes or with novel alleles.

Overall, my thesis provides insight into how genotyping *PRDM9* can be improved by avoiding reliance on mapping reads to the GRCh38 reference. These approaches—utilizing *k*-mers, long-read sequencing, and graph-based reference genomes—can be further refined and adapted to improve variant identification for additional polymorphic and repetitive genomic regions.

Acknowledgements

First and foremost, thank you to my two supervisors Dr. Philip Awadalla and Dr. Jared Simpson for their endless support and encouragement these past seven years. Being given the opportunity to learn and practice how to science in two supportive lab environments while repeatedly making mistakes has helped me grow as a researcher and as a person. Thank you for sending me to conferences and workshops, and for encouraging my teaching side quest. Philip, thank you for your guidance, your insights into the biology side of things, and your dad jokes. Jared, thank you for your insights into the computer science side of things, teaching me good coding practices and pushing me to tackle C++, and sharing pictures of your dog. I am honored to be your first student; you are an incredible mentor and you will only improve with every student you take under your wing. Thank you to my supportive committee members Dr. Quaid Morris and Dr. Trevor Pugh for your advice, encouragement, and outside perspectives over the years.

Thank you to Awadalla lab members past and present: Armande and Elyssa who started and completed this journey with me; Michelle and Jasmina for your friendship and support; Kim, Nick, Ido, and Tom for your lively conversations and feedback; postdocs Hilary, Fabien, Isabel, and Dave for your wisdom and advice; Elias for your kind words and belief in my potential; and Vanessa, Mawussé, and Marie-Julie for your help and expertise. Likewise, thank you to Simpson lab members Alister, Joanna, Paul, Sabiq, and Jonathan for shenanigans; Richard for rants and coffee breaks; Mike for your banter; and Phil and Ina for your expertise in all things wet lab. Both labs have been wonderful places to learn amongst great people. Thank you to Cindy for always stopping to say hi while keeping me sane and organized, and to Polina for your friendly assistance as I wrap up. Thank you to the OICR lunch crew for making work so enjoyable.

Thank you to the incredible educational scientists who have supported me in my exploration into teaching: Dr. Johanna Carroll and Dr. Martina Steiner for opportunities with TAing and developing course material, and Dr. Michelle Brazas for opportunities with Bioinformatics.ca and TorBUG. Thank you to Dr. Greg Wilson for getting me interested in teaching in the first place through Software Carpentry and for providing advice over the years. Teaching has been a welcome distraction and avenue for professional development during my PhD.

Thank you to the scientists who gave me a chance when I did not have the experience or external funding: Dr. Guang Sun at Memorial University for hiring me as a research assistant for my first research job, and Dr. Carrie Shemanko at the University of Calgary for letting me join your lab and use your graduate students to learn wet lab skills even though I only got a C- in your course. These experiences were pivotal in helping me focus and improve as I finished my bachelor's degree and discovered my love of research. Thank you to Dr. David Mutch at the University of Guelph for guiding me during and after my master's degree and for

encouraging me during the switch from wet lab to dry. An honorable mention to MoGen for rejecting me the first time I applied to this program; this led me down the path of bioinformatics, which as it turns out is far more enjoyable than bench research. Thank you to Dr. Gary Bader for letting me geek out over the ONT MinION during recruitment day the second time I applied to this program and for encouraging me to pursue fully computational research when I did not think I had the skills to do so.

Thank you to my MoGen friends Amanda, Ellen, and Tim for years of memories at grad socials, trivia nights, Jackbox drawing games, and shared PhD-or-bust support. Thank you to my long-distance friends in Calgary, Guelph, and beyond who have been my cheerleaders for so long. To Kimber for your friendship that has not ceased since grade three, for appreciating the way I drew cows sitting down, for our silly reunions, your Toronto visits, your humor, and your endless support when I need it most. To Amanda for teaching me research skills in Calgary, singing Disney karaoke in Ottawa, and for your support and sharing your experiences about this rollercoaster called grad school.

Thank you to my parents for supporting me every year I have been in school (I promise grade 27 is the last), for providing a safe haven in St. John's for summer and Christmas escapes, and for your constant love and encouragement. Thank you to my Gibling Siblings Leah and Colin for your thoughtful and humorous encouragement, and to my new brother-in-law Hamza for delicious home-cooked meals. Thank you to my extended family for always cheering me on.

Thank you especially to my little family that I have curated in Toronto. Thank you to Rob for being my rock, my hype man, my counselor, my partner, and my friend. The last three years have been a rollercoaster to say the least, but I am so thankful to have had you by my side for the ride. I cannot wait to see what is next for us. Thank you to my little buddy Jinx for your soft purrs and head butts and for not deleting too much code while walking on my keyboard.

Finally, at the risk of sounding smug, thank you to myself for making it this far. Seven years is a long time; long enough for nine committee meetings, three topic courses, one graduate CompSci course, two departmental seminar talks, attending six conferences, two new anxiety medications, three panic attacks, nine different mental health counselors, one postponed reclassification exam, the loss of my teenagehood cat Patrick, a global pandemic, two stolen bikes, living in three different apartments, four cell phones, attending 9.5 weddings, the release of 18 Marvel movies, 17 blood donations, two new Adele albums, attending 22 concerts, and 21 roundtrip plane rides, to name a few. It has been an interesting journey.

But I hear the music, I feel the beat
And for a moment when I'm dancing
I am free.

—Florence Welch

Table of Contents

List of Tables

List of Figures

List of Appendices

List of Abbreviations

1000GP	1000 Genomes Project
1000GP-SV	1000GP Structural Variant
ASE	Allele-Specific Expression
BAC	Bacterial Artificial Chromosome
bam	Binary Alignment/Map
CCS	Circular Consensus Sequencing
CLR	Continuous Long Reads
CNV	Copy Number Variant
CYP	Cytochrome P450
DNA	Deoxyribonucleic Acid
GA4GH	Global Alliance for Genomics and Health
gaf	Graphical Alignment Format
gam	Graphical Alignment/Map
gfa	Graphical Fragment Assembly
GIAB	Genome in a Bottle
GRCh35/37/38	Genome Reference Consortium human (genome build) 35/37/38
hg19	Human genome (reference) 19
HGP	Human Genome Project
HiFi	High Fidelity
HLA	Human Leukocyte Antigen
HMM	Hidden Markov Model
HPRC	Human Pangenome Reference Consortium
HPRC++	HPRC + 1000GP-SV + GIAB
IGV	Integrative Genomics Viewer
indel	Insertion or Deletion
IUPAC	International Union of Pure and Applied Chemistry
KIR	Killer Ig-Like Receptors
LINE	Long Interspersed Nuclear Element
MHC	Major Histocompatibility Complex
nCATS	Nanopore Cas9-Targeted Sequencing
NCBI	National Center for Biotechnology Information
NDJ	Nondisjunction
NK	Natural Killer
OHS	Ontario Health Study
OICR	Ontario Institute for Cancer Research
ONT	Oxford Nanopore Technologies
PacBio	Pacific BioSciences
PCAWG	Pan-Cancer Analysis of Whole Genomes
PCR	Polymerase Chain Reaction
POA	Partial-Order Alignment
PRDM9	PR/SET Domain 9
RNA	Ribonucleic Acid
SINE	Short Interspersed Nuclear Element
SNP/SNV	Single Nucleotide Polymorphism/Variant
STR	Short Tandem Repeat
SV	Structural Variant
T2T	Telomere-to-Telomere
TE	Transposable Element
UV	Ultraviolet
vcf	Variant Call Format
ZMW	Zero-Mode Waveguide

Chapter 1

Introduction

1.1 The human reference genome and genomic variance

The completion of the Human Genome Project (HGP) in 2004 produced a reference for what the human genome encompasses and how it is organized (International Human Genome Sequencing Consortium 2004). Using DNA from predominantly nine anonymous individuals, large DNA segments of 100–200kb were inserted into bacterial artificial chromosomes (BACs) which were shotgun Sanger sequenced and assembled into overlapping contiguous segments. These **contigs** were assigned to chromosomes using existing genetic maps and assembled together based on overlapping sequences. Whole-genome shotgun sequencing was performed using DNA from multiple additional samples to assist with assembly (Lander et al. 2001). Finishing the draft genome involved the process of closing numerous gaps by experimental and computational means, as well as eliminating duplication and assembly artifacts (International Human Genome Sequencing Consortium 2004). The resulting genome, build 35, has been continuously updated with corrections, gap closures, and additional variants stored as alternate contigs. The current build, **GRCh38**, represents 95% of the human genome (Schneider et al. 2017) with largely heterochromatic regions that remain absent.

Despite its widespread use—including by the Celera Genomics company for generating a whole human genome sequence in competition with the HGP (Venter et al. 2001)—there are criticisms and concerns about the data present in the current reference genome. GRCh38 is not the genome of a single person, it does not have equal representation across multiple populations, it does not represent alleles with the highest frequency, it does not represent ancestral alleles, it does not represent the least damaging variants, and it does not contain the longest structural variants; it is neither a consensus nor a canonical reference (Schneider et al. 2017). Instead, it is a mosaic of individuals and sequences and is reflective of the enhancements over the years to the original genome sequence compiled by the HGP. Approximately 70% of the sequence came from one man of admixed African and European ancestry; 23% is from 10 individuals, including six Europeans and one East Asian; and 7% from more than 50

individuals and cell lines (https://www.ncbi.nlm.nih.gov/grc/help/faq). The resulting reference is a haploid genome of approximately 57% European, 37% African, and 6% East Asian ancestry (Green et al. 2010; Miga and Wang 2021). In an effort to increase the representation of known variants, The Genome Reference Consortium (GRC) included 261 **alternate haplotypes** for 178 polymorphic regions throughout the GRCh38 build in addition to the primary reference sequences (Church et al. 2015).

Nevertheless, the reference genome is the predominant source of comparison for newly sequenced genomes in order to identify regions where variation exists. Researchers and clinicians use GRCh37 or GRCh38 as the basis for all genomic coordinates and annotations; genes, regulatory features, and variants are all assigned precise genomic locations based on their position in the reference genome (Schneider et al. 2017).

1.1.1 Variation in the human genome

While an estimated 99.9% of the genome is shared across humankind, the small percentage of variants can have a profound impact on phenotypes and diseases. There are many types of variants, ranging in size from a single base to several thousand bases long, that arise from mutations in genomes. The completion of the HGP has allowed for genome-wide research into these variants and has shed light into how large structural variants affect megabases of DNA.

An astonishing half of the human genome is composed of repetitive sequences (Lander et al. 2001). Much of these repeats originate from **transposable elements** (TEs), which are sections of DNA that can relocate within a genome. Though many TEs stopped mobilization 37 million years ago (Pace and Feschotte 2007), variation in **repetitive sequences** exists because some are still active in humans, such as long and short interspersed nuclear elements (LINEs and SINEs). Germline insertions of the SINE Alu and the LINE L1 are estimated to occur as frequently as one in 21 births (Xing et al. 2009) and one in 95 births (Ewing and Kazazian 2010), respectively. **Structural variants** (SVs) encompass long stretches of DNA, generally longer than 50bp, and include large insertions and deletions, segmental duplications, copy number variants (CNVs), translocations, and inversions (Mahmoud et al. 2019). SVs arise during the processes of DNA recombination, replication, and repair (Carvalho and Lupski

2016). Repetition in the human genome contributes to errors in these processes. For example, homologous recombination can occur between TEs that are dispersed far apart due to the high homology between the repetitive sequences. Many CNVs are in fact formed by such **nonallelic homologous recombination** (Stankiewicz and Lupski 2002). SVs are implicated in many human disorders owing to the losses, gains, or truncations of important genes or regulatory regions that may occur in affected regions (Lupski 1998; Weischenfeldt et al. 2013), although not all SVs are problematic (Weischenfeldt et al. 2013).

Single nucleotide polymorphisms or **variants** (SNPs or SNVs) are variants of a single DNA base. SNPs can be bi-allelic, with two known variants, or multi-allelic, with three or four variants. SNPs are always germline variants, but SNVs can be somatic. **Indels**, short insertions and deletions generally less than 50bp, are sequences absent from the reference but present in a genome, or missing from a genome but present in the reference, respectively. While also contributing to complex traits and diseases, SNPs and indels are frequently responsible for many classic Mendelian disorders, where a change in a single nucleotide can cause a missense or nonsense mutation that detrimentally affects proper gene translation or protein formation, or where a mutation has negative effects in a regulatory region (Thomas and Kejariwal 2004). Mutations can originate from environmental effects such as UV radiation or from mistakes in DNA replication and repair. Damage to nucleotides can cause erroneous base pairing, resulting in a nucleotide change during the next round of replication in place of the damaged base. DNA polymerase sometimes slips during replication due to the presence of small repeat sequences known as **short tandem repeats** (STRs), increasing the number of repeats and contributing to instability of these **microsatellite** regions. Though there are molecular mechanisms in place to repair damaged DNA, these processes are not always sufficient, leading to the incorporation of mutations in the genome. Nonsynonymous mutations can be under positive or negative selection, while synonymous mutations are generally under the influence of genetic drift, leading to the eventual loss or fixation of the variant in the genome.

The 1000 Genomes Project (1000GP), a large-scale consortium that sequenced and analyzed over 2,500 individuals from 26 populations across five continents, has been an invaluable source of information about variants in the human genome (The 1000 Genomes Project Consortium 2015). The project revealed that the average human genome generally differs at

4.1–5 million locations relative to the reference genome, and the vast majority of these differences are SNPs and indels. Additionally, the typical genome contains 2,100–2,500 SVs that affect nearly 20Mb of sequence. The collection of 1000GP variants has been used to test and validate methodology and software for calling variants from genomic data, though the project has been superseded by larger-scale collaborations such as gnomAD (Karczewski et al. 2020). There are many approaches that can be taken to identify existing variants as well as to discover novel variants not yet described.

1.1.2 Identifying variants

Though variant detection has been progressing since the early years of genetic research, the advancement and affordability of modern high-throughput sequencing technology has greatly increased variant discovery, improving our understanding of genomic structure and contributions to genetic diseases. The modern process of identifying variants relies on quality sequencing data, assembly of or alignment to a genome, and software to identify differences relative to the human reference genome.

1.1.2.1 Sequencing technologies

High-throughput **Sanger** sequencing employed by the HGP is still used to obtain highly accurate reads of targeted sequences. Due to the tedious and time consuming nature of the preparation and sequencing process, it is not feasible for genome-wide analyses. Instead, next generation and third generation sequencing technology delivers high-throughput generation of quality genomic data with continually decreasing costs. Illumina short-read sequencing is currently the industry standard for accurate genome-wide sequencing studies, but Pacific Biosciences and Oxford Nanopore Technologies long-read sequencing are becoming more popular as the technologies improve and costs decrease.

Illumina sequencing employs a sequencing-by-synthesis methodology. Briefly, DNA is broken into fragments, sequencing adapters are attached to the ends that allow each strand of the fragment to bind to flow cells, and DNA is clonally amplified to enrich the sequencing signal. Fluorescently tagged nucleotides that are blocked at the 3’ end are released and added to the sequencing primers bound to the DNA fragments. After image detection of the newly

incorporated fluorescent bases, the nucleotides are unblocked to allow for sequencing to continue in a controlled base-by-base manner, allowing for accurate base calling over many cycles (Goodwin et al. 2016). Illumina reads are generally 100–250bp long; longer reads are not viable due to the accumulation of sequencing errors in individual read clones as the cycles progress that affect read accuracy. The paired-end nature of sequencing DNA fragments from both ends provides contextual information about the expected distance between a pair of reads, given an estimated length distribution of the starting fragments. Illumina sequencing accuracy is very high—around 99.9%—with a bias towards incorporating mismatch base errors, particularly at the ends of the reads.

Pacific Biosciences (PacBio) performs single-molecule sequencing using zero-mode waveguides (ZMWs) that restrict the amount of light that penetrate the bottoms of wells on a flow cell. The ZMWs allow for the capture of single nucleotides at a time as fluorescently labeled nucleotides are incorporated to DNA fragments by polymerases fixed to the bottom of the wells. Cameras capture the fluorophores before they are cleaved away for the next sequencing cycle (Goodwin et al. 2016). The original approach, continuous long read (CLR) sequencing, processed each DNA fragment once and had an accuracy of around 90% (Rhoads and Au 2015). The newer **circular consensus sequencing** (CCS; also known as high-fidelity or HiFi) approach sequences each DNA fragment multiple times and generates a consensus sequence given the base calls from all of the passes. CCS functions by use of two hairpin adaptors attached to the ends of a double stranded fragment, resulting in a circular single stranded template that can be continuously sequenced. HiFi sequencing has an accuracy of over 99% (Hon et al. 2020). In terms of read length, CLR sequencing can produce very long reads, over 60kb, while the HiFi approach is limited to reads of 10–25kb due to the circular nature of the template strand (Hon et al. 2020). PacBio sequencing is prone to indel errors, but the unbiased distribution of these errors means they can be mitigated with high depth of coverage.

Oxford Nanopore Technologies (ONT) sequencing functions uniquely by performing direct molecule sequencing instead of sequencing by synthesis. Single strands of DNA are pushed through bioengineered nanopores embedded in membranes by attached motor proteins. As DNA moves through the pore, the nucleotides partially block the electrical current passing

through the pore, which is constantly being sampled. At any given time approximately five nucleotides are present in a single nanopore, and the composition of these five bases cause characteristic perturbations to the electrical current. The electrical current signal is transformed into a sequence of predicted nucleotides using neural networks trained on known DNA sequences (Wick et al. 2019). Benefits of ONT sequencing are the extreme read lengths that can be generated, which have reached over four million bases (https://nanoporetech.com/applications/whole-genome-sequencing), and the ability to directly detect base modifications such as 5-methylcytosine (Simpson et al. 2017). Due to the inconsistent speed at which DNA travels through the pore, it is not always possible to detect certain stretches of bases, in particular homopolymers or stretches of a repeated single base. ONT is therefore more prone to indel sequencing errors. Accuracy has traditionally been around 90%, but recent advances in technology have increased it to near-Illumina levels of 99% (Sereika et al. 2022).

In general, short reads are highly accurate but not long enough to directly detect SVs or resolve highly repetitive regions (**Figure 1.1 A**). Long reads, on the other hand, can be long enough to span the full length of a SV or repetitive region while aligning to unique sequences flanking one or both sides to help identify where in the genome the read originated and how many repeats are in a region (**Figure 1.1 B**). Though more costly than using only one sequencing platform for a project, there is a benefit to combining both short- and long-read sequencing: SVs and repetitive regions can be resolved by the long reads while the short reads improve overall accuracy at less complex regions. Using a combination of sequencing technologies has led to the discovery of many SVs (Chaisson et al. 2019).

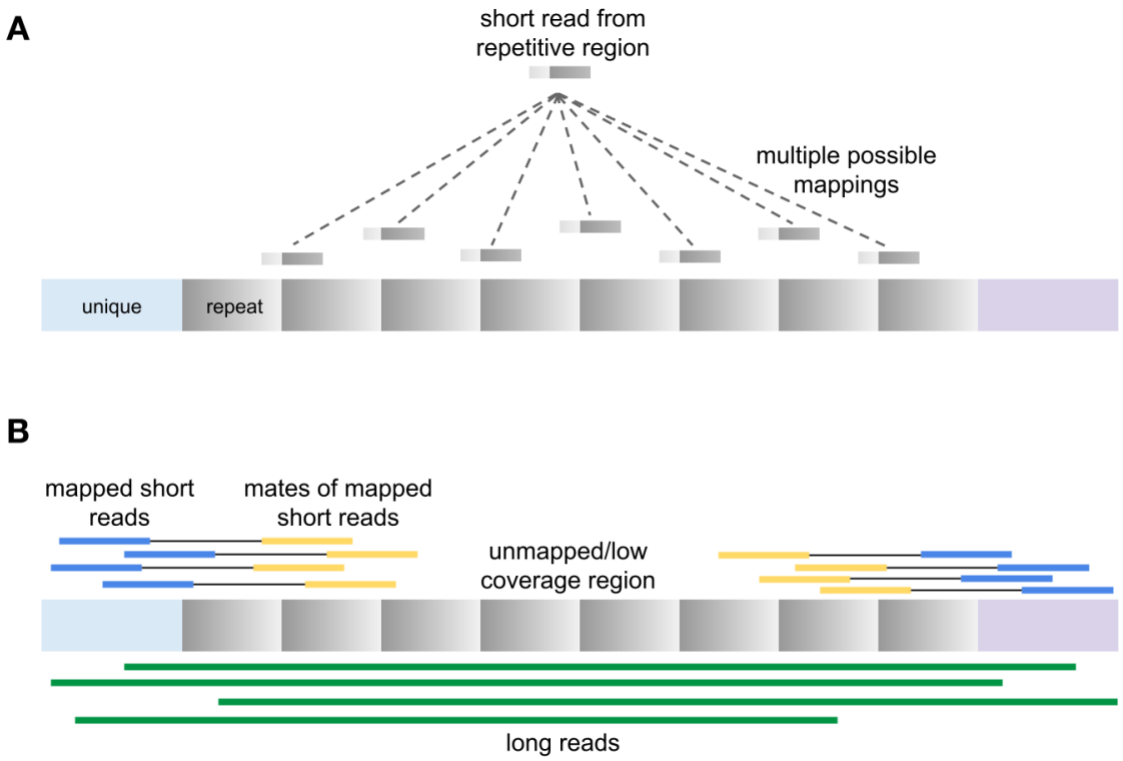


Figure 1.1: Mapping short and long reads to repetitive genomic regions. A) Short reads originating from a repetitive region of the genome can align equally well to many locations and are generally ignored due to ambiguity about their origin. **B)** Short reads (blue) can map to the unique sequences flanking the repetitive region, and their mates (yellow) might be mapped within the repetitive region based on the estimated average read fragment length. Even with the extension of read mates, there are often repetitive regions with low or no coverage. In contrast, long reads (green) that map to the unique flanking sequences can extend far into or across the entire repetitive region.

1.1.2.2 Genome assembly, resequencing, and read alignment

Once genome sequencing data have been generated, there are two main approaches for determining where the reads originated: 1) de novo assembly and 2) read alignment to a reference genome (also called resequencing). While both methods can be used for identifying genomic variants, there are strengths and weaknesses associated with each approach.

De novo genome assembly is the process of determining a genome from scratch by compiling DNA reads and fragments with overlapping sequences into contigs (Chaisson et al. 2015). Due to heterozygous alleles, repetitive loci, and sequencing errors, the overlaps are not fully linear

and branch out into a graph structure (discussed in greater detail in section **1.4**). Linear sequences can be produced by determining unambiguous paths through the graph structure and condensing the assembly into representative **consensus sequences** for contigs and scaffolds. There are numerous software applications for performing de novo assemblies, including `Velvet` (Zerbino and Birney 2008), `ABySS` (Simpson et al. 2009), and `SGA` (Simpson and Durbin 2012). Variants can then be called by aligning the assembly to another or by aligning contigs from the assembly to the reference genome (Li 2012; Zhang et al. 2020). Some de novo assemblers, such as `Cortex` (Iqbal et al. 2012), provide variant calling functions directly by maintaining heterozygous sites instead of condensing down into a single consensus sequence. Since the reference genome is not used during de novo assembly, reads originating from regions not represented in the reference are maintained. A total of 3.2Mb of non-reference sequence was identified using `Cortex` to generate three multi-sample population-specific de novo assemblies using 164 individuals from the 1000GP short-read sequencing data (Iqbal et al. 2012). Performing de novo assembly is effective at identifying structural variants, particularly when long sequencing reads are used as they can better resolve complex regions (Sohn and Nam 2018). However, in addition to the high computational resources required to generate the complex assembly structures, there are limitations with using de novo assembly approaches for variant detection; unequal genome coverage and the inability for short reads to span long repetitive regions leads to ambiguity, incorrect arrangements, and gaps in many genomic regions (Sohn and Nam 2018).

Resequencing is the process of using an existing reference genome as a guide for compiling sequencing reads. This involves **read mapping**, which is the process of identifying the location of a read sequence within the reference genome, and **read alignment**, which is the process of matching the read sequence to the reference sequence base-by-base. Broadly speaking, read mapping uses a **seed-and-extend** process by which small subsequences (seeds) within the reads are located in the reference **index**, which is a concise data structure that stores the locations of genome substrings. When a seed matches an indexed substring of the reference, the full read-to-reference alignment can be computed (extended) after extracting the reference substring around the initial seed match. Imperfect alignments are allowed to account for sequencing errors and variants within the reads. Several software tools have been developed

that perform both the mapping and alignment steps, including `Bowtie` (Langmead et al. 2009), `BWA` (Li 2013), and `NovoAlign` (https://www.novocraft.com/products/novoalign). RNA-sequencing aligners are similar but account for transcript splicing, whereby reads are mapped to the reference with gaps at intron locations without penalization (Trapnell et al. 2009; Dobin et al. 2013).

Resequencing has been the preferred approach for variant calling due to the reduced computational demand compared to de novo assembly. The reference genome has been fundamental in discovery of many variants, particularly SNPs and indels (Schneider et al. 2017). Resequencing has uncovered SVs as well, especially in understudied populations, but is more difficult with short-read technology. Nevertheless, there is a desire to improve variant analysis to better identify complex SVs.

1.1.2.3 Variant calling approaches

At the fundamental level, variant calling works by identifying differences between the aligned or assembled sequencing reads compared to another genome, usually the reference genome. Polymorphisms are discussed relative to the reference: an allele observed in a sample but not observed in the reference is considered an **alternate allele**. Germline variants are expected to be observed in 50% or 100% of the reads, depending on whether the variant is heterozygous or homozygous, respectively. Mismatches occurring at low frequencies are considered sequencing errors or somatic variants and are ignored for the purposes of germline variant calling. SNPs and small indels are relatively easy to call for both short- and long-read sequencing since the full variant is contained within many individual reads. Several software tools exist for calling small variants: the software `GATK` contains numerous tools for data preparation and calling variants, including the widely used `HaplotypeCaller` (DePristo et al. 2011; Poplin et al. 2018b); `FreeBayes` (Garrison and Marth 2012) `and Platypus` (Rimmer et al. 2014) perform SNP and indel calling; and `Dindel` (Albers et al. 2011) is specific for identifying indels.

SVs can be detected in short paired-end reads by identifying anomalies in the expected **fragment lengths** and **read orientations**, and by the presence of **split reads** (**Figure 1.2**)

(Mahmoud et al. 2019). Deletions can be inferred when the distance between a pair of reads mapped to the reference, the observed fragment length, is longer than the expected fragment length, which can be estimated by averaging the observed lengths for all read pairs. Some mapped reads will also appear to be split across the reference with breakpoints indicating the deletion site. Insertions can likewise be inferred from regions where the observed fragment length is shorter than expected, but depending on how long the insertion is, the precise endpoints may be difficult to discern. Interchromosomal translocations are inferred when reads in a pair are mapped to different chromosomes or when a read is split and mapped to two chromosomes, while intrachromosomal translocations are indicated by read pair orientations facing away from each other in a left-reverse, right-forward manner (instead of the expected left-forward, right-reverse orientation). Tandem duplications also have the incorrect left-reverse, right-forward read orientation. Inversions and inverted duplications are inferred when read-pair orientation is in the same direction (i.e. left-forward, right-forward or left-reverse, right-reverse). In addition to paired-end read patterns, coverage can be used to help identify duplications and deletions as these will see increases and decreases in read coverage relative to flanking regions, respectively. The detection of TE variants follows similar split read and read orientation information, and can also be discovered by searching the genome for known TE sequences or small motifs (Goerner-Potvin and Bourque 2018). Some short-read SV detection software includes `BreakDancer` (Chen et al. 2009), `DELLY` (Rausch et al. 2012), `LUMPY` (Layer et al. 2014), and `Manta` (Chen et al. 2016).

As mentioned earlier, SVs are easier to call with long reads due to their ability to often span the entire length of the variant while also providing positional context in the sequence flanking the SV. In contrast, SV detection from short reads has low sensitivity and a high false discovery rate, up to 85%, and attempts to reduce false discovery often lead to missing true variants (Sedlazeck et al. 2018). Long-read SV callers include `NanoSV` (Cretu Stancu et al. 2017), `SMRT-SV` (Huddleston et al. 2017), and `Sniffles` (Sedlazeck et al. 2018). TE-specific callers that search for TE motifs present within long reads but not within the reference genome include `TLDR` (Ewing et al. 2020) and `PALMER` (Zhou et al. 2020).

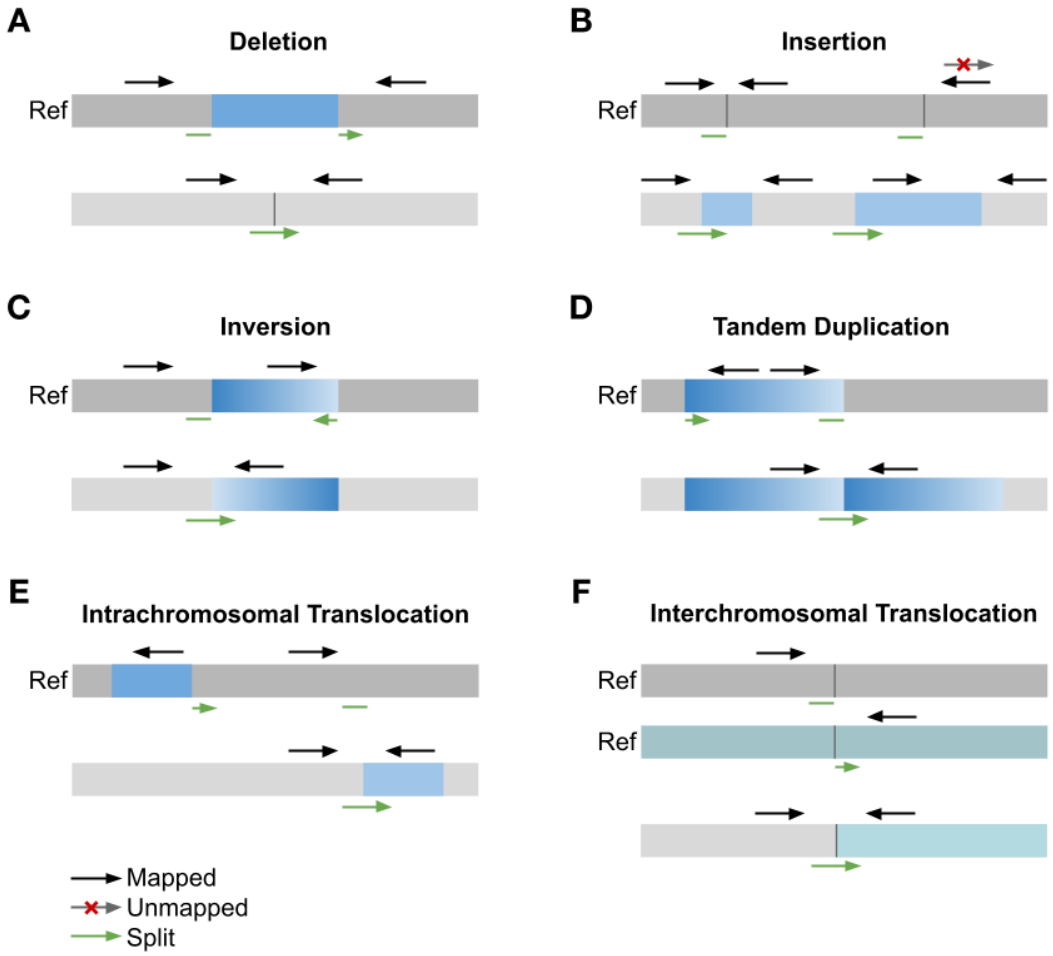


Figure 1.2: Detection of structural variants using short reads. SVs are named relative to the reference genome (top bar per figure). **A)** Deletion: longer observed fragment lengths than expected and the presence of split reads. **B)** Insertion: shorter observed fragment lengths than expected (left), or for large insertions, an unmapped read mate (right). In addition, partially mapped split reads may occur. **C)** Inversion: incorrect read pair orientations (facing the same direction) or split reads with one end mapped in the opposite direction. **D)** Tandem duplication: incorrect read pair orientations (facing outwards) or split reads with the 3' end facing the 5' end. **E)** Intrachromosomal translocation: incorrect read pair orientations (facing outwards) or split reads with the 3' end facing the 5' end. **F)** Interchromosomal translocation: read pairs map on separate chromosomes or reads are split between separate chromosomes. Image inspired by Mahmoud et al. (2019) and the Integrative Genomics Viewer (`IGV`) user guide (https://software.broadinstitute.org/software/igv/AlignmentData).

The majority of variants discovered have been deposited and curated in several databases, allowing researchers and clinicians to search and learn about the nature and frequency of the variants and any phenotypic or clinical implications they may have. Databases such as dbSNP exist for SNPs and indels (Sherry et al. 1999), while dbVar and DGVa compile SVs (Lappalainen et al. 2013). Several repositories curate a broad range of variants, such as dbGAP (Tryka et al. 2014), ClinVar (Landrum et al. 2014), OMIM (Amberger et al. 2015), gnomAD (Karczewski et al. 2020), and EVA (Cezard et al. 2022). Locus-specific databases exist for the human leukocyte antigen (HLA) and killer-cell immunoglobulin-like receptor (KIR) genes (Robinson et al. 2013) and for cytochrome P450 (CYP) genes (Gaedigk et al. 2018). Disease-specific databases also exist, such as CIViC (Griffith et al. 2017) and COSMIC (Tate et al. 2019) for cancer variants.

1.1.3 Differences in variation across human populations

Several large-scale projects have cataloged human variation across diverse populations since the completion of the reference genome: 1,064 cell lines from 54 populations for the Human Genome Diversity Project (Cann et al. 2002), 270 samples from four populations for the International HapMap Project (International HapMap Consortium 2005), 2,504 samples from 26 populations for the 1000 Genomes Project (The 1000 Genomes Project Consortium 2015), 184 predominantly European samples for the Personal Genome Project (Mao et al. 2016), 300 samples from 142 populations for the Simons Diversity Project (Mallick et al. 2016), and over 500,000 predominantly European samples for the UK BioBank project (Sudlow et al. 2015). Numerous population-specific studies have also been performed, including in the Netherlands (Francioli et al. 2014; Hehir-Kwa et al. 2016), Japan (Nagasaki et al. 2015), Iceland (Gudbjartsson et al. 2015; Beyter et al. 2021), Denmark (Maretty et al. 2017), and Africa (Sherman et al. 2019).

These large-scale population studies have uncovered the types of variants found in human genomes and have provided an idea of how frequently they are observed, which can differ across populations. Common variants are generally found in all populations, while rare variants tend to be confined to closely related populations, namely populations in close geographical proximity or populations with recent admixture (The 1000 Genomes Project Consortium

2015). 1,434 of over 68,000 identified SVs from the 1000GP varied greatly in allele frequency among populations, several of which were associated with known sites under positive selection (Sudmant et al. 2015). The 1000GP also found that 86% of variants identified were found in a single continental population. The majority of the total number of variants observed in the 1000GP cohort were very rare with a global frequency of < 0.5%, but these only comprised 1–4% of an individual genome. Relative to people of African ancestry, individuals of other ancestries have fewer variants owing to the reduced population sizes and bottleneck effects of ancestral groups migrating out of Africa (**Figure 1.3**). Africans have on average 27% more heterozygous deletions than other populations (Sudmant et al. 2015).

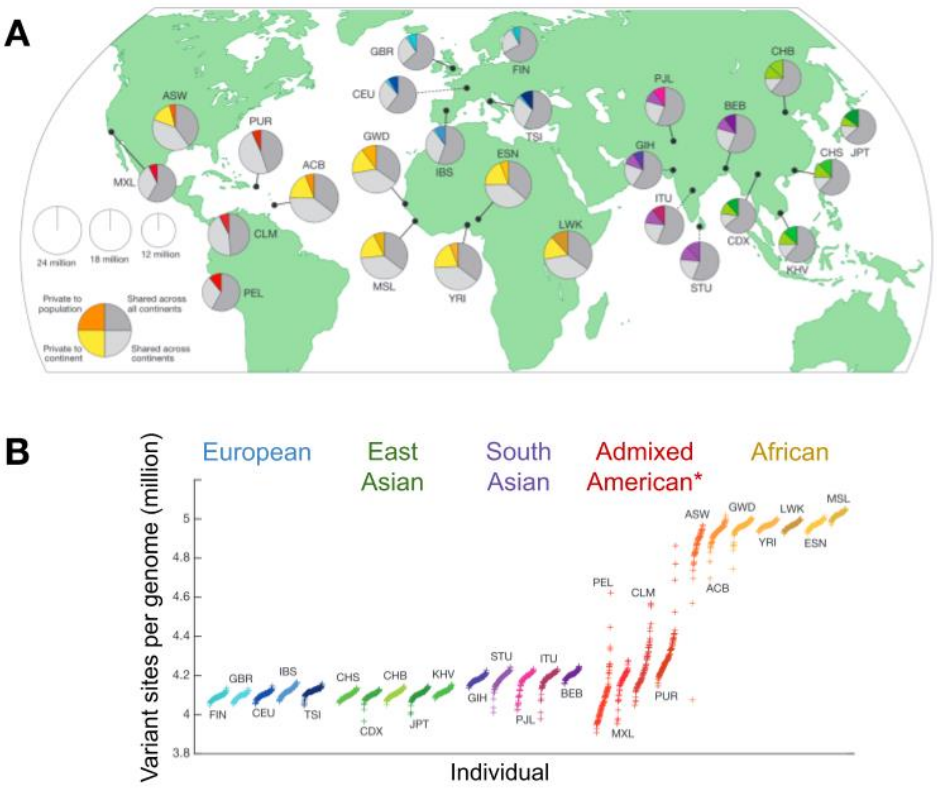


Figure 1.3: Overview of variants found in different populations by the 1000 Genomes Project. A) Distribution of variant frequencies across 26 populations in five continents. Pie graphs represent the proportion of variants shared across all continents (dark grey), shared across some continents (light grey), found only in the continental population (light color), and found only in the specific population (dark color). African populations have higher proportions of population-specific variants. Figure from The 1000 Genomes Project Consortium (2015), published under a Creative Commons License (CC BY-NC-SA 3.0). **B)** Number of variant sites per individual, grouped by population and continent. African populations have more variant sites than other populations. Population abbreviations are explained in **Appendix Table 1**. *Note: The 1000GP refers to this continental population as "American", but in this thesis it is referred to as "Admixed American" for clarification as the populations are of admixed European, African, and Indigenous American ancestries. Figure modified from The 1000 Genomes Project Consortium (2015), published under a Creative Commons License (CC BY-NC-SA 3.0).

Despite all the sequencing and variant detection studies performed on thousands of individuals, genomic data is still overwhelmingly European (Popejoy and Fullerton 2016; Sirugo et al. 2019). Given the clinical implications of disease-causing and medication-metabolizing variants, it is imperative that the scientific community expands research into underrepresented

populations in a way that respects and promotes cultural autonomy (Claw et al. 2018; Jackson et al. 2019).

1.2 Issues with the reference genome

As valuable as the reference genome has been with identifying an enormous number of variants in numerous studies, the mosaic composition of the reference can lead to biases in genomic analyses. This includes a propensity towards calling reference alleles over variant alleles, an inability to identify SVs, and overlooking clinically relevant reference alleles that occur at low frequencies in most populations.

1.2.1 Reference bias

When analyzing resequencing experiments, reads align better to the reference genome when there are fewer mismatches. Reads with an alternate SNP, for example, will have at least one mismatch relative to the reference, and any further sequencing errors in the read may result in the aligner mapping the read incorrectly or being unable to map the read entirely. During variant calling, this unfortunately leads to a **reference bias**, or an overestimation of reference alleles compared to alternate alleles that are present in the sequencing data. Strong reference biases tend to occur at SNP sites where the flanking sequence is shared in another genomic region (Degner et al. 2009), but the absence of long stretches of DNA in the reference also prevents the discovery of many types of SVs due to the inability to map reads from affected regions. In a study of 910 individuals of African descent, it was estimated that the collective African genome length is approximately 10% longer than GRCh38, which corresponds to 296,485,284 bp along 125,715 separate contigs (Sherman et al. 2019). 387 of these additional contigs originate from 315 protein-coding genes, meaning sequences with potentially functional effects have historically been overlooked. The longest contig, unable to be localized to a specific chromosome, was over 152,000bp long and was found in 11 samples. In addition, about 40% of the 296.5 Mb of novel sequence was previously identified in Korean or Chinese populations, demonstrating how GRCh38 does not accurately reflect all populations.

Reference bias can negatively affect several types of genomic analyses. When calling variants, reference bias can result in a heterozygous site being called as homozygous. One study aimed to quantify reference bias in the 1000GP variant calls for five HLA genes. The 1000GP called genotypes by aligning short-read sequencing data to the hg19 reference (an older build prior to GRCh38 roughly equivalent to GRCh37) and using a consensus of variant calling software calls to assign genotypes to each sample (The 1000 Genomes Project Consortium 2015). Comparing genotype concordance between the Illumina calls and calls from an HLA Sanger sequencing panel featuring the same samples (Gourraud et al. 2014), Brandt et al. (2015) found that an average of 18.6% of the SNPs disagreed between the two technologies with the 1000GP calls biased towards overestimating the reference allele. In addition, reference bias was higher for HLA genes with high levels of polymorphism and the bias was correlated with SNPs showing higher levels of heterozygosity among the samples. Reference bias is stronger with shorter read fragments, leading to issues in accurately identifying variants in paleogenomic studies where ancient DNA samples are shortened through degradation (Günther and Nettelblad 2019). Reference bias also has a particularly strong effect on **allele-specific expression** (ASE) analyses, where precise comparisons of RNA transcript expression are required to determine if one allele is favored over the other at a given heterozygous site (Degner et al. 2009). A bias towards the reference allele can incorrectly skew expression estimates and distort genuine instances of ASE.

Several methods have developed to account for or reduce reference bias: filtering out SNPs in regions with low mappability (Günther and Nettelblad 2019), mapping data to both parental genomes if available, increasing the number of mismatched bases allowed during read alignment to the reference (Stevenson et al. 2013), using a polymorphism-aware alignment tool (Wu and Nacu 2010), and modifying the reference to contain both reference and variant alleles (Vijaya Satya et al. 2012; Pandey et al. 2013; van de Geijn et al. 2015; Peyrégne et al. 2019). For non-ASE analyses, masking polymorphic SNPs by converting the base to a universal ‘N’ can also be performed. However, reducing reference bias by modifying the sequencing reads or the reference genome can result in the loss of informative reads. In addition, some methods that use alternate alleles only consider biallelic SNPs (Prüfer 2018; Günther and Nettelblad 2019), ignoring the numerous multi-allelic SNPs, indels and structural variants that are also subject to reference bias.

1.2.2 Reference minor alleles

Due to the mosaic nature of the reference genome, not all variant sites are represented with the most frequent alleles. **Reference minor alleles** (also called rare reference alleles) are generally ignored during variant calling because the software does not report reference alleles; only alternate alleles are identified, even if the alternate allele is very high in frequency (Ferrarini et al. 2015; Magi et al. 2015; Karthikeyan et al. 2017). This can have a negative impact on clinical studies if the reference minor allele is pathogenic. About two million GRCh37 reference alleles are minor alleles within the 1000GP populations (Ballouz et al. 2019), over 96,000 of which have allele frequencies less than 1% (Magi et al. 2015). In addition, nearly 20,000 of these rare reference alleles have never been observed in healthy individuals, and 70,000 have only been observed in a heterozygous manner. Many were, however, found in the GENCODE (Harrow et al. 2012) and ENCODE (ENCODE Project Consortium 2012) databases with pathogenic associations, suggesting that these rare reference alleles could be important loss- or gain-of-function mutations.

In a study that sequenced 13 diploid and two haploid genomes of diverse populations to identify structural variants, of the SVs considered either shared among all samples or present in at least half, an astonishing 95.4% and 66.7% were not observed in GRCh38, respectively (Audano et al. 2019). Instead, the minor allele was represented in the reference. Additionally, for a small number of SVs (< 1%) the reference allele was not observed in any samples at all, indicating that the reference genome represents a very minor allele, or possibly an error, at these loci. There was a lot of enrichment for repetitive regions such as STRs, variable number tandem repeats, and TEs in the shared and major SVs identified, which reflects the difficulty of sequencing repetitive regions experienced by the HGP. Nevertheless, these are potentially important regions that researchers want to be able to analyze.

1.3 Improving reference-based variant calling

How can the current practice of calling variants from the reference genome be improved? Advancements in sequencing technology and variant calling algorithms have helped increase the number of variants identified while reducing false calls. Several approaches have also been suggested to improve the reference genome itself: creating a consensus reference, creating an ancestral allele reference, using population-specific references, or switching to the newly released and complete Telomere-to-Telomere reference genome.

1.3.1 Using long-read sequencing to elucidate repetitive regions

Long-read sequencing has revolutionized SV calling by producing reads that are often able to fully span large SVs. With older implementations of long-read technology, Illumina sequencing was often also performed in order to error-correct the reads and improve accuracy. However, with the latest ONT and PacBio HiFi methodologies, highly accurate reads can be obtained without the need for additional sequencing (Hon et al. 2020; Sereika et al. 2022). Both ONT and PacBio HiFi sequencing have been used to elucidate clinically relevant SVs (Miao et al. 2018; Hu et al. 2020). Long-read sequencing improves SV calling and alignment across repetitive regions, as well as haplotype phasing of variants (Sedlazeck et al. 2018). Researchers taking advantage of long-read sequencing have recently released sequences for the entire centromere in chromosome Y (Jain et al. 2018), all of chromosome X (Miga et al. 2020), and finally the full human genome (Nurk et al. 2022), providing much-needed insight into the traditionally hard to study heterochromatic centromeres and telomeres.

While the benefits of long-read sequencing are clear, there have already been hundreds of thousands of genomes sequenced with short-read technology, and it is not feasible to resequence every sample. In addition to continuing long-read sequencing studies, research should also push towards improving short-read variant calling methods to extract additional information from the existing large inventory of sequenced genomes.

1.3.2 The utility of *k*-mers in genome analyses

The full length of a sequencing read can impair analyses if those reads contain sequencing errors as they might lead to incorrect read mapping or variant identification. To circumvent this, reads are often subset into ***k*-mers**, which are short substrings of DNA that are *k* bases long. With a *k* less than the length of a read, it is likely that several *k*-mers exist without the sequencing errors that are present in the read. There is a fixed number of *k*-mers that can be generated using the four nucleotides, yet interestingly, even when *k* is small (< 15bp), not all possible *k*-mers are observed in a given genome due to negative selection against these sequences (Hampikian and Andersen 2007; Georgakopoulos-Soares et al. 2021). Understanding the expected count distribution of *k*-mers for a genome is useful for identifying erroneous reads and correcting them or removing them from downstream analyses, since rarely occurring *k*-mers likely contain sequencing errors (Kelley et al. 2010; Marçais and Kingsford 2011).

k-mers are used in many other genomic applications. The **de Bruijn graph** (discussed in greater detail in section **1.4**) is a structure built from overlapping *k*-mers that can be used to perform de novo genome assembly. The structure allows for quick identification of paths or ways to travel through the graph and construction of genome contigs (Compeau et al. 2011). The **colored de Bruijn graph** is an extension of the de Bruijn graph that was developed for multi-sample genome assembly and for genotyping variants with the software `Cortex` (Iqbal et al. 2012). `Cortex` performs de novo assembly of multiple genomes simultaneously and preserves sample sequence identity by coloring (i.e. labeling) the graph components that correspond to haplotype paths. If one of the paths is the reference genome, the graph can be used for variant calling at sites where the sample of interest diverges from the reference path.

Precise read mapping is not always necessary for reference-based analyses. The software `kallisto` performs **pseudoalignment** to speed up RNA-sequencing quantification (Bray et al. 2016). This is done by identifying from which transcript a read originated instead of precisely aligning the read to a specific coordinate, which is enough for accurate transcript quantification. `kallisto` generates a de Bruijn graph using *k*-mers present in a **transcriptome**, the collection of transcripts for a particular organism, with colored paths

through the graph corresponding to the different known transcripts. The set of transcripts that match *k*-mers from the reads are identified in the de Bruijn graph. This allows for reads to be assigned to transcripts without precise alignment. Transcript quantification can then be estimated from all of the pseudoalignments through an expectation-maximization algorithm. Using *k*-mers to account for sequencing errors and determine sequence composition without precise alignment could be an interesting avenue for improving variant calling in repetitive regions of the genome.

1.3.3 Reference genome modifications

The major histocompatibility complex (MHC), which contains several genes with important roles in the response of the immune system, is the most polymorphic region in the human genome. The locus was originally represented in the reference genome as a mixture of haplotypes due to its large size of over 4Mb. In the NCBI35 build (three builds prior to GRCh38), the mosaic composition was replaced with a single haplotype, and beginning with build GRCh37, an additional seven alternate MHC sequences were included as separate contigs. These sequences were obtained by Sanger sequencing eight cell lines originating from consanguineous individuals, providing long-range homozygous sequences with accurate HLA haplotypes (Horton et al. 2008). Additional alternate loci in addition to the MHC were curated and included in GRCh38, which contains a total of 261 alternate haplotypes across 178 regions (**Figure 1.4**). In theory, these additional sequences allow for better read mapping in polymorphic regions because reads that differ greatly from the reference sequence may map better to an alternate haplotype. While the inclusion has improved alignment in many studies (Schneider et al. 2017), several read aligners ignore these contigs to avoid issues with **multimapping**, or mapping a read to multiple locations in the genome (Church et al. 2015). Additionally, there will likely need to be tens of thousands of alternate sequences to account for all of the polymorphisms found in genomes across populations, especially now that large novel SVs are being discovered more frequently. The GRCh38.p14 build currently has 90 novel contig patches ready for inclusion in the next major update (https://www.ncbi.nlm.nih.gov/grc/human).

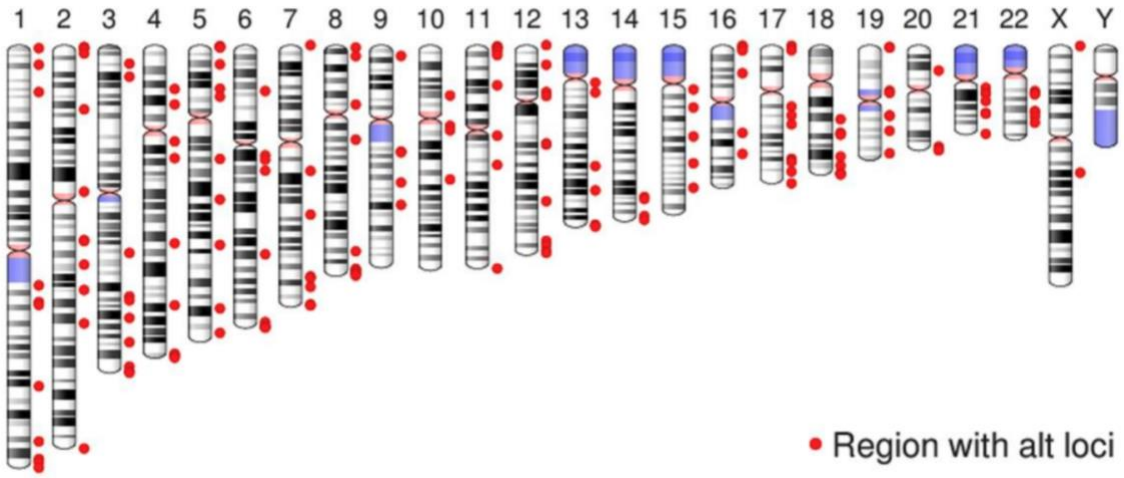


Figure 1.4: Additional haplotype sequences in the GRCh38 reference genome. GRCh38 contains sequences for 261 alternate haplotypes within 178 regions (red circles). In addition, patches are released for sequence corrections and novel sequences intended for inclusion in the next major update, where novel patches are converted into alternate loci. Figure from Schneider et al. (2017), published under a Creative Commons License (CC BY 4.0).

Instead of adding additional contigs to the reference, there have been proposals for changing the reference content to better reflect a generic genome. A **major allele** or **consensus reference** would contain alleles with the highest frequency at polymorphic sites (Dewey et al. 2011; Ballouz et al. 2019). By containing all major alleles, it is hypothesized that reference bias would be reduced in several cases, improving variant detection and downstream analyses. One study observed reduced genotyping errors and improved variant detection when using a major allele reference generated from GRCh37 and the major alleles in European 1000GP samples to align a family quartet with Northern European ancestry (Dewey et al. 2011). In another study, a major allele reference developed by considering allele frequencies for all the 1000GP variants was found to reduce false negative variant calls by 8% and false positive calls by 30% while also identifying new variants that were missed when using the unmodified hg19 reference (Karthikeyan et al. 2017).

An **ancestral allele reference** would contain alleles also found in our recent ancestors instead of alleles uniquely derived in humans (Balasubramanian et al. 2011). This reference would be population neutral, which should prevent unnecessary bias against populations underrepresented by the current reference. An ancestral allele reference would also remain

stable over time, meaning any updates would merely be corrections and not involve novel variants identified in modern humans. There would, however, be difficulties developing a complete ancestral allele reference since ancestral alleles have not been resolved at all polymorphic sites.

1.3.4 The Telomere-to-Telomere reference genome

Even with two decades of updates since the HGP released the first human genome draft sequence, the reference genome has not been able to represent heterochromatic sequences owing to the difficulties in sequencing these highly repetitive regions. In 2022, the **Telomere-to-Telomere** (T2T) consortium published the complete sequence of a human genome, save for chromosome Y, shedding light into the elusive telomeric and centromeric regions, ampliconic gene arrays, and ribosomal DNA loci (Nurk et al. 2022). Genome assembly was greatly simplified by using the entirely homozygous CHM13 genome as haplotype phasing was not required. A Y chromosome from scientist Leonid Peshkin was later sequenced and added to the T2T reference (Pennisi 2022; Rhie et al. 2022). Analyses using the T2T genome as a reference have demonstrated improvements in calling variants across the genome and in clinically relevant genes, decreases in unmapped Illumina reads, improved read coverage uniformity, fewer read mismatches for all five 1000GP continental populations for both short- and long-read sequencing, and decreases in false positives and false negatives for alignment of Illumina, PacBio Hifi, and ONT reads (Aganezov et al. 2022). The T2T reference clearly has incredible potential to provide more accurate variant calling in future studies.

Though more complete, switching to using the T2T genome as the gold-standard human reference is not a quick or simple endeavor owing to the presence of sequences in T2T not found in GRCh38. Coordinates will need to be remapped and databases for genes, regulatory regions, and variants will need to be updated with the new coordinates to ensure consistent representation with regards to decades of existing analyses (Church 2022).

1.3.5 Population-specific reference genomes

As sequencing costs decreased tremendously after the completion of the HGP, the sequencing and assembly of individual human genomes became increasingly popular. The first two

individual genomes described were those of scientists Craig Venter (Levy et al. 2007) and James Watson (Wheeler et al. 2008), both of European ancestry. Individual genomes from non-European populations have since been sequenced, using short-read sequencing or a combination of short- and long-read sequencing, including Chinese (Kidd et al. 2008; Wang et al. 2008; Li et al. 2010; Shi et al. 2016; Zook et al. 2016), Japanese (Kidd et al. 2008), Yoruba (Bentley et al. 2008; Kidd et al. 2008; McKernan et al. 2009; Li et al. 2010), Korean (Ahn et al. 2009; Kim et al. 2009; Cho et al. 2016; Seo et al. 2016), Khoisan and Bantu (Schuster et al. 2010), Mongolian (Bai et al. 2014), and Ashkenazi (Zook et al. 2016; Shumate et al. 2020) individuals. Ancient genomes have also been sequenced from a Neandertal (Green et al. 2010), an individual from the Saqqaq culture in Greenland (Rasmussen et al. 2010), and an ancient Siberian (Fu et al. 2014).

Some of the individual genomes sequenced have been assembled and annotated to some degree with the intention of being used as a **population-specific reference**. Using the fully annotated Ashkenazi Ash1 reference, sequencing reads from an individual estimated to have 66% Ashkenazi ancestry had fewer unmapped reads (0.5%) and around one million fewer SNPs identified after variant calling, compared to using GRCh38 as the reference genome (Shumate et al. 2020). Having fewer variants means more bases matched the reference used and were thus not penalized during read alignment. During annotation with all known genes from the CHESS database (Pertea et al. 2018), it was observed that 11 of the 108 genes that initially failed to map to Ash1 actually mapped to different chromosomes, suggesting the occurrence of translocations relative to GRCh38.

The use of multiple individuals by the HGP resulted in the interruption of some haplotypes at BAC boundaries, erroneously representing the sequences of some SVs (Church et al. 2015). A benefit of having a single-person reference genome is the preservation of haplotypes and better representation of SVs. The Chinese HX1 genome had over 20,000 insertions and deletions reflecting over 10Mb of sequences relative to GRCh38 (Shi et al. 2016). The Korean-specific reference KOREF_C, assembled from a single individual and augmented at sites of SNPs and indels with Korean major alleles, contained almost 10,000 SVs, with around 93% of the insertions and 70% of the deletions not represented in public databases (Cho et al. 2016). The

use of long-read sequencing in both of these genomes was imperative in detecting many SVs and maintaining long haplotypes.

Switching to a series of population-specific genomes would likely improve read mapping for many individuals. However, there are many technical questions that need to be considered for this approach to become standard practice. How many reference genomes should be maintained? Is one per continental population sufficient, or should there be references for individual populations as well? To which reference should an admixed individual be aligned? Is it beneficial to align samples to multiple population-specific reference genomes to determine the best fit? Finally, how would variants be described across population references if the coordinates differ due to deviations in SVs? Adopting the use of a single reference genome that represents multiple populations is perhaps a more feasible alternative.

1.4 Towards the use of graph-based reference genomes

The current human reference genome can only represent so much variation, and continuously adding alternate contigs to the reference will at some point become too cumbersome to use effectively. Recently there has been much discussion around using a human **pangenome** reference, which would contain variants from multiple populations in a non-discriminatory manner to reduce reference bias and improve genomic representation of understudied populations. The term pangenome was first used to describe an idea for collecting genomic and transcriptomic mutations from multiple types of common tumors (Sigaux 2000) and later for a core set of shared genes amongst all strains of a bacterial species (Tettelin et al. 2005). Pangenomes were initially predominantly used in the microbiology and metagenomics fields (Vernikos et al. 2015) and were collections of linear sequences from related strains. In recent years, pangenomes have been used in many species. However, species with larger genomes and a high number of individuals sampled become too large to maintain as a collection of linear sequences.

If variants from multiple populations are to be represented equally within a reference genome, a non-linear structure is the way forward (Church et al. 2015). A **graph** structure is one such

way to represent different variants at polymorphic sites. At the basic level, a mathematical graph consists of **nodes** connected by **edges** that depict relationships. When used in genomics, a **graph-based reference genome** generally has nucleotide sequences stored in nodes while the edges direct the flow of the sequence. Variation is represented by a divergence in the graph where alternate paths can be taken: SNPs form a "bubble" where one node connects to two or more nodes, indels involve edges that pass over nodes, inversions are represented with edges that connect to the end of a node to reverse the sequence, and repeats are represented with edges that connect to the beginning of a previous node (**Figure 1.5**). There are several names for this style of reference that are roughly equivalent while varying in precise structure, including reference graph, genome graph, pangenome graph, variant graph, and population reference graph.

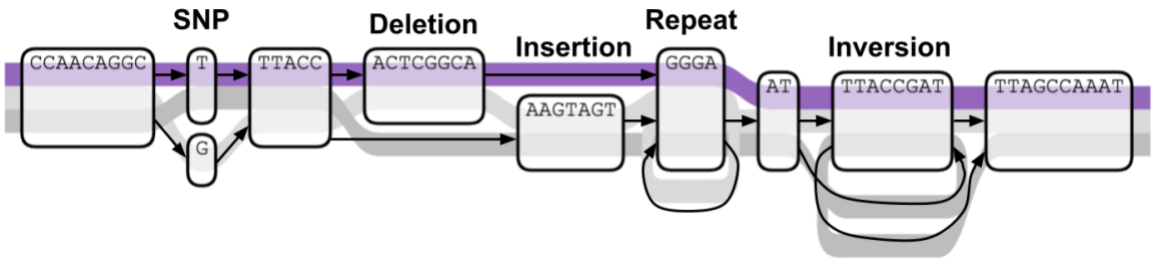


Figure 1.5: Example of a graph-based reference genome. The graph has nodes (boxes) that represent DNA sequences, which are connected by edges (arrows) that direct the flow of the sequence. The path (traversal through the nodes) through the graph for the reference is in purple, and two alternate paths are shown in grey. The alternate paths contain variants that are labeled relative to the reference sequence: SNP, two or more possible paths; deletion, an edge skips over a node; insertion, an extra node not in the reference path; repeat, an edge loops back to the beginning of the node; inversion, the edge from the previous node connects to the end of the node, and a second node leaves the start of the node and connects to the start of the next node. Image modified after generation with Sequence Tube Maps (https://vgteam.github.io/sequenceTubeMap).

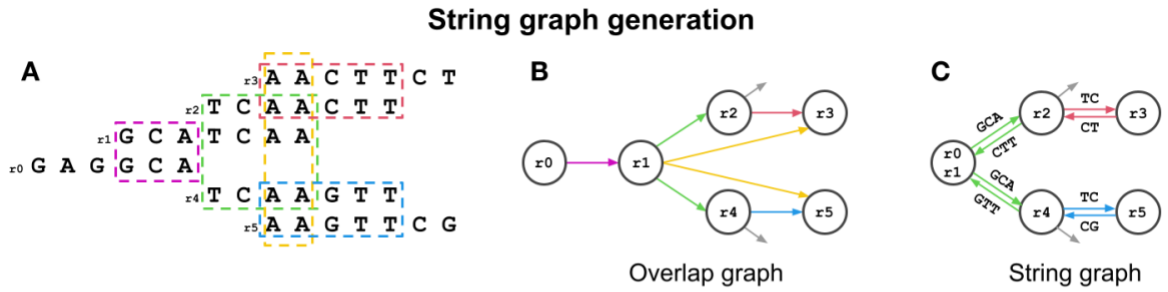
1.4.1 The history of graphs in genomics

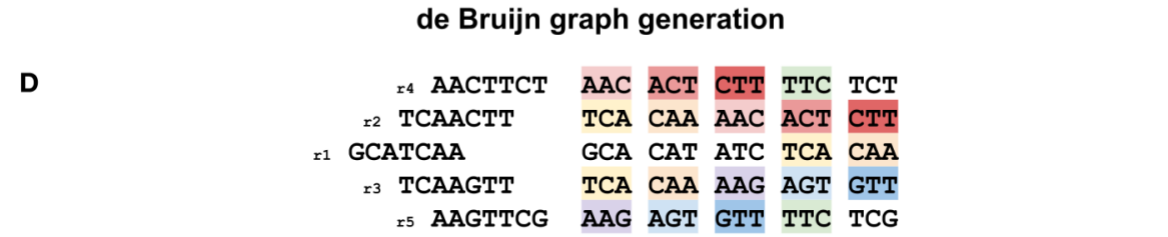
Though the focus on using graph-based reference genomes is relatively new, the graph structure has been used for several years to represent different kinds of genomic data, including phylogeny reconstruction, de novo genome assembly, splicing isoforms, and multiple

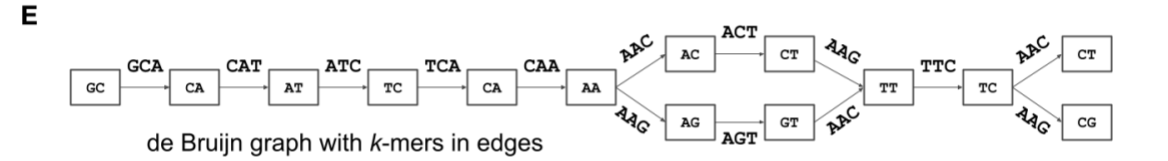
sequence alignments. Graphs used in genomics can be generally split into two categories, overlap graphs and sequence graphs.

Overlap graphs have historically been used for genome assembly (**Figure 1.6 A-B**). The overlap refers to how nodes, typically representations of reads, share overlapping sequences, which are represented with edges that connect the nodes with overlapping sequences. The overlaps are found by performing pairwise comparisons of all possible read pairs. Early genome assembly software produced overlap graphs from sequencing reads that were then arranged into contigs via edges that connected overlapping nodes (Myers Jr 2016). However, the early assemblers did not handle repeat units very well, if at all. A **string graph** is a type of overlap graph proposed as a way to reduce the number of edges in an overlap graph without changing the overall connections (Myers 2005). Edges are removed from the string graph if they are **transitive**, meaning they connect one node to another while skipping over an intermediate node while edges exist that connect the first node to the intermediate node and the intermediate node to the third node (e.g. the edge from `r1`↔`r3` can be removed if there are edges from `r1`↔`r2`↔`r3`). Transitive edges are therefore redundant and not needed. In addition to the removal of transitive edges, consecutive nodes with single edges between them are concatenated into a single node (**Figure 1.6 C**). While formally defined in 2005, the compression properties of string graphs were originally outlined ten years prior (Myers 1995), leading to the development of the `Celera` string-graph assembler that was used in the Celera Genomics human genome assembly that competed with the HGP (Myers et al. 2000; Venter et al. 2001). A **de Bruijn graph** is a type of overlap graph built by deconstructing sequencing reads into *k*-mers and connecting *k*-mers with overlapping sequences (**Figure 1.6 D**). This is similar to how reads with overlapping sequences are connected in string graphs, but does not require performing pairwise comparisons of all of the reads. The *k*-mers of a de Bruijn graph can be contained within the edges of the graph (**Figure 1.6 E**), as used by the assembler `Euler` (Pevzner et al. 2001), or within the nodes (**Figure 1.6 F**), as used by the assembler `Cortex` (Iqbal et al. 2012). The first proposed use of the de Bruijn graph in genomics was in the late 1980s for assembly after sequencing-by-hybridization, an old short-read sequencing technology that used very short oligos (~8bp) on microarrays that hybridized to small (~200bp) DNA fragments (Khrapko et al. 1989; Compeau et al. 2011). The hybridization oligos would

form *8*-mer edges of the de Bruijn graph, allowing for computationally efficient reconstruction of the DNA fragments sequenced (Pevzner 1989). In 2001 the de Bruijn graph-based assembler `Euler` was developed that made significant improvements in assembling repetitive sequences over the `Celera` string graph assembler, which masked repetitive regions because they could not be resolved (Pevzner et al. 2001). In addition, the nature of deconstructing sequencing reads into *k*-mers removed the need for all possible pairwise comparisons of the reads. In defence of the string graph, Myers (2005) stated that it is more space efficient than the de Bruijn graph, and that sequencing reads did not need to be broken down into *k*-mers for efficient genome assembly. String graphs have typically been used for longer Sanger sequencing while de Bruijn graphs became more popular with the explosion of next-generation short-read sequencing studies due to the faster compute time (Myers Jr 2016). Interestingly, string graph-based assembly approaches have become more prevalent again with long-read sequencing due to the higher error rates that hinder accurate *k*-mer representation in de Bruijn graphs.


String graph generation
A
r3 A A C T T C T
r2 T C A A C T T
r1 G C A T C A A
r0 G A G G C A
r4 T C A A G T T
r5 A A G T T C G
B
r0
r1
r2
r3
r4
r5
Overlap graph
C
r0
r1
r2
r3
r4
r5
GCA
CTT
TC
CT
GCA
GTT
TC
CG
String graph


de Bruijn graph generation
D
r4 AACTTCT AAC ACT CTT TTC TCT
r2 TCAACTT TCA CAA AAC ACT CTT
r1 GCATCAA GCA CAT ATC TCA CAA
r3 TCAAGTT TCA CAA AAG AGT GTT
r5 AAGTTCG AAG AGT GTT TTC TCG


E
GC
GCA
CA
CAT
AT
ATC
TC
TCA
CA
CAA
AA
AAC
AC
ACT
CT
AAG
AAG
AG
AGT
GT
AAC
TT
TTC
TC
AAC
CT
AAG
CG
de Bruijn graph with k-mers in edges

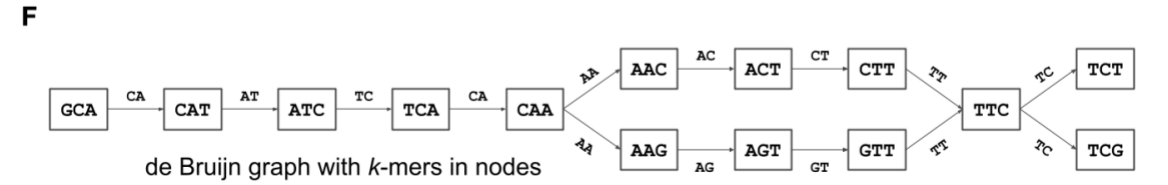

F
GCA
CA
CAT
AT
ATC
TC
TCA
CA
CAA
AA
AAC
AC
ACT
CT
CTT
TT
AA
AAG
AG
AGT
GT
GTT
TT
TTC
TC
TCT
TC
TCG
de Bruijn graph with k-mers in nodes

Figure 1.6: Types of overlap graphs used to represent genomic assemblies. A) Six sequencing reads with overlapping sections between reads outlined in dotted squares. **B)** The overlap graph has nodes representing sequencing reads and edges representing the overlaps between the connected reads. Edge colors match the dotted square overlaps in A). All possible overlaps between the prefix (beginning) of one read and the suffix (end) of another read are given an edge. There is no edge between `r2`↔`r4` because the overlap is in the prefixes for both reads, and there is no edge between `r2`↔`r5`, for example, because the overlap (yellow) in `r2` is in the middle of the read instead of the beginning or end. The two yellow edges are transitive as they skip over nodes (e.g. edge between `r1`↔`r3`) while representing the same information as two other edges connecting the nodes (e.g. `r1`↔`r2`↔`r3`). Gray arrows represent connections to nodes not depicted. **C)** The string graph is a more concise representation of an overlap graph. Transitive edges have been removed, and the remaining edges are bidirectional and labeled with the unmatched prefix of the upstream read and the unmatched suffix of the downstream read. The consecutive nodes `r0` and `r1` have been concatenated because there is a single edge (magenta) between them. Nodes `r2` and `r3` cannot be concatenated because there is also an edge connecting `r2` to another node (not depicted in this figure). **D)** de Bruijn graphs are generated by breaking down all sequencing reads into *k*-mers (*3*-mers in this example). This increases the amount of sequence data to represent, but *k*-mers shared among reads (colors) are represented as single nodes. **E)** Edge-centric de Bruijn graph with *k*-mers stored in the edges. **F)** Node-centric de Bruijn graph with *k*-mers stored in the nodes.

While the term **sequence graph** has been used for various types of graphs over the years, including an early de Bruijn graph (Idury and Waterman 1995), today it is generally used to describe a graph with **blunt edges** in that there are no overlapping sequences shared between nodes; edges instead direct the flow of information through the graph (**Figure 1.7 A–B**). Sequence graphs are used in graph aligners such as `vg` (Garrison et al. 2018) and `GraphAligner` (Rautiainen and Marschall 2020) and are typically the structure of modern graph-based reference genomes. The sequence graph was first introduced in 1989 as a method for performing alignment of related sequences and reconstructing phylogenies (Hein 1989). Hein's sequence graph used graph edges to represent nucleotides and was generated as part of the dynamic programming process for sequence alignment. Similarly, the **partial-order alignment** (POA) **graph** was developed to represent and perform multiple sequence alignments (Lee et al. 2002). The graph allows for all equally viable alignments to be represented without the degeneracy caused by condensing a pairwise alignment to a consensus sequence before iteratively aligning additional sequences. Unlike Hein's sequence graph, POA graphs represent single nucleotides in the nodes (**Figure 1.7 C**). **Splicing graphs** were

developed to simplify visualization of all known transcript isoforms that result from alternative splicing. Interestingly, the first splicing graphs were initially constructed as de Bruijn graphs and then simplified into a sequence graph where, given sufficient expressed sequence tag coverage, nodes represented exons or alternative splicing units, and edges were weighted by the number of expressed sequence tags that supported the splice junction (Heber et al. 2002) (**Figure 1.7 D**). Splicing graphs are still used in modern RNA-sequencing analysis software such as `MISO`, where they are used to generate sashimi plots (Katz et al. 2010).

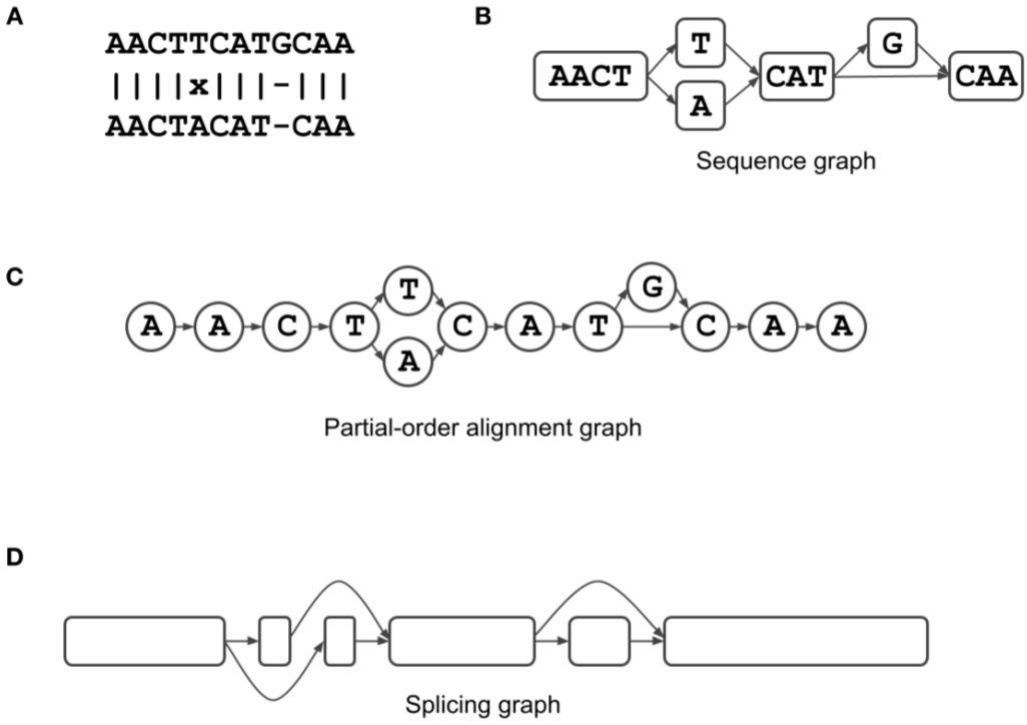


Figure 1.7: Types of sequence graphs used to represent DNA sequences. A) Alignment of two similar sequences that contains one base mismatch and one indel. Sequence graphs can be used to condense the information represented in alignments, such as the graphs in B) and C). **B)** The sequence graph has nodes of variable length corresponding to the length of the DNA contained within. Branching to multiple nodes occurs at sites of polymorphisms, while edges skip over nodes at sites of indels. **C)** The partial-order alignment graph is similar to the sequence graph, but has single-base nodes for alignment purposes. **D)** The splicing graph represents exons (or splicing units) in nodes while the edges represent the observed or known sites of alternative splicing.

The first representation of a pangenome as a graph structure instead of a collection of linear sequences was in 2009, where an Arabidopsis pangenome was generated with the alignment tool `GenomeMapper` (Schneeberger et al. 2009). Paten et al. (2014) proposed a sequence graph structure to represent the current reference and individual genome assemblies using a hierarchy of references, making updating the reference collection easier since new variants would be added to the bottom of the hierarchy and higher levels would be unaffected with new additions. Software tools have since been developed that allow users to create graph-based reference genomes (discussed further in section **1.4.2**). Reference graphs do not necessarily need to be generated by the user in order to take advantage of the structure; software for read alignment and variant calling have been developed that utilize the graph structure internally. In a graph-adjacent manner, Sirén et al. (2011; 2017) developed the generalized compressed suffix array (`GCSA` and `GCSA2`) that uses a multiple sequence alignment of several genomes converted into a finite automaton which is indexed in a way that allows string searches for any recombinant sequence obtained from the multiple sequence alignment. `Cortex` uses a de novo assembly approach for identifying sites of heterozygosity within an individual genome, as well as for distinguishing repeated elements from polymorphic sites when using several genomes (Iqbal et al. 2012). The software makes use of a modified de Bruijn graph by coloring the nodes and edges that correspond to specific samples. The software `BWBBLE` augments a standard reference by replacing SNP reference alleles with the **IUPAC nucleotide code** for ambiguous bases (Cornish-Bowden 1985) and appending indels and SVs at the end of the reference with flanking sequences for positional context (Huang et al. 2013). `HISAT2` is a variant-aware modification of the `HISAT` spliced-read aligner that converts the standard reference to a graph reference with around 14.5 million variants from dbSNP (Sherry et al. 1999) and uses several local indexes to speed up string searches during alignment (Kim et al. 2019). Reads from repetitive regions that would normally align to multiple locations in the genome are instead aligned to a single generated repeat sequence, decreasing the total number of alignments reported while still retaining information about mappings to repetitive regions. Using a linear reference and a variant call format (vcf) file of known variants, the software `FORGe` scores variants in terms of how much read alignment to those loci is expected to improve while considering factors such as SNP density, allele frequencies, and whether including the variant

results in additional repetitive loci (Pritt et al. 2018). The top-scoring variants from `FORGe` can then be used during the indexing stage of the companion aligner `HISAT2`.

1.4.2 Constructing human pangenome graphs

Though some pangenomes have been represented as de Bruijn graphs (Marcus et al. 2014), the majority make use of the sequence graph structure, which is more compact in size and avoids the redundancy of overlapping *k*-mers. Two common methods to construct a graph are: 1) using a linear reference sequence and augmenting it with known variants, and 2) performing a multiple sequence alignment of alleles or haplotypes and from that generating a partial-order alignment graph. A standardized way to represent a sequence graph is the **graphical fragment assembly** (gfa) format, which describes the structure in terms of segments, links, and paths. In its simplest form, each **segment** represents a node and consists of a name followed by the nucleotide sequence; **links** represent edges and contain the node names to be connected; and **paths** represent specific traversals through the graph nodes that signify sequences of interest (e.g. allele variants) and consist of the path name, the names of the series of nodes to be traversed, and the orientation of the sequence. The variant graph (`vg`) toolkit is comprehensive software that has been developed to generate, align, augment, visualize, and perform variant calling with reference graphs (Garrison et al. 2018).

The first human pangenomes described were region specific and focused on polymorphic genes and loci with medical relevance. Dilthey et al. (2015) represented the major histocompatibility complex (MHC) in a graph-based reference (referred to as a population reference graph by the authors) using the primary and alternate MHC sequences from GRCh37 and SNPs from the first phase of the 1000GP. The Global Alliance for Genomics and Health (GA4GH) coalition led a Human Genome Variation Map pilot project that built reference graphs for five polymorphic regions: the MHC, the killer cell immunoglobulin-like receptors (KIR) locus, the spinal muscular atrophy locus, and the *BRCA1* and *BRCA2* genes (Novak et al. 2017). The submitted graphs were constructed using at least two alternate sequences, and some additionally contained 1000GP variants.

1.4.3 Alignment and variant calling using graph-based references

Mapping software used with linear references is unable to map reads to graph-based references due to the diverging paths and numerous ways the nodes can be traversed. The actual algorithm for aligning a linear sequence to a graph structure was described 20 years ago by Lee et al. (2002) in their partial-order alignment paper, but the difficulty lies with finding the optimal mapping of a read given a large reference graph. Graph-aware aligners have been developed to tackle this task by using graph-friendly indexes for the seed step and by prioritizing which paths to explore during the extend step. Graph-based read aligners include the aforementioned `vg` (Garrison et al. 2018), `minigraph` (Li et al. 2020), and `GraphAligner` (Rautiainen and Marschall 2020). Alignments are generally stored as **graph alignment/map** (gam) or **graph alignment format** (gaf) files, which are similar to bam and sam files in terms of layout but include information on the series of nodes to which a read aligns. Mapping to a reference graph has been shown to result in a higher percentage of perfectly mapped reads compared to mapping to the current linear reference (Garrison et al. 2018).

Using a graph-based reference structure should reduce reference bias in variant calling because multiple alleles of common variants throughout the human population are incorporated. The fundamental variant calling approach is the same: identify locations where reads contain alleles not present in the reference graph. The software `gramtools` handles nested variants, such as SNPs within different MHC haplotypes. Similar to `BWBBLE` (see section **1.4.1**), `gramtools` encodes known variants into the linear reference, but adds the variants in place and marks them with positional anchors instead of adding them to the end of the reference sequence (Maciuca et al. 2016). `vg` performs SNP and SV variant calling and genotyping by generating a read pileup at each position and noting sites of potential variants not represented in the graph. The reference graph is augmented with these variants and, with enough evidence from the pileup, genotypes are emitted for each site of variation, along with genotypes for heterozygous reference sites (e.g. reads align to both SNP alleles represented in the graph). A reference path must be declared in the graph to assign a reference allele at each site of divergence. An alternate allele is called if the reads have evidence of a variant not present in the reference (Novak et al. 2017).

Numerous reference graphs built for specific genomic regions or for the entire human have genome resulted in improved read mapping, including perfectly mapped reads, reduced reference bias, and improved variant calling (Dilthey et al. 2015; Novak et al. 2017; Hickey et al. 2020; Li et al. 2020; Sirén et al. 2021; Hickey et al. 2022; Liao et al. 2022; Sibbesen et al. 2022). These improvements are particularly seen at the continental population level, indicating a reduction in ethnic bias due to lack of representation in GRCh38 (Novak et al. 2017). In addition, using a reference graph and alignment with `vg` was found to be very effective at reducing both SNP and indel reference bias in ancient DNA that might occur due to DNA degradation and base conversion over time (Martiniano et al. 2020). Pangenomes have also allowed better SV detection in the highly complex genomes of many plant species (Cao et al. 2011; Alonso-Blanco et al. 2016; Gao et al. 2019). Despite all of the genomic variants that have been curated from hundreds of thousands of individuals, identifying new variants still largely relies on reference-based methods. Recent research suggests taking advantage of accurate long reads and generating reference graphs during genome assembly as an effective way to include relevant SVs instead of using reference-based variant calling methods that can introduce biases to the variants included in a reference graph generated from the reference sequence and a list of variants (Hickey et al. 2022).

1.4.4 Reference graph considerations

Computational resources increase with larger reference graphs, particularly when variants are close in proximity. In the GA4GH graph comparison study, **graph compression**, the ratio of the length of the reference sequence to the total length of the graph, was positively correlated with variant calling accuracy (Novak et al. 2017). Therefore, the more additional variants present in the graph, the poorer variant calling became due to issues with accurately mapping reads. Incorporating all known variants into a genome graph is not desirable due to increasing mapping ambiguity and computational resources needed to store and analyze the graph (Pritt et al. 2018). A high density of variants in a graph can make graph index generation unfeasible by `vg`, particularly as the number of possible traversals through the graph grows exponentially with the number of sites of variation (Garrison et al. 2018; Garrison 2019). The `vg` developers instead suggest **pruning** the graph to remove some of the densely packed variants and replacing those sequences with the linear reference alleles. It has been recommended that only

variants with at least a 0.1% minor allele frequency be contained within a graph to reduce the effects of having too many variants (Garrison et al. 2018; Martiniano et al. 2020).

Reference graphs may contain cyclic regions as a concise representation of genomic repeats. Since the algorithms used in partial-order alignment are unable to handle cycles, this restricts many graph aligners to working on only acyclic reference graphs. `vg`, however, is able to handle alignment to cyclic regions by transforming the region into a directed acyclic graph and performing alignment on the linearized graph subset (Garrison et al. 2018). This is done by copying the nodes involved in the cycle a sufficient number of times to allow for seed mapping (Garrison 2019).

As with the proposed alternative reference genomes, determining a standard coordinate system for a graph reference will not be trivial. How will coordinates be decided when large SVs add and remove sequence from certain graph traversals? Moving to a graph-based reference could use GRCh38 as a backbone, allowing for some familiarity with coordinates and genomic annotation (Schneider et al. 2017).

1.5 The polymorphic repetitive gene *PRDM9*

Given that half of the human genome is derived from repetitive TEs, there are numerous regions that could benefit from improved representation during genome resequencing and variant calling analyses. I chose to focus on the gene *PRDM9* as a proof-of-concept for my thesis work on improving variant calling for repetitive and polymorphic regions of the genome. The gene has a fascinating repetitive structure and dozens of alleles have been observed with varying population-specific frequencies. Additionally, important clinical implications in cancer and other disorders have previously been described with regards to the functional effects of *PRDM9*.

1.5.1 The role of *PRDM9* in meiotic recombination

PRDM9 (PR/SET Domain 9) is a member of the PRDM family of transcriptional regulation genes (Fumasoni et al. 2007). The gene is composed of a PR/SET domain with epigenetic methyltransferase functions, a KRAB domain for protein-protein interactions, and a zinc finger array that binds DNA (**Figure 1.8 A**) (Ponting 2011). The PRDM9 protein is an important contributor to chromosomal recombination during meiosis, which does not occur at random locations throughout the genome but rather clusters in regions called **recombination hotspots** (Myers et al. 2008; Pratto et al. 2014). The zinc finger array in the most frequent allele, allele A, binds to the 13bp DNA motif CCNCCNTNNCCNC, which is overrepresented at recombination hotspot locations (Myers et al. 2008). The PR/SET domain is known to trimethylate both K4 and K36 in histone 3 (Powers et al. 2016), opening chromatin and leading to the recruitment of proteins for recombination, including the topoisomerase SPO11 (Romanienko and Camerini-Otero 2000; Pratto et al. 2014). Finally, the KRAB domain is believed to be implicated in the recruitment and binding of proteins that spatially align the DNA hotspot with recombination proteins (Parvanov et al. 2017). With the protein machinery in place, double stranded DNA breaks are initiated by SPO11 and the process of recombination begins (**Figure 1.8 B**) (Pratto et al. 2014).

While the PR/SET and KRAB domains are conserved across *PRDM9* alleles, the DNA-binding domain is variable in both the number and the composition of individual zinc finger repeats. The tandem C2H2 zinc finger repeats, each 84bp long, are similar in sequence owing to the cysteine and histidine amino acids that are crucial in forming the projecting finger structure (**Figure 1.8 C**). However, there is variability in the individual zinc finger sequences, particularly in the region that corresponds to the amino acids that physically bind DNA (**Figure 1.8 D**). The major allele has 13 zinc fingers, and the zinc fingers at positions 8–12 are typically responsible for determining the DNA binding motif (Berg et al. 2010).

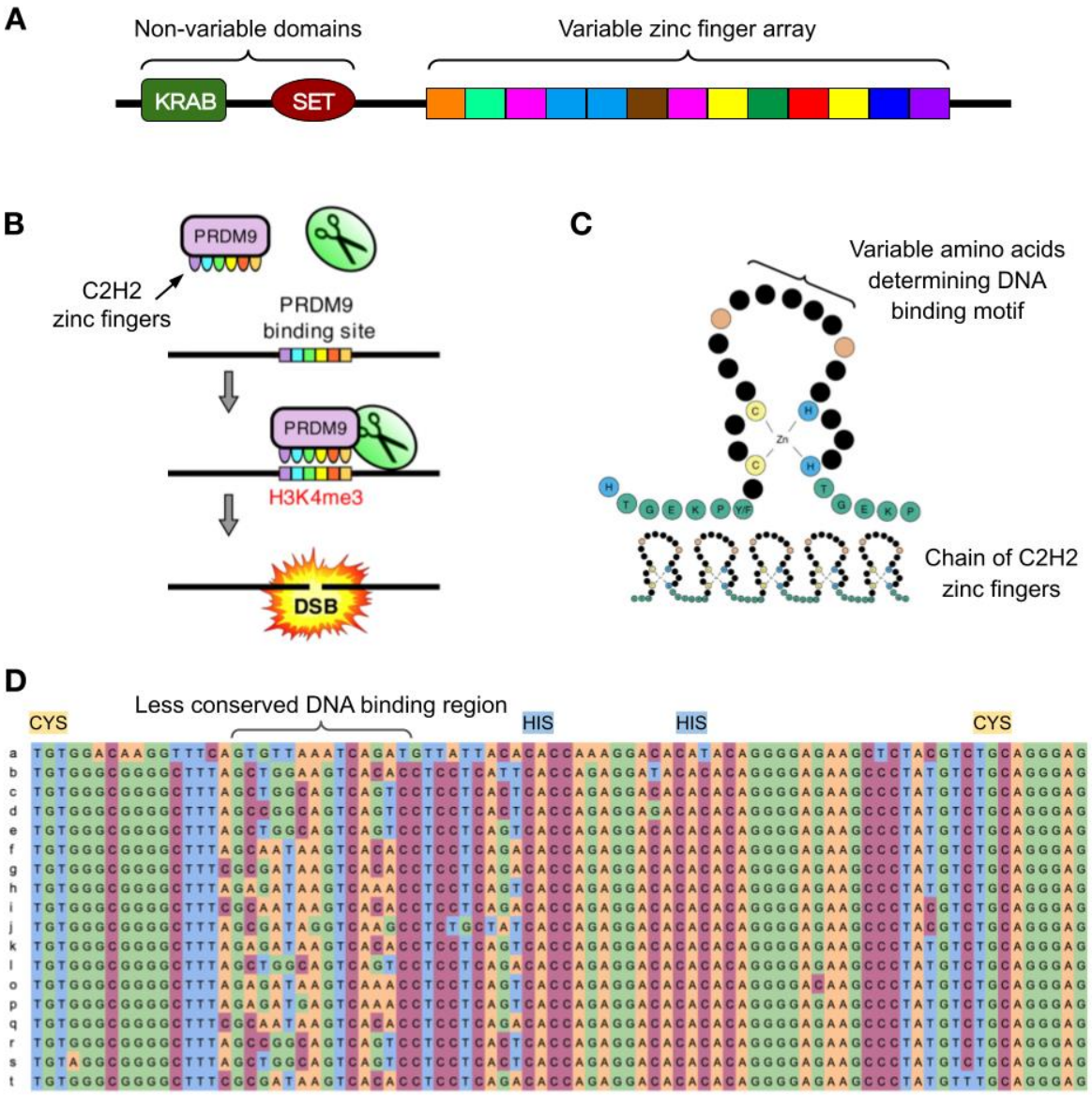


Figure 1.8: The role of *PRDM9* in initiating meiotic recombination. A) The three functional domains of *PRDM9*: KRAB, recombination protein binding and recruitment; SET, trimethylation of HK4 and HK36; zinc finger repeat, DNA binding. Each colored block in the zinc finger array represents a different zinc finger repeat in D). **B)** Zinc finger repeats determine the binding sites of PRDM9 proteins, where the PR/SET domain induces histone trimethylation to recruit and initiate proteins involved in double-stranded DNA breaks and recombination during meiosis. Figure modified from Brick et al. (2012); reproduced with permission from Springer Nature. **C)** Peptide structure of C2H2 zinc finger repeats. Figure modified from Knight and Shimeld (2001); reproduced with permission from Springer Nature. **D)** DNA sequences for zinc finger repeats from the alleles described by Berg et al. (2010). Locations of conserved cysteine (yellow) and histidine (blue) amino acids and the less conserved DNA binding region are indicated above the sequences.

1.5.2 *PRDM9* variability and evolution

Dozens of alleles have been identified since it was first determined that *PRDM9* played an important role in meiotic recombination (Oliver et al. 2009; Thomas et al. 2009; Baudat et al. 2010; Berg et al. 2010; Kong et al. 2010; Parvanov et al. 2010; Berg et al. 2011; Ponting 2011; Borel et al. 2012; Hussin et al. 2013; Jeffreys et al. 2013; Vergés et al. 2017; Alleva et al. 2021; Beyter et al. 2021; Wang et al. 2021). However, the majority of research on DNA binding motifs and population frequencies has been focused on alleles described by Baudat et al. (2010) and Berg et al. (2010). Among these 29 alleles, there are between eight and 18 zinc finger repeats in the variable DNA-binding domain (**Figure 1.9**). The major allele, A, has a frequency of 86% in Europeans and 50% in Africans, and there is greater allelic diversity overall in Africans compared to Europeans (Berg et al. 2010). As mentioned earlier, allele A binds to the DNA motif CCNCCNTNNCCNC; allele C, on the other hand, binds to the motif CCNCNNTNNNCNTNNC (Kong et al. 2010), leading to a different recombination map in individuals with C alleles, namely people of African descent (Kong et al. 2010; Hinch et al. 2011).

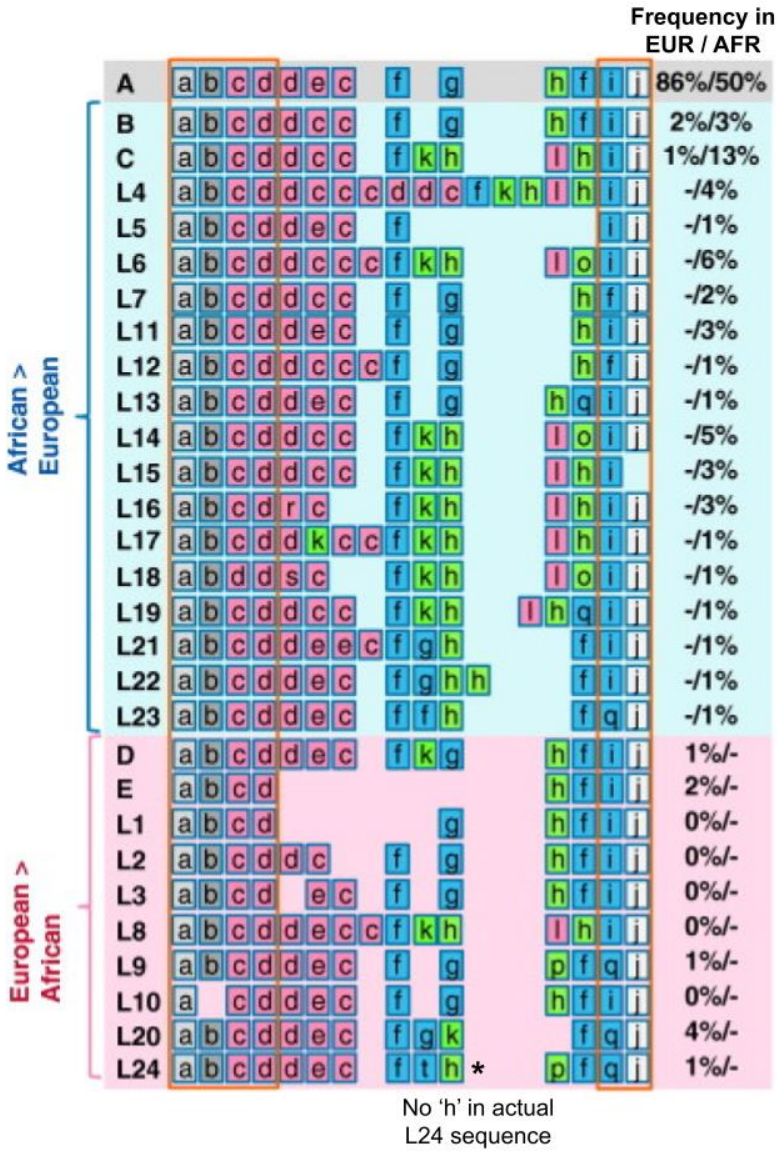


Figure 1.9: Comparison of *PRDM9* allele frequencies in European and African populations. Alleles (names on the left) are depicted by labeled boxes indicating the composition of the zinc finger domain. Allele A (grey highlight) is the most frequent in both populations, though with a much higher frequency in Europeans (left percentages) than Africans (right percentages). The remaining alleles are grouped by those with higher frequencies in Africans (blue box) or Europeans (pink box). Africans have greater allelic diversity than Europeans. Figure modified from Ponting (2011); reprinted with permission with Elsevier. Figure uses allele names and zinc finger content from Berg et al. (2010). *Note: Allele L24 is incorrectly depicted in this figure and should not have an 'h' zinc finger repeat.

PRDM9 is one of the fastest evolving genes in the human genome, with changes occurring to the zinc finger repeat domain that lead to new alleles while the PR/SET and KRAB domains remain conserved (Ponting 2011). Evolution of the gene has been so rapid that the chimpanzee, the closest relative to humans, has a different recombination landscape (Ptak et al. 2005), and several species lack a functional copy of the gene entirely, including dogs (Oliver et al. 2009). There is evidence that the zinc finger array has experienced positive selection, deletions, insertions, and gene conversions, but it is not yet fully understood how *PRDM9* is evolving so rapidly (Ponting 2011). Individual hotspots are lost over time likely through **gene conversion**; when recombination occurs at sites heterozygous for a hotspot and a coldspot, the coldspot DNA can be used as a template to repair the double-stranded break within the hotspot, resulting in the conversion of the original hotspot sequence to that of the closely homologous coldspot. The persistence of recombination hotspots within the human genome, instead of their depletion, suggests there are changes in the motifs to which PRDM9 binds that results in new hotspots. Recent research suggests new *PRDM9* alleles result in a limitation of the total number of binding sites for more effective use (Baker et al. 2022). Strong recombination hotspots are lost more quickly than sites of weak PRDM9 binding, which leads to a reduction in the number of strong binding sites and an increase in weaker binding sites for a given *PRDM9* allele. The number of symmetric double strand breaks also decreases, increasing selection against a given allele. New *PRDM9* variants act against gene conversion by restoring binding sites, which results in more symmetrical binding to a smaller genomic proportion, which is suggested to drive *PRDM9* evolution. Gene conversion is also believed to be responsible for the mutability of minisatellites, in part due to template switching of the DNA strand used for replication (Jurka and Gentles 2006). A single template switch during mitotic DNA replication or repair can result in the formation of novel somatic *PRDM9* variants, whereas the more complex novel variants observed in sperm cells involve multi-step template switching events, often requiring both parental alleles (Jeffreys et al. 2013; Alleva et al. 2021). Alleva et al. (2021) classified 71 *PRDM9* alleles as either A-type or C-type based on the similarity of predicted binding sites for each allele to the known binding motifs for alleles A and C. A-type alleles tended to be around the same length as allele A, and likewise for C-type alleles and allele C. Running template switching simulations showed that the generation of a novel A-type allele from two C-type parental alleles or vice versa is highly unlikely due to the large number of switches that would

be required, suggesting that mutational drift of two original alleles resulted in the variants observed across human populations today.

Despite a much lower frequency than the major allele A, GRCh38 represents *PRDM9* with allele B (a prime example of a reference minor allele described in section **1.2.2**). While there is only a single SNP difference between the two alleles, this results in the majority of samples having a mismatch in reads that align to the zinc finger domain. Recently, the GRC included a novel patch scaffold for allele A in their GRCh38.p14 release (NCBI Reference Sequence ID NW_025791779.1). This can potentially improve read alignment and variant calling if the alternate loci and patch scaffolds are taken into consideration, but given the different lengths and SVs of the numerous alleles, perhaps *PRDM9* can be better represented with a graph-based reference.

1.5.3 Clinical implications of *PRDM9* variants

PRDM9 variants have been associated with numerous clinical disorders both in patients and in the parents of affected individuals. It has been observed that variant *PRDM9* alleles are significantly increased in the mothers of children with Trisomy 21 of maternal origin (with self-reported Caucasian ancestry), suggesting a possible association between variant alleles and nondisjunction (NDJ) of chromosome 21 (Oliver et al. 2016). Given a known association between reduced or absent recombination and NDJ in chromosome 21 (Sherman et al. 1991; Oliver et al. 2008), the reduction in recombination may be associated with fewer binding sites for non-A alleles. L24 and L9 were the most frequently observed alleles associated with meiosis I NDJ of chromosome 21 with no recombination on the chromosome, and both alleles had fewer predicted binding sites on chromosome 21 (Oliver et al. 2016). Relatedly, a nonsynonymous point mutation within the zinc finger array was found in a study of infertile Japanese men (Irie et al. 2009).

Two studies have found an overabundance of non-A *PRDM9* alleles in the parents of children with acute lymphoblastic leukemia, particularly for alleles with a ‘k’ zinc finger repeat that is absent from the major allele A (Hussin et al. 2013; Woodward et al. 2014). For pediatric cancers, inheriting cancer-inducing mutations is unlikely as such mutations that cause disease

early in life would be expected to be selected against and removed from the population (Hussin et al. 2013). Finding variant *PRDM9* alleles in parents instead points to the role of parental genetics in the development of pediatric cancers, which has been previously observed in childhood leukemogenesis (Greaves 2006). Despite its meiosis-specific function, *PRDM9* expression has been observed in some cancer cell lines (Feichtinger et al. 2012). A large-scale exploration of cancerous tissue from the PanCancer Analysis of Whole Genomes (PCAWG) project uncovered aberrant *PRDM9* gene expression in tissues from 32 types of cancer (Ang Houle et al. 2018). Additionally, cancer samples expressing *PRDM9* had an enrichment of somatic breakpoints near recombination hotspots. Neither study that observed *PRDM9* expression in cancer attempted to determine which *PRDM9* alleles were present in the samples with aberrant gene expression, but elucidating the *PRDM9* genotypes could provide valuable information about potential variant-specific cancer associations.

Importantly, clinical implications are not limited to rare *PRDM9* alleles. Allele A was found to be the most frequent allele in healthy parents of neurofibromatosis type 1 patients, a disorder caused by small- or large-scale deletions of the *NF1* gene on chromosome 17 (Hillmer et al. 2016). The binding motif for allele A has been found near sites of non-allelic homologous recombination, which is believed to follow the same mechanisms as allelic homologous recombination (Dittwald et al. 2013), including locations near *NF1*. Sperm cells with variant *PRDM9* alleles were observed to have a reduced frequency of unequal meiotic exchanges associated with the disorder Charcot-Marie-Tooth type 1A, suggesting a protective function of variant alleles against the disease (Berg et al. 2010). DNA breakpoints associated with Hunter syndrome and with Potocki-Lupski/Smith-Magenis syndrome are also located near *PRDM9* A-allele binding motifs (Pratto et al. 2014).

Fully understanding the zinc finger array composition of variants is important with regards to clinical treatments for diseases associated with *PRDM9*. Most studies examining the clinical impacts of *PRDM9* assume the A allele binding motif is in use, though the C allele binding motif is also assessed in studies with samples known to have the C allele. If PRDM9 is confirmed to have direct implications in genome instability in tumor tissue by means of protein expression and DNA binding experiments, the gene or protein is in a unique position to act as a lower-risk target for therapeutic intervention owing to the meiotic-specific expression of the

gene. Reducing *PRDM9* expression should be limited to tumor cells in females and tumor and testes cells in males since non-gametic cells typically do not express *PRDM9*. However, if *PRDM9* variants are to be fully examined for clinical implications and therapeutic targets, a better genotyping approach needs to be developed owing to the difficulty of studying the polymorphic repetitive gene with short-read sequencing data.

1.6 Thesis rationale

Though the advancement of long-read sequencing technologies has improved variant calling for large SVs and repetitive regions of the genome, it will take time for research practices to fully adopt the technology and perform wide-scale sequencing projects on large population cohorts. Despite the shortcomings of short-read sequencing for SV detection, there is an abundance of short-read data currently available for hundreds of thousands of samples that may never be resequenced with PacBio or ONT technology. Can applications be developed to use the accurate short-read sequencing data available and better estimate the genotypes of polymorphic repetitive regions of the genome? For samples that have been long-read sequenced, can the process of identifying precise known variants be simplified into a single software application? Can a flexible reference genome that models known variants be used to improve read mapping to repetitive regions of the genome? My thesis aims to develop and assess three methodologies to address these questions using the *PRDM9* variable zinc finger repeat domain as a proof-of-concept region.

For **Chapter 2**, I developed two genotyping models for use with short-read sequencing that make use of *k*-mer information instead of precise alignment of full-length reads. I describe the process of developing the models and testing them with simulated sequencing data for both haploid and diploid genotyping of known *PRDM9* alleles. I tested the models on three real samples with easy-to-discern genotypes and described the issues encountered with higher sequencing error rates within the *PRDM9* variable domain.

For **Chapter 3**, I developed two genotyping models for use with long-read sequencing data that make use of the haplotype information preserved within long reads. These models rely on

the realignment of and consensus sequence generation from long reads that span the full variable zinc finger repeat array of *PRDM9*. Using 101 samples that have been both short- and long-read sequenced, I assessed the performance of the long-read models on genotyping *PRDM9* and then used these genotypes to validate the results from my short-read genotyping models.

For **Chapter 4**, I constructed and described five different topologies for reference graphs containing known *PRDM9* variants and assessed the graphs in terms of how well they represent the known alleles. I aligned simulated and real short-read sequencing data to the reference graphs and assessed how well the reads mapped to each graph compared to mapping to the GRCh38 reference.

Finally in **Chapter 5**, I summarize my findings and outline the computational models I have developed and the data I have curated. Possible avenues of further research exploration are discussed, and I finish by outlining best-practice recommendations for sequencing polymorphic repetitive regions of the genome such as *PRDM9*.

Chapter 2

Probabilistic models that use *k*-mer information for short-read genotyping in repetitive polymorphic genomic regions

2.1 Background

Identifying genomic variants typically involves alignment of sequencing reads to a reference genome, followed by using software to identify differences between the reads and the reference. `GATK` Best Practices for identifying short germline variants (SNVs and indels) in a set of samples currently dictate read mapping, marking duplicates, recalibrating base quality scores, calling variants for each sample, consolidating all of the variants, performing joint genotyping on all samples, and finally filtering the variant calls (Van der Auwera and O'Connor 2020). `HaplotypeCaller`, a Hidden Markov Model (HMM) approach, is the tool recommended by `GATK` to call these short variants, but numerous other callers exist, such as `FreeBayes` (Garrison and Marth 2012) and `Platypus` (Rimmer et al. 2014). Variant callers have also been developed specifically to identify CNVs (Boeva et al. 2012; Li et al. 2012; Fowler et al. 2016; Johansson et al. 2016) and SVs (Iqbal et al. 2012; Rausch et al. 2012; Layer et al. 2014; Chen et al. 2016). Subsequent software to annotate the identified differences can then be used to identify known variants, such as SNPs and indels from databases like dbSNP (Smigielski et al. 2000), GENCODE (Harrow et al. 2012), ClinVar (Landrum et al. 2018), and gnomAD (Karczewski et al. 2020). While this mapping-based approach is highly effective for most regions of the human genome, some areas remain problematic. Repetitive regions of the genome are prone to reduced read mapping due to there being several locations to which a read from a repeat-rich locus can align. Such multimapping reads are often ignored in downstream analyses since their exact location in the genome cannot be determined.

The variable zinc finger domain in the gene *PRDM9* is highly repetitive and polymorphic, resulting in reduced mapping and difficulty aligning short sequencing reads owing to their inability to span the full repetitive region. For reads that do manage to be mapped, due to the numerous different arrangements of its zinc finger array, genotyping *PRDM9* in the traditional manner is a twofold problem: the counts and locations of SNVs and indels first need to be

identified, and then these variants need to be cross-referenced with the variants from a list of known alleles. Instead, *PRDM9* is typically genotyped via realignment to known zinc finger or allele sequences after Sanger sequencing (Berg et al. 2010, 2011; Jeffreys et al. 2013; Hussin et al. 2013) or, more recently, after PacBio HiFi sequencing (Alleva et al. 2021). In both cases, the long sequences allow for a clear understanding of where zinc fingers have been inserted or deleted, and ambiguity about aligning these reads to the reference genome is greatly reduced. However, Sanger sequencing is time consuming and labor intensive and numerous datasets have already been short-read sequenced; resequencing these samples with either Sanger or PacBio HiFi sequencing is not realistically feasible. It would therefore be incredibly useful to have an accurate method of genotyping *PRDM9* from Illumina data, but the repetitive nature of the gene makes accurate mapping of reads to the reference genome difficult.

***k*-mers** are used in many aspects of genomic analysis. During de novo genome assembly, *k*-mers can be used to construct de Bruijn graphs, from which contigs can be determined (Sohn and Nam 2018). *k*-mers are useful because the shorter lengths mean many *k*-mers will not have sequencing errors that are present in the reads. There are several tools that identify rarely occurring *k*-mers, which are likely the result of sequencing errors, and apply corrections to sequencing reads containing these *k*-mers to improve accuracy of downstream analyses (Kelley et al. 2010; Marçais and Kingsford 2011). Understanding the distribution of *k*-mer counts is thus particularly useful. *k*-mer counts are often used to estimate genome size (Kurtz et al. 2008), to differentiate organisms in metagenomic analyses without the need for de novo assembly (Dubinkina et al. 2016), and to annotate and identify genomic repeats such as transposons (Price et al. 2005; Kurtz et al. 2008) and CNVs. Additionally, *k*-mer count profiles have been used in the RNA-sequence quantification software `Sailfish` (Patro et al. 2014) and `kallisto` (Bray et al. 2016) to avoid computationally heavy read alignment by comparing counts of *k*-mers found in reads and in RNA transcripts.

The differences in zinc finger repeat copy numbers among different *PRDM9* alleles is somewhat akin to identifying TE repeats, and the differentiation of similar zinc finger repeats is akin to estimating the abundance of related RNA transcripts. Given that there are *k*-mer counting tools for both of these analyses, I hypothesized that *k*-mer counts from Illumina sequencing data could be used to determine *PRDM9* genotypes. In particular, the zinc finger

copy number differences that differentiate many *PRDM9* alleles suggested that using a *k*-mer counting method would be a better approach than an HMM-based approach like `HaplotypeCaller`. In this chapter I detail the development of two *k*-mer based models to genotype polymorphic repetitive regions of the genome from short-read sequencing data, using the *PRDM9* zinc finger array as a proof-of-concept locus. One model uses *k*-mer count information in order to differentiate copy-number differences of zinc finger repeats amongst different *PRDM9* alleles. The other model uses *k*-mer distance information to differentiate the arrangement of specific zinc finger repeats. Both models are compared, as well as combined, by genotyping simulated short sequencing reads from known *PRDM9* alleles. In addition, the models are tested on three real sequencing data sets with easy-to-discern genotypes.

2.2 Results

2.2.1 Collecting all known *PRDM9* allele variants

Zinc finger and allele sequence information was collected for 36 *PRDM9* alleles described in a review article by Ponting (2011) and from a publication by former lab member Hussin et al. (2013). This list of known alleles is referred to as the ***PRDM9*-36** list (**Figure 2.1**). The allele list was later found to contain an incorrect sequence for allele L24 due to a typo in the review relative to the information from the original article (Berg et al. 2010). The *PRDM9*-36 list therefore contains 35 correct allele sequences from Berg et al. (2010) and Hussin et al. (2013) and one sequence that has not actually been observed in humans. The list is also missing one allele described by Berg et al. (2010). Since this list was primarily used to generate simulated sequencing data for experimentation, I was not concerned about results including the incorrect allele sequence. Additional alleles were later curated from a variety of publications and from the **NCBI Nucleotide database** (https://www.ncbi.nlm.nih.gov/nucleotide/), generating a larger list of 106 alleles hereafter referred to as the ***PRDM9*-106** list. A study by Jeffreys et al. (2013) identified a large number of alleles observed in sperm cells or as somatic mutations in blood. These additional alleles, which I refer to as **blood/sperm alleles**, were added to the

PRDM9-106 list to generate the all-encompassing ***PRDM9*-642** list of allele sequences. Full details about allele compilation are described in **Methods 2.4.1**.

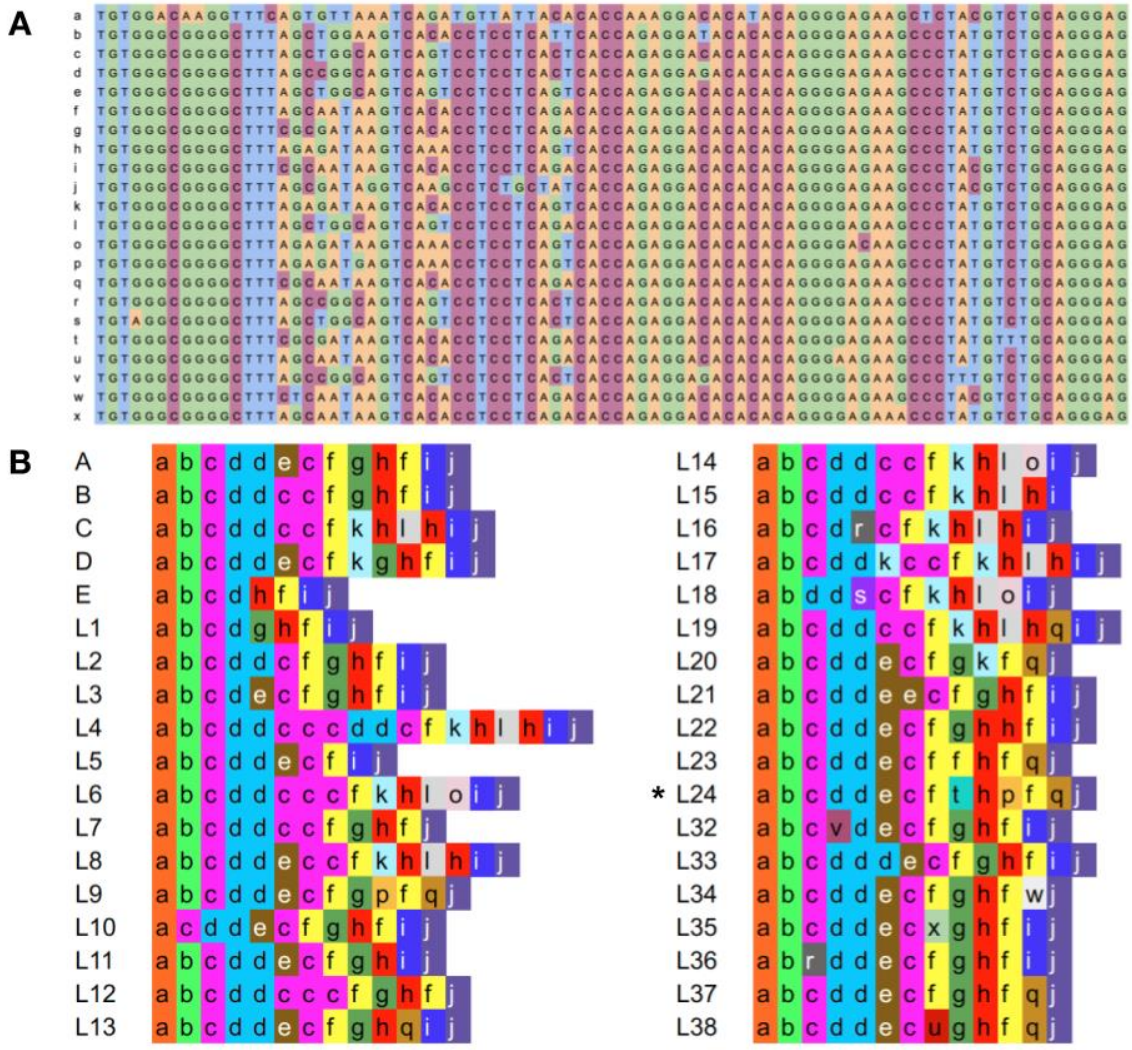


Figure 2.1: *PRDM9* allele variants and zinc finger compositions. A) Multiple sequence alignment of zinc finger repeats found in the *PRDM9*-36 list of alleles. Zinc finger names are lowercase letters to the left of the sequences. **B)** Composition of the zinc finger repeat array in each *PRDM9*-36 allele. Each colored block represents a zinc finger repeat. *Note: the composition for allele L24 incorrectly includes an 'h' zinc finger, a typo from Ponting (2011).

2.2.2 Simulating short-read sequencing data to develop and assess genotyping models

Many sequencing data simulators have been built, allowing for the development and evaluation of bioinformatics software and for validation of predicting genotypes and haplotypes (Escalona et al. 2016). Using simulated sequencing data allows for the assessment of models because the ground truth is known. In this case, simulated data from known *PRDM9* alleles provide a set of sequencing reads with known origins and thus known genotypes. All known alleles and genotypes can be assessed without needing to find real-life samples of every variant, which is particularly difficult given the rarity of most *PRDM9* alleles (see **Chapter 1** section **1.5.2**). Using simulations allows me to control the sequencing error rate, depth of coverage, and read and fragment lengths of the data. Since the goal is to use *k*-mer information from the short reads, it is important to understand the effect of increased read coverage on the practicality of *k*-mer counts. In addition, controlling the sequencing error rate allows me to examine the loss of informative *k*-mers. Error-laden *k*-mers will not be observed in the known allele sequences, but an error-rate that is too high might convert a *k*-mer expected in one allele into a *k*-mer only observed in a different allele, adding confusion to the true genotype of a sample. By using different combinations of these properties in simulated data, I can better define the parameters at which my models are effective.

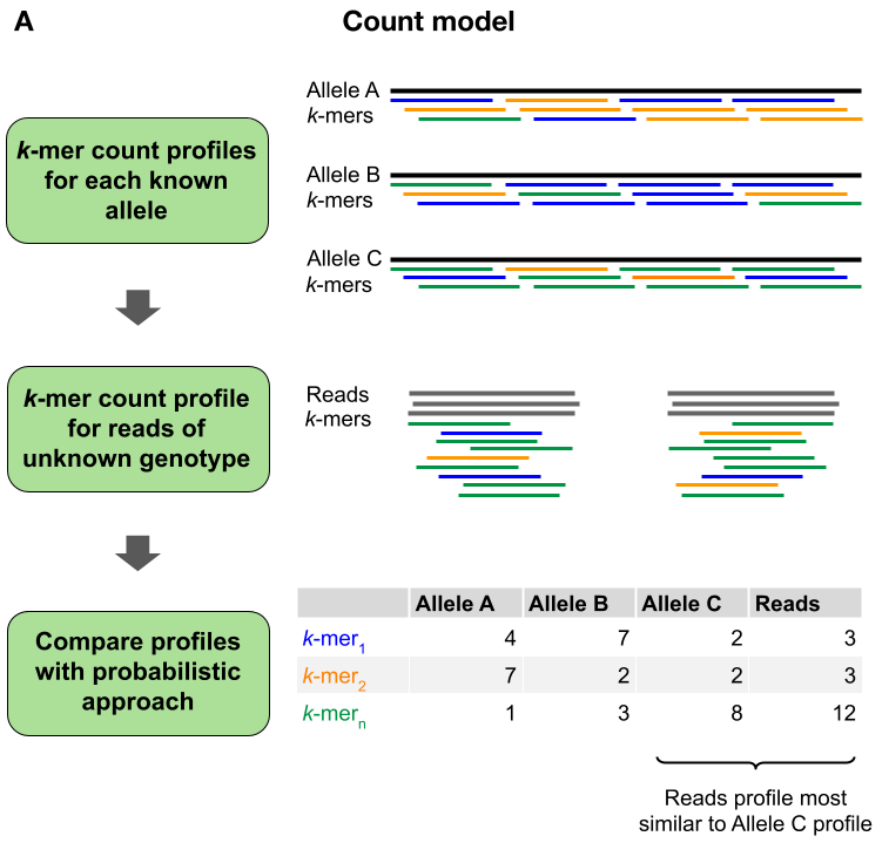
The **primary haploid simulation set** consisted of 100 simulated replicates per *PRDM9*-36 allele for all combinations of coverage (20X, 40X, 60X, 80X, or 100X) and sequencing error rates (0%, 0.1%, or 1%). The different *PRDM9* zinc finger arrays with 10kb upstream and 10kb downstream sequences (***PRDM9* + 10kb flanks**) were used to generate the simulations, which were 100bp paired-end reads with a mean fragment length of 250bp and a fragment length standard deviation of 50bp. The **primary diploid simulation set** was similarly generated; all 666 possible combinations of *PRDM9*-36 alleles were simulated under the same combinations of coverage and error rate. Additional haploid simulation sets were generated at different read lengths and fragment lengths, and without flanking sequences around the zinc finger array for initial model testing (**Methods 2.4.2**).

2.2.3 Developing short-read models to genotype polymorphic repetitive regions of the genome

Two main genotyping models were developed that use *k*-mer information from sequencing reads to call alleles and genotypes: the count model and the distance model. The **count model** (**Methods 2.4.3.1**) compares *k*-mer count profiles between the sequencing reads of a sample and of each known allele provided in a list (**Figure 2.2 A**). It uses a Poisson distribution to do so, modeling the number of times a *k*-mer is observed in the set of sequencing reads, assuming a fixed error rate. The model allows for the calculation of the allele that maximizes the likelihood across all *k*-mers expected from the set of known allele sequences, which is called as the haploid genotype for the sample. To call diploid genotypes, the *k*-mer count profiles from both alleles in a genotype are summed to generate a genotype *k*-mer count profile, and all possible genotypes that can be generated from the submitted list of alleles are assessed. To make the sample *k*-mer count profiles comparable to the allele *k*-mer count profiles, the allele and genotype *k*-mer count profiles are adjusted by a value $\boldsymbol{\lambda}$ to account for the sequencing depth and error rate of the read set. λ can be viewed as the expected ***k*-mer coverage** of the reads; it is a correction for the effect of breaking reads into short *k*-mer segments and for the effect of losing a proportion of *k*-mers to sequencing errors. λ can be estimated from a calculation using sequencing depth of coverage, read length, and the sequencing error rate, which I hereafter refer to as the **coverage** method. For cases where depth of sequencing coverage or sequencing error rate are not known, λ can be estimated from the counts of *k*-mers unique to the 10kb regions flanking the *PRDM9* zinc finger array (flank *k*-mers). Estimating λ from the average or median counts of flank *k*-mers in the reads are hereafter referred to as the **flank mean** and **flank median** methods, respectively.

The **distance model** (**Methods 2.4.3.2**) uses the outermost *k*-mers in a pair of sequencing reads in order to estimate sequencing fragment lengths (**Figure 2.2 B**). Since paired-end reads are sequenced from both ends of a library fragment, locating the outermost *k*-mers from a read in the sequence of an allele gives an idea of how long the fragment is for that read pair. The insertion or deletion of a zinc finger repeat in the allele being tested shifts the observed fragment length away from the expected fragment length. Assuming a normal distribution of

sequencing fragment lengths, the outermost *k*-mers from each read pair are identified in the sequences of all provided alleles, and the distance between the *k*-mer pairs in the allele sequence is used to determine the likelihood of the reads originating from that allele. If there is more than one possible distance between the *k*-mer pair in an allele sequence (i.e. at least one *k*-mer occurs more than once in the allele, such as a *k*-mer occurring in a zinc finger repeat present more than once in an allele), different methods can be used to calculate the likelihood: using the sum of probabilities for all fragment lengths (**sum** method), using the average probability for all fragment lengths (**mean** method), using the geometric mean of the probabilities for all fragment lengths (**geomean** method), or considering only the fragment length that gives the maximum probability (**max** method). The allele that results in the highest likelihood is considered the haploid genotype. For diploid calling, probabilities are determined separately for each allele in a genotype and then averaged to obtain a genotype probability, as is done in `GATK HaplotypeCaller`. Likelihoods are calculated for all possible genotypes that can be formed from the submitted list of alleles.



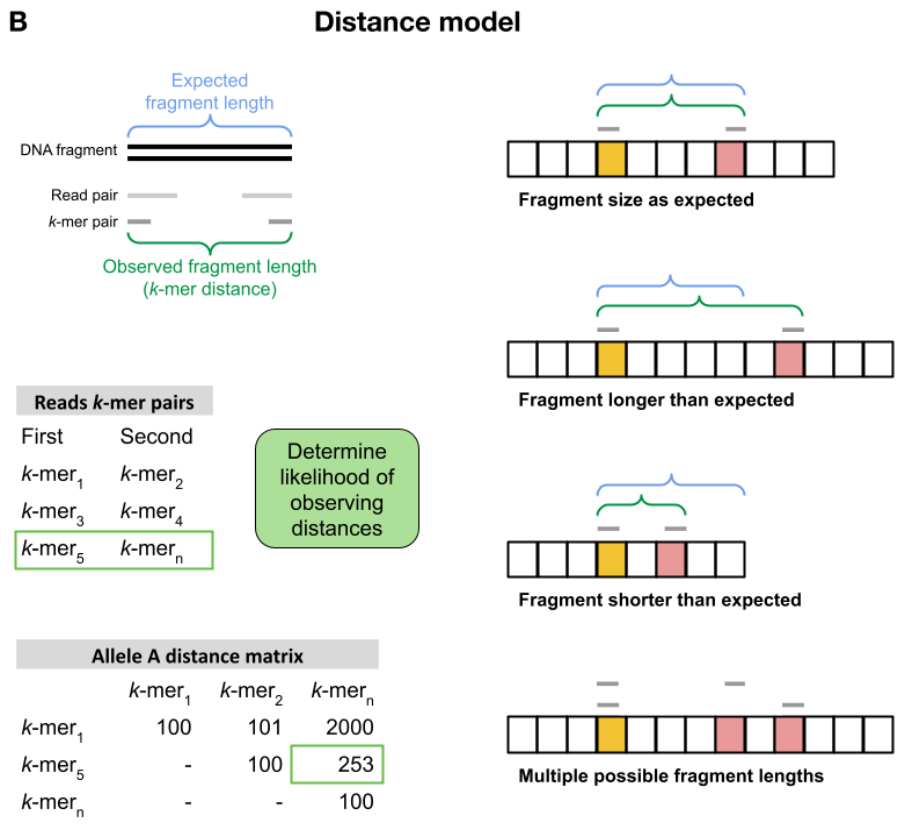


Figure 2.2: The short-read genotyping models. A) The count model. *k*-mer count profiles are generated for all known allele sequences and for the set of sequencing reads from the sample to be genotyped. The sample *k*-mer count profile is compared to the count profile for each allele sequence by modeling the number of times *k*-mers would be observed in the reads given a fixed error rate, sequencing depth of coverage, and read and *k*-mer length. **B)** The distance model. The mean fragment size of sequenced fragments from a sample can be calculated from the distance between read pairs, providing an expected fragment size (blue bracket) for the sample. The distance between the outermost *k*-mers (grey bars) in a read pair can be used as the observed fragment length (green bracket). By locating the *k*-mer pair in the known sequence of an allele, the observed fragment size can be calculated as the inclusive distance between them (from the start of the upstream *k*-mer to the end of the downstream *k*-mer) and compared to the expected fragment length. An observed fragment shorter or longer than expected suggests the arrangement of zinc finger arrays is different from the arrangement in the originating allele. When multiple fragment lengths are possible due to a *k*-mer occurring more than one in an allele sequence, the likelihood is determined via one of four methods that consider either all *k*-mer distances (mean, geomean, and sum methods) or the distance that results in the highest likelihood (max method).

2.2.4 The *k*-mer count model is effective at calling haploid genotypes from simulated reads

The **F1 score** is a measure of accuracy that can be used to assess model performance and is calculated as the harmonic mean of **precision** and **recall**. Accuracy in the traditional sense, the number of true positives and true negatives divided by the total number of observations, is strongly biased when using imbalanced class sizes, whereas the F1 score is not (Shung 2018). With the *PRDM9*-36 list of alleles, there will only ever be one true positive and 35 or 665 true negatives for haploid allele or diploid genotype calling, respectively. The F1 score was therefore chosen to determine accuracy of my genotyping models. Across the 100 replicates for each simulation, average F1 scores were calculated at the allele or genotype level as well as over all alleles or genotypes (**Methods 2.4.3.4**).

Average F1 scores for haploid calling of the *PRDM9*-36 alleles under the **count-coverage** model were above 0.93 for most *k*-mer lengths. Both higher coverage and lower sequencing error rates resulted in higher scores, nearly reaching 1 at 80X and 100X coverage (**Figure 2.3**). This was expected given the importance of having enough reads from deep sequencing coverage to provide adequate *k*-mer counts, as well as the loss of *k*-mers and effective *k*-mer coverage with increased sequencing error rates. F1 scores also peaked using *k*-mer lengths in the high 80s and low 90s, but dropped to near 0 when the length of *k* became very close to the 100bp length of the reads, likely due to the increased probability of containing a sequencing error and the reduced *k*-mer coverage with longer *k*-mers. The values of *k* that produced the highest F1 scores were generally in the 80s for simulations with 0% or 0.1% errors, and in the low 60s to mid 70s for 1% errors (**Appendix Table 2.1**).

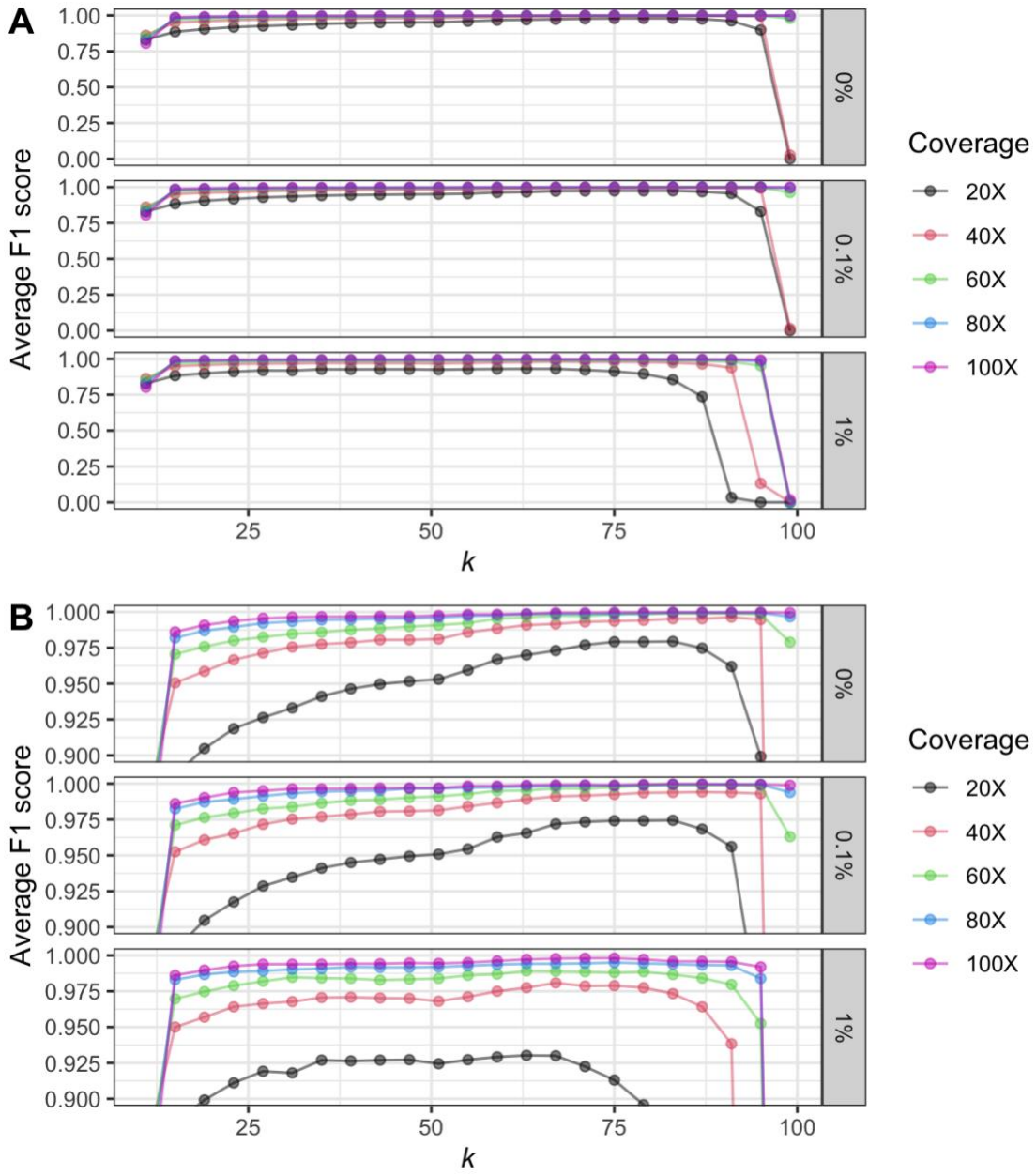


Figure 2.3: Haploid allele calling performance of the count-coverage model. Average F1 scores across all replicates and all *PRDM9*-36 alleles from the primary haploid simulation set, called using the *k*-mer count profile model with the coverage method of estimating λ and tested using different *k*-mer lengths. The simulations were generated with varying coverages (colors) and sequencing error rates (rows). F1 scores improved with higher coverage and lower error rates, but decreased at very high values of *k*. **A**) Full range of F1 scores. **B**) F1 scores over 0.9.

Despite having very high F1 scores when averaged across all alleles, the scores were not consistent when averaged for each allele individually (**Figure 2.4**). Some alleles had high F1 scores across most values of *k* and all coverages, such as L18, L24, L32, L34, L35, and L38, which each have a unique zinc finger ('s', 't', 'v', 'w', 'x', and 'u', respectively). Other high-scoring alleles have unique consecutive zinc finger motifs, such as allele L10 and its unique zinc finger pairings of 'ac'. As well, both the shortest allele E and the longest allele L4, with 8 and 18 zinc fingers, respectively, had consistently high F1 scores. Other alleles had low F1 scores at lower coverages and only peaked at high values of *k*, even under error-free conditions. For allele A, the most frequently observed allele in humans, the highest F1 score for 20X error-free simulations was 0.91 and occurred at $k = 83$.

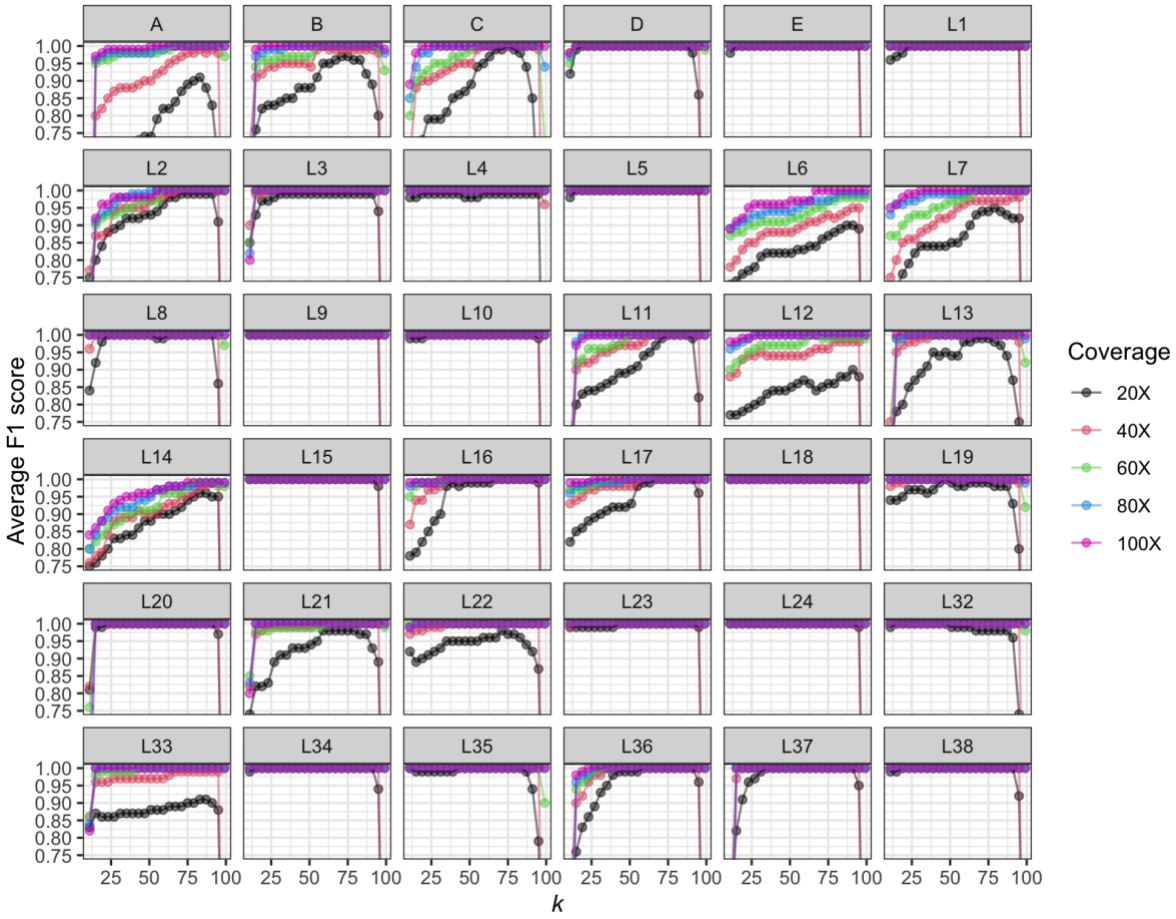


Figure 2.4: Per-allele haploid calling performance of the count-coverage model. Per-allele average F1 scores across all replicates from the error-free subset of the primary haploid simulation set, called using the *k*-mer count profile model with the coverage method of estimating λ and tested using different *k*-mer lengths. The simulations were generated with varying coverages (colors); only error-free results are shown. F1 scores differed amongst the different alleles, with some alleles resulting in high average F1 scores at all coverages and values of *k*, and others resulting in reduced average F1 scores at low coverages and low and very high values of *k*.

2.2.4.1 Effect of read length, fragment length, and flanking sequences on model accuracy

An ideal genotyping model would be effective for many read lengths and would not be biased towards or against longer or shorter sequencing fragments. To assess the robustness of the count-coverage model in this respect, I looked at the effect of read and fragment length for reads simulated from just the zinc finger array of *PRDM9* or from the zinc finger array plus

the 10kb regions flanking the region. Using 100X error-free simulations, 50 replicates of paired-end reads for each *PRDM9*-36 allele were generated for both of the following: read lengths from 50bp to 500bp in increments of 25, and fragment lengths from 200bp to 500bp in increments of 50. For the variable read lengths, all simulations used an average fragment length of 250bp and a standard deviation of 50bp, and for the variable fragment lengths, all simulations used read lengths of 100bp and a fragment standard deviation of 50bp (**Methods 2.4.3.5**).

Surprisingly, shorter reads (**Figure 2.5 A**) and shorter fragment lengths (**Figure 2.5 B**) were strongly correlated with higher F1 scores averaged across all alleles. These biases were stronger at lower values of *k* and became less severe as *k* increased. The fragment length bias was particularly unexpected because the count model treats mates from paired-end reads independently. The length of the allele sequence determines the number of possible starting positions from which reads can be simulated. Longer reads also means fewer reads are necessary to provide the same amount of coverage as shorter reads. Additionally, fewer reads means there is more dependence among *k*-mers, as well as the possibility that not enough *k*-mers are being independently sampled to accurately distinguish alleles. To see if increasing the length of the allele sequences used to simulate the reads reduced the effects of read and fragment length, the simulations with 10kb flanks around the zinc finger array were analyzed. For these simulations, read length had much less of an effect on average F1 scores, and the bias from fragment lengths was nearly gone. By including 10kb flanking sequences around the region of interest, I am able to account for bias in read coverage that originates from simulating reads from a small allele sequence, allowing the count model to function independently of read or fragment length.

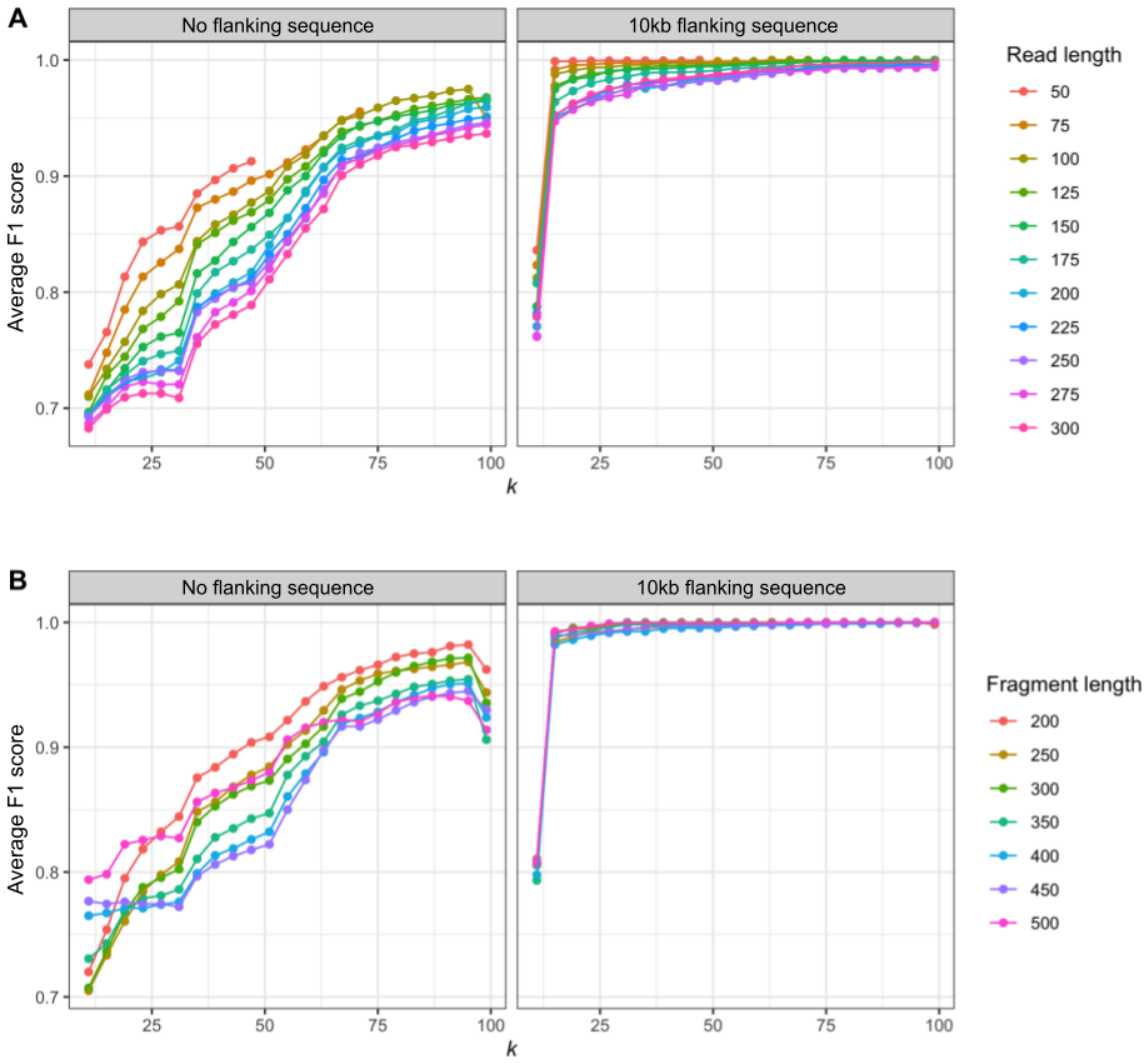


Figure 2.5: Effect of read and fragment length on the count-coverage model performance. Average F1 scores across 50 replicates and all *PRDM9*-36 alleles for 100X error-free simulated haploid reads using different read and fragment lengths and genotyped using the count model with the coverage method of estimating λ and tested using different *k*-mer lengths. **A)** Simulations generated from allele sequences without (left) and with (right) 10kb flanks surrounding the zinc finger array using different read lengths (colors) but controlled 250bp fragment lengths. Without flanks, average F1 scores varied substantially with read length, with a bias of higher F1 scores with shorter reads. With flanks, a small length bias remained at low values of *k*, but it was substantially reduced. **B)** Simulations generated from allele sequences without (left) and with (right) 10kb flanks using different fragment lengths (colors) but controlled 100bp read lengths. Without flanks, average F1 scores varied substantially with fragment length, with a bias towards shorter reads. With flanks, the bias was almost entirely removed.

2.2.4.2 Different methods of calculating λ

The count-coverage model uses the probability mass function of a Poisson distribution to model the *k*-mer count profiles from the sequencing reads, which is parameterized by the allele *k*-mer count profiles and uses the estimated parameter λ, which is in turn dependent on the *k*-mer counts of the allele under consideration. It is easy to calculate λ when the exact coverage and sequencing error rate values are known, as was the case when the values used to simulate the reads were known. However, knowing the exact values is unrealistic when working with real sequencing data. I therefore assessed if λ could be estimated directly from the counts of *k*-mers in the 10kb flanking sequences. Considering only *k*-mers that occur uniquely in the flanking sequences (i.e. have a count of one in the reference sequence and are located entirely within either flanking sequence), both the mean or median counts of these flank *k*-mers that are observed in the sequencing reads were calculated (**Methods 2.4.3.1**). Though these two methods (**flank mean** and **flank median**) performed nearly as well as the coverage method at high simulated coverage rates, the coverage method performed the best overall while the flank median method performed the worst at lower coverages and higher error rates (**Figure 2.6**). The low performance of the median method is largely due to the median *k*-mer count being zero at low coverage using long *k*-mers.

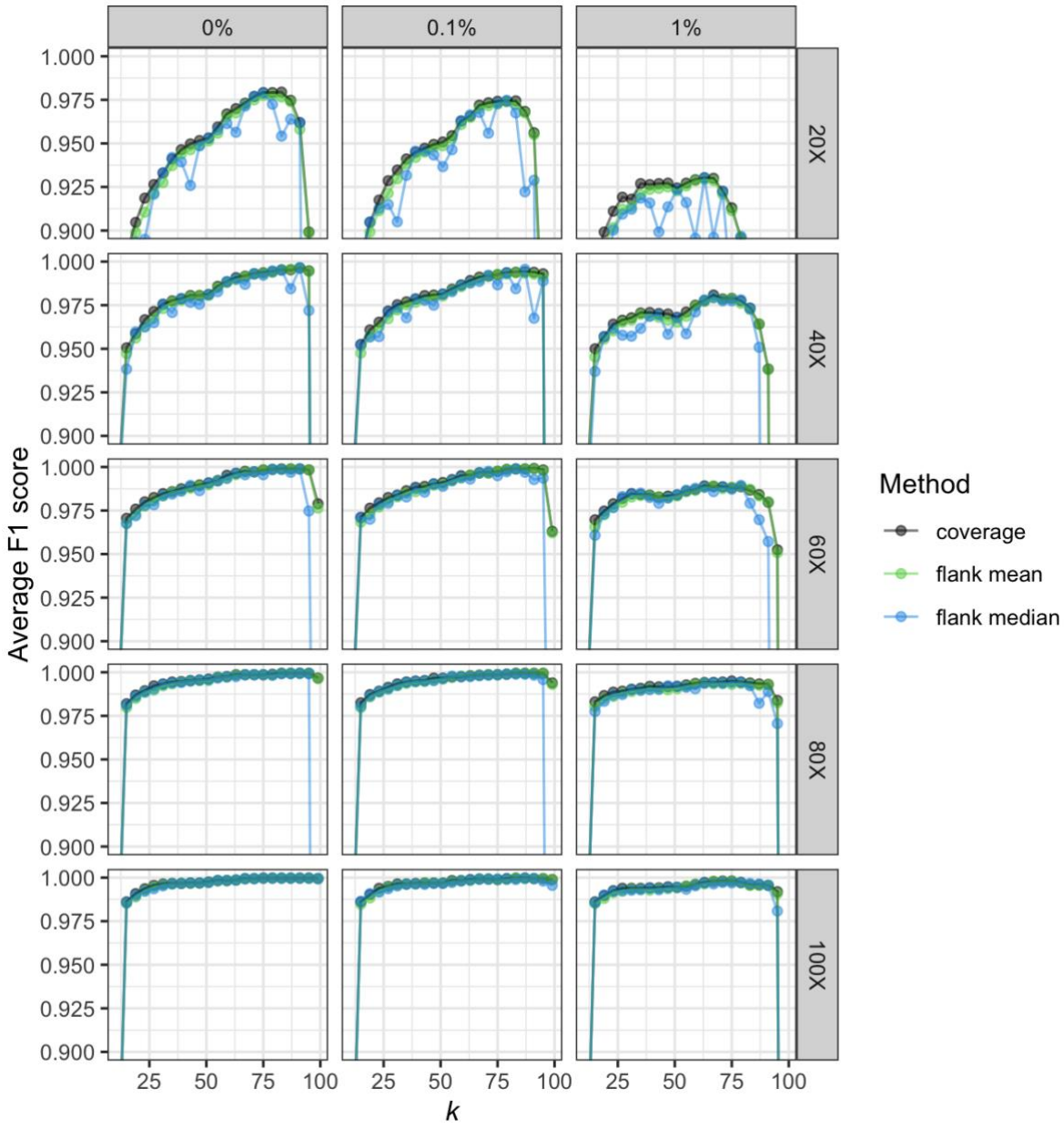


Figure 2.6: Comparison of λ estimation methods used in the count model. Average F1 scores across all replicates and all *PRDM9*-36 alleles from the primary haploid simulation set, genotyped using the *k*-mer count profile model with the coverage, mean flank, or median flank methods of estimating λ (colors) and tested using different *k*-mer lengths. 100 replicates of 100bp paired-end reads at varying coverages (rows) and sequencing error rates (columns) were used. At low coverage, F1 scores were slightly higher for the coverage method than the two flank-related models, and the improvement was more pronounced at higher error rates. The improvement diminished with higher coverage as the methods became comparable.

The software `Cortex` (Iqbal et al. 2012) describes a different way to estimate *k*-mer coverage (λ), which I modified to generate my count-coverage λ estimation formula. The `Cortex` formula was compared to my modification using all replicates of the 20X and 100X simulations from the primary haploid simulation set (**Methods 2.4.3.6**). The original and modified formulas gave identical results for error-free simulations, as expected since an error rate of 0 means the chance of correctly sequencing every base in a *k*-mer is 1 for both formulas (**Methods 2.4.3.1**). The modified formula performed slightly better (though barely noticeable) for the 0.1% error rate simulations and substantially better for the 1% error rate simulations, particularly as the value of *k* increased (**Figure 2.7 A**). The original `Cortex` formula underestimates λ and resulted in lower F1 scores.

It would be convenient if the *k*-mer information from a sample could be utilized without the need to estimate λ at all. Kirk et al. (2018) published a long non-coding RNA classification method that uses the **Pearson correlation** coefficient to determine similarity between a pair of *k*-mer count profiles without consideration of sequencing error rate or depth of coverage. I hypothesized that a modification of this approach could also work for *PRDM9* genotyping. Across all replicates of the 20X and 100X error-free simulations from the primary haploid simulation set, I compared the average F1 scores resulting from using the Pearson correlation to those resulting from using the Poisson distribution in the count-coverage model. At 100X coverage, the Pearson model performed slightly better than the Poisson model at low-to-mid-range values of *k*, but performed comparably at higher values of *k* for low error rates and slightly worse at high error rates. At 20X coverage, however, The Poisson model outperformed the Pearson model at most *k*-mer lengths, only performing worse at the lowest and highest values of *k* (**Figure 2.7 A**).

2.2.4.3 Calling alleles with a pair Hidden Markov Model

`GATK HaplotypeCaller` uses a **pair HMM** to identify SNVs and indels in a haplotype-aware manner (Poplin et al. 2018b). While this has been a successful approach for calling small variants, its utility to call repeat variants is reduced because pair HMMs do not have an innate ability to account for copy number differences, as is desired for repetitive genomic regions. Given the popularity of `HaplotypeCaller`, I assessed how well genotyping the *PRDM9*

zinc finger array would work by using a pair HMM. The HMM and the count-coverage model were used to genotype 50 replicates of 20X error-free reads simulated from the *PRDM9*-36 alleles without 10kb flanks (**Methods 2.4.3.6**). Considering all alleles together, the HMM underperformed substantially relative to the count-coverage model (**Figure 2.7 B**), though there were a few alleles for which the HMM approach was comparable or better (**Figure 2.7 C**). The count model with the modified formula for estimating λ was therefore used for the remaining analyses instead of the original Cortex formula, the Pearson correlation method, or the HMM.

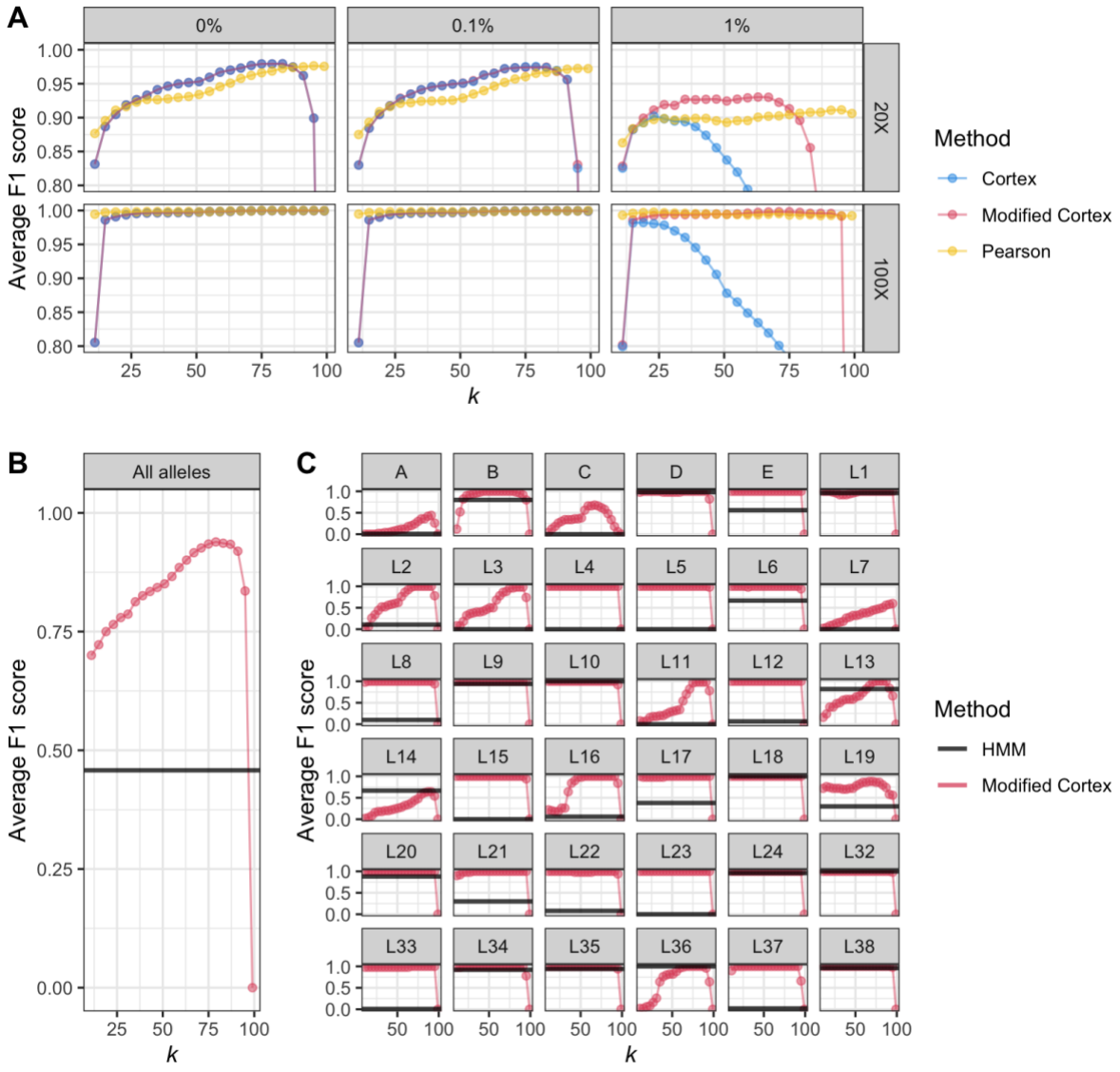


Figure 2.7: Comparison of different methods for determining allele likelihoods. A) Different approaches to calculating allele likelihoods: calculating λ with the original `Cortex` formula, calculating λ with the modified formula used by the count-coverage model, and using a Pearson correlation of the *k*-mer count profiles instead of the Poisson model (colors). Analyses used the 20X and 100X (rows) simulated reads under different sequencing error rates (columns) from the primary haploid simulation set and tested using different *k*-mer lengths. The two λ calculation approaches were identical or comparable at low sequencing error rates, but the modified formula outperformed with the 1% error rate. The Pearson correlation model was comparable to or outperformed the modified formula at high coverage, but underperformed at most values of *k* at low coverage. **B)** Calling alleles using the modified Cortex (count-coverage) model compared to using a pair HMM using 50 replicates of 20X error-free simulations per *PRDM9*-36 allele without 10kb flanks. At all but one value of *k* tested, the modified λ formula approach (red) outperformed the HMM approach (black). **C)** Per-allele F1 scores for the data from B). The HMM approach was comparable to or outperformed the modified λ formula approach for 11 alleles, but severely underperformed for the majority of the alleles.

2.2.5 The *k*-mer count model is unable to distinguish diploid genotypes

Haploid allele calls are not sufficient for use with human sequencing data; a usable model needs to be able to identify diploid genotypes. To call diploid genotypes with the count model, the *k*-mer count profiles from both alleles in a genotype were summed, since the *k*-mer counts from a sample are aggregate data that cannot be accurately split into haplotypes. The read *k*-mer count profile was compared to genotype *k*-mer count profiles for all 666 possible genotypes from the list of *PRDM9*-36 alleles (**Methods 2.4.3.1**). As with the haploid calling assessment, average F1 scores increased with coverage and with higher values of *k*. However, average F1 scores were not nearly as high as they were for the haploid simulations (**Figure 2.8 A**). The highest average F1 score obtained was 0.7481, from the error-free 100X reads at $k = 99$. The 20X reads were only called at a maximum average F1 score of 0.5272 (**Appendix Table 2**).

Many individual genotypes actually had very high average F1 scores across all 100 replicates (**Figure 2.8 B**). For error-free 100X simulations called with $k = 51$, 23.6% of the genotypes had an average F1 score of 0, 70.9% had an average F1 score of at least 0.75, and 21.3% had

an average F1 score of 1. Unfortunately, the most frequent genotype, A/A, performed poorly with an average F1 score of 0.43 even under high coverage error-free conditions.

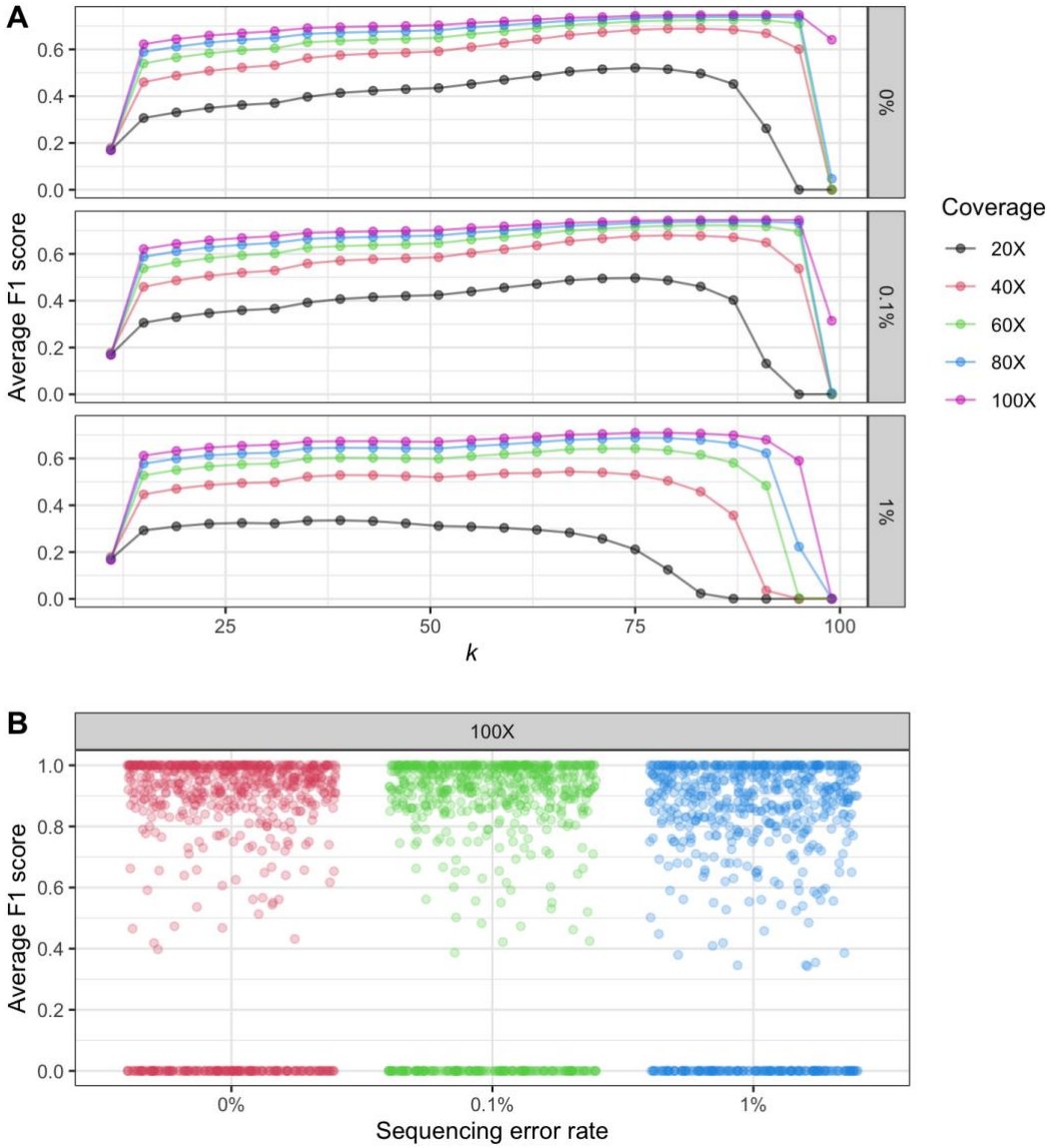


Figure 2.8: Diploid genotype calling performance of the count-coverage model. A) Average F1 scores across all replicates and all 666 possible *PRDM9*-36 genotypes from the primary diploid simulation set, called using the *k*-mer count profile model with the coverage method of estimating λ and tested using different *k*-mer lengths. The simulations were generated with varying coverages (colors) and sequencing error rates (rows). Average F1 scores were substantially lower than they were for the haploid simulations. Compared to the haploid calls, diploid F1 scores increased more gradually with increasing values of *k*, but shared a similar sharp drop as *k* neared the simulated read length. **B)** Average F1 scores per genotype across 100 simulations of 100X coverage simulations, scored using $k = 51$. Though several genotypes had average F1 scores of 0, most had scores of ≥ 0.75, and many had average scores of 1.

Given the huge increase in the number of possible genotypes when moving from haploid calls to diploid calls (36 to 666), it is not surprising that genotyping becomes more difficult. To investigate what might be contributing to the difficulties with calling A/A genotypes, the A/A *k*-mer count profiles were compared to all other genotype count profiles across a large range of *k*-mer lengths. For all profiles using $k \leq 137$, A/A had the exact *k*-mer count profile of at least one other genotype. Since the simulated reads are only 100bp, this means the A/A genotype could not be called uniquely at any *k*-mer length I tested.

The uniqueness of genotype count profiles was compared to the uniqueness of allele count profiles (**Methods 2.4.3.7**). Despite there being unique haploid profiles for all *PRDM9*-36 alleles at $k \geq 3$, no value of $k < 303$ produced unique *k*-mer count profiles for all diploid profiles. This is because two different genotypes could contain the same copy number of zinc finger repeats and thus the same *k*-mer counts (**Figure 2.9**).

A

***k*-mer**	**Allele *k*-mer counts**			**Genotype *k*-mer counts**	
	A	**L2**	**L21**	**A/A**	**L2/L21**
AGTCA	13	11	15	26	26
AGTCC	5	4	6	10	10
AGTGT	13	12	14	26	26

B

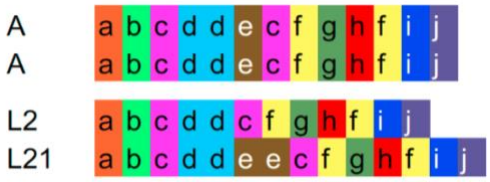


Figure 2.9: Diploid genotype *k*-mer count profiles are not all unique. A) A small subset of 5-mer counts for alleles A, L2, and L21, and for genotypes A/A and L2/L21. Though the three allele *k*-mer count profiles are unique, the two genotype profiles, which are sums of the corresponding allele profiles, are identical. **B)** Zinc finger repeat content for the A/A and L2/L21 genotypes. The copy numbers of all zinc finger repeats in the two genotypes are identical; there are no unique 5-mers spanning the 'de', 'dc', or 'ee' zinc finger repeat junctions that could be used to differentiate the two genotypes.

2.2.6 The *k*-mer distance model is nearly as effective as the *k*-mer count model at calling simulated data

Another potential source of valuable information provided by *k*-mers is the distance between a pair of *k*-mers in a known allele sequence. As introduced in **Results 2.2.3**, the outermost pair of *k*-mers from a sequencing fragment can be located in a known allele sequence and the distance between the *k*-mers in the allele sequence can be compared to the expected read fragment length. The distance model was developed to mitigate the issues with identical genotype *k*-mer count profiles. The model was first tested with the primary haploid simulation set at different values of *k* (**Methods 2.4.3.2**). The max method of dealing with multiple *k*-mer pair distances within an allele sequence resulted in the highest F1 scores, followed by the mean and median methods, which were comparable (**Figure 2.10**). The geomean method performed the worst. For all but the geomean method, the scores peaked when *k* values reached the 90s and only slightly dropped at the highest *k* values for reads simulated with errors. F1 scores were at least 0.83 when calling with the max method for all combinations of coverage and error rate (**Appendix Table 2.1**).

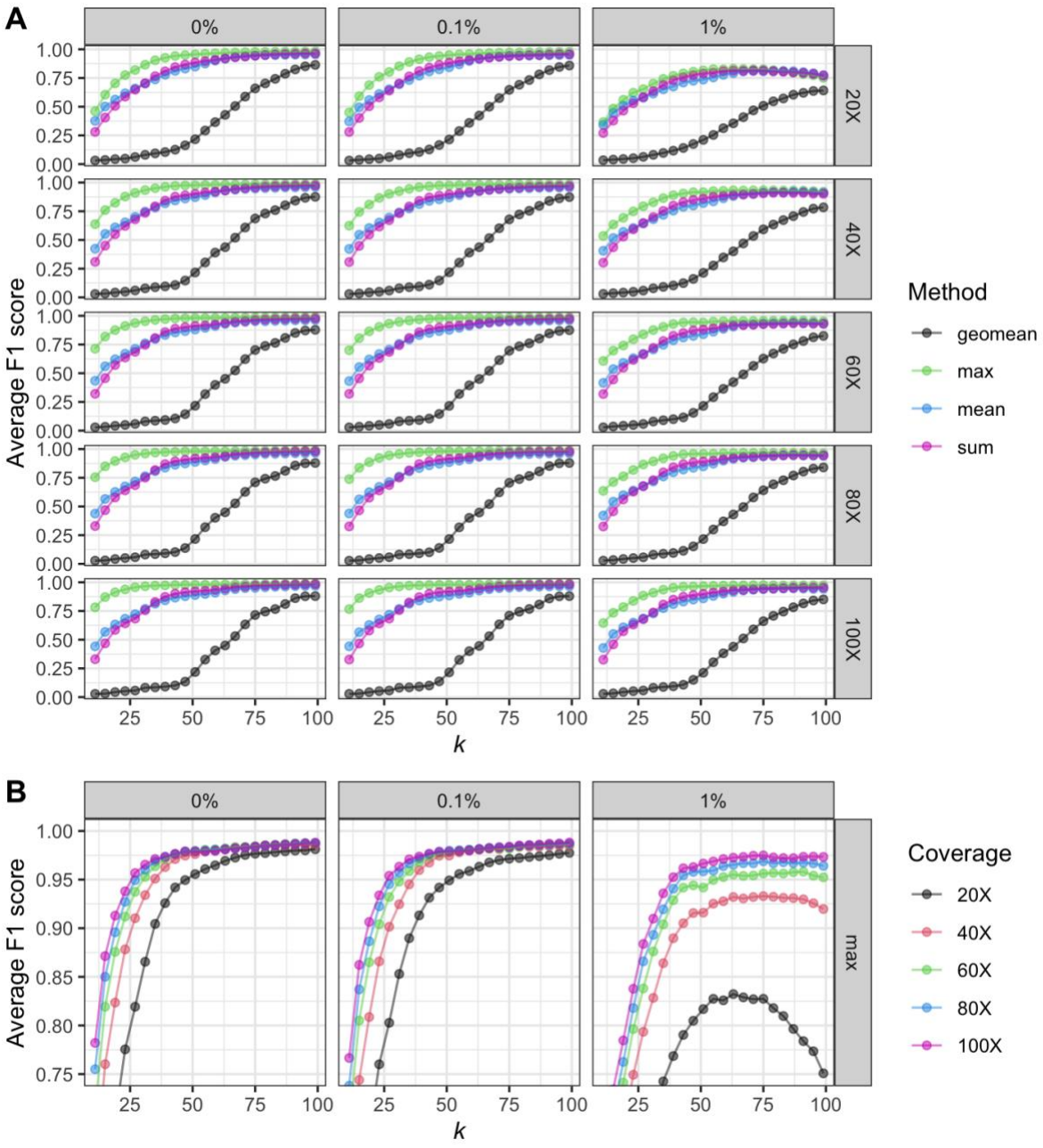


Figure 2.10: Haploid allele calling performance of the distance model. Average F1 scores across all replicates and all *PRDM9*-36 alleles for the primary haploid simulation set, called using the *k*-mer distance model and tested using different *k*-mer lengths. **A)** Four different methods (colors) of dealing with multiple distances per sequencing fragment were tested at different sequencing error rates (columns) and coverages (rows). **B)** Average F1 scores obtained using the distance-max method, which resulted in the best results overall. In general, scores improved with higher coverage (colors) and lower error rates (columns), and did not decrease at very high values of *k* for the error-free simulations.

For haploid allele calling, the count-coverage model performed better at lower values of *k* than the distance-max model, particularly for higher error rates. The distance model does not account for the sequencing error rate, which makes genotyping under high error rate conditions less accurate. At higher values of *k*, however, the distance model was comparable to or better than the count model. Unlike the count model, F1 scores for the distance model did not drop sharply as *k* neared the length of the reads, particularly for the error-free and 0.1% sequence error simulations. Once *k*-mer lengths near the length of the sequencing read, there is reduced *k*-mer coverage using the count model and therefore less information to be obtained from the count profiles. The distance model does not rely on *k*-mer coverage. In fact, *k*-mers become more unique as they get longer, meaning determining their positions within an allele sequence becomes more precise. Both models had lower F1 scores at higher values of *k* for the simulations with a 1% sequencing error rate because longer *k*-mers are more likely to contain sequencing errors (**Figure 2.11**).

The distance model was then tested using the primary diploid simulation set. At low coverages, F1 scores were generally higher than they were for the count model and did not drop as steeply at higher values of *k* (**Figure 2.11**). Since each allele has unique arrangements of zinc finger repeats, and since the two alleles in the diploid genotypes are assessed by the distance model as independent haplotypes instead of combined, the model is not prone to being unable to distinguish between genotypes like the count model. Even when two alleles differ by only a single SNP, the distance model is likely able to use read fragments that span the SNP to distinguish between the two alleles. Overall, the highest average F1 score obtained using the distance model for diploid calling was 0.7565, from the error-free 100X reads at $k = 99$. The 20X reads were called at a maximum average F1 score of 0.7124 (**Appendix Table 2**).

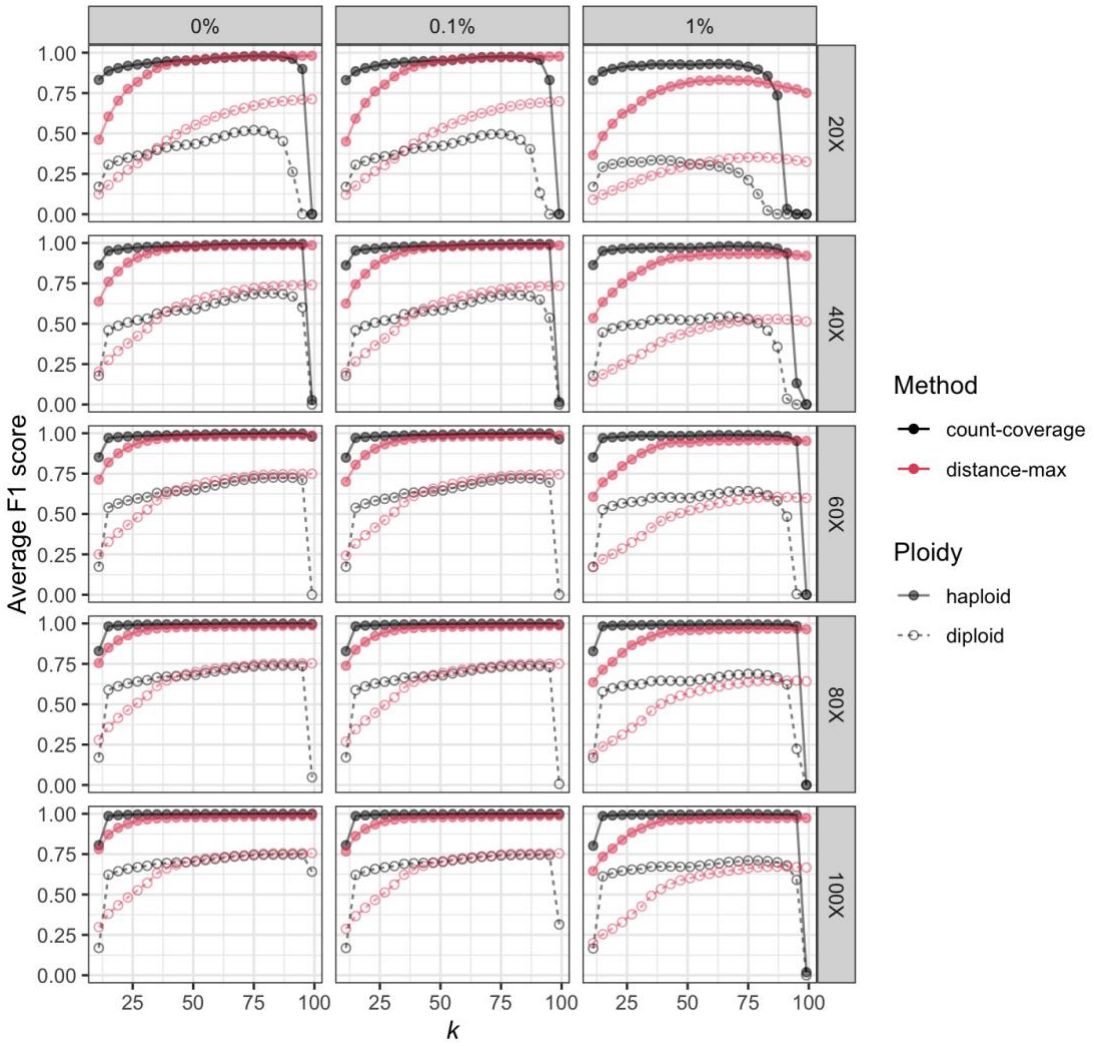


Figure 2.11: Haploid allele and diploid genotype calling performance of the distance-max model compared to the count-coverage model. Average F1 scores across all replicates and all *PRDM9*-36 alleles or genotypes for the primary haploid (solid circle) and diploid (open circle) simulation sets at different error rates (columns) and coverages (rows) and tested using different *k*-mer lengths. The simulations were called using the *k*-mer distance model with the max (red) method of dealing with multiple distances per sequencing fragment, and are compared to the *k*-mer count-coverage model (black). For haploid calls, the distance-max and count-coverage models performed similarly at the middle *k* range at low coverage and low error rates, but the count-coverage model outperformed the distance-max model at higher sequencing error rates. For diploid calls, the distance-max model outperformed the count-coverage model at low error rates and coverages, the models were comparable at high coverages, and the count-coverage model outperformed the distance-max model at high error rates for all but the longest *k*-mers.

2.2.7 Combining the *k*-mer count and distance models performs as well as or better than either method alone

The count and distance models have different strengths and weaknesses. The distance model inherently does not take into account sequencing error rates, whereas the count model does. On the other hand, the distance model is less likely to have identical distance profiles amongst genotypes, whereas the count model is unable to discriminate between two genotypes with identical *k*-mer count profiles. Combining the models was hypothesized to boost genotyping accuracy. For each replicate in each combination of coverage, sequencing error, and *k*-mer length, I multiplied the haploid likelihoods for each allele from both models to obtain a combined per-allele probability. All submethods of the count model (coverage, flank mean, and flank median) were combined with all submethods of the distance model (geomean, max, mean, and sum) for a total of 12 combined models tested. F1 scores were then averaged across all replicates for and across all alleles as before (**Methods 2.4.3.3**).

The different combined models performed remarkably similarly, though the models using the distance geomean method resulted in the lowest average F1 scores (**Figure 2.12 A**). Combined models using the count-flank mean or count-flank median methods did not perform as well as those that used the count-coverage method. The count-coverage & distance-max model combination resulted in the best F1 scores overall, performing noticeably better than the distance-max model alone (**Figure 2.12 B**). Compared to the count-coverage model, the combined model performed slightly better at low coverage and higher error rates, and was comparable to the count model at higher coverage. Perfect F1 scores were called for higher coverages with lower error rates. The lowest maximum F1 score observed was 0.9414 for 20X simulations with a 1% error rate; all other conditions had maximum scores over 0.98 (**Appendix Table 2.1**).

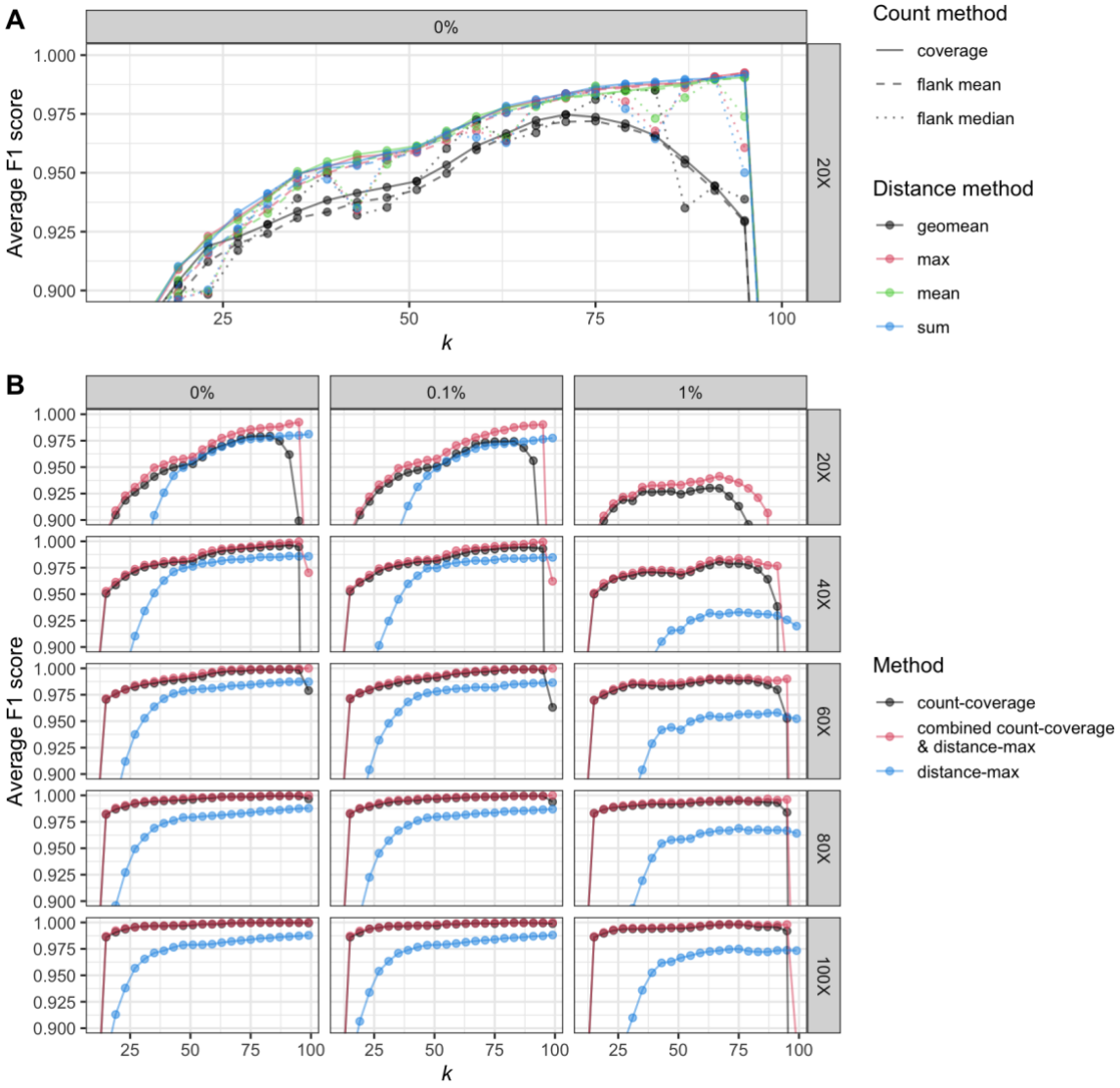


Figure 2.12: Haploid allele calling performance of combining the count and distance models. Average F1 scores across all replicates and all *PRDM9*-36 alleles for the primary haploid simulation set at different error rates (columns) and coverages (rows), called with combinations of count and distance methods and tested using different *k*-mer lengths. **A)** Results for all 12 model combinations on 20X error-free simulated reads, visualized by the count method (line type) and distance method (color) components. Combinations that used distance geomean had the lowest F1 scores. The combinations using the other distance methods performed very similarly. Combinations using count-coverage (solid line) performed better than those that used the flank mean or flank median methods for the count model. **B)** Full simulation set results for the best-performing distance, count, and combined models (color). The combined count-coverage & distance-max model performed as good as or better than the count-coverage and distance-max models.

Similar results were observed for calling diploid genotypes, but at a lower average F1 score, as was observed with the diploid results for count and distance model calls. In addition, while the combined count-coverage & distance-max model performed as good as or better than the count and distance models alone at high error rates and high coverages, the distance-max model generally had the highest F1 scores for low coverage and low sequencing error rate simulations (**Figure 2.13**). The highest average F1 score obtained was 0.7620, from the error-free 100X reads at k = 99. The 20X reads were only called at a maximum average F1 score of 0.6557 (**Appendix Table 2**).

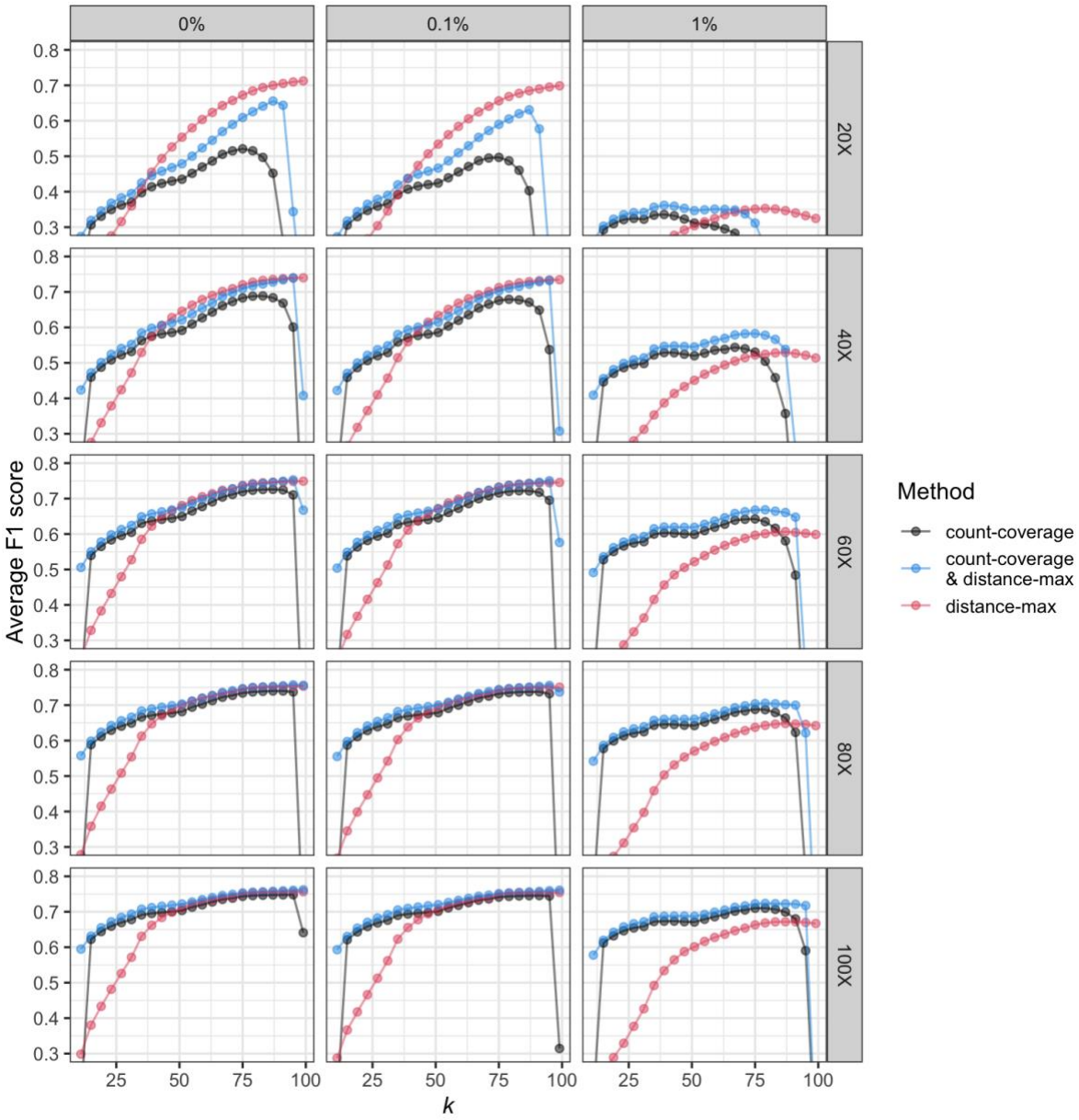


Figure 2.13: **Combining the count and distance models for genotyping diploid simulations.** Average F1 scores across all replicates and all 666 *PRDM9*-36 genotypes for the primary diploid simulation set at different error rates (columns) and coverages (rows), genotyped with the count-coverage (black) and distance-max (red) models, as well as a combination of the two (blue) and tested using different *k*-mer lengths. In general, the combined count-coverage & distance-max model performed as good as or better than the count-coverage and distance-max models alone for the high coverage and high error rate simulations, but the distance-max model performed the best for low coverage simulations.

2.2.8 Testing the short-read genotyping models on real sequencing data

Though useful for controlling data properties while developing and testing models, simulated reads are not perfect representations of real sequencing data. To assess the impact of real sequencing errors and unequal depth of coverage, I tested the models using the Genome in a Bottle (GIAB) Ashkenazi trio (Zook et al. 2016). These samples were chosen because they appeared to have common *PRDM9* genotypes very similar to the reference genotype after inspection of PacBio HiFi reads in `IGV` (**Methods 2.4.4.1**). Samples HG002 and HG003 both had a homozygous C > G SNV compared to the reference allele B at a position in the second 'c' zinc finger that corresponds to the SNV that differentiates allele A from allele B (GRCh38 chr5:23527130). These two samples did not have any other SNVs or indels, strongly suggesting that they have the genotype A/A. In addition to having the same SNV as HG002 and HG003, HG004 had a heterozygous C > T SNV at a position in the 'i' zinc finger that corresponds to the differentiation of allele A and allele L37 (GRCh38 chr5:23527668), strongly suggesting this sample has the genotype A/L37 (**Figure 2.14**).

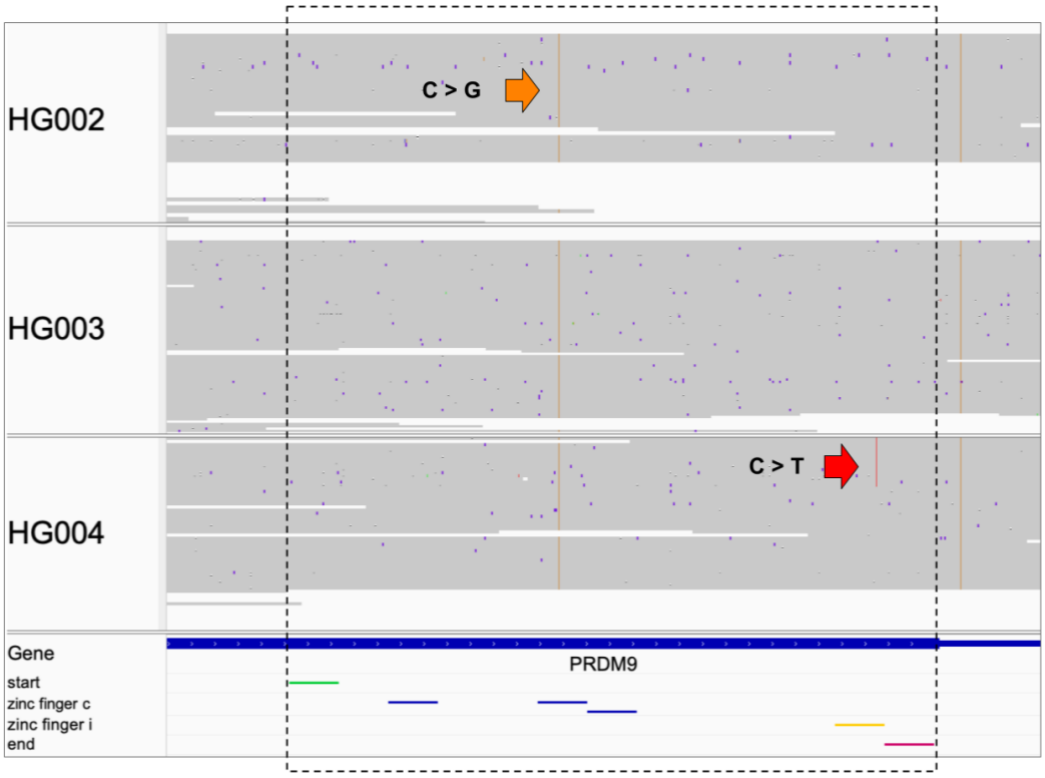


Figure 2.14: Genotyping the Ashkenazi trio with PacBio HiFi reads. `IGV` view of alignments to GRCh38 for the Ashkenazi trio (HG002, HG003, and HG004) PacBio HiFi reads from GIAB. The view is centered on the final exon in *PRDM9* containing the zinc finger region, which for the reference allele B starts with zinc finger 'a' (green) and ends with zinc finger 'j' (pink), shown as motifs in the bottom track. Also depicted in the bottom track are the locations for three 'c' zinc fingers (blue) and one 'i' zinc finger (yellow). Each thin horizontal grey line in the three alignment tracks is a single HiFi read aligned to GRCh38. Insertions are depicted by short vertical purple bars and deletions by thin horizontal black lines breaking up the grey reads. The lack of consistency for indel locations in the majority of the reads suggests these indels are small PacBio sequencing errors as opposed to true variants in the samples. It can therefore be assumed that all three samples have *PRDM9* genotypes with zinc finger arrays the same length as the reference allele B. Mismatched bases to the reference are colored relative to the nucleotide observed in the reads. Long vertical bars of color indicate that the mismatch occurs in several reads and is likely a SNV. The first SNVs occurring in the second 'c' zinc finger in all samples (long vertical orange lines) is characteristic of the A allele, which is a single C > G SNP compared to the reference allele B at that position. This suggests that HG002 and HG003 have homozygous *PRDM9* genotypes of A/A. The second SNV (short vertical red line) above the 'i' zinc finger (yellow) is indicative of the single C > T SNP that differentiates alleles A and L37. The presence of this SNV in roughly half of the reads in HG004 suggests this sample has the heterozygous genotype A/L37. The third SNV present in all three samples is located outside of the zinc finger array and does not affect allelic identity.

The Ashkenazi trio had both ~300X 150bp paired-end read and ~60X 250bp paired-end read Illumina data available. Genotype A/L37 has a unique *k*-mer count profile at most values of *k*, and A/A has a unique profile at $k \geq 138$, so it should theoretically be possible to call the genotypes of these three samples with either data set using the count-coverage model with a large enough *k*-mer size. Since the sequencing error rates and precise depths of coverage for the region were unknown, estimates were determined by calculating the average values of each from the reads mapped to the *PRDM9* + 10kb flank region. I assessed the coverage, flank mean, and flank median methods for estimating λ for the count model across a range of *k*-mer lengths. The genotypes were then ranked in descending order of likelihood, whereby the genotype with the highest likelihood had a rank of 1 (**Methods 2.4.4.1**).

None of the samples were able to be correctly genotyped (**Figure 2.15**). At best, the true genotype for HG004 was ranked eighth when using the 300X 2x150bp reads at *k*-mer lengths of 123, 127, and 131. While it was expected that the extremely high coverage would result in good genotyping performance given the simulation results, it was not expected that the longer 2x250bp reads would perform much worse, with the best rank being 96 for HG004, given that more genotypes have unique *k*-mer count profiles as *k* increases and longer reads allow for longer *k*-mers. The inability to correctly genotype the samples may be due to an increased accumulation of sequencing errors in the longer reads, and coupled with the lower sequencing coverage of the data set (~60X compared to ~300X), it is likely that there were not enough informative *k*-mers for the models to perform as well as the shorter-read data sets. Additionally, the flank mean and flank median methods of estimating λ underperformed compared to the coverage method. This could be explained by an increase in sequencing errors across the zinc finger array relative to the flanking sequences, which would result in an overestimation of λ when using either of the flank methods because the increase in zinc finger region errors would lower zinc finger *k*-mer counts relative to what would be expected given the less error-prone flank *k*-mer counts.

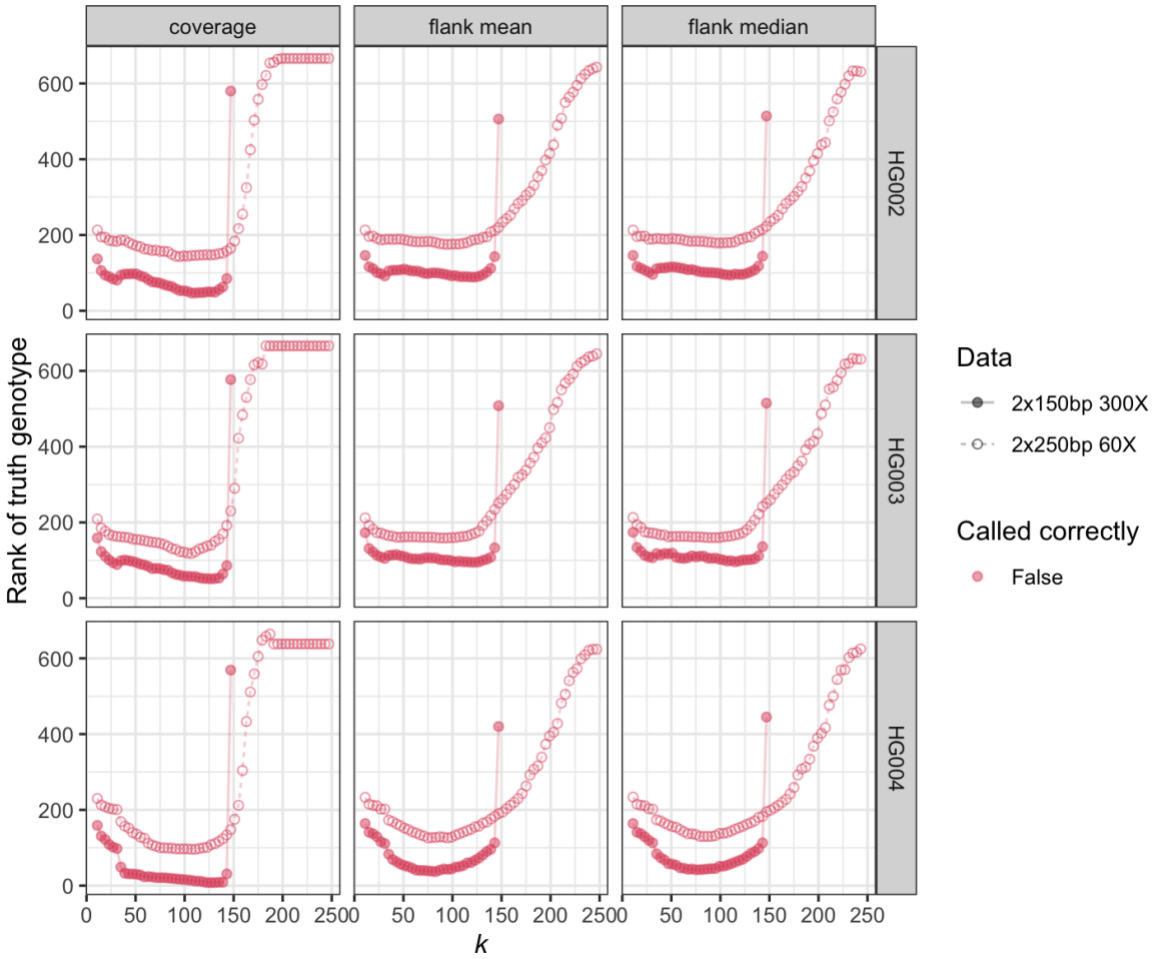


Figure 2.15: Assessing different methods of calculating λ for the count model on real sequencing data. Genotyping results for three different methods for estimating λ with the count model (columns) using the 2x150bp 300X (solid circles) and 2x250bp 60X (open circles) read sets for the GIAB Ashkenazi trio samples (rows). The ranking of the true genotype for each sample is plotted as a function of the length of *k*-mer used in the count model, where a rank of 1 would indicate the true genotype was called with the highest likelihood of all 666 possible *PRDM9*-36 genotypes. No samples were called correctly under any condition. In general, better ranks were achieved for the coverage method compared to the flank mean and flank mean methods, and the 2x250bp read data set had higher ranks for the true genotypes and was thus harder to call than the 2x150bp read data.

2.2.8.1 The *PRDM9* zinc finger array is more prone to sequencing errors than the flanking regions

Error rate plays a large part in the accuracy of both the count and distance models, as well as in the combination of the two. The sequencing error rate is uneven across the genome,

particularly in repetitive regions. I hypothesized that there is an increased sequencing error rate in the *PRDM9* zinc finger region relative to the 10kb flanking sequences. To assess this hypothesis, I counted *71*-mers throughout the *PRDM9* + 10kb region for both the reference allele B and the HG003 60X reads (**Methods 2.4.4.2**). Plotting the count of the unique allele B *k*-mers found in the sample reads allowed for observation of locations where expected *k*-mer counts were low. While the counts fluctuated across the full region of interest, there were two types of notable drops in *k*-mer coverage: in the zinc finger region, where counts appeared to be about half of the flanking region counts; and in the flanking regions, where there were sharp drops to 0 (**Figure 2.16 A**). Comparing these locations to the `IGV` alignments of HG003 to the GRCh38 reference (allele B), the positions with a count of 0 matched locations of homozygous SNVs in the flank regions, while the uneven dips in *k*-mer counts along the zinc finger region matched the high density of sequencing errors observed in the alignments (**Figure 2.16 B**). Reads with SNVs or sequencing errors result in erroneous *k*-mers, explaining the large drops in *k*-mer coverage and suggesting the reads aligning to the zinc finger region do have a higher error rate than those in the flanks. While some *k*-mers (286) in the zinc finger region that occur more than once in the reference sequence are not included in the plot and would appear as having a count of 0, the reads in the zinc finger region have a higher error rate than the flanking regions as observed by the increase in reads with colored sequencing errors in the read alignments. Additionally, overall coverage in the read coverage track from the `IGV` alignments was fairly uniform except around SNVs and within the zinc finger region, where coverage dropped for the 60X reads. This is often observed in repetitive regions of the genome (Van der Auwera and O'Connor 2020) and may occur because reads are mapping to other regions of the genome, or because they are left unmapped due to ambiguity from the repeats. Interestingly, coverage increased in the zinc finger region for the 300X reads, perhaps due to reads mapping to the repetitive region that should have mapped elsewhere in the genome.

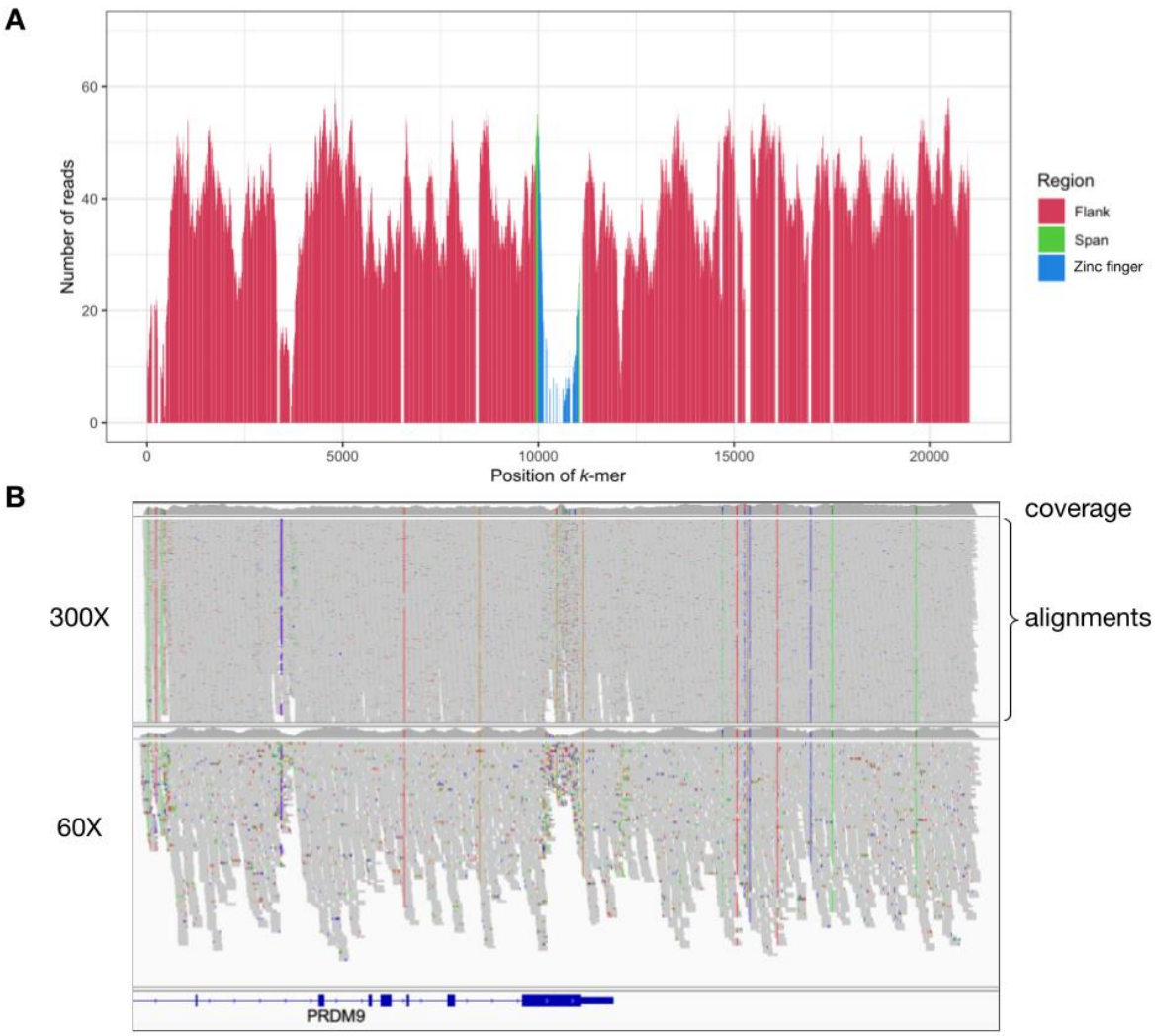


Figure 2.16: Distribution of *k*-mer coverage across the *PRDM9* zinc finger and flanking regions. A) Number of HG003 60X 2x250bp Illumina sequencing reads containing unique reference allele B *71*-mers starting at each position across the *PRDM9* zinc finger region + 10k flanks. *k*-mers are color coded as occurring entirely within the flank region (red) or zinc finger region (blue), or partially spanning both (green). **B)** `IGV` alignment tracks for the 300X and 60X reads for HG003. A) and B) are arranged such that the start and end positions of the region of interest are aligned as closely as possible. The gaps in the flanking regions of the *k*-mer count plot in A) correspond with the SNVs observed in B), appearing as colored vertical lines in `IGV`. The presence of a SNV in the reads relative to the reference genome means there is no coverage of the reference *k*-mer for the next *k* positions (e.g. 71 positions for *71*-mers). There is a substantial decrease in *71*-mer counts in the zinc finger region relative to the flanking regions. Additionally, there is an increase in sequencing errors (colored specks) in the zinc finger region relative to the flanking regions, and a drop of coverage across the zinc finger region for the 60X reads.

A known contributor to sequencing errors is the GC content of read fragments. Regions with both high and low proportions of GC bases have reduced representation in Illumina sequencing data, and PCR amplification is believed to be the primary cause of this bias (Benjamini and Speed 2012). Even though the Ashkenazi trio data were prepared with PCR-free libraries, I wanted to understand the effect of GC content on *k*-mer coverage. The proportion of GC bases in 250bp windows across the *PRDM9* + 10kb region of reference allele B was determined for the HG003 2x250bp reads, as was the distribution of GC content of *k*-mers in the flank regions for all Ashkenazi trio samples. GC content for allele B was somewhat higher across the zinc finger region than it was across the flanks (**Figure 2.17 A**). In addition, flank *k*-mers with higher counts in the sequencing reads tended to have lower GC content (**Figure 2.17 B**). Taken together, GC content might partially explain the increase in sequencing errors and the drop in *k*-mer coverage across the zinc finger region.

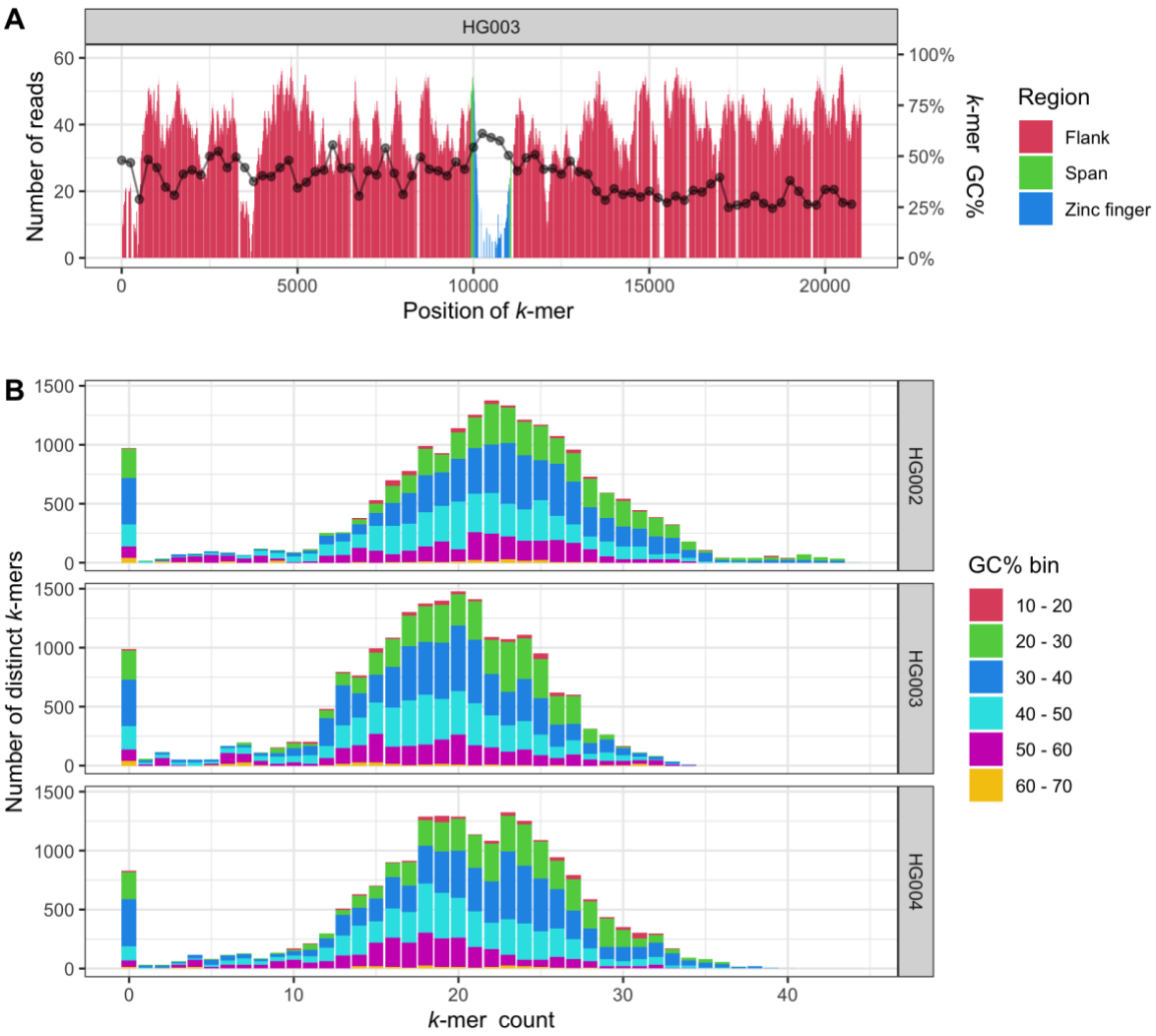


Figure 2.17: Effect of GC content across the zinc finger region. A) Number of Illumina sequencing reads containing unique reference allele B *71*-mers starting at each position across the *PRDM9* + 10k region for the HG003 2x250bp reads. *k*-mers are colored as occurring entirely within the flank region (red) or zinc finger region (blue), or spanning both (green). GC content as a proportion of 250bp bins is plotted across the region (black dots) and is higher in the zinc finger region than in the flanks. The gaps in the flanking regions of the *k*-mer count plot correspond to SNVs present in the reads relative to allele B. **B)** Flanking *71*-mer count distributions and the proportion of those *71*-mers for each GC% bin (lower value is inclusive, higher value is exclusive) for HG002, HG003, and HG004. *k*-mers with higher counts tended to have a slightly lower GC%.

Given that the relationship between base coverage and *k*-mer coverage is a function of *k* and the error rate, I expected the ratio between base depth and *k*-mer coverage to be constant throughout the genome. Instead, plotting the *k*-mer count divided by the base depth of the Illumina reads for HG003 resulted in a severe drop across the zinc finger region, with a ratio averaging around 0.8 to around 0.3, not including sharp drops to 0 at SNV positions (**Figure 2.18 A**). This drop was not observed in the long PacBio HiFi reads, where the ratio remained around an average of 0.9 throughout the flanks and zinc finger region (**Figure 2.18 B**). This further demonstrates problems with sequencing the *PRDM9* zinc finger domain with short-read technology, which may be caused by the higher GC% within the zinc finger region or by difficulties with mapping to this repetitive region. Unfortunately, this has a negative impact on *k*-mer counts and subsequently genotyping based on *k*-mer count profiles.

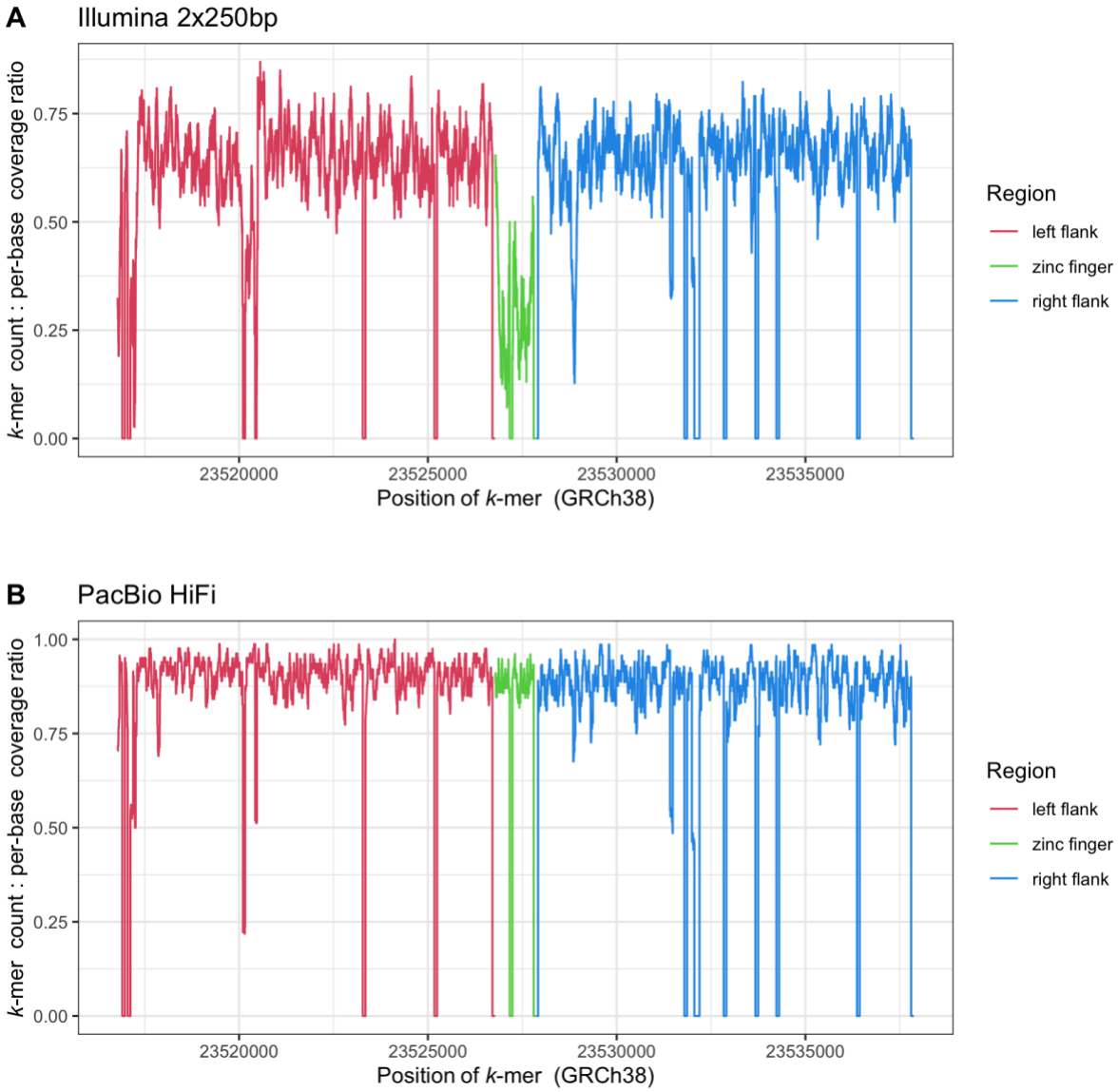


Figure 2.18: The ratio of *k*-mer counts to per-base read coverage for HG003 short and long reads. *k*-mer counts divided by per-base read coverage for HG003. Proportions per base are colored as occurring entirely within the flank region (red or green) or the zinc finger region (blue). Steep drops to zero indicate the presence of a SNV relative to GRCh38. **A)** The ratio for Illumina 2x250bp reads has a large drop throughout the zinc finger region at approximately half of the average across the flanks, indicating the reads mapping to the variable zinc finger domain are much more error-prone than those that map to the flanking regions. **B)** The ratio for PacBio HiFi reads is consistent throughout both flanking regions and the zinc finger region.

2.2.8.2 Performing error corrections on sequencing reads improves *k*-mer coverage

Read correction is frequently used on long reads to improve the high sequencing error rate (Fu et al. 2019) and on short reads to improve genome assembly (Heydari et al. 2017). I hypothesized that correcting error-prone Illumina reads would improve genotyping with my short-read models. I tested two programs with correction applications, `SPAdes` (Prjibelski et al. 2020) and `GraphAligner` (Rautiainen and Marschall 2020), as well as an in-house read correction script called `AlignCorrect` (Jared Simpson, unpublished). The 60X 2x250bp Illumina reads for HG003 were corrected and I compared *k*-mer coverage plots to those from the original raw reads (**Methods 2.4.4.3**). `GraphAligner` overcorrected the reads by removing true SNVs in the flanking regions, while `SPAdes` threw out 1.4% of the reads, reducing overall *k*-mer coverage. `AlignCorrect` was the most effective at improving *k*-mer coverage and reducing errors in the zinc finger region without removing too many reads (**Figure 2.19**).

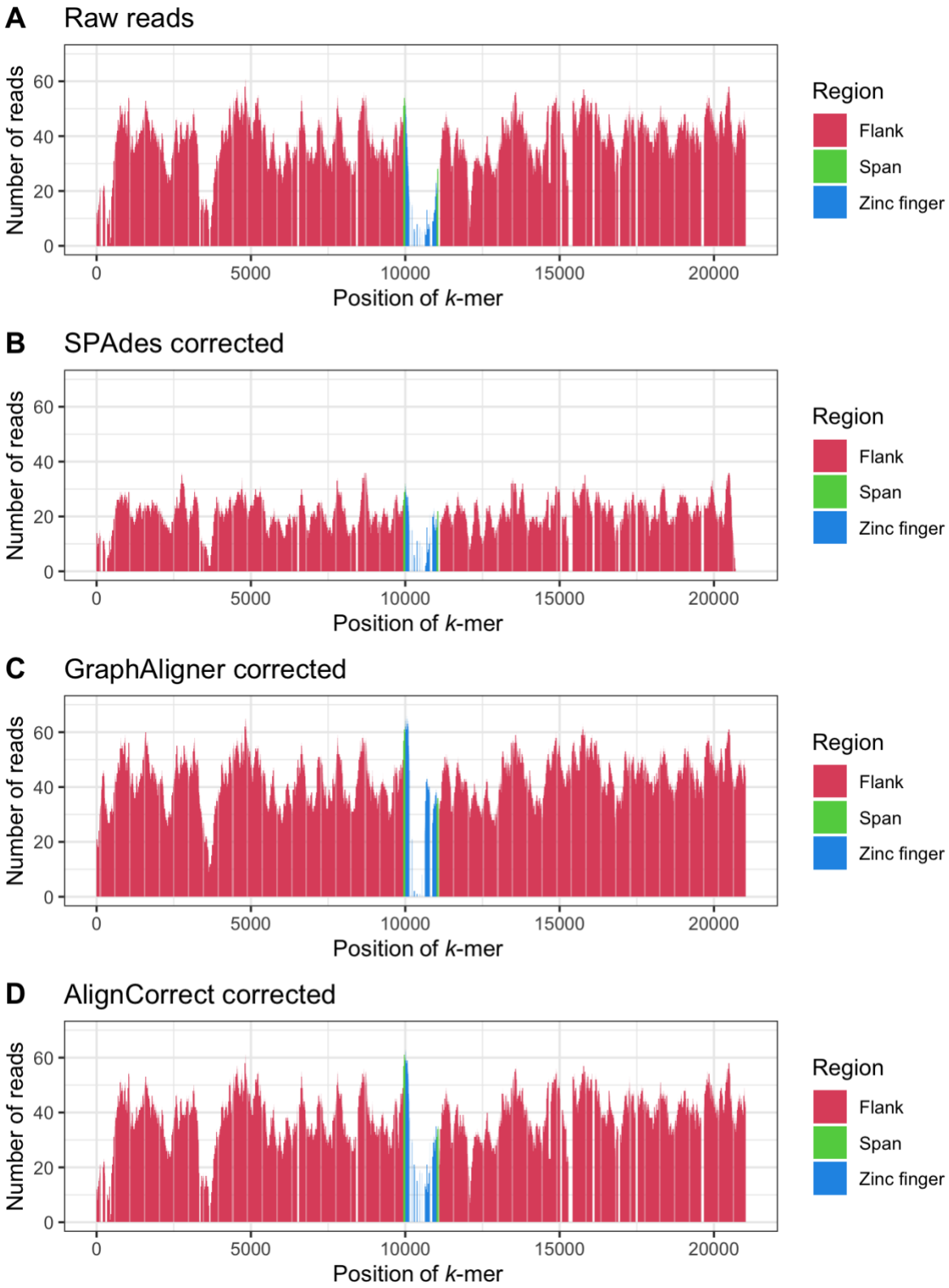

A
Raw reads
B
SPAdes corrected
C
GraphAligner corrected
D
AlignCorrect corrected
Number of reads
0
20
40
60
0
5000
10000
15000
20000
Position of k-mer
Region
Flank
Span
Zinc finger

Figure 2.19: The effect of correcting reads on the distribution of *k*-mer coverage across the zinc finger region for HG003. The number of Illumina sequencing reads containing unique reference allele B *71*-mers starting at each position across the *PRDM9* zinc finger region + 10k flanks for HG003. *k*-mers are colored as occurring entirely within the flank region (red) or zinc finger region (blue), or spanning both (green). The sharp drops in the flanking regions of the *k*-mer count plot correspond to SNVs present in the reads relative to the B allele. **A)** *k*-mer coverage for raw (uncorrected) reads. There is a substantial drop in coverage in the zinc finger domain, suggesting the reads are error prone in this region. **B)** *k*-mer coverage after read correction with `SPAdes`. Coverage throughout the flanking regions dropped substantially due to the software discarding many reads. **C)** *k*-mer coverage after read correction with `GraphAligner`. Coverage is slightly increased across the zinc finger region, but the true SNVs in the flanking regions have been incorrectly adjusted, as evident by the lack of sharp drops in *k*-mer coverage. **D)** *k*-mer coverage after read correction with `AlignCorrect`. Coverage across the zinc finger region increased, while the coverage and SNVs present in the flanking regions remained relatively the same. Overall, `AlignCorrect` produced the best read correction improvements in sequencing error rates.

HG003 reads corrected with `AlignCorrect` were then remapped to GRCh38. Comparing the raw and corrected reads in `IGV`, there was a noticeable reduction in errors across the zinc finger region while read coverage appeared to be the same (**Figure 2.20**). Read corrections were then performed on the remaining Ashkenazi samples and the count models were rerun using the corrected data. Only the count-coverage model resulted in genotypes being called correctly (ranked first): HG003 and HG004 were called using both data sets, whereas HG002 was only called with the 60X 2x250bp reads (**Figure 2.21**). For the 300X 2x150bp reads, only one length of *k*-mer resulted in correct calls for each of HG003 and HG004, whereas between four and eight *k*-mer lengths resulted in correct calls using the 60X 2x250bp reads. Successful *k*-mer lengths were 143 for the shorter reads and between 119 to 191 for the longer reads. Shorter *k*-mers were not expected to be successful given the results of the simulations, but it was again unexpected that the 2x250 reads did not have success with *k*-mers closer to the length of the reads. The remaining sequencing errors in the zinc finger region of the corrected reads may still be hampering longer *k*-mers.

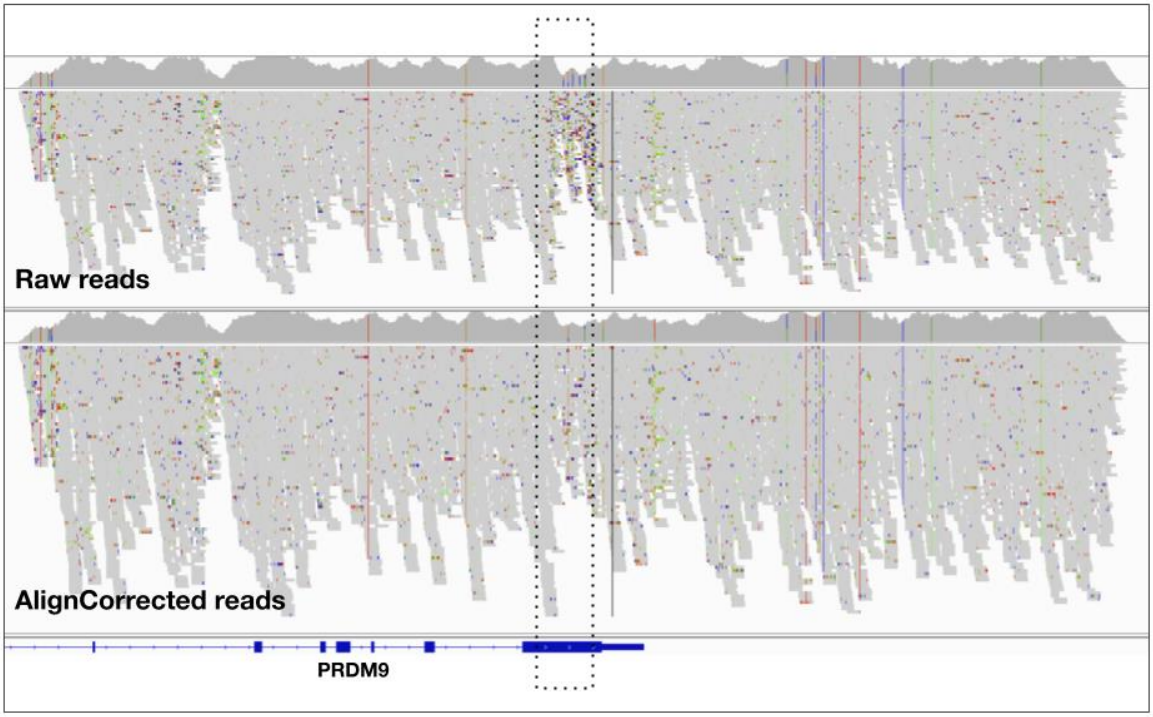


Figure 2.20: Reduction of sequencing errors after performing read correction. HG003 2x250bp Illumina reads aligned to the *PRDM9* + 10k flank region in GRCh38 as viewed in `IGV`. The grey horizontal lines in the alignment tracks are sequencing reads, the long vertical colored bars are SNVs, and the small multicolored specks are sequencing errors and/or somatic point mutations. The bottom track indicates the position of *PRDM9* exons (thick blue bars on the blue line) and the 3' untranslated region (tapered blue bar after the last exon). The zinc finger domain is located in the final exon (dotted box). The raw reads (top alignment track) had many sequencing errors clustered in the zinc finger region, visualized as multicolored specks. The reads after correction with `AlignCorrect` (bottom alignment track) showed great improvement in reducing sequencing errors in the zinc finger region.

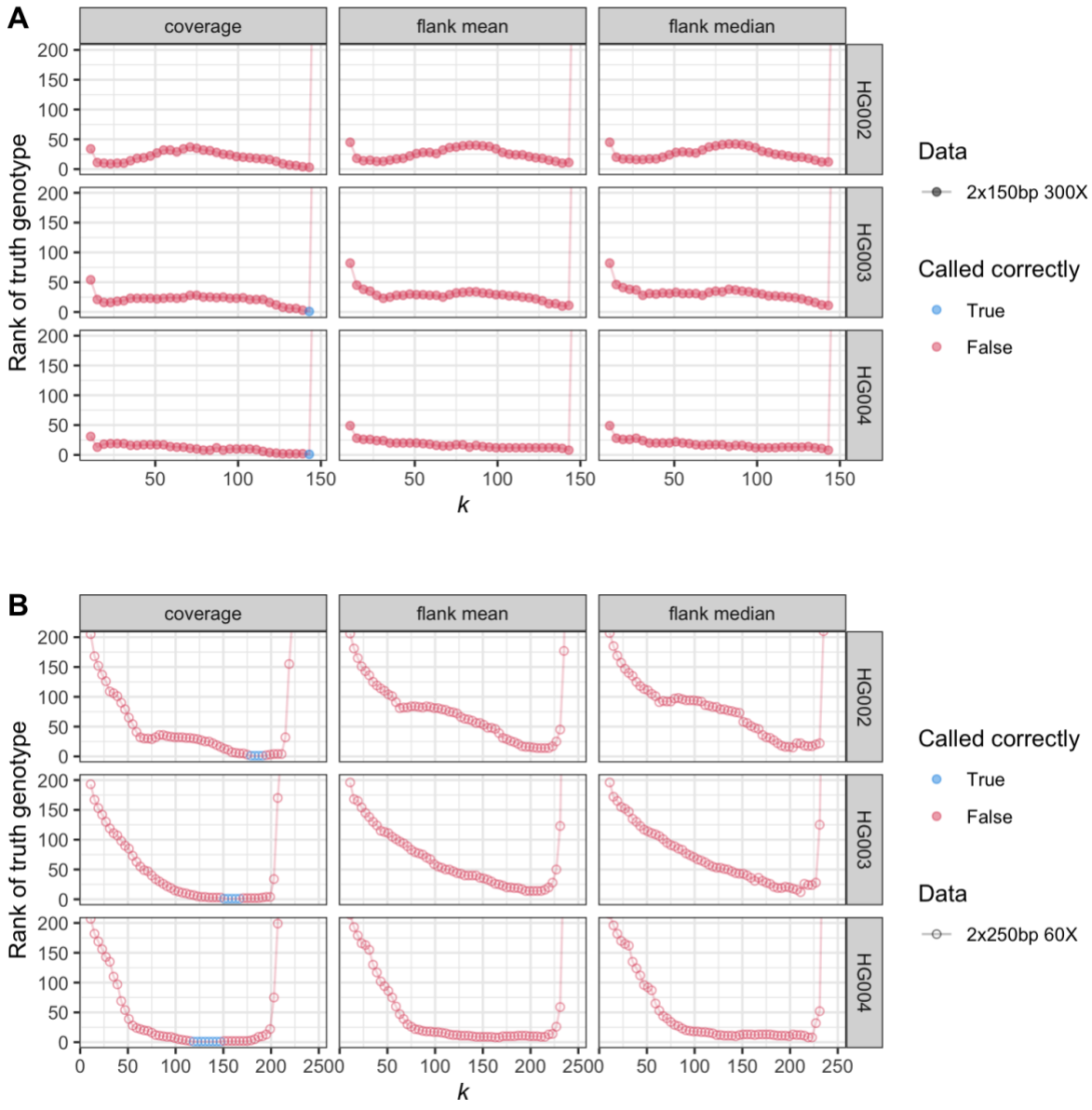


Figure 2.21: The effect of correcting sequencing reads prior to genotyping real sequencing data. Genotyping results for three different methods for estimating λ with the count model (columns) for the GIAB Ashkenazi trio samples (rows). The ranking of the true genotype for each sample is plotted as a function of the length of *k*-mer used in the count model, where a rank of 1 would indicate the true genotype was called with the highest likelihood of all 666 possible *PRDM9*-36 genotypes. Ranks up to 200 are plotted to better visualize the correct calls. **A)** The 2x150bp 300X reads. Two samples were correctly genotyped using the coverage method at k = 143. **B)** The 2x250bp 60X reads. All three samples were correctly genotyped using the coverage method for a small number of *k*-mer lengths.

2.2.8.3 Examining the range of λ values that result in correct genotypes for the count model

Given that the true value of λ is unknown and is likely affected by GC% and local sequencing errors, I manually supplied the count model with a wide range of λ values to see which, if any, resulted in the correct genotypes having the highest likelihoods for the Ashkenazi trio samples. To provide a baseline for how the manual λ results should ideally look, I first tested simulated 100X data for genotypes A/A and A/L37 using 10 replicates of each sequencing error rate from the primary diploid simulation set, representing the two genotypes observed in the trio samples. The λ values tested (0.5–70 in increments of 0.5) were above and below what λ would be if it was estimated using the count-coverage model for the data (**Methods 2.4.4.4**). For both $k = 51$ and $k = 71$, both genotypes were called correctly using several λ values at all three simulated error rates (**Figure 2.22**). For these simulations, λ estimated from the flank mean method was almost spot on with the λ estimated from the coverage method. Both of these λ estimations were a value that resulted in a correct genotype call for both A/A and A/L37 at $k = 71$, and for A/L37 for $k = 51$. Results were consistent for 10 different replicates of simulated reads.

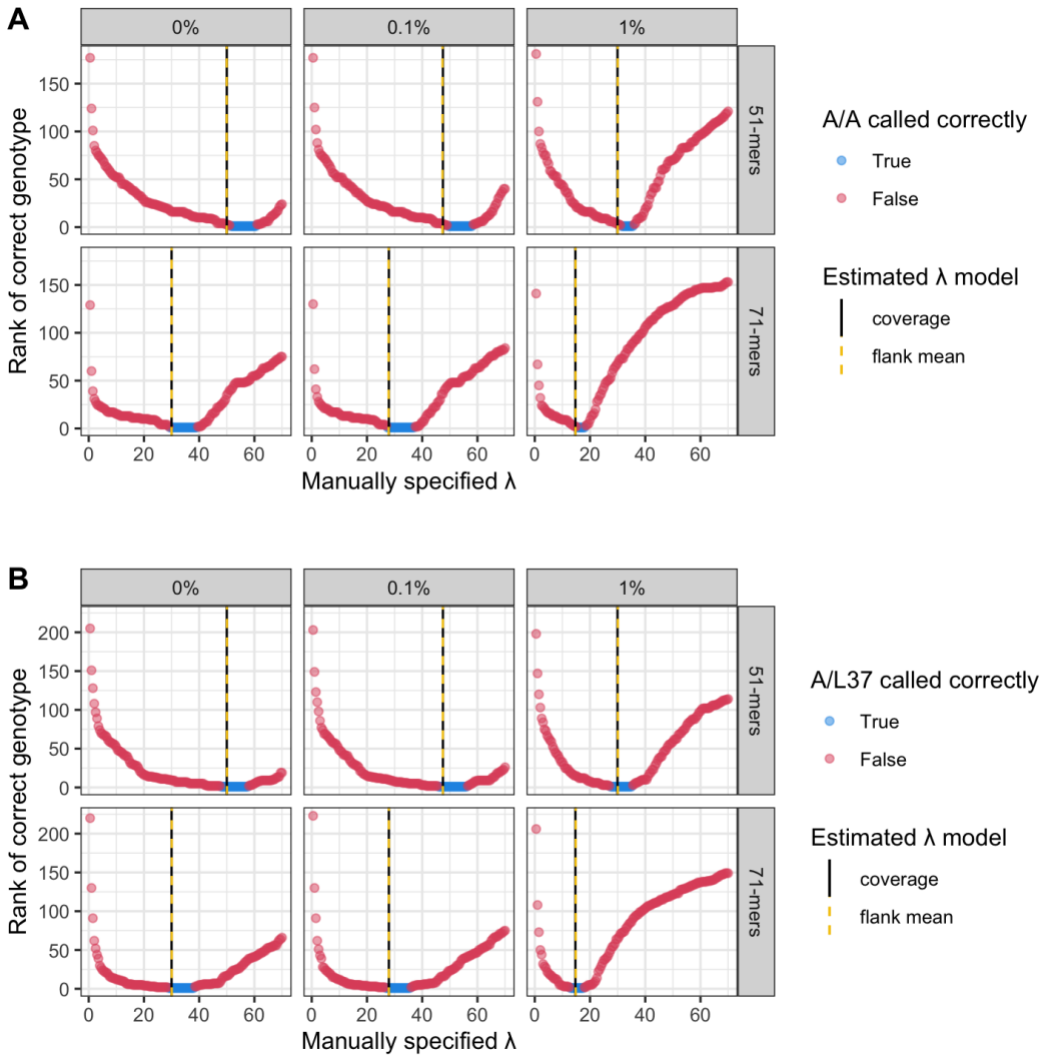


Figure 2.22: Diploid calling performance of the count-coverage model on simulated A/A and A/L37 genotypes using a range of manually specified λ values. One replicate of 100X simulated reads at various error rates (columns) for genotypes A/A and A/L37 were genotyped by manually specifying a range of λ values from 0.5 to 70 in increments of 0.5, using k = 51 or k = 71 (rows). The rank of the correct genotype is dependent on the specified λ value, with points colored when the correct genotype was ranked first and thus called correctly by the model (blue) or not (red). The vertical lines represent two nearly overlapping values corresponding to prior methods of estimating λ: from the mean flank k-mer counts (yellow), and from the coverage method calculated from the corresponding sequencing error rates (black). **A)** For the A/A simulations, the range of values that resulted in correct calls did not always intersect with the prior methods of estimating λ when k = 51 was used, reflecting the low rate at which A/A was called in the broad-range diploid simulation results. **B)** In all cases, the A/L37 simulations were called at a range that intersected the prior methods of estimating λ, as is expected given that in the simulated diploid genotyping results, A/L37 was correctly genotyped with an average F1 score of up to 0.91.

Using the count model on raw and corrected Ashkenazi trio 300X 2x150bp reads downsampled to 60X (performed by and available from GIAB; see **Methods 2.4.4.4**), I specified a range of λ values that approximately corresponded to different error rates between 0% and 1%, calculated using a coverage value of 60X. The total number of λ values tested differed for each *k*-mer length, with a minimum of 61 ($k = 131$) and a maximum of 167 ($k = 51$). Even with five different *k*-mer lengths tested, none of the samples were correctly called at any of the λ values. The best ranks for correct genotypes were 10, 23, and 3 for HG002, HG003, and HG004, respectively (**Figure 2.23 A**). After correcting the trio reads, there were several values of λ that resulted in the correct genotype being ranked first (**Figure 2.23 B**). These values of λ fell between the value estimated by the flank mean method and the value estimated by the coverage method with a sequencing error rate of 1%, suggesting the error rate of these samples is still quite high within the zinc finger region even with the corrections. The λ values calculated from the coverage and flank mean methods for the simulated A/A and A/L37 genotypes were nearly overlapping each other (see **Figure 2.22**); the spread of the λ values for these calculations using the Ashkenazi data shows the difficulty in determining the best way to estimate λ from real sequencing data.

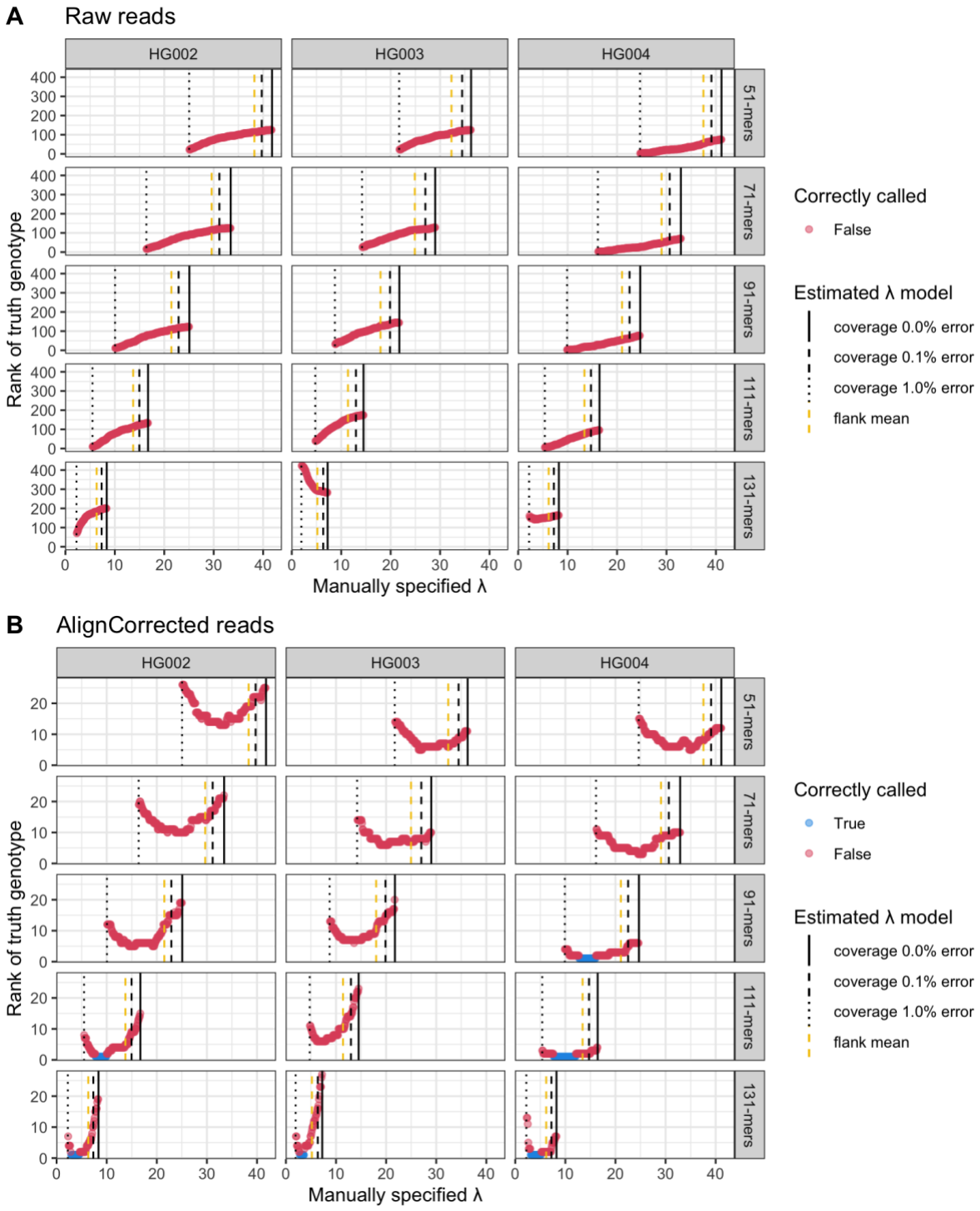

A
Raw reads
HG002
HG003
HG004
51-mers
71-mers
91-mers
111-mers
131-mers
Rank of truth genotype
Manually specified λ
Correctly called
False
Estimated λ model
coverage 0.0% error
coverage 0.1% error
coverage 1.0% error
flank mean
B
AlignCorrected reads
HG002
HG003
HG004
51-mers
71-mers
91-mers
111-mers
131-mers
Rank of truth genotype
Manually specified λ
Correctly called
True
False
Estimated λ model
coverage 0.0% error
coverage 0.1% error
coverage 1.0% error
flank mean

Figure 2.23: Effect of read correction on genotyping the Ashkenazi trio with the count-coverage model using a range of manually specified λ values. 2x150bp Illumina reads downsampled to 60X coverage from the Ashkenazi trio samples (columns) were genotyped using the count model at different values of *k* (rows) over a specified range of λ values. The λ range was determined per *k*-mer length tested, starting from the value estimated by the coverage model for a 1% error rate and ending at the value estimated by the coverage model for a 0% error rate. The rank of the correct genotype is plotted as dependent on the specified λ value, with dots colored when the correct genotype was ranked first and thus called correctly (blue), or ranked worse and not called correctly (red). The vertical lines represent four values corresponding to prior methods of estimating λ: from the mean flank *k*-mer counts (dashed yellow), and from the coverage method using a sequencing error rate of 0% (solid black), 0.1% (dashed black), and 1% (dotted black). **A)** Using the raw reads, genotypes were not called correctly for any sample at any λ for any length of *k* tested. **B)** Using the reads corrected with `AlignCorrect`, correct genotypes were called for all samples under at least one condition, and longer *k*-mers improved calls. In all cases where the sample was correctly genotyped, the λ values fell between the values that correspond to estimating λ with the flank mean method and from the coverage method with sequencing error rate of 1%.

To see if a length bias was affecting the results, I examined which genotypes were being called at λ less than and greater than a λ resulting in a correct genotype for the HG003 corrected 2x250bp reads. It appears that when λ is too small, there is a bias towards calling a genotype with longer alleles, and when λ is too large, there is a bias towards calling a genotype with shorter alleles (**Figure 2.24**).

A

k-mer length	λ	Genotype ranked first	Genotype ranked second
131	7.7	A/L33	A/A
	8.7	A/A	L3/L33
	9.7	A/L5	A/A
171	3.5	A/L33	A/A
	5.5	A/A	A/L33
	7.0	A/L5	A/A
211	2.1	A/L21	A/A
	2.9	A/A	A/L21
	3.8	A/L5	A/A

B

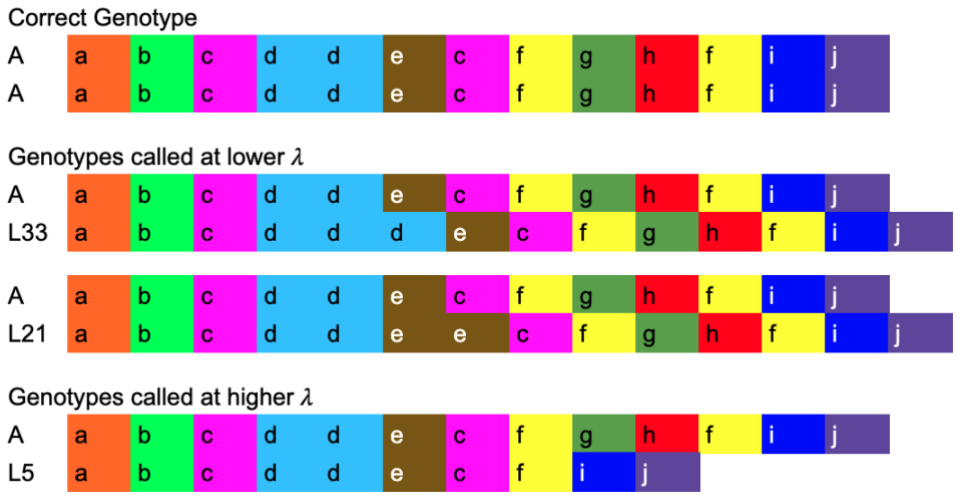


Figure 2.24: Comparison of first- and second-ranked genotypes called at different λ values. Genotyping on HG003 60X 2x250bp reads after correcting sequencing errors with `AlignCorrect` was performed with the count model at a variety of manually specified λ values. **A)** Genotypes ranked first and second at λ values that resulted in the correct genotype (A/A) being called first (yellow), along with the genotypes resulting from representative λ values higher and lower than the λ that resulted in a correct call. When not called first, the correct genotype was ranked second in all cases. **B)** Depiction of the correct *PRDM9* genotype for HG003 (two A alleles of 13 zinc finger repeats) and incorrect genotypes called when the λ value was less than or higher than the λ that resulted in the correct call. Lower λ values resulted in longer alleles being called, whereas higher values resulted in shorter alleles being called.

Overall, the results from genotyping the Ashkenazi trio Illumina data highlight the importance of estimating λ correctly. This is unfortunately difficult due to the error-prone reads that are sequenced from the repetitive *PRDM9* zinc finger region when using short-read technology. Correcting these reads in a manner that considers several known allele sequences, such as with the `AlignCorrect` method, has shown to reduce errors across the zinc finger region while still maintaining important SNVs for distinguishing genotypes, resulting in the correct genotypes being called for the two genotypes present in the trio samples.

2.3 Discussion

Repetitive regions of the genome remain difficult to sequence and study due to the limitations of short sequencing reads, namely being unable to map uniquely to the reference genome in many regions. To improve genotype calling in polymorphic repetitive regions of the genome using short-read data, I developed two models that use *k*-mer information instead of relying on precise realignments. By focusing on *k*-mers instead of mapping coordinates, the mappability issue with short reads is less important because the copy number of each *k*-mer directly informs the genotype inference. The count-coverage model takes into account estimates for sequencing error and depth of coverage in order to adjust the expected *k*-mer counts for a particular sample. While this model worked very well for simulated haploid samples using the *PRDM9*-36 list of alleles, many diploid genotypes had identical *k*-mer count profiles for $k < 303$, meaning they would never be able to be distinguished with this model alone when using Illumina reads, which are typically 100–250bp long. The distance model was developed to address this shortcoming, but as it is currently written, it does not have a way to factor in sequencing errors when calculating genotype likelihoods. Nevertheless, the distance-max model worked reasonably well for simulations with low sequencing error rates, and outperformed the count model when *k* approached the length of the reads. Results from combining the two models were at least as good as or better than the count model alone for the haploid simulations, and often as good for diploid simulations, showing value in both approaches of using *k*-mer information.

In addition to the lack of using error rates in determining allele and genotype likelihoods, the distance model also only considers one *k*-mer per read (the outermost 5' *k*-mer), whereas the count model considers all *k*-mers in the reads. Using more *k*-mer pairs might improve results for the distance model, but it would not be advisable to use all *k*-mers in the reads given the increased error rate that occurs at the ends of Illumina reads. Incorporating an estimated sequencing error rate with or without the use of some additional *k*-mer pairs would likely be able to improve accuracy of the distance model.

Both the count and distance models have limitations in the assumptions they make. The count model assumes that *k*-mer counts follow a Poisson distribution, which implies the *k*-mers are independent. However, since all *k*-mers were used in sliding 1bp windows, this assumption is not true, as subsequent *k*-mers in a sequence are highly related to the *k*-mers immediately upstream. The distance model assumes that paired-end sequencing reads come from sequence fragments that follow a normal distribution in terms of length. While the true distribution is close to normally distributed, the practice of size selection during library preparation often means the tail ends of the distribution are sharply cut and thus the assumption does not strictly hold true.

Even though the *PRDM9*-36 alleles do not have unique *k*-mer count profiles for all genotypes until $k = 303$, the number of genotypes with unique count profiles increases substantially as *k* increases. At $k = 73$, for example, there are only six pairs of identical profiles, and only two pairs at $k = 138$. For an unknown sample with 150bp reads, the correct genotype could in theory be narrowed down to two possibilities because the count model outputs a ranked list of all possible genotype likelihoods instead of a single call. Ideally the count model should be improved to a point where false calls are greatly reduced. While some callers use population frequencies in calling variants, I have chosen not to follow this approach because I want my models to be able to identify rare alleles and not choose the more common genotype in a tiebreaker situation.

Moving from simulated data to real sequencing data proved to be challenging. The Illumina reads for the GIAB Ashkenazi trio were highly erroneous in the *PRDM9* zinc finger region, much more so than in the 10kb flanking regions. The A/A and A/L37 genotypes of the samples

could not be called on the raw reads, despite testing with many different *k*-mer lengths and ways of estimating λ. It was possible to genotype the samples, however, after performing read correction on the data. Specifically, correcting the reads by considering all known allele sequences allowed for minimal corrections to be made while preserving true SNVs relative to the GRCh38 reference. In order to define the parameters that resulted in correct genotype calls, I had to manually supply a range of λ values, demonstrating how difficult it is to determine the best estimate of λ using the coverage model. Additional λ estimate methods were developed that use the *k*-mer counts from the flank alleles, but they were not as effective as the coverage model in estimating λ and correctly genotyping the trio. One possible way to improve genotyping of real data with the count-coverage model could be to select a range of λ values likely to be accurate (e.g. values estimated based on sequencing error rates of 0.1% and 1%) and average the likelihoods for each allele or genotype across that range, then rank the calls on the averaged likelihoods.

Overall, the count and distance models show promise in genotyping difficult regions of the genome with short-read data. These results are a proof-of-concept for the potential benefits of using *k*-mer information to better elucidate genomic variants in repetitive loci.

2.4 Methods

2.4.1 Compiling sequences from known *PRDM9* variants

The zinc finger content of 29 *PRDM9* alleles were collected from a review (Ponting 2011), along with zinc finger content and sequences for seven additional alleles from a study by a previous lab member (Hussin et al. 2013). The DNA sequences for all 36 alleles were pieced together using the zinc finger sequences to generate a list of known alleles referred to as the ***PRDM9*-36** list. While intending to summarize the 29 alleles initially described by Berg et al. (2010), the review incorrectly added an ‘h’ zinc finger to the array in allele L4, changing the zinc finger content from ‘abcddecftpfqj’ to ‘abcddecft**h**pfqj’ (Figure 4 in Ponting 2011). Upon

discovery that one of the allele sequences was incorrect, a thorough check of *PRDM9* variants described in publications and deposited in the NCBI Nucleotide database was performed.

A total of 15 publications were found to describe at least one *PRDM9* allele variant (Oliver et al. 2009; Thomas et al. 2009; Baudat et al. 2010; Berg et al. 2010; Kong et al. 2010; Parvanov et al. 2010; Berg et al. 2011; Ponting 2011; Borel et al. 2012; Hussin et al. 2013; Jeffreys et al. 2013; Vergés et al. 2017; Alleva et al. 2021; Beyter et al. 2021; Wang et al. 2021). All available sequence information was downloaded, copied, or manually typed into files. Descriptions of amino acids or mutation locations were attempted to be reconstructed into DNAs sequences. The sequences were searched for known zinc finger motifs, which were replaced with known zinc finger names when possible. All remaining 84bp strings were given temporary zinc finger names.

A multi-fasta file of 128 *PRDM9* sequences was downloaded from the NCBI Nucleotide database after the following search in December 2021:

```
(PRDM9) AND "Homo sapiens"[porgn:__txid9606] NOT "patch" NOT "scaffold" NOT "assembly" NOT "chain" NOT "ZCWPW1"
```

The NCBI sequences were searched for known zinc finger motifs which were replaced with zinc finger names. Any sequence upstream of zinc finger ‘a’ was removed, since all alleles start with zinc finger ‘a’. As well, the motif GATGAGTAA (or the shorter motif GATGAG) and all bases downstream were removed, as this is the sequence immediately following the zinc finger array in the GRCh38 reference. Accession numbers were cross-referenced with accessions mentioned in publications and the sequences were compared between the two when possible.

Unique allele and zinc finger sequences were identified and those with well-known names were sorted alphabetically, using publication dates to decide orders for different alleles or zinc fingers given the same name. For example, the zinc finger name ‘u’ was used for different sequences by both Berg et al. (2011) and Hussin et al. (2013), so the former sequence was ordered first because it was described in a 2011 publication whereas the other was published in 2013 (see **Chapter 5 section 5.1.5** for more details). Sequences only found in the NCBI

Nucleotide database were ordered after all the publication sequences. **Standardized names** were given to each unique zinc finger and allele sequence to ease computational parsing and remove confusion around inconsistent naming strategies. Allele names start with 'P' and have three digits (e.g. allele A is P001), whereas zinc finger names start with 'Z' and have three digits (e.g. zinc finger a is Z001). Alleles and zinc fingers only described in Jeffreys et al. (2013) as being observed in sperm cells or as somatic mutations in blood (blood/sperm alleles) were removed to form a list of 106 known alleles observed as germline variants in the human population (the ***PRDM9*-106** list). The complete list that includes the blood/sperm alleles is referred to as the ***PRDM9*-642** list. Specific details about curating the lists can be found online (https://github.com/hgibling/PRDM9-Variants).

2.4.2 Simulating short-read sequencing data

For each allele in the *PRDM9*-36 list, a total of fifteen sets of simulations were generated for combinations of coverage (20X–100X in increments of 20) and sequencing error rates (0%, 0.1%, and 1%). Each set consisted of 100 replicates of 100bp paired-end reads with a mean fragment length of 250bp and a fragment length standard deviation of 50bp. The simulations used the allele-specific zinc finger array sequences along with 10kb flanking sequences from GRCh38 (chr5:23516673–23526673 for the left flank; chr5:23527764–23537764 for the right flank). Simulations were generated with `wgsim` (Li et al. 2009) using seeds from one to 100 to ensure unique simulations for each replicate. This collection of *PRDM9*-36 allele simulations is referred to as the **primary haploid simulation set**. Additional haploid simulation sets were generated as well for initial model testing: paired-end reads of different lengths (25bp–500bp in increments of 25) with a fixed fragment length of 250bp and fragment length standard deviation of 50bp, and paired-end reads of different fragment lengths (200bp–500bp in increments of 50bp) with a fixed read length of 100bp and fragment length standard deviation of 50bp. Both of these conditions were generated for all combinations of 20X and 100X coverage with 0%, 0.1%, and 1% simulated sequencing errors, simulated from allele sequences with and without 10kb flanking sequences around the *PRDM9* zinc finger array, with 100 replicates for each condition. These additional sets were generated with `wgsim` in the same manner as the primary haploid simulation set.

To obtain a set of reads from a diploid genotype, I determined the number of reads required to be simulated from each allele in the genotype to obtain the specified coverage, averaged that value, and provided as input a fasta file containing both alleles to perform the simulations with `wgsim`. The read parameters for generating this **primary diploid simulation set** were otherwise the same as those used in generating the primary haploid simulation set (paired end 100bp reads with a mean fragment length of 250bp and a fragment length standard deviation of 50bp, coverage of 20X, 40X, 60X, 80X, or 100X, and a sequencing error rate of 0%, 0.1%, or 1%).

2.4.3 Short-read genotyping models

2.4.3.1 Probabilistic *k*-mer counting model

The **count** model was developed to genotype short-read data using *k*-mer counts. To call haploid genotypes, for each allele from a list of known variants, the model determines how many times *k*-mers of different lengths occurred, generating **allele *k*-mer count profiles** across a range of *k* values. The model also counts *k*-mers from the supplied set of sequencing reads for a sample, generating a **read *k*-mer count profile**. Then by using a list of *k*-mers from the union of all *k*-mers in the allele *k*-mer count profiles (e.g. all *k*-mers in the *PRDM9*-36 alleles without 10kb flanks), the likelihood of observing the read *k*-mer count profile given the *k*-mer count profile of an allele is calculated using a modification of the probability mass function of a Poisson distribution:

$$L(a) = P(R|a) = \prod_i \frac{(c_i\lambda)^{r_i} \cdot e^{-c_i\lambda}}{r_i!}$$

where a = the allele of interest, R = the vector of read counts, i = the index of a vector K of union *k*-mers from the allele count profiles, c_i = the count of *k*-mer K_i from the allele count profile, r_i = the count of *k*-mer K_i from the read count profile, e = Euler's number, and λ = the estimated *k*-mer coverage. The model assumes independence between *k*-mers as a simplification, though the *k*-mers in a single read are inherently related.

I developed three different methods to estimate λ: a modification of a formula from the `Cortex` software (Iqbal et al. 2012) that takes into account the effect of sequencing errors (the **coverage** method), calculating the mean count of *k*-mers unique to the flanking sequences in the read *k*-mer count profile (the **flank mean** method), and calculating the median flanking *k*-mer count in the read *k*-mer count profile (the **flank median** method). The original *k*-mer coverage formula used by `Cortex` is:

$$\lambda = (l - k + 1) \cdot \frac{v}{l} \cdot (1 - \varepsilon k)$$

where l = the read length, k = the *k*-mer length, v = the average sequencing read coverage, and ε = the substitution error rate. The **modification** of this formula used in the count-coverage model is:

$$\lambda = (l - k + 1) \cdot \frac{v}{l} \cdot (1 - \varepsilon)^k$$

because it was found to be better at estimating λ than the original `Cortex` formula (see **Results 2.2.4.2**). The term $(1 - \varepsilon)^k$ represents the chance of correctly sequencing every base in the *k*-mer, which is required to properly observe it in the allele.

There are two instances where c_i in the modified probability mass function can equal zero. The first case is when a *k*-mer present in the reads is present in some allele, but not in the allele being tested, as would be the case when the reads were not simulated from the allele being tested. The second case is when a read contains a sequencing error that results in a *k*-mer having a point mutation that makes the erroneous *k*-mer match a *k*-mer from a different allele than the one from which the read originated. In the first instance, $r_i \simeq c_i\lambda$ of the correct allele, whereas in the second instance, $r_i \simeq 1$. In both of these cases, for all methods, I substituted $c_i\lambda$ with a $\boldsymbol{\lambda}$ **error value** of 1 to avoid obtaining a probability of 0 from the Poisson probability mass function, as would be the case if c_i was left as 0.

Since the count model assesses the set of sequencing reads against the full list of provided alleles, the read *k*-mer count profile is compared to each allele *k*-mer count profile. The model outputs a ranked list of alleles in descending order of likelihood.

To use this approach to call diploid genotypes (pairs of alleles), I generated all possible **genotype *k*-mer count profiles** by summing all possible pairs of allele *k*-mer count profiles. The value of λ was divided in half to account for each allele representing half of the coverage. The diploid count model determines likelihoods for all possible pairings of the input list of alleles (i.e. all possible genotypes).

2.4.3.2 Probabilistic *k*-mer distance model

The distance model was developed to genotype short-read data using the distances between paired *k*-mers as estimates of read fragment lengths. The model scores a set of paired-end sequencing reads using the probability density function of a normal distribution, assuming that the read fragment lengths are normally distributed within a sample:

$$L(a) = P(F|A) = \prod_i \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{-\frac{(f_i-\mu)^2}{2\sigma^2}}$$

where a = the allele of interest, F = the vector of read fragment lengths, A = the vector of outer distances between pairs of *k*-mers in the allele, i = the index of A, f_i = the observed length of fragment F_i (the distance in the allele between the pair of *k*-mers from the ends of a read pair), μ = the mean expected fragment length, and σ = the standard deviation of the expected fragment length.

A dictionary was generated to store all possible non-negative distances between all *k*-mer pairs for each known allele. To determine the observed distance between a read pair, I considered the outermost *k*-mer pair from the reads and looked up the possible distances in the dictionary. It is possible that a *k*-mer pair can be observed more than once in a given allele. In these cases, all possible fragment lengths were considered (f_i = [f_{i1}, f_{i2}, …, f_{in}]) using one of four different approaches: adding all likelihoods for all fragment lengths (**sum**), averaging the likelihoods from all fragment lengths (**mean**), using the geometric mean of the likelihoods from all fragment lengths (**geomean**), or considering only the fragment length that gave the highest likelihood (**max**).

As before, all input alleles were tested for a sample and the allele with the highest likelihood was considered the called allele. For diploid genotyping, the model determines likelihoods for all possible pairings of the input list of alleles, and the average of the likelihoods from each allele in the genotype is used to calculate the genotype likelihoods:

$$L(f_i|g_j) = \prod_i \left(\frac{L(f_i|h_1)}{2} + \frac{L(f_i|h_2)}{2}\right)$$

where f_i = read fragment i, g_j = genotype j, h_1 = the first haplotype (allele) in g_j, and h_2 = the second haplotype in g_j.

2.4.3.3 Combining the count and distance models

After generating likelihoods using both the count and distance models, the likelihoods per allele or genotype can be multiplied to obtain a list of likelihoods for a combination of the two methods. Since there are three different methods of estimating λ for the count model, and the distance model has four methods of handling multiple possible *k*-mer distances, there are a total of 12 possible combinations of the two models, all of which were assessed.

2.4.3.4 Assessing model performance with simulated data

The primary haploid and diploid simulation sets both had 100 replicates for each sequencing condition (combination of coverage and sequencing error rate) for each allele or genotype from the *PRDM9*-36 list. Both sets were used to assess the full array of methods from the count model (three), the distance model (four), and the combination of the two models (12), and all were assessed under a range of *k*-mer lengths (11–99 in increments of four). For each model, the allele or genotype with the highest likelihood was called as the genotype for the sample. To assess how well the models performed, the F1 metric was used as a measure of accuracy:

$$F1 = 2 \cdot \frac{precision \cdot recall}{precision + recall}$$

where

$$precision = \frac{true\ positives}{true\ positives + false\ positives}$$

$$recall = \frac{true\ positives}{true\ positives + false\ negatives}$$

Using haploid allele calling as an example, if the called allele for a read set was correct (i.e. the allele that was used to simulate the reads had the highest likelihood), the allele was considered a **true positive**. If the read set was called as an allele other than the one used to generate the reads, the allele that generated the reads was considered a **false negative** and the allele called with the highest likelihood was considered a **false positive**. All other alleles that did not generate the sequencing reads and did not have the highest likelihood were considered **true negatives**. If the correct allele shared the highest likelihood with one or more incorrect alleles, the correct allele was considered a true positive while the incorrect alleles were considered false positives. Therefore, for each simulated set of reads, there was at most one true positive or one false negative allele, because reads were only simulated for one allele. The F1 score was therefore used instead of accuracy because of the large difference in true positive and true negative group sizes.

F1 scores were calculated for each individual set of sequencing reads and were then averaged across the 100 replicates of simulations under the same conditions. For example, the F1 scores from all 100 20X error-free allele A simulations assessed with *71*-mers were averaged to provide an overall F1 score for calling allele A under these conditions. Once the F1 scores were averaged for all simulation sets, they were averaged across all alleles from the same sequencing conditions. Finally, all 36 allele-level averages were averaged to get an overall average F1 score for the condition at hand (e.g. the 20X error-free simulations from *PRDM9*-36 alleles genotyped with *71*-mers). The same process was used to assess diploid genotypes,

with each genotype being assigned as a true or false positive or negative, and F1 scores being averaged across all 100 simulations per genotype, then averaged across all 666 genotypes.

Considering all fifteen simulation conditions (five different coverages and three different error rates) and all values of *k* assessed (23), a total of 345 overall average F1 scores were generated for both haploid and diploid simulations, along with 12,420 allele-level and 229,770 genotype-level average F1 scores.

2.4.3.5 Assessing read and fragment length biases in the count model

The 100X error-free reads simulated from alleles with and without 10kb flanks at varying read and fragment lengths were used to assess potential length biases in the count genotyping model. Average F1 scores were calculated for each set across a range of *k*-mer lengths (11–99 in increments of four) as described in **Methods 2.4.3.4**.

2.4.3.6 Additional approaches to determine likelihoods for the count model

Two additional methods to determine *PRDM9*-36 likelihoods using the count model were tested:

1) Using the original `Cortex` formula for estimating λ instead of the modified formula (see **Methods 2.4.3.1**)
2) Calculating the Pearson correlation coefficient between the vector of read *k*-mer counts and the vectors of *k*-mer counts for each allele instead of using the modified Poisson probability mass function, where the allele that resulted in the largest Pearson correlation coefficient was considered the called allele.

The 20X and 100X simulations from the primary haploid simulation set were used to test each approach and average F1 scores were calculated as previously described (**Methods 2.4.3.4**).

A pair HMM was tested as a genotyping model (**Figure 2.25**). The HMM had five states: **Begin**, **End**, **Align**, **Insert**, and **Delete** (Durbin, R., Eddy, S., Krogh, A., & Mitchison, G. 1998). The begin and end states were silent and thus had no emission probabilities, while the insert and delete states each had one emission (gap) with a probability of 1. The align state had

two possible emissions, match or mismatch, with probabilities of 0.9999999999 and 0.0000000001, respectively. The mismatch emission probability reflects the sequencing base error rate and was given a very small value to avoid multiplication by zero for the error-free simulations assessed. The transition probabilities reflect other Illumina sequencing error rates: δ = the gap opening rate (0.0001%), ε = the gap extension rate (0.0001%), and τ = the termination rate (1%).

The begin state had transitions to the align, insert, and end states, with transition probabilities of $1 - \delta - \tau$, δ, and τ, respectively. The align state had transitions to itself ($1 - 2\delta - \tau$) and the insert (δ), delete (δ), and end (τ) states. The insert state had transitions to itself (ε) and to the align ($1 - \varepsilon - \tau$) and end (τ) states, while the delete state had transitions to itself (ε) and to the align ($1 - \varepsilon$) state. The model did not allow for transitions between the begin and delete states or between the delete and end states because a deletion at the start or end of a sequencing read is not possible.

All reads from a given simulation replicate were aligned to all *PRDM9*-36 allele sequences using the pair HMM. The **forward algorithm** was used to determine the probability of observing a single read from an allele by summing over all possible alignments between the read and the allele. The probabilities from each read in a simulation set were then multiplied to obtain an overall likelihood of observing the set of reads from a given allele, and the allele with the highest likelihood was called as the allele for that simulation. 50 replicates of the 20X error-free simulations without 10kb flanks were used to test the HMM and average F1 scores were calculated as previously described (**Methods 2.4.3.4**).

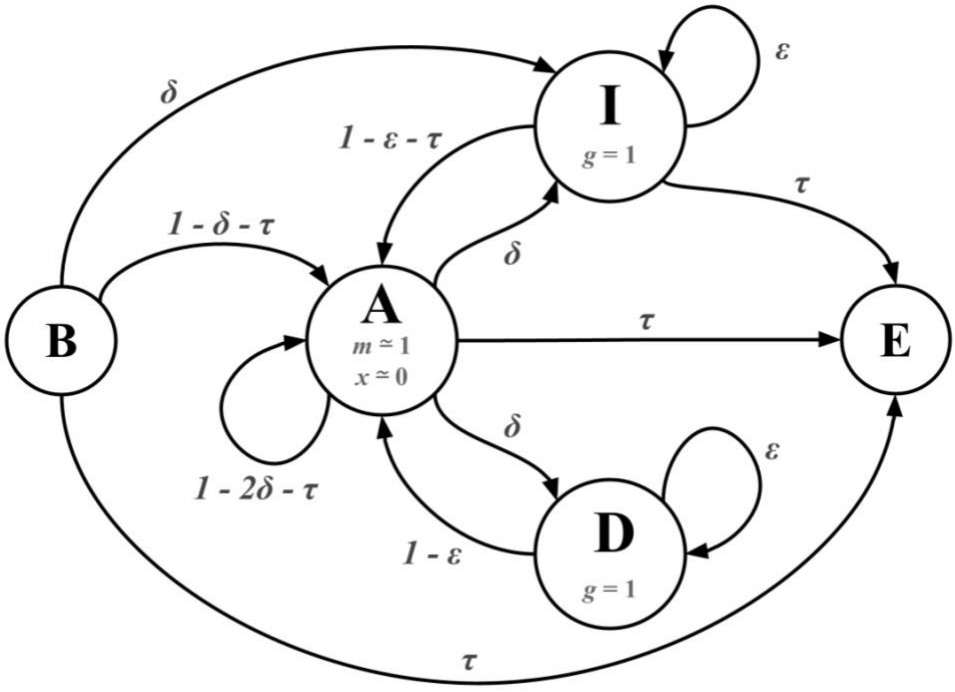


Figure 2.25: Modified pair HMM used to call *PRDM9* alleles. States are represented by circles: *B*, Begin; *A*, Align; *I*, Insert (in the read sequence); *D*, Delete (in the read sequence); and *E*, End. Transition probabilities are represented by arrows and reflect Illumina sequencing error rates: δ, gap opening rate (0.000001); ε, gap extension rate (0.000001); and τ, termination rate (0.01). There are no arrows connecting *B* to *D* or *D* to *E* since a deletion at the end of a read is not possible. *B* and *E* are silent states and have no emission probabilities, while *I* and *D* each have one emission (*g*, gap) with a probability of 1, and *A* has two possible emissions (*m*, match; *x*, mismatch) with probabilities of ~ 1 and ~ 0, respectively, for error-free simulations.

2.4.3.7 Comparison of allele *k*-mer count profiles

To determine if each *PRDM9*-36 allele had a unique *k*-mer count profile, I compared the full count profiles for all 630 possible pairings of alleles (not including same allele pairings, e.g. A vs A) at all values of *k* from one to 400. Count profiles for two alleles were considered non-unique if all counts for all *k*-mers matched between the two at a particular *k*-mer length. The same was done for all 221,445 possible pairings of diploid genotypes.

2.4.4 Initial testing of real sequencing data

2.4.4.1 Obtaining the Ashkenazi trio samples and sample genotypes

Both short- and long-read bam files for the GIAB Ashkenazi Jewish trio (HG002, HG003, HG004) were downloaded (**Appendix Table 3**). There were two types of Illumina data: 2x250bp at ~50x coverage and 2x150bp at ~300x coverage. Bam files were also available for the 2x150bp data downsampled by GIAB to ~60X. The long-read data were from whole-genome PacBio Hifi sequencing. Both the Illumina and PacBio bams were subset to contain only reads aligned to the *PRDM9* + 10kb flanks region.

Viewing the PacBio alignments in `IGV` (Robinson et al. 2011), two of the samples had the most frequent genotype A/A, but HG004 had a heterozygous SNP at a position that indicated it had the genotype A/L37 (see **Figure 2.15**). These were considered to be the truth genotypes for the samples because many of the highly accurate HiFi reads spanned the full length of the repetitive zinc finger region. I then used the count model with varying *k*-mer lengths (11–147 in increments of four) to see if the genotypes called with the highest likelihoods matched the truth genotypes for the samples. All three methods of estimating λ were assessed. For the coverage method, estimates depth of coverage across the *PRDM9* + 10kb flank region were calculated per sample using `samtools coverage -Hr chr5:23516673-23537764`, and estimates for sequencing error rate were calculated using `samtools stats` with a text file containing the *PRDM9* + 10kb coordinates in GRCh38 given to the `-t` parameter (Li et al. 2009). The 2x250bp data sets had estimated coverages closer to 60X than 50X as described by GIAB.

2.4.4.2 Assessing differences in *k*-mer coverage between the zinc finger and flanking regions

Counts were determined for *71*-mers for each 2x250bp Ashkenazi trio sample and for reference allele B using `jellyfish count -m 71 -s 100` (specifying an initial hash size of 100) and `jellyfish dump` (Marçais and Kingsford 2011). For each uniquely occurring *71*-mer

across the *PRDM9* + 10kb flanks in allele B, the number of reads containing that *k*-mer was determined for the trio samples.

The proportion of GC bases in the allele B *71*-mers was calculated across 250bp windows to assess the potential of a skewed GC proportion causing higher sequencing error rates in the zinc finger region. GC% was also determined for all *71*-mers from the flanking sequences, using bins of < 20%, 20–29%, 30–39%, 40–49%, 50–59%, 60–69%, and ≥ 70%.

Finally, the ratio of *71*-mer counts to per-base read depth was also determined across the *PRDM9* + 10kb region for both the PacBio HiFi and 2x250bp Illumina reads for each sample.

2.4.4.3 Performing error corrections on short reads

Two published software tools with short-read error-correction functionalities were applied to the reads of each of the trio samples: `SPAdes` (Prjibelski et al. 2020) and `GraphAligner` (Rautiainen and Marschall 2020). Read correction with `SPAdes` used the parameter `--only-error-correction`. For `GraphAligner`, a sequence graph (see **Chapter 1** section **1.4.1**) was generated for the GRCh38 reference *PRDM9* sequence (allele B zinc finger region + 10kb flanks) and modified to have nodes that are no longer than 1,024bp using the command `vg mod -X 1024` (Garrison et al. 2018). The parameters `-g GRCh38-flank10k-chop1024.gfa -x vg --corrected-out` were used to specify the graph name, the graph style to use for correction, and to output the corrected reads instead of an alignment, respectively.

In addition, an in-house read-correction script called `AlignCorrect` (Jared Simpson, unpublished) was applied to the reads. `AlignCorrect` works by aligning each read to a supplied list of allele sequences, in this case the *PRDM9*-36 alleles with 10kb flanks. For each read, the closest allele is determined using the lowest edit distance value between the read and the allele. If the edit distance is below a specified threshold (40 in this case), the read sequence is replaced with the allele sequence subset where the read best aligned. In cases where two or more alleles have the same best edit distance, the allele appearing first in the provided list is used. This allows for the correction of sequencing errors and somatic mutations while allowing

for mismatches and indels relative to the reference to be checked against non-reference alleles as potential non-reference variants; variants in the reads also observed in the closest matching allele are not "corrected" as they would be if only the reference allele sequence was used.

To compare the different methods of correcting reads, all *71*-mer counts in the corrected reads were collected with `Jellyfish` as before. The *k*-mer counts in the reads were plotted for the unique *71*-mers in allele B (i.e. *71*-mers that occur only once in the allele B + 10kb flank sequence). Plots for allele B *k*-mer counts were generated for the raw, `SPAdes`-corrected, `GraphAligner`-corrected, and `AlignCorrect`-corrected HG003 2x250bp reads.

2.4.4.4 Using manually specified λ estimates for the count model

To see at which values of λ a sample could be called correctly, I supplied a large range of values that surrounded the value estimated by the count-coverage method for the corresponding sample parameters of coverage, read length, and error rate. The estimates used the modified `Cortex` *k*-mer coverage equation (see **Methods 2.4.3.1**). For a specific λ value in the range provided, I calculated what the approximate error rate was by rearranging that formula:

$$\varepsilon = 1 - \left(\frac{\lambda}{(l - k + 1) \cdot \frac{v}{r}}\right)^{\frac{1}{k}}$$

where ε = the substitution error rate, λ = the specified *k*-mer coverage, *l* = the read length, *k* = the *k*-mer length, and *v* = the average sequencing read coverage.

Ranges of manually supplied λ values were tested on the 100X simulations for genotypes A/A and A/L37 from the primary diploid simulation set using *k*-mer lengths of 51 and 71. In addition, a range of λ values was provided to test the Ashkenazi trio 2x150bp 60X data (downsampled from 300x) data with several *k*-mer lengths (51–131 in increments of 20). For the trio, I estimated what λ would be with a 0%, 0.1%, and 1% sequencing error rate and 60X coverage using the modified `Cortex` formula, and chose the range to be from the lowest calculated λ from the error rate and coverage combinations to the highest, in increments of 0.1. This approach was repeated after correcting the 2x150bp 60X reads. For each condition, the

genotype likelihoods were ranked such that the genotype with the highest likelihood was ranked first. The ranks for genotypes A/A and A/L37 were assessed to see if the trio samples were called correctly.

Chapter 3

A haplotype-based model for long-read genotyping and validation of the short-read genotyping models

3.1 Background

Long-read sequencing technologies allow for better mappability of reads to the reference, which is highly advantageous for studying repetitive regions of the genome. Long reads have been used to generate more accurate de novo genome assemblies (Koren et al. 2018), close gaps in the human reference genome (Nurk et al. 2022), target complex regions of the genome (Buermans et al. 2017), improve genome phasing (Cretu Stancu et al. 2017), and identify SVs missed by Illumina sequencing (Chaisson et al. 2019). One such technology, PacBio HiFi sequencing, is able to produce reads with an average of 13.5kb in length that are 99.8% accurate, thanks to the continuous sequencing of individual DNA molecules multiple times and subsequent formation of a consensus sequence for each read (Wenger et al. 2019). Such highly accurate reads long enough to span the full zinc finger array of *PRDM9* would greatly improve read mapping and variant identification for the polymorphic gene.

While numerous programs exist to call variants in short-read sequencing data, they are not able to be used with long-read data due to the different sequencing error profiles of the two read types, in terms of both the higher error rate and the propensity for small indels over base errors observed in PacBio reads (De Coster et al. 2021). In addition, because short-read models usually work in small window sizes, they are unable to take advantage of the haplotype information available within long reads (Edge and Bansal 2019). As such, many long-read specific variant callers phase reads into haplotypes as a variant-calling strategy. Some examples are `Medaka` (https://github.com/nanoporetech/medaka) and `Clair3` (Zheng et al. 2021) for small variants, and `PRINCESS` (Mahmoud et al. 2021) which is able to call SNVs and SVs. `Longshot` uses phasing information in combination with a pair HMM to account for alignment uncertainty and calculate genotype likelihoods (Edge and Bansal 2019). Some tools do not rely on phasing, such as `DeepVariant` which calls SNVs and small indels using a convolutional neural network trained on human data (Poplin et al. 2018a). `Sniffles` is an

alignment-based SV caller that identifies potential SVs in each read and then clusters SV calls on supporting reads (Sedlazeck et al. 2018). Another alignment-based caller is `cuteSV`, which does not rely on high coverage in order to call SVs (Jiang et al. 2020). Performing de novo assembly of long reads into a genome and then using the assembly in comparison to GRCh38 is another option for calling variants, particularly when using diploid assembly software such as `hifiasm` (Cheng et al. 2021) as a way to preserve haplotypes instead of condensing heterozygous sites into a single consensus sequence.

Calling *PRDM9* genotypes is in part a SNV-identification problem and a CNV-identification problem, and would be greatly simplified with haplotype phasing information, as evident by how haploid allele calling was more accurate than diploid genotype calling for simulated data (see **Chapter 2 Results 2.2.4–2.2.5**). However, since the zinc finger repeat region is not extremely long and the zinc finger repeat insertions and deletions usually occur in 84bp blocks, complex methods that can identify messy chromosomal breakpoints are likely not needed. Instead, a simple realignment-based model that uses edit distances to known sequences to determine variant likelihoods or a model that generates consensus sequences from a POA graph could be feasible.

In this chapter I detail the development of two long-read models used to validate the genotype calls of the short-read models described in **Chapter 2**. In contrast to the short subsequences of information provided by *k*-mers, one model relies on the realignment of long reads to known *PRDM9* allele sequences, taking advantage of the ability for reads to span the full length of the zinc finger array and removing ambiguity of where each repeat occurs in the variable domain. The other model uses a POA graph to generate one or two consensus sequences that best summarize the long reads and further checks for potential novel alleles. I validate the accuracy of these models by comparing the calls to a thorough genotyping-by-eye approach in `IGV`. I then use the results from the models to validate the accuracy of the short-read models developed in **Chapter 2**.

3.2 Results

3.2.1 Collecting samples with both long-read and short-read sequencing data

Both whole-genome Illumina and whole-genome PacBio HiFi were collected from multiple publicly available sources: 44 samples from The Human PanGenome Reference Consortium (HPRC) (Wang et al. 2022) year 1 freeze, four samples from the 1000 GP Structural Variant (1000GP-SV) consortium (Fairley et al. 2020), and five samples from the GIAB consortium (Zook et al. 2016). Collectively, I refer to these as the **HPRC++** dataset. Additionally, access was granted to 50 samples from the Ontario Health Study (OHS) cohort (Kirsh et al. 2022). The samples from the **OHS** were selected based on estimated ethnicity and the genotype of the downstream SNP rs6889665, which is associated with longer *PRDM9* alleles and the presence of the 'k' zinc finger, and is not present in either the reference or the most frequent allele (Hinch et al. 2011). Both whole-genome Illumina and targeted PacBio Hifi sequencing was performed on the OHS samples. Full sample collection and preparation steps are described in **Methods 3.4.1** and **3.4.2**. One sample from each dataset was removed for having insufficient reads after filtering, leaving 52 HPRC++ and 49 OHS samples to analyze. Summative demographics and characteristics for the samples are described in **Table 3.1**.

Table 3.1: *PRDM9* information for the HPRC++ and OHS datasets. Two datasets with short- and long-read sequencing data were used to validate the short-read genotyping models from **Chapter 2**: HPRC++ (HPRC, 1000GP-SV, and GIAB) and OHS. Admixed American* and European ancestry were not assessed in OHS samples; samples not clustered with the African, East Asian, or South Asian groups were classified as "Other". The alternate allele for SNP rs6889665 is associated with longer non-reference alleles and/or alleles containing the 'k' zinc finger. *PRDM9* data were determined by visualizing the PacBio HiFi alignments to GRCh38 in `IGV` and confirmed after final genotypes were assigned (**Results 3.4.3.3–3.4.3.4**). *Note: The 1000GP refers to this continental population as "American", but in this thesis it is referred to as "Admixed American" for clarification as the populations are of admixed European, African, and Indigenous American ancestries. Additionally, three Ashkenazi samples in the HPRC++ cohort are grouped as European.

Characteristic			Number of samples		
			HPRC++	OHS	Total
Total sample size			52	49	101
Estimated ethnicity	**African**		22	25	47
	Admixed American		17	NA	17
	East Asian		8	10	18
	European		4	NA	4
	South Asian		1	11	12
	Other		NA	3	3
***PRDM9* genotype zygosity**	**Heterozygous**		28	13	41
	Homozygous		24	36	60
rs6889665 genotype	**0/0**		36	8	44
	0/1		14	33	47
	1/1		2	8	10
Number of rare *PRDM9* alleles (non-A/non-P001)		**0**	24	24	48
		1	21	11	32
		2	7	14	21
Number of *PRDM9* alleles shorter than the reference allele (B/P002)		**0**	51	44	95
		1	1	2	3
		2	0	3	3
Number of *PRDM9* alleles longer than the reference allele (B/P002)		**0**	33	40	73
		1	16	3	19
		2	3	6	9

3.2.1.1 Filtering noisy reads resulting from PCR laddering during amplification for targeted PacBio Hifi sequencing

Due to the repetitive nature of the *PRDM9* zinc finger array and the necessary PCR amplification steps during preparation for targeted PacBio sequencing, several of the resulting reads from the OHS samples showed signs of **laddering**. This is an issue with repetitive regions and has been observed in targeted *PRDM9* sequencing previously (Alleva et al. 2021), as well as in other repetitive genomic regions (Briggs et al. 2012). A filtering pipeline was developed to remove noisy reads and retain those most likely to be true reflections of the zinc finger region (**Methods 3.4.2.3**). Briefly, reads were removed if they did not span the full zinc finger repetitive region, if they contained indels at positions found in < 25% of the reads, or if they were not of the expected read length (2,041bp including PCR primers) ± a multiple of the length of a zinc finger repeat (84bp). This is a highly conservative filtering strategy, removing an average of 23,997 or 80% of the reads from the samples (minimum: 58%; maximum: 94%), but the extremely high depth of coverage from the targeted HiFi sequencing meant there were ample reads still remaining that spanned the full zinc finger repeat region (mean: 6,063; minimum: 1,026; maximum: 21,832).

3.2.2 Determining the truth genotypes of the samples

In order to validate the accuracy of the realignment and consensus models, the **truth genotypes** for the samples needed to be determined. As mentioned in **Results 3.2.2**, the current best strategy is to align the reads to all known allele sequences. I generated a fasta file containing all known allele sequences (zinc finger region + 10kb flanks) to use as a ***PRDM9*-specific reference** that would allow for easier visualization of the alignments to several alleles in an `IGV` session, particularly for samples with an allele shorter or longer than the reference allele (**Figure 3.1**). The *PRDM9*-106 list of known alleles (see **Chapter 2 Results 2.2.1**) was used to generate the *PRDM9*-specific reference to ensure an up-to-date list of possible variants was available for the ethnically diverse samples.

The HPRC++ samples were first filtered to keep only reads than fully spanned the *PRDM9* zinc finger domain region to concentrate the reads of interest. Each HPRC++ and OHS sample

was realigned to the *PRDM9*-specific reference. The two alleles with the most aligned reads were used to generate a **genotype-specific reference** for each sample, to which reads were again realigned. The proportions of reads for each allele in the genotype-specific alignments were then determined. If an allele had at least 25% of the reads aligned to it, it was conditionally considered as part of the genotype. The genotype-specific alignments were then viewed in `IGV` for visual confirmation for each sample. If the reads had no SNVs or indels present in at least 25% of the reads, the allele was confirmed as part of the truth genotype. If there were SNVs or indels in a sufficient number of reads in consistent locations, the allele was flagged as novel (e.g. Novel_Similar_P001), as was the case for seven samples (**Methods 3.4.3**). The truth genotypes are listed in column two in **Table 3.2** and **Table 3.3** for the HPRC++ and OHS samples, respectively.

Figure 3.1: Effect of *PRDM9* zinc finger indels on long-read alignments. PacBio HiFi reads from sample HG01891 aligned to different references as visualized in `IGV`. Alignments are centered between the starting (green bar) and ending (pink bar) zinc fingers in *PRDM9* (bottom panel). The grey horizontal lines are sequencing reads, the long vertical colored bars are SNVs, and the purple blocks are insertions. The black vertical line in the middle depicts the centre of the alignment view (i.e. not a variant). **A)** Alignments to the GRCh38 reference. The presence of multiple SNVs and 84bp insertions suggests the genotype is not A/A (P001/P001) and consists of at least one allele that is longer than the reference allele. Certain SNV and insertions are in consistent locations across many reads (dark grey arrows). However, other elements are not in consistent locations and it is unclear if these are due to misalignment of some reads or if they are specific to only a few reads as either sequencing errors or somatic mutations (light grey arrows). **B)** Alignments to allele P015 (L6) in the genotype-specific reference genome for HG01891. The reads do not have any large indels or SNVs within the zinc finger region, suggesting these reads map perfectly to allele P015.

3.2.3 Developing models for genotyping long-read data

Genotyping *PRDM9* in samples using long reads is easier than with short reads because most long reads span the entire zinc finger region, which is 1,596bp for the longest known *PRDM9*-106 allele. For alleles that are the same length as the reference allele, it is a matter of determining where SNVs occur in the alignments and comparing them to the sequences of other zinc fingers located at that position in other alleles, which is not trivial given the potential for two alleles to be the same length but have vastly different zinc finger arrangements. For alleles longer or shorter than the reference allele, however, this becomes even less clear as the zinc finger indels do not always line up neatly; the high similarity of sequence at the starts and ends of zinc fingers results in multiple ways for an aligner to place the indel (see **Figure 3.1A**). Additionally, given the large number of *PRDM9* variants, it is not trivial to determine which SNVs and indels present in a sample belong to which allele. *PRDM9* is typically genotyped by realigning long sequencing reads to known allele sequences, where the allele or alleles that resulted in the best alignment are called as the sample genotype.

To simplify this process, I developed two different genotyping models that take advantage of long reads. The first was the **realignment model** that assigns likelihoods to each possible genotype based on the differences between sequencing reads and known allele sequences (**Figure 3.2**). Briefly, all reads are realigned to each allele provided in a fasta file. The **edit distances** for each alignment are used to calculate the probability of observing each read from each possible genotype. A total of four models were tested to determine the best way to calculate likelihood using edit distances: the **mismatch model**, which only considers the bases that did not match in the alignment; the **match-mismatch model**, which considers all bases in the alignment; and each model with a **modified edit distance** scoring approach, where indels were only considered to have an edit distance of one instead of a value equal to the length of the indel. For both datasets, there was no difference between using the match-mismatch or mismatch model, but the edit distance method did matter: the regular edit distance gave better results for the HPRC++ samples while the modified edit distance gave better results for the OHS samples, possibly due to the issues with PCR laddering. The results described herein were obtained using the match-mismatch with regular edit distance model for the HPRC++

dataset and the match-mismatch with modified edit distance model for the OHS dataset (**Methods 3.4.4.1**).

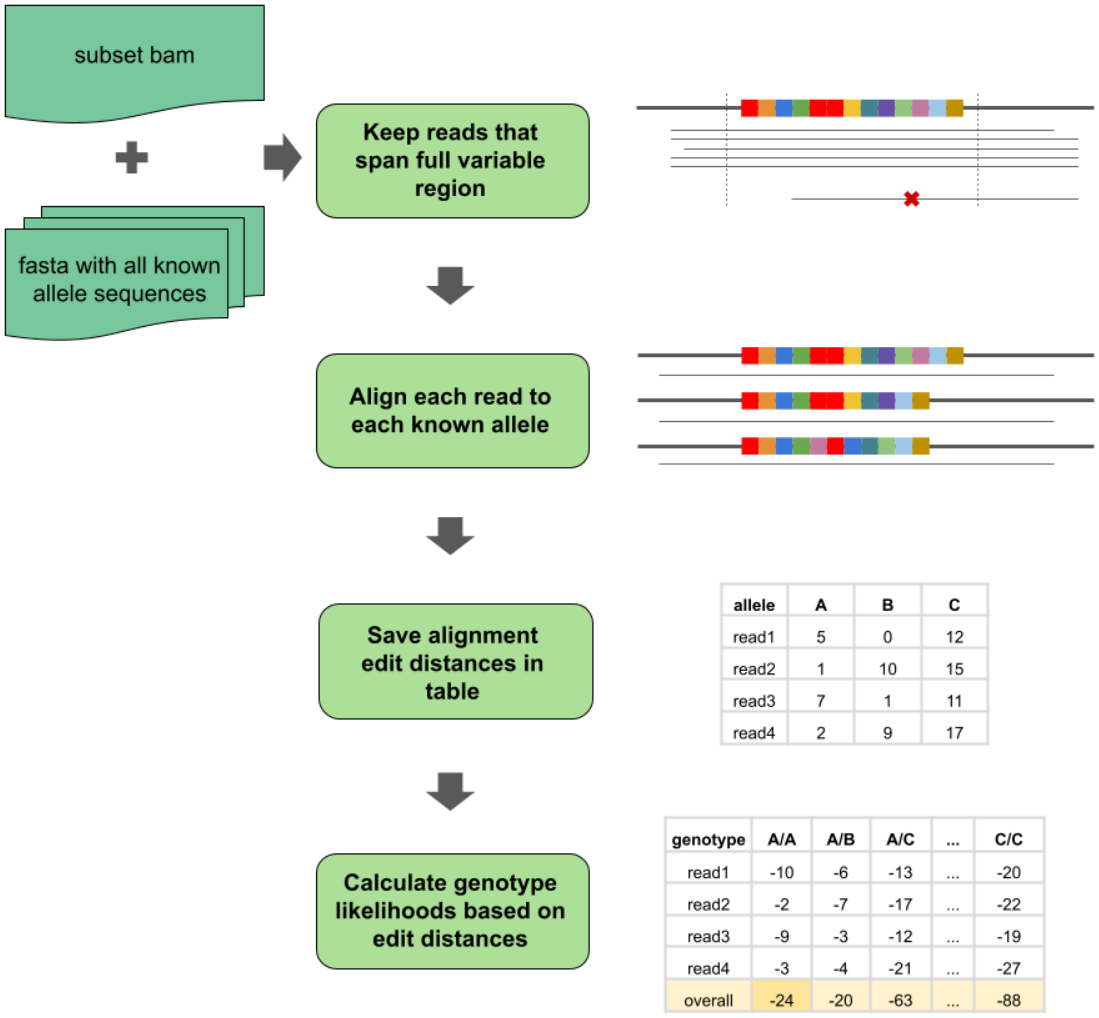


Figure 3.2: The realignment model for genotyping with long reads. Reads are first filtered to retain only those that span into the 50bp flanks around the zinc finger region (GRCh38 coordinates). Each read is realigned to all known *PRDM9*-106 alleles and edit distances were determined. Using the edit distances from each allele in a genotype, all possible genotypes were assigned a likelihood.

The second approach developed was a **consensus model** that generates a POA graph (introduced in **Chapter 1** section **1.4.1**) of the long reads that span the full length of the zinc finger domain and then determines one or two consensus sequences that best summarize the graph (**Figure 3.3**). The number of consensus sequences is dependent on the allele frequencies of polymorphic loci in the POA graph: a base needs to have a frequency of at least 0.25 in the sequencing reads to be considered valid. The model then compares the consensus sequences to all known alleles to flag any novel alleles. If a consensus sequence does not match any allele, the call is tagged as "**novel similar to**" the allele closest in edit distance to the consensus sequence (e.g., Novel_Similar_A). If more than one allele has an equal edit distance to the consensus sequence, the allele occurring first in the provided fasta file is used in the novel tag (**Methods 3.4.4.2**).

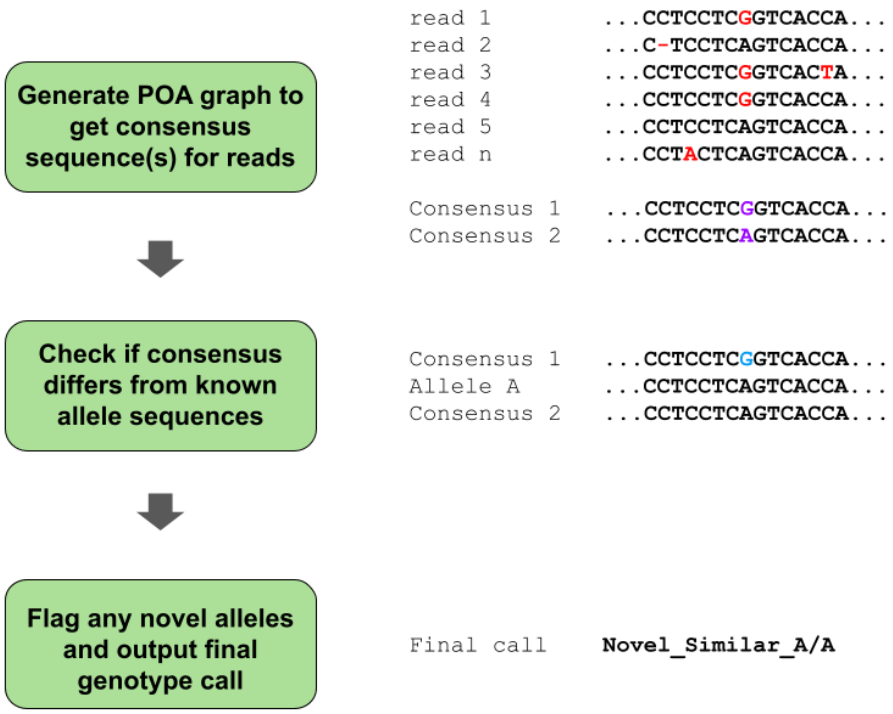


Figure 3.3: The consensus model for genotyping with long reads. A POA graph is generated using all of the reads that fully span the zinc finger repeat region, which is then condensed down into one or two consensus sequences. Each consensus sequence is then compared to each known allele sequence. If a consensus sequence is not a perfect match to a known allele, it is flagged as novel with a similarity to the allele with the smallest edit distance to the consensus sequence.

3.2.4 The realignment and consensus models are effective at genotyping and identifying novel alleles using long-read data

The realignment and consensus models were developed to simplify the genotyping process for highly polymorphic genomic regions when long-read data are available. The efficacy of the models was determined by comparing the model calls to the meticulously determined truth genotypes for the HPRC++ and OHS samples (**Methods 3.4.4.3**). Given the ethnic diversity of the samples, the number of alternative rs6889665 alleles within the samples, and the preliminary visualizations of alignments in `IGV`, several different *PRDM9*-106 alleles were expected to be identified. Additionally, given the rapid evolution of *PRDM9* (Ponting 2011) and the identification of 27 new alleles recently described (Alleva et al. 2021; see **Chapter 5** section **5.1.5** for why their reported number of 32 is incorrect), it was anticipated that some novel alleles could be present in the samples as well.

Both the realignment and consensus models were very successful at genotyping the HPRC++ samples. Since the realignment model is unable to flag novel alleles, calls for samples with novel alleles were considered correct if the genotype contained the **base allele** of the novel tag (e.g. P001 in Novel_Similar_P001). All 52 HPRC++ samples were called correctly using both models, and the consensus sequence correctly flagged all seven novel alleles identified in the truth genotypes (**Table 3.2**). The consensus model also did very well for the 49 OHS samples: 47 samples were called correctly, including flagging the one novel allele, while two heterozygous samples were incorrectly called as homozygous with one correct allele from the truth heterozygous genotype (**Table 3.3**). There was less success in genotyping the OHS samples with the realignment model: 24 samples were called correctly, including the base allele of the novel allele, while 25 homozygous samples were incorrectly called as heterozygous by the realignment model with one allele correctly in the truth homozygous genotype. Interestingly, while the miscalled genotypes involved seven different alleles, there were only three incorrect additional alleles: P025 (L16), P034, and P081. None of these incorrect alleles are particularly similar to the truth alleles; the average edit distance between the correct and incorrect pairings was 60 (minimum: 2; maximum: 98).

Overall, the consensus model successfully genotyped 99/101 or 98% of the samples while the realignment model successfully genotyped 76/101 or 75% of the samples. Notably, all samples had at least one allele correct in the called genotypes. In addition, all of the incorrect calls were for the OHS samples, which is likely due to the PCR laddering and subsequent filtering process for reads, making these samples arguably more difficult to call.

Table 3.2: Genotyping performance of the long-read models on the HPRC++ samples. Truth genotypes and genotypes called by the realignment and consensus models for the HPRC++ samples using the *PRDM9*-106 list of alleles. Truth genotypes were determined by realigning samples to *PRDM9*-specific and genotype-specific references and visualization in `IGV`. All samples were called correctly (green) by the consensus model. The realignment model also called all samples correctly, taking into consideration that the base alleles of the novel alleles were called (yellow) since the realignment model is unable to identify novel alleles.

Sample	Truth genotype	Genotype calls	
		Realignment model	Consensus model
HG001	P001/P001	P001/P001	P001/P001
HG002	P001/P001	P001/P001	P001/P001
HG003	P001/P001	P001/P001	P001/P001
HG004	P001/P035	P001/P035	P001/P035
HG005	P001/P001	P001/P001	P001/P001
HG006	P001/P001	P001/P001	P001/P001
HG007	P001/P001	P001/P001	P001/P001
HG00438	P001/P001	P001/P001	P001/P001
HG00514	P001/P001	P001/P001	P001/P001
HG00621	P001/P001	P001/P001	P001/P001
HG00673	P001/P001	P001/P001	P001/P001
HG00731	P001/P028	P001/P028	P001/P028
HG00732	P001/P001	P001/P001	P001/P001
HG00733	P001/P001	P001/P001	P001/P001
HG00735	P001/P003	P001/P003	P001/P003
HG00741	Novel_Similar_P001/P001	P001/P001	Novel_Similar_P001/P001
HG01071	P001/P029	P001/P029	P001/P029
HG01106	P001/P033	P001/P033	P001/P033
HG01109	P001/P607	P001/P607	P001/P607
HG01175	Novel_Similar_P015/P001	P001/P015	Novel_Similar_P015/P001
HG01243	P001/P001	P001/P001	P001/P001
HG01258	P001/P001	P001/P001	P001/P001
HG01358	Novel_Similar_P002/P033	P002/P033	Novel_Similar_P002/P033
HG01361	P001/P023	P001/P023	P001/P023
HG01891	P015/P015	P015/P015	P015/P015
HG01928	P001/P042	P001/P042	P001/P042
HG01952	P001/P001	P001/P001	P001/P001
HG01978	P001/P001	P001/P001	P001/P001
HG02055	P001/P001	P001/P001	P001/P001
HG02080	P001/P001	P001/P001	P001/P001
HG02145	P003/P026	P003/P026	P003/P026
HG02148	P001/P003	P001/P003	P001/P003
HG02257	P001/P015	P001/P015	P001/P015

HG02572	Novel_Similar_P009/P016	P009/P016	Novel_Similar_P009/P016
HG02622	P001/P028	P001/P028	P001/P028
HG02630	P001/P001	P001/P001	P001/P001
HG02717	P001/P003	P001/P003	P001/P003
HG02723	P001/P001	P001/P001	P001/P001
HG02818	P001/P003	P001/P003	P001/P003
HG02886	P001/P001	P001/P001	P001/P001
HG02970	P001/P002	P001/P002	P001/P002
HG03453	P001/P001	P001/P001	P001/P001
HG03486	P001/P027	P001/P027	P001/P027
HG03492	P001/P001	P001/P001	P001/P001
HG03516	P001/P028	P001/P028	P001/P028
HG03540	P001/P035	P001/P035	P001/P035
HG03579	Novel_Similar_P015/P001	P001/P015	Novel_Similar_P015/P001
NA18906	Novel_Similar_P613/P035	P035/P613	Novel_Similar_P613/P035
NA19030	P035/P611	P035/P611	P035/P611
NA19240	Novel_Similar_P613/P003	P003/P613	Novel_Similar_P613/P003
NA20129	P001/P003	P001/P003	P001/P003
NA21309	P001/P035	P001/P035	P001/P035

Table 3.3: Genotyping performance of the long-read models on the OHS samples. Truth genotypes and genotypes called by the realignment and consensus models for the OHS samples using the *PRDM9*-106 list of alleles. Truth genotypes were determined by realigning samples to *PRDM9*-specific and genotype-specific references and visualization in `IGV`. The consensus model called 47 samples correctly (green) and miscalled two heterozygous samples as homozygous (blue). The realignment model called 24 samples correctly, taking into consideration that the base allele of the novel allele was called (yellow) since the realignment model is unable to identify novel alleles. The remaining 25 samples were homozygous but incorrectly called as heterozygous (pink) by the realignment model.

Sample	Truth genotype	Genotype calls	
		Realignment model	Consensus Model
AWA4634-584	P029/P029	P029/P029	P029/P029
AWA4634-802	P021/P021	P021/P081	P021/P021
AWA4634-1285	P003/P616	P003/P616	P003/P616
AWA4634-1383	P001/P001	P001/P025	P001/P001
AWA4634-1463	P003/P003	P003/P034	P003/P003
AWA4634-1522	P001/P001	P001/P081	P001/P001
AWA4634-1594	P003/P003	P003/P081	P003/P003
AWA4634-1660	P002/P002	P002/P081	P002/P002
AWA4634-1921	P001/P002	P001/P002	P001/P001
AWA4634-2087	P001/P003	P001/P003	P001/P003
AWA4634-2096	P001/P001	P001/P081	P001/P001
AWA4634-2198	P001/P001	P001/P025	P001/P001
AWA4634-2683	P001/P001	P001/P081	P001/P001
AWA4634-2802	P001/P002	P001/P002	P002/P002
AWA4634-3383	P001/P001	P001/P001	P001/P001
AWA4634-3640	P001/P001	P001/P081	P001/P001
AWA4634-3825	P016/P016	P016/P016	P016/P016
AWA4634-3975	P001/P002	P001/P002	P001/P002
AWA4634-4005	P003/P003	P003/P025	P003/P003
AWA4634-4206	P001/P016	P001/P016	P001/P016
AWA4634-4551	P001/P001	P001/P001	P001/P001
AWA4634-4646	P001/P024	P001/P024	P001/P024
AWA4634-4893	P001/P001	P001/P081	P001/P001
AWA4634-5177	P003/P003	P003/P025	P003/P003
AWA4634-5213	P020/P020	P020/P034	P020/P020
AWA4634-5509	P001/P001	P001/P001	P001/P001
AWA4634-5567	P001/P001	P001/P001	P001/P001
AWA4634-5623	P001/P001	P001/P081	P001/P001
AWA4634-5752	P001/P001	P001/P001	P001/P001
AWA4634-5968	P001/P001	P001/P001	P001/P001
AWA4634-6046	P016/P016	P016/P025	P016/P016
AWA4634-6247	P001/P001	P001/P001	P001/P001

AWA4634-6273	P002/P002	P002/P002	P002/P002
AWA4634-6515	P001/P001	P001/P081	P001/P001
AWA4634-6735	P001/P001	P001/P001	P001/P001
AWA4634-6962	P003/P616	P003/P616	P003/P616
AWA4634-8312	P001/P003	P001/P003	P001/P003
AWA4634-8653	P001/P001	P001/P081	P001/P001
AWA4634-8881	P001/P001	P001/P081	P001/P001
AWA4634-8948	P001/P001	P001/P081	P001/P001
AWA4634-9128	P001/P001	P001/P025	P001/P001
AWA4634-9316	P001/P029	P001/P029	P001/P029
AWA4634-9323	P001/P001	P001/P081	P001/P001
AWA4634-9326	P001/P616	P001/P616	P001/P616
AWA4634-9339	P001/P029	P001/P029	P001/P029
AWA4634-9347	P024/P024	P024/P081	P024/P024
AWA4634-9378	Novel_Similar_P016/P001	P001/P016	Novel_Similar_P016/P001
AWA4634-9381	P001/P001	P001/P081	P001/P001
AWA4634-9407	P001/P001	P001/P081	P001/P001

3.2.5 The short-read *k*-mer models are successful in calling over 40% of the samples

The short-read models were developed and fine-tuned using simulated sequencing data (see **Chapter 2**), which are not a fully realistic representation of true sequencing data, particularly in repetitive parts of the genome like the *PRDM9* zinc finger region. The 101 samples across the HPRC++ and OHS datasets were sequenced with both short-read Illumina and long-read PacBio Hifi technology, providing a set of samples that are ethnically diverse with a variety of *PRDM9* alleles that can be used to assess the short-read models with the accurate long reads. Validation of the short-read models was determined by comparing the genotypes obtained from the short-read models to the truth genotypes. Since the short-read models are unable to identify novel alleles, only the base alleles from the truth genotypes were assessed in the comparisons.

Using the Illumina reads for each sample, the three *k*-mer count methods, four distance methods, and 12 combinations of count and distance methods were all assessed over a wide range of *k* values. For the count-coverage model and for all of the distance models, estimates of sequencing coverage, error rate, mean fragment length, and fragment length standard deviation were determined across the *PRDM9* zinc finger + 10kb flank region for each

HPRC++ sample, or across the full read length for the OHS samples (which were shorter due to targeted sequencing). After assessing the raw reads, the models were rerun on the samples after correction with `AlignCorrect` (Jared Simpson, unpublished; see **Chapter 2 Methods 2.4.4.2**). Results are broken down by three criteria: the number of samples that were genotyped correctly, the number of samples where one allele in the truth genotype was present in the called genotype, and the number of samples where the truth genotype was in the top 10 called genotypes ranked by likelihoods (**Methods 3.4.5**).

For the HPRC++ dataset (**Table 3.4**), the count model did not call any samples correctly, and at best could call one allele in the genotype for 21 samples with the coverage method. The distance-max model performed best, calling 13 samples correctly and calling one allele in an additional 38 samples. None of the combined models were able to call any samples, but several were able to call one allele in 21 of the samples. Running the short-read models after read correction improved results for the count and combined models: 18 samples were called correctly by count-flank median, and 20 for the combined coverage-max and coverage-mean methods. Read corrections impaired results for the distance model, with the max and mean methods each only able to call 7 samples correctly. Overall, 23/52 (44%) samples were called correctly by at least one short-read model.

For the OHS dataset (**Table 3.5**), the count-coverage model called 15 samples correctly and called one genotype allele for 43 samples. The distance-mean and distance-geomean methods called 13 samples correctly, while the combined count-coverage & distance-max model called 17 samples correctly. Interestingly, running the short-read models after correcting the reads impaired most of the models: none of the count and combined models were able to correctly call any genotypes. At best, the count-flank median method called one allele in 20 samples, and the combined count-coverage & distance-geomean method called one allele in 24 samples. Read corrections minimally improved results for the distance model, with the mean method able to call 19 samples correctly. The difference in the effect of read correction between the two datasets was unexpected, but given the higher coverage across the *PRDM9* + 10kb flank region for OHS samples (~64X) compared to HPRC++ samples (~32X), it is possible that the OHS reads were initially easier to call (see **Chapter 2 Results 2.2.4**) but then overcorrected. Overall, 20/49 (41%) samples were called correctly by at least one short-read model.

Table 3.4: Genotyping performance of the short-read models on HPRC++ samples. All short-read models and methods were run on the HPRC++ raw and corrected Illumina data using the *PRDM9*-106 list of alleles across a wide range of *k*-mer lengths. Results are the number of samples (out of 52) from the best outcome (i.e. the value of *k* that provided the best results). The best method per model (light yellow) and the overall best model and method (dark yellow) are highlighted for both the raw and corrected reads. Correcting the reads resulted in a substantial improvement in overall performance for most of the short-read models on the HPRC++ samples.

Model	Method	Number of samples (raw reads)			Number of samples (corrected reads)		
		Called correctly	One allele correct	Correct genotype in top 10	Called correctly	One allele correct	Correct genotype in top 10
count	**coverage**	0	21	6	17	38	39
	flank mean	0	17	0	16	39	34
	flank median	0	18	0	18	39	35
distance	**geomean**	5	29	14	6	26	18
	max	13	38	30	7	32	28
	mean	9	34	25	7	35	20
	sum	6	29	20	3	26	19
combined	**coverage & geomean**	0	18	1	18	42	40
	coverage & max	0	21	7	20	40	39
	coverage & mean	0	21	4	20	41	40
	coverage & sum	0	21	8	18	40	38
	flank mean & geomean	0	18	0	17	41	36
	flank mean & max	0	20	1	16	41	36
	flank mean & mean	0	20	1	16	41	36
	flank mean & sum	0	21	2	16	39	35
	flank median & geomean	0	18	0	15	44	35
	flank median & max	0	21	1	17	42	37
	flank median & mean	0	20	1	17	42	36
	flank median & sum	0	20	1	17	41	37

Table 3.5: Genotyping performance of the short-read models on OHS samples. All short-read models and methods were run on the OHS raw and corrected Illumina data using the *PRDM9*-106 list of alleles across a wide range of *k*-mer lengths and compared to the truth genotypes. Results are the number of samples (out of 49) from the best outcome (i.e. value of *k* that provided the best results). The best method is highlighted (light yellow) for each model, and the overall best model and method are highlighted (dark yellow) for both the raw and corrected reads. Correcting the reads resulted in poorer overall performance of the short-read models on the OHS samples.

Model	Method	Number of samples (raw reads)			Number of samples (corrected reads)		
		Called correctly	One allele correct	Correct genotype in top 10	Called correctly	One allele correct	Correct genotype in top 10
count	**coverage**	15	43	22	0	18	11
	flank mean	14	45	21	0	18	11
	flank median	14	46	21	0	20	14
distance	**geomean**	18	40	22	18	40	23
	max	8	24	21	9	29	19
	mean	18	46	22	19	46	23
	sum	0	12	4	0	13	6
combined	**coverage & geomean**	14	42	22	0	24	14
	coverage & max	17	41	22	0	20	11
	coverage & mean	16	42	22	0	22	13
	coverage & sum	11	41	22	0	17	10
	flank mean & geomean	11	45	21	0	21	14
	flank mean & max	14	45	21	0	18	13
	flank mean & mean	12	45	21	0	19	13
	flank mean & sum	9	44	21	0	20	11
	flank median & geomean	11	45	21	0	22	14
	flank median & max	14	46	21	0	19	14
	flank median & mean	12	46	21	0	21	14
	flank median & sum	9	44	21	0	19	13

3.2.6 The *k*-mer count model is able to call most samples using long-read data

The short read models use *k*-mer information and are not overly affected by read length, as long as the reads are not too short (see **Chapter 2 Results 2.2.4.1**). While the distance model relies on paired-end reads, which are not available with long-read sequencing technology, in theory the count model should work with long-read data. As a proof of concept, I ran the count model on the HPRC++ and OHS PacBio HiFi data, testing the coverage, flank mean, and flank median methods of estimating λ. The 10kb flank sequences were used with the HPRC++ samples, while the OHS data used shortened flanks (716bp left and 233bp right) due to the targeted sequencing (**Methods 3.4.5**). It worked surprisingly well: 49 of the 52 HPRC++ samples were called correctly by at least one method and at least one value of *k*. Similarly, 48 of the 49 OHS samples were called correctly under at least one condition, but only with the coverage or flank mean methods (**Figure 3.4**).

The flank mean method produced the best results, with a wider range of *k*-mer lengths resulting in a correct call for the majority of samples. The coverage method performed worse for the HPRC++ samples than for the OHS samples, which was not unexpected; the targeted OHS reads were of uniform length and sequenced to very high and consistent coverage, making it easier to estimate λ with the coverage method than for the low-coverage and variable-length HPRC++ reads. Due to the shorter targeted OHS reads, which only extended 716bp upstream and 233bp downstream, the flank median method was unable to estimate λ because most of the *k*-mers from the 10kb flanks had counts of zero, leading to a mean count and thus λ estimate of zero. For the OHS samples, most calls were made with longer *k*-mers; only three and five samples were called using $k < 200$ with the coverage and flank mean methods, respectively. The lowest value of *k* able to call a sample correctly was 61. On the other hand, nine, 25, and 31 HPRC++ samples were able to be called using *21*-mers with the coverage, flank mean, and flank median methods, respectively.

Overall, these results support the utility of counting *k*-mers as a genotyping strategy and suggest that the main issue with the model is due to the short reads lacking information for lengthy repetitive regions of the genome such as *PRDM9*.

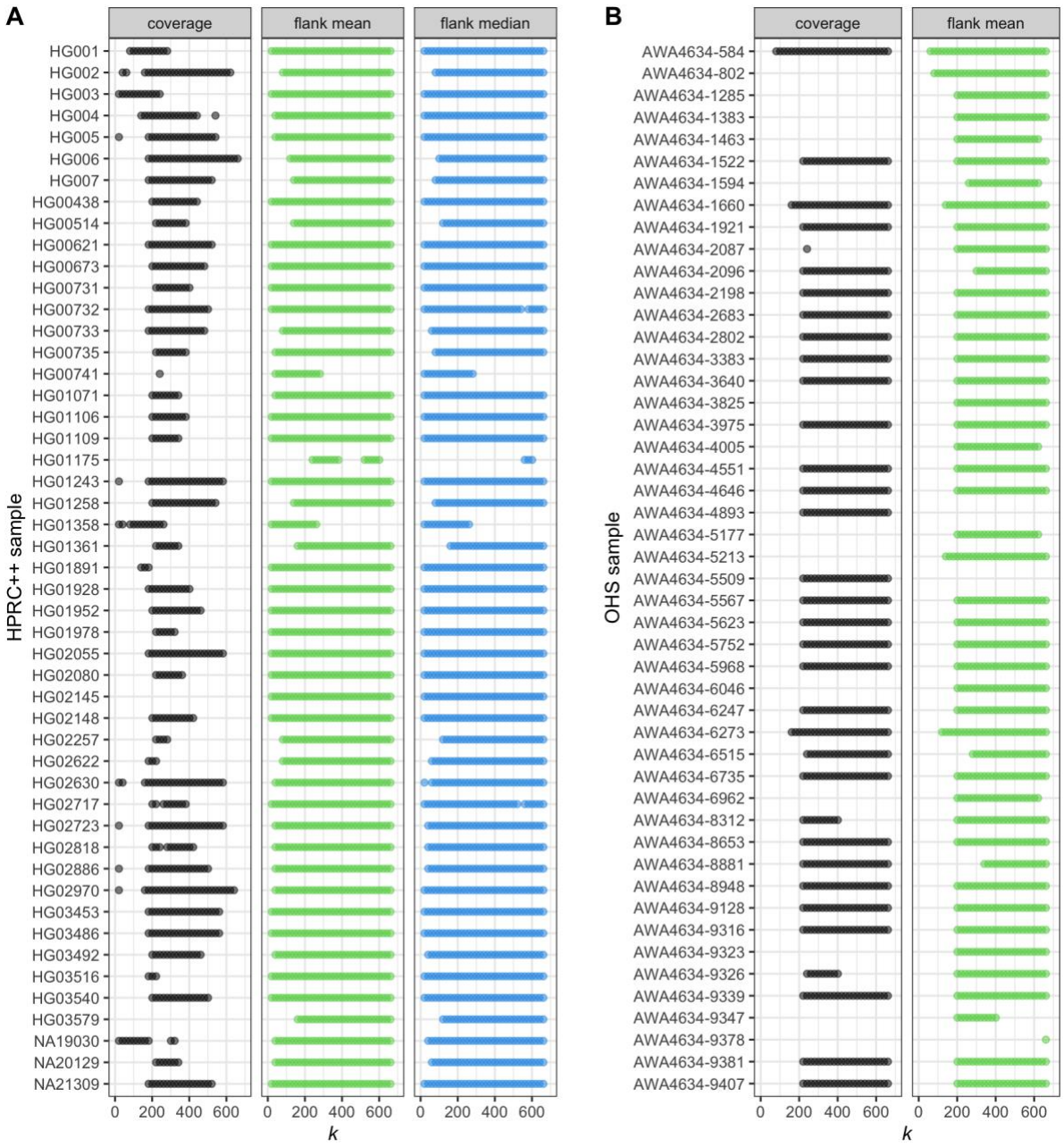


Figure 3.4: Performance of the *k*-mer count models on HPRC++ and OHS long-read data. The HPRC++ and OHS PacBio HiFi reads were used by the count model across a wide range of *k*-mer lengths, testing all three methods of estimating λ: coverage (black), flank mean (green), and flank median (blue). **A)** 49 HPRC++ samples were called by at least one method. The coverage, flank mean, and flank median methods were able to call 46, 49, and 49 samples, respectively. **B)** 48 OHS samples were called by at least one method. The coverage and flank mean methods were able to call 34 and 46 samples, respectively. The flank median method was not able to call any samples.

3.2.7 The consensus model identified novel alleles identified using the long-read data

PRDM9 is one of the fastest evolving genes in humans (Ponting 2011). It is reasonable to expect to see previously undescribed alleles as sequencing technologies improve and as more and more samples from a variety of ancestral backgrounds are assessed. Eight alleles were flagged as novel after genotyping the HPRC++ and OHS samples with the consensus model, and viewing the genotype-specific alignments in `IGV` showed that seven of the novel alleles were unique.

Alleva et al. (2021) described 13 instances where called alleles in their samples turned out to be from the Jeffreys (2013) list of blood/sperm alleles (see **Chapter 2 Results 2.2.1**). The HPRC++ and OHS samples with novel alleles were re-genotyped with the consensus model using the *PRDM9*-642 list of allele sequences, which contains the *PRDM9*-106 alleles plus the additional blood/sperm alleles. One allele initially flagged as novel was determined to be allele P315, a blood/sperm allele, and was present in two samples.

To further characterize the novel alleles, I searched for known zinc finger sequences within the novel allele consensus sequences (**Methods 3.4.4.3**). Three samples had novel arrangements of known zinc finger repeats, while the remaining three samples each had one unique novel zinc finger repeat. One of these novel zinc fingers had a 5bp deletion, which is quite unusual given the conservation of the 84bp repeat size, but was confirmed in `IGV` in both the Illumina and PacBio HiFi reads (**Figure 3.5**). Overall, five novel alleles were identified and named P643-P647, along with three new zinc finger repeats named Z100-Z102 (**Table 3.6**). A potential novel allele was also identified in OHS sample AWA4634-9378; no new name was given, however, considering how the extreme PCR laddering issues observed could have incorrectly or preferentially amplified certain DNA templates.

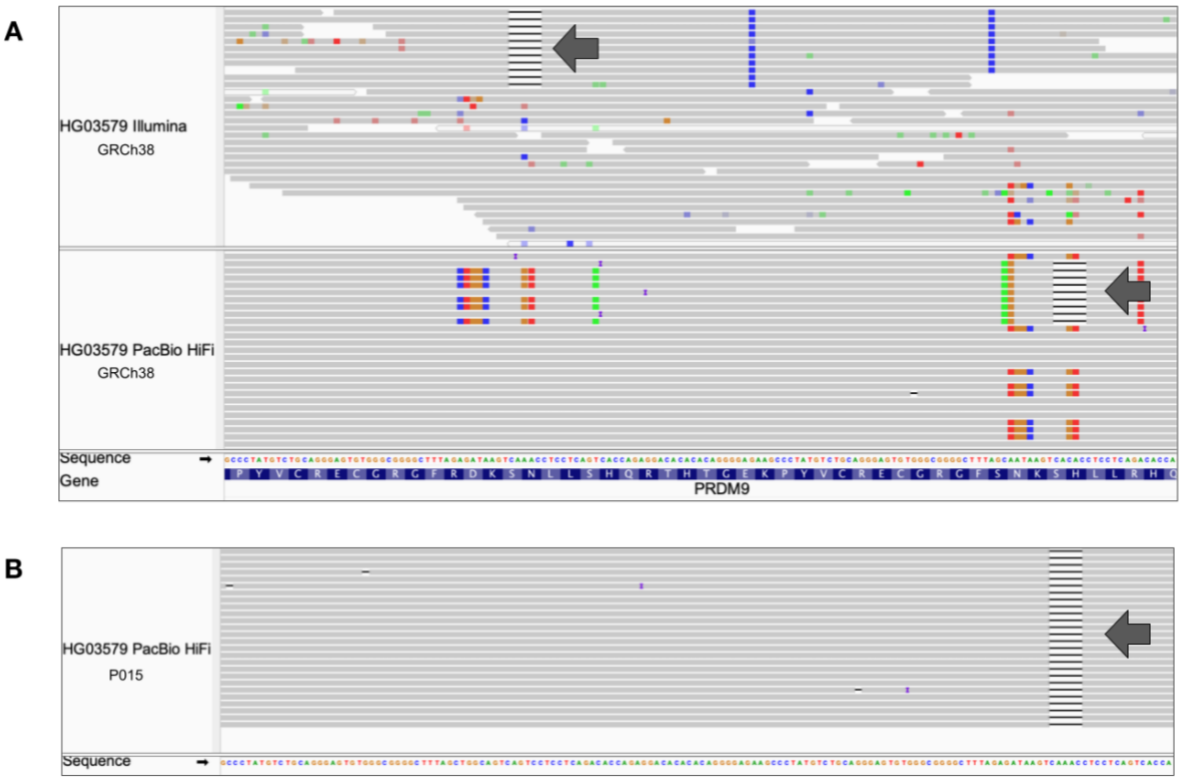


Figure 3.5: Unusual 5bp deletion in a novel *PRDM9* allele identified by the consensus model. Read alignments of HPRC++ sample HG03579 showing a heterozygous 5bp deletion (arrows) within the *PRDM9* zinc finger array. **A)** Illumina (top) and PacBio HiFi (bottom) read alignments to GRCh38. Differences in location of the deletion is the result of multiple possible alignments of short reads to the reference, owing to the high similarity of the start and end sequences of individual zinc finger repeats. **B)** PacBio HiFi alignments to allele P015, part of the genotype-specific reference for sample HG03579.

Table 3.6: Novel alleles in the HPRC++ and OHS samples identified by the consensus model. The consensus sequence of the novel allele was first searched within the list of *PRDM9*-642 allele sequences, which contains blood/sperm alleles, where two samples were identified as having the blood/sperm allele P315 instead of Novel_Similar_P613. The remaining novel allele sequences were scanned for zinc finger repeat motifs to identify novel combinations of zinc fingers and novel zinc finger sequences. Three samples each had a unique novel zinc finger sequence (denoted "ZNOV"; yellow), one of which contained a 5bp deletion. The other three samples each had unique novel arrangements of zinc finger repeats. New standardized names for the novel alleles and zinc fingers are shown in bold in the last two columns. No new name was given for the potential novel allele from AWA4634-9378 due to the laddering issues from preparation for targeted PacBio sequencing.

Sample	Novel allele called by consensus model	Novel allele actually in *PRDM9*-642 list	New allele name for a novel combination of zinc finger repeats	New zinc finger name for a novel zinc finger repeat sequence
HG00741	Novel_Similar_P001	NA	**P643** Z001_Z002_Z003_Z004_ Z004_Z005_Z003_Z006_ Z007_ZNOV_Z006_Z009_ Z010	**Z100** TGTGGGTGGGGCTTTAGAGAT AAGTCAAACCTCCTCAGTCAC CAGAGGACACACACAGGGGAG AAGCCCTATGTCTGCAGGGAG
HG01175	Novel_Similar_P015	NA	**P644** Z001_Z002_Z003_Z004_ Z004_Z018_Z003_Z003_ Z006_Z011_Z008_Z012_ Z015_Z009_Z010	NA
HG01358	Novel_Similar_P002	NA	**P645** Z001_Z002_Z003_Z004_ Z004_Z003_Z003_Z006_ Z007_ZNOV_Z006_Z009_ Z010	**Z101** TGTGGGCGGGGCTTTAGAGAT AAGTCAAACCTCCTCAGTCAC CAAAGGACACACACAGGGGAG AAGCCCTATGTCTGCAGGGAG
HG02572	Novel_Similar_P009	NA	**P646** Z001_Z002_Z003_Z004_ Z005_Z003_Z004_Z004_ Z005_Z003_Z006_Z007_ Z008_Z006_Z009_Z010	NA
HG03579	Novel_Similar_P015	NA	**P647** Z001_Z002_Z003_Z004_ Z004_Z003_Z003_Z003_ Z006_Z011_Z008_Z012_ ZNOV_Z009_Z010	**Z102** TGTGGGCGGGGCTTTAGAGAT AAGTCTCCTCAGTCACCAGAG GACACACACAGGGGACAAGCC CTATGTCTGCAGGGAG
NA18906	Novel_Similar_P613	P315	NA	NA
NA19240	Novel_Similar_P613	P315	NA	NA
AWA4634-9378	Novel_Similar_P016	NA	**Potential novel** Z001_Z002_Z003_Z004_ Z004_Z003_Z003_Z003_ Z007_Z008_Z006_Z010	NA

3.2.8 The effect of the number of input alleles on genotyping calls

Providing a larger list of known alleles for genotyping will complicate analyses and make it more difficult for the models to correctly identify genotypes, particularly if the additional alleles are very similar to the existing ones. For the *PRDM9*-36 alleles, all allele *k*-mer count profiles were unique when using a value of *k* of at least 3, and all genotype *k*-mer count profiles were unique when using a value of *k* of at least 303. For the *PRDM9*-106 alleles, a minimum value of *k* of 62 or 401 was required to obtain unique allele *k*-mer count profiles or genotype *k*-mer count profiles, respectively. In addition to the need for higher values of *k*, the number of comparisons increases nearly exponentially as more alleles are added. Since the initial model development work with simulated reads only used the list of 36 alleles, the effect of the number of provided known alleles was assessed on both the HPRC++ and OHS datasets in comparison to the truth genotypes converted to equivalent *PRDM9*-36 allele names (**Table 3.7**). Samples that had a *PRDM9*-106 genotype allele without an equivalent *PRDM9*-36 allele were excluded, leaving 45 of the 49 HPRC++ samples and 46 of the 49 OHS samples (**Methods 3.4.6**).

Table 3.7: *PRDM9*-36 alleles and the equivalent *PRDM9*-106 names. No *PRDM9*-106 name exists for allele L24 because the L24 sequence in the *PRDM9*-36 list contains a typo (see **Chapter 2 Results 2.2.1**).

***PRDM9*-36 name**	***PRDM9*-106 name**	***PRDM9*-36 name**	***PRDM9*-106 name**
A	P001	L14	P023
B	P002	L15	P024
C	P003	L16	P025
D	P004	L17	P026
E	P005	L18	P027
L1	P010	L19	P028
L2	P011	L20	P029
L3	P012	L21	P030
L4	P013	L22	P031
L5	P014	L23	P032
L6	P015	L24	-
L7	P016	L32	P041
L8	P017	L33	P042
L9	P018	L34	P043
L10	P019	L35	P044
L11	P020	L36	P045
L12	P021	L37	P035
L13	P022	L38	P046

3.2.8.1 The allele list size had no effect on the realignment or consensus model calls

Converting the *PRMD9*-106 truth genotypes to the *PRDM9*-36 equivalents resulted in 45 HPRC++ samples having complete genotypes (i.e. both alleles in the genotype were in the *PRDM9*-36 list) and seven samples with one allele not in the *PRDM9*-36 list. Of the seven HPRC++ samples with novel alleles in the truth genotypes, three had complete *PRDM9*-36 genotypes. When genotyped with the shorter allele list, both the consensus and realignment models correctly called all 45 samples with complete genotypes. For the seven samples with incomplete *PRDM9*-36 genotypes, both models included the present allele in the genotype and

the consensus model flagged a second allele as novel, which is the expected outcome for an allele not present in the input fasta file. The consensus model also correctly flagged the novel alleles in the three samples with complete genotypes that contained a novel allele. Thus, there was no difference in performance for either the realignment or consensus model when using the shorter list of alleles to genotype the HPRC++ samples.

For the OHS samples, converting the *PRMD9*-106 truth genotypes to the *PRDM9*-36 equivalents resulted in 46 samples having complete genotypes, including the sample with a novel allele in the truth genotype, and three samples with one allele not in the *PRDM9*-36 list. The consensus model called 44 samples with complete *PRDM9*-36 genotypes correctly; the two samples called incorrectly were the same incorrectly called by the consensus model with the *PRDM9*-106 list of alleles (AWA4634-1921 and AWA4634-2802), which again had one allele correct. For the three samples with incomplete genotypes, the consensus model included the present allele in the genotype and flagged the other allele as novel, as expected. Of the 46 samples with complete *PRDM9*-36 genotypes, the realignment model called 22 samples correctly and 24 samples with one allele correct. The 24 samples with only one correct allele were also only partially called using the *PRDM9*-106 list of alleles. The realignment model also called one allele correctly for the three samples with incomplete *PRDM9*-36 genotypes. Again, there was no difference in performance for either the realignment or consensus model when using the *PRDM9*-36 list of alleles to genotype the OHS samples (**Table 3.8**).

Table 3.8: Effect of known allele list size on the long-read genotyping model results. Changing the number of known *PRDM9* alleles for long-read genotyping had no effect on the number of HPRC++ or OHS samples called correctly. The samples with one correct allele using the shorter list of alleles had truth genotypes with one allele not present in the *PRDM9*-36 list.

Long-read genotyping model	*PRDM9* list	Number of correct alleles		Number of samples	
				HPRC++ (out of 52)	OHS (out of 49)
Realignment	**106**		**2**	52	24
			1	0	25
	36	**Complete genotype**	**2**	45	22
			1	0	24
		Incomplete genotype	**1**	7	3
Consensus	**106**		**2**	52	47
			1	0	2
	36	**Complete genotype**	**2**	45	44
			1	0	2
		Incomplete genotype	**1**	7	3

3.2.8.2 The short-read models generally improved with the shorter allele list size

In contrast to the long-read models, using the shorter list of *PRDM9* alleles generally improved the results for the short-read model calls, particularly for the HPRC++ dataset. A larger proportion of HPRC++ samples were called correctly or had the correct genotype in the top 10 calls across all model methods and for both raw and corrected reads, more than double for some short-read methods (**Figure 3.6 A**). For the OHS dataset, using the shorter list resulted in a slightly larger proportion of samples called correctly or in the top 10 calls for the count and combined models on corrected reads only (**Figure 3.6 B**). In contrast, most of the model methods had slightly fewer proportions of correct calls or calls in the top 10 for the raw reads. Since the long-read models were more effective than the short-read models in genotyping samples with the longer list of *PRDM9* alleles, the lack of improvement using the shorter list

of alleles with the long-read models is not surprising. In contrast, having fewer possible genotypes to begin with made it easier for the short-read models to genotype the samples, particularly since the majority of samples had genotypes with alleles found in the shorter list. If more samples had alleles specific to the *PRDM9*-106 list, there would likely be less of an improvement in the short-read model calls using the shorter list of alleles.

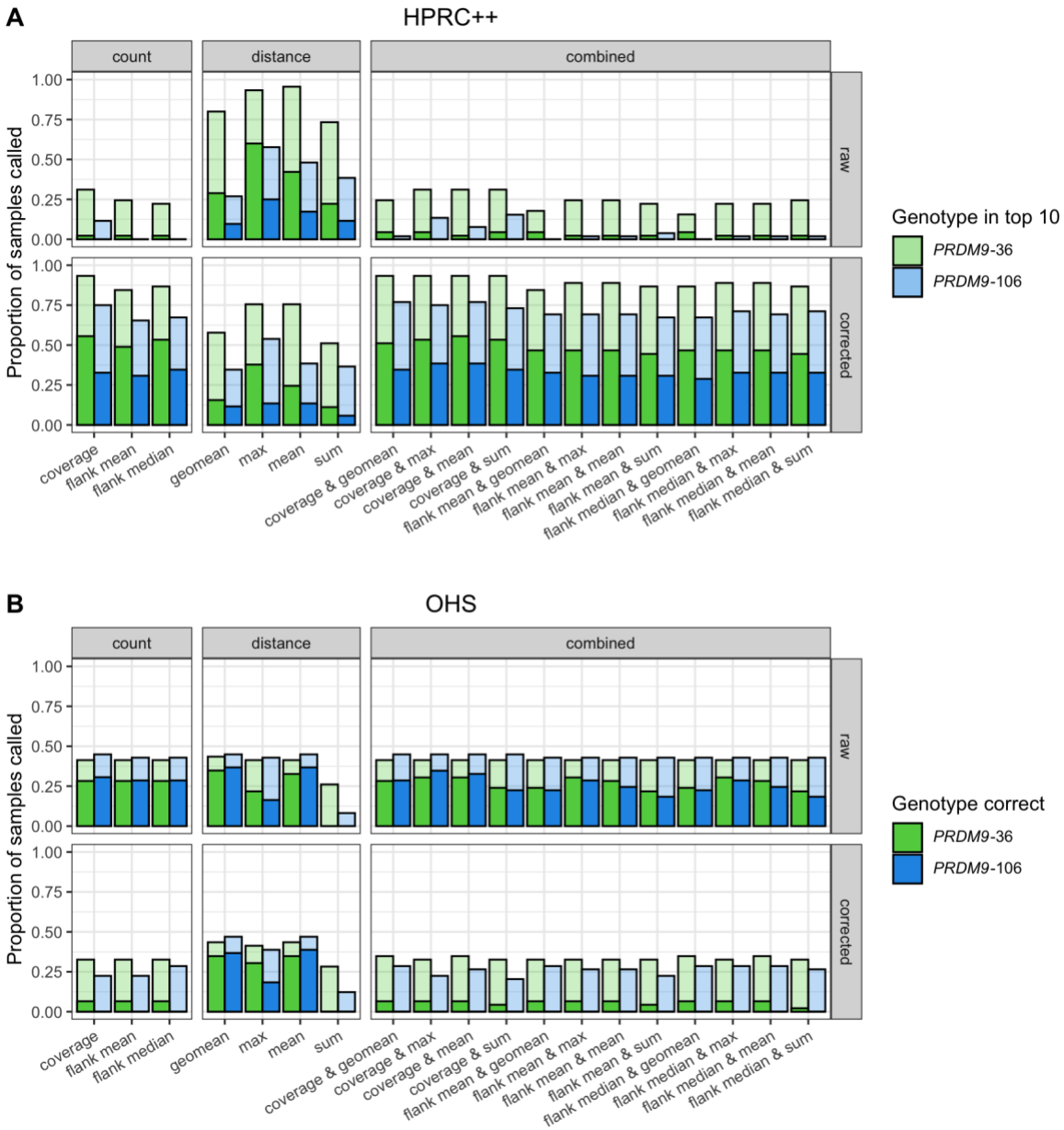


Figure 3.6: Effect of the known allele list size on the short-read genotyping model calls. Correct genotype in the top 10 (pale colors) or called correctly (solid colors) using either the *PRDM9*-36 (green) or *PRDM9*-106 (blue) list of alleles. All counts, distance, and combined methods were tested (columns) for both raw and corrected reads (rows). **A)** HPRC++ samples. The smaller list of alleles resulted in a higher proportion of samples called correctly or with the correct genotype in the top 10 calls for most methods. **B)** OHS samples. Using the smaller list of alleles resulted in a slight increase in the proportions of samples called correctly or with the correct genotype in the top 10 calls.

3.3 Discussion

While long PacBio HiFi reads can alleviate mappability issues in repetitive regions of the genome, they are slightly more prone to errors than short reads, specifically to indels. This may be problematic depending on the repeat lengths in a given region. The realignment and consensus models were developed to simplify the *PRDM9* genotyping process for long reads, which required realignment to a list of known alleles, additional variant-calling steps, and visual confirmation in `IGV` for variants. Even though viewing alignments to *PRDM9*-specific or genotype-specific references simplified visual confirmations compared to viewing alignments to GRCh38, it still remains tedious and there is often uncertainty if a particular variant is at a high enough frequency to be considered a polymorphism or a sequencing error. Though there is the option to use already available SNV and SV calling tools after realigning samples, the full *PRDM9* genotyping process requires multiple steps. The realignment and consensus models, however, assess all possible genotypes given a list of known alleles in a single step. The consensus model is additionally able to flag novel alleles.

Though simpler and unable to provide likelihood values for called genotypes, the consensus model performed better than the realignment model. Considering both the HPRC++ and OHS datasets together, the consensus model correctly genotyped 99/101 or 98% of the samples while the realignment model successfully genotyped 76/101 or 75% of the samples. There was a noticeable difference in performance of the realignment model between the two datasets, with all HPRC++ samples called correctly while the OHS dataset only had 24/49 (49%) samples called correctly. The lower performance on the OHS dataset is likely PCR- laddering related. One possibility is that the filtering approach for cleaning up the laddered reads potentially removed reads representing true alleles in the sample genotypes. It is also possible that during PCR, shorter fragments or fragments with laddering were preferentially amplified, resulting in an incorrect overabundance of repeats in the samples that could not be adjusted in silico.

The long reads were used to determine truth genotypes for the samples that could then be used to validate performance for the short-read models. 23/52 (44%) of the HPRC++ samples and 20/49 (41%) of the OHS samples were called correctly by at least one short-read model, and

47/52 (90%) of the HPRC++ samples and 23/49 (47%) of the OHS samples had the correct genotype ranked 10th or better by at least one short-read mode, with the assumption that the OHS truth genotypes determined from the PacBio reads were correct. While the percentage of correctly called samples is not as high for the short-read models as it was for the long-read models, the results are still supportive for using models that utilize *k*-mer information for genotyping purposes. The results for running the *k*-mer count model on the PacBio HiFi reads were also very supportive: 49/52 (94%) HPRC++ and 48/49 (98%) OHS samples were called at least once after being tested with the different λ estimation methods and range of *k*-mer lengths. In general, longer *k*-mers resulted in more samples being called, again emphasizing the limitations of short-read sequencing for genotyping repetitive regions of the genome.

The two long-read models and the three short-read models are proofs of concept for alternative ways of genotyping highly polymorphic and repetitive regions of the genome. The short-read models essentially ignore alignment information, so long as the input set of reads were initially mapped to the region of interest. The long-read models do rely on alignment, but either relative to all known variants or relative to all long sequencing reads from the sample. Nevertheless, there are ways to improve the models. As discussed in **Chapter 2 Discussion 2.3**, the count-distance model would greatly benefit from accounting for sequencing errors, and could also improve calls by considering more than one *k*-mer pair per read fragment. Overall, the consensus model shows great ability to genotype long-read data and can be utilized even when samples have experienced significant PCR laddering artifacts with appropriate filtering of noisy reads. The model demonstrates a one-step approach to genotyping samples with many possible variants and providing a long-range, full haplotype assessment of the reads while also checking for novel alleles. The short-read count and distance models show promise in genotyping difficult regions of the genome with *k*-mer information, but remain a proof of concept until performance can be improved.

3.4 Methods

3.4.1 Preparing publicly available data

Data were collected from three publicly available sources that had performed both PacBio HiFi and Illumina whole-genome sequencing on samples: the HPRC, the 1000GP-SV consortium, and the GIAB consortium (**Appendix Table 4**). The HPRC (Wang et al. 2022) year 1 freeze data (Human Pangenome Reference Consortium 2021) included 44 samples representing five continental populations, the majority of which are understudied. The 1000GP-SV consortium (Fairley et al. 2020) provided four more samples, and the GIAB consortium (Zook et al. 2016) provided an additional five samples. Collectively, I refer to these samples as the **HPRC++ dataset** (**Table 3.9**).

Most samples had aligned bams available and details about sample sequencing and subsequent read processing can be found in each dataset publication. Briefly, for PacBio sequencing, both the HPRC and 1000GP-SV samples were aligned with `winnowmap` using the `-ax map-pb` option (Jain et al. 2020), while the GIAB samples were aligned with `pbmm2` using the `--preset CCS` option (https://github.com/PacificBiosciences/pbmm2). For Illumina sequencing, the GIAB samples were aligned with `NovoAlign` (https://www.novocraft.com/products/novoalign/), while the HPRC and 1000GP-SV samples were aligned with `BWA-MEM` (Li 2013). All samples were aligned to the GRCh38 reference genome.

Five HPRC++ samples (HG00514, HG00731, HG00732, HG02970 and NA19030) did not have aligned bams for PacBio data. Instead, the unmapped bams (three for HG00514 and NA19030, four for HG02970, five for HG00731, and nine for HG00732) were downloaded, fastq files were generated with `samtools fastq` (Li et al. 2009) and merged per sample, and the reads were mapped to GRCh38 with `minimap2` using the `-ax map-pb` option (Li 2018). An additional three samples (HG002, HG005, and NA21309) did not have aligned bams for Illumina data. The fastq files available for each of these samples (two for HG002 and NA21309, and 48 for HG005) were combined into a single fastq file per sample and then aligned to GRCh38 with `BWA-MEM`.

Aligned bam files for both Illumina and PacBio data were then subset to contain only reads that aligned to the ***PRDM9* zinc finger array + 10kb flanks region** (GRCh38 chr5:23516673–23537764; the zinc finger array alone is located at chr5:23526673–23527764) using `samtools view`. One sample did not have sufficient coverage of PacBio reads (NA19239), leaving a total of 52 samples to analyze. The rs6889665 genotype was determined by viewing alignments in `IGV` with an allele fraction threshold of 0.3.

Table 3.9: The HPRC++ sample demographics. Each sample is listed with its originating cohort, its 1000GP population and continental population, and its rs6889665 genotype as determined by viewing the PacBio HiFi reads in `IGV`. Sample NA19239 was not analyzed due to a lack of sufficient PacBio HiFi reads. Cohorts: GIAB, Genome in a Bottle; HPRC, Human Pangenome Consortium; HPRC_PLUS, additional samples in the HPRC cohort not sequenced by the HPRC team; and 1000GP-SV, 1000 Genomes Project Structural Variant consortium. Continental populations: AFR, African; AMR, Admixed American*; EAS, East Asian; EUR, European; and SAS, South Asian. Populations: CEU, CEPH (Utah Residents with Northern and Western European ancestry); CHS, Southern Han Chinese; PUR, Puerto Rican; CLM, Colombian; PEL, Peruvian; ACB, African-Caribbean; KHV, Kinh Vietnamese; GWD, Gambian; ESN, Esan; MSL, Mende; PJL, Punjabi: YRI, Yoruba; LWK, Luhya; and ASW, African-American Southwest (United States). *Note: The 1000GP refers to this continental population as "American", but in this thesis it is referred to as "Admixed American" for clarification as the populations are of admixed European, African, and Indigenous American ancestries. Additionally, three Ashkenazi samples are grouped as European.

Sample ID	Cohort	Continental Population	Population	rs6889665 Genotype
HG001	GIAB	EUR	CEU	0/0
HG002	HPRC_PLUS	EUR	Ashkenazi	0/0
HG003	GIAB	EUR	Ashkenazi	0/0
HG004	GIAB	EUR	Ashkenazi	0/0
HG005	HPRC_PLUS	EAS	Han Chinese	0/0
HG006	GIAB	EAS	Han Chinese	0/0
HG007	GIAB	EAS	Han Chinese	0/0
HG00438	HPRC	EAS	CHS	0/0
HG00514	1000GP-SV	EAS	CHS	0/0
HG00621	HPRC	EAS	CHS	0/0
HG00673	HPRC	EAS	CHS	0/0
HG00731	1000GP-SV	AMR	PUR	0/1
HG00732	1000GP-SV	AMR	PUR	0/0
HG00733	HPRC_PLUS	AMR	PUR	0/0
HG00735	HPRC	AMR	PUR	0/1
HG00741	HPRC	AMR	PUR	0/0
HG01071	HPRC	AMR	PUR	0/0
HG01106	HPRC	AMR	PUR	0/0
HG01109	HPRC_PLUS	AMR	PUR	0/0
HG01175	HPRC	AMR	PUR	0/1
HG01243	HPRC_PLUS	AMR	PUR	0/0
HG01258	HPRC	AMR	CLM	0/0
HG01358	HPRC	AMR	CLM	0/0
HG01361	HPRC	AMR	CLM	0/1
HG01891	HPRC	AFR	ACB	1/1
HG01928	HPRC	AMR	PEL	0/0

HG01952	HPRC	AMR	PEL	0/0
HG01978	HPRC	AMR	PEL	0/0
HG02055	HPRC_PLUS	AFR	ACB	0/0
HG02080	HPRC_PLUS	EAS	KHV	0/0
HG02145	HPRC_PLUS	AFR	ACB	1/1
HG02148	HPRC	AMR	PEL	0/1
HG02257	HPRC	AFR	ACB	0/1
HG02572	HPRC	AFR	GWD	0/0
HG02622	HPRC	AFR	GWD	0/1
HG02630	HPRC	AFR	GWD	0/0
HG02717	HPRC	AFR	GWD	0/1
HG02723	HPRC_PLUS	AFR	GWD	0/0
HG02818	HPRC_PLUS	AFR	GWD	0/1
HG02886	HPRC	AFR	GWD	0/0
HG02970	HPRC_PLUS	AFR	ESN	0/0
HG03453	HPRC	AFR	MSL	0/0
HG03486	HPRC_PLUS	AFR	MSL	0/0
HG03492	HPRC_PLUS	SAS	PJL	0/0
HG03516	HPRC	AFR	ESN	0/1
HG03540	HPRC	AFR	GWD	0/0
HG03579	HPRC	AFR	MSL	0/1
NA18906	HPRC_PLUS	AFR	YRI	0/0
NA19030	HPRC_PLUS	AFR	LWK	0/1
NA19239	1000GP-SV	AFR	YRI	0/0
NA19240	HPRC_PLUS	AFR	YRI	0/1
NA20129	HPRC_PLUS	AFR	ASW	0/1
NA21309	HPRC_PLUS	AFR	Masai	0/0

3.4.2 Preparing Ontario Health Study PacBio HiFi data

3.4.2.1 Sample selection

A total of 50 samples were accessed from the **OHS** cohort (Kirsh et al. 2022) under approval of the University of Toronto Ethics Committee (Human Participant Ethics Protocol Submission #20345) and the Canadian Partnership for Tomorrow's Health (application #DAO-041621/2020-11-BSL-01-HMWSP). Samples were selected based on two criteria: the genotype of SNP rs6889665 and on estimated ethnicity. SNP rs6889665, which is associated with longer *PRDM9* alleles (Hinch et al. 2011), was imputed for all OHS samples. Continental ethnicity was estimated by comparing all of the OHS samples to select 1000GP samples via

clustering of a principal component analysis of microarray genotyping data (performed by Vanessa Bruat at OICR; **Figure 3.7**). Once samples were annotated with their continental population cluster and imputed for rs6889665 (performed by Jasmina Uzunović at OICR), 90 samples were chosen representing a mix of ethnicities and primarily had homozygous alternate or heterozygous rs6889665 genotypes. I then chose 50 samples based on DNA quality as assessed by the OHS team. The final samples consisted of 25 clustering with African ancestry, 10 with East Asian ancestry, 11 with South Asian ancestry, and three clustering with "Other" ancestry, which was mainly European or admixed (**Table 3.10**).

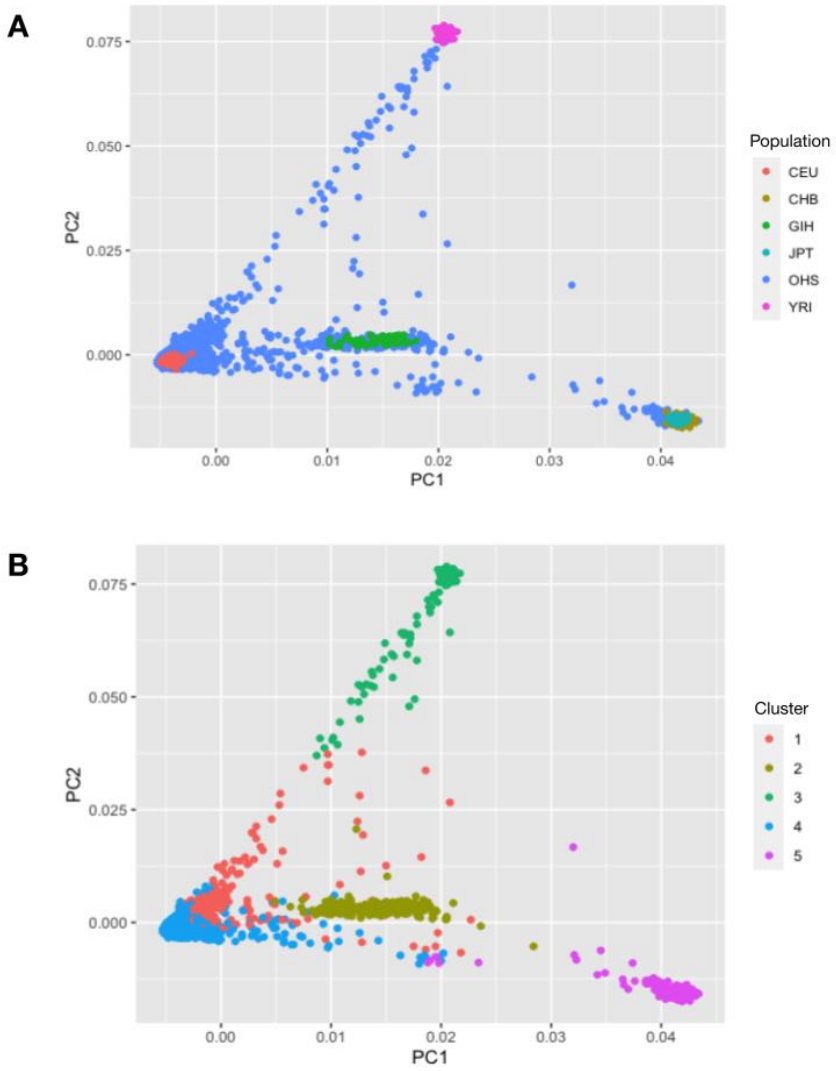


Figure 3.7: Clustering of OHS samples by inferred ancestry. A) Principal component analysis of genome-wide microarray genotyping data was performed on OHS samples and selected 1000GP populations (colors): CEU, Utah Residents (CEPH) with Northern and Western European ancestry; CHB, Han Chinese in Beijing, China; GIH, Gujarati Indians in Houston, Texas; JPT, Japanese in Tokyo, Japan; and YRI, Yoruba in Ibadan, Nigeria. **B**) All samples from A) were clustered into five groups (colors) using *k*-means clustering. OHS samples were imputed for SNP rs6889665 and sample selection aimed to balance ancestry groups and the presence of rs6889665 alternate alleles. The final samples were: 25 samples from cluster 3 (African ancestry), 10 samples from cluster 5 (East Asian ancestry), 11 samples from cluster 2 (South Asian ancestry), and three samples from clusters 1 and 4 ("Other" ancestry, mainly European or admixed). African samples had any rs6889665 genotype, while the East and South Asian samples had at least one rs6889665 alternate allele, and the remaining samples had two rs6889665 alternate alleles. Plots in A) and B) generated by Vanessa Bruat (OICR).

Table 3.10: The OHS sample demographics. Each sample is listed with its estimated 1000GP continental population and its rs6889665 genotype, which was imputed from genome-wide genotypes. Sample AWA4634-806 was not analyzed further due to a lack of sufficient PacBio HiFi reads. Estimated continental populations: AFR, African; EAS, East Asian; and SAS, South Asian. Samples that did not cluster with these three populations were given the broad label of "Other".

Sample ID	Estimated continental population	rs6889665 genotype	Sample ID	Estimated continental population	rs6889665 genotype
AWA4634-584	AFR	0/1	**AWA4634-5213**	AFR	0/1
AWA4634-802	AFR	0/1	**AWA4634-5509**	AFR	0/0
AWA4634-806	EAS	1/1	**AWA4634-5567**	AFR	0/0
AWA4634-1285	Other	1/1	**AWA4634-5623**	SAS	0/1
AWA4634-1383	EAS	0/1	**AWA4634-5752**	AFR	0/1
AWA4634-1463	AFR	1/1	**AWA4634-5968**	AFR	0/1
AWA4634-1522	SAS	0/1	**AWA4634-6046**	EAS	0/1
AWA4634-1594	AFR	1/1	**AWA4634-6247**	AFR	0/0
AWA4634-1660	EAS	0/1	**AWA4634-6273**	AFR	1/1
AWA4634-1921	AFR	0/0	**AWA4634-6515**	EAS	0/1
AWA4634-2087	SAS	0/1	**AWA4634-6735**	AFR	0/1
AWA4634-2096	SAS	0/1	**AWA4634-6962**	Other	1/1
AWA4634-2198	AFR	0/1	**AWA4634-8312**	SAS	0/1
AWA4634-2683	EAS	0/1	**AWA4634-8653**	AFR	0/1
AWA4634-2802	AFR	0/0	**AWA4634-8881**	EAS	0/1
AWA4634-3383	AFR	0/0	**AWA4634-8948**	SAS	0/1
AWA4634-3640	AFR	0/1	**AWA4634-9128**	EAS	0/1
AWA4634-3825	EAS	1/1	**AWA4634-9316**	AFR	0/0
AWA4634-3975	AFR	0/0	**AWA4634-9323**	SAS	0/1
AWA4634-4005	Other	1/1	**AWA4634-9326**	SAS	0/1
AWA4634-4206	EAS	0/1	**AWA4634-9339**	SAS	0/1
AWA4634-4551	AFR	0/1	**AWA4634-9347**	AFR	0/1
AWA4634-4646	AFR	0/1	**AWA4634-9378**	EAS	0/1
AWA4634-4893	AFR	0/1	**AWA4634-9381**	SAS	0/1
AWA4634-5177	AFR	1/1	**AWA4634-9407**	SAS	0/1

3.4.2.2 DNA sequencing

Genomic DNA was extracted with either Qiagen FlexiGene or Qiasymphony kits and diluted to working concentrations with low-Tris-EDTA buffer (performed by the OHS team). For short-read sequencing (performed by the Genomics department at OICR), 100ng of DNA per sample was sheared using a Covaris E220 Focused-Ultrasonicator. Whole-genome libraries were generated using a KAPA HyperPrep Kit. Libraries were sequenced to a target of 40X

coverage on an Illumina NovaSeq 6000 system, generating paired end 151bp reads. Sequencing data preprocessing (performed by Mawussé Agbessi at OICR) used `Picard` tools (Broad Institute 2019): read group IDs were added with `FastqToSam`, adapters were marked with `MarkIlluminaAdapters`, and interleaved fastq files were generated with `SamToFastq`. Reads were aligned to GRCh38 with `BWA-MEM` and merged with `Picard MergeBamAlignment` to generate one final bam file per sample.

For targeted PacBio HiFi long-read sequencing (performed by Genome Quebec), a 2,041bp product containing the zinc finger repeat domain of *PRDM9* plus 716bp upstream and 233bp downstream flanks (GRCh38 chr5:23525957–23527997) was amplified using the primers HsPrdm9-F3 (TGTAAGGAATGACACTGCCCTGA) and HsPrdm9-R1 (ATGTCCCCCGAACACTTACAGAA), originally described by Baudat et al. (2010). The F3 and R1 primers were tagged with sequences ACACTGACGACATGGTTCTACA and TACGGTAGCAGAGACTTGGTCT, respectively. Using 1µl of genomic DNA (diluted 1:30) and Invitrogen Platinum SuperFi II DNA Polymerase, PCR amplification ran for 30 cycles with 98°C melting, 60°C annealing, and 72°C amplifying stages. Sample barcodes and PacBio adapters were then added with a second round of PCR (15 cycles using the same temperatures as the first round) using the Roche FastStart High Fidelity PCR kit. Each amplicon was then quantified with the Life Technologies Quant-iT PicoGreen dsDNA Assay kit. The libraries were prepared using 932ng of purified amplicons, following the PacBio Barcoded Universal Primers for Multiplexing Amplicons Template Preparation and Sequencing protocol. The DNA Damage Repair, End Repair and Adapter Ligation steps were performed as described using the SMRTbell Template Prep Kit 2.0. Primer annealing was performed with sequencing primer v4 at a final concentration of 1nM and the Sequel II 2.1 polymerase was bound at 0.5nM. The libraries underwent an AMPure bead cleanup following the SMRT Link calculator procedure. Sequencing occurred on a PacBio Sequel II instrument at a loading concentration of 120pM following the diffusion loading protocol, using the Sequel II Sequencing kit 2.0, SMRT Cells 8M, and 10 hour movies with a pre-extension time of 30 minutes.

The resulting subreads were compiled into HiFi consensus reads with PacBio SMRT Link (performed by Genome Quebec). I then aligned the reads for each sample to GRCh38 with `minimap2` using the `-ax map-hifi` option. Aligned bam files for both Illumina and

PacBio data were then subset to contain only reads that aligned within the *PRDM9* zinc finger array + 10kb flanks region using `samtools view -hb chr5:23516673-23537764`.

3.4.2.3 Filtering PacBio HiFi reads affected by PCR laddering

Due to the PCR amplification required for the targeted PacBio sequencing, laddering of the PCR products occurred and was evident in the HiFi reads as there were many deletions and insertions scattered throughout the reads (**Figure 3.8**). A filtering pipeline was performed to remove noisy reads and retain those most likely to be true reflections of the zinc finger region (**Figure 3.9**). First, reads were removed if they did not span the full repetitive zinc finger region. Then a list of all indels > 10bp was compiled for each sample. If a specific indel (based on start coordinates and length) was found in < 25% of the reads, the reads with those indels were removed. Reads were also removed if they contained any indels > 10bp that were not a multiple of 84bp (the length of a zinc finger repeat). Read length counts were then determined for each sample. From the `ggpmisc` R package (https://github.com/aphalo/ggpmisc), the `find_peaks` function with the parameter `span=21` was used to determine if the reads were overwhelmingly of one or two lengths by identifying the local maxima in the distribution of read lengths. Maxima that were not equal to the expected read length ± a multiple of 84 were ignored. The expected read length was 2042bp, which includes a single base of adapter that matches the reference, as the read lengths were calculated from CIGAR strings output by `minimap2`. Reads were kept only if they were the same length as a maximum identified in the sample. Finally, the filtering pipeline output a file of read names to keep and bams were filtered by passing the file to the `samtools view -N` argument. After this stringent filtering, one sample (AWA4634-806) did not have enough remaining reads, leaving 49 samples to analyze.

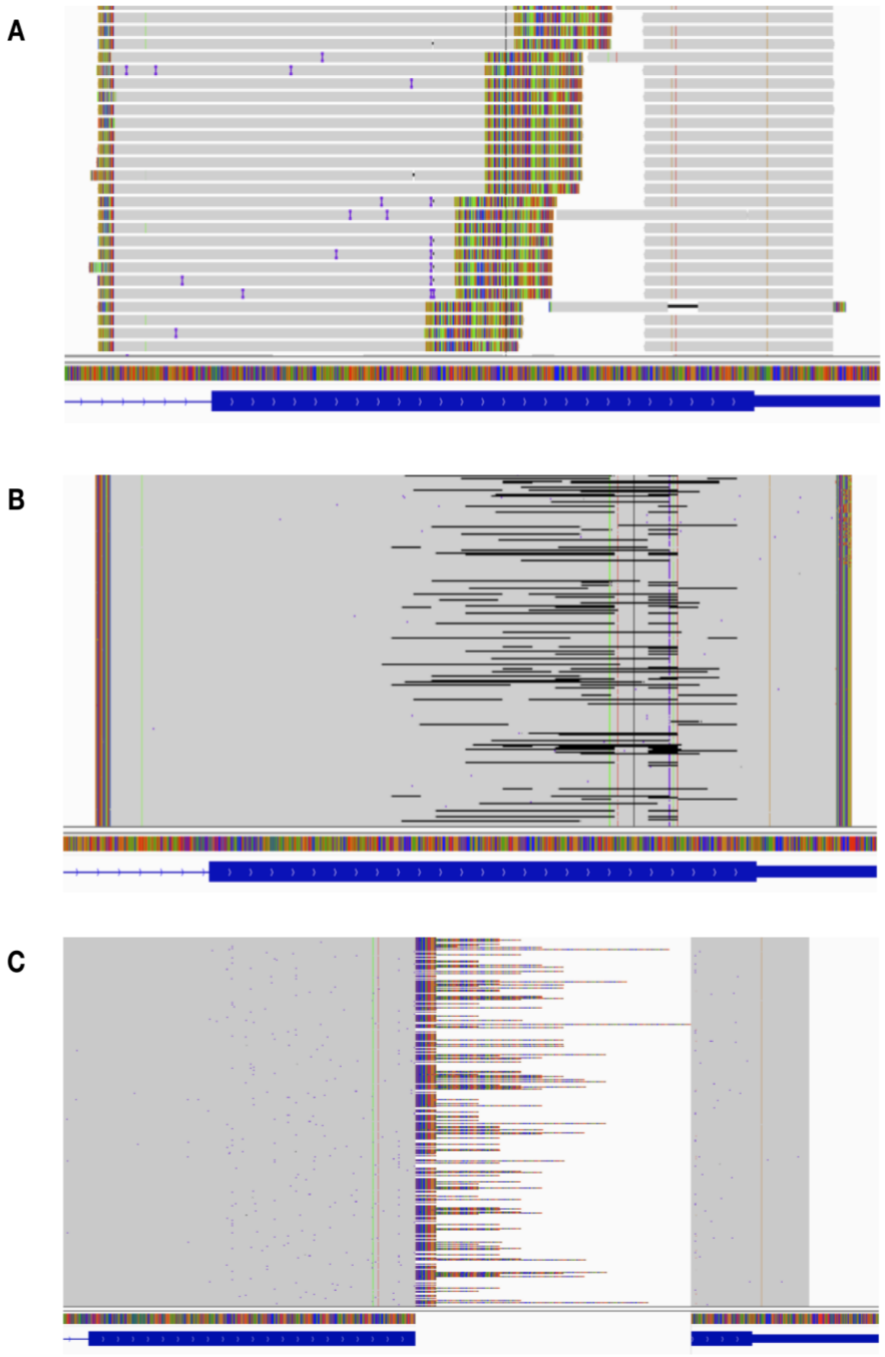


Figure 3.8: PCR laddering in the OHS targeted PacBio HiFi reads. Laddering caused by PCR amplification during library preparation resulted in gains and losses of zinc finger repeats within the zinc finger region in *PRDM9* exon 11 (thick blue horizontal bar). **A)** Truncated reads that do not fully span the targeted region (sample AWA4634-1383). **B)** Reads with variable lengths of deletions (sample AWA4634-4005). **C)** Reads with variable lengths of insertions (sample AWA4634-6926). Most of the indels were roughly multiples of 84, the length of each zinc finger repeat found in *PRDM9*.

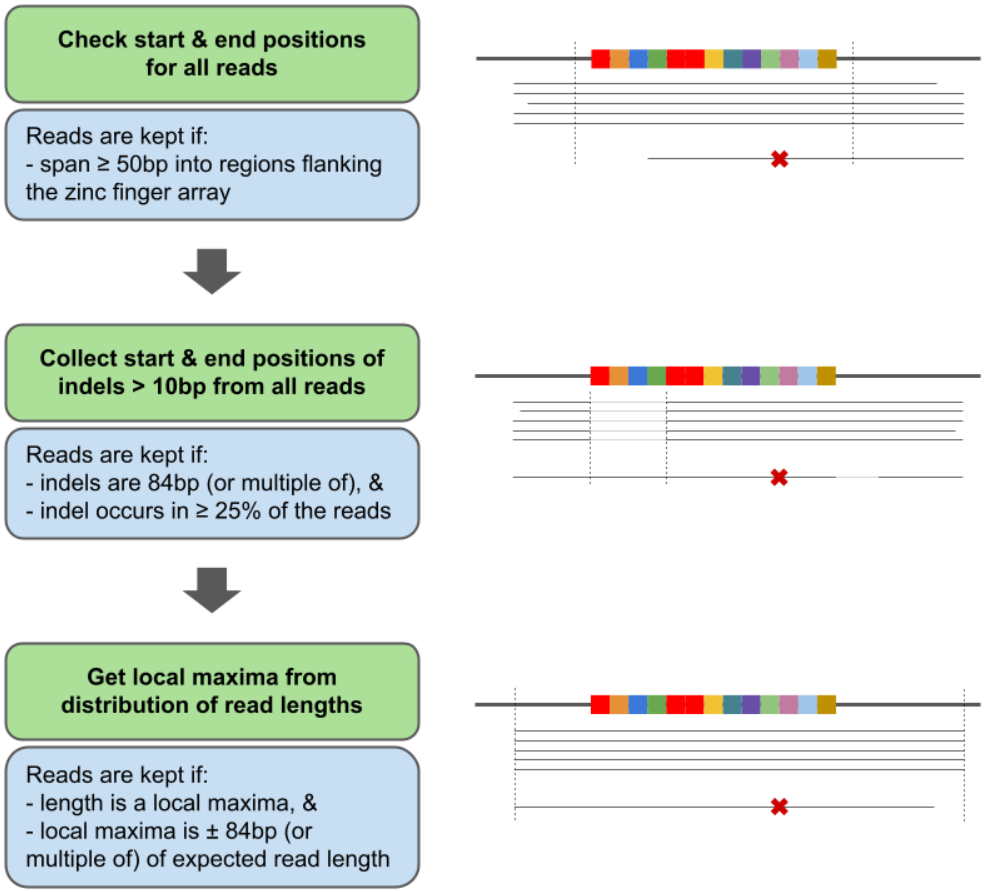


Figure 3.9: Overview of the long-read filtering pipeline to remove noisy reads from PCR laddering. OHS samples underwent a filtering pipeline to remove reads most likely to be products of PCR laddering. First, the read must span the full length of the zinc finger array plus 50bp flanking regions. Then, if the read has an indel, it must be 84bp long or a multiple thereof, and the indel must occur in at least 25% of the usable reads. Finally, the read must be the same length as a local maximum read length determined for that sample, as long as the maximum is the expected read length ± 84bp or a multiple thereof.

3.4.3 Obtaining truth genotypes with *PRDM9*-specific references

Truth genotypes were assigned for each sample in both datasets using the PacBio HiFi reads. The reads from the bams were extracted using `samtools fastq` and then realigned with `minimap2 -ax map-hifi` to a ***PRDM9*-specific reference**, which was a fasta file with the *PRDM9*-106 allele sequences (zinc finger region + 10kb flanks). The two alleles with the most reads aligned were then used to generate **genotype-specific references** for each sample, to which reads were aligned again. The proportion of reads aligned to each allele in the

genotype-specific realignments were then calculated. Alleles were considered to be part of the sample genotype if at least 25% of the reads aligned to them.

These alignments were then viewed in `IGV` to visually check for any variants. If the reads had no SNVs or indels consistently present, the allele was confirmed as part of the truth genotype. If there were SNVs in at least 25% of the reads, the allele was flagged as "novel similar" (e.g. if the reads best aligned to P001 but there were SNVs present, the allele was flagged as Novel_Similar_P001 in the truth genotype). Two samples had an 84bp deletion in one of their resulting genotype alleles. To check if there were alleles with a smaller edit distance to the reads that would be a better fit for use as the base for the novel allele, the deleted sequence was removed and at sites of SNVs observed in IGV, the nucleotide in the allele was changed to the variant base. The resulting modified sequence was then compared to each known allele and the allele with the smallest edit distance was used as the base allele and tagged as novel. These final genotypes were considered the truth genotypes.

3.4.4 Long-read genotyping methods

3.4.4.1 The realignment model

A long-read genotyping model was developed that uses realignment and edit distances of reads to all known allele sequences. The model determines genotypes by considering all pairings of haplotypes (*PRDM9* alleles), similar to the `GATK HaplotypeCaller` (Poplin et al. 2018b):

$$L(r_i|g_j) = \prod_i \left(\frac{L(r_i|h_1)}{2} + \frac{L(r_i|h_2)}{2}\right)$$

where r_i = read i, g_j = genotype j, h_1 = the first haplotype (allele) in g_j, and h_2 = the second haplotype in g_j.

A bam file for a sample and a fasta file of known allele sequences are provided to the realignment model. Reads are checked if they span the full variable region of interest (*PRDM9*

zinc finger domain) plus at least 50bp flanks. Reads that do not fully span the region are ignored. Each remaining read is then aligned to each allele sequence and the **edit distances** for each alignment are calculated with the `edlib` (Šošić and Šikić 2017) python wrapper `align` function using the parameters `mode="HW"` (infix or "**glocal**" alignment, where the entire read is aligned to a substring of the target sequence) and `task="path"`. Given the alignment between a read and an allele, the probability that the sequences are identical is the product of all the mismatches being sequencing errors. Therefore, likelihoods are calculated per read for each allele using either the **mismatch** or the **match-mismatch** model:

$$L(r_i|h_j)mismatch = \varepsilon^{d_j}$$

$$L(r_i|h_j)match\text{-}mismatch = (1-\varepsilon)^{l-d_j} \cdot \varepsilon^{d_j}$$

where i_i = sequencing read i, h_j = haplotype (allele) j, ε = the sequencing error rate, d_j = the edit distance between read r_i and haplotype h_j, and l = the length of the alignment of read r_i. The read likelihoods were then multiplied to obtain an overall likelihood for each genotype using the above equation.

Instead of using an HMM like `HaplotypeCaller` uses, a simpler edit distance model was used because of the high accuracy from the long haplotypes provided by PacBio HiFi sequencing. A **modified edit distance** calculation was tested as well, where indels were only considered to have an edit distance of 1, instead of an edit distance equal to the length of the indel. Error rates of 0.1% and 1% were both tested. While there was no difference between using the match-mismatch or mismatch model, the best results were obtained with an error rate of 1%, and with the regular edit distance for the HPRC++ samples or the modified edit distance for the OHS samples.

3.4.4.2 The consensus model

The consensus model uses the `abPOA` (Gao et al. 2021) python wrapper `pyabpoa` to generate a POA graph of the reads from which either one or two consensus sequences that best represent the data are determined. The parameters `out_cons=True`, `out_msa=False`, `max_n_cons=2`, and `min_freq=0.25` were used. A consensus sequence is called if at

least 25% of the reads support it. Each consensus sequence is then compared to each known allele sequence to identify the genotype. If a consensus sequence does not perfectly match any known allele, the allele with the lowest edit distance to the consensus sequence is used and flagged as a **novel allele**. If two or more alleles have equally low edit distances to the consensus sequence, the allele that occurs first in the provided list is used as the base allele in the novel flag. No likelihood is provided for the consensus model calls.

3.4.4.3 Validating calls from the long-read genotyping models

The genotype calls from the two long-read models were compared to the truth genotypes. If the models called a different allele not present in the truth genotype, reads were realigned to that allele and visualized in `IGV` to double check that the truth allele was correct. Samples with novel alleles were subsequently genotyped with the realignment and consensus models again using the *PRDM9*-642 list of alleles (see **Chapter 2 Results 2.2.1**) to see if any were blood/sperm alleles. If the models updated the genotype to switch the novel allele to one observed from the *PRDM9*-642 list, the sample was realigned to a genotype-specific reference that included the corresponding blood/sperm allele and the call was then confirmed by visualizing the new genotype-specific alignments in `IGV`. If the genotype still had a novel allele after comparison to the *PRDM9*-642 list, the consensus sequence for the novel allele was further characterized by searching within it for all known zinc finger repeat sequences found in *PRDM9*-642 alleles to piece together the zinc finger array. Novel alleles and novel zinc finger repeats were given standardized names in continuation with the *PRDM9*-642 list.

3.4.5 Validating calls from the short-read genotyping models

To assess all aspects of the short-read genotyping models, the Illumina data for each sample from both datasets were genotyped by the three count methods, the four distance methods, and the 12 combinations of count and distance methods. I tested several *k*-mer lengths: 11–231 in increments of four for sample HG005, which had 250bp reads, and 11–131 in increments of four for the other samples, which had 150bp or 151bp reads. For the count-coverage and all of the distance models, `samtools stats` (Li et al. 2009) was used to estimate sequencing

coverage, error rates, mean fragment length, and fragment standard deviation across the *PRDM9* zinc finger + 10kb flank region for each sample.

The genotype with the highest likelihood from each model call was compared to the long-read truth genotypes. For truth genotypes with a novel flag, only the **base allele** (e.g. P001 in Novel_Similar_P001) was compared to the short-read calls, since the short-read models cannot detect novel alleles. Both the raw reads and the reads corrected with `AlignCorrect` (see **Chapter 2 Methods 2.4.4.2**) were assessed.

The count model was also assessed by using long-read data to call genotypes, after removing soft clips from the OHS samples with `ngsutilsj bam-removeclipping` (Breese and Liu 2013) to remove the targeted PacBio adapter sequences. The coverage, flank mean, and flank median methods of estimating λ were used on the PacBio HiFi data with a range of k values less than the length of the shortest *PRDM9*-106 allele (21–661 in increments of 20). For the coverage method, sequencing coverage and error rate were estimated with `samtools stats` across the entire targeted region (OHS samples) or across the *PRDM9* zinc finger + 10kb flank region (HPRC++ samples), while read length used the mode length of all reads per sample. For the flank mean and median methods, 10kb flank sequences were provided for the HPRC++ samples, while 716bp left and 233bp right flank sequences were provided for the OHS samples due to the shorter targeted reads. The called genotypes from the count models were compared to the truth genotypes to determine success of the model.

3.4.6 Testing genotyping accuracy with different numbers of input alleles

The short- and long-read genotyping models were rerun on all samples using the *PRDM9*-36 list of known alleles to see if accurate genotyping was in part dependent on the number of alleles tested. Results were compared to the truth genotypes converted to *PRDM9*-36 alleles.

Chapter 4

Improving read alignment in polymorphic genomic regions using graph-based reference genomes

4.1 Background

Read alignment is traditionally performed with a reference genome, which is considered a "gold standard" representation of a genome. The GRCh38 reference genome is a mosaic of different individuals, is haploid, and does not include known variants within the main chromosome contigs. Although GRCh38 provides 261 alternate haplotypes for 178 polymorphic regions of the genome in the form of alternate contigs, they are often ignored by researchers to simplify read alignment and subsequent variant calling, in part because alignment tools generally penalize reads that map to multiple locations (Church et al. 2015). The lack of variant information in the reference also leads to a reference bias during read alignment, where reads map better if they contain a reference allele compared to those with a variant allele. Reads that are highly divergent from the reference genome are often lost or misaligned and therefore not utilized properly in downstream variant analyses, leading to a **reference** or **mapping bias** of overestimating reference allele frequencies (Vijaya Satya et al. 2012). Regions of the genome that are highly repetitive are less mappable than less complex genomic regions due to the lack of unique sequence in the reference (Schwartz et al. 2011).

Graph-based references have been proposed as a better representation of genomes. Small indels, SNVs, and large SVs can all be represented in a single graph structure, allowing for numerous known variants to be present in one reference with the possibility of improving read alignment. Several studies have shown such an improvement in alignment (Dilthey et al. 2015; Novak et al. 2017; Hickey et al. 2020; Li et al. 2020; Sirén et al. 2021; Hickey et al. 2022; Liao et al. 2022; Sibbesen et al. 2022), but there has not yet been extensive research into how the topologies of reference graphs affect read alignment accuracy, particularly for complex topologies that represent all known variants of a specific polymorphic region.

One challenge with switching to a graph-based reference is the adaptation of alignment software. Many aligners use a **seed-and-extend** strategy, where small subsequences (seeds) from a read are searched for and mapped before extending alignment in both directions (Ye et al. 2015). A match for the seed sequence is searched within the reference genome index to find the location of the seed sequence in the reference. After locating the seeds to the reference, the rest of the read is aligned by dynamic programming to local subsequences of the reference. If multiple distinct mapping or alignment locations are found, the program typically chooses the best alignment and outputs a mapping quality representing its confidence that the selected alignment is correct. Although the algorithm for aligning a sequence to a graph structure was determined two decades ago (Lee et al. 2002), mapping a read to a graph-based reference is complicated by the numerous possible ways one can traverse through the nodes, and mapping software needs to prioritize which paths should be assessed for potential refinement (extension) of the read alignment. Using reference graphs is therefore more complicated and more computationally intensive than using a linear reference. Nonetheless, a handful of graph aligners have been developed to further develop research into the practicality of using graph-based references.

In this chapter I construct different graph topologies for the polymorphic *PRDM9* zinc finger region to determine how graph topologies affect read alignment. I assess how well different topologies represent the *PRDM9*-36 alleles and whether read alignment scores improve when reads are mapped to the graphs relative to GRCh38 while comparing three different graph aligners. Finally, I use the best performing graph topology and software to assess improvements in read alignment for the HPRC++ Illumina data.

4.2 Results

4.2.1 Constructing *PRDM9* reference graphs with different topologies

Current research has shown that graph-based reference structures improve read mapping for many parts of the genome, and there are many software tools available to generate and work with reference graphs. The structural components of these graphs, however, have not been studied in depth. Given the high polymorphism of *PRMD9,* a graph structure seems beneficial for mapping allelic variation, but could become quite complicated with many possible paths to traverse through the nodes. I hypothesized that read alignment accuracy deteriorates with more complex graphs. To quantify the effect graph topology has on read alignment accuracy, five different versions of reference graphs were constructed using the *PRDM9*-36 (see **Chapter 2 Results 2.2.1**) allele sequences: allele-msa, allele-msa-acyclic, allele-stacked, zinc-finger-msa, and zinc-finger-loop (**Methods 4.4.1**).

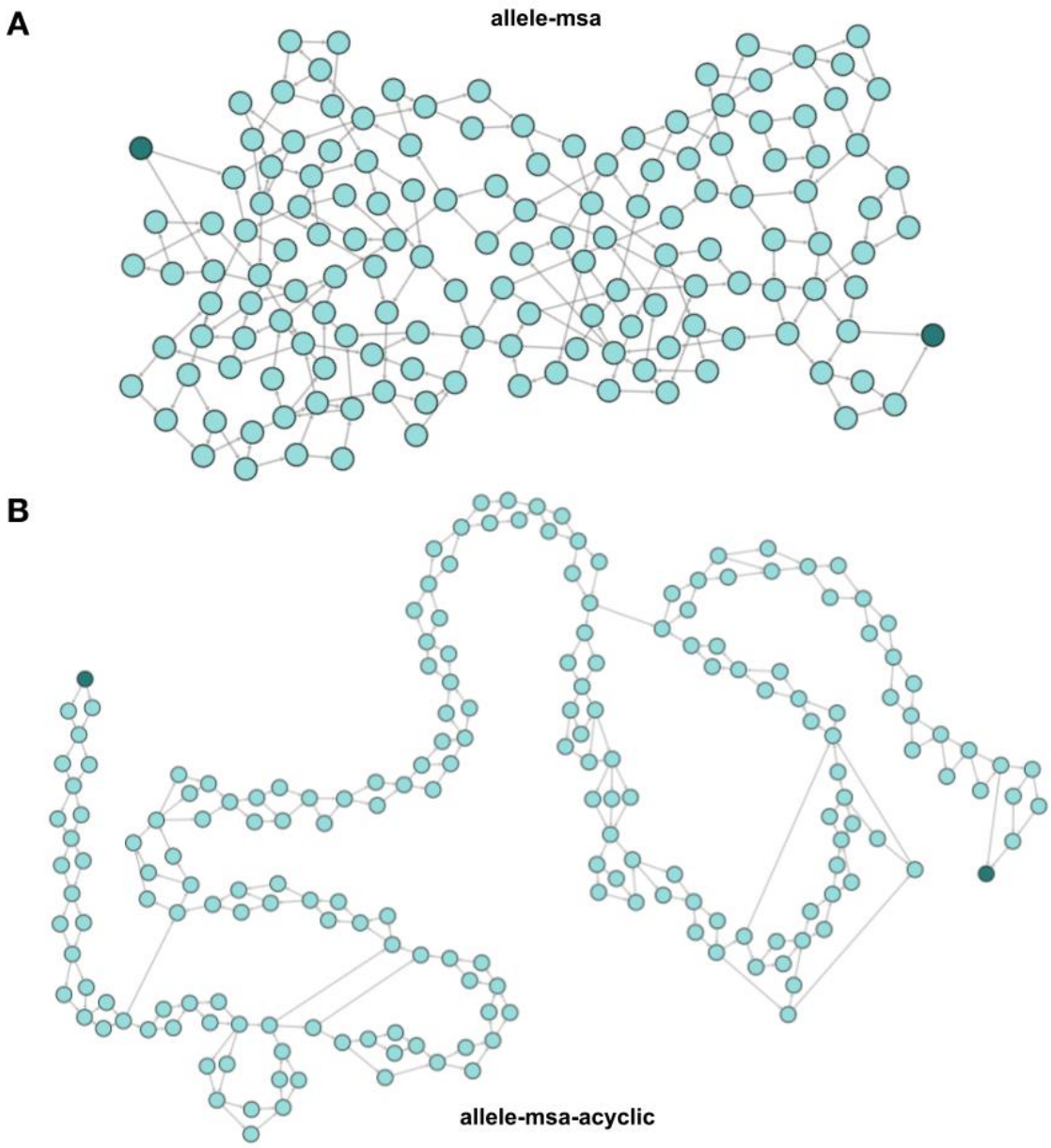
The **allele-msa** graph was constructed from a multiple sequence alignment of all 36 alleles (zinc finger array + 10kb flanks). The alignment was condensed into a single node at sites with identical nucleotides (**Figure 4.1 A**). This is similar to how a POA graph is generated, which condenses multiple sequence alignments into single-base nodes (Lee et al. 2002). The software used to generate the allele-msa graph unexpectedly produced a cyclic graph, meaning one or more edges looped back to connect a downstream node to an upstream node. To see how the cyclic aspect affected alignment, the **allele-msa-acyclic** graph was also constructed for comparison (**Figure 4.1 B**).

The **allele-stacked** graph was constructed by placing the full zinc finger array sequence for each allele into separate nodes, along with additional nodes for the left and right 10kb flank sequences. The left flank node was connected to each allele node, and each allele node was connected to the right flank node (**Figure 4.1 C**).

The **zinc-finger-loop** graph was constructed by placing each individual zinc finger sequence into separate nodes. Edges were used to connect nodes if a pair of zinc fingers appear

sequentially in the array of a known allele. A node for the left flank was then added with a connection to the ‘a’ zinc finger node, since all of the *PRDM9*-36 alleles have arrays that start with ‘a’. Finally, a right flank node was added, and edges from the ‘i’ and ‘j’ zinc finger nodes were added to connect to the to the right flank node, since all *PRDM9*-36 alleles end with either ‘i’ or ‘j’ (**Figure 4.1 D**).

Finally, the **zinc-finger-msa** graph was constructed from a multiple sequence alignment of the 24 individual zinc finger repeat sequences that are observed in the *PRDM9*-36 alleles. As with the allele-msa-graph, the zinc-finger-msa graph had common motifs amongst zinc fingers condensed into single nodes. An edge was added to connect the last node to the first node to allow for traversal through multiple zinc finger sequences. Nodes for the left and right flank sequences were then added, with edges connecting the left flank node to the original first node in the graph, and connecting the original last node to the right flank node (**Figure 4.1 E**).


A
allele-msa
B
allele-msa-acyclic

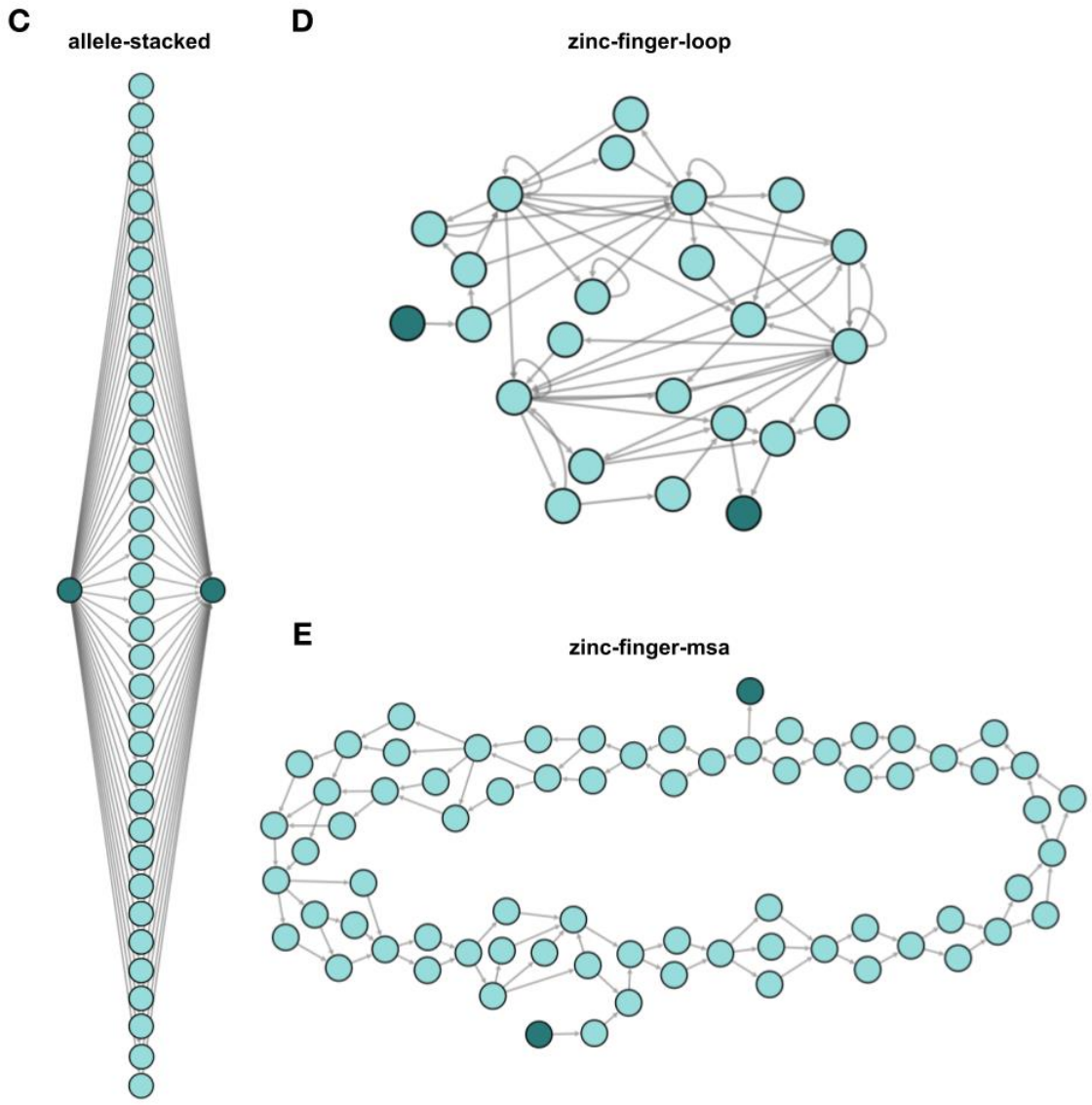


Figure 4.1: *PRDM9* reference graph topologies. Different graph structures for representing the *PRDM9*-36 alleles were constructed to assess the effect of reference graph topology on read alignment. Nodes containing the left and right flank sequences are depicted in dark teal. Nodes are not to scale relative to nucleotide content. Graph images generated with the Cytoscape (Shannon et al. 2003) Assembly app (https://apps.cytoscape.org/apps/assembly). **A)** The allele-msa graph, created from a multiple sequence alignment of all the allele sequences. **B)** The allele-msa-acyclic graph, an acyclic version of the allele-msa graph. **C)** The allele-stacked graph, where each node contains the entire zinc finger array sequence array for one allele. **D)** The zinc-finger-loop graph, where each node contains the entire sequence of an individual zinc finger repeat. **E)** The zinc-finger-msa graph, created from a multiple alignment of all the individual zinc finger sequences.

Seven topology characteristics were determined for each graph: 1) the number of nodes. 2) the number of edges, 3) the highest node degree (i.e. the highest number of edges connected to a single node), 4) the average degree of the nodes (i.e. the edge:node ratio), 5) the total length of sequences in all the nodes combined, 6) the graph compression (i.e. the ratio of the GRCh38 reference sequence length to the total number of bases in the graph), and 7) whether the graph was cyclic (**Methods 4.4.1**). These characteristics, summarized in **Table 4.1**, give an idea of the **complexity** or connectivity of each graph topology. More complex graphs have a higher average degree and tend to have a lot of smaller nodes representing SNVs. A positive correlation between graph compression and variant calling accuracy has previously been reported (Novak et al. 2017), where less overall sequence represented in the nodes resulted in better read alignments for the regions tested.

Table 4.1: Characteristics of different *PRDM9* reference graph topologies. Per-graph summary of the number of nodes and edges, the highest degree for nodes (i.e. highest number of edges connected to a single node), the average degree of nodes (i.e. the edge:node ratio), the total length of sequences in the nodes, the graph compression (i.e. the reference sequence length:total graph length ratio), and whether the graph is cyclic or not. Values were calculated on graphs prior to final modifications (see **Methods 4.4.1**), except for the single-base average degree, which was calculated on graphs modified to have single-base nodes and 1bp flanking sequences.

Graph topology name and structure	Number of nodes	Number of edges	Highest degree	Average degree		Total length	Graph compression	Cyclic
				Original	Single base			
allele-msa Multiple sequence alignment of the allele sequences	150	225	6	1.50	1.08	21,089	1.000	Yes
allele-msa-acyclic Acyclic multiple sequence alignment of the allele sequences	201	276	5	1.37	1.05	21,683	0.973	No
allele-stacked Each allele sequence in an individual node	38	72	36	1.89	1.00	59,564	0.354	No
zinc-finger-loop Each zinc finger repeat sequence in an individual node	24	61	14	2.54	1.02	21,848	0.965	Yes
zinc-finger-msa Multiple sequence alignment of the zinc finger repeat sequences	71	100	6	1.41	1.18	20,156	1.046	Yes

4.2.2 Read alignment accuracy to reference graphs is dependent on graph topology and alignment software

4.2.2.1 Reference graphs with a high density of variants have reduced short path accuracy

During an assessment of reference graphs by the GA4GH (https://www.ga4gh.org), different graphs were developed for a few polymorphic regions and were compared in terms of how accurate and unique read alignment was, along with different characteristics of the graphs. One

metric examined was **short path accuracy**, which is a measure of *k*-mer precision and *k*-mer recall (Novak et al. 2017). ***k*-mer precision** is the proportion of graph *k*-mers also found in the list of allele *k*-mers, or a measure of how much the various traversals through the graph deviate from known allele sequences. ***k*-mer recall** is the proportion of allele *k*-mers also found in the list of graph *k*-mers, or a measure of how completely the graph represents the different alleles. It was hypothesized that the graphs would have better *k*-mer recall compared to GRCh38 due to their ability to incorporate all known *PRDM9* alleles, but worse *k*-mer precision due to the numerous possible traversals through the graph nodes.

The series of nodes traversed through a graph that represents a particular sequence is known as a **path**. To examine short path accuracy for the different *PRDM9* reference graph topologies, I obtained the union of *k*-mers from all 36 alleles, then obtained graph-specific *k*-mers by traversing all possible paths in each graph, using lengths of *k* from one to 99 in increments of two. Since the majority of *k*-mers would be originating from the 20kb of linear flanking sequences that were identical amongst the graphs and GRCh38, the graphs were modified to remove the left and right flanking nodes prior to *k*-mer collection. A pseudo-F1 score was calculated to summarize the *k*-mer precision and recall values at each value of *k*. The graph F1 scores were compared to the F1 scores for the *PRDM9* zinc finger array sequence (without 10kb flanks) in GRCh38 (**Methods 4.4.2**).

The allele-msa, allele-msa-dag, allele-stacked, and zinc-finger-loop graphs had perfect *k*-mer recall for all values of *k*, meaning they completely represented all 36 allele sequences (**Figure 4.2 A**). Two graphs, allele-stacked and zinc-finger loop, also had perfect *k*-mer precision, giving them perfect F1 scores of 1 across all *k*-mer lengths. For the allele-msa and allele-msa-dag graphs, *k*-mer precision lowered as the value of *k* increased, owing to the numerous possible traversals through a region with a high density of connected nodes that spell out *k*-mers not actually observed in any known allele sequence (**Figure 4.2 C**). Interestingly, the zinc-finger-msa graph started with perfect *k*-mer precision and recall, then *k*-mer precision dropped as *k* increased, and finally *k*-mer recall began to drop at $k = 59$ until both precision and recall reached zero at $k = 93$. The zinc-finger-msa graph was highly connected, and a single pass through the cyclic graph corresponded to 84bp. As *k* increased to be longer than a single zinc finger repeat, the number of possible paths through the graph became very large,

generating an extremely high number of possible *k*-mers (**Figure 4.2 B**). The drop in *k*-mer recall was unexpected since all 36 alleles were used to generate the graph. This could be the result of an issue with the graph creation or perhaps with difficulty in obtaining all of the possible *k*-mers at such large lengths. In contrast, the two graphs with perfect *k*-mer precision were strategically designed to have long nodes that only connected with other nodes if necessitated by the arrangement of zinc finger repeats (zinc-finger-loop) or to connect to the left and right flanks (zinc-finger-loop and allele-stacked). Since the longest *k*-mer tested (99) was not much longer than the nodes representing full 84bp zinc finger repeats in the zinc-finger-loop graph, the *k*-mers could not loop through several nodes in a way that could generate a novel sequence.

GRCh38, on the other hand, had perfect *k*-mer precision and decreasing *k*-mer recall as *k* increased. It did not contain any *k*-mers that would not be observed in the sequence of any *PRDM9* allele because GRCh38 represents the sequence for allele B and the reference can only be traversed in a single path. In terms of *k*-mer recall, GRCh38 had a slightly lower value due to its inability to represent any of the *k*-mers unique to non-reference alleles.

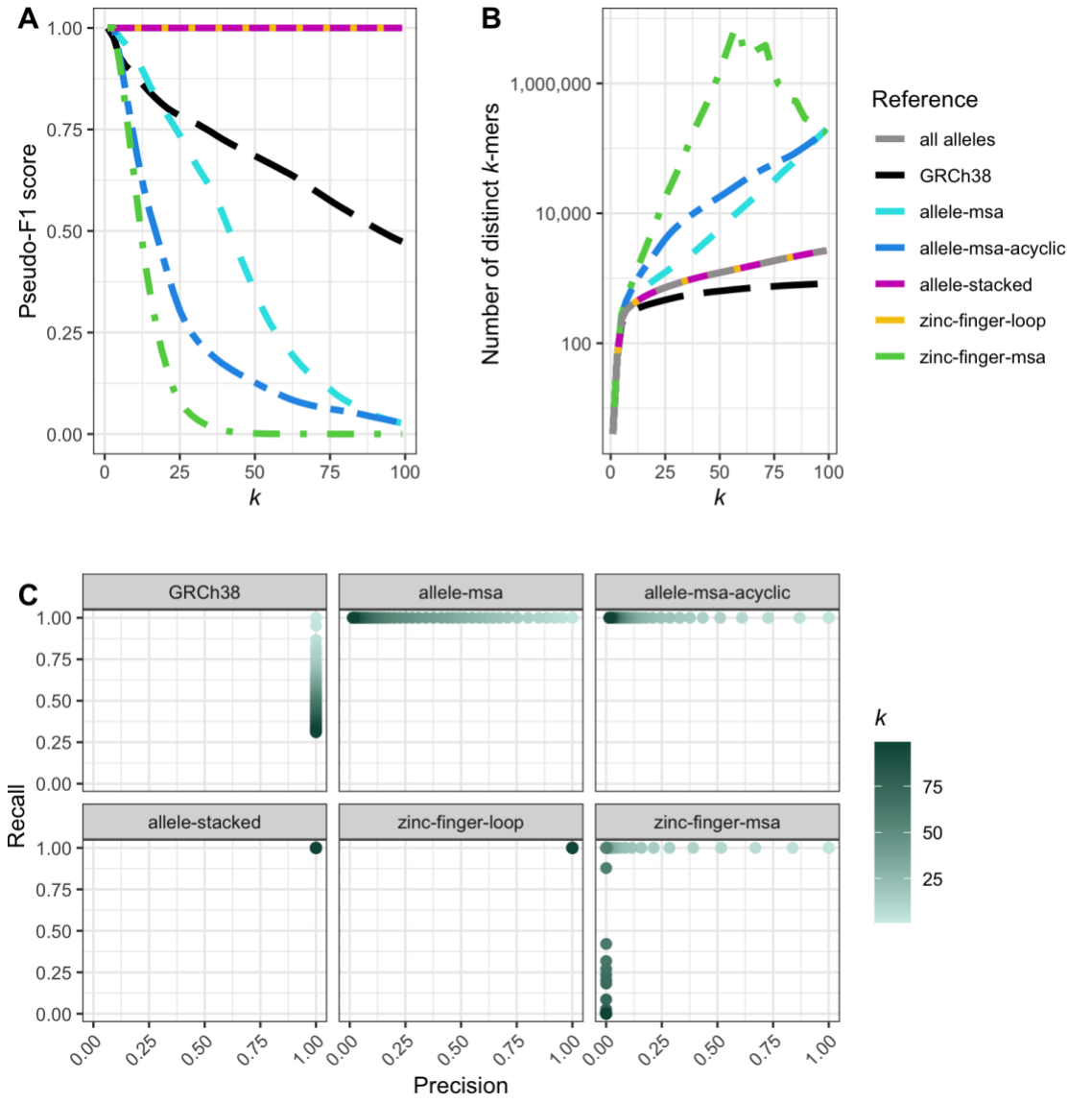


Figure 4.2: Short path accuracy for different *PRDM9* reference graph topologies. A) Pseudo-F1 scores for *k*-mer precision and *k*-mer recall for the different graph topologies (colors) and GRCh38 across a range of *k*-mer lengths. *k*-mer precision is the proportion of *k*-mers found in graph paths also found in the union of allele sequence *k*-mers, and *k*-mer recall is the proportion of allele *k*-mers also found in the paths of a graph. **B)** Number of unique *k*-mers per graph topology, GRCh38, or the union of all allele *k*-mers (colors). The zinc-finger-msa graph had substantially more *k*-mers than the other topologies. **C)** *k*-mer precision-recall plots for each graph topology and GRCh38 across different values of *k* (shades of teal). The allele-stacked and zinc-finger-loop graphs both had perfect *k*-mer precision and perfect *k*-mer recall. GRCh38 had perfect *k*-mer precision while *k*-mer recall decreased with increasing *k*. The allele-msa and allele-msa-acyclic graphs had perfect *k*-mer recall while *k*-mer precision decreased with increasing *k*. The zinc-finger-msa graph started with perfect *k*-mer precision and recall, but decreased first in precision and then in recall as *k* increased.

4.2.2.2 Declaring expected fragment lengths improves read alignment to *PRDM9* graphs with the `vg` aligner

A few software applications have been developed that enable the mapping of sequencing reads to graph-based reference genomes. Three were tested on the *PRDM9* reference graphs to compare read alignment accuracy: `vg` (Hickey et al. 2020), `GraphAligner` (Rautiainen and Marschall 2020), and `minigraph` (Li et al. 2020). All three aligners provide an alignment metric for **read divergence**, which is the proportion of bases in the read that do not match the corresponding reference sequence. To get a comparable metric for aligning reads to the traditional reference, the *PRDM9* + 10kb flanks sequence from GRCh38 was converted to a graph format. Conceptually this can be thought of as a single-node graph, though technically the sequence was segmented due to restrictions in how the index can represent long nodes. The resulting "**flat graph**" had only one possible traversal since there were no variants included. The 100X simulated reads from the primary haploid simulation set (see **Chapter 2 Methods 2.4.2**) were mapped by each aligner to the GRCh38 reference graph and to each graph topology. The **difference in divergence** was calculated for each read by subtracting the divergence value obtained when mapping the read to the graph from the divergence value obtained when mapping the read to the GRCh38 flat graph. If the difference was positive, this meant the read mapped better to the graph than to GRCh38, while a negative difference meant the read mapped better to GRCh38 than to the graph.

`GraphAligner`, having been built for aligning long reads, does not use paired-end information when mapping short reads. While `minigraph` does incorporate fragment length into mapping considerations, the only related adjustable parameter is for maximum fragment length. For `vg`, however, there are three parameters for handling paired-end reads that affect how expected fragment lengths and standard deviations are determined: 1) specifying the expected values that remain fixed throughout the alignment process (**fixed**), 2) specifying the expected values but letting `vg` adjust them as alignment progresses (**variable**), and 3) not specifying the expected values and letting the program determine them during the alignment process (**not specified**). I assessed these three methods to initially determine the optimal parameters for using `vg` before comparing the software to `GraphAligner` and `minigraph`

(**Methods 4.4.2.2**). Unfortunately, `vg` was unable to generate an index for or perform alignments to the allele-msa graph due to its high complexity.

Given that expected fragment lengths can help map ambiguous reads when their mate is able to be placed on a reference genome, it was hypothesized that declaring the expected fragment length distribution would improve read mapping over letting `vg` calculate or adjust the values, which may be incorrect since reads are originating from a highly repetitive region and might map to multiple locations. The average difference in divergence (which was calculated only from reads that had a difference in divergence > 0) per allele was variable within a given graph and alignment setting (**Figure 4.3 A**). Variability across alleles was expected since some alleles are more divergent from the reference allele. Considering all alleles, the results for the allele-msa-acyclic, allele-stacked, zinc-finger-loop, and zinc-finger-msa graphs were very similar to each other. When comparing the different fragment length treatments, the fixed method was more consistent among the alleles, whereas the variable and not specified methods resulted in a wider spread of average difference in divergence amongst the alleles, with some alleles having negative differences and hence resulting in worse alignments than GRCh38, which was not expected. Results were more consistent for the error-free simulations than the erroneous simulations, particularly for the fixed model. Interestingly, the distribution of average differences in distribution did not shift towards zero for the simulations with higher error rates, as would be expected if an erroneous read no longer matched to a *PRDM9* sequence in the graph. When considering the proportion of reads that had a difference in divergence, the three fragment length models had very similar results (**Figure 4.3 B**). The overall proportion of affected reads was only about 1% for error-free reads and slightly higher for the erroneous reads. An even smaller proportion of reads did not map to the graphs (~0.15%), with the variable and not specified fragment models of alignment resulting in more unmapped reads.

Overall, the fixed fragment model was the most consistent across alleles and simulated error rates. Given the repetitive nature of the *PRDM9* zinc finger array, it is possible that `vg` is incorrectly estimating the average fragment length and standard deviation when it performs the calculations itself during alignment since many reads can align equally well to multiple positions. The fixed model was used with `vg` for the remaining graph alignment analyses.

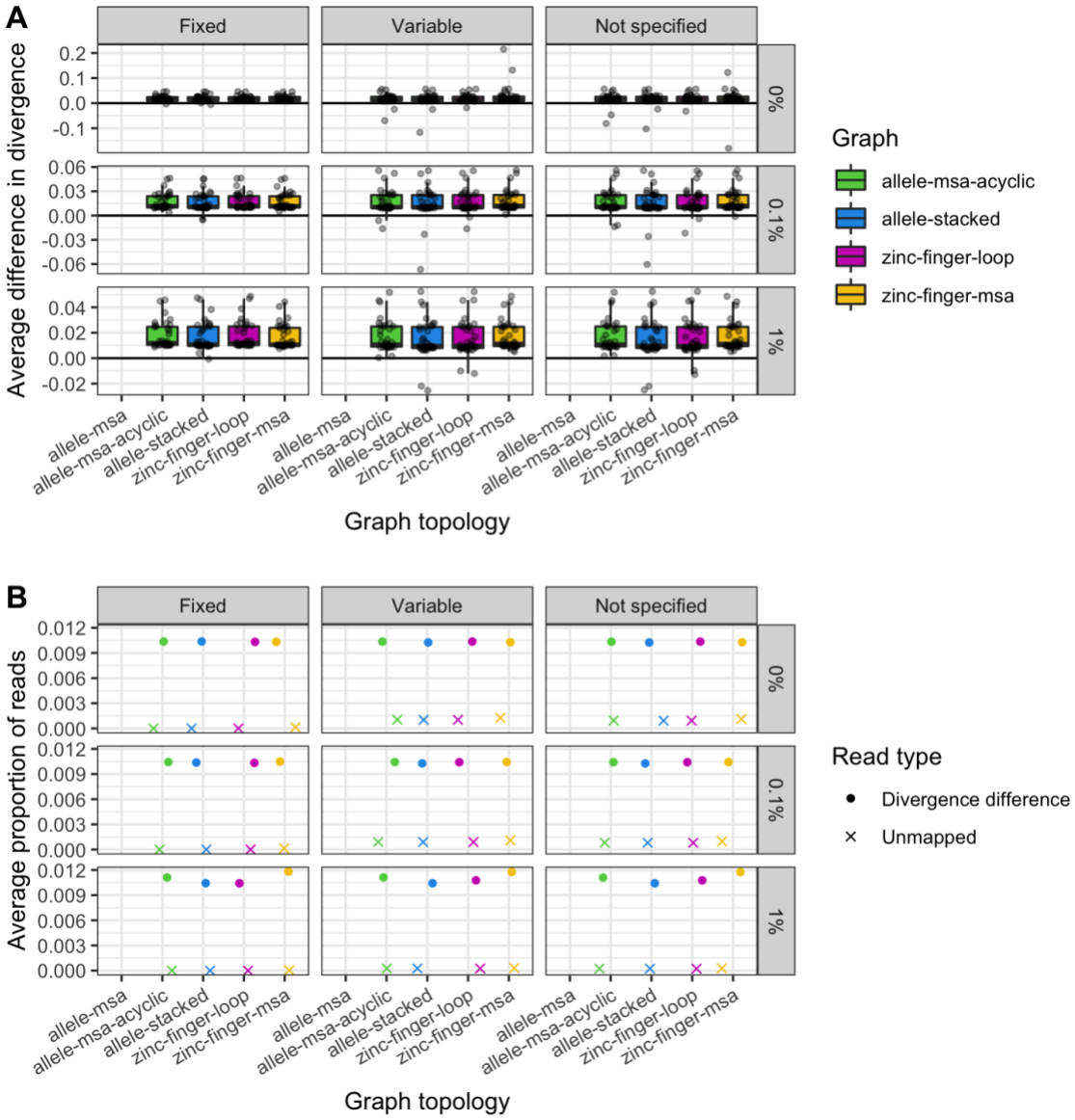


Figure 4.3: Performance of different paired-end read treatments during alignment by `vg`. The difference in read divergence between `vg` alignments of 100 iterations of simulated 100X paired-end reads to GRCh38 and to different *PRDM9* graph topologies (colors) was calculated for each read. Simulations were performed with different sequencing error rates (rows). Three methods of handling paired-end reads (columns) were assessed: fixed, where an expected fragment length mean and standard deviation are provided to `vg` and used as is; variable, where the provided expected fragment length mean and standard deviation are used as a starting point and adjusted by `vg` as alignment progresses; and not specified, where `vg` calculates the expected fragment length mean and standard deviation during alignment. `vg` was unable to align to the allele-msa graph due to indexing issues. **A)** Differences in divergence averaged across reads for each allele. **B)** Proportion of reads either unmapped or with a difference in divergence between alignment to GRCh38 and the graph, averaged across all alleles.

4.2.2.3 Graph-based references are prone to spurious alignments

The *PRDM9* allele represented in GRCh38 is allele B. Given that the reference graphs have perfect *k*-mer recall, there should be no differences in alignment when a read simulated from allele B is mapped to a graph compared to when it is mapped to GRCh38. Due to the high similarity of many *PRDM9* alleles, where some are differentiated by only a single SNV, it is possible that a sequencing error could inadvertently cause a read to better align to a graph path that does not match the allele of origin for a read. The result of this scenario would be a better-than-expected alignment to the graph and thus a lower divergence value. These **spurious improvements** in alignments, when a read aligns better to a graph than to the allele from which it was sequenced, are only expected to be observed in reads simulated with sequencing errors, assuming aligners are capable of finding optimal alignments. **Spurious impairments**, when a read aligns worse to a graph than to the allele from which it was sequenced, are not expected to be observed as long as all of the complete original allele sequences are fully represented in the graph. To assess how frequently these unexpected alignments occurred, I compared alignments to each of the graph topologies to **self alignments**, where reads were aligned to the allele sequence from which they were simulated (e.g. reads simulated from allele A were aligned to the allele A sequence). Only error-free simulations were considered to obtain a clear picture of any potential biases in read mapping, since it is not known what the divergence values should be for the reads simulated with sequencing errors. The difference in divergence was again calculated for each read and averaged for and across all alleles (**Methods 4.4.2.2**).

All aligners resulted in some spurious alignments for all graph topologies. The differences in divergence were highest and most variable for `vg` alignments and lowest for `minigraph` alignments (**Figure 4.4 A**). Improvements had differences in divergence of more than 0.2 on average for some alleles when aligned by `vg` to the allele-msa-acyclic, allele-stacked, zinc-finger-loop, and zinc-finger-msa graphs, while impairments had differences of over 0.25. This corresponds to a staggering 20–25% of the read being differently aligned between the graphs and the originating allele sequence. Given that error-free simulated reads were assessed, only a small proportion of spurious alignments was expected for all aligners, if any. The proportion of spuriously aligned reads, however, was over 10% for `minigraph` alignments compared to around 0.01% for alignments with `vg` (**Figure 4.4 B**). Alignments by `GraphAligner` fell

between those by `vg` and `minigraph`; differences in divergence peaked around 0.1 in both directions for alignments to the allele-msa and allele-msa-cyclic graphs, with at most 3.1% of reads impaired when aligned to the allele-msa graph and 1.2% improved when aligned to the zinc-finger-loop graph. Interestingly, no reads had spurious alignments to the zinc-finger-msa graph when mapped with `GraphAligner`. Using `minigraph` lead to variable proportions of unmapped reads to each of the different graphs, from 0.00001% to 3.3%, whereas `vg` and `GraphAligner` only left reads unmapped to the zinc-finger-msa graph, corresponding to 0.01% and 5% of the reads, respectively.

The larger and more variable average differences in divergence for `vg` alignments seem to be a factor of the very small proportion of affected reads as opposed to a particular deficiency in the alignment algorithm. Since only error-free reads were examined, neither type of spurious alignment was expected. One explanation for the spurious impairments is that the aligners are not able to efficiently find the optimal alignment to the graphs, which is quite plausible given the high density of variants in some of the tested topologies. Spurious improvements, on the other hand, are more difficult to explain, given that the linear self references should be easier to map to compared to the graph references. Computationally speaking, it is possible that some optimal alignments are missed in favor of faster processing times.

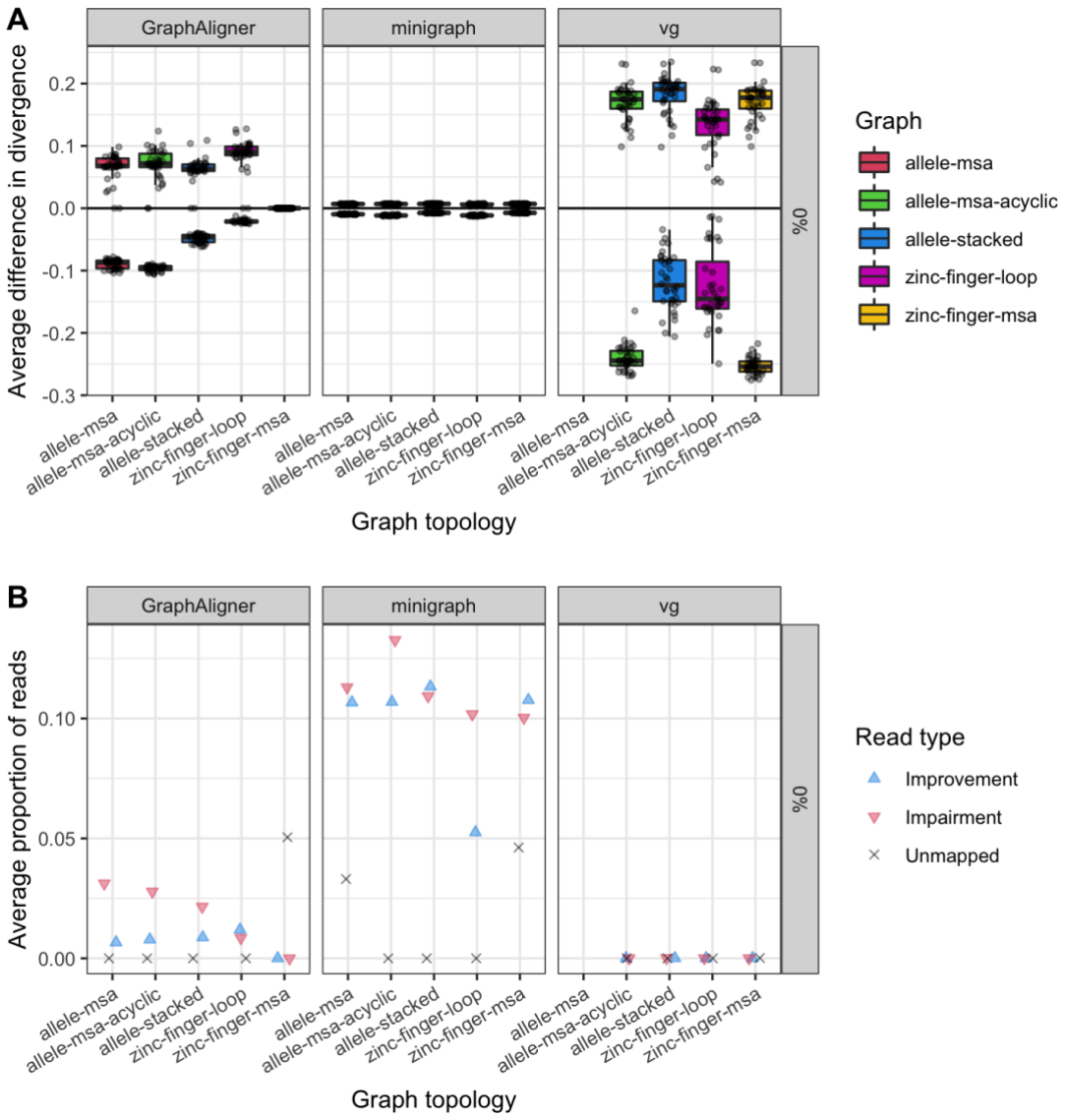


Figure 4.4: Spurious improvements and impairments in read alignments to the different *PRDM9* graph topologies. The spurious difference in read divergence between self alignments of simulated 100X error-free paired-end reads and alignments to different *PRDM9* graph topologies (colors) was calculated for each read, testing different graph alignment software (columns). Spurious improvements (positive differences) are when reads have better alignment to a graph than to the originating allele sequence, while spurious impairments (negative differences) are when reads have worse alignment to a graph than to the originating allele sequence. `vg` was unable to align to the allele-msa graph due to indexing issues. **A)** Average spurious differences for and across alleles, with separate boxplots for improvements and impairments. **B)** Average proportion of reads across all alleles either unmapped or with spurious improvements or impairments.

4.2.2.4 Read alignments by `vg` improve when paths for known alleles are embedded in the *PRDM9* graph structures

When constructing the different graph topologies, all of the graphs had **embedded paths** for each *PRDM9*-36 allele, meaning the specific series of nodes that represented each allele was part of the graph structure file provided to the alignment software. For graphs with a high density of variants, the presence of paths can theoretically help aligners identify optimal mappings for reads by narrowing down the number of possible paths to explore for alignment. Since the allele-msa, allele-msa-acyclic, and the zinc-finger-msa graphs all had lower *k*-mer precision compared to GRCh38, it was hypothesized that having embedded path information would improve alignments to these graphs. To see if paths affected read mapping for each of the aligners to the different graphs, the path information was removed from each graph topology and alignments to graphs and self references were performed again (**Methods 4.2.2.2**).

Removing paths had little to no effect on `GraphAligner` and `minigraph` in terms of both the average spurious difference in divergence (**Figure 4.5 A**) and in the proportion of unmapped and spuriously mapped reads (**Figure 4.5 B**). For `vg`, removing paths in the graphs decreased the average spurious difference in divergence substantially for all graphs, all the way down to zero for three of the topologies. However, removing paths also led to an increase in unmapped reads by `vg` for the allele-msa-acyclic, zinc-finger-loop, and zinc-finger-msa graphs, affecting as many as 1.2% of the reads. Given that these three graphs have a high density of variants and thus numerous possible traversals through the nodes, it is logical that removing embedded paths for known alleles would make it more difficult for `vg` to map some reads. The allele-stacked graph, with only 36 possible traversals, was understandably unaffected in terms of unmapped reads after removing the embedded paths. Though not required for `GraphAligner` or `minigraph`, using reference graphs with embedded paths for known alleles appears to be important for aligning reads with `vg`.

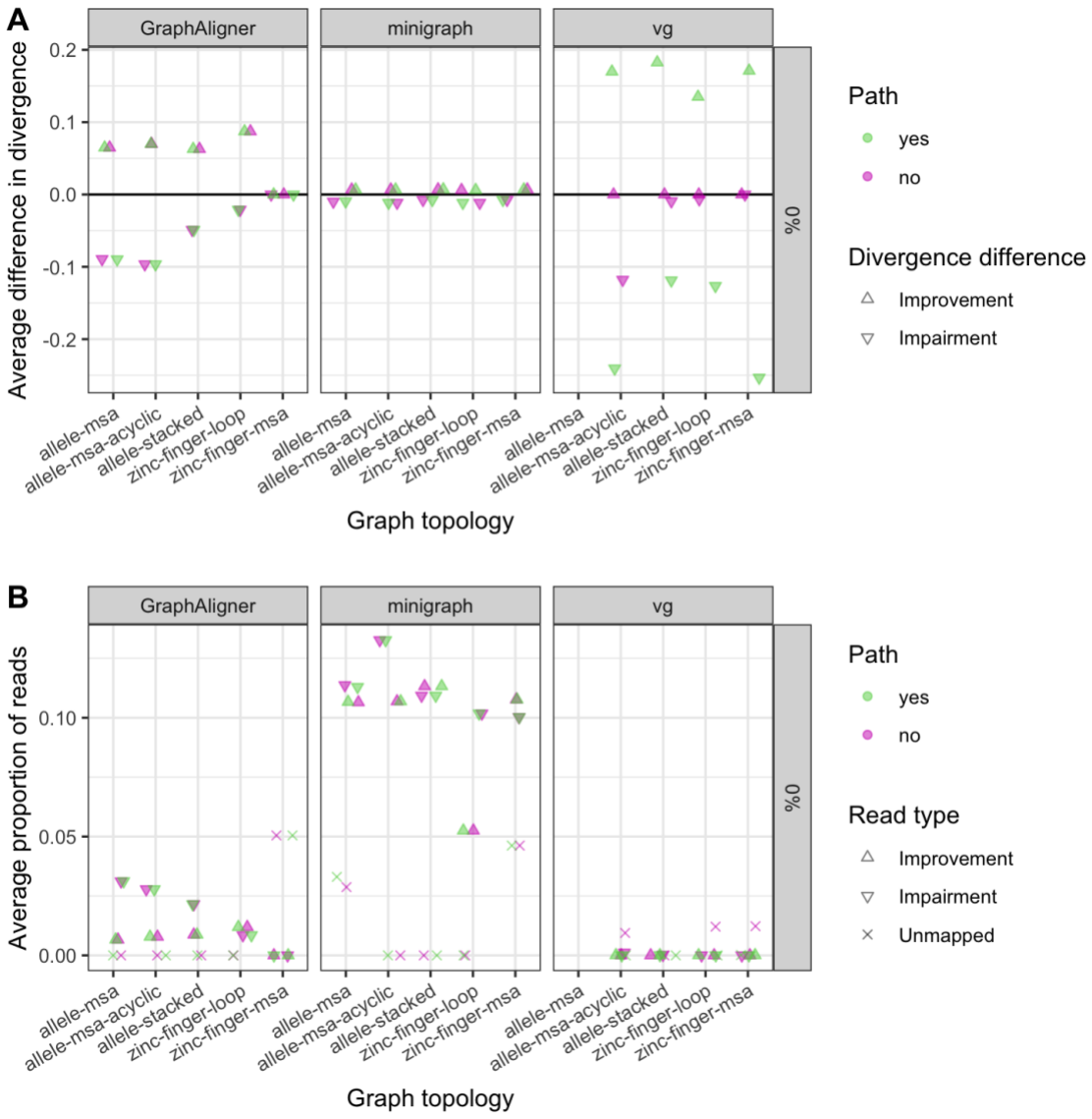


Figure 4.5: Effect of embedded paths in read alignments to the *PRDM9* reference graph topologies. The spurious differences in read divergence between self alignments of simulated 100X error-free paired-end reads and alignments to different *PRDM9* graph topologies was calculated for each read, testing different graph alignment software (columns) and versions of the graphs with and without embedded paths for known alleles (colors). Spurious improvements (positive differences) are when reads have better alignment to a graph than to the originating allele sequence, while spurious impairments (negative differences) are when reads have worse alignment to a graph than to the originating allele sequence. `vg` was unable to align to the allele-msa graph due to indexing issues. **A)** Spurious differences in divergence averaged across all alleles. **B)** Average proportion of reads across all alleles either unmapped or with spurious improvements or impairments.

4.2.2.5 Using *PRDM9* reference graphs results in better read alignment accuracy compared to the GRCh38 reference

Having compared read alignment to the graph references and to the self references for each *PRDM9*-36 allele, the reads that experienced spurious alignments can be identified and removed from the error-free simulations. This then allows for comparisons between using the *PRDM9* graphs and using GRCh38 as a reference to be performed without mistaking spurious differences for genuine differences in read alignment. I reassessed the average differences in read divergence after removing the spuriously aligned reads identified for each allele by each aligner (**Methods 4.2.2.2**).

For `vg`, all graphs resulted in improved alignments relative to GRCh38 when paths were embedded as part of the graph structures. Without paths, alignments to allele-msa-acyclic were impaired, while alignments to the other three graphs were slightly higher than the corresponding graphs with paths. The average proportion of reads with true improvements after mapping to graphs with `vg` was fairly low, and there were only unmapped reads when the graphs did not have any embedded path information. Interestingly, alignments to all five topologies by `minigraph` and to the allele-msa, allele-msa-acyclic, and allele-stacked graphs by `GraphAliner` were actually worse overall compared to using GRCh38 as a reference (**Figure 4.6 A**). While the average difference in divergence was less than 0.02 for the `minigraph` alignments, more than 12% of reads were affected for each graph (**Figure 4.6 B**). Removing embedded paths from the graphs again had little to no effect on true alignment differences by `GraphAligner` or `minimap`.

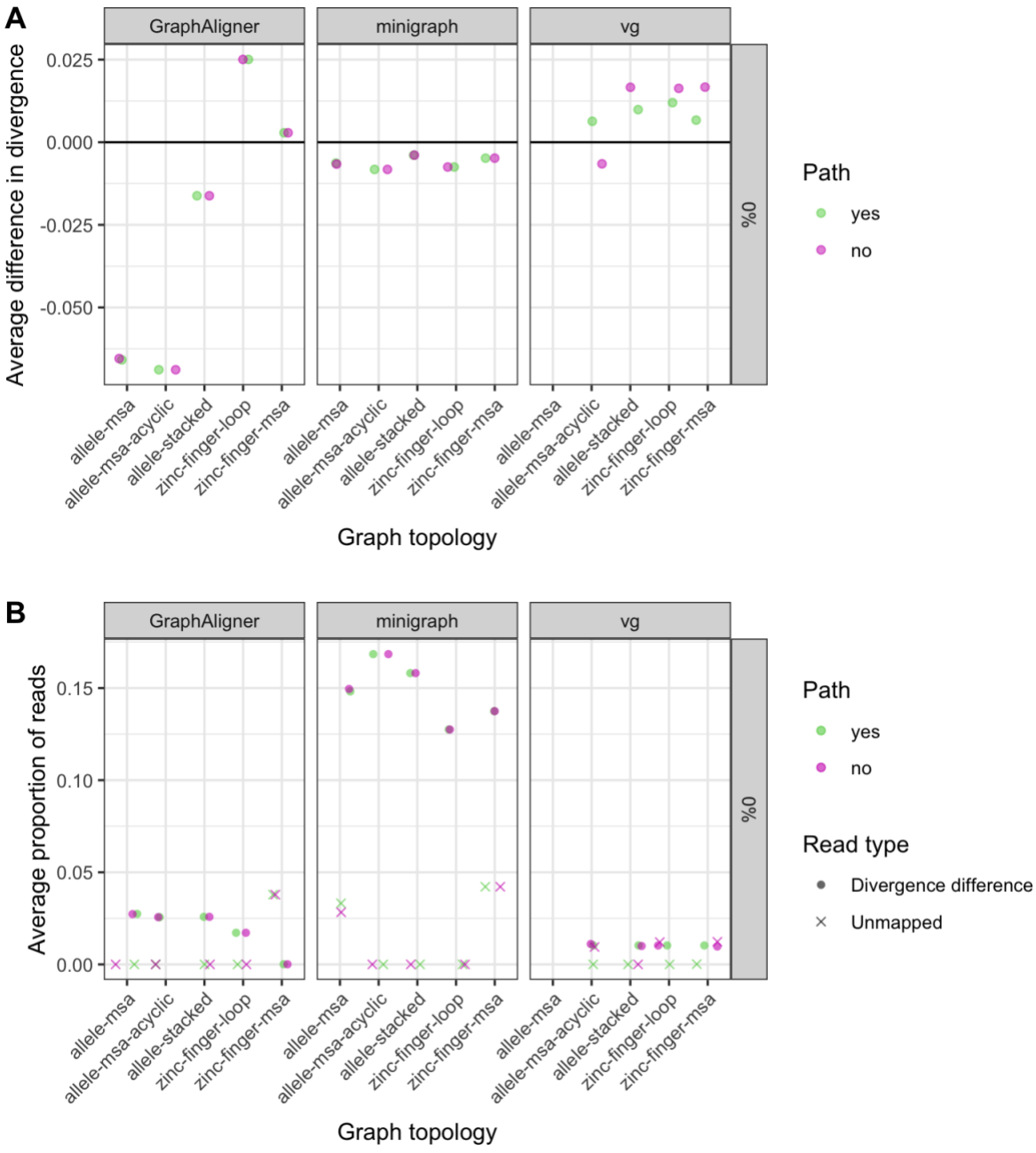


Figure 4.6: True alignment improvements to the different *PRDM9* graph topologies. The difference in read divergence between GRCh38 alignments of simulated 100X error-free paired-end reads and alignments to different *PRDM9* graph topologies was calculated for each read, testing different graph alignment software (columns) and versions of the graphs with and without paths (colors). Alignments were performed after correcting for spurious alignments by removing reads with divergence differences between self references and the graphs. `vg` was unable to align to the allele-msa graph due to indexing issues. **A)** Corrected divergence differences averaged across alleles. **B)** Average proportion of reads across alleles either unmapped or with differences in divergence.

The small proportion of reads with differences in divergence is not surprising given the total 20kb of flanking sequence compared to the lengths of the zinc finger repeat array sequences, which ranged from 672 to 1,512bp. To get a better idea of the impact of graph aligners specifically within the highly polymorphic variable region, the alignments were subset to contain only read pairs where both mates were sequenced within the zinc finger region or up to 100bp away. The average improvement in alignment for the zinc finger region remained about the same as before, which was as expected since the averages were only calculated using reads that had a difference, and the majority of the reads without a difference likely originated from the flanking regions (**Figure 4.7 A**). However, the proportion of reads with differences in divergence rose substantially to about 20% for the `vg` alignments (**Figure 4.7 B**), which is about ten-fold higher than when considering all flanking reads as well.

Overall, aligning to the allele-stacked graph with embedded paths using `vg` and providing fixed fragment length expectations resulted in the best improvements in alignments compared to mapping to GRCh38, taking into consideration the low proportion of unmapped reads. Though it has the second highest average edge:node ratio among the five *PRDM9* graph topologies, the allele-stacked graph has only 36 possible traversals, leading to perfect *k*-mer precision and *k*-mer recall. Aligning to the allele-stacked graph with `vg` had a very low proportion of both unmapped reads and reads with spurious alignments, while still providing an overall improvement in alignments compared to using GRCh38 as a reference.

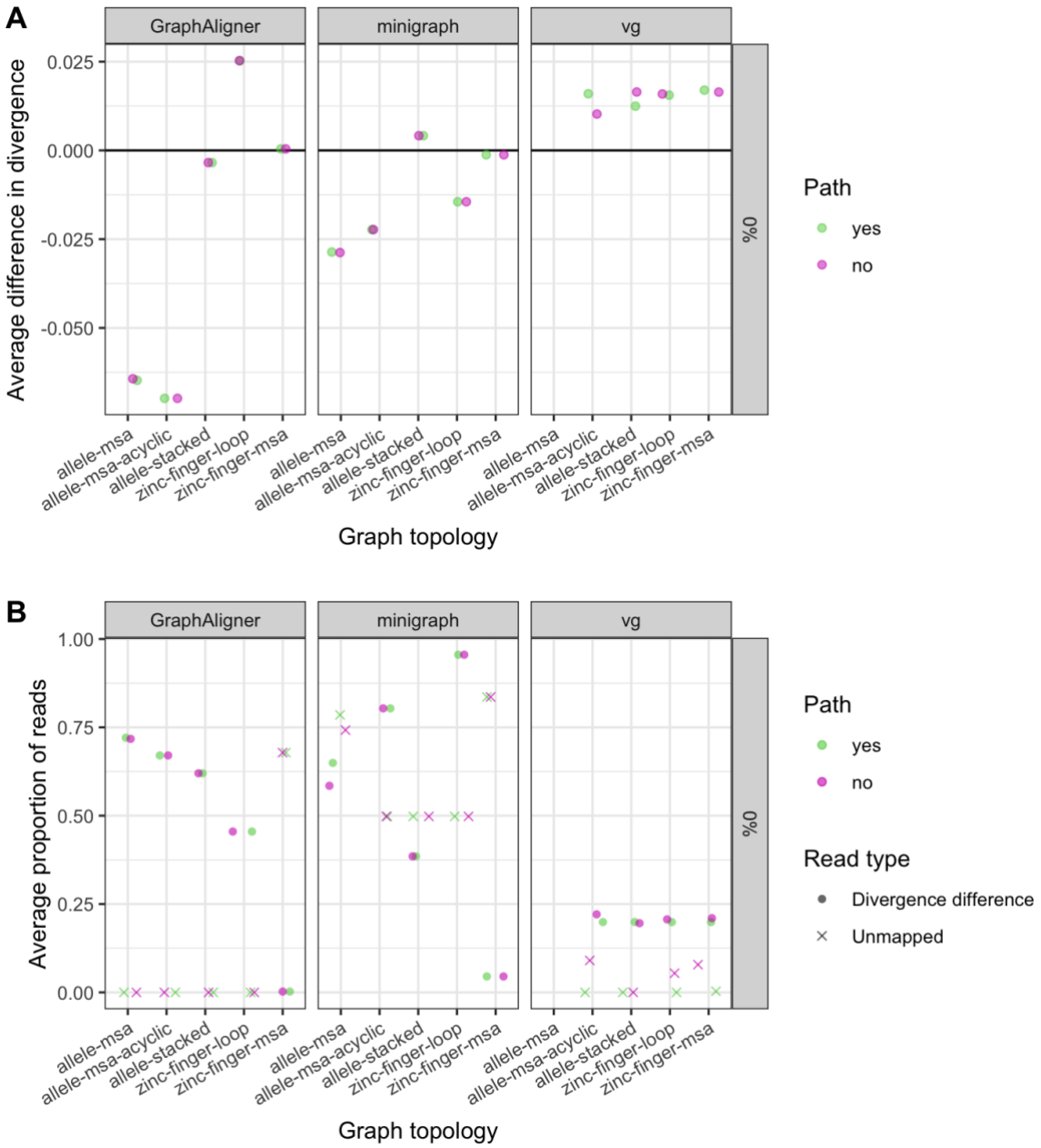


Figure 4.7: True alignment improvements to the different *PRDM9* graph topologies in the zinc finger region alone. The difference in read divergence between GRCh38 alignments of simulated 100X error-free paired-end reads and alignments to different *PRDM9* graph topologies was calculated for each read, testing different graph alignment software (columns) and versions of the graphs with and without paths (colors). Data were analyzed after subsetting the reads to only read pairs that had both mates sequenced from or within 100bp of the zinc finger region, and then correcting for spurious alignments by removing reads with divergence differences between the graphs and self references. `vg` was unable to align to the allele-msa graph due to indexing issues. **A)** True divergence differences averaged across alleles. **B)** Average proportion of reads across alleles either unmapped or with differences in divergence.

4.2.3 HPRC++ samples from African or Admixed American populations have greater improvements in alignment to the *PRDM9* graph than samples from other populations

Given the diversity in *PRDM9* alleles found in the HPRC++ sample genotypes, and the identification of several novel alleles, it was hypothesized that there would be a noticeable improvement in read alignment when using a reference graph over the traditional reference. Using the HPRC++ Illumina dataset (see **Chapter 3 Methods 3.4.1**), bam files subset to contain reads that aligned to the *PRDM9* + 10kb flank region were aligned to both the allele-stacked graph and to the GRCh38 flat graph with `vg`, using sample-specific expected mean fragment lengths and standard deviations as a fixed parameter. The difference in deviation between the graph and GRCh38 references were again calculated for each read and then averaged for each sample, for just reads that originally aligned to within 100bp of the zinc finger array region as well as for all reads across the *PRDM9* + 10kb flanks (**Methods 4.4.3**).

All samples had positive average differences in divergence, which ranged from 0.0041 to 0.0390. When grouped by continental population, the African and the Admixed American samples had the highest median differences in divergence across the full *PRDM9* + 10kb flank region (0.0168 and 0.0169, respectively), followed by the East Asian, European, and South Asian samples, although there was only one sample with South Asian ancestry (**Figure 4.8 A**). The Admixed American samples had the highest median difference in divergence when considering only reads mapping to the zinc finger array region (0.0245), and all populations had higher median differences compared to when reads from the flank regions were included, though the increase was quite small (0.00527 on average).

Seven samples were flagged and confirmed to have novel alleles not present in the *PRDM9*-106 list (see **Chapter 3 Results 3.2.7**). To see if these samples had greater improvement in read alignment, the samples were grouped into those that had a novel allele as part of their genotype and those that did not. There was a wider range of average differences in divergence for the novel group for reads including the flanking regions, but the median differences per group were fairly similar at 0.0158 and 0.0162, respectively (**Figure 4.8 B**). When considering

only the zinc finger region reads, the novel group had a slightly higher median improvement than the non-novel group at 0.0297 compared to 0.0207, respectively.

Across all samples, the improvements in alignment were fairly small, with differences in divergence of 0.017 and 0.022 for the *PRDM9* + 10kb flank and zinc finger array only reads, respectively. This is, however, higher than the improvements averaged across the simulated data (**Figure 4.8 C**). The HPRC++ average proportion of reads with differences in divergence was also higher than the corrected error-free simulations at 1.48% for all reads and 22.3% for just the zinc finger array reads (**Figure 4.8 D**).

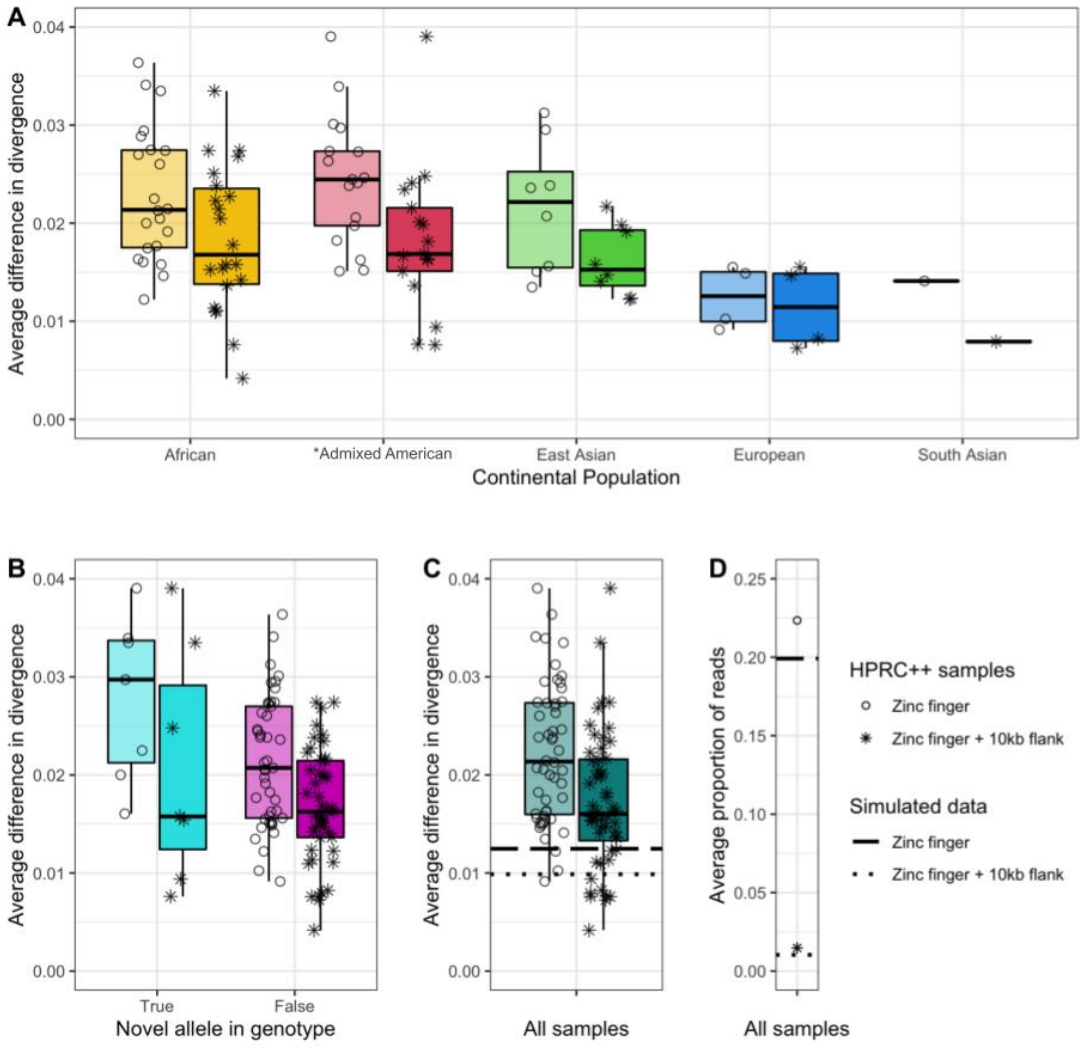


Figure 4.8: Improvement in read alignments for the HPRC++ samples using a *PRDM9* reference graph compared to GRCh38. Reads from the original HPRC++ bams that aligned to the *PRDM9* + 10kb flank region were realigned with `vg` to the allele-stacked graph and to the GRCh38 flat graph, both with embedded paths. The difference in read divergence was calculated for each read and averaged for each sample. Differences were averaged for reads across the full region and for only reads that originally aligned to or within 100bp of the zinc finger repeat array (shapes). **A)** Average improvements in alignment to the *PRDM9* reference graph for samples grouped by continental ancestry (colors). **B)** Average improvements in alignment to the *PRDM9* reference graph for samples grouped by the presence of a genotypic novel allele detected by the consensus genotyping model (colors). **C)** Average improvements in alignment to the *PRDM9* reference graph for all HPRC samples, compared to the average improvements for the corrected error-free simulated data. **D)** Average proportion of HPRC++ reads with a difference in divergence across all samples, compared to the average proportion for the corrected error-free simulated data.

4.3 Discussion

There are several ways to construct a reference graph to represent a particular region of the genome, but the topology can have an effect on read alignment accuracy. Calculating the short path accuracy, or *k*-mer precision and recall, is one way to determine how well a graph represents the known variants for a particular region. High values are desirable for both since perfect *k*-mer precision means the graph does not have paths that spell out sequences not observed in the variants of interest, and perfect *k*-mer recall means the graph represents all known variants within one reference structure. Two graph topologies I tested had both perfect *k*-mer precision and recall: allele-stacked and zinc-finger-loop. Both characterized by longer nodes of at least 84bp, there were fewer opportunities for reads to traverse incorrect paths than for the allele-msa, allele-msa-acyclic, and zinc-finger-msa graphs. A graph with a high density of small variants may have numerous edges connecting the nodes, leading to several possible traversals through the nodes that represent sequences not actually observed. This poses two problems: it becomes more difficult for an aligner to identify the correct path through the nodes, and it might provide an opportunity for a read with a sequencing error to have a better alignment to an incorrect path.

While the seed-and-extend strategy of read alignment has been optimized numerous times for mapping to the linear reference genome, read mapping to graph-based references is a more recent application and is inherently more complicated, particularly for reference graphs with topologies that are cyclic or that include high densities of branching variants. After a seed has been found in the reference, there could be more than one possible path through which the alignment can be extended. Aligners need to decide through which nodes to traverse, adding to the time and complexity of the alignment process. Graphs with fewer possible paths reduce this constraint. Since aligners allow for a threshold of mismatches when aligning reads to account for genetic variants, somatic mutations, and sequencing errors within the reads, an incorrect base match between the read and the reference will not necessarily cancel the alignment extension. A read with a sequencing error or somatic mutation might end up mapping to an incorrect location because the affected base happens to cause the read to map to a different site in the reference, particularly in repetitive regions. In reference graphs with

separate nodes for known single nucleotide variants, there are more opportunities for erroneous bases to better align to nodes outside the correct path. While this might not cause many problems if the read is generally in the right location, it can still reduce variant calling accuracy in lower coverage regions due to spurious alignments.

Spurious improvements in alignment would only be expected for a small number of reads with sequencing errors, and only when aligned to a graph topology that is highly connected with a high density of small nodes. I did not expect any spurious improvements for the simulated error-free reads I used to assess mapping to the different graph topologies; the fact that there were suggests an inability for the alignment software to find the optimal alignment when mapping reads to a self reference. This seems counterintuitive given that the self allele has a single possible path, but it could be a reflection of a sacrifice in accuracy by the software in favor of faster alignments. For example, `vg map` has a default setting for attempting to align up to 64 of the best seed chain candidates for paired-end reads; if an adequate alignment is not found after searching 64 series of seeds that identify potential alignment locations, alignment of the read is abandoned. In addition, while trying to rescue the mate of an aligned read by considering alignments within an expected fragment length distance, both `vg` and the linear reference aligner `bwa mem` try up to 64 or 50 times, respectively, before abandoning the mate and leaving it unmapped.

Spurious impairments were also not expected for the error-free simulations, nor would they be expected for reads with sequencing errors. If a read does not map perfectly to a set of nodes in a graph, it should not map better to a self reference if the self reference is fully represented within the graph. The observation of reads with spurious impairments in alignment in my assessment suggests the aligners are not always able to find the optimal alignment in a graph. This is logical given the complexity and numerous possible paths of many reference graphs and has been observed before (Garrison et al. 2018). In theory, embedding the paths for known variants within the graph structure would provide aligners with useful information about the most likely paths to explore that lead to optimal alignments. Interestingly, while `vg` utilizes path information from graph references when performing alignments, `minigraph` and `GraphAligner` do not.

The graph-alignment software used to map reads also has an effect on alignment accuracy. `minigraph` does not actually perform base-level alignments, but rather provides the traversed path as an alignment metric (Li et al. 2020). This can lead to confusion of where to align a read if there are similar paths. As well, since `minigraph` was built for assessing large structural variants as opposed to small indels and SNVs, it has trouble aligning to small variants since the aligner cannot seed the initial alignments. It is very likely that the reduced ability of `minigraph` to align the simulated reads in my analyses accurately is due to the close similarity of numerous *PRDM9* alleles. In addition, since the alignments are not at the base level, the divergence values provided after alignment are estimations (Li 2018). `GraphAligner`, on the other hand, was developed for the alignment of long reads. When finding seeds, only those that match a sequence entirely within a single node are explored, meaning seeds that span multiple nodes are missed (Rautiainen and Marschall 2020). Since long reads usually map in part to a linear section of a graph, at least one seed is usually able to match within a single node and allow for alignment of the rest of the read. For short reads, however, it is more difficult to find seeds in a single node when so many of the nodes are small and highly connected.

Since spuriously aligned reads cannot be identified and removed from real data when the genotype is unknown, using a graph topology and aligner that result in a very low percentage of spuriously affected reads is ideal. The allele-stacked graph resulted in the best alignments overall, improving read divergence over aligning to GRCh38 and resulting in the lowest proportions of unmapped or spuriously aligned reads. These results were somewhat surprising due to the low graph compression and the second-highest average degree across the nodes, but with only 36 possible traversals through the graph, the topology seems to have prevented misalignments better than the higher-density topologies, likely due to the preservation of allele haplotypes. When using the allele-stacked graph and `vg` to perform read alignment on the HPRC++ samples, higher averages in improvement were observed for samples with Admixed American, African, or East Asian ancestry relative to European samples. Africans are known to have a more diverse range of *PRDM9* alleles than Europeans and have a higher frequency of longer alleles such as allele C, but little research has previously been done on *PRDM9* genotypes in Admixed American or Asian populations. The improvements observed for the

Admixed American population were on par with those for the African individuals, providing interesting insight into the utility of *PRDM9* reference graphs for this population. Of the four European samples, three had the most common genotype A/A (P001/P001), but all had an average improvement in divergence. Allele A has a one SNV difference from allele B, the allele represented in GRCh38, so it is not surprising to see a small increase for these samples. The higher median difference in divergence for samples with a novel allele in their genotype suggests the novel alleles are more differentiated from the reference allele B than the *PRDM9*-106 alleles found in the HPRC++ samples.

The *PRDM9* graph topologies tested were constructed to represent all *PRDM9*-36 allele sequences. In theory this is the most desirable outcome, but in practice aligning to all of the graphs resulted in spurious read alignments, some likely because of graph complexity. It is possible to prune graphs to remove nodes in areas with a high density of variants, and doing so might further improve read alignments. It would be interesting to test additional topologies, including ones built from the *PRDM9*-106 list of alleles. As well, it would be interesting to use a *PRDM9* graph as the initial reference for read alignment to see how many reads that were unmapped or misaligned to GRCh38 had better alignments to the graph-based reference.

While I only assessed one genomic region, the results for *PRDM9* read mapping to reference graphs can likely be generalized to other problematic regions of the genome. Repeat sequences can be represented with cycles in a graph, where an edge leaving the end of a node loops back to the start of that node or another. If the repeats are fairly short, such as for short tandem repeats or minisatellite repeats, the number of possible paths through the cyclic region can become very extensive. Additionally, since some graph aligners cannot handle cyclic graphs, alternative acyclic topologies would need to be considered. Likewise, regions with high densities of SNPs will likely need to be pruned to reduce the number of possible paths. Alternatively, a structure that to some degree retains haplotype information, similar to the allele-stacked topology, could be beneficial.

4.4 Methods

4.4.1 Constructing *PRDM9* reference graphs

Five different structures or topologies of reference graphs were constructed to represent the *PRDM9*-36 alleles using either simple bash commands or the software `vg`, the variant graph toolkit (Garrison et al. 2018). Each graph was generated as a gfa file and then adjusted and indexed with `vg`. Adjustments included normalizing the graph, sorting the node IDs, and chopping node lengths to a maximum of 1024bp (required for indexing), using the command `vg mod -nc -X 1024`. Both `xg` and `gcsa` indexes were generated for the graphs using `vg index`. Two versions of each graph were generated, one with path information embedded in the graph structure and one without. Values for the number of nodes and edges, the edge:node ratio, the highest and average degree, the total length, and whether the graph is cyclic were determined with `vg stats -zlAD` and the default `gfastats` (Formenti et al. 2022) output.

The **allele-msa** graph used `vg msga -b B` to generate a multiple sequence alignment from a fasta file containing all of the *PRDM9*-36 allele sequences including the 10kb flanks. Allele B was the base sequence and identical sequences in the multiple sequence alignment were collapsed into single nodes to generate a structure similar to the POA graph (Lee et al. 2002).

The **allele-msa-acyclic** graph was an acyclic modification of the allele-msa graph. Along with the parameters used above, `-w 100000 -J 100` were used to adjust the band width and band jump parameters to force the graph to be acyclic.

The **allele-stacked** graph was generated in bash by making a node for each allele zinc finger array sequence. Nodes were also created for the left and right 10kb flanking sequences. Edges were specified to connect the left flank node to each of the allele nodes, and to connect each of the allele nodes to the right flank node.

The **zinc-finger-msa** graph was constructed from a multiple sequence alignment from a fasta file containing the sequences of each individual zinc finger repeat observed in the *PRDM9*-36

alleles, using `vg msga -b b` to initiate the alignment with the 'b' zinc finger. The resulting POA graph was then adjusted to add an edge from the last node to the first node to allow for consecutive traversals of individual zinc finger paths. Nodes for the left and right flanks were added and connected with edges to the first and from the last zinc finger nodes, respectively.

The **zinc-finger-loop** graph was built by giving each zinc finger sequence its own node and then adding edges between the nodes if the pair of zinc fingers occurs consecutively in any known *PRDM9*-36 allele sequence. Nodes for the left and right flanks were then added, along with an edge from the left flank node to the 'a' zinc finger node, and from the 'i' and 'j' zinc finger nodes to the right flank node.

For testing purposes, "flat" linear graphs for each *PRDM9*-36 allele were also generated by supplying the allele + 10kb flank sequence as the reference sequence to `vg construct`, creating **self references**. These graphs consisted of multiple nodes after adjusting the node size to be no more than 1,024bp, but each had only one possible path for traversal. The flat graph for allele B functioned as the graph representation of GRCh38.

4.4.2 Comparing reference graph topologies

4.4.2.1 Assessing short path accuracy

Graphs were modified to remove the first and last nodes containing a combined 20kb of flanking sequence. For *k*-mer lengths from one to 99 in increments of two, lists of *k*-mers were gathered from each *PRDM9*-36 allele sequence (variable zinc finger domain only), from the five *PRDM9* graph topologies, and from GRCh38 using `vg kmers`. The list of *k*-mers for each topology and for GRCh38 was compared to the union of *k*-mers from all alleles: ***k*-mer precision** was calculated as the number of graph *k*-mers that were also in the list of allele *k*-mers, and ***k*-mer recall** was calculated as the number of allele *k*-mers that were also in the list of graph *k*-mers. Across all values of *k*, a **pseudo-F1 score** was calculated for each graph and for GRCh38:

$$Pseudo\text{-}F1 = 2 \cdot \frac{kmer\ precision\ \cdot\ kmer\ recall}{kmer\ precision\ +\ kmer\ recall}$$

4.4.2.2 Assessing read alignment to the graph-based references

The 100X reads from the primary haploid simulation set (**Methods 2.4.2**) were used to assess read alignment to each graph and to the linear reference, comparing three different graph aligners: `vg map`, `minigraph` (Li et al. 2020); and `GraphAligner` (Rautiainen and Marschall 2020). Alignments with `vg` were performed in three ways: 1) without any optional parameters (not specified), 2) with the `-I 5000:250:50:0:1` parameter to provide the expected fragment length and standard deviation while letting `vg` adjust them (variable), and 3) with parameters `-U -I 5000:250:50:0:1` to specify the expected fragment statistics without letting `vg` adjust them (fixed). The values correspond to a maximum fragment length of 5000, an expected average length of 250, an expected standard deviation of 50, flip orientation, and forward direction. Alignments with `minigraph` used the parameter `-x sr` to specify short reads. The **divergence** tag in the resulting **graphical alignment format** (**gaf**) files was used to quantify how well the alignments matched the graph sequences, where divergence is the proportion of bases in the read that did not match the reference.

All reads were aligned to each of the five graph topologies, to the GRCh38 flat graph, and to each *PRDM9*-36 allele self reference. For each graph, the divergence value was subtracted from the divergence value for the read aligned to GRCh38 to get a value for the **difference in divergence**. Considering only reads that aligned to both the graph and to GRCh38 and that had a non-zero difference in divergence, the differences were then averaged for each allele and each graph, along with the proportion of reads that did not align to the graph. Finally, the differences were averaged across all alleles for each graph. Positive values indicated an **improvement** in read alignment when mapped to the graph compared to GRCh38, and negative values indicated an **impairment** in alignment.

Differences in divergence values were also calculated between the allele-specific self references and the graphs. **Spurious improvements** were identified as reads that had a positive difference in divergence, while **spurious impairments** were identified as reads that had a negative difference in divergence. Finally, **true improvements** in alignment were calculated between the graph and GRCh38 alignments after removing reads with either type of spurious

alignment. The true improvement assessment was repeated using only reads that aligned to or within 100bp of the zinc finger repeat array.

4.4.3 Assessing graph-based alignment for the HPRC++ samples

For each HPRC++ sample (see **Chapter 3 Methods 3.4.1**), the subset bam file of reads aligned to the *PRDM9* + 10kb region were realigned to the allele-stacked graph using `vg map`. The average length and standard deviation of read fragments were calculated for each sample (**Chapter 3 Methods 3.4.3.5**) and used as a fixed parameter during alignment. Each sample was also realigned to the GRCh38 flat graph and differences in divergence were calculated for each read in the same manner as the simulated data. The differences were then averaged across all reads that had a difference relative to GRCh38. The analysis was performed again on a further subset of reads whereby both mates mapped to or within 100bp of the zinc finger repeat array.

Chapter 5

Summary and future directions

5.1 Summary of major contributions

The objective of this thesis was to investigate and develop methodologies to improve allele calling in highly polymorphic and repetitive regions of the genome, using *PRDM9* as a proof-of-concept region. I developed short-read genotyping models to improve the information obtained from Illumina sequencing, which has historically been limited in repetitive regions due to non-unique mapping of reads. To validate my short-read models, I also developed long-read genotyping models to speed up and improve allele calling from PacBio HiFi reads. I also assessed the effect of non-linear reference genomes on aligning short reads by exploring different topologies of graph-based reference structures. I have collated information on different *PRDM9* alleles and the zinc finger repeats found within the variable array, and I have provided this *PRDM9* variant database and the software for my genotyping models as open-source resources for further research and development.

5.1.1 Short-read genotyping models

In **Chapter 2** I developed two models, count and distance, that utilize *k*-mer information from short sequencing reads to genotype *PRDM9* alleles. The models are intended to make use of reads mapped to a region of interest without relying on precise base-to-base alignments. Focusing on *k*-mers within the reads instead of on sequence alignment reduces mappability and bias issues traditionally associated with short reads because *k*-mer copy numbers are used to determine the genotypes.

The **count model** compares *k*-mer count profiles between the sequencing reads of a sample and each known allele. It uses a Poisson distribution to model the number of times a *k*-mer is observed in the set of sequencing reads, assuming a fixed error rate. The model allows for the calculation of the allele pair that maximizes the likelihood across all *k*-mers expected from the set of known allele sequences, which is called as the genotype. Three different methods were used to estimate the shaping parameter λ, which corresponds to *k*-mer coverage: 1) a

calculation using depth of coverage, sequencing error, read length, and *k*-mer length; 2) the average observed count of *k*-mers in the flanking sequences outside the region of interest; and 3) the median observed count of *k*-mers in the flanking sequences outside the region of interest. The count model worked very well for calling haploid alleles from simulated reads but was unable to fully distinguish diploid genotypes because some genotypes had identical *k*-mer count profiles.

The **distance model** uses the outermost *k*-mers in a pair of sequencing reads to estimate sequencing fragment lengths. Assuming a normal distribution of fragment lengths, the outermost *k*-mers from each read pair are located in the sequences of a known allele, and the distance between the *k*-mer pairs in the allele sequence is used to determine the likelihood of the reads originating from that allele. The insertion or deletion of a zinc finger repeat in the allele being tested shifts the observed fragment length away from the expected fragment length. Four different ways of handling multiple possible fragment lengths in the alleles were tested: 1) using the sum of probabilities for all fragment lengths, 2) using the average probability for all fragment lengths, 3) using the geometric mean of the probabilities for all fragment lengths, or 4) considering only the fragment length that gives the highest probability. The distance model was not as successful as the count model in genotyping haploid alleles for low to medium-length *k*-mers, but outperformed the count model when *k* neared the length of the reads. For calling diploid genotypes, the distance model generally outperformed the count model for error-free simulations, but the lack of accounting for sequencing errors during likelihood calculations meant the distance model underperformed when reads were simulated with errors.

I also assessed the performance of **combining** the count and distance models, with the hypothesis that using both could amplify the strengths of each model. The combination was generally as good as or better than the count model alone for both haploid and diploid genotype calling.

When genotyping real sequencing data, it was found that correcting the reads for sequencing errors generally improved calling results. Among the HPRC++ and OHS samples, approximately 43% were called correctly by at least one short-read model, and 69% had the

correct genotype called in the top 10 likelihoods by at least one model. While the count-coverage and distance-max models were frequently the best at genotyping real samples, they were on occasion outperformed by the count-flank median or distance-mean models. Providing multiple approaches to calculate statistics during genotyping will allow for further development of the models as parameters are fine tuned.

5.1.2 Long-read genotyping models

In **Chapter 3** I developed two models for genotyping *PRDM9* that use long-read data to speed up and improve over manual approaches (e.g. genotyping by eye in `IGV`). The **realignment** model aligns each long read to each known allele sequence and uses the edit distance between the two sequences to calculate genotype likelihoods. The **consensus** model generates a POA graph using all of the sequencing reads, then from the graph determines the one or two consensus sequences that best summarize the alignment. The consensus model is also able to flag novel alleles observed in a sample by comparing the consensus sequences to all known allele sequences and identifying the allele with the lowest edit distance to the consensus if there is no exact match. While both models called all 52 HPRC++ samples correctly, the consensus model greatly outperformed the realignment model by correctly calling 47 of the 49 OHS samples in comparison to 24 samples called correctly by the latter model. A total of five novel alleles (and one potential novel allele conservatively not named due to issues with PCR laddering during targeted PacBio HiFi library preparation) were identified by the consensus model and verified in `IGV`. Three of these novel alleles also had novel zinc finger repeat sequences, while the other two had novel arrangements of known zinc finger repeats. Additionally, two samples had a blood/sperm allele as part of their genotype. The consensus model is a straight-forward approach that can accurately determine genotypes from long sequencing reads and identify novel alleles in a single step.

5.1.3 Graph-based reference genomes

In **Chapter 4** I explored how different topologies affected read alignment to graph-based references. I generated five different *PRDM9* graphs that all contained variants for the *PRDM9*-36 list of alleles that varied in terms of node length, average node degree, and

cyclicity. I looked at short-path accuracy, which assessed how well the graphs represented the *k*-mers from alleles and how many additional sequences they spelled out from all possible traversals through the nodes. Two topologies had perfect *k*-mer recall and *k*-mer precision, meaning they had no additional sequences represented through path traversals and represented all known allele *k*-mers. The other three had perfect *k*-mer recall but reduced *k*-mer precision due to the numerous possible path traversals.

I then assessed how well simulated short reads aligned to the graphs in comparison to GRCh38. Mapping simulated reads to the graphs resulted in some spurious read alignments: spurious improvements occurred when reads aligned better to the graph than to the allele sequence from which the reads were simulated, and spurious impairments occurred when reads aligned worse to the graph than the originating allele sequence. These spurious alignments were observed even with error-free simulated reads, suggesting that there are limitations in how well the graph alignment software can find optimal mappings.

The allele-stacked graph, which consisted of separate nodes for the zinc finger region of each allele attached to the shared left and right flanking nodes, had perfect *k*-mer precision and recall, and resulted in the best alignments for simulated reads, taking into consideration both the proportion of reads with spurious alignments and the proportion of reads that did not map to the graph. The overall improvement in alignment was fairly small, but there was a noticeable increase in improvement for African and Admixed American HPRC++ samples compared to European samples. Additionally, samples that had novel allele genotypes as identified by the realignment model in **Chapter 3** had greater improvements in alignment when using the allele-stacked graph over GRCh38.

5.1.4 Open-source genotyping software: `gbkc` and `gblr`

Both the short- and long-read genotyping models have been written as open-source software and are available on github for use or modification by other researchers. The short-read models are developed as C++ software called **`gbkc`** for **genotyping by *k*-mer counts** (https://github.com/hgibling/gbkc). `gbkc` contains four modules: `count`, `distance`, `check-profiles`, and `get-genotypes`.

The `count` module is used to run the count models and requires:

1) a fasta file of alleles for the region of interest
2) one or two fasta or fastq files containing the sequencing reads to be analyzed
3) a fasta file containing the two flanking sequences
4) the values of *k* to be tested
5) the read length
6) the estimated sequencing error rate
7) the average sequencing depth of coverage
8) the method to use for calculating likelihoods
9) whether scoring is haploid or diploid.

Optional parameters available are:

1) the penalty likelihood to assign when a read *k*-mer is not observed in the allele
2) a manual value for λ
3) how many of the top scores to print to file
4) the name of the output file
5) the number of threads to use.

The `distance` module is used to run the distance models. It has the same required parameters as the `count` module except it does not require a fasta file for the flanking sequences, and the same optional parameters for output file name, number of top scores to print, and the number of threads to use. It also has required parameters for:

1) the mean fragment length
2) the fragment length standard deviation
3) the fragment length to use when a *k*-mer pair is not observed in an allele.

In order to combine the likelihoods from the count and distance models, likelihoods need to be printed for all alleles or genotypes and manually added together, since they are in log space. This is not currently implemented within `gbkc`, but writing a function or module to combine the likelihoods would be straightforward.

The `check-profiles` module is used to generate and compare *k*-mer count profiles. It requires a fasta file of alleles of interest and the range of *k* values to test, and has options to generate diploid count profiles, print out the difference between count profiles if they are not identical, or print the individual count profiles. Finally, the `get-genotypes` module is used for generating all possible genotypes from a list of alleles provided as a fasta file.

The long-read genotyping models have been developed as Python software called **`gblr`** for **genotyping by long reads** (https://github.com/hgibling/gblr). `gblr` contains two scripts: `gap-filter.py` and `gblr.py`.

The `gap-filter.py` script is used to filter bam files to determine which reads fully span the region of interest and do not have indel artifacts. It requires:

1) a bam file of reads to be analyzed
2) the genomic coordinates for the region of interest
3) the flank tolerance or how many bases past the region of interest to which the read should align
4) the gap tolerance or the maximum size of ignored indels
5) the read threshold or the proportion of reads that must have a specific indel to not be considered an artifact
6) a length multiple or expected value for the indel size whereby indels greater than the gap threshold but not a multiple of the length multiple will be ignored.

An optional parameter is a length tolerance which allows for indels to be within a certain number of bases of the length multiple. The script outputs the names of the reads to be kept for downstream analyses, which can be supplied to `samtools view -N` to subset the bam file.

The `gblr.py` script is used to run the realignment and consensus models. It requires:

1) a fasta file of the alleles or haplotypes of interest (including flanking sequences)
2) a bam file of the sequencing reads to be analyzed
3) genomic coordinates for the region of interest
4) the length of the flanking sequences

5) the flank tolerance or how many bases past the region of interest to which the read should align
6) an estimate of the sequencing error rate
7) the scoring model to use
8) the edit distance model to use.

Optional parameters are:

1) the number of scores to print to file
2) print the allele edit distance table to `stderr`
3) print the consensus sequence(s) to file.

There is also a quick count mode which simply counts the number of reads that best align to each allele. The `gblr.py` script also filters for reads that span the entire length of the region of interest, so the `gap-filter.py` script is only necessary for removing noisy long reads such as those with PCR laddering artifacts as observed in **Chapter 3**. The additional filtering step of retaining only reads that are the same length as the majority of the reads (local maxima filtering) was performed separately in R, but would be easy to implement within the `gap-filter.py` script.

5.1.5 The *PRDM9* standardized nomenclature and variant database

The initial *PRDM9* variants described in early literature were given names relative to the population in which they were found: AA1 to AA11 in African Americans, CH1 to CH3 for Han Chinese, and M1 and M2 for the alleles found in multiple populations (Parvanov et al. 2010). Variants were then renamed or given single-letter names from A to F, H, I, and K. The current standard for naming has been assigning a number after the letter L, with L49 being the most recent. In 2021, Alleva et al. named novel alleles they identified as M1-M32. Alleles M4, M5, and M10, however, had zinc finger arrangements already identified and named in published literature as L25, L32, and L26, respectively. In addition, M20 and M21 had zinc finger arrangements already deposited in the NCBI Nucleotide database in 2016, as X5

(KU721986.1) and X14 (KU886573.1), respectively. The names M1 and M2 were also used by Parvanov et al. (2010) to describe alleles A and B, respectively. Similarly, allele L37 described by Hussin et al. (2013) has the same zinc finger arrangement as L26, which was described by Berg et al. (2011). There have also been typos in the literature: an extra 'h' zinc finger in allele L24 (Ponting 2011), and a 'c' zinc finger instead of a 'd' in allele I (Borel et al. 2012). Incorrect sequences were also deposited to NCBI Nucleotide database deposits: JQ044376.1 has sequence corresponding to a 'w' zinc finger instead of an 'x' for allele L35, and JQ044372.1 has sequence corresponding to an 'i' zinc finger instead of a 'q' for allele L38 (Hussin et al. 2013). Alleva et al. (2021) also describe how 13 alleles previously only seen in sperm or as a somatic blood mutation were observed in their samples, but six had already been described and named in published literature or the NCBI Nucleotide database: Av:0128 as L31, Av:0024 as L33, Av:0001 as L39, Av:0017 as L44, Av:0003 as X13 (KU886572.1), and Cv:0212 as X8 (KU721989.1).

The names and sequences for the individual zinc fingers found in *PRDM9* are similarly confusing. Originally named 'A' to 'T' in Berg et al. (2010), Ponting (2011) used lowercase letters to distinguish zinc finger names from allele names. Hussin et al. (2013) described four new zinc finger sequences 'u' to 'x', but Berg et al. (2010) had already used 'U' and 'V' to describe zinc fingers with different sequences. Jeffreys et al. (2013) used lowercase letters, single digit numbers, and 10 symbols ('!', '@', '£', '$', '%', '&', '§', '*', ':', and '±') to provide names for the numerous zinc fingers observed in somatic mutations and sperm cells. Alleva et al. (2021) used '!' before capital letters to name their novel zinc fingers and added ':' before the Berg et al. (2010) zinc finger names. For the Jeffreys et al. (2013) names, Alleva et al. (2021) changed the symbols to capital letters and added '|' in front.

The lack of a central repository of *PRDM9* variants has clearly made it difficult for researchers to keep track of which names have been used and which arrangement of zinc finger repeats have already been described. While having short names for alleles and zinc fingers is desirable for computational parsing, it is arguably more important to remove inconsistencies and duplicated names. I propose a nomenclature for providing **standardized names** to current alleles and zinc finger repeats with a system that is easy to expand as new variants are discovered. Allele names follow the pattern 'P###' and zinc fingers follow the pattern 'Z###';

allele A and zinc finger ‘a’ have the standardized names P001 and Z001, respectively. Alphabetical order of current names was used to assign standardized names, using publication dates as tie breakers when the same name was provided to different sequences. Variants found only in the NCBI Nucleotide database were named last. If more than 999 variants are eventually described, an additional leading zero can simply be added to the standardized names to maintain a consistent nomenclature.

The *PRDM9*-106 list of variants observed as part of diploid genotypes is not composed of consecutively named variants, but rather it starts with P001 and ends with P642. While generating two separate naming systems was initially considered to differentiate the germline variants observed in the human population and the blood/sperm alleles, the finding that some variants originally only observed in sperm or as somatic mutations in blood were recently observed as part of a diploid germline genotype (Alleva et al. 2021) suggests it would be confusing to rename such variants if this were to occur again. I therefore provide two lists: one of only variants observed as part of diploid genotypes (*PRDM9*-106) and one with the addition of the blood/sperm alleles (*PRDM9*-642). The lists, comparison of names and sequences, and code used to obtain the lists are available online (https://github.com/hgibling/PRDM9-Variants).

5.2 Future directions

The work presented in this thesis provides a starting point for further developing these genotyping models. Additional features can be assessed in order to improve genotyping accuracy. Other polymorphic regions of the genome that are tricky to genotype with short-read sequencing can also be tested. In addition, more can be learned about *PRDM9* variants, particularly in understudied populations, and in terms of aberrant expression in cancers.

5.2.1 Further exploration of *PRDM9* alleles

Non-European populations are chronically understudied in genetics research (Popejoy and Fullerton 2016; Claw et al. 2018; Sirugo et al. 2019). Interestingly, African samples have been

included in early *PRDM9* research and recombination studies due to the limited *PRDM9* variability in Europeans (Berg et al. 2010). There is therefore a decent understanding of *PRDM9* variants and allele frequencies in people with African ancestry. Specific European populations that have been studied are Hutterites, a founder population in North America (Baudat et al. 2010); Icelanders (Kong et al. 2010); and French-Canadians (Hussin et al. 2013). Non-European and non-African populations have traditionally only been studied in small sample sizes: 16 Han Chinese and 16 Mexican Americans (Parvanov et al. 2010); 74 Han Chinese or Japanese in Tokyo individuals (Kong et al. 2010); 33 British Indians of primarily Gujarati descent (Berg et al. 2011); 27 Moroccan individuals; and 11 individuals of various Hispanic, Asian, and Native American ancestry (Hussin et al. 2013). A recent study involved 1,030 Chinese individuals (Wang et al. 2021).

The most comprehensive study to date looked at 720 samples across seven populations: 120 Yoruba in Ibadan, Nigeria; 120 Luhya in Webuye, Kenya; 114 Toscani in Italy; 100 Finnish in Finland; 70 Peruvian in Lima, Peru; 120 Han Chinese in Beijing, China; and 108 Punjabi in Lahore, Pakistan (Alleva et al. 2021). Several novel alleles were identified in this study, along with interesting frequencies regarding specific alleles. The Han Chinese population had the third highest level of heterozygous *PRDM9* genotypes despite having a low number of allele variants, likely due to the enrichment of allele B (P002) compared to other populations. Allele D (P004) was similarly enriched in Finnish samples. Previous studies generally assumed that allele C (P003) was an African-specific variant, but Alleva et al. show it is actually present in many populations while being depleted in Europeans.

Due to the low frequency of many *PRDM9* variants, novel alleles and interesting population-specific frequencies will only be uncovered with large sample sizes from diverse populations. This is evident from the study by Alleva et al. (2021), but was also seen in Baudat et al. (2010), where allele I (P008) was only observed in Hutterites. There are still additional populations lacking *PRDM9* statistics in South America, Africa, the Middle East, and Indigenous North America and Australia, to name a few. The HPRC plans on sequencing 350 individuals in total from diverse populations, meaning additional high-quality whole-genome PacBio HiFi data will soon be available for accurate *PRDM9* genotyping. A meta-analysis examining *PRDM9*

genotypes from all available populations should be performed to obtain updated allelic frequencies.

5.2.2 Genotyping tumor samples with aberrant *PRDM9* expression

Given that the PRDM9 protein functions in signaling sites for homologous recombination during meiosis, gene expression is not expected outside of testes and fetal ovaries (Sun et al. 2015). Gene expression of *PRDM9* has, however, been observed in different cancer cell lines and over 30 different types of cancerous tissue (Feichtinger et al. 2012; Ang Houle et al. 2018), though the specific alleles being expressed were not identified. There is an enrichment of structural variant breakpoints near sites of PRDM9 binding, suggesting that aberrant expression of the protein in somatic cells may lead to genomic instability (Ang Houle et al. 2018). Since alleles A (P001) and C (P003) bind to different DNA motifs (Hinch et al. 2011; Alleva et al. 2021), potential implications of aberrant chromosomal breaks in cancer tissue could be variant dependent. Determining the *PRDM9* genotypes for cancers with aberrant expression could therefore show interesting associations with certain alleles. Running the short-read genotyping models on the PCAWG RNA-sequencing data used by Ang Houle et al. (2018) would be an interesting avenue for exploration into the role of specific *PRDM9* variants in cancer development and progression.

5.2.3 Genotyping additional polymorphic regions of the genome

PRDM9 is but one genomic region with high polymorphism that is of interest to researchers. Regions that could greatly benefit from improved short- or long-read genotyping, and from read mapping to graph-based reference genomes, include those with highly similar repeats or a high density of SNPs. Additionally, there should be some curation of well-known alleles or haplotypes to provide a basis on which novel variants can be identified. The 4Mb **MHC** on chromosome 6 is the most polymorphic locus in the human genome (Horton et al. 2008) and contains over 150 protein-coding genes, including the **HLA** genes which encode proteins that present foreign antigens to immune cells. Balancing selection acting on HLA genes due to variation among pathogen genomes has resulted in over 34,000 alleles across the 21 HLA genes described to date (https://www.ebi.ac.uk/ipd/imgt/hla/about/statistics), allowing for

effective immune responses to pathogens. Along with foreign antigen presentation to trigger immune responses, MHC proteins are known to present cancer-specific peptides to T cells as well, leading to a targeted destruction of tumor cells (Schumacher and Schreiber 2015). There is great interest in developing personalized immunotherapies for cancer treatments, but the high level of polymorphism makes mapping short sequencing reads to the MHC locus very difficult.

The **KIR** genes are another group of immune system genes and are involved in recognition of self cells by natural killer (NK) cells (Barrow and Trowsdale 2008). Located on chromosome 19, the region is variable regarding the gene variants present in a haplotype (Norman et al. 2016), as well as in the number of genes and gene copy numbers due to the non-allelic homologous recombination that occurs in the region (Wright 2020). The region is considered the second-most polymorphic after the MHC region: individual genes have between 34 and 228 alleles, and a total of 1,535 alleles across all genes have been described to date (https://www.ebi.ac.uk/ipd/kir/about/statistics). While some cancer cells downregulate expression of tumor antigens as a means to evade the HLA immune response, NK cells can recognize and kill such cells (Middleton and Gonzelez 2010), making the KIR genes of interest to cancer researchers. KIR alleles also have an influence on hematopoietic stem cell transplants, along with HLA genes. While matches between donor and recipient HLA variants are desired, it is thought that having KIR variants that activate NK cells is actually beneficial for patients (Wright 2020). Genotyping KIR variants with better accuracy could lead to improvements in transplant outcomes for leukemia patients.

In addition to immune response genes, there is great interest in better understanding the **CYP** genes involved in drug metabolism (Hoffman et al. 2001). CYP enzymes metabolize around 75% of known drugs (Zhao et al. 2021). The over 350 polymorphisms across all 57 CYP genes affect the metabolic efficiency of a given drug across patients and has led to the classifications of poor, intermediate, extensive, and ultra-rapid metabolizers (Zhao et al. 2021). Given a standard dosage, poor metabolizers may experience toxic effects while ultra-rapid metabolizers may not experience any benefit at all. Identifying and understanding the polymorphisms associated with CYP genes is therefore extremely important for determining optimal therapeutic treatments for many conditions (Zhou et al. 2009).

5.2.4 Modifications to the genotyping models

As discussed in **Chapter 2**, there are modifications to the genotyping models that can be made to hopefully improve accuracy. The distance model currently does not account for sequencing errors, and only one pair of *k*-mers is assessed from each read pair. Increasing the number of *k*-mers explored and modeling the error rate during the likelihood calculations could be avenues to explore. Aligning the read pairs to allele sequences while using an HMM to handle multiple possible fragment sizes could also be tested instead using *k*-mers. A different approach for modeling sequencing error rates could be used for the count method, whereby a multinomial model would account for the observation of *k*-mers with a hamming distance of one relative to the *k*-mers observed in allele count profiles. One assumption made by the count model in order to utilize the Poisson distribution is that *k*-mers are occurring independently, which is not the case. Any given *k*-mer is dependent on its neighboring *k*-mers, and this is particularly important when estimating sequencing error rates from low-frequency *k*-mers. For example, regions of the genome that are difficult to sequence with short-read technology have reduced read alignment due to mapping ambiguity. *k*-mers from regions with low sequencing coverage may be mistaken as erroneous, but considering all *k*-mers within the context of individual reads could allow for improved or dynamic error rate estimates.

As more *PRDM9* allele variants are uncovered, the number of possible genotypes increases nearly exponentially. Other regions of interest mentioned in section **5.2.3** also have large numbers of known variants, and generating simulated data to assess the genotyping models for these regions might not be feasible given the number of genotypes that need to be assessed. One possible modification would be to call **functional classes** of variants instead of precise genotypes. For example, CYP alleles can be labeled as having poor, intermediate, extensive, or ultra-rapid metabolism. *k*-mer characteristics of each class could be determined, allowing for a classification of metabolic function for a given sample, such as poor/poor or intermediate/ultra-rapid. This could provide actionable information regarding a genotype without needing to identify specific alleles amongst a large number of candidates.

5.2.5 Additional exploration using graph-based references

This thesis only assessed three graph-specific alignment software applications: `vg`, `GraphAligner`, and `minigraph`. As is common with bioinformatics software, updates to these tools have since been made and can be assessed for further improvements in read alignment for *PRDM9* variants. The developers of `vg` have released two additional mapping algorithms alongside the all-purpose `map` algorithm: `giraffe` (Sirén et al. 2021) and `mpmap` (Sibbesen et al. 2022). `giraffe` is a short-read mapper that is haplotype aware, making use of the embedded paths to a greater degree than the original mapping algorithm. `mpmap` is a multi-path aligner that provides a collection of non-linear sub-graph alignments instead of linear paths through the graph. Though intended for use with RNA transcript alignments, `mpmap` can be used to align reads to a graph with structural variants due to the similarity of SVs and variable exon inclusion in RNA transcripts, which might translate well for the copy number differences of zinc finger repeats in *PRDM9*. While newer versions of `minigraph` perform base-level alignments (the version tested in **Chapter 4** did not), the software is still intended for alignment to graphs with large SVs as opposed to graphs with SNVs and indels, meaning the aligner is still unlikely to be suitable for *PRDM9* graphs. `GraphAligner` underperformed in my analyses because it was designed for long-read alignment. It would therefore be interesting to assess differences in PacBio HiFi read alignments to the *PRDM9* graphs compared to GRCh38, and to examine if any long reads that were unable to map to GRCh38 were rescued when mapped to the *PRDM9* graphs, as might be the case for reads from samples with highly divergent *PRDM9* alleles.

Additional tools may also be developed that perform sequence-to-graph alignment as the field continues to grow. Currently, several aligners use graphs as an intermediate step during alignment (Alser et al. 2021), and some tools use graphs to perform variant detection and genotyping (e.g. Letcher et al. 2021), but there are relatively few aligners specifically designed to align reads to a graph structure directly and provide an output of the alignments. `HISAT2` can be used for both DNA and RNA alignment and has been used successfully in HLA typing (Kim et al. 2019). `Graph Genome Pipeline` was developed by the Seven Bridges company and performs read alignment and variant calling as part of its workflow (Rakocevic

et al. 2019), outputting alignments to a bam file. Another tool is `PaSGAL` which performs alignments to sequence graphs much like `vg`, though it currently only works for acyclic graphs (Jain et al. 2019). `GraphChainer` is a long-read graph aligner that claims to be faster than `GraphAligner` (Ma et al. 2022), and it would be interesting to include it in assessments for mapping PacBio reads to the *PRDM9* graphs. Finally, aligners that map to de Bruijn graphs (e.g. Heydari et al. 2018) have been available for many years now due to the utility of constructing de Bruijn graphs during genome assembly. It would be interesting to explore the use of representing genome references as de Bruijn graphs instead of as sequence graphs as used in this thesis.

5.3 Recommendations for genotyping *PRDM9* and other polymorphic repetitive regions of the genome

While Sanger sequencing has long been the gold standard for generating highly accurate reads, the process is time consuming, tedious, limited in the length of sequences producible, and not well-suited to scaling up for large data sets. High-throughput short-read sequencing is cost effective and scales well, but lacks the ability to provide consistent sequencing coverage and mappability across repetitive genomic regions. Long-read sequencing is able to generate reads of great length, allowing for unique mapping across repetitive regions. As technology is continuously improved, high-throughput long-read sequencing is becoming more affordable and the once high error rates plaguing the technology have been drastically reduced. Long-read whole-genome or PCR-free targeted sequencing is therefore recommended when wanting to genotype repetitive polymorphic regions of the genome.

PacBio HiFi sequencing has a low sequencing error rate owing to the circular sequencing of the same read repeatedly, providing a consensus sequence for each fragment. However, as observed in the OHS samples in **Chapter 3**, targeted PacBio sequencing can result in substantial artifacts in the reads. The targeted sequencing used in this thesis first requires PCR amplification, where laddering of molecules can occur due to the repetitive nature of *PRDM9*. This results in the overamplification of fragments with missing or added repeats and thus

fragments of incorrect lengths. Whole-genome PacBio HiFi sequencing is far more likely to provide high-quality data as long as there is sufficient coverage, though at a higher cost if only one genomic region is to be analyzed.

ONT also provides long-read sequencing via direct-molecule sequencing of reads as they pass through nanopores. Historically the 10–15% error rate has necessitated the use of consensus polishing software (Wick et al. 2019; Dohm et al. 2020). With continuous improvements to the sequencing chemistry and base-calling software, however, ONT currently claims to have a raw read accuracy of 99%, bringing it almost on par with Illumina sequencing error rates (Sereika et al. 2022). If whole-genome sequencing is not desired, there are also methods for targeted ONT sequencing that do not require PCR amplification, though they do require a larger starting amount of DNA. **Adaptive sampling** is a method of enriching specific DNA sequences by an ONT platform. Individual pores are computationally controlled to eject DNA molecules being sequenced if they do not match predefined genomic regions, saving time and resources to focus on the region of interest (Martin et al. 2022). A different method called **nanopore Cas9-targeted sequencing** (nCATS**)** uses Cas9 to target genomic regions of interest prior to nanopore sequencing. The ends of DNA molecules are dephosphorylated prior to site-specific cuts by Cas9 and guide RNA, allowing for the preferential ligation of ONT sequencing adapters and thus region-specific sequencing (Gilpatrick et al. 2020). Additional methods are sure to be developed as interest grows in PCR-free targeted long-read sequencing, for both ONT and PacBio technologies.

5.4 Conclusions

PRDM9 remains a fascinating gene in terms of both functional biology and of the allelic variants discovered to date. The variable zinc finger repeat array is a showcase of SNVs, 84bp indels, and small-scale rearrangements of repeated elements, and serves as an excellent example of a genomic region prone to reduced mappability and increased reference bias due to short sequencing reads not spanning the full length of the array. Though recent developments in long-read technology offer better variant identification for repetitive genes like *PRDM9*,

short-read sequencing is still immensely popular and cost effective, and hundreds of thousands of samples in large cohorts have already been sequenced with this technology. The continued development of bioinformatics tools and approaches to improve variant identification such as the models I describe in this thesis will allow for better genotyping and read mapping for *PRDM9* and for many additional polymorphic repetitive regions of the genome.

References

Aganezov S, Yan SM, Soto DC, Kirsche M, Zarate S, Avdeyev P, Taylor DJ, Shafin K, Shumate A, Xiao C, et al. 2022. A complete reference genome improves analysis of human genetic variation. *Science* **376**: eabl3533.

Ahn S-M, Kim T-H, Lee S, Kim D, Ghang H, Kim D-S, Kim B-C, Kim S-Y, Kim W-Y, Kim C, et al. 2009. The first Korean genome sequence and analysis: Full genome sequencing for a socio-ethnic group. *Genome Res* **19**: 1622–1629.

Albers CA, Lunter G, MacArthur DG, McVean G, Ouwehand WH, Durbin R. 2011. Dindel: Accurate indel calls from short-read data. *Genome Res* **21**: 961–973.

Alleva B, Brick K, Pratto F, Huang M, Camerini-Otero RD. 2021. Cataloging Human PRDM9 Allelic Variation Using Long-Read Sequencing Reveals PRDM9 Population Specificity and Two Distinct Groupings of Related Alleles. *Front Cell Dev Biol* **9**.

Alonso-Blanco C, Andrade J, Becker C, Bemm F, Bergelson J, Borgwardt KM, Cao J, Chae E, Dezwaan TM, Ding W, et al. 2016. 1,135 Genomes Reveal the Global Pattern of Polymorphism in Arabidopsis thaliana. *Cell* **166**: 481–491.

Alser M, Rotman J, Deshpande D, Taraszka K, Shi H, Baykal PI, Yang HT, Xue V, Knyazev S, Singer BD, et al. 2021. Technology dictates algorithms: recent developments in read alignment. *Genome Biol* **22**: 249.

Amberger JS, Bocchini CA, Schiettecatte F, Scott AF, Hamosh A. 2015. OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders. *Nucleic Acids Res* **43**: D789–D798.

Ang Houle A, Gibling H, Lamaze FC, Edgington HA, Soave D, Fave M-J, Agbessi M, Bruat V, Stein LD, Awadalla P. 2018. Aberrant PRDM9 expression impacts the pan-cancer genomic landscape. *Genome Res* **28**: 1611–1620.

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.

Bai H, Guo X, Zhang D, Narisu N, Bu J, Jirimutu J, Liang F, Zhao X, Xing Y, Wang D, et al. 2014. The Genome of a Mongolian Individual Reveals the Genetic Imprints of Mongolians on Modern Human Populations. *Genome Biol Evol* **6**: 3122–3136.

Baker Z, Przeworski M, Sella G. 2022. Down the Penrose stairs: How selection for fewer recombination hotspots maintains their existence. bioRxiv 10.1101/2022.09.27.509707.

Balasubramanian S, Habegger L, Frankish A, MacArthur DG, Harte R, Tyler-Smith C, Harrow J, Gerstein M. 2011. Gene inactivation and its implications for annotation in the era of personal genomics. *Genes Dev* **25**: 1–10.

Ballouz S, Dobin A, Gillis JA. 2019. Is it time to change the reference genome? *Genome Biol* **20**: 159.

Barrow AD, Trowsdale J. 2008. The extended human leukocyte receptor complex: diverse ways of modulating immune responses. *Immunol Rev* **224**: 98–123.

Baudat F, Buard J, Grey C, Fledel-Alon A, Ober C, Przeworski M, Coop G, de Massy B. 2010. PRDM9 is a major determinant of meiotic recombination hotspots in humans and mice. *Science* **327**: 836–840.

Benjamini Y, Speed TP. 2012. Summarizing and correcting the GC content bias in high-throughput sequencing. *Nucleic Acids Res* **40**: e72–e72.

Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, Hall KP, Evers DJ, Barnes CL, Bignell HR, et al. 2008. Accurate Whole Human Genome Sequencing using Reversible Terminator Chemistry. *Nature* **456**: 53–59.

Berg IL, Neumann R, Lam K-WG, Sarbajna S, Odenthal-Hesse L, May CA, Jeffreys AJ. 2010. PRDM9 variation strongly influences recombination hot-spot activity and meiotic instability in humans. *Nat Genet* **42**: 859.

Berg IL, Neumann R, Sarbajna S, Odenthal-Hesse L, Butler NJ, Jeffreys AJ. 2011. Variants of the protein PRDM9 differentially regulate a set of human meiotic recombination hotspots highly active in African populations. *Proc Natl Acad Sci U S A* **108**: 12378–12383.

Beyter D, Ingimundardottir H, Oddsson A, Eggertsson HP, Bjornsson E, Jonsson H, Atlason BA, Kristmundsdottir S, Mehringer S, Hardarson MT, et al. 2021. Long-read sequencing of 3,622 Icelanders provides insight into the role of structural variants in human diseases and other traits. *Nat Genet* **53**: 779–786.

Boeva V, Popova T, Bleakley K, Chiche P, Cappo J, Schleiermacher G, Janoueix-Lerosey I, Delattre O, Barillot E. 2012. Control-FREEC: a tool for assessing copy number and allelic content using next-generation sequencing data. *Bioinformatics* **28**: 423–425.

Borel C, Cheung F, Stewart H, Koolen DA, Phillips C, Thomas NS, Jacobs PA, Eliez S, Sharp AJ. 2012. Evaluation of PRDM9 variation as a risk factor for recurrent genomic disorders and chromosomal non-disjunction. *Hum Genet* **131**: 1519–1524.

Brandt DYC, Aguiar VRC, Bitarello BD, Nunes K, Goudet J, Meyer D. 2015. Mapping Bias Overestimates Reference Allele Frequencies at the HLA Genes in the 1000 Project Phase I Data. *G3 GenesGenomesGenetics* **5**: 931–941.

Bray NL, Pimentel H, Melsted P, Pachter L. 2016. Near-optimal probabilistic RNA-seq quantification. *Nat Biotechnol* **34**: 525–527.

Breese MR, Liu Y. 2013. NGSUtils: a software suite for analyzing and manipulating next-generation sequencing datasets. *Bioinformatics* **29**: 494–496.

Brick K, Smagulova F, Khil P, Camerini-Otero RD, Petukhova GV. 2012. Genetic recombination is directed away from functional genomic elements in mice. *Nature* **485**: 642–645.

Briggs AW, Rios X, Chari R, Yang L, Zhang F, Mali P, Church GM. 2012. Iterative capped assembly: rapid and scalable synthesis of repeat-module DNA such as TAL effectors from individual monomers. *Nucleic Acids Res* **40**: e117.

Broad Institute. 2019. Picard toolkit. *Broad Inst GitHub Repos.* https://broadinstitute.github.io/picard.

Buermans HPJ, Vossen RHAM, Anvar SY, Allard WG, Guchelaar H-J, White SJ, den Dunnen JT, Swen JJ, van der Straaten T. 2017. Flexible and Scalable Full-Length CYP2D6 Long Amplicon PacBio Sequencing. *Hum Mutat* **38**: 310–316.

Cann HM, de Toma C, Cazes L, Legrand M-F, Morel V, Piouffre L, Bodmer J, Bodmer WF, Bonne-Tamir B, Cambon-Thomsen A, et al. 2002. A Human Genome Diversity Cell Line Panel. *Science* **296**: 261–262.

Cao J, Schneeberger K, Ossowski S, Günther T, Bender S, Fitz J, Koenig D, Lanz C, Stegle O, Lippert C, et al. 2011. Whole-genome sequencing of multiple Arabidopsis thaliana populations. *Nat Genet* **43**: 956–963.

Carvalho CMB, Lupski JR. 2016. Mechanisms underlying structural variant formation in genomic disorders. *Nat Rev Genet* **17**: 224–238.

Cezard T, Cunningham F, Hunt SE, Koylass B, Kumar N, Saunders G, Shen A, Silva AF, Tsukanov K, Venkataraman S, et al. 2022. The European Variation Archive: a FAIR resource of genomic variation for all species. *Nucleic Acids Res* **50**: D1216–D1220.

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.

Chaisson MJP, Wilson RK, Eichler EE. 2015. Genetic variation and the de novo assembly of human genomes. *Nat Rev Genet* **16**: 627–640.

Chen K, Wallis JW, McLellan MD, Larson DE, Kalicki JM, Pohl CS, McGrath SD, Wendl MC, Zhang Q, Locke DP, et al. 2009. BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. *Nat Methods* **6**: 677–681.

Chen X, Schulz-Trieglaff O, Shaw R, Barnes B, Schlesinger F, Källberg M, Cox AJ, Kruglyak S, Saunders CT. 2016. Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics* **32**: 1220–1222.

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.

Cho YS, Kim H, Kim H-M, Jho S, Jun J, Lee YJ, Chae KS, Kim CG, Kim S, Eriksson A, et al. 2016. An ethnically relevant consensus Korean reference genome is a step towards personal reference genomes. *Nat Commun* **7**: 13637.

Church DM. 2022. A next-generation human genome sequence. *Science* **376**: 34–35.

Church DM, Schneider VA, Steinberg KM, Schatz MC, Quinlan AR, Chin C-S, Kitts PA, Aken B, Marth GT, Hoffman MM, et al. 2015. Extending reference assembly models. *Genome Biol* **16**: 13.

Claw KG, Anderson MZ, Begay RL, Tsosie KS, Fox K, Garrison NA. 2018. A framework for enhancing ethical genomic research with Indigenous communities. *Nat Commun* **9**: 2957.

Compeau PEC, Pevzner PA, Tesler G. 2011. How to apply de Bruijn graphs to genome assembly. *Nat Biotechnol* **29**: 987–991.

Cornish-Bowden A. 1985. Nomenclature for incompletely specified bases in nucleic acid sequences: recommendations 1984. *Nucleic Acids Res* **13**: 3021–3030.

Cretu Stancu M, van Roosmalen MJ, Renkens I, Nieboer MM, Middelkamp S, de Ligt J, Pregno G, Giachino D, Mandrile G, Espejo Valle-Inclan J, et al. 2017. Mapping and phasing of structural variation in patient genomes using nanopore sequencing. *Nat Commun* **8**: 1326.

De Coster W, Weissensteiner MH, Sedlazeck FJ. 2021. Towards population-scale long-read sequencing. *Nat Rev Genet* **22**: 572–587.

Degner JF, Marioni JC, Pai AA, Pickrell JK, Nkadori E, Gilad Y, Pritchard JK. 2009. Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data. *Bioinformatics* **25**: 3207–3212.

DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M, et al. 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* **43**: 491–498.

Dewey FE, Chen R, Cordero SP, Ormond KE, Caleshu C, Karczewski KJ, Whirl-Carrillo M, Wheeler MT, Dudley JT, Byrnes JK, et al. 2011. Phased Whole-Genome Genetic Risk in a Family Quartet Using a Major Allele Reference Sequence. *PLoS Genet* **7**: e1002280.

Dilthey A, Cox C, Iqbal Z, Nelson MR, McVean G. 2015. Improved genome inference in the MHC using a population reference graph. *Nat Genet* **47**: 682–688.

Dittwald P, Gambin T, Szafranski P, Li J, Amato S, Divon MY, Rodríguez Rojas LX, Elton LE, Scott DA, Schaaf CP, et al. 2013. NAHR-mediated copy-number variants in a clinical population: Mechanistic insights into both genomic disorders and Mendelizing traits. *Genome Res* **23**: 1395–1409.

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**: 15–21.

Dohm JC, Peters P, Stralis-Pavese N, Himmelbauer H. 2020. Benchmarking of long-read correction methods. *NAR Genomics Bioinforma* **2**: lqaa037.

Dubinkina VB, Ischenko DS, Ulyantsev VI, Tyakht AV, Alexeev DG. 2016. Assessment of k-mer spectrum applicability for metagenomic dissimilarity analysis. *BMC Bioinformatics* **17**: 38.

Durbin, R., Eddy, S., Krogh, A., & Mitchison, G. 1998. Pairwise alignment using HMMs. In *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*, pp. 81–100, Cambridge University Press, Cambridge doi:10.1017/CBO9780511790492.005.

Edge P, Bansal V. 2019. Longshot enables accurate variant calling in diploid genomes from single-molecule long read sequencing. *Nat Commun* **10**: 4660.

ENCODE Project Consortium. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**: 57–74.

Escalona M, Rocha S, Posada D. 2016. A comparison of tools for the simulation of genomic next-generation sequencing data. *Nat Rev Genet* **17**: 459–469.

Ewing AD, Kazazian HH. 2010. High-throughput sequencing reveals extensive variation in human-specific L1 content in individual human genomes. *Genome Res* **20**: 1262–1270.

Ewing AD, Smits N, Sanchez-Luque FJ, Faivre J, Brennan PM, Richardson SR, Cheetham SW, Faulkner GJ. 2020. Nanopore Sequencing Enables Comprehensive Transposable Element Epigenomic Profiling. *Mol Cell* **80**: 915-928.e5.

Fairley S, Lowy-Gallego E, Perry E, Flicek P. 2020. The International Genome Sample Resource (IGSR) collection of open human genomic variation resources. *Nucleic Acids Res* **48**: D941–D947.

Feichtinger J, Aldeailej I, Anderson R, Almutairi M, Almatrafi A, Alsiwiehri N, Griffiths K, Stuart N, Wakeman JA, Larcombe L, et al. 2012. Meta-analysis of clinical data using human meiotic genes identifies a novel cohort of highly restricted cancer-specific marker genes. *Oncotarget* **3**: 843–853.

Ferrarini A, Xumerle L, Griggio F, Garonzi M, Cantaloni C, Centomo C, Vargas SM, Descombes P, Marquis J, Collino S, et al. 2015. The Use of Non-Variant Sites to Improve the Clinical Assessment of Whole-Genome Sequence Data. *PLOS ONE* **10**: e0132180.

Formenti G, Abueg L, Brajuka A, Brajuka N, Gallardo-Alba C, Giani A, Fedrigo O, Jarvis ED. 2022. Gfastats: conversion, evaluation and manipulation of genome sequences using assembly graphs. *Bioinformatics* **38**: 4214–4216.

Fowler A, Mahamdallie S, Ruark E, Seal S, Ramsay E, Clarke M, Uddin I, Wylie H, Strydom A, Lunter G, et al. 2016. Accurate clinical detection of exon copy number variants in a targeted NGS panel using DECoN. 1:20. doi: 10.12688/wellcomeopenres.10069.1.

Francioli LC, Menelaou A, Pulit SL, van Dijk F, Palamara PF, Elbers CC, Neerincx PBT, Ye K, Guryev V, Kloosterman WP, et al. 2014. Whole-genome sequence variation, population structure and demographic history of the Dutch population. *Nat Genet* **46**: 818–825.

Fu Q, Li H, Moorjani P, Jay F, Slepchenko SM, Bondarev AA, Johnson PLF, Aximu-Petri A, Prüfer K, de Filippo C, et al. 2014. Genome sequence of a 45,000-year-old modern human from western Siberia. *Nature* **514**: 445–449.

Fu S, Wang A, Au KF. 2019. A comparative evaluation of hybrid error correction methods for error-prone long reads. *Genome Biol* **20**: 26.

Fumasoni I, Meani N, Rambaldi D, Scafetta G, Alcalay M, Ciccarelli FD. 2007. Family expansion and gene rearrangements contributed to the functional specialization of PRDM genes in vertebrates. *BMC Evol Biol* **7**: 187.

Gaedigk A, Ingelman-Sundberg M, Miller NA, Leeder JS, Whirl-Carrillo M, Klein TE, PharmVar Steering Committee. 2018. The Pharmacogene Variation (PharmVar) Consortium: Incorporation of the Human Cytochrome P450 (CYP) Allele Nomenclature Database. *Clin Pharmacol Ther* **103**: 399–401.

Gao L, Gonda I, Sun H, Ma Q, Bao K, Tieman DM, Burzynski-Chang EA, Fish TL, Stromberg KA, Sacks GL, et al. 2019. The tomato pan-genome uncovers new genes and a rare allele regulating fruit flavor. *Nat Genet* **51**: 1044–1051.

Gao Y, Liu Y, Ma Y, Liu B, Wang Y, Xing Y. 2021. abPOA: an SIMD-based C library for fast partial order alignment using adaptive band. *Bioinformatics* **37**: 2209–2211.

Garrison E. 2019. Graphical pangenomics. Thesis, University of Cambridge https://www.repository.cam.ac.uk/handle/1810/294516 (Accessed October 19, 2022).

Garrison E, Marth G. 2012. Haplotype-based variant detection from short-read sequencing. *arXiv* 10.48550/arXiv.1207.3907.

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.

Georgakopoulos-Soares I, Yizhar-Barnea O, Mouratidis I, Hemberg M, Ahituv N. 2021. Absent from DNA and protein: genomic characterization of nullomers and nullpeptides across functional categories and evolution. *Genome Biol* **22**: 245.

Gilpatrick T, Lee I, Graham JE, Raimondeau E, Bowen R, Heron A, Downs B, Sukumar S, Sedlazeck FJ, Timp W. 2020. Targeted nanopore sequencing with Cas9-guided adapter ligation. *Nat Biotechnol* **38**: 433–438.

Goerner-Potvin P, Bourque G. 2018. Computational tools to unmask transposable elements. *Nat Rev Genet* **19**: 688–704.

Goodwin S, McPherson JD, McCombie WR. 2016. Coming of age: ten years of next-generation sequencing technologies. *Nat Rev Genet* **17**: 333–351.

Gourraud P-A, Khankhanian P, Cereb N, Yang SY, Feolo M, Maiers M, Rioux JD, Hauser S, Oksenberg J. 2014. HLA Diversity in the 1000 Genomes Dataset. *PLOS ONE* **9**: e97282.

Greaves M. 2006. Infection, immune responses and the aetiology of childhood leukaemia. *Nat Rev Cancer* **6**: 193–203.

Green RE, Krause J, Briggs AW, Maricic T, Stenzel U, Kircher M, Patterson N, Li H, Zhai W, Fritz MH-Y, et al. 2010. A Draft Sequence of the Neandertal Genome. *Science* **328**: 710–722.

Griffith M, Spies NC, Krysiak K, McMichael JF, Coffman AC, Danos AM, Ainscough BJ, Ramirez CA, Rieke DT, Kujan L, et al. 2017. CIViC is a community knowledgebase for expert crowdsourcing the clinical interpretation of variants in cancer. *Nat Genet* **49**: 170–174.

Gudbjartsson DF, Helgason H, Gudjonsson SA, Zink F, Oddson A, Gylfason A, Besenbacher S, Magnusson G, Halldorsson BV, Hjartarson E, et al. 2015. Large-scale whole-genome sequencing of the Icelandic population. *Nat Genet* **47**: 435–444.

Günther T, Nettelblad C. 2019. The presence and impact of reference bias on population genomic studies of prehistoric human populations. *PLoS Genet* **15**: e1008302.

Hampikian G, Andersen T. 2007. Absent sequences: nullomers and primes. *Pac Symp Biocomput* 355–366.

Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, Aken BL, Barrell D, Zadissa A, Searle S, et al. 2012. GENCODE: The reference human genome annotation for The ENCODE Project. *Genome Res* **22**: 1760–1774.

Heber S, Alekseyev M, Sze S-H, Tang H, Pevzner PA. 2002. Splicing graphs and EST assembly problem. *Bioinformatics* **18**: S181–S188.

Hehir-Kwa JY, Marschall T, Kloosterman WP, Francioli LC, Baaijens JA, Dijkstra LJ, Abdellaoui A, Koval V, Thung DT, Wardenaar R, et al. 2016. A high-quality human reference panel reveals the complexity and distribution of genomic structural variants. *Nat Commun* **7**.

Hein J. 1989. A new method that simultaneously aligns and reconstructs ancestral sequences for any number of homologous sequences, when the phylogeny is given. *Mol Biol Evol* **6**: 649–668.

Heydari M, Miclotte G, Demeester P, Van de Peer Y, Fostier J. 2017. Evaluation of the impact of Illumina error correction tools on de novo genome assembly. *BMC Bioinformatics* **18**: 374.

Heydari M, Miclotte G, Van de Peer Y, Fostier J. 2018. BrownieAligner: accurate alignment of Illumina sequencing data to de Bruijn graphs. *BMC Bioinformatics* **19**: 311.

Hickey G, Heller D, Monlong J, Sibbesen JA, Sirén J, Eizenga J, Dawson ET, Garrison E, Novak AM, Paten B. 2020. Genotyping structural variants in pangenome graphs using the vg toolkit. *Genome Biol* **21**: 35.

Hickey G, Monlong J, Novak A, Eizenga JM, Consortium HPR, Li H, Paten B. 2022. Pangenome Graph Construction from Genome Alignment with Minigraph-Cactus. *bioRxiv* 2022.10.06.511217.

Hillmer M, Wagner D, Summerer A, Daiber M, Mautner V-F, Messiaen L, Cooper DN, Kehrer-Sawatzki H. 2016. Fine mapping of meiotic NAHR-associated crossovers causing large NF1 deletions. *Hum Mol Genet* **25**: 484–496.

Hinch AG, Tandon A, Patterson N, Song Y, Rohland N, Palmer CD, Chen GK, Wang K, Buxbaum SG, Akylbekova EL, et al. 2011. The landscape of recombination in African Americans. *Nature* **476**: 170–175.

Hoffman SMG a, Nelson DR b, Keeney DS c. 2001. Organization, structure and evolution of the CYP2 gene cluster on human chromosome 19. *Pharmacogenetics* **11**: 687–698.

Hon T, Mars K, Young G, Tsai Y-C, Karalius JW, Landolin JM, Maurer N, Kudrna D, Hardigan MA, Steiner CC, et al. 2020. Highly accurate long-read HiFi sequencing data for five complex genomes. *Sci Data* **7**: 399.

Horton R, Gibson R, Coggill P, Miretti M, Allcock RJ, Almeida J, Forbes S, Gilbert JGR, Halls K, Harrow JL, et al. 2008. Variation analysis and gene annotation of eight MHC haplotypes: The MHC Haplotype Project. *Immunogenetics* **60**: 1–18.

Hu L, Liang F, Cheng D, Zhang Z, Yu G, Zha J, Wang Y, Xia Q, Yuan D, Tan Y, et al. 2020. Location of Balanced Chromosome-Translocation Breakpoints by Long-Read Sequencing on the Oxford Nanopore Platform. *Front Genet* **10**.

Huang L, Popic V, Batzoglou S. 2013. Short read alignment with populations of genomes. *Bioinformatics* **29**: i361–i370.

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.

Human Pangenome Reference Consortium. 2021. Year 1 Sequencing Data. *Hum Pangenome Ref Consort GitHub Repos.*
https://github.com/human-pangenomics/HPP_Year1_Data_Freeze_v1.0.

Hussin J, Sinnett D, Casals F, Idaghdour Y, Bruat V, Saillour V, Healy J, Grenier J-C, Malliard T de, Busche S, et al. 2013. Rare allelic forms of PRDM9 associated with childhood leukemogenesis. *Genome Res* **23**: 419–430.

Idury RM, Waterman MS. 1995. A new algorithm for DNA sequence assembly. *J Comput Biol J Comput Mol Cell Biol* **2**: 291–306.

International HapMap Consortium. 2005. A haplotype map of the human genome. *Nature* **437**: 1299–1320.

International Human Genome Sequencing Consortium. 2004. Finishing the euchromatic sequence of the human genome. *Nature* **431**: 931–945.

Iqbal Z, Caccamo M, Turner I, Flicek P, McVean G. 2012. De novo assembly and genotyping of variants using colored de Bruijn graphs. *Nat Genet* **44**: 226–232.

Irie S, Tsujimura A, Miyagawa Y, Ueda T, Matsuoka Y, Matsui Y, Okuyama A, Nishimune Y, Tanaka H. 2009. Single-Nucleotide Polymorphisms of the PRDM9 (MEISETZ) Gene in Patients With Nonobstructive Azoospermia. *J Androl* **30**: 426–431.

Jackson L, Kuhlman C, Jackson F, Fox PK. 2019. Including Vulnerable Populations in the Assessment of Data From Vulnerable Populations. *Front Big Data* **2**.

Jain C, Dilthey A, Misra S, Zhang H, Aluru S. 2019. Accelerating Sequence Alignment to Graphs. *bioRxiv* 10.1101/651638v1.

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.

Jain M, Olsen HE, Turner DJ, Stoddart D, Bulazel KV, Paten B, Haussler D, Willard HF, Akeson M, Miga KH. 2018. Linear assembly of a human centromere on the Y chromosome. *Nat Biotechnol* **36**: 321–323.

Jeffreys AJ, Cotton VE, Neumann R, Lam K-WG. 2013. Recombination regulator PRDM9 influences the instability of its own coding sequence in humans. *Proc Natl Acad Sci* **110**: 600–605.

Jiang T, Liu Y, Jiang Y, Li J, Gao Y, Cui Z, Liu Y, Liu B, Wang Y. 2020. Long-read-based human genomic structural variation detection with cuteSV. *Genome Biol* **21**: 189.

Johansson LF, van Dijk F, de Boer EN, van Dijk-Bos KK, Jongbloed JDH, van der Hout AH, Westers H, Sinke RJ, Swertz MA, Sijmons RH, et al. 2016. CoNVaDING: Single Exon Variation Detection in Targeted NGS Data. *Hum Mutat* **37**: 457–464.

Jurka J, Gentles AJ. 2006. Origin and diversification of minisatellites derived from human Alu sequences. *Gene* **365**: 21–26.

Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, Collins RL, Laricchia KM, Ganna A, Birnbaum DP, et al. 2020. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**: 434–443.

Karthikeyan S, Bawa PS, Srinivasan S. 2017. hg19K: addressing a significant lacuna in hg19-based variant calling. *Mol Genet Genomic Med* **5**: 15–20.

Katz Y, Wang ET, Airoldi EM, Burge CB. 2010. Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nat Methods* **7**: 1009–1015.

Kelley DR, Schatz MC, Salzberg SL. 2010. Quake: quality-aware detection and correction of sequencing errors. *Genome Biol* **11**: R116.

Khrapko K r., P. Lysov Y, Khorlyn A a., Shick V v., Florentiev V l., Mirzabekov A d. 1989. An oligonucleotide hybridization approach to DNA sequencing. *FEBS Lett* **256**: 118–122.

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.

Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. 2019. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat Biotechnol* **37**: 907–915.

Kim J-I, Ju YS, Park H, Kim S, Lee S, Yi J-H, Mudge J, Miller NA, Hong D, Bell CJ, et al. 2009. A highly annotated whole-genome sequence of a Korean individual. *Nature* **460**: 1011–1015.

Kirk JM, Kim SO, Inoue K, Smola MJ, Lee DM, Schertzer MD, Wooten JS, Baker AR, Sprague D, Collins DW, et al. 2018. Functional classification of long non-coding RNAs by k-mer content. *Nat Genet* **50**: 1474–1482.

Kirsh VA, Skead K, McDonald K, Kreiger N, Little J, Menard K, McLaughlin J, Mukherjee S, Palmer LJ, Goel V, et al. 2022. Cohort Profile: The Ontario Health Study (OHS). *Int J Epidemiol* dyac156.

Knight RD, Shimeld SM. 2001. Identification of conserved C2H2 zinc-finger gene families in the Bilateria. *Genome Biol* **2**: research0016.1-research0016.8.

Kong A, Thorleifsson G, Gudbjartsson DF, Masson G, Sigurdsson A, Jonasdottir A, Walters GB, Jonasdottir A, Gylfason A, Kristinsson KT, et al. 2010. Fine-scale recombination rate differences between sexes, populations and individuals. *Nature* **467**: 1099–1103.

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.

Kurtz S, Narechania A, Stein JC, Ware D. 2008. A new method to compute K-mer frequencies and its application to annotate large repetitive plant genomes. *BMC Genomics* **9**: 517.

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.

Landrum MJ, Lee JM, Benson M, Brown GR, Chao C, Chitipiralla S, Gu B, Hart J, Hoffman D, Jang W, et al. 2018. ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res* **46**: D1062–D1067.

Landrum MJ, Lee JM, Riley GR, Jang W, Rubinstein WS, Church DM, Maglott DR. 2014. ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res* **42**: D980-985.

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

Lappalainen I, Lopez J, Skipper L, Hefferon T, Spalding JD, Garner J, Chen C, Maguire M, Corbett M, Zhou G, et al. 2013. DbVar and DGVa: public archives for genomic structural variation. *Nucleic Acids Res* **41**: D936-941.

Layer RM, Chiang C, Quinlan AR, Hall IM. 2014. LUMPY: a probabilistic framework for structural variant discovery. *Genome Biol* **15**: R84.

Lee C, Grasso C, Sharlow MF. 2002. Multiple sequence alignment using partial order graphs. *Bioinformatics* **18**: 452–464.

Letcher B, Hunt M, Iqbal Z. 2021. Gramtools enables multiscale variation analysis with genome graphs. *Genome Biol* **22**: 259.

Levy S, Sutton G, Ng PC, Feuk L, Halpern AL, Walenz BP, Axelrod N, Huang J, Kirkness EF, Denisov G, et al. 2007. The Diploid Genome Sequence of an Individual Human. *PLOS Biol* **5**: e254.

Li H. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv* 10.48550/arXiv.1303.3997

Li H. 2012. Exploring single-sample SNP and INDEL calling with whole-genome de novo assembly. *Bioinformatics* **28**: 1838–1844.

Li H. 2018. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**: 3094–3100.

Li H, Feng X, Chu C. 2020. The design and construction of reference pangenome graphs with minigraph. *Genome Biol* **21**: 265.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**: 2078–2079.

Li J, Lupat R, Amarasinghe KC, Thompson ER, Doyle MA, Ryland GL, Tothill RW, Halgamuge SK, Campbell IG, Gorringe KL. 2012. CONTRA: copy number analysis for targeted resequencing. *Bioinforma Oxf Engl* **28**: 1307–1313.

Li R, Li Y, Zheng H, Luo R, Zhu H, Li Q, Qian W, Ren Y, Tian G, Li J, et al. 2010. Building the sequence map of the human pan-genome. *Nat Biotechnol* **28**: 57–63.

Liao W-W, Asri M, Ebler J, Doerr D, Haukness M, Hickey G, Lu S, Lucas JK, Monlong J, Abel HJ, et al. 2022. A Draft Human Pangenome Reference. *bioRxiv* 10.1101/2022.07.09.499321.

Lupski JR. 1998. Genomic disorders: structural features of the genome can lead to DNA rearrangements and human disease traits. *Trends Genet* **14**: 417–422.

Ma J, Cáceres M, Salmela L, Mäkinen V, Tomescu AI. 2022. Chaining for Accurate Alignment of Erroneous Long Reads to Acyclic Variation Graphs. *bioRxiv* 10.1101/2022.01.07.475257.

Maciuca S, del Ojo Elias C, McVean G, Iqbal Z. 2016. A Natural Encoding of Genetic Variation in a Burrows-Wheeler Transform to Enable Mapping and Genome Inference. In *Algorithms in Bioinformatics* (eds. M. Frith and C.N. Storm Pedersen), *Lecture Notes in Computer Science*, pp. 222–233, Springer International Publishing, Cham.

Magi A, D'Aurizio R, Palombo F, Cifola I, Tattini L, Semeraro R, Pippucci T, Giusti B, Romeo G, Abbate R, et al. 2015. Characterization and identification of hidden rare variants in the human genome. *BMC Genomics* **16**: 340.

Mahmoud M, Doddapaneni H, Timp W, Sedlazeck FJ. 2021. PRINCESS: comprehensive detection of haplotype resolved SNVs, SVs, and methylation. *Genome Biol* **22**: 268.

Mahmoud M, Gobet N, Cruz-Dávalos DI, Mounier N, Dessimoz C, Sedlazeck FJ. 2019. Structural variant calling: the long and the short of it. *Genome Biol* **20**: 246.

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.

Mao Q, Ciotlos S, Zhang RY, Ball MP, Chin R, Carnevali P, Barua N, Nguyen S, Agarwal MR, Clegg T, et al. 2016. The whole genome sequences and experimentally phased haplotypes of over 100 personal genomes. *GigaScience* **5**: 42.

Marçais G, Kingsford C. 2011. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* **27**: 764–770.

Marcus S, Lee H, Schatz MC. 2014. SplitMEM: a graphical algorithm for pan-genome analysis with suffix skips. *Bioinformatics* **30**: 3476–3483.

Maretty L, Jensen JM, Petersen B, Sibbesen JA, Liu S, Villesen P, Skov L, Belling K, Theil Have C, Izarzugaza JMG, et al. 2017. Sequencing and de novo assembly of 150 genomes from Denmark as a population reference. *Nature* **548**: 87–91.

Martin S, Heavens D, Lan Y, Horsfield S, Clark MD, Leggett RM. 2022. Nanopore adaptive sampling: a tool for enrichment of low abundance species in metagenomic samples. *Genome Biol* **23**: 11.

Martiniano R, Garrison E, Jones ER, Manica A, Durbin R. 2020. Removing reference bias and improving indel calling in ancient DNA data analysis by mapping to a sequence variation graph. *Genome Biol* **21**: 250.

McKernan KJ, Peckham HE, Costa GL, McLaughlin SF, Fu Y, Tsung EF, Clouser CR, Duncan C, Ichikawa JK, Lee CC, et al. 2009. Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding. *Genome Res* **19**: 1527–1541.

Miao H, Zhou J, Yang Q, Liang F, Wang D, Ma N, Gao B, Du J, Lin G, Wang K, et al. 2018. Long-read sequencing identified a causal structural variant in an exome-negative case and enabled preimplantation genetic diagnosis. *Hereditas* **155**: 32.

Middleton D, Gonzelez F. 2010. The extensive polymorphism of KIR genes. *Immunology* **129**: 8–19.

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.

Miga KH, Wang T. 2021. The Need for a Human Pangenome Reference Sequence. *Annu Rev Genomics Hum Genet* **22**: 81–102.

Myers EW. 2005. The fragment assembly string graph. *Bioinforma Oxf Engl* **21** Suppl 2: ii79-85.

Myers EW. 1995. Toward simplifying and accurately formulating fragment assembly. *J Comput Biol J Comput Mol Cell Biol* **2**: 275–290.

Myers EW, Sutton GG, Delcher AL, Dew IM, Fasulo DP, Flanigan MJ, Kravitz SA, Mobarry CM, Reinert KHJ, Remington KA, et al. 2000. A Whole-Genome Assembly of Drosophila. *Science* **287**: 2196–2204.

Myers Jr EW. 2016. A history of DNA sequence assembly. *It - Inf Technol* **58**: 126–132.

Myers S, Freeman C, Auton A, Donnelly P, McVean G. 2008. A common sequence motif associated with recombination hot spots and genome instability in humans. *Nat Genet* **40**: 1124–1129.

Nagasaki M, Yasuda J, Katsuoka F, Nariai N, Kojima K, Kawai Y, Yamaguchi-Kabata Y, Yokozawa J, Danjoh I, Saito S, et al. 2015. Rare variant discovery by deep whole-genome sequencing of 1,070 Japanese individuals. *Nat Commun* **6**: 8018.

Norman PJ, Hollenbach JA, Nemat-Gorgani N, Marin WM, Norberg SJ, Ashouri E, Jayaraman J, Wroblewski EE, Trowsdale J, Rajalingam R, et al. 2016. Defining KIR and HLA Class I Genotypes at Highest Resolution via High-Throughput Sequencing. *Am J Hum Genet* **99**: 375–391.

Novak AM, Hickey G, Garrison E, Blum S, Connelly A, Dilthey A, Eizenga J, Elmohamed MAS, Guthrie S, Kahles A, et al. 2017. Genome Graphs. *bioRxiv* 10.1101/101378.

Nurk S, Koren S, Rhie A, Rautiainen M, Bzikadze AV, Mikheenko A, Vollger MR, Altemose N, Uralsky L, Gershman A, et al. 2022. The complete sequence of a human genome. *Science* **376**: 44–53.

Oliver PL, Goodstadt L, Bayes JJ, Birtle Z, Roach KC, Phadnis N, Beatson SA, Lunter G, Malik HS, Ponting CP. 2009. Accelerated Evolution of the Prdm9 Speciation Gene across Diverse Metazoan Taxa. *PLoS Genet* **5**.

Oliver TR, Feingold E, Yu K, Cheung V, Tinker S, Yadav-Shah M, Masse N, Sherman SL. 2008. New Insights into Human Nondisjunction of Chromosome 21 in Oocytes. *PLoS Genet* **4**: e1000033.

Oliver TR, Middlebrooks C, Harden A, Scott N, Johnson B, Jones J, Walker C, Wilkerson C, Saffold S-H, Akinseye A, et al. 2016. Variation in the Zinc Finger of PRDM9 is Associated with the Absence of Recombination along Nondisjoined Chromosomes 21 of Maternal Origin. *J Syndr Chromosome Abnorm* **2**: 115.

Pace JK, Feschotte C. 2007. The evolutionary history of human DNA transposons: Evidence for intense activity in the primate lineage. *Genome Res* **17**: 422–432.

Pandey RV, Franssen SU, Futschik A, Schlötterer C. 2013. Allelic imbalance metre (Allim), a new tool for measuring allele-specific gene expression with RNA-seq data. *Mol Ecol Resour* **13**: 740–745.

Parvanov ED, Petkov PM, Paigen K. 2010. Prdm9 Controls Activation of Mammalian Recombination Hotspots. *Science* **327**: 835.

Parvanov ED, Tian H, Billings T, Saxl RL, Spruce C, Aithal R, Krejci L, Paigen K, Petkov PM. 2017. PRDM9 interactions with other proteins provide a link between recombination hotspots and the chromosomal axis in meiosis. *Mol Biol Cell* **28**: 488–499.

Paten B, Novak A, Haussler D. 2014. Mapping to a Reference Genome Structure. *arXiv* 10.48550/arXiv.1404.5010

Patro R, Mount SM, Kingsford C. 2014. Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms. *Nat Biotechnol* **32**: 462–464.

Pennisi E. 2022. Most complete human genome yet is revealed. *Science* **376**: 15–16.

Pertea M, Shumate A, Pertea G, Varabyou A, Breitwieser FP, Chang Y-C, Madugundu AK, Pandey A, Salzberg SL. 2018. CHESS: a new human gene catalog curated from thousands of large-scale RNA sequencing experiments reveals extensive transcriptional noise. *Genome Biol* **19**: 208.

Pevzner PA. 1989. l-Tuple DNA sequencing: computer analysis. *J Biomol Struct Dyn* **7**: 63–73.

Pevzner PA, Tang H, Waterman MS. 2001. An Eulerian path approach to DNA fragment assembly. *Proc Natl Acad Sci* **98**: 9748–9753.

Peyrégne S, Slon V, Mafessoni F, de Filippo C, Hajdinjak M, Nagel S, Nickel B, Essel E, Le Cabec A, Wehrberger K, et al. 2019. Nuclear DNA from two early Neandertals reveals 80,000 years of genetic continuity in Europe. *Sci Adv* **5**: eaaw5873.

Ponting CP. 2011. What are the genomic drivers of the rapid evolution of PRDM9? *Trends Genet* **27**: 165–171.

Popejoy AB, Fullerton SM. 2016. Genomics is failing on diversity. *Nature* **538**: 161–164.

Poplin R, Chang P-C, Alexander D, Schwartz S, Colthurst T, Ku A, Newburger D, Dijamco J, Nguyen N, Afshar PT, et al. 2018a. A universal SNP and small-indel variant caller using deep neural networks. *Nat Biotechnol* **36**: 983–987.

Poplin R, Ruano-Rubio V, DePristo MA, Fennell TJ, Carneiro MO, Auwera GAV der, Kling DE, Gauthier LD, Levy-Moonshine A, Roazen D, et al. 2018b. Scaling accurate genetic variant discovery to tens of thousands of samples. *bioRxiv* 10.1101/201178.

Powers NR, Parvanov ED, Baker CL, Walker M, Petkov PM, Paigen K. 2016. The Meiotic Recombination Activator PRDM9 Trimethylates Both H3K36 and H3K4 at Recombination Hotspots In Vivo. *PLoS Genet* **12**.

Pratto F, Brick K, Khil P, Smagulova F, Petukhova GV, Camerini-Otero RD. 2014. Recombination initiation maps of individual human genomes. *Science* **346**: 1256442.

Price AL, Jones NC, Pevzner PA. 2005. De novo identification of repeat families in large genomes. *Bioinformatics* **21**: i351–i358.

Pritt J, Chen N-C, Langmead B. 2018. FORGe: prioritizing variants for graph genomes. *Genome Biol* **19**: 220.

Prjibelski A, Antipov D, Meleshko D, Lapidus A, Korobeynikov A. 2020. Using SPAdes De Novo Assembler. *Curr Protoc Bioinforma* **70**: e102.

Prüfer K. 2018. snpAD: an ancient DNA genotype caller. *Bioinformatics* **34**: 4165–4171.

Ptak SE, Hinds DA, Koehler K, Nickel B, Patil N, Ballinger DG, Przeworski M, Frazer KA, Pääbo S. 2005. Fine-scale recombination patterns differ between chimpanzees and humans. *Nat Genet* **37**: 429–434.

Rakocevic G, Semenyuk V, Lee W-P, Spencer J, Browning J, Johnson IJ, Arsenijevic V, Nadj J, Ghose K, Suciu MC, et al. 2019. Fast and accurate genomic analyses using genome graphs. *Nat Genet* **51**: 354–362.

Rasmussen M, Li Y, Lindgreen S, Pedersen JS, Albrechtsen A, Moltke I, Metspalu M, Metspalu E, Kivisild T, Gupta R, et al. 2010. Ancient human genome sequence of an extinct Palaeo-Eskimo. *Nature* **463**: 757–762.

Rausch T, Zichner T, Schlattl A, Stütz AM, Benes V, Korbel JO. 2012. DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinforma Oxf Engl* **28**: i333–i339.

Rautiainen M, Marschall T. 2020. GraphAligner: rapid and versatile sequence-to-graph alignment. *Genome Biol* **21**: 253.

Rhie A, Nurk S, Cechova M, Hoyt SJ, Taylor DJ, Altemose N, Hook PW, Koren S, Rautiainen M, Alexandrov IA, et al. 2022. The complete sequence of a human Y chromosome. *bioRxiv* 2022.12.01.518724.

Rhoads A, Au KF. 2015. PacBio Sequencing and Its Applications. *Genomics Proteomics Bioinformatics* **13**: 278–289.

Rimmer A, Phan H, Mathieson I, Iqbal Z, Twigg SRF, Wilkie AOM, McVean G, Lunter G. 2014. Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. *Nat Genet* **46**: 912–918.

Robinson J, Halliwell JA, McWilliam H, Lopez R, Marsh SGE. 2013. IPD—the Immuno Polymorphism Database. *Nucleic Acids Res* **41**: D1234–D1240.

Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP. 2011. Integrative genomics viewer. *Nat Biotechnol* **29**: 24–26.

Romanienko PJ, Camerini-Otero RD. 2000. The Mouse Spo11 Gene Is Required for Meiotic Chromosome Synapsis. *Mol Cell* **6**: 975–987.

Schneeberger K, Hagmann J, Ossowski S, Warthmann N, Gesing S, Kohlbacher O, Weigel D. 2009. Simultaneous alignment of short reads against multiple genomes. *Genome Biol* **10**: R98.

Schneider VA, Graves-Lindsay T, Howe K, Bouk N, Chen H-C, Kitts PA, Murphy TD, Pruitt KD, Thibaud-Nissen F, Albracht D, et al. 2017. Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. *Genome Res* **27**: 849–864.

Schumacher TN, Schreiber RD. 2015. Neoantigens in cancer immunotherapy. *Science* **348**: 69–74.

Schuster SC, Miller W, Ratan A, Tomsho LP, Giardine B, Kasson LR, Harris RS, Petersen DC, Zhao F, Qi J, et al. 2010. Complete Khoisan and Bantu genomes from southern Africa. *Nature* **463**: 943–947.

Schwartz S, Oren R, Ast G. 2011. Detection and Removal of Biases in the Analysis of Next-Generation Sequencing Reads. *PLoS ONE* **6**: e16685.

Sedlazeck FJ, Rescheneder P, Smolka M, Fang H, Nattestad M, von Haeseler A, Schatz MC. 2018. Accurate detection of complex structural variations using single-molecule sequencing. *Nat Methods* **15**: 461–468.

Seo J-S, Rhie A, Kim J, Lee S, Sohn M-H, Kim C-U, Hastie A, Cao H, Yun J-Y, Kim J, et al. 2016. De novo assembly and phasing of a Korean human genome. *Nature* **538**: 243–247.

Sereika M, Kirkegaard RH, Karst SM, Michaelsen TY, Sørensen EA, Wollenberg RD, Albertsen M. 2022. Oxford Nanopore R10.4 long-read sequencing enables the generation of near-finished bacterial genomes from pure cultures and metagenomes without short-read or reference polishing. *Nat Methods* **19**: 823–826.

Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. 2003. Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Res* **13**: 2498–2504.

Sherman RM, Forman J, Antonescu V, Puiu D, Daya M, Rafaels N, Boorgula MP, Chavan S, Vergara C, Ortega VE, et al. 2019. Assembly of a pan-genome from deep sequencing of 910 humans of African descent. *Nat Genet* **51**: 30–35.

Sherman SL, Takaesu N, Freeman SB, Grantham M, Phillips C, Blackston RD, Jacobs PA, Cockwell AE, Freeman V, Uchida I, et al. 1991. Trisomy 21: Association between reduced recombination and nondisjunction. *Am J Hum Genet* **49**: 608–620.

Sherry ST, Ward M, Sirotkin K. 1999. dbSNP—Database for Single Nucleotide Polymorphisms and Other Classes of Minor Genetic Variation. *Genome Res* **9**: 677–679.

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.

Shumate A, Zimin AV, Sherman RM, Puiu D, Wagner JM, Olson ND, Pertea M, Salit ML, Zook JM, Salzberg SL. 2020. Assembly and annotation of an Ashkenazi human reference genome. *Genome Biol* **21**: 129.

Shung KP. 2018. Accuracy, Precision, Recall or F1? *Medium*. https://towardsdatascience.com/accuracy-precision-recall-or-f1-331fb37c5cb9 (Accessed December 5, 2022).

Sibbesen JA, Eizenga JM, Novak AM, Sirén J, Chang X, Garrison E, Paten B. 2022. Haplotype-aware pantranscriptome analyses using spliced pangenome graphs. *bioRxiv* 2021.03.26.437240.

Sigaux F. 2000. Cancer genome or the development of molecular portraits of tumors. *Bull Acad Natl Med* **184**: 1441–1447; discussion 1448-1449.

Simpson JT, Durbin R. 2012. Efficient de novo assembly of large genomes using compressed data structures. *Genome Res* **22**: 549–556.

Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJM, Birol İ. 2009. ABySS: A parallel assembler for short read sequence data. *Genome Res* **19**: 1117–1123.

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.

Sirén J. 2017. Indexing Variation Graphs. In *2017 Proceedings of the Ninteenth Workshop on Algorithm Engineering and Experiments (ALENEX), Proceedings*, pp. 13–27, Society for Industrial and Applied Mathematics.

Sirén J, Monlong J, Chang X, Novak AM, Eizenga JM, Markello C, Sibbesen JA, Hickey G, Chang P-C, Carroll A, et al. 2021. Genotyping common, large structural variations in 5,202 genomes using pangenomes, the Giraffe mapper, and the vg toolkit. *bioRxiv* 10.1101/2020.12.04.412486.

Sirén J, Välimäki N, Mäkinen V. 2011. Indexing Finite Language Representation of Population Genotypes. In *Algorithms in Bioinformatics* (eds. T.M. Przytycka and M.-F. Sagot), *Lecture Notes in Computer Science*, pp. 270–281, Springer, Berlin, Heidelberg.

Sirugo G, Williams SM, Tishkoff SA. 2019. The Missing Diversity in Human Genetic Studies. *Cell* **177**: 26–31.

Smigielski EM, Sirotkin K, Ward M, Sherry ST. 2000. dbSNP: a database of single nucleotide polymorphisms. *Nucleic Acids Res* **28**: 352–355.

Sohn J, Nam J-W. 2018. The present and future of de novo whole-genome assembly. *Brief Bioinform* **19**: 23–40.

Šošić M, Šikić M. 2017. Edlib: a C/C ++ library for fast, exact sequence alignment using edit distance. *Bioinformatics* **33**: 1394–1395.

Stankiewicz P, Lupski JR. 2002. Genome architecture, rearrangements and genomic disorders. *Trends Genet* **18**: 74–82.

Stevenson KR, Coolon JD, Wittkopp PJ. 2013. Sources of bias in measures of allele-specific expression derived from RNA-seq data aligned to a single reference genome. *BMC Genomics* **14**: 536.

Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J, Downey P, Elliott P, Green J, Landray M, et al. 2015. UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. *PLOS Med* **12**: e1001779.

Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, Huddleston J, Zhang Y, Ye K, Jun G, Hsi-Yang Fritz M, et al. 2015. An integrated map of structural variation in 2,504 human genomes. *Nature* **526**: 75–81.

Sun F, Fujiwara Y, Reinholdt LG, Hu J, Saxl RL, Baker CL, Petkov PM, Paigen K, Handel MA. 2015. Nuclear Localization of PRDM9 and Its Role in Meiotic Chromatin Modifications and Homologous Synapsis. *Chromosoma* **124**: 397–415.

Tate JG, Bamford S, Jubb HC, Sondka Z, Beare DM, Bindal N, Boutselakis H, Cole CG, Creatore C, Dawson E, et al. 2019. COSMIC: the Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Res* **47**: D941–D947.

Tettelin H, Masignani V, Cieslewicz MJ, Donati C, Medini D, Ward NL, Angiuoli SV, Crabtree J, Jones AL, Durkin AS, et al. 2005. Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: Implications for the microbial "pan-genome." *Proc Natl Acad Sci* **102**: 13950–13955.

The 1000 Genomes Project Consortium. 2015. A global reference for human genetic variation. *Nature* **526**: 68–74.

Thomas JH, Emerson RO, Shendure J. 2009. Extraordinary Molecular Evolution in the PRDM9 Fertility Gene. *PLOS ONE* **4**: e8505.

Thomas PD, Kejariwal A. 2004. Coding single-nucleotide polymorphisms associated with complex vs. Mendelian disease: Evolutionary evidence for differences in molecular effects. *Proc Natl Acad Sci* **101**: 15398–15403.

Trapnell C, Pachter L, Salzberg SL. 2009. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25**: 1105–1111.

Tryka KA, Hao L, Sturcke A, Jin Y, Wang ZY, Ziyabari L, Lee M, Popova N, Sharopova N, Kimura M, et al. 2014. NCBI's Database of Genotypes and Phenotypes: dbGaP. *Nucleic Acids Res* **42**: D975-979.

van de Geijn B, McVicker G, Gilad Y, Pritchard JK. 2015. WASP: allele-specific software for robust molecular quantitative trait locus discovery. *Nat Methods* **12**: 1061–1063.

Van der Auwera GA, O'Connor BD. 2020. *Genomics in the Cloud*. O'Reilly Media, Inc.

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.

Vergés L, Vidal F, Geán E, Alemany-Schmidt A, Oliver-Bonet M, Blanco J. 2017. An exploratory study of predisposing genetic factors for DiGeorge/velocardiofacial syndrome. *Sci Rep* **7**: 40031.

Vernikos G, Medini D, Riley DR, Tettelin H. 2015. Ten years of pan-genome analyses. *Curr Opin Microbiol* **23**: 148–154.

Vijaya Satya R, Zavaljevski N, Reifman J. 2012. A new strategy to reduce allelic bias in RNA-Seq readmapping. *Nucleic Acids Res* **40**: e127.

Wang J, Wang W, Li R, Li Y, Tian G, Goodman L, Fan W, Zhang J, Li J, Zhang J, et al. 2008. The diploid genome sequence of an Asian individual. *Nature* **456**: 60–65.

Wang T, Antonacci-Fulton L, Howe K, Lawson HA, Lucas JK, Phillippy AM, Popejoy AB, Asri M, Carson C, Chaisson MJP, et al. 2022. The Human Pangenome Project: a global resource to map genomic diversity. *Nature* **604**: 437–446.

Wang Y, Guo T, Ke H, Zhang Q, Li S, Luo W, Qin Y. 2021. Pathogenic variants of meiotic double strand break (DSB) formation genes PRDM9 and ANKRD31 in premature ovarian insufficiency. *Genet Med* **23**: 2309–2315.

Weischenfeldt J, Symmons O, Spitz F, Korbel JO. 2013. Phenotypic impact of genomic structural variation: insights from and for human disease. *Nat Rev Genet* **14**: 125–138.

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* **37**: 1155–1162.

Wheeler DA, Srinivasan M, Egholm M, Shen Y, Chen L, McGuire A, He W, Chen Y-J, Makhijani V, Roth GT, et al. 2008. The complete genome of an individual by massively parallel DNA sequencing. *Nature* **452**: 872–876.

Wick RR, Judd LM, Holt KE. 2019. Performance of neural network basecalling tools for Oxford Nanopore sequencing. *Genome Biol* **20**: 129.

Woodward EL, Olsson ML, Johansson B, Paulsson K. 2014. Allelic variants of PRDM9 associated with high hyperdiploid childhood acute lymphoblastic leukaemia. *Br J Haematol* **166**: 947–949.

Wright PA. 2020. Killer-cell immunoglobulin-like receptor assessment algorithms in haemopoietic progenitor cell transplantation: current perspectives and future opportunities. *HLA* **95**: 435–448.

Wu TD, Nacu S. 2010. Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics* **26**: 873–881.

Xing J, Zhang Y, Han K, Salem AH, Sen SK, Huff CD, Zhou Q, Kirkness EF, Levy S, Batzer MA, et al. 2009. Mobile elements create structural variation: Analysis of a complete human genome. *Genome Res* **19**: 1516–1526.

Ye H, Meehan J, Tong W, Hong H. 2015. Alignment of Short Reads: A Crucial Step for Application of Next-Generation Sequencing Data in Precision Medicine. *Pharmaceutics* **7**: 523–541.

Zerbino DR, Birney E. 2008. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res* **18**: 821–829.

Zhang L, Zhou X, Weng Z, Sidow A. 2020. De novo diploid genome assembly for genome-wide structural variant detection. *NAR Genomics Bioinforma* **2**: lqz018.

Zhao M, Ma J, Li M, Zhang Y, Jiang B, Zhao X, Huai C, Shen L, Zhang N, He L, et al. 2021. Cytochrome P450 Enzymes and Drug Metabolism in Humans. *Int J Mol Sci* **22**: 12808.

Zheng Z, Li S, Su J, Leung AW-S, Lam T-W, Luo R. 2021. Symphonizing pileup and full-alignment for deep learning-based long-read variant calling. *bioRxiv* 2021.12.29.474431.

Zhou S-F, Liu J-P, Chowbay B. 2009. Polymorphism of human cytochrome P450 enzymes and its clinical impact. *Drug Metab Rev* **41**: 89–295.

Zhou W, Emery SB, Flasch DA, Wang Y, Kwan KY, Kidd JM, Moran JV, Mills RE. 2020. Identification and characterization of occult human-specific LINE-1 insertions using long-read sequencing technology. *Nucleic Acids Res* **48**: 1146–1163.

Zook JM, Catoe D, McDaniel J, Vang L, Spies N, Sidow A, Weng Z, Liu Y, Mason CE, Alexander N, et al. 2016. Extensive sequencing of seven human genomes to characterize benchmark reference materials. *Sci Data* **3**: 160025.

Appendix

Appendix Table 1: Populations from the 1000 Genomes Project. Abbreviations, names, and locations for populations from the 1000GP (Zook et al. 2016). *Note: The 1000GP refers to this continental population as "American", but in this thesis it is referred to as "Admixed American" for clarification as the populations are of admixed European, African, and Indigenous American ancestries. The Barbadian and African-American populations are admixed with African and European ancestry.

Continental population	Population abbreviation	Population name	Population origin
African (AFR)	ESN	Esan	Esan in Nigeria
	GWD	Gambian	Gambian in Western Division
	LWK	Luhya	Luhya in Webuye, Kenya
	MSL	Mende	Mende in Sierra Leone
	YRI	Yoruba	Yoruba in Ibadan, Nigeria
	ACB	Barbadian	African Caribbean in Barbados
	ASW	African-American SW	People with African Ancestry in Southwest USA
***Admixed American (AMR)**	CLM	Columbian	Colombians in Medellin, Colombia
	MXL	Mexican-American	People with Mexican Ancestry in Los Angeles, CA, USA
	PEL	Peruvian	Peruvians in Lima, Peru
	PUR	Puerto Rican	Puerto Ricans in Puerto Rico Puerto
East Asian (EAS)	CDX	Dai Chinese	Chinese Dai in Xishuangbanna, China
	CHB	Han Chinese	Han Chinese in Beijing, China
	CHS	Southern Han Chinese	Southern Han Chinese
	JPT	Japanese	Japanese in Tokyo, Japan
	KHV	Kinh Vietnamese	Kinh in Ho Chi Minh City, Vietnam
European (EUR)	CEU	CEPH	Utah residents (CEPH) with Northern and Western European ancestry
	GBR	British	British in England and Scotland
	FIN	Finnish	Finnish in Finland
	IBS	Spanish	Iberian Populations in Spain
	TSI	Tuscan	Toscani in Italia
South Asian (SAS)	BEB	Bengali	Bengali in Bangladesh
	GIH	Gujarati	Gujarati Indians in Houston, TX, USA
	ITU	Telugu	Indian Telugu in the UK
	PJL	Punjabi	Punjabi in Lahore, Pakistan
	STU	Tamil	Sri Lankan Tamil in the UK

Appendix Table 2: Maximum F1 scores for haploid allele calling and diploid genotype calling using the *k*-mer count, distance, and combined short-read models. F1 scores were averaged across all replicates of all *PRDM9*-36 alleles from the primary haploid simulation set. *k* range indicates the value(s) at which the maximum average F1 score was obtained. Results are for the best method for each model (count-coverage, distance-max, and combined count-coverage & distance-max).

Model and Ploidy	Coverage	Error	Maximum average F1 score	*k* range
Count haploid	20X	0%	0.9794	83
		0.1%	0.9744	83
		1%	0.9303	63
	40X	0%	0.9964	91
		0.1%	0.9942	87
		1%	0.9808	67
	60X	0%	0.9989	83–91
		0.1%	0.9992	91
		1%	0.9892	63
	80X	0%	0.9994	83–95
		0.1%	0.9994	83–95
		1%	0.9950	75
	100X	0%	0.9997	75–95
		0.1%	0.9997	83–87
		1%	0.9981	71–75
Count diploid	20X	0%	0.5209	75
		0.1%	0.4968	75
		1%	0.3357	39
	40X	0%	0.6887	83
		0.1%	0.6792	79
		1%	0.5438	67
	60X	0%	0.7259	87
		0.1%	0.7218	83
		1%	0.6424	75
	80X	0%	0.7399	87
		0.1%	0.7376	87
		1%	0.6879	75

	100X	0%	0.7480	95
		0.1%	0.7455	91
		1%	0.7100	79
Distance haploid	20X	0%	0.9812	99
		0.1%	0.9775	99
		1%	0.8323	63
	40X	0%	0.9860	91
		0.1%	0.9847	99
		1%	0.9330	75
	60X	0%	0.9872	91–99
		0.1%	0.9864	99
		1%	0.9581	91
	80X	0%	0.9878	99
		0.1%	0.9870	99
		1%	0.9688	75
	100X	0%	0.9878	99
		0.1%	0.9881	99
		1%	0.9749	75
Distance diploid	20X	0%	0.6701	99
		0.10%	0.6526	99
		1%	0.2651	71
	40X	0%	0.7008	99
		0.10%	0.6946	99
		1%	0.4319	99
	60X	0%	0.7108	99
		0.10%	0.7067	99
		1%	0.5417	87
	80X	0%	0.7154	99
		0.10%	0.7125	99
		1%	0.5908	87
	100X	0%	0.7184	99
		0.10%	0.7161	99
		1%	0.6194	91

Combined haploid	20X	0%	0.9925	95
		0.10%	0.9903	95
		1%	0.9414	67
	40X	0%	0.9997	95
		0.10%	0.9994	95
		1%	0.9839	75
	60X	0%	1.0000	99
		0.10%	1.0000	99
		1%	0.9906	71, 79
	80X	0%	1.0000	95–99
		0.10%	1.0000	99
		1%	0.9967	87
	100X	0%	1.0000	79–99
		0.10%	1.0000	95–99
		1%	0.9983	71–75
Combined diploid	20X	0%	0.6557	87
		0.10%	0.6309	87
		1%	0.3617	39
	40X	0%	0.7399	95
		0.10%	0.7323	95
		1%	0.5827	75
	60X	0%	0.7527	95
		0.10%	0.7503	95
		1%	0.6683	79
	80X	0%	0.7573	95
		0.10%	0.7558	95
		1%	0.7052	79
	100X	0%	0.7620	99
		0.10%	0.7609	99
		1%	0.7236	79

Appendix Table 3: URLS for publicly available GIAB Ashkenazi trio files used in analyses. All URLs link to GRCh38-aligned bams.

Sample ID	Dataset	Read length	Technology	File URL
HG002	GIAB Ashkenazi	2x150bp	Illumina	https://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/giab/data/AshkenazimTrio/HG002_NA24385_son/NIST_HiSeq_HG002_Homogeneity-10953946/NHGRI_Illumina300X_AJtrio_novoalign_bams/HG002.GRCh38.300x.bam
	GIAB Ashkenazi	2x250bp	Illumina	https://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/giab/data/AshkenazimTrio/HG002_NA24385_son/NIST_Illumina_2x250bps/novoalign_bams/HG002.GRCh38.2x250.bam
	GIAB Ashkenazi	15–20kb	PacBio	https://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/giab/data/AshkenazimTrio/HG002_NA24385_son/PacBio_CCS_15kb_20kb_chemistry2/GRCh38/HG002.SequelII.merged_15kb_20kb.GRCh38.duplomap.bam
HG003	GIAB Ashkenazi	2x150bp	Illumina	https://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/giab/data/AshkenazimTrio/HG003_NA24149_father/NIST_HiSeq_HG003_Homogeneity-12389378/NHGRI_Illumina300X_AJtrio_novoalign_bams/HG003.GRCh38.300x.bam
	GIAB Ashkenazi	2x250bp	Illumina	https://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/giab/data/AshkenazimTrio/HG003_NA24149_father/NIST_Illumina_2x250bps/novoalign_bams/HG003.GRCh38.2x250.bam
	GIAB Ashkenazi	15–20kb	PacBio	https://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/giab/data/AshkenazimTrio/HG003_NA24149_father/PacBio_CCS_15kb_20kb_chemistry2/GRCh38/HG003.GRCh38.consensusalignments.bam
HG004	GIAB Ashkenazi	2x150bp	Illumina	https://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/giab/data/AshkenazimTrio/HG004_NA24143_mother/NIST_HiSeq_HG004_Homogeneity-14572558/NHGRI_Illumina300X_AJtrio_novoalign_bams/HG004.GRCh38.300x.bam
	GIAB Ashkenazi	2x250bp	Illumina	https://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/giab/data/AshkenazimTrio/HG004_NA24143_mother/NIST_Illumina_2x250bps/novoalign_bams/HG004.GRCh38.2x250.bam
	GIAB Ashkenazi	15–20kb	PacBio	https://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/giab/data/AshkenazimTrio/HG004_NA24143_mother/PacBio_CCS_15kb_20kb_chemistry2/GRCh38/HG004.GRCh38.consensusalignments.bam

Appendix Table 4: URLS for publicly available HPRC++ sample files used in analyses. All URLs link to GRCh38-aligned bams expect for: HG00514, HG00731, HG00732, HG02970 and NA19030 (PacBio HiFi unmapped bams), and HG002, HG005, and NA21309 (Illumina fastqs). The HPRC_PLUS cohort refers to samples not sequenced by but provided by the HPRC.

Sample ID	Dataset	Cohort	Technology	File URL
HG001	HPRC++	GIAB	Illumina	https://ftp.sra.ebi.ac.uk/vol1/run/ERR323/ERR3239334/NA12878.final.cram
	HPRC++	GIAB	PacBio	https://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/giab/data/NA12878/PacBio_SequelII_CCS_11kb/HG001_GRCh38/HG001_GRCh38.haplotag.RTG.trio.bam
HG002	HPRC++	HPRC_PLUS	Illumina	https://s3-us-west-2.amazonaws.com/human-pangenomics/index.html?prefix=working/HPRC_PLUS/HG002/raw_data/Illumina/child/HG002_HiSeq30x_subsampled_R1.fastq.gz
	HPRC++	HPRC_PLUS	Illumina	https://s3-us-west-2.amazonaws.com/human-pangenomics/index.html?prefix=working/HPRC_PLUS/HG002/raw_data/Illumina/child/HG002_HiSeq30x_subsampled_R2.fastq.gz
	HPRC++	HPRC_PLUS	PacBio	https://s3-us-west-2.amazonaws.com/human-pangenomics/index.html?prefix=working/HPRC_PLUS/HG002/analysis/aligned_reads/hifi/GRCh38/HG002_aligned_GRCh38_winnowmap.sorted.bam
HG003	HPRC++	GIAB	Illumina	https://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/giab/data/AshkenazimTrio/HG003_NA24149_father/NIST_HiSeq_HG003_Homogeneity-12389378/NHGRI_Illumina300X_AJtrio_novoalign_bams/
	HPRC++	GIAB	PacBio	https://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/giab/data/AshkenazimTrio/HG003_NA24149_father/PacBio_CCS_15kb_20kb_chemistry2/GRCh38/HG003.GRCh38.consensusalignments.bam
HG004	HPRC++	GIAB	Illumina	https://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/giab/data/AshkenazimTrio/HG004_NA24143_mother/NIST_HiSeq_HG004_Homogeneity-14572558/NHGRI_Illumina300X_AJtrio_novoalign_bams/
	HPRC++	GIAB	PacBio	https://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/giab/data/AshkenazimTrio/HG004_NA24143_mother/PacBio_CCS_15kb_20kb_chemistry2/GRCh38/HG004.GRCh38.consensusalignments.bam
HG005	HPRC++	HPRC_PLUS	Illumina	https://s3-us-west-2.amazonaws.com/human-pangenomics/working/HPRC_PLUS/HG005/raw_data/Illumina/child/5A1-24481579/5A1_S5_L001_R1_001.fastq.gz
	HPRC++	HPRC_PLUS	Illumina	https://s3-us-west-2.amazonaws.com/human-pangenomics/working/HPRC_PLUS/HG005/raw_data/Illumina/child/5A1-24481579/5A1_S5_L001_R2_001.fastq.gz
	HPRC++	HPRC_PLUS	Illumina	https://s3-us-west-2.amazonaws.com/human-pangenomics/working/HPRC_PLUS/HG005/raw_data/Illumina/child/5A1-24481579/5A1_S5_L002_R1_001.fastq.gz
	HPRC++	HPRC_PLUS	Illumina	https://s3-us-west-2.amazonaws.com/human-pangenomics/working/HPRC_PLUS/HG005/raw_data/Illumina/child/5A1-24481579/5A1_S5_L002_R2_001.fastq.gz
	HPRC++	HPRC_PLUS	Illumina	https://s3-us-west-2.amazonaws.com/human-pangenomics/working/HPRC_PLUS/HG005/raw_data/Illumina/child/5A2-24481580/5A2_S6_L001_R1_001.fastq.gz
	HPRC++	HPRC_PLUS	Illumina	https://s3-us-west-2.amazonaws.com/human-pangenomics/working/HPRC_PLUS/HG005/raw_data/Illumina/child/5A2-24481580/5A2_S6_L001_R2_001.fastq.gz
	HPRC++	HPRC_PLUS	Illumina	https://s3-us-west-2.amazonaws.com/human-pangenomics/working/HPRC_PLUS/HG005/raw_data/Illumina/child/5A2-24481580/5A2_S6_L002_R1_001.fastq.gz
	HPRC++	HPRC_PLUS	Illumina	https://s3-us-west-2.amazonaws.com/human-pangenomics/working/HPRC_PLUS/HG005/raw_data/Illumina/child/5A2-24481580/5A2_S6_L002_R2_001.fastq.gz
	HPRC++	HPRC_PLUS	Illumina	https://s3-us-west-2.amazonaws.com/human-pangenomics/working/HPRC_PLUS/HG005/raw_data/Illumina/child/5B1-24481581/5B1_S7_L001_R1_001.fastq.gz

HPRC++	HPRC_PLUS	Illumina	https://s3-us-west-2.amazonaws.com/human-pangenomics/working/HPRC_PLUS/HG005/raw_data/Illumina/child/5B1-24481581/5B1_S7_L001_R2_001.fastq.gz
HPRC++	HPRC_PLUS	Illumina	https://s3-us-west-2.amazonaws.com/human-pangenomics/working/HPRC_PLUS/HG005/raw_data/Illumina/child/5B1-24481581/5B1_S7_L002_R1_001.fastq.gz
HPRC++	HPRC_PLUS	Illumina	https://s3-us-west-2.amazonaws.com/human-pangenomics/working/HPRC_PLUS/HG005/raw_data/Illumina/child/5B1-24481581/5B1_S7_L002_R2_001.fastq.gz
HPRC++	HPRC_PLUS	Illumina	https://s3-us-west-2.amazonaws.com/human-pangenomics/working/HPRC_PLUS/HG005/raw_data/Illumina/child/5B2-24481582/5B2_S8_L001_R1_001.fastq.gz
HPRC++	HPRC_PLUS	Illumina	https://s3-us-west-2.amazonaws.com/human-pangenomics/working/HPRC_PLUS/HG005/raw_data/Illumina/child/5B2-24481582/5B2_S8_L001_R2_001.fastq.gz
HPRC++	HPRC_PLUS	Illumina	https://s3-us-west-2.amazonaws.com/human-pangenomics/working/HPRC_PLUS/HG005/raw_data/Illumina/child/5B2-24481582/5B2_S8_L002_R1_001.fastq.gz
HPRC++	HPRC_PLUS	Illumina	https://s3-us-west-2.amazonaws.com/human-pangenomics/working/HPRC_PLUS/HG005/raw_data/Illumina/child/5B2-24481582/5B2_S8_L002_R2_001.fastq.gz
HPRC++	HPRC_PLUS	Illumina	https://s3-us-west-2.amazonaws.com/human-pangenomics/working/HPRC_PLUS/HG005/raw_data/Illumina/child/5C1-24481583/5C1_S9_L001_R1_001.fastq.gz
HPRC++	HPRC_PLUS	Illumina	https://s3-us-west-2.amazonaws.com/human-pangenomics/working/HPRC_PLUS/HG005/raw_data/Illumina/child/5C1-24481583/5C1_S9_L001_R2_001.fastq.gz
HPRC++	HPRC_PLUS	Illumina	https://s3-us-west-2.amazonaws.com/human-pangenomics/working/HPRC_PLUS/HG005/raw_data/Illumina/child/5C1-24481583/5C1_S9_L002_R1_001.fastq.gz
HPRC++	HPRC_PLUS	Illumina	https://s3-us-west-2.amazonaws.com/human-pangenomics/working/HPRC_PLUS/HG005/raw_data/Illumina/child/5C1-24481583/5C1_S9_L002_R2_001.fastq.gz
HPRC++	HPRC_PLUS	Illumina	https://s3-us-west-2.amazonaws.com/human-pangenomics/working/HPRC_PLUS/HG005/raw_data/Illumina/child/5C2-24481584/5C2_S10_L001_R1_001.fastq.gz
HPRC++	HPRC_PLUS	Illumina	https://s3-us-west-2.amazonaws.com/human-pangenomics/working/HPRC_PLUS/HG005/raw_data/Illumina/child/5C2-24481584/5C2_S10_L001_R2_001.fastq.gz
HPRC++	HPRC_PLUS	Illumina	https://s3-us-west-2.amazonaws.com/human-pangenomics/working/HPRC_PLUS/HG005/raw_data/Illumina/child/5C2-24481584/5C2_S10_L002_R1_001.fastq.gz
HPRC++	HPRC_PLUS	Illumina	https://s3-us-west-2.amazonaws.com/human-pangenomics/working/HPRC_PLUS/HG005/raw_data/Illumina/child/5C2-24481584/5C2_S10_L002_R2_001.fastq.gz
HPRC++	HPRC_PLUS	Illumina	https://s3-us-west-2.amazonaws.com/human-pangenomics/working/HPRC_PLUS/HG005/raw_data/Illumina/child/5D1-24481585/5D1_S11_L001_R1_001.fastq.gz
HPRC++	HPRC_PLUS	Illumina	https://s3-us-west-2.amazonaws.com/human-pangenomics/working/HPRC_PLUS/HG005/raw_data/Illumina/child/5D1-24481585/5D1_S11_L001_R2_001.fastq.gz
HPRC++	HPRC_PLUS	Illumina	https://s3-us-west-2.amazonaws.com/human-pangenomics/working/HPRC_PLUS/HG005/raw_data/Illumina/child/5D1-24481585/5D1_S11_L002_R1_001.fastq.gz
HPRC++	HPRC_PLUS	Illumina	https://s3-us-west-2.amazonaws.com/human-pangenomics/working/HPRC_PLUS/HG005/raw_data/Illumina/child/5D1-24481585/5D1_S11_L002_R2_001.fastq.gz
HPRC++	HPRC_PLUS	Illumina	https://s3-us-west-2.amazonaws.com/human-pangenomics/working/HPRC_PLUS/HG005/raw_data/Illumina/child/5D2-24481586/5D2_S12_L001_R1_001.fastq.gz
HPRC++	HPRC_PLUS	Illumina	https://s3-us-west-2.amazonaws.com/human-pangenomics/working/HPRC_PLUS/HG005/raw_data/Illumina/child/5D2-24481586/5D2_S12_L001_R2_001.fastq.gz
HPRC++	HPRC_PLUS	Illumina	https://s3-us-west-2.amazonaws.com/human-pangenomics/working/HPRC_PLUS/HG005/raw_data/Illumina/child/5D2-24481586/5D2_S12_L002_R1_001.fastq.gz
HPRC++	HPRC_PLUS	Illumina	https://s3-us-west-2.amazonaws.com/human-pangenomics/working/HPRC_PLUS/HG005/raw_data/Illumina/child/5D2-24481586/5D2_S12_L002_R2_001.fastq.gz

	HPRC++	HPRC_PLUS	Illumina	https://s3-us-west-2.amazonaws.com/human-pangenomics/working/HPRC_PLUS/HG005/raw_data/Illumina/child/5F1-24481575/5F1_S1_L001_R1_001.fastq.gz
	HPRC++	HPRC_PLUS	Illumina	https://s3-us-west-2.amazonaws.com/human-pangenomics/working/HPRC_PLUS/HG005/raw_data/Illumina/child/5F1-24481575/5F1_S1_L001_R2_001.fastq.gz
	HPRC++	HPRC_PLUS	Illumina	https://s3-us-west-2.amazonaws.com/human-pangenomics/working/HPRC_PLUS/HG005/raw_data/Illumina/child/5F1-24481575/5F1_S1_L002_R1_001.fastq.gz
	HPRC++	HPRC_PLUS	Illumina	https://s3-us-west-2.amazonaws.com/human-pangenomics/working/HPRC_PLUS/HG005/raw_data/Illumina/child/5F1-24481575/5F1_S1_L002_R2_001.fastq.gz
	HPRC++	HPRC_PLUS	Illumina	https://s3-us-west-2.amazonaws.com/human-pangenomics/working/HPRC_PLUS/HG005/raw_data/Illumina/child/5F2-24481576/5F2_S2_L001_R1_001.fastq.gz
	HPRC++	HPRC_PLUS	Illumina	https://s3-us-west-2.amazonaws.com/human-pangenomics/working/HPRC_PLUS/HG005/raw_data/Illumina/child/5F2-24481576/5F2_S2_L001_R2_001.fastq.gz
	HPRC++	HPRC_PLUS	Illumina	https://s3-us-west-2.amazonaws.com/human-pangenomics/working/HPRC_PLUS/HG005/raw_data/Illumina/child/5F2-24481576/5F2_S2_L002_R1_001.fastq.gz
	HPRC++	HPRC_PLUS	Illumina	https://s3-us-west-2.amazonaws.com/human-pangenomics/working/HPRC_PLUS/HG005/raw_data/Illumina/child/5F2-24481576/5F2_S2_L002_R2_001.fastq.gz
	HPRC++	HPRC_PLUS	Illumina	https://s3-us-west-2.amazonaws.com/human-pangenomics/working/HPRC_PLUS/HG005/raw_data/Illumina/child/5L1-24481577/5L1_S3_L001_R1_001.fastq.gz
	HPRC++	HPRC_PLUS	Illumina	https://s3-us-west-2.amazonaws.com/human-pangenomics/working/HPRC_PLUS/HG005/raw_data/Illumina/child/5L1-24481577/5L1_S3_L001_R2_001.fastq.gz
	HPRC++	HPRC_PLUS	Illumina	https://s3-us-west-2.amazonaws.com/human-pangenomics/working/HPRC_PLUS/HG005/raw_data/Illumina/child/5L1-24481577/5L1_S3_L002_R1_001.fastq.gz
	HPRC++	HPRC_PLUS	Illumina	https://s3-us-west-2.amazonaws.com/human-pangenomics/working/HPRC_PLUS/HG005/raw_data/Illumina/child/5L1-24481577/5L1_S3_L002_R2_001.fastq.gz
	HPRC++	HPRC_PLUS	Illumina	https://s3-us-west-2.amazonaws.com/human-pangenomics/working/HPRC_PLUS/HG005/raw_data/Illumina/child/5L2-24481578/5L2_S4_L001_R1_001.fastq.gz
	HPRC++	HPRC_PLUS	Illumina	https://s3-us-west-2.amazonaws.com/human-pangenomics/working/HPRC_PLUS/HG005/raw_data/Illumina/child/5L2-24481578/5L2_S4_L001_R2_001.fastq.gz
	HPRC++	HPRC_PLUS	Illumina	https://s3-us-west-2.amazonaws.com/human-pangenomics/working/HPRC_PLUS/HG005/raw_data/Illumina/child/5L2-24481578/5L2_S4_L002_R1_001.fastq.gz
	HPRC++	HPRC_PLUS	Illumina	https://s3-us-west-2.amazonaws.com/human-pangenomics/working/HPRC_PLUS/HG005/raw_data/Illumina/child/5L2-24481578/5L2_S4_L002_R2_001.fastq.gz
	HPRC++	HPRC_PLUS	PacBio	https://s3-us-west-2.amazonaws.com/human-pangenomics/working/HPRC_PLUS/HG005/analysis/aligned_reads/hifi/GRCh38/HG005_aligned_GRCh38_winnowmap.sorted.bam
HG006	HPRC++	GIAB	Illumina	https://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/giab/data/ChineseTrio/HG006_NA24694-huCA017E_father/NA24694_Father_HiSeq100x/NHGRI_Illumina100X_Chinesetrio_novoalign_bams/HG006.GRCh38_full_plus_hs38d1_analysis_set_minus_alts.100x.bam
	HPRC++	GIAB	PacBio	https://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/giab/data/ChineseTrio/HG006_NA24694-huCA017E_father/PacBio_MtSinai/PacBio_minimap2_bam/HG006_PacBio_GRCh38.bam
HG007	HPRC++	GIAB	Illumina	https://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/giab/data/ChineseTrio/HG007_NA24695-hu38168_mother/NA24695_Mother_HiSeq100x/NHGRI_Illumina100X_Chinesetrio_novoalign_bams/HG007.GRCh38_full_plus_hs38d1_analysis_set_minus_alts.100x.bam
	HPRC++	GIAB	PacBio	https://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/giab/data/ChineseTrio/HG

				007_NA24695-hu38168_mother/PacBio_MtSinai/PacBio_minimap2_bam/HG007_PacBio_GRCh38.bam
HG00438	HPRC++	HPRC	Illumina	https://s3-us-west-2.amazonaws.com/human-pangenomics/working/HPRC/HG00438/raw_data/Illumina/child/HG00438.final.cram
	HPRC++	HPRC	PacBio	https://s3-us-west-2.amazonaws.com/human-pangenomics/working/HPRC/HG00438/analysis/aligned_reads/hifi/GRCh38/HG00438_aligned_GRCh38_winnowmap.sorted.bam
HG00514	HPRC++	1000GP-SV	Illumina	https://ftp.sra.ebi.ac.uk/vol1/run/ERR398/ERR3988781/HG00514.final.cram
	HPRC++	1000GP-SV	PacBio	https://ftp.sra.ebi.ac.uk/vol1/run/ERR498/ERR4982328/HG00514-hifi-r54329U_20200717_234302-A01.bam
	HPRC++	1000GP-SV	PacBio	https://ftp.sra.ebi.ac.uk/vol1/run/ERR498/ERR4982327/HG00514-hifi-r54329U_20200715_193257-A01.bam
	HPRC++	1000GP-SV	PacBio	https://ftp.sra.ebi.ac.uk/vol1/run/ERR498/ERR4982329/HG00514-hifi-r54329U_20200717_234302-B01.bam
HG00621	HPRC++	HPRC	Illumina	https://s3-us-west-2.amazonaws.com/human-pangenomics/working/HPRC/HG00621/raw_data/Illumina/child/HG00621.final.cram
	HPRC++	HPRC	PacBio	https://s3-us-west-2.amazonaws.com/human-pangenomics/working/HPRC/HG00621/analysis/aligned_reads/hifi/GRCh38/HG00621_aligned_GRCh38_winnowmap.sorted.bam
HG00673	HPRC++	HPRC	Illumina	https://s3-us-west-2.amazonaws.com/human-pangenomics/working/HPRC/HG00673/raw_data/Illumina/child/HG00673.final.cram
	HPRC++	HPRC	PacBio	https://s3-us-west-2.amazonaws.com/human-pangenomics/working/HPRC/HG00673/analysis/aligned_reads/hifi/GRCh38/HG00673_aligned_GRCh38_winnowmap.sorted.bam
HG00731	HPRC++	1000GP-SV	Illumina	https://ftp.sra.ebi.ac.uk/vol1/run/ERR324/ERR3241754/HG00731.final.cram
	HPRC++	1000GP-SV	PacBio	https://ftp.sra.ebi.ac.uk/vol1/run/ERR384/ERR3840018/HG00731-hifi-r54329U_20190531_173856-A01.bam
	HPRC++	1000GP-SV	PacBio	https://ftp.sra.ebi.ac.uk/vol1/run/ERR384/ERR3840021/HG00731-hifi-r54329U_20190906_202809-A01.bam
	HPRC++	1000GP-SV	PacBio	https://ftp.sra.ebi.ac.uk/vol1/run/ERR384/ERR3840020/HG00731-hifi-r54329U_20190604_223930-B01.bam
	HPRC++	1000GP-SV	PacBio	https://ftp.sra.ebi.ac.uk/vol1/run/ERR384/ERR3840017/HG00731-hifi-r54329U_20190528_230410-A01.bam
	HPRC++	1000GP-SV	PacBio	https://ftp.sra.ebi.ac.uk/vol1/run/ERR384/ERR3840019/HG00731-hifi-r54329U_20190531_173856-B01.bam
HG00732	HPRC++	1000GP-SV	Illumina	https://ftp.sra.ebi.ac.uk/vol1/run/ERR324/ERR3241755/HG00732.final.cram
	HPRC++	1000GP-SV	PacBio	https://ftp.sra.ebi.ac.uk/vol1/run/ERR386/ERR3861391/HG00732-hifi-r54329U_20190607_183639-E01.bam
	HPRC++	1000GP-SV	PacBio	https://ftp.sra.ebi.ac.uk/vol1/run/ERR386/ERR3861392/HG00732-hifi-r54329U_20190703_201135-C01.bam
	HPRC++	1000GP-SV	PacBio	https://ftp.sra.ebi.ac.uk/vol1/run/ERR498/ERR4987503/HG00732-hifi-r54329U_20200528_195409-A01.bam
	HPRC++	1000GP-SV	PacBio	https://ftp.sra.ebi.ac.uk/vol1/run/ERR386/ERR3861390/HG00732-hifi-r54329U_20190607_183639-D01.bam
	HPRC++	1000GP-SV	PacBio	https://ftp.sra.ebi.ac.uk/vol1/run/ERR386/ERR3861388/HG00732-hifi-r54329U_20190604_223930-A01.bam
	HPRC++	1000GP-SV	PacBio	https://ftp.sra.ebi.ac.uk/vol1/run/ERR386/ERR3861393/HG00732-hifi-r54329U_20190830_234003-B01.bam
	HPRC++	1000GP-SV	PacBio	https://ftp.sra.ebi.ac.uk/vol1/run/ERR498/ERR4987504/HG00732-hifi-r64076_20200601_233545-A01.bam
	HPRC++	1000GP-SV	PacBio	https://ftp.sra.ebi.ac.uk/vol1/run/ERR386/ERR3861389/HG00732-hifi-r54329U_20190607_183639-C01.bam
	HPRC++	1000GP-SV	PacBio	https://ftp.sra.ebi.ac.uk/vol1/run/ERR498/ERR4987505/HG00732-hifi-r64076_20200601_233545-B01.bam
HG00733	HPRC++	HPRC_PLUS	Illumina	https://s3-us-west-2.amazonaws.com/human-pangenomics/working/HPRC_PLUS/HG00733/raw_data/Illumina/child/HG00733.final.cram
	HPRC++	HPRC_PLUS	PacBio	https://s3-us-west-2.amazonaws.com/human-pangenomics/working/HPRC_PLUS/HG00733/analysis/aligned_reads/hifi/GRCh38/HG00733_aligned_GRCh38_winnowmap.sorted.bam

HG00735	HPRC++	HPRC	Illumina	https://s3-us-west-2.amazonaws.com/human-pangenomics/working/HPRC/HG00735/raw_data/Illumina/child/HG00735.final.cram
	HPRC++	HPRC	PacBio	https://s3-us-west-2.amazonaws.com/human-pangenomics/working/HPRC/HG00735/analysis/aligned_reads/hifi/GRCh38/HG00735_aligned_GRCh38_winnowmap.sorted.bam
HG00741	HPRC++	HPRC	Illumina	https://s3-us-west-2.amazonaws.com/human-pangenomics/working/HPRC/HG00741/raw_data/Illumina/child/HG00741.final.cram
	HPRC++	HPRC	PacBio	https://s3-us-west-2.amazonaws.com/human-pangenomics/working/HPRC/HG00741/analysis/aligned_reads/hifi/GRCh38/HG00741_aligned_GRCh38_winnowmap.sorted.bam
HG01071	HPRC++	HPRC	Illumina	https://s3-us-west-2.amazonaws.com/human-pangenomics/working/HPRC/HG01071/raw_data/Illumina/child/HG01071.final.cram
	HPRC++	HPRC	PacBio	https://s3-us-west-2.amazonaws.com/human-pangenomics/working/HPRC/HG01071/analysis/aligned_reads/hifi/GRCh38/HG01071_aligned_GRCh38_winnowmap.sorted.bam
HG01106	HPRC++	HPRC	Illumina	https://s3-us-west-2.amazonaws.com/human-pangenomics/working/HPRC/HG01106/raw_data/Illumina/child/HG01106.final.cram
	HPRC++	HPRC	PacBio	https://s3-us-west-2.amazonaws.com/human-pangenomics/working/HPRC/HG01106/analysis/aligned_reads/hifi/GRCh38/HG01106_aligned_GRCh38_winnowmap.sorted.bam
HG01109	HPRC++	HPRC_PLUS	Illumina	https://s3-us-west-2.amazonaws.com/human-pangenomics/working/HPRC_PLUS/HG01109/raw_data/Illumina/child/HG01109.final.cram
	HPRC++	HPRC_PLUS	PacBio	https://s3-us-west-2.amazonaws.com/human-pangenomics/working/HPRC_PLUS/HG01109/analysis/aligned_reads/hifi/GRCh38/HG01109_aligned_GRCh38_winnowmap.sorted.bam
HG01175	HPRC++	HPRC	Illumina	https://s3-us-west-2.amazonaws.com/human-pangenomics/working/HPRC/HG01175/raw_data/Illumina/child/HG01175.final.cram
	HPRC++	HPRC	PacBio	https://s3-us-west-2.amazonaws.com/human-pangenomics/working/HPRC/HG01175/analysis/aligned_reads/hifi/GRCh38/HG01175_aligned_GRCh38_winnowmap.sorted.bam
HG01243	HPRC++	HPRC_PLUS	Illumina	https://s3-us-west-2.amazonaws.com/human-pangenomics/working/HPRC_PLUS/HG01243/raw_data/Illumina/child/HG01243.final.cram
	HPRC++	HPRC_PLUS	PacBio	https://s3-us-west-2.amazonaws.com/human-pangenomics/working/HPRC_PLUS/HG01243/analysis/aligned_reads/hifi/GRCh38/HG01243_aligned_GRCh38_winnowmap.sorted.bam
HG01258	HPRC++	HPRC	Illumina	https://s3-us-west-2.amazonaws.com/human-pangenomics/working/HPRC/HG01258/raw_data/Illumina/child/HG01258.final.cram
	HPRC++	HPRC	PacBio	https://s3-us-west-2.amazonaws.com/human-pangenomics/working/HPRC/HG01258/analysis/aligned_reads/hifi/GRCh38/HG01258_aligned_GRCh38_winnowmap.sorted.bam
HG01358	HPRC++	HPRC	Illumina	https://s3-us-west-2.amazonaws.com/human-pangenomics/working/HPRC/HG01358/raw_data/Illumina/child/HG01358.final.cram
	HPRC++	HPRC	PacBio	https://s3-us-west-2.amazonaws.com/human-pangenomics/working/HPRC/HG01358/analysis/aligned_reads/hifi/GRCh38/HG01358_aligned_GRCh38_winnowmap.sorted.bam
HG01361	HPRC++	HPRC	Illumina	https://s3-us-west-2.amazonaws.com/human-pangenomics/working/HPRC/HG01361/raw_data/Illumina/child/HG01361.final.cram
	HPRC++	HPRC	PacBio	https://s3-us-west-2.amazonaws.com/human-pangenomics/working/HPRC/HG01361/analysis/aligned_reads/hifi/GRCh38/HG01361_aligned_GRCh38_winnowmap.sorted.bam
HG01891	HPRC++	HPRC	Illumina	https://s3-us-west-2.amazonaws.com/human-pangenomics/working/HPRC/HG01891/raw_data/Illumina/child/HG01891.final.cram
	HPRC++	HPRC	PacBio	https://s3-us-west-2.amazonaws.com/human-pangenomics/working/HPRC/HG01891/analysis/aligned_reads/hifi/GRCh38/HG01891_aligned_GRCh38_winnowmap.sorted.bam

HG01928	HPRC++	HPRC	Illumina	https://s3-us-west-2.amazonaws.com/human-pangenomics/working/HPRC/HG01928/raw_data/Illumina/child/HG01928.final.cram
	HPRC++	HPRC	PacBio	https://s3-us-west-2.amazonaws.com/human-pangenomics/working/HPRC/HG01928/analysis/aligned_reads/hifi/GRCh38/HG01928_aligned_GRCh38_winnowmap.sorted.bam
HG01952	HPRC++	HPRC	Illumina	https://s3-us-west-2.amazonaws.com/human-pangenomics/working/HPRC/HG01952/raw_data/Illumina/child/HG01952.final.cram
	HPRC++	HPRC	PacBio	https://s3-us-west-2.amazonaws.com/human-pangenomics/working/HPRC/HG01952/analysis/aligned_reads/hifi/GRCh38/HG01952_aligned_GRCh38_winnowmap.sorted.bam
HG01978	HPRC++	HPRC	Illumina	https://s3-us-west-2.amazonaws.com/human-pangenomics/working/HPRC/HG01978/raw_data/Illumina/child/HG01978.final.cram
	HPRC++	HPRC	PacBio	https://s3-us-west-2.amazonaws.com/human-pangenomics/working/HPRC/HG01978/analysis/aligned_reads/hifi/GRCh38/HG01978_aligned_GRCh38_winnowmap.sorted.bam
HG02055	HPRC++	HPRC_PLUS	Illumina	https://s3-us-west-2.amazonaws.com/human-pangenomics/working/HPRC_PLUS/HG02055/raw_data/Illumina/child/HG02055.final.cram
	HPRC++	HPRC_PLUS	PacBio	https://s3-us-west-2.amazonaws.com/human-pangenomics/working/HPRC_PLUS/HG02055/analysis/aligned_reads/hifi/GRCh38/HG02055_aligned_GRCh38_winnowmap.sorted.bam
HG02080	HPRC++	HPRC_PLUS	Illumina	https://s3-us-west-2.amazonaws.com/human-pangenomics/working/HPRC_PLUS/HG02080/raw_data/Illumina/child/HG02080.final.cram
	HPRC++	HPRC_PLUS	PacBio	https://s3-us-west-2.amazonaws.com/human-pangenomics/working/HPRC_PLUS/HG02080/analysis/aligned_reads/hifi/GRCh38/HG02080_aligned_GRCh38_winnowmap.sorted.bam
HG02145	HPRC++	HPRC_PLUS	Illumina	https://s3-us-west-2.amazonaws.com/human-pangenomics/working/HPRC_PLUS/HG02145/raw_data/Illumina/child/HG02145.final.cram
	HPRC++	HPRC_PLUS	PacBio	https://s3-us-west-2.amazonaws.com/human-pangenomics/working/HPRC_PLUS/HG02145/analysis/aligned_reads/hifi/GRCh38/HG02145_aligned_GRCh38_winnowmap.sorted.bam
HG02148	HPRC++	HPRC	Illumina	https://s3-us-west-2.amazonaws.com/human-pangenomics/working/HPRC/HG02148/raw_data/Illumina/child/HG02148.final.cram
	HPRC++	HPRC	PacBio	https://s3-us-west-2.amazonaws.com/human-pangenomics/working/HPRC/HG02148/analysis/aligned_reads/hifi/GRCh38/HG02148_aligned_GRCh38_winnowmap.sorted.bam
HG02257	HPRC++	HPRC	Illumina	https://s3-us-west-2.amazonaws.com/human-pangenomics/working/HPRC/HG02257/raw_data/Illumina/child/HG02257.final.cram
	HPRC++	HPRC	PacBio	https://s3-us-west-2.amazonaws.com/human-pangenomics/working/HPRC/HG02257/analysis/aligned_reads/hifi/GRCh38/HG02257_aligned_GRCh38_winnowmap.sorted.bam
HG02572	HPRC++	HPRC	Illumina	https://s3-us-west-2.amazonaws.com/human-pangenomics/working/HPRC/HG02572/raw_data/Illumina/child/HG02572.final.cram
	HPRC++	HPRC	PacBio	https://s3-us-west-2.amazonaws.com/human-pangenomics/working/HPRC/HG02572/analysis/aligned_reads/hifi/GRCh38/HG02572_aligned_GRCh38_winnowmap.sorted.bam
HG02622	HPRC++	HPRC	Illumina	https://s3-us-west-2.amazonaws.com/human-pangenomics/working/HPRC/HG02622/raw_data/Illumina/child/HG02622.final.cram
	HPRC++	HPRC	PacBio	https://s3-us-west-2.amazonaws.com/human-pangenomics/working/HPRC/HG02622/analysis/aligned_reads/hifi/GRCh38/HG02622_aligned_GRCh38_winnowmap.sorted.bam
HG02630	HPRC++	HPRC	Illumina	https://s3-us-west-2.amazonaws.com/human-pangenomics/working/HPRC/HG02630/raw_data/Illumina/child/HG02630.final.cram
	HPRC++	HPRC	PacBio	https://s3-us-west-2.amazonaws.com/human-pangenomics/working/HPRC/HG02630/analysis/aligned_reads/hifi/GRCh38/HG02630_aligned_GRCh38_winnowmap.sorted.bam

HG02717	HPRC++	HPRC	Illumina	https://s3-us-west-2.amazonaws.com/human-pangenomics/working/HPRC/HG02717/raw_data/Illumina/child/HG02717.final.cram
	HPRC++	HPRC	PacBio	https://s3-us-west-2.amazonaws.com/human-pangenomics/working/HPRC/HG02717/analysis/aligned_reads/hifi/GRCh38/HG02717_aligned_GRCh38_winnowmap.sorted.bam
HG02723	HPRC++	HPRC_PLUS	Illumina	https://s3-us-west-2.amazonaws.com/human-pangenomics/working/HPRC_PLUS/HG02723/raw_data/Illumina/child/HG02723.final.cram
	HPRC++	HPRC_PLUS	PacBio	https://s3-us-west-2.amazonaws.com/human-pangenomics/working/HPRC_PLUS/HG02723/analysis/aligned_reads/hifi/GRCh38/HG02723_aligned_GRCh38_winnowmap.sorted.bam
HG02818	HPRC++	HPRC_PLUS	Illumina	https://s3-us-west-2.amazonaws.com/human-pangenomics/working/HPRC_PLUS/HG02818/raw_data/Illumina/child/HG02818.final.cram
	HPRC++	HPRC_PLUS	PacBio	https://s3-us-west-2.amazonaws.com/human-pangenomics/working/HPRC_PLUS/HG02818/analysis/aligned_reads/hifi/GRCh38/HG02818_aligned_GRCh38_winnowmap.sorted.bam
HG02886	HPRC++	HPRC	Illumina	https://s3-us-west-2.amazonaws.com/human-pangenomics/working/HPRC/HG02886/raw_data/Illumina/child/HG02886.final.cram
	HPRC++	HPRC	PacBio	https://s3-us-west-2.amazonaws.com/human-pangenomics/working/HPRC/HG02886/analysis/aligned_reads/hifi/GRCh38/HG02886_aligned_GRCh38_winnowmap.sorted.bam
HG02970	HPRC++	HPRC_PLUS	Illumina	https://s3-us-west-2.amazonaws.com/human-pangenomics/working/HPRC_PLUS/HG02970/raw_data/Illumina/child/HG02970.final.cram
	HPRC++	HPRC_PLUS	PacBio	https://s3-us-west-2.amazonaws.com/human-pangenomics/working/HPRC_PLUS/HG02970/raw_data/PacBio_HiFi/m64043_200229_115457.ccs.bam
	HPRC++	HPRC_PLUS	PacBio	https://s3-us-west-2.amazonaws.com/human-pangenomics/working/HPRC_PLUS/HG02970/raw_data/PacBio_HiFi/m64043_200306_190054.ccs.bam
	HPRC++	HPRC_PLUS	PacBio	https://s3-us-west-2.amazonaws.com/human-pangenomics/working/HPRC_PLUS/HG02970/raw_data/PacBio_HiFi/m64043_200310_162218.ccs.bam
	HPRC++	HPRC_PLUS	PacBio	https://s3-us-west-2.amazonaws.com/human-pangenomics/working/HPRC_PLUS/HG02970/raw_data/PacBio_HiFi/m64043_200312_183358.ccs.bam
HG03453	HPRC++	HPRC	Illumina	https://s3-us-west-2.amazonaws.com/human-pangenomics/working/HPRC/HG03453/raw_data/Illumina/child/HG03453.final.cram
	HPRC++	HPRC	PacBio	https://s3-us-west-2.amazonaws.com/human-pangenomics/working/HPRC/HG03453/analysis/aligned_reads/hifi/GRCh38/HG03453_aligned_GRCh38_winnowmap.sorted.bam
HG03486	HPRC++	HPRC_PLUS	Illumina	https://s3-us-west-2.amazonaws.com/human-pangenomics/working/HPRC_PLUS/HG03486/raw_data/Illumina/child/HG03486.final.cram
	HPRC++	HPRC_PLUS	PacBio	https://s3-us-west-2.amazonaws.com/human-pangenomics/working/HPRC_PLUS/HG03486/analysis/aligned_reads/hifi/GRCh38/HG03486_aligned_GRCh38_winnowmap.sorted.bam
HG03492	HPRC++	HPRC_PLUS	Illumina	https://s3-us-west-2.amazonaws.com/human-pangenomics/working/HPRC_PLUS/HG03492/raw_data/Illumina/child/HG03492.final.cram
	HPRC++	HPRC_PLUS	PacBio	https://s3-us-west-2.amazonaws.com/human-pangenomics/working/HPRC_PLUS/HG03492/analysis/aligned_reads/hifi/GRCh38/HG03492_aligned_GRCh38_winnowmap.sorted.bam
HG03516	HPRC++	HPRC	Illumina	https://s3-us-west-2.amazonaws.com/human-pangenomics/working/HPRC/HG03516/raw_data/Illumina/child/HG03516.final.cram
	HPRC++	HPRC	PacBio	https://s3-us-west-2.amazonaws.com/human-pangenomics/working/HPRC/HG03516/analysis/aligned_reads/hifi/GRCh38/HG03516_aligned_GRCh38_winnowmap.sorted.bam
HG03540	HPRC++	HPRC	Illumina	https://s3-us-west-2.amazonaws.com/human-pangenomics/working/HPRC/HG03540/raw_data/Illumina/child/HG03540.final.cram

	HPRC++	HPRC	PacBio	https://s3-us-west-2.amazonaws.com/human-pangenomics/working/HPRC/HG03540/analysis/aligned_reads/hifi/GRCh38/HG03540_aligned_GRCh38_winnowmap.sorted.bam
HG03579	HPRC++	HPRC	Illumina	https://s3-us-west-2.amazonaws.com/human-pangenomics/working/HPRC/HG03579/raw_data/Illumina/child/HG03579.final.cram
	HPRC++	HPRC	PacBio	https://s3-us-west-2.amazonaws.com/human-pangenomics/working/HPRC/HG03579/analysis/aligned_reads/hifi/GRCh38/HG03579_aligned_GRCh38_winnowmap.sorted.bam
NA18906	HPRC++	HPRC_PLUS	Illumina	https://s3-us-west-2.amazonaws.com/human-pangenomics/working/HPRC_PLUS/NA18906/raw_data/Illumina/child/NA18906.final.cram
	HPRC++	HPRC_PLUS	PacBio	https://s3-us-west-2.amazonaws.com/human-pangenomics/working/HPRC_PLUS/NA18906/analysis/aligned_reads/hifi/GRCh38/NA18906_aligned_GRCh38_winnowmap.sorted.bam
NA19030	HPRC++	HPRC_PLUS	Illumina	https://s3-us-west-2.amazonaws.com/human-pangenomics/working/HPRC_PLUS/NA19030/raw_data/Illumina/child/NA19030.final.cram
	HPRC++	HPRC_PLUS	PacBio	https://s3-us-west-2.amazonaws.com/human-pangenomics/working/HPRC_PLUS/NA19030/raw_data/PacBio_HiFi/m64136_200608_190329.ccs.bam
	HPRC++	HPRC_PLUS	PacBio	https://s3-us-west-2.amazonaws.com/human-pangenomics/working/HPRC_PLUS/NA19030/raw_data/PacBio_HiFi/m64136_200610_011635.ccs.bam
	HPRC++	HPRC_PLUS	PacBio	https://s3-us-west-2.amazonaws.com/human-pangenomics/working/HPRC_PLUS/NA19030/raw_data/PacBio_HiFi/m64136_200611_074225.ccs.bam
NA19239	HPRC++	1000GP-SV	Illumina	https://ftp.sra.ebi.ac.uk/vol1/run/ERR323/ERR3239454/NA19239.final.cram
	HPRC++	1000GP-SV	PacBio	https://ftp.sra.ebi.ac.uk/vol1/run/ERR496/ERR4968414/NA19239_20191205_CLEE_m64039_190823_190750.ccs.bam
NA19240	HPRC++	HPRC_PLUS	Illumina	https://s3-us-west-2.amazonaws.com/human-pangenomics/working/HPRC_PLUS/NA19240/raw_data/Illumina/child/NA19240.final.cram
	HPRC++	HPRC_PLUS	PacBio	https://s3-us-west-2.amazonaws.com/human-pangenomics/working/HPRC_PLUS/NA19240/analysis/aligned_reads/hifi/GRCh38/NA19240_aligned_GRCh38_winnowmap.sorted.bam
NA20129	HPRC++	HPRC_PLUS	Illumina	https://s3-us-west-2.amazonaws.com/human-pangenomics/working/HPRC_PLUS/NA20129/raw_data/Illumina/child/NA20129.final.cram
	HPRC++	HPRC_PLUS	PacBio	https://s3-us-west-2.amazonaws.com/human-pangenomics/working/HPRC_PLUS/NA20129/analysis/aligned_reads/hifi/GRCh38/NA20129_aligned_GRCh38_winnowmap.sorted.bam
NA21309	HPRC++	HPRC_PLUS	Illumina	https://s3-us-west-2.amazonaws.com/human-pangenomics/working/HPRC_PLUS/NA21309/raw_data/Illumina/child/AATGACGC-AATGACGC_S5_L001_R1_001.fastq.gz
	HPRC++	HPRC_PLUS	Illumina	https://s3-us-west-2.amazonaws.com/human-pangenomics/working/HPRC_PLUS/NA21309/raw_data/Illumina/child/AATGACGC-AATGACGC_S5_L001_R2_001.fastq.gz
	HPRC++	HPRC_PLUS	PacBio	https://s3-us-west-2.amazonaws.com/human-pangenomics/working/HPRC_PLUS/NA21309/analysis/aligned_reads/hifi/GRCh38/NA21309_aligned_GRCh38_winnowmap.sorted.bam