Design and application of methods for genome inference

Inaugural-Dissertation

zur Erlangung des Doktorgrades
der Mathematisch-Naturwissenschaftlichen Fakultät
der Heinrich-Heine-Universität Düsseldorf

vorgelegt von

Jana Ebler
aus Oberkirch

Düsseldorf, Dezember, 2022

aus dem Institut für Medizinische Biometrie und Bioinformatik
der Heinrich-Heine-Universität Düsseldorf

Gedruckt mit der Genehmigung der
Mathematisch-Naturwissenschaftlichen Fakultät der
Heinrich-Heine-Universität Düsseldorf

Berichterstatter:

1. Prof. Dr. Tobias Marschall

2. Prof. Dr. Gunnar W. Klau

Tag der mündlichen Prüfung: 21. Juni 2023

Statement

I declare under oath that I have produced my thesis independently and without any undue assistance by third parties under consideration of the “Principles for the Safeguarding of Good Scientific Practice at Heinrich Heine University Düsseldorf”.

Düsseldorf, December 2022

Jana Ebler

Abstract

Humans are diploid and carry two copies of their DNA, packaged into chromosomes. Other species, including many plants, are polyploid and carry more than two copies of each chromosome. Phasing describes the process of inferring the exact sequences of these chromosomal copies, called haplotypes, based on sequencing data. While tools for accurately phasing diploid genomes exist already, phasing polyploids is still challenging.

In the first part of this thesis, a new algorithm for polyploid phasing is introduced and applied to sequencing data of a tetraploid potato genome. Next, it is demonstrated how the new PacBio Circular Consensus Sequencing (CCS) technology simplifies alignment-based phasing by providing accurate long reads that enable variant calling and phasing based on a single sequencing technology, removing the necessity of an additional short-read dataset. In addition, CCS reads enable reference-free *de novo* assembly of individual haplotypes on the scale of chromosomes that include structural variation typically missed by alignment-based phasing methods. Such haplotype sequences enable the construction of pangenome graphs that provide a representation of the genetic diversity of the contained samples.

In the second part of this thesis, a new genotyping method, PanGenie, is presented, which leverages a pangenome graph in order to infer genotypes of genetic variants from short-read sequencing data, without requiring time consuming read alignments. It improves genotyping accuracy of structural variants over traditional alignment-based short-read genotyping methods, which often perform worse due to poor reference alignments in these regions.

The third part of this thesis describes several applications of PanGenie. It presents results of structural variant genotyping across a large cohort of human samples based on pangenome representations generated by the HGSVC and HPRC consortia. Results show that PanGenie is able to genotype structural variants previously inaccessible by other short-read based methods, enabling the inclusion of such variants into Quantitative trait locus (QTL) analyses. Furthermore, it is demonstrated how SNP genotypes produced by PanGenie across the cohort samples can be used to detect carriers of rare inversions.

Kurzfassung

Der Mensch ist diploid und trägt daher zwei Kopien seiner DNA, die in Chromosomen verpackt sind. Andere Arten, darunter viele Pflanzen, sind polyploid und tragen mehr als zwei Kopien jedes Chromosoms. Haplotypisierung beschreibt die Rekonstruktion der Sequenzen dieser Kopien, die sogenannten Haplotypen, auf der Grundlage von Sequenzierdaten. Während bereits einige Methoden zur Haplotypisierung diploider Genome existieren, ist die Haplotypisierung von polyploiden Organismen immer noch eine Herausforderung.

Im ersten Teil dieser Arbeit wird ein neuer Algorithmus zur Haplotypisierung von polyploiden Genomen vorgestellt und auf Sequenzierdaten eines tetraploiden Kartoffelgenoms angewendet. Anschließend wird gezeigt, dass die neue Circular Consensus Sequenziermethode (CCS) von PacBio die Alignment-basierte Haplotypisierung vereinfacht, da sie lange Reads mit geringen Fehlerraten liefert, die sowohl zur Detektion von Varianten als auch zur eigentlichen Haplotypisierung verwendet werden können. Dadurch ist kein zusätzlicher Datensatz mit kurzen Reads mehr notwendig. Darüber hinaus ermöglichen CCS-Reads die referenzfreie *de novo* Assemblierung individueller Haplotypen ganzer Chromosomen. Solche Haplotypsequenzen schließen neben kurzen Varianten auch strukturelle Varianten ein, die von Alignment-basierten Methoden meist nicht miteinbezogen werden. Somit ermöglichen sie die Konstruktion von Pangenom-Graphen, die eine detaillierte Beschreibung der genetischen Variabilität einer Art darstellen.

Im zweiten Teil dieser Arbeit wird eine neue Genotypisierungsmethode namens PanGenie vorgestellt, die einen Pangenom-Graphen nutzt, um Genotypen genetischer Varianten aus kurzen Sequenzierreads abzuleiten, ohne dass zeitaufwändige Read-Alignments berechnet werden müssen. PanGenie liefert genauere Genotypen für strukturelle Varianten als bereits existierende Methoden für kurze Reads, die aufgrund von schlechten Referenz-Alignments in den entsprechenden Regionen oft schlecht abschneiden.

Im dritten Teil dieser Arbeit werden verschiedene Anwendungen von PanGenie vorgestellt. Es werden Ergebnisse der Genotypisierung struktureller Varianten in einer großen Kohorte menschlicher Genomen diskutiert, für die die Pangenom-Graphen der HGSVC- und HPRC-Projekte verwendet wurden. Die Ergebnisse verdeutlichen, dass PanGenie in der Lage ist, strukturelle Varianten zu genotypisieren, die zuvor mit anderen Methoden nicht zugänglich waren. Dies ermöglicht es, solche Varianten in “Quantitative Trait Locus”-Analysen einzubeziehen. Außerdem wird demonstriert, wie die von PanGenie über die Kohorte hinweg erzeugten SNP-Genotypen zur Erkennung von Trägern seltener Inversionen

verwendet werden können.

Acknowledgments

First and foremost, I would like to thank my supervisor Tobias Marschall for his great support during my PhD, and also during my bachelor's and master's projects, enabling me to do great research from an early stage. I also want to thank my colleagues (listed in alphabetical order) Ali Ghaffaari, Christina Gros, Daniel Dörr, Fawaz Dabbaghieh, Hufsah Ashraf, Hugo Magalhães, Konstantinn Bonnet, Maryam Ghareghani, Mikko Rautiainen, Peter Ebert, Rebecca Serra Mari, Samarendra Pani and Shilpa Garg for creating a great atmosphere at and outside of work. It really helped me to endure the hard and challenging periods during my PhD. I also want to thank my family and friends for always supporting me. Lastly, I would like to thank Fawaz Dabbaghieh, Hufsah Ashraf, Hugo Magalhães, Konstantinn Bonnet, Maryam Ghareghani, Mikko Rautiainen, Peter Ebert, Rebecca Serra Mari and Samarendra Pani for proof-reading my thesis.

Contents

Statement iii

Abstract v

Kurzfassung vii

Acknowledgments ix

Contents xi

List of Figures xv

List of Tables xvii

Introduction 1

1 Background 5
1.1 Genomes . . . 5
1.2 Inheritance and recombination . . . 6
1.3 Genetic variation . . . 7
1.4 DNA sequencing . . . 8
1.4.1 Massively parallel short-read sequencing . . . 9
1.4.2 Single molecule fluorescent sequencing . . . 10
1.4.3 Nanopore sequencing . . . 10
1.4.4 Comparison of sequencing technologies . . . 11
1.5 Sequence alignment . . . 12
1.6 Genotyping . . . 12
1.6.1 Short-read based genotyping . . . 13
1.6.2 Long-read based genotyping . . . 15
1.7 Phasing . . . 16
1.7.1 Evaluating phasing results . . . 17
1.8 Genome assembly . . . 18

1.9 Pangenomics 20
1.10 Mathematical background 21
1.10.1 Hidden Markov Models 22
1.10.2 Forward-Backward algorithm 22
1.11 File formats 23
1.11.1 VCF format 24
1.11.2 SAM/BAM format 25
1.11.3 FASTA/FASTQ format 25

2 Reference-based haplotype phasing **27**
2.1 Accurate polyploid phasing from long reads 27
2.1.1 Introduction 27
2.1.2 Phasing algorithm 29
2.1.3 Evaluation on artificial polyploid humans 31
2.1.4 Analysis of potato data 32
2.1.5 Discussion 36
2.2 Phasing small variants with circular consensus long reads 38
2.2.1 Introduction 38
2.2.2 Data generation and variant calling 39
2.2.3 Phasing small variants with CCS reads 39
2.2.4 Improved genome assembly with CCS reads 40
2.2.5 Discussion 41
2.3 Conclusion 43

3 PanGenie: Pangenome-based genome inference **45**
3.1 Introduction 45
3.2 Algorithm overview 47
3.2.1 Hidden Markov Model 48
3.3 Variant calling from haplotype-resolved assemblies 52
3.4 Constructing a pangenome reference 54
3.5 Comparison to existing genotyping methods 55
3.5.1 Evaluation metrics 56
3.5.2 Evaluation regions 58
3.5.3 Results 59
3.5.4 Resources 62
3.6 Accuracy in the Major Histocompatibility Complex 63
3.7 Genotyping larger cohorts 64
3.8 Comparison to gnomAD 67
3.9 LD analysis 68
3.10 Discussion and conclusions 69
3.10.1 Limitations and future directions 70

4 Application: Genotyping Large Cohorts **73**
4.1 HGSVC project . . . 75
4.1.1 Introduction . . . 75
4.1.2 Speeding up PanGenie for larger panels . . . 75
4.1.3 Variant calling from haplotype-resolved assemblies . . . 76
4.1.4 Genotyping SVs across a cohort of 3,202 individuals . . . 76
4.1.5 Added value from graph-based genotyping into short-read WGS data . 81
4.1.6 Discussion . . . 82
4.2 Identifying rare inversions . . . 84
4.2.1 Introduction . . . 84
4.2.2 Identifying potential inversion carriers using PanGenie . . . 84
4.2.3 Discussion . . . 85
4.3 HPRC project . . . 88
4.3.1 Introduction . . . 88
4.3.2 Pangenome construction and variant calling . . . 88
4.3.3 Genotyping SVs across a cohort of 3,202 individuals . . . 89
4.3.4 Discussion . . . 101
4.4 Conclusions . . . 104
4.4.1 Computing PanGenie's Forward-Backward algorithm more efficiently 105

Summary **109**

Bibliography **111**

A Appendices: Application and advances in haplotype phasing **127**
A.1 Accurate polyploid phasing from long reads . . . 127

B Appendices: PanGenie: Pangenome-based genome inference **131**

C Appendices: Application: genotyping large cohorts **147**
C.1 HGSVC project . . . 147
C.2 HPRC project . . . 155

D Code Availability **163**

Published articles **165**

E Published articles underlying this thesis **165**
E.1 Haplotype threading: accurate polyploid phasing from long reads 165
E.1.1 Authors 165
E.1.2 Contributions 165
E.1.3 Licence and copyright information 166
E.2 Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome 166
E.2.1 Authors 166
E.2.2 Contributions 167
E.2.3 Licence and copyright information 167
E.3 Fully phased human genome assembly without parental data using single-cell strand sequencing and long reads 168
E.3.1 Authors 168
E.3.2 Contributions 168
E.3.3 Licence and copyright information 169
E.4 Pangenome-based genome inference allows efficient and accurate genotyping across a wide spectrum of variant classes 169
E.4.1 Authors 169
E.4.2 Contributions 170
E.4.3 Licence and copyright information 170
E.5 Haplotype-resolved diverse human genomes and integrated analysis of structural variation 170
E.5.1 Authors 171
E.5.2 Contributions 171
E.5.3 License and copyright information 172
E.6 Recurrent inversion polymorphisms in humans associate with genetic instability and genomic disorders 172
E.6.1 Authors 172
E.6.2 Contributions 172
E.6.3 License and copyright information 173
E.7 A Draft Human Pangenome Reference 173
E.7.1 Authors 173
E.7.2 Contributions 173
E.7.3 License and copyright information 174

List of Figures

1.1 DNA and recombination 6
1.2 Genetic variants 7
1.3 Illumina sequencing 9
1.4 Single molecule real sequencing 10
1.5 Nanopore sequencing 11
1.6 Genotyping 13
1.7 Haplotype reconstruction 16
1.8 Computation of switch errors in polyploid setting 18
1.9 Pangenome representations 20
1.10 VCF format example (multiallelic) 24
1.11 VCF format example (biallelic) 25

2.1 MEC problem in polyploid setting 29
2.2 WhatsHap polyphase overview 30
2.3 Phasing of a potato genome 33
2.4 Fraction of phased variants in relation to heterozygosity in the potato genes . 34
2.5 Haplotype assemblies for the FRIGIDA gene 35
2.6 Phasing with CCS reads 41

3.1 Overview of PanGenie 47
3.2 PanGenie HMM 49
3.3 Callset statistics 52
3.4 Variant calling and graph construction 53
3.5 Overview of leave-one-out experiment 55
3.6 PanGenie evaluation metrics 57
3.7 Results of leave-one-out experiment (SNPs and small variants) 59
3.8 Results of leave-one-out experiment (midsize and large variants) 60
3.9 HLA genotyping 63
3.10 Cohort genotype filtering 65
3.11 Genotyping large cohorts 67
3.12 LD analysis 68
3.13 Nested variant representation 71

4.1 Unfiltered HGSVC genotypes . . . 78
4.2 Strictly filtered HGSVC genotypes . . . 79
4.3 Leniently filtered HGSVC genotypes . . . 82
4.4 Rare inversion on chromosome 15. . . . 86
4.5 Rare inversion chromosome 2 . . . 87
4.6 Decomposition of graph bubbles . . . 90
4.7 Traversal-based decomposition . . . 91
4.8 HPRC allele statistics . . . 92
4.9 HPRC leave-one-out experiment . . . 94
4.10 HPRC unfiltered and positive sets . . . 96
4.11 HPRC filtered set . . . 98
4.12 HPRC novel and known variants . . . 100

B.1 Evaluation example . . . 133
B.2 Adjusted precision/recall for NA12878 (non-repetitive regions) . . . 134
B.3 Adjusted precision/recall for NA12878 (STR/VNTR regions) . . . 135
B.4 Comparison to GIAB small variants for NA12878 . . . 136
B.5 Weighted genotype concordance for NA24385 (non-repetitive regions) . . . 137
B.6 Weighted genotype concordance for NA24385 (STR/VNTR regions) . . . 138
B.7 Adjusted precision/recall for NA24385 (non-repetitive regions . . . 139
B.8 Adjusted precision/recall for NA24385 (STR/VNTR regions) . . . 140
B.9 Adjusted F-score for NA24385 (non-repetitive regions) . . . 141
B.10 Adjusted F-score for NA24385 (STR/VNTR regions) . . . 142
B.11 Comparison to syndip benchmark SVs . . . 143
B.12 GIAB medically relevant SVs in our lenient set . . . 144
B.13 LD analysis for ABO . . . 144
B.14 LD analysis for CCDC91 . . . 145

C.1 Comparison PanGenie and Paragraph on HGSVC calls . . . 148
C.2 HGSVC genotyping results for SNVs . . . 149
C.3 HGSVC genotyping results for indels . . . 150
C.4 Filtering HGSVC SV insertions . . . 151
C.5 Filtering HGSVC SV deletions . . . 152
C.6 F_{ST} versus SV length for all superpopulations (deletions) . . . 153
C.7 F_{ST} versus SV length for all superpopulations (insertions) . . . 154
C.8 HPRC callset statistics SV deletions . . . 156
C.9 HPRC callset statistics SV insertions . . . 157
C.10 HPRC callset statistics SV others . . . 158
C.11 HPRC common SVs . . . 159
C.12 HPRC SVs in repeat regions . . . 159
C.13 Runtime and memory usage of PanGenie . . . 161

List of Tables

2.1 WhatsHap phasing performance on DeepVariant (CCS) callset 40

3.1 Runtime and memory usage of different genotypers 61
3.2 Number of variants before/after filtering 66

4.1 Number of HGSVC variants 77
4.2 Callset statistics for the HGSVC lenient set 81
4.3 HPRC number of variant alleles in HPRC callsets 95
4.4 HPRC medically relevant SV benchmarking 101

A.1 Phasing evaluation on artificial polyploid human 128
A.2 Phasing evaluation in/outside collapsing regions 129

B.1 Variant calling statistics 131
B.2 Variants in pangenome graph 132
B.3 Number of variants in repetitive and non-repetitive regions 132

Introduction

The deoxyribonucleic acid (DNA) is the carrier of hereditary information enabling the translation of genetic information that is necessary to maintain a living organism [8]. The discovery of its molecular structure marked a milestone in genetics, revealing the mechanism of DNA replication and gene expression [8, 197]. According to these findings, the DNA is composed of a sequence of four different nucleotide bases, organized in a double helix structure. In eukaryotes, the DNA is packaged into chromosomes and individuals carry several copies of their DNA, referred to as haplotypes. Humans are diploid and carry two copies of each chromosome. Other species, including many plants, are polyploid and carry more than two. Although these copies are very similar, they are not identical due to genetic variation present in the underlying DNA sequences. Various types of genetic variation exist and such mutations can have an impact on the expression of proteins that are encoded in the DNA. In this way, genetic variation can influence traits of an individual and in some cases cause diseases, including cancer [2, 25, 31, 62, 118, 133, 163, 168, 181, 193, 198, 201]. Therefore, studying the exact DNA sequence of an individual is important in biological and medical research. DNA sequencing is the key to analyzing the genome of an individual. The basic idea is to fragment its DNA into many small pieces and determine the DNA sequence of each of them, resulting in sets of so-called sequencing reads. Several sequencing technologies exist. They differ in terms of the length and accuracy of reads they produce. While short-read sequencing technologies produce very accurate reads of up to 300 bases in length, long-read sequencing technologies typically produce much longer (up to 100 kilo bases), but less accurate reads. Long reads are especially useful for reconstructing the haplotype sequences of an individual. Haplotype reconstruction is an active field of research [59, 98, 100, 145, 154]. Knowing the haplotype sequences of an individual is beneficial in many ways. It allows studying how combinations of variants impact phenotypes [100] and provides insights into allele-specific DNA methylation and gene expression [186]. Furthermore, haplotypes are very useful for the construction of pangenome graphs. Pangenomics is a relatively new field of research with the goal of developing data structures to capture DNA sequences and genetic variability of a species. A compact way of representing a pangenome is through a graph structure, constructed from known haplotype sequences of individuals of a species. In the long term, by providing a more accurate representation of complex genomic regions, such graphs offer the possibility of replacing the linear reference genome and improving the various downstream analyses that are currently based upon it. This the-

sis introduces several approaches for genome inference, including a long-read based method for haplotype reconstruction, a short-read based genotyping approach using a pangenome graph structure, as well as applications of these methods to various sequencing data sets.

Outline

Chapter 1 gives a detailed introduction to the biological concepts relevant to this thesis, as well as basic mathematical definitions that are underlying the proposed algorithms.

In Chapter 2, a new alignment-based algorithm for polyploid phasing using error-prone, long sequencing reads is presented. It uses cluster editing to group reads by their similarity and reconstructs the haplotypes based on a novel haplotype threading model. The focus is on an application of this algorithm to sequencing data of the tetraploid potato. This work was published in *Genome Biology* [166]. The second part of Chapter 2 presents the new PacBio Circular Consensus Sequencing technology. The focus of this section is on phasing of a diploid human sample based on these new reads. My contribution to this work was published as part of two *Nature Biotechnology* publications [145, 199].

Chapter 3 presents a new approach to genotyping genetic variants based on a pangenome graph and short-read sequencing data. The algorithm uses counts of allele specific k-mers computed from the reads of the sample to be genotyped, in combination with known haplotype paths represented in the pangenome to genotype the sample. The experiments demonstrate that leveraging the pangenome structure is especially helpful for accurately genotyping structural variants. The material presented in this chapter was published in *Nature Genetics* [49].

Chapter 4 presents three applications of the genotyping algorithm introduced in Chapter 3. The first application is genotyping variants detected from 64 haplotype-resolved assemblies across 3,202 human samples based on data generated by the Human Genome Structural Variation Consortium (HGSVC) [24, 46]. The results were published as part of a *Science* publication [46]. The second application presents an approach to detect rare inversions in human samples based on the HGSVC genotypes, described as part of a *Cell* publication [146]. The third application presents my contributions to the Human Pangenome Reference Consortium (HPRC) [90, 113]. Based on a new pangenome reference containing 88 haplotypes, 3,202 human samples were genotyped and genotypes of structural variants analyzed. The work is currently under revision and publicly available as part of a preprint [113].
The publications underlying this thesis are listed below. First authorship is denoted by *.

- S. D. Schrinner*, R. Serra Mari*, **J. Ebler***, M. Rautiainen, L. Seillier, J. J. Reimer, B. Usadel, T. Marschall, and G. W. Klau. Haplotype threading: accurate polyploid phasing from long reads. *Genome Biology*, 21(1):1–22, 2020

- A. M. Wenger*, P. Peluso*, W. J. Rowell, P.-C. Chang, R. J. Hall, G. T. Concepcion, **J. Ebler**, A. Fungtammasan, A. Kolesnikov, N. D. Olson, et al. Accurate circular consensus

long-read sequencing improves variant detection and assembly of a human genome. *Nature Biotechnology*, 37(10): 1155–1162, 2019

- D. Porubsky*, P. Ebert*, P. A. Audano, M. R. Vollger, W. T. Harvey, P. Marijon, **J. Ebler**, K. M. Munson, M. Sorensen, A. Sulovari, et al. Fully phased human genome assembly without parental data using single-cell strand sequencing and long reads. *Nature Biotechnology*, 39(3):302–308, 2021

- **J. Ebler***, P. Ebert, W. E. Clarke, T. Rausch, P. A. Audano, T. Houwaart, Y. Mao, J. O. Korbel, E. E. Eichler, M. C. Zody, et al. Pangenome-based genome inference allows efficient and accurate genotyping across a wide spectrum of variant classes. *Nature Genetics*, 54(4):518–525, 2022

- P. Ebert*, P. A. Audano*, Q. Zhu*, B. Rodriguez-Martin*, D. Porubsky, M. J. Bonder, A. Sulovari, **J. Ebler**, W. Zhou, R. Serra Mari, et al. Haplotype-resolved diverse human genomes and integrated analysis of structural variation. *Science*, 372(6537), 2021

- D. Porubsky*, W. Höps*, H. Ashraf*, P. Hsieh, B. Rodriguez-Martin, F. Yilmaz, **J. Ebler**, P. Hallast, F. A. M. Maggiolini, W. T. Harvey, et al. Recurrent inversion polymorphisms in humans associate with genetic instability and genomic disorders. *Cell*, 185(11):1986–2005, 2022

- W.-W. Liao*, M. Asri*, **J. Ebler***, D. Doerr, M. Haukness, G. Hickey, S. Lu, J. K. Lucas, J. Monlong, H. J. Abel, et al. A draft human pangenome reference. *bioRxiv*, 2022

Chapter 1

Background

This chapter provides an overview of the biological concepts that are important in this thesis, as well as the computational problems arising when analyzing biological data. It provides definitions and explanations of the problems underlying Chapters 2, 3 and 4.

1.1 Genomes

Genetic information is encoded in deoxyribonucleic acid (DNA) sequences. A DNA molecule consists of two polynucleotide strands that are wound around the same axis so as to form a double helix [8, 197]. Each strand is composed of a sequence of nucleotides, which consist of a sugar molecule (desoxyribose), a phosphate group and a base [8]. In order to form a DNA strand, a sugar molecule of one nucleotide is linked to the phosphate group of the next, forming a backbone of alternating sugar phosphate molecules [8]. The two strands of a DNA molecule are connected by hydrogen bonds formed between pairs of bases [8]. Four bases can occur in a DNA sequence: adenine, thymine, guanine or cytosine. In order to form hydrogen bonds, adenine can be paired with thymine only and guanine with cytosine [8]. As a consequence, the two strands forming a DNA molecule are complementary [8]. Figure 1.1a provides an illustration of a DNA molecule.

The order of nucleotides along the DNA strands encodes genetic information [8]. Certain sections within the DNA strands, so-called protein-coding genes, provide instructions on the construction of proteins. During gene expression, nucleotide sequences of genes are translated into amino acid sequences of the encoded proteins [8]. The genome of an organism describes the complete set of information encoded in its DNA [8].

In eukaryotes, the DNA is typically packaged into dense structures, called chromosomes, by means of specialized proteins. Human cells contain 22 autosomes and two sex chromosomes. Since humans are diploid organisms, each cell carries two copies of each autosome, one inherited from the mother and the other inherited from the father [8]. These two versions of a chromosome are homologous meaning their structure and shape are the same, however, their DNA sequences are not identical as each chromosome carries genetic vari-

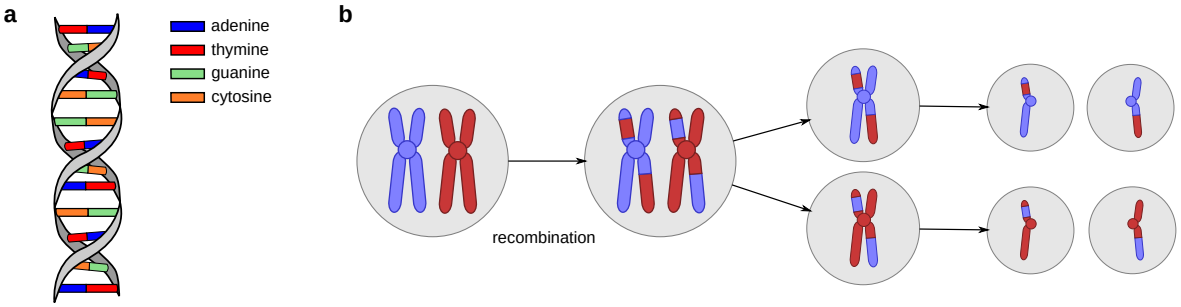


Figure 1.1: DNA and recombination. a Double helix structure of the DNA. The two complementary strands are connected by hydrogen bonds between the nucleotides. Adenine can be paired with thymine only and cytosine with guanine. **b** Meiosis. After chromosomes have replicated in order to consist of two identical sister chromatids, recombination takes place during which segments are exchanged between homologous chromosomes. Subsequently, four haploid gametes are formed in two steps of cell division.

ation (see Section 1.3). While the two sex chromosomes in females are homologous (two X chromosomes), this is not the case for males, as they carry one X chromosome, inherited from the mother, and one Y chromosome which is inherited from the father [8].

The number of chromosomes differs across different species. While the total number of chromosomes is 46 for humans (counting all copies), it is 48 for gorilla (*Gorilla gorilla*) and chimpanzee (*Pan troglodytes*), 40 for mouse (*Mus musculus*) and 42 for rat (*Mus norvegicus*), which are all diploid [36, 140]. Other species are polyploid and carry more than two copies of their homologous chromosomes. Examples are bread wheat (*Triticum aestivum*) which is hexaploid and thus carries six homologous copies, or potato (*Solanum tuberosum*) which is tetraploid (4 copies) [45, 196].

1.2 Inheritance and recombination

During sexual reproduction, two gametes fuse in order to form a diploid organism. The process in which gametes are formed in the parental individuals is called meiosis. The gamets are haploid, which means that they contain only one copy of each chromosome. The first step of meiosis is the replication of the genetic material. As a result, each chromosome consists of two sister chromatids [8]. In the prophase of meiosis, each chromosome pairs with its homologous copy resulting in a structure that is called bivalent [8]. In this phase, recombination takes place between the chromosomes, which means that fragments are exchanged between the two homologous copies, such that the resulting chromosomes are combinations of paternal and maternal DNA [8]. Next, the cell divides to form two haploid daughter cells, each carrying one copy of each chromosome [8]. However, each of the chromosomes still consists of two sister chromatids. In a second cell division step, the sister chromatids are separated and again distributed to two daughter cells [8]. In this way, four haploid gametes are formed [8]. Figure 1.1b provides an illustration of meiosis. Fusion of a maternal and paternal gamete leads to a diploid organism, inheriting half of its chromosomes from the

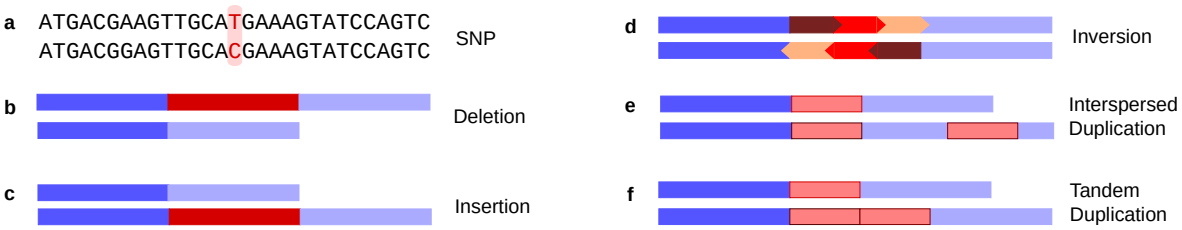


Figure 1.2: Genetic variants. Overview of different types of genetic variations. **a** Single Nucleotide Polymorphisms (SNPs) are mutations of a single base pair. **b,c** Segments deleted from the sequence or inserted into the sequence, are called deletions and insertions, respectively. **d** In case of an inversion, the sequence of a segment is inverted. **e, f** A segment is duplicated and occurs more than once in the sequence. In case of a tandem duplication, the copies are located adjacent to the segment.

mother and the other half from its father.

1.3 Genetic variation

Due to genetic variants, the DNA sequences of different individuals are not completely identical, even if they belong to the same species. Genetic variation defines genetic diversity in populations, but can also cause genetic diseases. Different types of genetic variants exist. Single nucleotide polymorphisms (SNPs) and single nucleotide variants (SNVs) are alterations of single bases in the nucleotide sequences of an individual (Figure 1.2a). While the term “SNP” is sometimes defined to refer to germline events that are present in at least 1% of a population, many authors do not apply this threshold [2, 88, 107]. Early uses of the term “SNV” often focused on somatic events [43, 66, 96, 107]. In this thesis, “SNP” will be used to refer to any single nucleotide germline substitution, regardless of its allele frequency.

SNPs are the most common type of variants. In a human sample, on average more than every 1000th genomic position carries a SNP. Some SNPs have been shown to be associated with diseases, including cancer. Specific SNPs located in microRNA binding sides can influence the susceptibility of humans to get certain cancers [133]. Another example is migraine, which is associated with specific SNPs located in multiple genes, including the TRPM8 or the LRP1 gene [25]. Furthermore, SNPs can be related to specific traits. Certain SNPs in human 5-HT-2A gene for example, were shown to be associated with anger- and aggression-related traits, with some being protective against suicidal behaviour [62].

Besides alternations of single bases, parts of the DNA sequence of an individual can be deleted, or additional sequence can be inserted into the DNA, resulting in deletions and insertions, respectively (Figure 1.2b,c). Depending on their length, such variants are typically classified as indels if the inserted or deleted sequence is less than 50 base pairs long. Longer variants ($\geq$ 50 bp) are called structural variants (SVs), and, besides insertions and deletions, also include other variant types, such as inversions or duplications. In case of an inversion, a segment in the genome is inverted (Figure 1.2d). In case of a duplication, several copies

of a segment are present in the genomic sequence of an individual. Repeated segments can either be non-adjacent (interspersed duplication, Figure 1.2e) or they are inserted adjacent to the original segment (tandem duplication, Figure 1.2f).

Besides SNPs, other types of genetic variants can have an influence on traits or diseases of individuals. An indel in the human CFTR gene for example, has been shown to cause cystic fibrosis [31]. More than 99.9% of the genetic variation in a human genome consists of SNPs and indels [2]. Due to their larger sizes, structural variants, however, affect more bases and are thus a major contributor to phenotypic variation and diseases of individuals [2, 181, 198]. Interpreting the functional consequences of SVs is more difficult than for other types of genetic variants, since they often occur in more complex, repetitive genomic regions [198]. Structural variants have been linked to various traits and diseases, like attention-deficit hyperactivity [201], autism [163, 168] or schizophrenia [118, 193], and play a major role in many cancers [112].

In diploid or polyploid individuals, either of the chromosomal copies can carry different genetic variants. The different versions of sequence segments that can be present at a variant locus are referred to as alleles. The sequence of alleles located on the same chromosomal copy is called haplotype [143].

1.4 DNA sequencing

Sequencing aims at determining the nucleotide sequence of an individual and is thus crucial for studying an individual's genome and the genetic variation it carries. Typically, the genome is fragmented into many small pieces prior to sequencing and sequences of each of these pieces are determined, resulting in so-called sequencing reads. These reads then provide the basis for downstream analyses, such as alignment to a reference genome, genotyping genetic variants, phasing or genome assembly.

The most relevant early sequencing methods were Sanger dideoxy synthesis and the Maxam-Gilbert chemical cleavage method [121, 164, 165]. The Maxam-Gilbert method is based on base-specific cleavages of radioactively labeled DNA segments [177]. Sanger dideoxy synthesis uses chain terminating nucleotides which, starting from a labeled primer, are incorporated into a newly formed strand by a polymerase and prevent sequence elongation [122, 177]. For each of the four bases, both methods produce labeled fragments of different length. Both methods then separate these fragments by size using electrophoresis [173]. By exposure to X-ray film, the sequence can be determined based on the ordering of fragments [173]. Although both sequencing methods have been improved further, the method by Sanger became the sequencing standard [177]. Instead of radioactive labeling of segments, fluorescent dyes were introduced which enabled sequence detection based on laser-induced fluorescent emission instead of using X-rays [122]. Sanger sequencing is still a useful method today, especially in scenarios where high throughput is not required [173]. However, sequencing larger genomes, like the human genome, requires a much more effi-

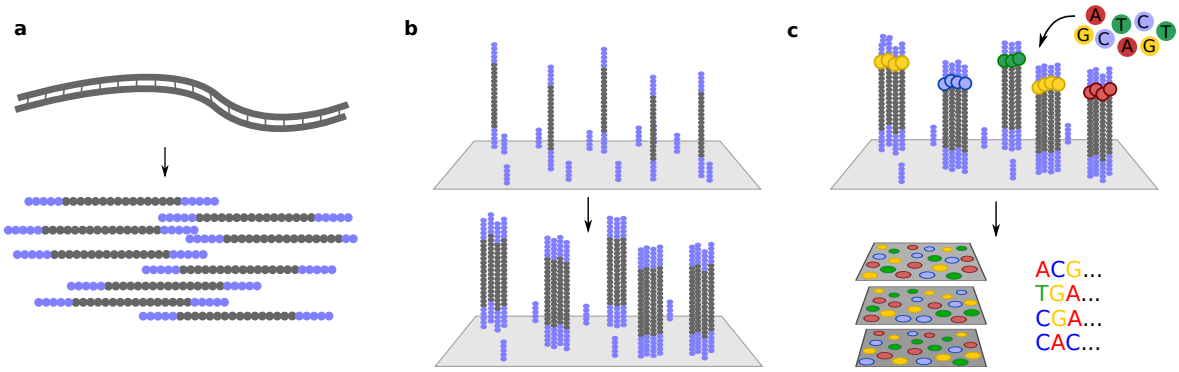


Figure 1.3: Illumina sequencing. **a** The DNA is randomly fragmented into short, single-stranded segments and adapters are added to each end of the resulting fragments. **b** Fragments are loaded into a flow cell with surface-bound oligonucleotides. Fragments bind and are amplified by PCR resulting in distinct clusters. **c** Fluorescently labeled nucleotides are used by polymerases to synthesize complementary strands. After a new base was incorporated into each strand, the flow cell is imaged. In this way, the sequence of each fragment can be determined step by step.

cient and scalable approach that offers high throughput.

1.4.1 Massively parallel short-read sequencing

Most short-read sequencing approaches are based on the concept of Sequencing by Synthesis (SBS). The underlying ideas are similar to Sanger sequencing, but steps are parallelized in order to generate higher throughput [122]. Most methods no longer use dideoxy terminators so that chain elongation proceeds while imaging the nucleotides incorporated [177]. Another difference to the original Sanger method is that most methods based on SBS only achieve shortened read fragments (300-500 bp) [177]. One of the most popular technologies is Illumina sequencing. The first step is library preparation [81]. The DNA is randomly fragmented into short, single stranded segments [81]. After adapter sequences ligated on the 5' and 3' ends, each resulting fragment is amplified using PCR [81] (Figure 1.3 A). The second step is cluster generation. The library is loaded into a flow cell which contains surface-bound oligonucleotides that are complementary to the adapter sequences [81]. Fragments are captured on the surface and amplified using bridge amplification, resulting in distinct, clonal clusters consisting of around 1,000 copies of each fragment [81, 177] (Figure 1.3B). The next step is to sequence the fragments based on Sequencing by Synthesis. A polymerase synthesizes complementary strands using fluorescently labeled nucleotides (dNTPs) [81]. Incorporation of each dNTP into the newly synthesized strands produces a light signal. Each of the four dNTPs emits a different wavelength which enables to identify which base was added to a cluster in each step. The flow cell is imaged after each newly added base. In this way, the sequences of each distinct cluster can be determined step by step [81] (Figure 1.3C). The resulting read sequences can then be used in downstream analyses, such as read alignment or k-mer counting.

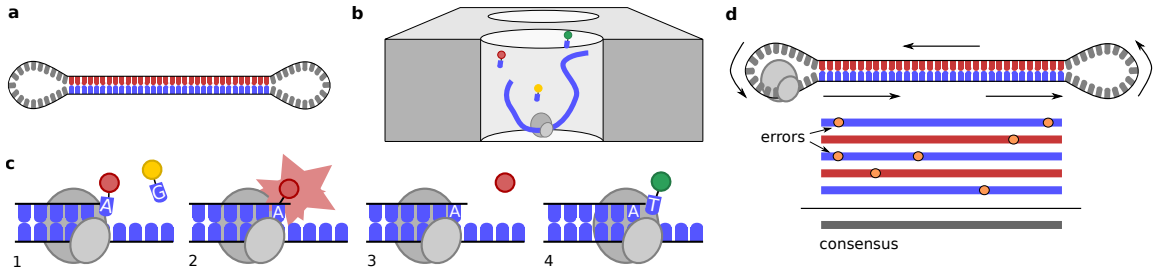


Figure 1.4: Single molecule real time sequencing. **a** A SMRTbell molecule consisting of a DNA fragment with hairpin adapters added to both ends. **b** The SMRTbell is loaded into a ZMW, which has a DNA polymerase bound at the bottom. **c** The sequence of the DNA fragment is determined by Sequencing by Synthesis. Fluorescently labeled nucleotides are incorporated into the strand by the polymerase and emit a light signal. **d** For circular consensus sequencing, the fragment is sequenced multiple times, producing multiple reads for the same segment. An accurate consensus sequence is formed which eliminates sequencing errors contained in each individual read.

1.4.2 Single molecule fluorescent sequencing

Pacific Biosciences (PacBio) developed a sequencing technology based on single molecule real time sequencing (SMRT). In contrast to Illumina sequencing, PacBio sequencing can generate reads of up to 50 kbp in length [177]. In order to sequence DNA, fragments are prepared by ligating hairpin adapters to both ends in order to create a circular, double-stranded molecule which is called SMRTbell [157] (Figure 1.4a). SMRTbells are then loaded into a SMRTcell, a chip containing up to 1 million sequencing units called Zero Mode Waveguides (ZMW) [157, 177]. In each ZMW, a polymerase is bound at the bottom which binds the hairpin adapters and replicates the DNA fragment [177] (Figure 1.4b). Fluorescently labeled nucleotides are provided for chain elongation and enable identifying the bases incorporated into the strand by the polymerase in each step [157] (Figure 1.4c). From the sequence of light signals emitted, the sequence of the underlying fragments can be determined in this way, resulting in continuous long reads (CLR). Compared to Illumina sequencing, PacBio CLR sequencing suffers from a much higher sequencing error rate (see Section 1.4.4). However, PacBio recently presented a method to generate highly accurate long reads by using Circular Consensus Sequencing (CCS) [199]. The circular nature of the SMRTbell molecule enables the polymerase to read the same fragment more than once, producing several reads of the same template segment. Afterwards, a single consensus sequence can be formed from these reads (Figure 1.4d). This allows to eliminate the high rate of sequencing errors within each individual read and thus results in a highly accurate, long-read sequence.

1.4.3 Nanopore sequencing

Nanopore-based sequencing of DNA, commercialized by Oxford Nanopore Technologies, produces sequencing reads reaching lengths over 100 kbp that are much longer than the

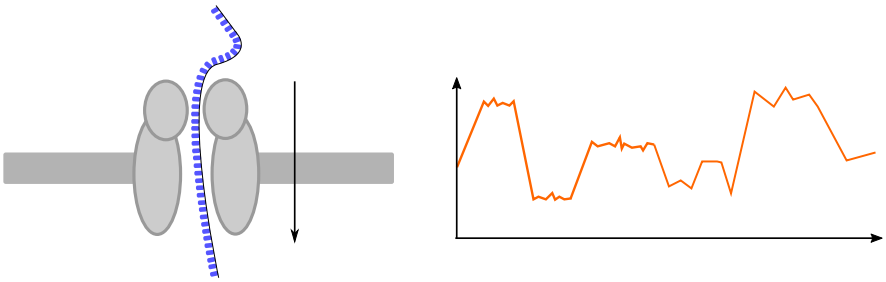


Figure 1.5: Nanopore sequencing. A DNA fragment passes through a nanopore embedded in an electrically resistant membrane. A detector measures the change of the current flowing through the pore. Based on characteristic changes of the current induced by each nucleotide, the sequence of the molecule can be determined.

ones produced by Illumina or PacBio [114]. Nanopore sequencing uses protein nanopores that are embedded in an electrically resistant membrane in order to determine the sequence of long DNA fragments [177]. A detector is connected to each nanopore and measures the electric current flowing through the pore [139]. DNA fragments pass the nanopores electrophoretically. Each base passing through the pore causes a change in current, generating a characteristic pattern which is known as "squiggle" [139] (Figure 1.5). These fluctuations are then processed into raw reads by a specific software and stored in FAST5 format [139]. A basecalling algorithm based on neural networks is used subsequently in order to decode the nucleotide sequences of the stored reads and converts them to FASTQ format [139].

1.4.4 Comparison of sequencing technologies

Depending on the sequencing machine used, Illumina sequencing produces paired-end reads of 150-300 bp in length [87]. These short reads are highly accurate, with error rates $< 0.1\%$ [83–86]. Due to their low error rates, short reads are well suited for detecting or genotyping SNPs and small indels. However, they are less useful for tasks that require longer sequence spans [192]. Especially repetitive genomic regions are difficult to access by these reads, as their short read length makes it hard to determine from which location of the genome they originated from. PacBio and Oxford Nanopore sequencing overcome these limitations since they produce much longer reads. However, they suffer from much higher error rates. PacBio CLR reads are typically between 5-60 kbp in length, with error rates between 8-15% [114, 157]. Nanopore sequencing can produce even longer reads (10-100 kbp, Ultra-long ONT reads reach lengths > 100 kbp) [114]. Read accuracies are on average 87-98% [114]. In contrast to short reads, long sequencing reads enable accessing more complex and repetitive regions of the genome. Due to their length, they can often span such regions, which makes it easier to determine their precise location when aligning them to a reference genome [114, 157]. This property makes long reads valuable for many applications, including calling or genotyping of structural variants, haplotype phasing or genome assembly [199]. PacBio CCS reads combine characteristics of short and long reads. Their read length is typically

between 10-30 kbp in length [114], while having an accuracy of $> 99\%$ (99.8% on average) [199]. Wenger et al. [199] compared read accuracy and mappability of PacBio CCS reads, Illumina reads and PacBio CLR data for human sample HG002 (NA24385). The longer read length of CCS reads increased their mappability. While 97.5% of reference genome version GRCh37 was mappable by CCS reads (excluding gap regions), Illumina reads could map to only 94.8% of the reference [199]. This enables to access medically relevant genes and other complex regions like HLA by CCS reads, which previously were not fully accessible by short reads [199]. In comparison to short reads, variant calling performance for small insertions and deletions (< 50 bp) is worse for PacBio CLR and Oxford Nanopore reads [48, 144, 171]. PacBio CCS reads, however, enable accurate small variant detection. While the accuracy for indel detection is comparable to Illumina reads (F-scores are 99.58% for CCS and 99.59% for Illumina data at coverage $35\times$ for HG003) [171], the accuracy of detecting SNPs with CCS reads is even better than for Illumina with the recently developed PEPPER-Margin-DeepVariant pipeline (F-scores are 99.90% for CCS and 99.63% for Illumina at coverage $35\times$ for HG003) [171]. Long reads are also favorable for structural variant detection. A comparison based on different datasets for human sample HG002 resulted in F-scores of around 96% for CCS reads, 94.6% for CLR data and only around 67.5% for Illumina [199].

1.5 Sequence alignment

A starting point for many downstream analyses are alignments of sequences to a reference genome. A reference genome is a DNA sequence used as a representative of the genome of a species, sometimes assembled from data of multiple individuals. Alignment aims at identifying the precise location in the genome from which a sequence originated. Due to genetic variants that are present in different individuals of the same species, the sequence of a query and the respective location in the reference genome are not always identical. Additionally, sequencing errors contained in reads or sequence segments resulting from DNA assembly (so-called contigs) lead to differences. Therefore, alignment requires approximate string matching algorithms in order to determine the location of a sequence in the reference genome. Many alignment methods are based on the Needleman-Wunsch algorithm [132] or the Smith-Waterman algorithm [178]. Sequence alignment is a powerful tool to compare DNA sequences. One example is aligning sequencing reads or assembly contigs to a reference genome. Multiple sequence alignment aims at computing alignments of more than two sequences, allowing to identify similarities and differences among multiple queries, even if reference genomes are not available.

1.6 Genotyping

Each chromosomal copy of an individual can carry genetic variants. Given a variant locus and possible allele sequences for this variant, genotyping describes the process of determin-

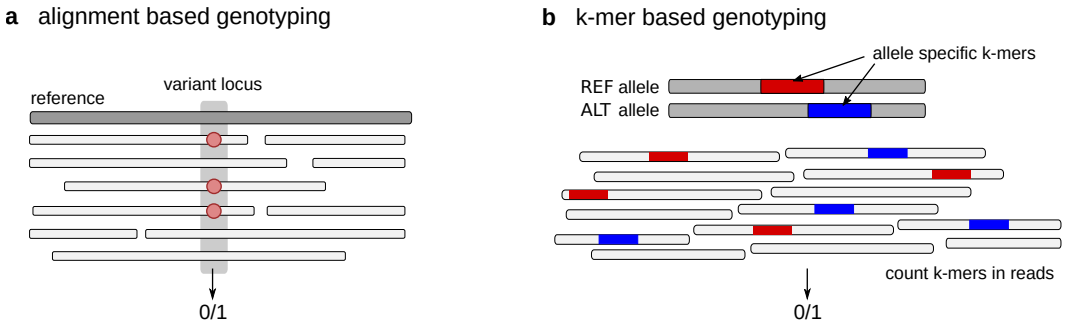


Figure 1.6: Genotyping. a Alignnment-based genotyping methods use sequence alignments spanning or aligning close to a variant locus in order to infer the genotype. Reads marked by a red circle carry the alternative allele, while the other reads carry the reference allele. This example illustrates a heterozygous ("0/1") genotype. **b** K-mer-based genotyping methods count allele-specific k-mers in the reads and infer genotypes based on the observed counts. The counts for the blue k-mer (specific to the reference allele) as well as the red k-mer (alternative allele) are the same, indicating a heterozygous genotype.

ing the most likely combination of alleles that a specific individual carries at the respective location in its genome. A common use case of genotyping is to determine genotypes of an individual for variants previously detected in other individuals. Genotyping is different from variant calling. While detection algorithms aim at finding locations of (possibly novel) genetic variants in the genome of an individual, genotyping methods focus on determining genotypes given a set of known alleles and typically cannot discover novel variation themselves. In this thesis, the same notation as used in VCF format is adapted (see Section 1.11.1 for details). Genotypes are represented as lists of alleles, separated by a "/". Typically, one refers to the reference allele as "0" and uses higher numbers to enumerate possible alternative alleles (see Section 1.11.1). In case of a biallelic locus in a diploid organism for example, the three possible genotypes are "0/0" (homozygous for reference allele), "0/1" (heterozygous) and "1/1" (homozygous for alternative allele).

In Chapter 3 of this thesis, a new approach to genotyping is presented which is based on short reads. This section here provides an overview of different computational approaches to genotyping from short and long sequencing data. The genotyping methods presented can be classified as either alignment-based or k-mer-based methods. Alignment-based approaches use alignments of sequencing reads to a linear reference genome or a graph structure representing possible reference and alternative alleles in order to infer genotypes (Figure 1.6a). K-mer-based approaches work with the raw reads and count allele-specific k-mers in the reads [49]. K-mers are short sequences of a fixed length k, whose counts in the reads provide evidence for the absence or presence of certain variant alleles (Figure 1.6b).

1.6.1 Short-read based genotyping

Various genotyping methods have been developed for short-read data. Many of them are based on alignments of short reads to a linear reference genome. GATK is a widely used

tool for detecting as well as genotyping SNPs and indels from short reads [39]. Its genotyping method is based on a Bayesian algorithm which computes likelihoods for all possible genotypes and can consider multiple samples simultaneously [39]. FreeBayes is a variant detection tool for SNPs and indels leveraging local haplotype information [60]. It uses a Bayesian model to compute genotype likelihoods based on the aligned sequencing reads and *a priori* expectations about the allele distributions within a population [60]. Platypus is another local haplotype-based small variant caller [158]. It uses local *de novo* assembly in order to increase the accuracy of variant detection and computes genotype likelihoods based on an expectation-maximization algorithm [158]. Besides methods that focus on small variants, several approaches to genotyping structural variants based on aligned short reads have been proposed. SVTyper is based on a maximum-likelihood Bayesian classification algorithm which infers SV genotypes from split-read and paired-end alignments [28]. DIGTYPER is a genotyping tool I developed previously. It uses maximum-likelihood estimation in order to compute genotype likelihoods for inversions and tandem duplications [47]. It considers the orientation of paired-end reads, the insert size and split-read alignments in order to find support for reference and alternative alleles [47].

Instead of using alignments to a linear reference genome, more recent genotyping approaches are based on graph structures that include possible alternative alleles in order to improve genotyping accuracy. This is especially useful for structural variants, since short reads often fail to properly align in the respective genomic regions, resulting in biased genotype estimates [26, 75, 175]. GraphTyper [52, 53] constructs a pangenome representation which includes alternative sequences and re-aligns sequencing reads to this graph. Genotypes for structural variants are computed by analyzing reads that align to SV breakpoints as well as alignment coverages [53]. Similarly, the SV genotyper Paragraph constructs a graph encoding reference and alternative alleles of input variants and performs local re-alignment of reads to this graph [26]. Genotype likelihoods are computed based on the resulting alignments [26]. Recently, genotyping of structural variants was demonstrated with the VG toolkit [75, 176]. Unlike other graph-based methods that locally re-align short reads previously mapped to a linear reference, short reads can be directly aligned to a pangenome graph in an efficient manner using the latest mapper Giraffe [176]. Genotypes are then computed based on the resulting alignments. While Giraffe was demonstrated to be fast on rather simple pangenome graphs excluding complex regions, we observed very high runtimes for more complex graphs (see Section 3.5.4).

Aligning reads to reference genomes or graph structures is time consuming (Chapter 3). Therefore, alternative approaches are k-mer-based. They use the raw, unaligned sequencing data in order to count allele-specific k-mers in the reads. Cortex, a very early graph-based tool, constructs a colored *de Bruijn* graph that can represent genomes of multiple samples. Genotype likelihoods are computed based on the structure of bubbles in the graph [89]. Dilthey et al. used a graph-based approach for genotyping the MHC region based on k-mer information. LAVA uses approximate k-mer matching in order to genotype known SNP loci

[172]. Dolle et al. introduced a data structure based on the Burrows-Wheeler transform (BWT) that can store sequencing data of a population. It can be used for fast genotyping of SNPs and indels by determining read support for allele-specific k-mer sequences [44]. BayesTyper is a k-mer-based tool for genotyping a wide range of genetic variants, including structural variants. Given variants to be genotyped, it constructs a graph representing reference and alternative alleles and computes genotype likelihoods based on counting allele specific k-mers in the sequencing reads of a sample [175]. While k-mer-based approaches are much faster than alignment-based methods for genotyping, they can struggle in more complex and repetitive regions of the genome that lack unique k-mers [49] (Chapter 3).

1.6.2 Long-read based genotyping

Besides methods that rely on short sequencing reads for genotyping, several approaches based on long reads exist. Genotyping SNPs and indels accurately from long reads is challenging due to the higher error rates of PacBio CLR and Oxford Nanopore reads (Section 1.4). However, due to their length, long reads provide access to more difficult regions that are otherwise inaccessible by short reads, enabling small variant detection also in such regions [48]. WhatsHap's genotyping algorithm is based on my previous work on genotyping. It formulates the model underlying the original WhatsHap phasing algorithm [143] (Section 1.7) in terms of a Hidden Markov Model in order to genotype SNPs and indels from long reads [48]. The Forward-Backward algorithm is used to compute genotype-likelihoods in a haplotype-aware manner by using the linkage information between heterozygous variants provided by the long reads in order to reduce the impact of sequencing errors [48]. Longshot can call and genotype SNPs and indels from noisy, long reads [50]. It uses a similar idea as WhatsHap and computes haplotype-aware genotypes based on performing haplotype assembly using HapCUT2 [50]. Furthermore, learning-based approaches were introduced in order to call and genotype SNPs and indels. Clair and DeepVariant can perform small variant detection and genotyping from noisy long reads using deep neural networks trained on read pileup data [117, 144]. NanoCaller and PEPPER-Margin-DeepVariant extend deep learning-based models by incorporating haplotype information for small variant detection [7, 171].

For structural variants, several tools exist which focus on the detection of structural variants from aligned long-read data, but also provide genotypes for their variant predictions [74, 91, 169]. SVs are typically identified based on signatures in the alignments, such as split-reads or gaps, and genotypes are inferred from the fraction of reads supporting reference or alternative alleles [74, 91, 169]. Several tools were developed specifically for genotyping known structural variant alleles in a new sample based on long reads. VaPoR genotypes known SVs by comparing read k-mers to the k-mer spectrum observed in reference and alternative alleles of a variant [205]. LRCaller genotypes SVs from Nanopore reads based on genotyping models using reference alignments and realignment to alternative alleles [13]. SV Jedi genotypes SVs from long reads based on alignments to reference

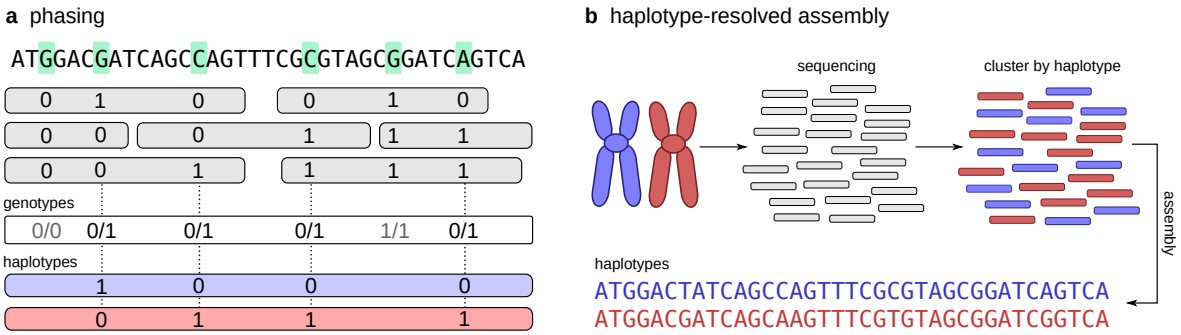


Figure 1.7: Haplotype reconstruction. **a** Read-based phasing. Reads are aligned to a reference genome, variants are genotyped and heterozygous variants are phased in order to determine which haplotype carries which allele ("0" = reference allele, "1" = alternative allele). **b** Haplotype-resolved assembly. Sequencing reads are clustered by haplotype and an assembly sequence for each haplotype is computed.

and alternative sequences of the input alleles [103]. Alignments are analyzed in order to determine the presence or absence of alleles in the data [103].

1.7 Phasing

Haplotype phasing aims at reconstructing the chromosomal copies of an individual based on genotype information and/or sequencing data.

Alignment-based phasing approaches use reads spanning at least two heterozygous variant locations in order to determine which alleles reside on the same haplotypes (Figure 1.7a). Such methods require reads aligned to a linear reference genome, as well as genotyped SNPs and indels. WhatsHap solves the minimum error correction problem (MEC) in order to partition aligned, long sequencing reads based on variant alleles they cover and reconstruct phased haplotype blocks from these read sets [143]. The HapCut algorithm constructs graphs from sequencing reads and computes max-cuts in these graphs in order to group reads by haplotype [10, 51]. HapCompass creates a graph based on sequencing reads and variants, and computes haplotypes based on spanning-trees [5]. Long reads are preferable for alignment-based phasing, since the distance between adjacent heterozygous variant positions often exceeds the length of short reads [204]. One limitation of alignment-based phasing is that most methods exclude structural variants, as high quality read alignments are often missing in regions where the genome of an individual is too different from the reference genome [204]. Also, even long reads might not cover adjacent variant positions. Therefore, alignment-based phasing usually leads to fragmented haplotype blocks and fails to provide haplotypes on the scale of a chromosome [204].

Population-based phasing methods use genotype data of large cohorts in order to reconstruct the haplotypes of an individual [19, 38, 58, 116]. The underlying idea is that individuals share haplotype segments due to common ancestry [58]. However, in contrast to read-based methods, population-based phasing is less powerful for rare variants and fails

for variants missing from the used reference panels [58].

The recent advances in long-read sequencing lead to improvements in methods for genome assembly and enable haplotype-resolved *de novo* assembly of individuals. Assembly-based phasing can provide haplotype sequences on the scale of a whole chromosome and overcomes many limitations of alignment-based and population-based phasing (see Section 1.8).

1.7.1 Evaluating phasing results

Several metrics are typically used to evaluate the quality of haplotype predictions produced by a tool, and in order to compare these phasing results to ground truth haplotypes. These metrics are used in Chapter 2 and are thus introduced in detail here.

Phase block N50

Especially read-based phasing tools are usually not able to generate a single, chromosome-scale haplotype prediction, but rather split their haplotype predictions into several phased blocks. Variants within the same block are phased relative to each other, but it remains unclear whether variant alleles from different phased blocks reside on the same haplotype or not. Cutting the phasing into several blocks is often necessary because there is not enough evidence in the data to connect phasing information between consecutive variants, for example because no read alignment exists that connects two variants [166].

The N50 metric is often used to evaluate the contiguity of phasing results [24, 166, 199], as well as the quality of genome assemblies [46, 59, 100, 145]. It is defined as the length of the shortest block such that 50% of the entire sequence length is contained in blocks with lengths longer or equal to this size [22]. More specifically, in order to compute the N50 of a set of phased blocks, one sorts them in descending order and, starting from the largest block, adds up block lengths until 50% of the total length of the underlying genomic sequence is reached. The length of the block at which the 50% threshold is reached is defined as the N50 value.

Switch errors and Hamming errors

The material presented in this section is re-used from my joint publication with Sven Schrinner and Rebecca Serra Mari on polyploid phasing, published in Genome Biology [166].

The switch error rate is a commonly used metric for evaluating phasing results [58, 143]. Given haplotype predictions and ground truth haplotypes, the switch error rate counts how many times the assignment between predicted and true haplotype blocks needs to be switched so as to reconstruct the true haplotypes from the predicted sequences. This metric has been mainly used in order to evaluate phasings for diploid individuals, but has also been extended to the polyploid case [12].

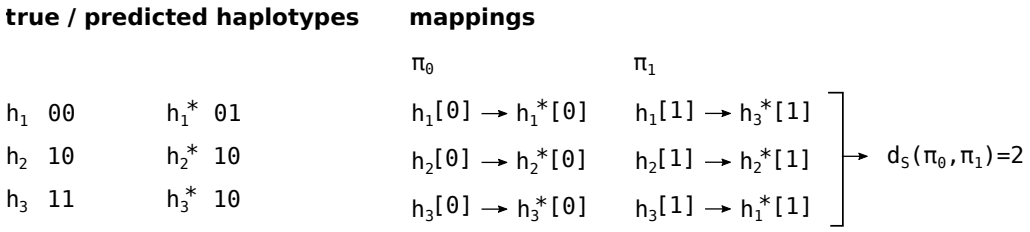


Figure 1.8: Computation of switch errors in polyploid setting. Shown are three ground truth haplotypes h_i and corresponding haplotype predictions h_i^* for two variant positions. On the right, the corresponding mappings π_0 and π_1 are shown. The mapping changes for h_1 and h_3 between positions 0 and 1 as a result of a haplotype switch between the two haplotypes. Therefore, $d_S(\pi_0, \pi_1) = 2$.

For ploidy k, let $h = \{h_1, ..., h_k\}$ be the set of ground truth haplotype sequences and $h^* = \{h_1^*, ..., h_k^*\}$ the predicted haplotype sequences. Let j be the number of variant positions and Π_j be defined as the set of one-to-one mappings between the true and predicted haplotypes h and h^*, such that for each $\pi \in \Pi_j$ it holds that $h_i[j] = h^*_{\pi(i)}[j]$ for all haplotypes h_i. The switch error rate is defined as:

$$\text{SER} = \min_{(\pi_1,...,\pi_m)\in\Pi_1\times...\times\Pi_m} \frac{1}{k(m-1)} \sum_{i=1}^{m-1} d_S\left(\pi_i, \pi_{i+1}\right)$$

where m is the number of variants and $d_S\left(\pi_i, \pi_{i+1}\right)$ the number of different mappings between π_i and π_{i+1}. See Figure 1.8 for an example on how to compute $d_S\left(\pi_i, \pi_{i+1}\right)$. If the genotype of h^* is not equal to the genotype of h for all variant positions, the set $\Pi_1 \times \ldots \times \Pi_m$ is empty and the switch error rate cannot be computed. When computing SER, we therefore only take those positions into consideration for which the genotypes of true and predicted haplotypes are the same.

Another metric that can be used to evaluate phasing results is the Hamming error rate. It is defined as:

$$\text{HE} = \min_{\sigma\in S_k} \frac{1}{k} \sum_{i=1}^{k} d_H(h_i, h^*_{\sigma(i)})$$

Here, S_k represents the permutation group on $\{1, ..., k\}$. $d_H()$ is the Hamming distance between two sequences.

As we noted in our publication [166], the Hamming error rate is more sensitive than the switch error rate. A single switch error in the middle of a haplotype block can lead to a maximal hamming error rate of 50%.

1.8 Genome assembly

Genome sequencing (Section 1.4) produces a high number of unordered sequencing reads each containing sequence information of a fragment of the individuals genome. Genome

assembly aims at assembling these sequence fragments in order to reconstruct the complete nucleotide sequence of an individual.

In 2001, the first drafts of a human genome sequence were published by the Human Genome Project and Celera Genomics [88, 191] marking a milestone in the field of genomics. However, the initial drafts were still incomplete. Since then, the sequence was continually improved, but only recently researchers were able to fill the remaining 8% of the human genome missing from previous versions [136]. The Telomere-to-Telomere (T2T) Consortium presented a first complete assembly of the human genome, providing gapless consensus sequences for all autosomes and chromosome X [136]. The advances in long-read sequencing made these recent breakthroughs possible, since the reads cover longer sequence spans which are necessary for large-scale genome assembly.

The latest human genome sequence is based on the homozygous CHM13hTERT cell line and is thus not yet haplotype-resolved [184]. Haplotype-resolved assembly attempts to assemble haplotypes separately instead of generating a single consensus sequence representing the genome of an individual (Figure 2.6b). Several approaches have been introduced that attempt to globally separate long-read data by haplotype prior to assembly, for example by additionally using parental data [98]. However, parental data is not always available. Other approaches combine long-read data with Hi-C data in order to obtain phased assemblies [59, 100], but phasing on scale of a chromosome remains problematic.

The PGAS pipeline aims to overcome these limitations by making use of Strand-seq data in order to produce chromosome-scale, phased assemblies of human genomes without using a reference genome [145]. It was used in order to generate assemblies for studies described in detail in Chapters 3 and 4, and will thus be briefly described here. In order to produce haplotype-resolved assemblies, the first step of the PGAS pipeline is to generate haplotype-unaware assemblies from long-read data [145]. The resulting contigs are then clustered using Strand-seq data, such that ideally, each resulting cluster corresponds to a chromosome [145]. Next, Strand-seq data and long reads are aligned to the clustered contigs in order to call SNPs, and SNPs are then phased in order to produce chromosome-scale haplotypes [145]. This phasing information is used in order to separate the long reads by haplotype [145]. Finally, a *de novo* assembly is computed for each haplotype based on the separated reads [145]. PGAS was demonstrated to produce high quality and highly contiguous assemblies, which are valuable for many downstream analyses and enable accurate phased variant detection, even for structural variants [46, 145].

Very recently, another method for haplotype-resolved assembly, Verkko, was introduced [154]. It improves and automates the assembly pipeline that was used by the T2T consortium in order to generate the CHM13 genome assembly. Verkko combines PacBio CCS data and Ultra-long Nanopore reads to compute complete assemblies of diploid genomes using an iterative, graph-based approach [154]. Applied to human sample HG002, Verkko assembled 20 out of 46 diploid chromosomes without any gaps with an accuracy of 99.9997% [154].

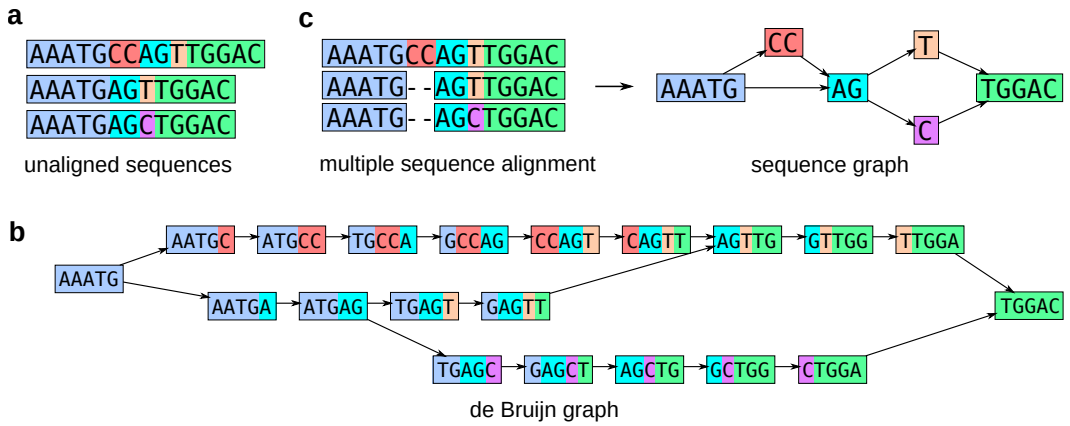


Figure 1.9: Pangenome representations. Overview of possible data structures used to represent a pangenome. **a** Unaligned collection of genomes. **b** *De Bruijn* graph representation. Each node corresponds to a k-mer sequence observed in the input genomes ($k = 5$). Two nodes are connected if they overlap by $k - 1$ bases. **c** Sequence graphs can be constructed from multiple sequence alignments of known genomes/haplotypes. Nodes corresponds to sequence segments, and edges connect nodes in a way that the resulting sequence reflects the input sequences.

1.9 Pangenomics

With advances in sequencing technologies and the development of methods for genome assembly, the number of sequenced and assembled genomes is steadily increasing [54, 187]. Typically, linear reference genomes are used in order to analyze sequencing data, e.g. by aligning the reads to it (Section 1.5). However, using linear reference genomes is problematic, especially when analyzing structural variants, due to the absence of alternative sequences from the reference genome [187]. This can cause a reference bias when aligning reads, which can lead to errors in downstream analyses [187]. Also, in a linear reference based setting, multiple genomes can only be compared indirectly via their relationship to the reference genome [54, 187]. Pangenomics aims at solving these issues by replacing the linear reference genome by structures that capture the full spectrum of sequence diversity inherent in the analyzed genomes [187]. The Computational Pan-Genomics Consortium defines the term *pangenome* to refer to "any collection of genomic sequences to be analyzed jointly or to be used as a reference" [187]. The aim is to provide a data structure that allows direct and unbiased comparisons of genomes [54].

Ideally, a pangenome should have the following properties. It should be constructable from different data types, such as sequencing reads, variants or assemblies, and should be easy to be updated as new datasets are available [187]. It should provide a coordinate system, positional access to genomic regions or variants and enable searching for specific sequences [187]. Also, the data structure should be fast and memory efficient [187].

Several data structures were proposed for representing pangenomes. In the simplest case, a collection of individual, unaligned genomic sequences can be interpreted as a pangenome

[54, 187] (Figure 1.9a). Furthermore, a pangenome can be constructed from a linear reference genome and variant alleles by inserting alternative alleles into the reference at their respective coordinates, resulting in a directed and acyclic graph [54]. The variant positions are represented in terms of bubble structures in these graphs [54]. Other approaches construct pangenomes based on *de Bruijn* graphs [77, 89, 120, 127]. In a *de Bruijn* graph, nodes correspond to k-mers and edges connect nodes if k-mers overlap by $k-1$ bases [54] (Figure 1.9b). Alternative approaches construct sequence graphs from multiple sequence alignments of assembled genomes or haplotypes [9, 61, 67, 104, 126]. In these graphs, nodes correspond to sequence fragments and edges connect nodes such that the resulting sequence reflects the genomic sequences from which the graph was constructed [54] (Figure 1.9c). Walks in these graphs correspond to recombinations of the input genomes and bubble structures represents variation between the genomes [54]. The linear reference sequence can additionally be incorporated as a linear path through such a graph, enabling projections to the reference genome [54].

Replacing the linear reference genome by a pangenome graph has been demonstrated to improve many downstream analyses, such as read alignment, variant calling or genotyping (Sections 1.6, 3, 4). Pangenome-based methods are especially important for studying regions difficult to access by sequencing reads, including repetitive regions and such that are poorly assembled in the linear reference genome [187]. Especially structural variants are often located in these regions and are therefore often missed by alignment-based methods [24, 46, 49, 198, 206]. Thus, pangenome-based methods provide the potential to detect such structural variants, genotype them across cohorts and study associations with diseases [46, 49, 187]. The field of pangenomics is still relatively young and actively researched. Larger consortia working in this area include the Human Pangenome Reference Consortium (HPRC) [90, 113] or the Human Genome Structural Variant Consortium (HGSVC) [24, 46]. The work I have contributed to these projects is presented in Chapters 3 and 4 of this thesis. It will be demonstrated how pangenomes improve genotyping performance of SNPs, indels and especially SVs, and that they enable accessing regions and variants that previously were inaccessible by purely short-read based approaches.

1.10 Mathematical background

The mathematical model behind PanGenie, the genotyping approach that is presented in Chapter 3, is a Hidden Markov Model (HMM). Hidden Markov Models describe stochastic processes and due to their flexibility, are widely used in many areas of research, such as speech recognition or bioinformatics [48, 57, 92, 111, 128, 150]. In this thesis, an HMM is used to compute genotype likelihoods for genetic variants based on sequencing data and a panel of known haplotype sequences. This section introduces the concept of Hidden Markov Models, provides a formal definition and demonstrates how to compute posterior state probabilities.

1.10.1 Hidden Markov Models

Hidden Markov Models consist of two discrete-time stochastic processes modeled over collections of random variables $(X_t)_{t\in\mathbb{N}^+}$ and $(Y_t)_{t\in\mathbb{N}^+}$ which take values on discrete state spaces $Q = \{q_1, \ldots, q_N\}$ (hidden states) and $V = \{v_1, \ldots v_M\}$ (observable states), respectively [11, 149]. Starting from any of the hidden states selected according to an initial state distribution π at time $t = 1$, the model can switch to a different hidden state with a certain probability defined by a transition probability matrix P. Hidden states produce outputs v_i according to an emission probability matrix B. While the sequence of emissions produced by an HMM, $Y_1 = v_{k_1}, \ldots, Y_T = v_{k_T}$, can be observed, the underlying sequence of hidden states, $X_1 = q_{k_1}, \ldots, X_T = q_{k_T}$, that produced them remains unknown. However, the emissions allow to draw conclusions about the hidden states (see Section 1.10.2). An important characteristic of Hidden Markov Models is the Markov property. It states that the probability of transitioning to the next state at time $t + 1$ only depends on the state in which the model is at time t, and anything that happened at earlier steps $< t$, does not have any further influence. Mathematically, this property is defined in terms of the equation shown below [149]:

$$P(X_{t+1} = q_{k_{t+1}} | X_1 = q_{k_1}, X_2 = q_{k_2}, \ldots, X_t = q_{k_t}) = P(X_{t+1} = q_{k_{t+1}} | X_t = q_{k_t})$$

A Hidden Markov Model can be formally defined based on the following elements [149]:

- a finite number of N hidden states: $Q = \{q_1, \ldots q_N\}$
- a finite number of M observable states (emissions): $V = \{v_1, \ldots v_M\}$
- transition probabilities: $P = \{p_{ij}\}$, such that $p_{ij} = P(X_{t+1} = q_j | X_t = q_i)$
- emission probabilities: $B = \{b_j(k)\}$, such that $b_j(k) = P(Y_t = v_k | X_t = q_j)$
- initial state distribution: $\pi = \{\pi_i\}$, such that $\pi_i = P(X_1 = q_i)$

1.10.2 Forward-Backward algorithm

The Forward-Backward algorithm determines the probability that the model was in a certain state q_i at a certain time t, given the full sequence of emissions. More formally, given a sequence of T observations $Y_1 = v_{k_1}, \ldots, Y_T = v_{k_T}$, it computes the probability: $P(X_t = q_i | Y_1 = v_{k_1}, \ldots, Y_T = v_{k_T})$, for all $t = 1, \ldots, T$ and all $q_i, i = 1, \ldots, N$. The algorithm consists of two steps, the computation of Forward probabilities, and the computation of Backward probabilities. Probabilities are later combined to obtain posterior state probabilities.

Forward Probabilities

The Forward probability is defined as: $P(Y_1 = v_{k_1}, \ldots, Y_t = v_{k_t}, X_t = q_i)$ [149]. It describes the probability of observing the first t emissions and being in state q_i at time t, and can be

computed recursively as shown below [149]:

$$\alpha_1(i) = \pi_i \cdot b_i(v_{k_1}) \qquad \text{for } i = 1, \ldots, N$$
$$\alpha_{t+1}(i) = b_i(v_{k_t}) \sum_{j=1}^{N} \alpha_t(j) \cdot p_{ji} \qquad \text{for } t = 1, \ldots, T \text{ and } i = 1, \ldots, N$$

Backward Probabilities

The Backward probability is defined as: $P(Y_{t+1} = v_{k_{t+1}}, \ldots, Y_T = v_{k_T} | X_t = q_i)$ [149]. It describes the probability of observing the partial sequence of emissions starting from $t+1$, given that at time t, the model is in state q_i, and can be computed using the following recursion [149]:

$$\beta_T(i) = 1 \qquad \text{for } i = 1, \ldots, N$$
$$\beta_t(i) = \sum_{j=1}^{N} b_j(v_{k_{t+1}}) \cdot p_{ij} \cdot \beta_t(j+1) \qquad \text{for } t = T-1, T-2, \ldots, 1 \text{ and } i = 1, \ldots, N$$

Posterior Probabilities

Finally, the probability to be in a state q_i given the full sequence of observations can be computed from the Forward and Backward probabilities. Using the definition of conditional probabilities, the sought probability is defined as:

$$P(X_t = q_i | Y_1 = v_{k_1}, \ldots, Y_T = v_{k_T}) = \frac{P(X_t = q_i, Y_1 = v_{k_1}, ..., Y_T = v_{k_T})}{P(Y_1 = v_{k_1}, ..., Y_T = v_{k_T})}$$

It holds that [11]:

$$\begin{aligned} P(Y_1 = v_{k_1}, ..., Y_T = v_{k_T}, X_t = q_i) &= P(Y_1 = v_{k_1}, ..., Y_t = v_{k_t}, X_t = q_i) \\ &\quad \cdot P(Y_{t+1} = v_{k_{t+1}}, ..., Y_T = v_{k_T} | X_t = q_i) \\ &= \alpha_t(i)\beta_t(i) \end{aligned}$$

Therefore, the posterior state probabilities for all $i = 1, \ldots, N$ and all $t = 1, \ldots, T$ can be computed as [11]:

$$P(X_t = q_i | Y_1 = v_{k1}, \ldots, Y_T = v_{kT}) = \frac{\alpha_t(i)\beta_t(i)}{\sum_j \alpha_t(j)\beta_t(j)}$$

1.11 File formats

This section provides an overview of the different file formats that are relevant in this thesis. It presents the VCF format which is used to represent genetic variation, the SAM/BAM format

```
##fileformat=VCFv4.2
##FILTER=<ID=PASS,Description="The variant passed all filters">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##INFO=<ID=AC,Number=A,Type=Integer,Description="Number of alternate alleles in genotypes">
#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT HG00438 HG00621 HG00673
chr1 594965 .  CCACC CT,CGGCC,CTGCC 60 .  AC=2,1,1 GT 0/1 1|0 2|3
```

Figure 1.10: VCF format example (multiallelic). Example of a multiallelic VCF file. Meta information lines start with "##" and provide general information on the VCF file [162]. The header line starts with a single "#" and defines the columns of the data lines [162]. Here, the VCF contains a single data line which defines a multiallelic variant for which three alternative alleles exist. The VCF provides genotype information on three samples. Genotypes of the second and third sample are phased. The example shown here was extracted from a VCF generated by our variant calling pipeline presented in Chapter 3.

used to store sequence alignments, and the FASTA/FASTQ format which is used to store genetic sequences, such as sequencing reads or reference genomes.

1.11.1 VCF format

The VCF format is a text format used to describe genetic variation relative to a linear reference genome. Files contain meta-information lines, a header line and data lines [162]. The meta information lines define "key = value" pairs that are later used in the data lines to describe variant information [162]. The header defines the columns for which information is provided in the data lines. The first one is the CHROM column, specifying the chromosome on which a variant is located, the second column, POS, specifies the exact coordinate. In the ID column, a string identifier can be provided for the variant record. The REF and ALT columns contain reference and alternative sequences of the variant. If the variant is multiallelic meaning more than one alternative allele exists at this position, the respective alternative allele sequences are provided as a comma-separated list in the ALT field. The QUAL and FILTER columns can be used to give information on the quality of the variant. Additional information on the locus can be provided in the INFO column in terms of "key = value" pairs previously defined in the meta information lines. If sample specific genotype information is available, an additional column for each sample is provided as well as a FORMAT column, which specifies the type of data and the order of the genotype information provided in the sample columns [162]. Genotype information typically includes a genotype of the sample, that describes which alleles the sample carries on its chromosomal copies. Such a genotype consists of a list of numbers, separated by "/" or "|", depending on whether the genotype is unphased or phased, respectively. The numbers give the index of the respective alleles. "0" stands for the reference allele, and a number $i > 0$ refers to the i-th alternative allele provided in the ALT column [162].

Examples for VCF files are shown in Figures 1.10 and 1.11. Both VCF files describe the same genetic variation but in different formats: the example shown in Figure 1.10 uses a single, multiallelic variant record, while the example shown in Figure 1.11 uses a biallelic

```
##fileformat=VCFv4.2
##FILTER=<ID=PASS,Description="The variant passed all filters">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##INFO=<ID=AC,Number=A,Type=Integer,Description="Number of alternate alleles in genotypes">
#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT HG00438 HG00621 HG00673
chr1 594965 .  CCACC CT 60 .  AC=2 GT 0/1 1|0 0|0
chr1 594965 .  CCACC CGGCC 60 .  AC=1 GT 0/0 0|0 1|0
chr1 594965 .  CCACC CTGCC 60 .  AC=1 GT 0/0 0|0 0|1
```

Figure 1.11: VCF format example (biallelic). Example of a VCF file in biallalic representation. Meta information lines start with "##" and provide general information on the VCF file [162]. The header line starts with a single "#" and defines the columns of the data lines [162]. The VCF file contains a separate data line for each of three possible alternative alleles at position chr1:594965. Genotypes indicates the presence ("1") or absence ("0") of the allele in the genome of a sample. The example shown here was extracted from a VCF generated by our variant calling pipeline presented in Chapter 3.

representation and contains a separate record for each alternative allele. In the biallelic version, "1" is used to indicate the presence of the respective alternative allele in a genotype, while "0" indicates absence. While both representations are equivalent, the biallelic version is often preferred because it simplifies downstream analyses.

1.11.2 SAM/BAM format

The SAM format is a text format that is used to describe sequence alignments [108]. It consists of a header section and an alignment section [108]. The header lines start with character "@" and define "TAG:VALUE" fields which are used in the alignment section to describe properties of the alignments [108]. In the alignment section, each line represents a linear alignment of a sequence segment [108]. Each line consists of at least 11 fields defining the alignment of the segment against the query sequence. The fields contain information such as the position of the alignment relative to the query sequence, the quality of the alignment or the CIGAR string, which describes the alignment itself by specifying matches, mismatches, insertions and deletions [108]. BAM files are SAM files compressed in BGZF format [108].

1.11.3 FASTA/FASTQ format

The FASTA format is used to store sequence information. It contains blocks describing sequence segments. Each such block starts with a description line (starting with ">") and is followed by lines of sequence data [131]. The FASTA format can be used to store nucleic acid sequences and amino acid sequences, which must be represented in the standard IUB/IUPAC codes [131]. The FASTQ format is an extension of the FASTA format which additionally allows to store position-wise quality scores for the sequences. Similarly to FASTA files, FASTQ files consist of blocks describing sequences. Each block consists of a header line (starting with "@"), lines of sequence data, a separator line containing a single "+", and position-wise quality scores (Phred scaled) [82]. In this work, FASTA/FASTQ files are

mainly used to store nucleic acid sequences, such as sequencing reads, assembly contigs or reference genomes. Nucleic acids are encoded using letter “A” for adenine, “C” for cytosine, “G” for guanine and “T” for thymine, as well as “N”, if sequence information is missing [131].

Chapter 2

Reference-based haplotype phasing

Phasing aims at reconstructing the haplotype sequences of diploid or polyploid individuals (Section 1.7). Assigning variant alleles to haplotypes is important in many applications, for example in order to identify selective pressures or subpopulations in population studies [1, 143, 188], in clinical genetics [179], or in order to link disease-causing SNPs to one another [69, 143].

Many phasing approaches are based on alignments of long sequencing reads to a reference genome (Section 1.7). The high sequencing error rates of these reads make accurate phasing of SNPs difficult, especially for highly variable genomes. The first part of this chapter introduces a new phasing algorithm designed for polyploid species and presents an application of this algorithm to data of the tetraploid potato. The second part of this chapter presents an application of the diploid phasing tool WhatsHap to the new, highly accurate PacBio Circular Consensus reads and demonstrates that these reads enable high quality phasing without requiring additional short-read data for variant detection.

2.1 Accurate polyploid phasing from long reads

This section presents an approach to polyploid phasing based on long sequencing reads. In this project, I shared first authorship with Rebecca Serra Mari and Sven Schrinner. The results of this study were published in Genome Biology [166]. My main contribution to this project was to analyze a tetraploid potato dataset using the new algorithm. Sections 2.1.1, 2.1.2, 2.1.3 and 2.1.4 re-use some material from this paper. Sections 2.1.2 and 2.1.3 summarize joint work with Sven Schrinner and Rebecca Serra Mari for this publication. Author contributions and publication details are provided in Section E.1.

2.1.1 Introduction

This section re-uses some material from [166].

Unlike diploid organisms, which carry two homologous copies of their autosomes, polyploid

species have more than two such copies. Especially plants are often polyploid (Section 1.7). Being able to phase such organisms is crucial for breeding and/or genome engineering, since it helps to improve yield qualities of important crops by improving associations between genotypes and phenotypes [101]. The phasing problem is more complicated in the polyploid setting than it is in the diploid case. When phasing biallelic, heterozygous variants in diploid organisms, it is sufficient to know the configuration of one haplotype, as the other one will be complementary. However, in polyploid organisms, this is not the case. Some of the haplotypes of an individual can be locally identical, which makes it difficult to reconstruct them. Autopolyploid species are especially challenging to phase and assemble since they occur as a result of genome duplications [170].

Many phasing approaches for the diploid setting solve the Minimum Error Correction (MEC) problem. The idea is to partition the reads into distinct sets that correspond to different haplotypes by minimizing the number of corrections that need to be applied to their sequences. However, this model is based on the assumption that haplotypes are different and thus cannot describe cases where haplotypes are locally identical. In such cases, sequencing errors in the reads can lead to wrong haplotype assignments. An example is provided in Figure 2.1. On the left, the four haplotypes of a polyploid individual are shown. The blue and pink haplotypes are identical. Given sequencing reads, the MEC approach aims at partitioning the reads into sets that correspond to the same haplotype, ideally leading to balanced partitions (middle). However, since two haplotypes are identical, the MEC model assigns most of the reads originating from these two haplotypes to one partition, and collects noisy reads in a separate, sparse partition in order to minimize the overall MEC score (right), leading to wrong haplotypes. This example demonstrates that approaches developed for diploid phasing are not always applicable in a polyploid setting. Also, many of these approaches become infeasible when generalized to polyploid genomes [17].

Several methods for polyploid phasing based on long sequencing reads already exist. HapCompass is based on spanning trees and the Minimum Weighted Edge Removal (MWER) [5, 6]. HapTree uses a maximum-likelihood approach in order to construct the most likely haplotypes given the aligned sequencing reads [12]. SDHap uses a semi-definite programming approach based on an approximate MEC criterion [34]. All three methods have been evaluated based on simulated data in a study by Motazedi et al.. Results show that none of them was useful in practice, either because they scaled poorly when applied to larger genomes, or because they produced inaccurate phasing results [130]. More recently, H-PoP was demonstrated to outperform these methods in terms of phasing accuracy and runtime [202]. It is based on Polyploid Balance Optimal Partition (PBOP) and aims to partition the sequencing reads by haplotype [202]. H-PoP's underlying model can be interpreted as a generalization of the MEC problem to the polyploid case. If genotyping information is available for input variants, these constraints can be added to the model, the resulting extension of H-PoP is referred to as H-PoPG.

A few other approaches have been suggested that do not scale well to whole-genome

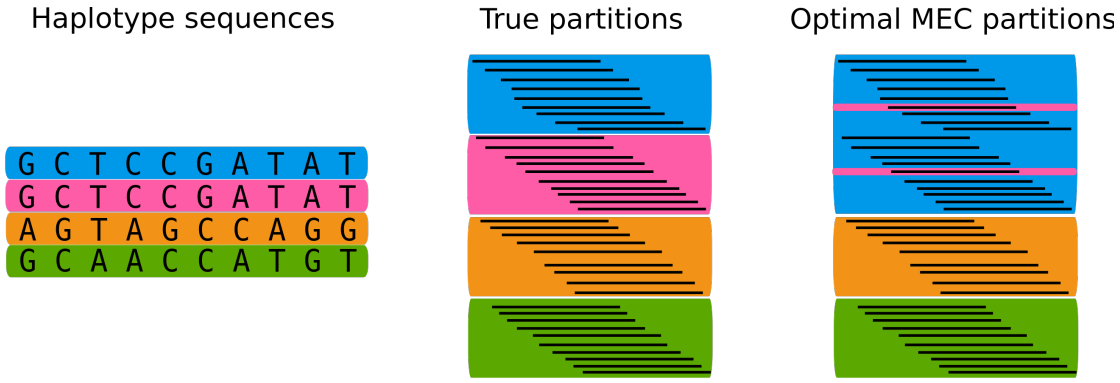


Figure 2.1: MEC problem in polyploid setting. Four haplotype sequences of a polyploid individual are shown on the left. The blue and pink haplotypes are identical. In the middle, the desired partitioning of sequencing data for the respective sample is shown. Ideally, reads originating from the four haplotypes can be clustered into sets of similar sizes. On the right, likely outcomes of clusters computed by the MEC model are shown. MEC fails to properly handle identical haplotype sequences, and combines reads from the blue and pink haplotypes into one big cluster. Noisy reads might end up in a sparse cluster, likely leading to wrongly reconstructed haplotype sequences. Figure taken from [166].

single-individual phasing, including PolyHarsh [73], which uses Gibbs Sampling, TriPoly [129], which requires family data, and SDA [23], which was designed to resolve multicopy duplications during genome assembly.

In the following, a new algorithm, `whatshap polyphase`, for polyploid phasing is introduced that overcomes the limitations mentioned above. It is designed to properly handle locally identical haplotypes by taking read coverage into account, and consists of two steps, a read clustering step and a newly established "threading" step. The focus of the following sections is on the application of `whatshap polyphase` to a tetraploid potato dataset, which demonstrates that the algorithm is scalable in practice and delivers high quality haplotype predictions that enable haplotype-resolved assembly of genes, providing biological insights into the genomes of polyploid species.

2.1.2 Phasing algorithm

The phasing algorithm presented here was mainly developed by Sven Schrinner and Rebecca Serra Mari. This subsection provides a summary of their work which was presented in our joint publication [166]. Some material from this publication is re-used. Formal definitions and mathematical details are skipped here and can be found in [166].

The input required by `whatshap polyphase` is a BAM file with aligned sequencing reads, a VCF file containing variants and unphased genotype information, and a reference genome. The first phase of `whatshap polyphase` groups input reads that likely belong to the same haplotype. For each pair of reads, the clustering step computes a similarity score that is positive if two reads originate from the same haplotype and negative if they stem from

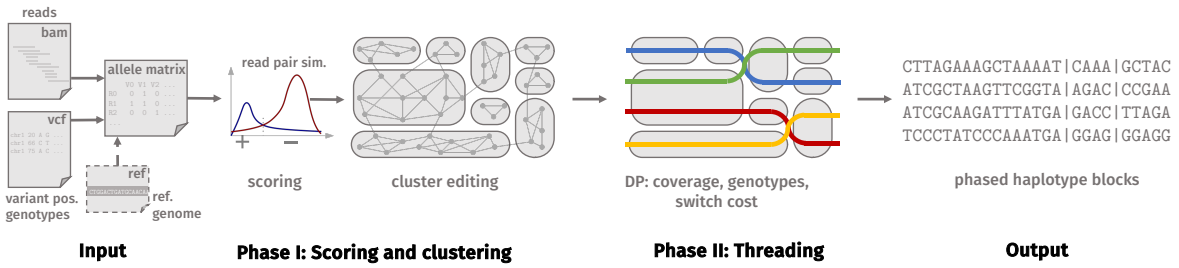


Figure 2.2: WhatsHap polyphase overview. Input. The inputs to the algorithm are aligned sequencing reads in a BAM file, genotyped variants in a VCF file and the corresponding reference genome. **Phase I.** Input reads are clustered based on their similarity using a cluster editing approach. **Phase II.** Haplotypes are reconstructed based on "haplotype threading" by assembling each haplotype through a sequence of read clusters. Haplotypes that are locally identical can be threaded through the same clusters. **Output.** As an output, the algorithm generates the predicted haplotype blocks. Figure taken from [166].

different haplotypes. A complete graph is constructed in which each sequencing read is represented as a node and each edge is labeled with the similarity of the underlying read pair. In order to generate read clusters, the algorithm uses cluster editing [203] to transform the graph into disjoint cliques. Since the cluster editing problem is NP-hard, a previously introduced heuristic [15] is used to efficiently compute a solution.

In the second phase of the algorithm, the actual phasing results are computed based on the read clusters using haplotype threading. Given the ploidy p of the data, p haplotypes need to be reconstructed. For each such haplotype, the idea is to assemble a sequence of clusters by choosing one cluster at each variant position that the haplotype is "threaded" through. The number of haplotypes that can be threaded through the same cluster at a position depends on the position-wise read coverage of the cluster. Since some haplotypes can be locally identical for polyploids, clusters might contain reads originating from different haplotype sequences that are identical or similar in the respective region. The coverage of the cluster helps to find out how many haplotypes are collapsed at a position and are thus allowed to use the same cluster. Haplotypes threaded through the same cluster at a variant position must also carry the same allele which is derived from the consensus sequence of the reads contained in the cluster. Therefore, another constraint that needs to be considered when choosing clusters is that the cluster selection needs to be compatible with the input genotypes. Additionally, the haplotype threading algorithm aims to minimize the number of haplotype switches between clusters so that haplotypes are encouraged to remain in the same cluster as long as possible. The cluster assignments are used in a final step in order to construct p haplotypes which are then output in terms of a phased VCF file. Figure 2.2 provides an overview of the phasing algorithm.

2.1.3 Evaluation on artificial polyploid humans

This section provides a summary on joint work of Sven Schrinner, Rebecca Serra Mari and myself published in [166]. Some material from this publication is re-used. I wrote the Snakemake pipeline that constructs the tetraploid datasets from simulated and real data, runs the phasing tools and computes phasing statistics. Sven Schrinner extended this pipeline for penta- and hexaploid cases. Rebecca Serra Mari evaluated and compared results in collapsing and non-collapsing regions.

We combined real PacBio CLR sequencing data for three human samples (NA19240, HG00514, HG0733) at different coverage levels in order to generate artificial tetraploid, pentaploid and hexaploid datasets for human chromosome 1. For all three samples, high-quality ground truth haplotype information is available [24]. In addition, equivalent simulated read datasets were produced for these samples using PBSIM [138] as a read simulator. `whatshap polyphase` and H-PoPG were evaluated on these datasets and results were compared to the ground truth haplotypes by computing the switch error rate and the hamming error rate (Section 1.7.1). Both tools use different strategies in order to define blocks of variants that they can phase together. H-PoPG maximizes the block length by cutting haplotypes only between consecutive variants that are not connected by any read alignment. `whatshap polyphase` implements different approaches to define phased blocks. Per default, block cuts are introduced between two adjacent positions whenever there are not enough reads to connect at least three of the haplotypes. In addition, a block is cut whenever at least two haplotypes switch clusters during the threading step, or whenever at least one of the haplotypes sharing the same cluster in a collapsed region switches to another one afterwards. In order to properly compare both tools, `whatshap polyphase` was run using the same phase block definition as H-PoPG. Across all datasets tested, `whatshap polyphase` produced more accurate results compared to H-PoPG when considering the switch error rate (Table A.1). Switch error rates were around 30-40% lower than for H-PoPG when using the same strategy to define phase blocks as H-PoPG. Switch-error rates are even lower when using a more sensitive definition of phased blocks, but more and shorter haplotype blocks are produced in turn (Table A.1). Compared to H-PoPG, `whatshap polyphase` works especially well in regions with locally identical haplotypes ("collapsing regions"). Here, the switch error rate was around 3.06 and 2.76 times higher for H-PoPG at coverages 40× and 80× using real reads, compared to the non-collapsing regions, where these factors were only 1.19 and 1.12, respectively, when using the same phase block definition as H-PoPG (Table A.2). In summary, the results demonstrate that `whatshap polyphase` produces high quality phasing results across different ploidies and improves over current state-of-the-art tools. Results especially improve in regions with identical haplotypes. This demonstrates the ability of `whatshap polyphase` to handle regions typically challenging for existing tools.

2.1.4 Analysis of potato data

This section re-uses material presented in [166].

In order to demonstrate a use case of our phasing algorithm, we applied it to real sequencing data of a tetraploid potato genome (*Solanum tuberosum*), for which paired-end short Illumina and long Oxford Nanopore reads were available. In a first step, we aligned the reads produced by the different technologies to the potato reference genome published by the Potato Genome Sequencing Consortium (PGSC) [71]. We observed unbalanced coverage distributions for the alignments, especially for the short Illumina reads, hinting towards a high number of structural variations and rearrangements being present in the data (Figure 2.3a), which confirms that the potato genome is highly heterogeneous [71]. Thus, the Illumina reads are ill-suited for reliable variant calling as their short length makes it more difficult to unambiguously align them to the reference. We therefore relied on the much longer Nanopore reads to identify SNPs that we could use for phasing. However, Oxford Nanopore reads typically come with high sequencing error rates, complicating the calling process. In order to obtain reliable variant positions and genotypes from these error-prone reads, we ran an error correction pipeline [153] that uses the short Illumina reads in order to reduce the number of sequencing errors in the long reads. Figure 2.3c shows an exemplary IGV [160] screenshot of uncorrected (top) and corrected reads (bottom) for the FRIGIDA-like protein 5 isoform X2 gene. Next, we ran minimap2 [106] to align the corrected Nanopore reads to the potato reference genome and called SNPs inside of all gene regions using FreeBayes [60] with additional parameters: `-p 4 --no-indels --no-mnsp --no-complex`. As base qualities are not produced during error correction and FreeBayes needed them in order to compute genotypes, we added a constant quality of 40 for all bases to the BAM file before calling SNPs. Finally, we ran `whatshap polyphase` in order to phase the variants with option `--verify-genotypes`. This option invokes an additional step prior to phasing, which re-genotypes all variants and only keeps those positions for which the computed genotype matches the one reported by FreeBayes. For determining the genotype of a position, we implemented a simple algorithm that calculates the fraction of reference and alternative alleles that cover a variant and compares it to the fractions that we would expect for all possible genotypes. We then assign the genotype whose expected fractions of reference and alternative alleles best match the ones observed in the data.

We focused on the potato genes [71] as they are biologically interesting for phasing. For the total of 36,274 genes containing heterozygous variants after calling and retyping, 91% could be (at least partially) phased by `whatshap polyphase`. On average, about 2.13 phased blocks were produced per gene. Furthermore, for each gene, we determined the number of phased variants inside the longest phased block. We observed that a large fraction of genes, including many long genes, can be fully phased, see Figure 2.3b. We also evaluated the percentage of phased genes in relation to their level of heterozygosity, but could not observe a strong dependency, see Figure 2.4.

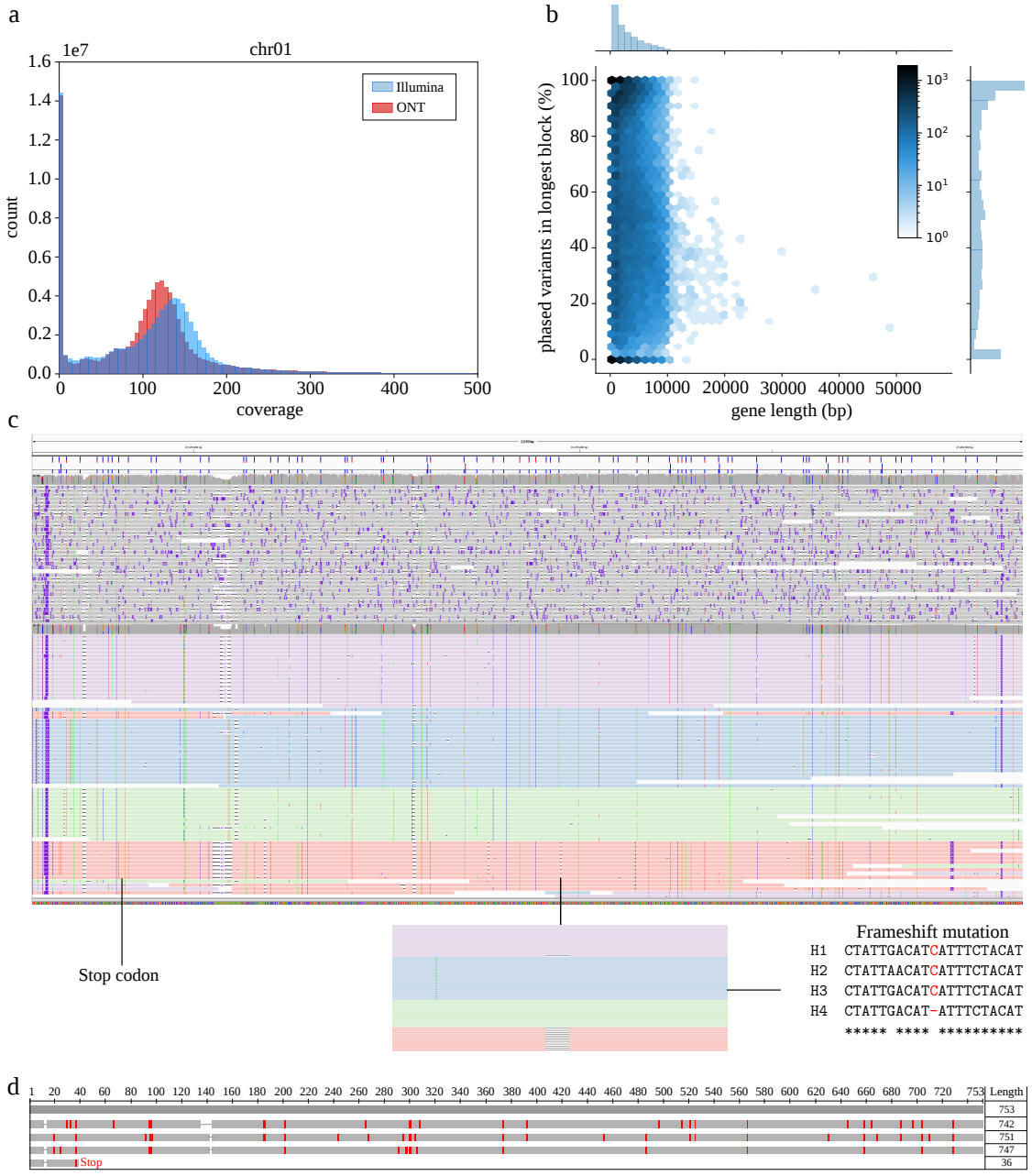


Figure 2.3: Phasing of a potato genome. **a** Per-base coverage distribution of Illumina and ONT MinION alignments on Chr01. **b** Fraction of phased variants in relation to gene length. The x-axis shows the gene length and the y-axis the percentage of phased variants in the longest block. Axis histograms and hexagons illustrate the distribution of data points. **c** IGV [160] screenshot showing alignments of uncorrected (top) and corrected MinION reads (bottom) of FRIGIDA-like protein 5 isoform X2 gene on Chr04. The corrected reads are colored (red, green, blue, purple) according to the haplotypes `whatshap polyphase` assigned them to. **d** Multiple sequence alignment of the ORFs detected in the four haplotype sequences. The uppermost gray sequence represents the reference, the others correspond to the four haplotypes (same order as in panel c). Figure taken from [166].

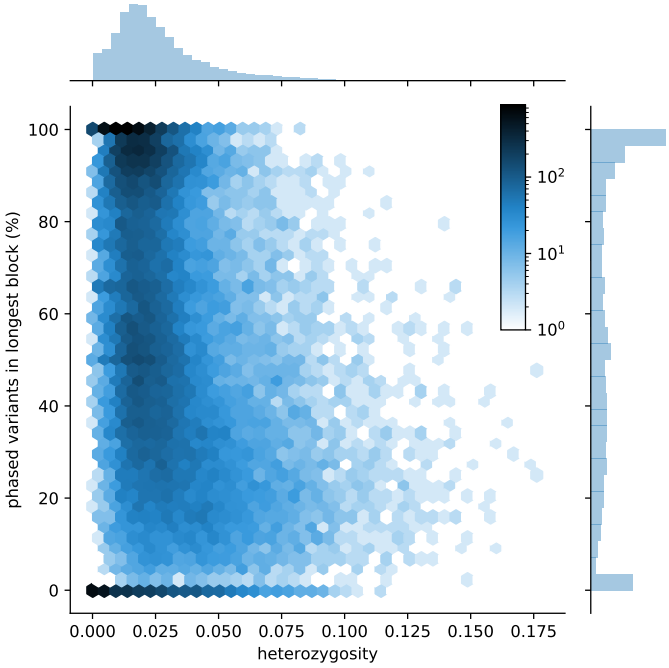


Figure 2.4: Fraction of phased variants in relation to heterozygosity in the potato genes. We determined the heterozygosity level of a gene as the fraction of heterozygous positions. The plot shows the fraction of phased variants (y-axis) in relation to heterozygosity level (x-axis). Axis histograms and hexagons illustrate the distribution of data points. Figure taken from [166].

We used the FRIGIDA-like protein 5 isoform X2 (accession: XP_015169713) gene as an example to demonstrate how `whatshap polyphase` enables haplotype-resolved assembly. We extracted the phasing of the longest phased block reported for this gene and separated the reads by haplotype as follows. We extended the command `whatshap haplotag`, which was previously implemented for the diploid version of WhatsHap, to the polyploid case. Given a phased VCF with predicted haplotypes and a BAM file with sequencing reads, we assign each read to the haplotype it is most similar to in terms of the alleles observed at variant positions in the read. This assignment is stored by tagging the respective sequences in the BAM file, which enables visualizing the haplotype clusters by programs like IGV [160]. Furthermore, we extended the subcommand `whatshap split` to higher ploidies, which can be used to split tagged reads by haplotype and store them in separate files. The reads shown in Figure 2.3c are colored according to the resulting haplotype assignments. In the next step, we separately ran wtdbg2 [161] (with options `-x ccs -g 1m`) on each haplotype-specific read set to produce local assemblies of the four haplotypes. Figure 2.5 shows a visualization of a multiple sequence alignment of these haplotypes. The red sequence corresponds to the reference genome. Figure 2.5a shows the whole graph, panel 2.5b the region corresponding to the FRIGIDA-like protein 5 isoform X2 gene.

We ran the NCBI ORFfinder [200] on each of the assemblies and detected a long open

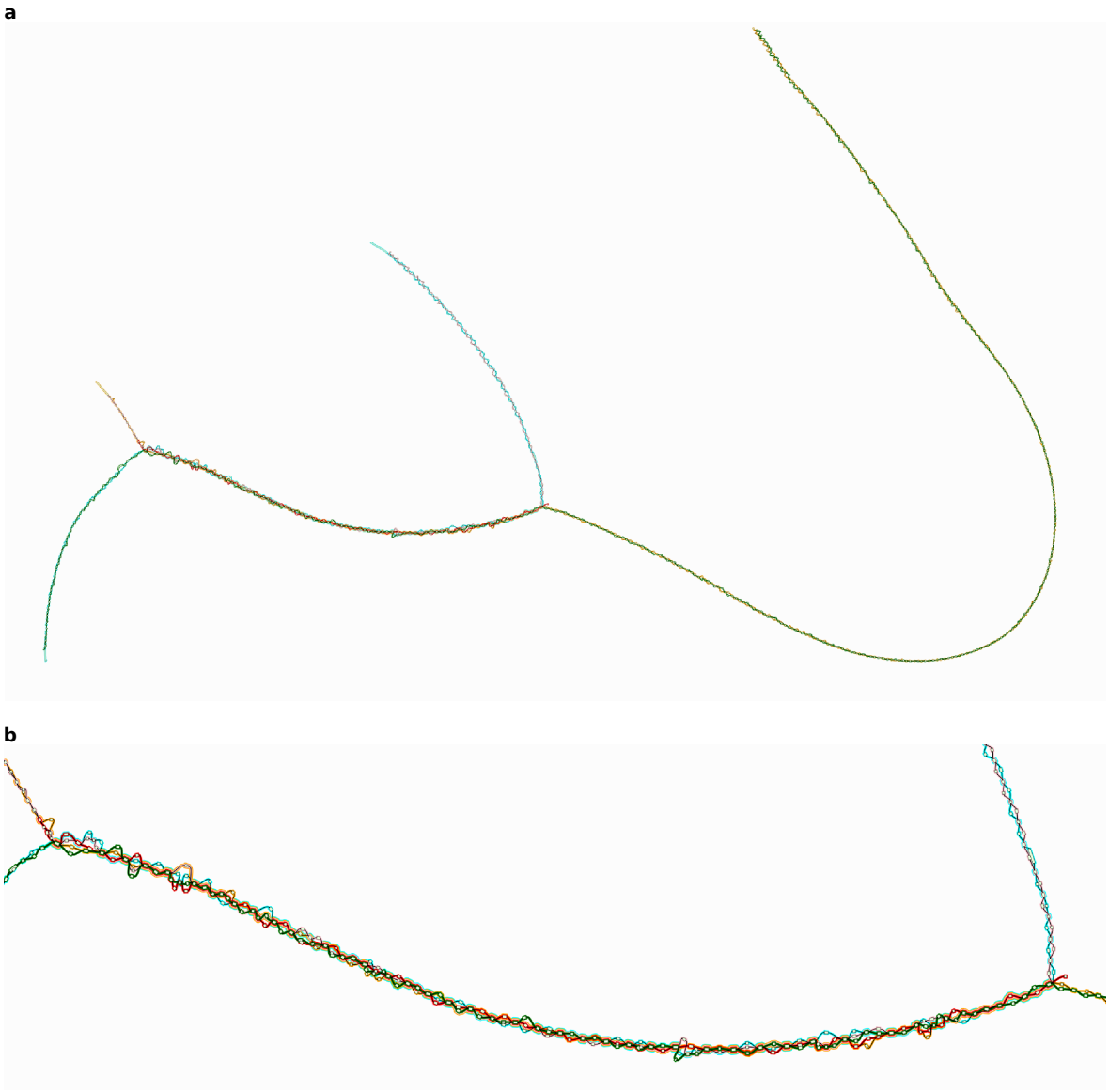


Figure 2.5: Haplotype assemblies for the FRIGIDA gene. We ran Reveal (`https://github.com/jasperlinthorst/reveal`) to produce a graph that represents an alignment of the local haplotype assemblies for the FRIGIDA gene and the corresponding reference sequence. We visualized this graph using GfaViz [64]. The red sequence corresponds to the reference genome. **a** shows the whole graph, **b** shows the part of the graph that corresponds to the FRIGIDA-like protein 5 isoform X2 gene. Figure taken from [166].

reading frame (ORF) in the first three haplotypes representing the FRIGIDA-like protein 5 isoform X2 gene coding sequence. For the fourth haplotype we could not detect a corresponding ORF, as the putative FRIGIDA gene in the fourth phase showed an early STOP codon highlighted in Figure 2.3c. Interestingly, the fourth phase carried an additional frameshift mutation shown in the inset of Figure 2.3c where only the phasing provides the information that this mutation is linked to the premature STOP codon, highlighting the necessity of (local) phasing to understand gene architecture. Using COBALT [141], we generated multiple sequence alignments of the amino acid sequences resulting from these three ORFs and the corresponding reference sequence (Figure 2.3d). The three sequences show an overall good alignment with the reference, with small differences that may serve as an input for functional follow-up studies.

2.1.5 Discussion

We introduced a new algorithm for polyploid phasing, `whatshap polyphase`, and demonstrated that it produces more accurate phasing results than state-of-the-art polyploid phasing methods. It is designed to handle (locally) identical haplotypes by taking coverage of read clusters into account. We applied `whatshap polyphase` to Oxford Nanopore reads produced from a tetraploid potato genome in order to show that the algorithm works well in practice and to demonstrate how phasing information enables haplotype-resolved assembly of the FRIGIDA gene. Being able to assemble the haplotypes is crucial to study these genomes, especially in order to increase yields for important food crops by studying associations of genotypes and phenotypes [14, 101]. However, there are limitations of the approach. Since it relies on a reference genome, the quality of the phasing results highly depends on the quality of the reference genome. Inaccurate or fragmented reference genomes lead to poor read alignments and subsequently to erroneous haplotype predictions in the respective genomic regions. A similar problem occurs if the genome is highly heterogeneous, like it is the case for potato. The potato genome contains a high number of genetic variants, including SVs and copy number variations [71]. Even if there was a complete, high quality reference genome, differences between the reference sequence and the genome of the individual would lead to a decreased mappability of reads in these regions. These limitations affect not only our approach, but all other alignment-based phasing methods. Due to these issues, phasing the complete potato genome was challenging. Therefore we focused only on the genes, as calling variants in the remaining regions took too long, due to the high number of variation present there. Another limitation of `whatshap polyphase` is that its runtime scales exponentially with increasing ploidy. While we demonstrated that the algorithm was fast up to ploidies of six, it will likely become slow as ploidies get higher. However, many industrially important plants have ploidies smaller or equal to six [14].

An important aspect that this study showed is that especially for highly variable genomes like the potato genome, an error correction step is crucial in order to enable accurate small variant detection from noisy long-read data, since short reads often cannot be unambigu-

ously aligned to a reference genome and are thus ill-suited for variant detection. However, Illumina data is still necessary in order to do error correction. Therefore, phasing currently requires short-read data and long-read data to be available for a sample, making studies more complex and costly. Highly accurate long-read sequencing technologies can provide a solution to this issue, enabling haplotype reconstruction purely based on long reads (see Section 2.2).

Despite all these limitations, we demonstrated that our polyploid phasing approach is very useful in practice. Even though we did not assemble chromosome-scale haplotypes, our method enables the construction of haplotype-resolved assemblies on a gene-level for a highly heterogeneous potato genome, which provides a basis for studying gene expression.

2.2 Phasing small variants with circular consensus long reads

In this section, the circular consensus long-read sequencing technology is presented. The focus is on how this new sequencing technique improves phasing of a diploid genome. The material presented here was published as part of a Nature Biotechnology publication [199]. Sections 2.2.1, 2.2.2 and 2.2.3 re-use material from this paper. Section 2.2.4 re-uses some material that I contributed to the PGAS paper, published in Nature Biotechnology [145]. See Sections E.2 and E.3 for details on author contributions and publication details.

2.2.1 Introduction

This section re-uses some material presented in [199].

Recent improvements in DNA sequencing have revolutionized biological sciences. Short-read sequencing technologies (Section 1.4) produce highly accurate reads but are limited in read length to less than 300 bp. This makes them well suited to detect and genotype small variants like SNPs and indels ($<$ 50 bp), but less useful for structural variant detection, genome assembly or haplotype phasing as they lack long range connectivity information. In contrast, long-read sequencing technologies like PacBio CLR sequencing or Oxford Nanopore sequencing are able to produce much longer reads ($>$ 10 kbp), but are less accurate (Section 1.4). Therefore, such reads are well suited for tasks like phasing or genome assembly, but less useful for small variant characterization.

The complementing characteristics of short and long-read sequencing technologies (Section 1.4) make population-scale studies more costly and complex, as several sequencing technologies need to be combined in order to fully analyze a sample. Recently, PacBio presented a new sequencing technology which can overcome these limitations. The idea is to derive an accurate consensus sequence for long PacBio reads from multiple passes of a single template molecule (Section 1.4). In this way, erroneous single-pass sequences are merged into a highly accurate read combining characteristics of short and long sequencing reads.

The PacBio CCS technology was first applied to the well-characterized human sample HG002 (NA24385), one of the benchmark samples of the Genome in a Bottle Project (GIAB) [208], and demonstrated to enable accurate SNP, indel and structural variant detection, as well as high quality genome assembly. As a part of this project, which was published in Wenger et al. and of which I am a co-author, I demonstrated how PacBio CCS reads enable accurate haplotype phasing of small variants. Previously, phasing pipelines consisted of a variant calling step which detected SNPs and indels from accurate short-read data of a sample, and a haplotyping step, in which variants were phased based on long sequencing reads (such as PacBio CLR or Oxford Nanopore reads). With the new technology, short reads are no longer required in order to call variants, as CCS reads are accurate enough to enable small variant detection. The following sections will present details of this analysis.

2.2.2 Data generation and variant calling

This section provides a summary on the data generation and small variant detection steps prior to phasing described in [199]. Some material from this publication is re-used. All analyses described here were performed by co-authors of this publication.

CCS reads were generated for sample HG002 (NA24385) with a read length of 13.5 ± 1.2 kbp and a median accuracy of 99.9%. Reads were mapped to reference genome GRCh37 with pbmm2 (v.0.10.0) resulting in an average coverage of 28×. In order to analyze which fraction of the genome is accessible by CCS reads, all genomic positions were determined that were covered by at least 10 reads. In addition, coverage-matched Illumina short reads were aligned. At the highest mapping quality value of 60, 97.5% of GRCh37 is mappable with CCS reads, while for short reads, this percentage is only 94.8%. The regions now accessible by CCS data include medically relevant genes previously difficult to be accessed by short reads, such as HLA class 1 and class 2 regions.

In a next step, SNPs and indels were detected from the aligned CCS reads using Google's DeepVariant [144] and variant calling performance was evaluated using the GIAB small variant benchmark set [208]. This resulted in a precision of 99.91% and a recall of 99.96% for SNPs, and 96.9% and 95.98% for indels, respectively. When using short Illumina reads instead for variant detection, precision was 99.96% and recall 99.94% for SNPs, and 99.6% and 99.4% for indels, respectively.

2.2.3 Phasing small variants with CCS reads

This section re-uses material presented in [199].

To determine whether CCS reads could provide both highly accurate variant calls and long-range information needed to generate haplotypes, we used WhatsHap [143] (version v.0.17) to phase the DeepVariant calls. Nearly all (99.64%) autosomal heterozygous variants were phased into 19,215 blocks with an N50 of 206 kbp (Table 2.1). We computed the theoretical phase block length distribution based on the GIAB benchmark phase set [208]. This was done by introducing cuts between heterozygous ground truth variants that are separated by more than the average CCS read length of 13.5 kbp. The phase block length distribution that we observed for the DeepVariant callset closely matches the theoretical distribution (Figure 2.6A). We furthermore computed the theoretical phase block length N50s for different read lengths for phasings produced in the same way (Figure 2.6B). Both experiments suggest that the phase block length is limited by the read length and the amount of variation in HG002, not by coverage or the quality of the variant calls.

Evaluated against the trio-phased variant calls in the GIAB ground truth, the switch error rate of our phased DeepVariant calls is 0.37% and the Hamming error rate is 1.91% (Table 2.1). To evaluate the depth of CCS read coverage required for phasing, we randomly subsampled from the full dataset (28× coverage) in steps of 10%. We phased the full DeepVariant

chromosome	heterozygous variants	% phased	phase blocks	hamming error rate (%)	switch errors	switch error rate (%)	phase block N50 (bp)
1	220,180	99.61%	1,585	1.53%	1,168	0.65%	225,534
2	212,809	99.62%	1,879	1.53%	373	0.21%	179,190
3	193,762	99.73%	1,312	1.63%	408	0.25%	259,761
4	199,451	99.70%	1,338	1.65%	547	0.33%	238,088
5	186,023	99.75%	1,115	1.06%	237	0.15%	277,697
6	177,458	99.71%	1,160	0.96%	303	0.20%	265,656
7	166,051	99.70%	1,048	2.23%	1,105	0.80%	246,748
8	153,941	99.71%	1,002	1.22%	322	0.25%	250,705
9	119,897	99.72%	778	1.30%	362	0.36%	207,951
10	141,433	99.72%	840	2.13%	344	0.29%	255,026
11	128,503	99.67%	948	1.24%	169	0.16%	203,073
12	135,470	99.72%	832	3.51%	229	0.20%	292,306
13	100,628	99.69%	638	2.19%	123	0.14%	244,289
14	93,645	99.68%	548	2.70%	520	0.66%	292,617
15	81,981	99.61%	609	0.71%	411	0.61%	188,168
16	87,697	99.71%	596	4.69%	455	0.63%	198,059
17	78,865	99.65%	569	3.06%	380	0.61%	209,363
18	74,575	99.68%	568	2.44%	95	0.15%	215,577
19	70,975	99.78%	345	2.17%	149	0.26%	283,264
20	61,413	99.65%	425	3.53%	165	0.33%	207,556
21	44,142	99.49%	257	4.29%	545	1.60%	178,353
22	38,604	99.71%	249	1.29%	87	0.28%	221,143
all	2,779,801	99.64%	19,215	1.91%	8,497	0.37%	206,063

Table 2.1: WhatsHap phasing performance on DeepVariant (CCS) callset. WhatsHap provides a highly complete phasing (99.64%) of heterozygous variants in the DeepVariant (CCS) callset that is concordant with the GIAB Trio/10X Genomics phasing benchmark set. Statistics are reported by WhatsHap with Hamming and switch error rates evaluated against the benchmark. Table taken from [199].

callset based on each of the resulting BAM files. We observed that the phase block N50 remains above 150 kbp down to 10-fold coverage (Figure 2.6C), demonstrating excellent phasing performances even on lower read coverages.

2.2.4 Improved genome assembly with CCS reads

This section describes work that I contributed to the PGAS paper [145] together with co-authors and re-uses material from this publication.

Recently, the PGAS pipeline was demonstrated to produce high quality, haplotype-resolved genome assemblies based on long reads and Strand-seq data [145] (see Section 1.8 for details on this pipeline). Using CCS data, assemblies reached N50s of 23.7 and 25.9 Mbp for the two haplotypes, respectively. A comparison of haplotype-resolved genome assemblies generated from PacBio CCS, PacBio CLR and Oxford Nanopore data for sample HG00733

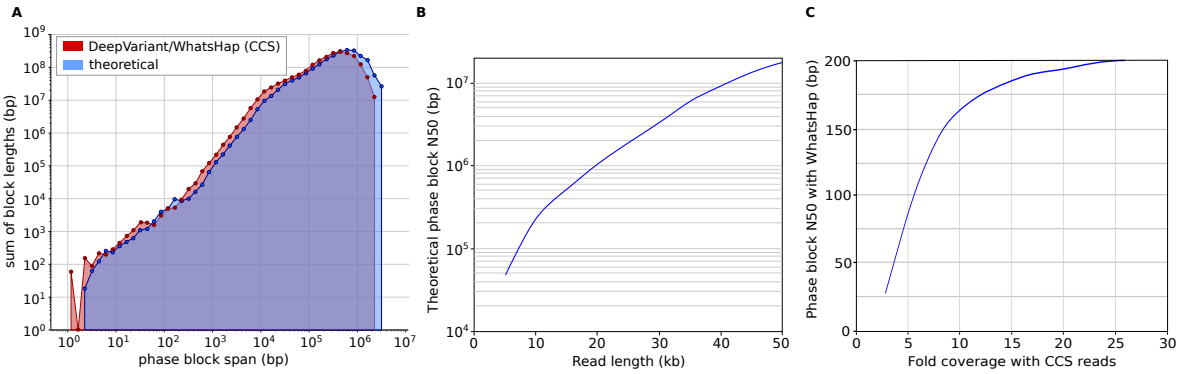


Figure 2.6: Phasing with CCS reads. A Phasing of heterozygous DeepVariant variant calls with WhatsHap, compared to the theoretical phasing of HG002 with 13.5 kbp reads. **B** Theoretical phase block N50 in HG002 at different read lengths. To model the phase blocks achievable with a given read length, cuts were introduced between heterozygous variants in the GIAB trio-phased HG002 variant callset that are separated by more than the read length, which effectively assumes that adjacent heterozygous variants separated by less than the read length can be phased. **C** Phase block N50 for phasing of the 28-fold DeepVariant (CCS) callset with WhatsHap, subsampling in steps of 10%. Figures taken from [199].

with PGAS showed that assemblies generated from CCS data are the most accurate. In order to analyze them, we aligned haplotype-resolved assemblies for HG00733 as well as CCS-based assemblies for the parents (HG00731 and HG00732) to reference genome version GRCh38 using minimap2 [106]. We generated variant calls for each haplotype from these alignments and created a merged callset across samples in order to produce a phased VCF with five samples (HG00733-CCS, HG00733-CLR, HG00733-ONT, HG00731-CCS and HG00732-CCS). This callset contained 10,697,583 variants in total. Genotypes for HG00733 across the three assemblies were consistent for 46% of all variants. Most differences came from variants for which genotypes of CLR and CCS assemblies agree but disagree with the ONT-based genotypes, caused likely by errors in the ONT assemblies. In contrast, we found only 106,270 (0.99%) such variants for CLR and 25,586 (0.24%) for CCS. Furthermore, we analyzed Mendelian consistencies by taking the two parent samples into account. For ONT we found 5,601,071 (52.36%) Mendelian errors, for CLR 469,127 (4.39%) and for CCS we observed 131,281 (1.23%) errors. These experiments demonstrate that CCS reads enable the generation of highly contiguous haplotype-resolved genome assemblies on the scale of a whole chromosome. Compared to CLR or ONT-based assemblies, these assemblies are more accurate and enable the reconstruction of high quality haplotypes of a human sample.

2.2.5 Discussion

The recently developed PacBio CCS technology produces long sequencing reads with low sequencing error rates. The experiments presented demonstrate that PacBio CCS reads enable accurate small variant detection, while being able to access more regions of the genome than short-read data. While variant calling accuracy for SNPs was very similar for both

sequencing technologies, PacBio CCS reads performed slightly worse when calling indels because of the different error profiles that variant callers could not handle as well as for short reads. Recent experiments with the PEPPER-Margin-DeepVariant pipeline show that newer methods are able to overcome these limitations (Section 1.4.4). However, even with the older version of DeepVariant used in the study described in this chapter, high quality phasing results could be achieved with CCS data. The resulting phase block distribution closely matches the theoretical limit, defined mainly by the amount of variation in the sample and the read length, rather than read coverage or the quality of the variant calls [199]. Evaluation based on ground truth haplotypes resulted in very low switch error rates, which shows that read-based phasing based on CCS data enables accurate haplotype predictions. Results nicely demonstrate that CCS reads indeed combine many characteristics of short and long reads. They enable excellent small variant detection providing high quality calls for phasing, such that short-read data is no longer needed for haplotype reconstruction. Furthermore, CCS reads enable high quality genome assembly of human samples that are more accurate than such based on CLR data or ONT reads, which provides the basis for reconstructing chromosomal-scale haplotypes in a reference-free manner. This allows to construct more contiguous haplotypes that include structural variants or regions still inaccessible by read alignments that are currently missed by alignment-based phasing methods. The recently introduced PacBio Revio Platform[1], allows CCS based sequencing on larger scales, providing higher throughput (360 Gb/day instead of 24 Gb/day) at lower costs.

[1] `https://www.pacb.com/revio/`

2.3 Conclusion

This chapter presented different methods and applications of reference-based phasing. First, the polyphase algorithm was described which accurately phases polyploid genomes. It can handle regions where haplotypes are locally identical and improves over current state-of-the art approaches, especially in such collapsed regions. We applied the algorithm to ONT data of a tetraploid potato and used the results in order to generate haplotype-resolved assemblies of the FRIGIDA gene. Second, we demonstrated that accurate, long PacBio CCS reads enable accurate phasing that no longer depends on additional short-read datasets for variant calling. Both of these applications show how phasing is useful in practice, but also demonstrate the limitations of alignment-based phasing methods: they struggle in regions that are too different from the reference genome due to high variability in the sample or because some complex genomic regions are poorly assembled in the reference genome. Therefore, alignment-based phasing methods typically only phase SNPs and indels, but exclude structural variants. As a result, haplotype predictions are fragmented, incomplete and exclude complex, but biologically interesting genomic regions.

PacBio CCS reads can help to overcome the limitation of mapping-based phasing approaches: they provide the basis for highly accurate, chromosome-scale assemblies of haplotypes in a reference-free manner that are able to cover structural variants and complex genomic regions inaccessible to alignment-based methods. Therefore, alternative approaches to polyploid haplotype reconstruction have recently been developed that are based on *de novo* assembly using CCS reads [170, 183]. For diploid genomes, the PGAS pipeline can generate accurate haplotype assemblies reaching N50s of over 25 Mbp per haplotype when combining PacBio CCS reads with Strand-seq data [145]. Besides allowing to study the haplotypes of the assembled sample itself, such chromosome-scale haplotype predictions enable the construction of accurate and complete pangenomes that can replace the linear reference genome. In this way, they provide the potential to improve many downstream analyses, such as genotyping new, yet unassembled samples.

Chapter 3

PanGenie: Pangenome-based genome inference

This chapter introduces an approach that leverages a pangenome in order to improve short-read based genotyping of various types of genetic variation. The work was published in Nature Genetics [49] and this chapter presents an extended version of this publication of which I am the first author. All sections presented in this chapter re-use material that I contributed to this work. Parts of Sections 3.3, 3.6 and 3.9 also summarize work contributed by co-authors of this publication. See Section E.4 for author contributions and publication details.

3.1 Introduction

This section re-uses material presented in [49].

Recent single-molecule, long-read sequencing technologies have enabled breakthroughs in producing *de novo* haplotype-resolved genome assemblies [46, 59, 98, 145] (Section 1.8 and 2.2). Major efforts are under way [90, 113] to generate hundreds of human genome assemblies, with the intention of deriving a variation-aware pangenome representation that replaces the current linear reference genome, GRCh38 (Section 1.9). Such pangenome references provide the potential to improve the analysis of complex genomic regions currently difficult to access with the linear reference. Although long-read technologies are rapidly advancing, generating haplotype-resolved assemblies is still relatively slow and costly, and thus, does not yet scale to large study cohorts consisting of tens of thousands of samples. Due to their low cost, short reads are a more practical approach for such settings in the foreseeable future.

Diploid organisms have two copies of each autosomal chromosome, each of which carries genetic variation. The process of determining whether a known variant allele is located on none, one or both of these copies is referred to as genotyping (Section 1.6). Variant genotyping is an essential step in genetic studies, enabling population analysis, quantita-

tive trait locus studies or trait association analysis. Large studies have produced comprehensive catalogs of human variation ranging from single-nucleotide polymorphisms (SNPs) and indels (insertions and deletions up to 49 bp in size) to larger structural variants (SVs) [2, 24, 32, 181], and many such variants have been linked to diseases and other traits [33, 118, 163, 168, 193, 201].

Widely used genotyping methods for sequencing data [28, 39, 60, 152, 158] are based on short-read alignments to a reference genome or pangenome graphs, which include possible alternative alleles [26, 52, 53, 75, 95, 151, 176] (Section 1.6). Graph-based approaches have been shown to improve genotyping accuracy over methods that rely on a linear reference genome. However, aligning sequencing reads is time-consuming even for linear reference genomes, where mapping 30$\times$ short-read sequencing data of a single human sample takes around 100 CPU hours. This problem is amplified when transitioning to graph-based pangenome references, where the read-mapping problem is even more computationally expensive.

A much faster alternative is to genotype known variants based on k-mers, short sequences of a fixed length k, in the raw sequencing reads without alignment to a reference. Counts of reference- and allele-specific k-mers allow fast and accurate genotyping of various types of genetic variation [40, 44, 89, 172, 175, 182] (Section 1.6). However, these methods can struggle in repetitive and duplicated regions of the genome not covered by unique k-mers. This is especially problematic for SVs, which are often located in repeat-rich or duplicated regions of the genome [24, 198] that are generally difficult to access by short-read sequencing [206].

This problem has been addressed previously by leveraging long-range connectivity information from sequencing reads [189]. In a similar manner, haplotype-resolved assemblies of known samples could improve k-mer-based genotyping, especially in difficult-to-access regions of large diploid genomes, but methods for this have so far been lacking. Known haplotypes have been used to construct population-based reference panels to phase small variants (Li–Stephens model) [111] as well as impute missing genotypes [18, 35, 78, 125], but accurate reference panels that include SVs are still lacking.

In this chapter, PanGenie (for Pangenome-based Genome Inference) is introduced, an algorithm that makes use of haplotype information from an assembly-derived pangenome representation in combination with read k-mer counts to efficiently genotype a wide spectrum of variants. That is, our method can leverage short and longer linkage disequilibrium (LD) structures inherent in the assemblies to infer the genome of a new sample for which only short reads are available. PanGenie bypasses read mapping and is entirely based on k-mers, which allows it to rapidly proceed from the input short reads to a final callset including SNPs, indels and SVs, enabling analysis of variants typically not accessible in short-read workflows – including many deletions $<$ 1 kbp and most insertions $\geq$ 50 bp. We applied our method to genotype variants called from haplotype-resolved assemblies of 11 individuals, revealing a substantial advance in terms of runtime, genotyping accuracy and number of

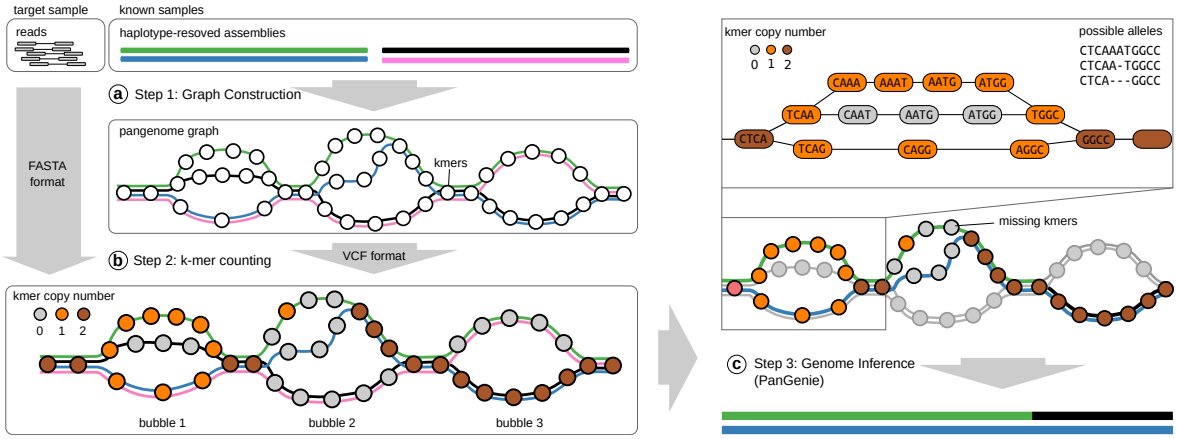


Figure 3.1: Overview of PanGenie. **a** Step 1: variants are called from haplotype-resolved assemblies of a set of known samples and a pangenome graph is constructed, which represents variants as bubbles and contains one path per haplotype. **b** Step 2: the k-mers (represented by circles) contained in the graph are counted in the short-read sequencing data of the target sample to be genotyped. The color of the nodes indicates copy number estimates for the k-mers. **c** Step 3: PanGenie uses k-mer counts and haplotype paths to infer the unknown genome. For the first bubble, k-mer counts suggest that the sample carries the alleles of the green and blue haplotypes. The second bubble is poorly covered by k-mers; however, linkage to adjacent bubbles can be used to infer the two local haplotype paths. Figure taken from [49].

accessible variants.

3.2 Algorithm overview

This section and all its subsections re-use material presented in [49].

We call variants from haplotype-resolved assemblies (see Section 3.3) of several samples and construct a pangenome graph (see Section 3.4) in which these variants are represented as bubbles and each haplotype as a path (Figure 3.1, Step 1). This graph is given as input to PanGenie, together with short-read sequencing data of a new sample to be genotyped. The k-mers contained in the graph are counted in the reads and k-mers unique to bubble regions are identified (Step 2 in Figure 3.1). PanGenie combines two sources of information to genotype bubbles: read k-mer counts and the already known haplotype sequences. The distribution of k-mer counts along the allele paths of a bubble can provide evidence for the genotype of the sample. Figure 3.1 (right panel) provides an example: k-mers corresponding to the second allele of the first bubble are absent from the reads, indicating that the individual carries the alleles of the green and blue haplotypes. However, bubbles may be poorly covered by k-mers or no unique k-mers may exist in repetitive regions of the genome. Such positions cannot be reliably genotyped based on the k-mer counts alone, but known haplotypes can help to infer genotypes based on neighboring bubbles. An example is shown in Figure 3.1

(right panel): the second bubble is poorly covered by k-mers. However, k-mer information of the two neighboring bubbles indicates that the sample carries the alleles covered by the green and blue haplotypes.

For genotyping, we integrate information from k-mer counts and haplotypes by constructing a Hidden Markov Model (HMM), which models the unknown genome as a mosaic of the provided haplotypes and reconstructs it based on the read k-mer counts observed in the sample's sequencing reads (see Section 3.2.1). Hidden states represent pairs of haplotype paths that can be chosen at each bubble position and emit counts for the unique k-mers of the respective region. State transitions between adjacent bubbles correspond to recombination events. Using the Forward–Backward algorithm, genotype likelihoods are computed for each bubble, from which a genotype is derived.

3.2.1 Hidden Markov Model

The input to PanGenie consists of short-read sequencing reads of the sample to be genotyped as well as a fully phased, multi-sample VCF file encoding the pangenome graph (see Section 1.11.1 for details on the VCF format). Each variant record in this VCF represents a bubble in the graph. Further details on how our input VCFs look like and how we construct them are provided later in Section 3.4.

Identifying unique k-mers

Sets of bubbles that are less than the k-mer size apart (we use k = 31) are combined and treated as a single bubble. The alleles corresponding to such a combined bubble are defined by the haplotype paths in the respective region.

For each bubble position v, we determine a set of k-mers, $kmers_v$, that uniquely characterize the region. This is done by counting all k-mers along haplotype paths in the pangenome graph using Jellyfish [119] (v.2.2.10), and then determining k-mers for each bubble that occur at most once within a single allele sequence and are not found anywhere outside the variant bubble. We additionally counted all k-mers of the graph in the sequencing reads. This allows us to compute the mean k-mer coverage of the data, k_{cov}, which we use later to compute emission probabilities (Section 3.2.1).

Hidden states and transitions

We assume to be given N haplotype paths $H_i, i = 1, ..., N$, through the graph. Furthermore, for each bubble $v, v = 1, ..., M$, we are given a vector of k-mers, $kmers_v$ that uniquely characterize the alleles of a bubble. We assume some (arbitrary) order of the elements in $kmers_v$ and refer to the i-th k-mer as $kmers_v[i]$. In addition, we are given sequencing data of the sample to be genotyped and corresponding k-mer counts for all k-mers in $kmers_v$. For each bubble v, we define a set of hidden states $\eta_v = \{H_{v,i,j} \mid i, j \leq N\}$ which contains a state for each possible pair of the N given haplotype paths in the graph. Each such state $H_{v,i,j}$

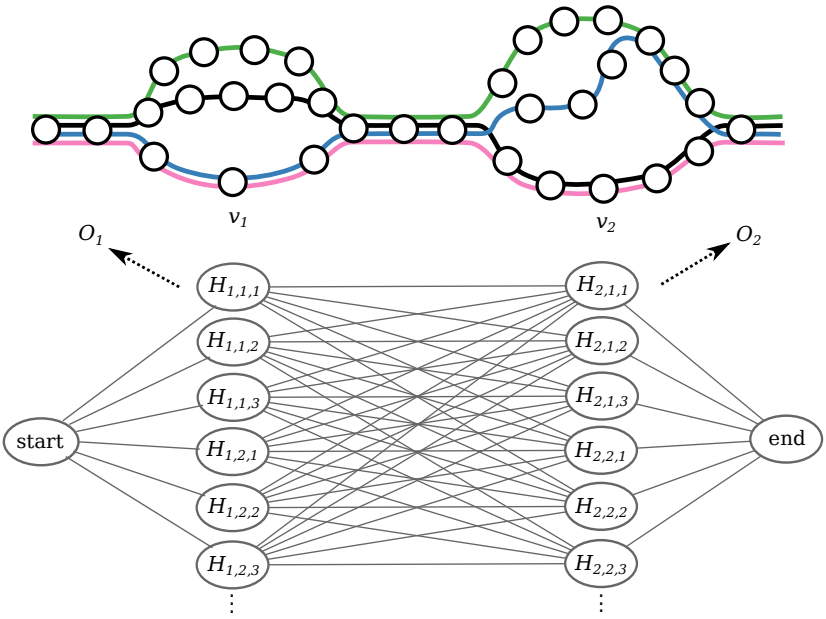


Figure 3.2: PanGenie HMM. Graphical illustration of the HMM underlying PanGenie. For each variant bubble in the pangenome graph, a set of hidden states is defined which represents all possible pairs of haplotype paths. Connections between states represent possible transitions. Hidden states emit counts for the unique k-mers of a bubble.

induces an assignment of copy numbers to all k-mers in $kmers_v$. We define a vector $a_{v,i,j}$ such that the k-th position contains the copy number assigned to the k-th k-mer in $kmers_v$:

$$a_{v,i,j}[k] = \begin{cases} 0 & kmers_v[k] \notin H_i \cup H_j \\ 1 & kmers_v[k] \in H_i \setminus H_j \\ 1 & kmers_v[k] \in H_j \setminus H_i \\ 2 & kmers_v[k] \in H_i \cap H_j \end{cases} \quad \forall k = 1, \ldots, |kmers_v|$$

The idea here is that we expect to see copy number 2 for all k-mers occurring on both selected haplotype paths. In case only one of the haplotypes contains a k-mer, its copy number must be 1 and k-mers that do not appear in any of the two paths must have copy number 0. See the leftmost bubble in the graph shown in Figure 3.1 for an example: for the green and blue haplotypes, the expected copy number of all orange k-mers is 1, since each of them is covered by exactly one of these two haplotypes. The gray k-mers are covered by neither the green nor blue haplotype, therefore we expect to see copy number 0 for all of them. The brown k-mers are carried by both the green and blue haplotype. Therefore, the expected copy number of these k-mers is 2.

From each state $H_{v,i,j} \in \eta_v$ that corresponds to bubble position v, there is a transition to each state corresponding to the next position, $v + 1$. In addition, there is a $start$ state, from which there is a transition to each state of the first bubble, and an end state, to which there is a transition from each state that corresponds to the last bubble.

Figure 3.2 presents a graphical illustration of the model. The hidden states are shown as circles below the variant bubbles and connections represent possible transitions between the states.

Transition probabilities

Transition probabilities are computed following the Li-Stephens model [111]. Given a recombination rate r, the effective population size N_e and the distance x (in basepairs) between two ascending bubbles $v-1$ and v, we define:

$$d = x \cdot \frac{1}{1000000} \cdot 4r \cdot N_e$$

We compute the Li-Stephens transition probabilities as:

$$p_r = (1 - exp(-\frac{d}{N})) \cdot \frac{1}{N}$$
$$q_r = exp(-\frac{d}{N}) + p_r$$

Finally, the transition probability from state $H_{v-1,k,l}$ to state $H_{v,i,j}$ is computed as shown below:

$$P(H_{v,i,j}|H_{v-1,k,l}) = \begin{cases} q_r \cdot q_r & i = k \text{ and } j = l \\ q_r \cdot p_r & i = k \text{ and } j \neq l \\ q_r \cdot p_r & i \neq k \text{ and } j = l \\ p_r \cdot p_r & i \neq k \text{ and } j \neq l \end{cases} \tag{3.1}$$

Observable states

Each hidden state $H_{v,i,j} \in \eta_v$ outputs a count for each k-mer in $kmers_v$. Let O_v be a vector of length $|kmers_v|$ for bubble v such that $O_v[k]$ contains the observed k-mer count of the k-th k-mer in the sequencing reads. To define the emission probabilities, we first need to model the distribution of k-mer counts for each copy number, $P(O_v[k]|a_{v,i,j}[k] = c), c = 0, 1, 2$. For copy number 2, we use a Poisson distribution with mean λ which we set to the mean k-mer coverage k_{cov} that we compute from the k-mer counts of all graph k-mers. Similarly, we approximate the k-mer count distribution for copy number 1 in terms of a Poisson distribution with mean $\lambda/2$. For copy number 0, we need to model the erroneous k-mers that arise from sequencing errors. This is done using a Geometric distribution, the parameter p of which we choose based on the mean k-mer coverage as shown below:

$$p = \begin{cases} 0.99 & k_{cov} < 10 \\ 0.95 & 10 \leq k_{cov} < 20 \\ 0.9 & 20 \leq k_{cov} < 40 \\ 0.8 & k_{cov} \geq 40 \end{cases}$$

Finally, we compute the emission probability for a given state and given observed read k-mer counts as shown below, making the assumption that the k-mer counts are independent.

$$P(\mathcal{O}_v|H_{v,i,j}) = \prod_{l=1}^{|kmers_v|} P(\mathcal{O}_v[l] \mid a_{v,i,j}[l])$$

Forward-Backward algorithm

In this model, the hidden states represent possible genotypes a sample could have at each bubble position given the haplotype paths. Genotype likelihoods can be computed based on the Forward-Backward algorithm. The initial distribution of our HMM is such that we assign probability 1 to the $start$ state and 0 to all others. Forward probabilities $\alpha_v()$ are computed in the following way:

$$\alpha_0(start) = 1$$

For states corresponding to bubbles $v = 1, ..., M$, the Forward probabilities are computed as shown below.

$$\alpha_v(H_{v,i,j}) = \sum_{H_{v-1,s,t} \in \eta_{v-1}} \alpha_{v-1}(H_{v-1,s,t}) \cdot P(H_{v,i,j}|H_{v-1,s,t}) \cdot P(\mathcal{O}_v|H_{v,i,j}) \qquad \forall i, j$$

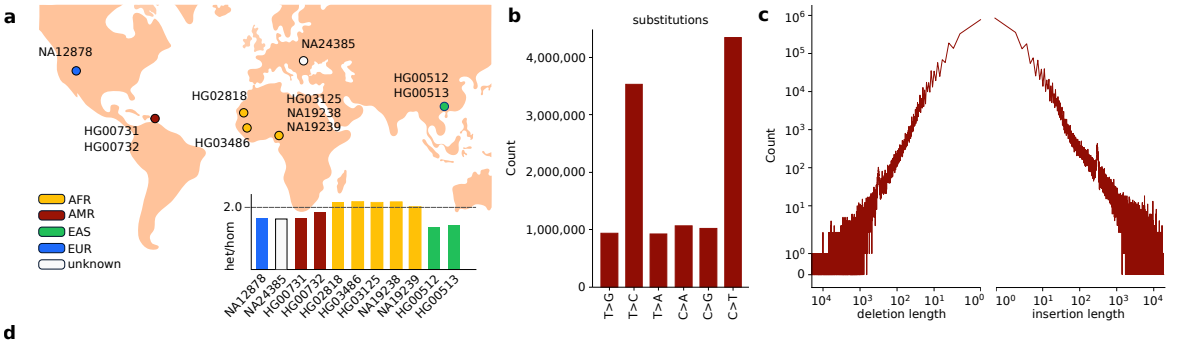
The transition probabilities are computed as described above, except for transitions from the *start* state to all states in the first column, which we assume to have uniform probabilities. Backward probabilities are computed in a similar manner. We set:

$$\beta_M(end) = 1$$

For $v = 1, ..., M-1$, we compute them as:

$$\beta_v(H_{v,i,j}) = \sum_{H_{v+1,s,t} \in \eta_{v+1}} \beta_{v+1}(H_{v+1,s,t}) \cdot P(H_{v+1,s,t}|H_{v,i,j}) \cdot P(\mathcal{O}_{v+1}|H_{v+1,s,t}) \qquad \forall i, j$$

Finally, posterior probabilities for the states can be computed:



d

	SNPs	small (1-19 bp)			midsize (20-49 bp)			large (>= 50 bp)		
		deletions	insertions	complex	deletions	insertions	complex	deletions	insertions	complex
variants	12,095,177	1,811,123	1,922,163	–	80,027	110,882	–	41,499	108,929	–
bubbles in graph	11,556,580	819,445	810,298	597,044	12,720	20,300	87,392	4,397	18,325	52,272

Figure 3.3: Callset statistics. **a** Overview of the samples for which variants were called from haplotype-resolved assemblies as well as their het:hom ratios. Color corresponds to the population from which the samples originate. **b** The number of different substitutions reported for all samples. **c** Length distribution of insertions and deletions across all samples (in basepairs). **d** Total number of distinct variant alleles detected across all 11 samples (first row), as well as the number of bubbles in the corresponding pangenome graph (second row). We distinguished small (1-19 bp), midsize (20-49 bp) and large ($\geq$ 50 bp) variants. Biallelic bubbles were classified as SNPs, insertions or deletions; complex corresponds to all remaining bubbles with more than two branches resulting from inserting overlapping variant calls into the graph. Figure taken from [49].

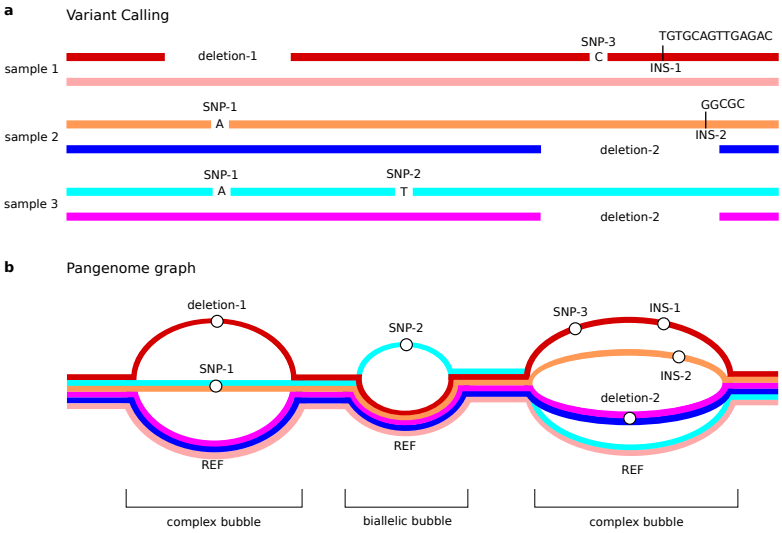
$$P(H_{v,i,j}|O_1, O_2, \ldots, O_M) = \frac{\alpha_v(H_{v,i,j}) \cdot \beta_v(H_{v,i,j})}{\sum_{h \in \eta_v} \alpha_v(h)\beta_v(h)}$$

Since different paths can cover the same allele, several hidden states at a bubble position v can correspond to the same genotype. Also, the alleles in a genotype are unordered, therefore states $H_{v,i,j}$ and $H_{v,j,i}$ always lead to the same genotype. In order to compute genotype likelihoods, we sum up the posterior probabilities for all states that correspond to the same genotype. In this way, we can compute genotype likelihoods for all genotypes at a bubble position, based on which a genotype prediction can be made by choosing the genotype with highest probability.

3.3 Variant calling from haplotype-resolved assemblies

This section re-uses material presented in [49]. The generation of the assemblies with PGAS was performed by Peter Ebert, a co-author of this publication.

We used a development version of the PGAS pipeline [46, 145] (parameter settings v.13) to generate haplotype-resolved assemblies of 14 individuals including 3 mother-father-child trios (Figure 3.3a; samples include: Yoruban trio: NA19238, NA19239, NA19240; Puerto Rican trio: HG00731, HG00732, HG00733; southern Han Chinese trio: HG00512, HG00513,



VCF

	ALT	sample 1	sample 2	sample 3
variant 1	deletion-1	1|0	0|0	0|0
variant 2	SNP-1	0|0	1|0	1|0
variant 3	SNP-2	0|0	0|0	1|0
variant 4	deletion-2	0|0	0|1	0|1
variant 5	SNP-3	1|0	0|0	0|0
variant 6	INS-1	1|0	0|0	0|0
variant 7	INS-2	0|0	1|0	0|0

VCF

	ALT	sample 1	sample 2	sample 3
bubble 1	deletion-1, SNP-1	1|0	2|0	2|0
bubble 2	SNP-2	0|0	0|0	1|0
bubble 3	SNP-3:INS-1, INS-2, deletion-2	1|0	2|3	0|3

Figure 3.4: Variant calling and graph construction. **a** Shown are haplotype-resolved assemblies for three samples and corresponding variant calls made relative to a reference genome. On the right, we show how these variants are represented in a VCF file (simplified). The VCF file is biallelic and contains one record per (distinct) variant allele detected across the assemblies. **b** Shown is the pangenome representation of the variants detected in panel a). Variants are represented as bubble structures. Sets of overlapping variants are merged into a single multiallelic bubble (see first and last bubble for examples). Each haplotype can be represented as a path through the graph. We represent the pangenome in terms of a multiallelic VCF file containing a record for each bubble and alleles corresponding to the branches of the bubble (right). We keep track of which callset variants each branch of the bubble was constructed from as illustrated in the VCF representation. In this way, we can later convert genotypes derived for a bubble back to genotypes for each individual variant inside of a bubble. Note that our VCFs contain the actual allele sequences in their "ALT" column, we replaced them by their IDs in this figure for simplicity. Figure taken from [49].

HG00514; and NA12878, HG02818, HG03125, NA24385 and HG03486) and called variants on each haplotype of all autosomes and chromosome X. The three child samples (HG00733, HG00514 and NA19240) were used only for quality control and filtering, and thus were not part of our final callset/graph. For each sample, we separately mapped contigs of each haplotype to the reference genome (GRCh38). This was done using minimap2 [106] (v.2.18) with parameters `-cx asm20 -m 10000 -z 10000,50 -r 50000 -end-bonus=100 -O 5,56 -E 4,1 -B 5 -cs`. In the next step, we called variants on each haplotype using paftools (`https://github.com/lh3/minimap2/tree/master/misc`, v.2.18) with default parameters. We generated a biallelic VCF file containing variant calls made across all 11 unrelated samples (Figure 3.4a). If a region was not covered by any contig alignment in a sample, or the sample had multiple overlapping contig alignments, we set all its genotypes in this region to missing ("./."), because it is unclear what the true genotype alleles are in this case. Furthermore, we removed variants from our callset for which $>$ 20% of the samples had missing genotype information. The remaining regions covered 91.8% (2.8 Gbp) of chromosomes 1–22 and chromosome X. Of the 8.2% of regions not covered, 48.3% were gaps in GRCh38 and 24.0%

were centromeres.

We computed the Mendelian consistency for the Puerto Rican (HG00731, HG00732, HG00-733), Chinese (HG00512, HG00513, HG00514) and Yoruban (NA19238, NA19239, NA19240) trios and observed that 97.9%, 96.8% and 97.6% of all variants were consistent with Mendelian laws, respectively. We removed a variant from our callset if there was a Mendelian conflict in at least one of the three trios. We show the number of variants in our final callset and the intermediate stages of variant calling in the first three columns of Table B.1.

We computed the transition:transversion (ti:tv) ratio for SNPs and the heterozygous:homozygous (het:hom) ratio as quality control measures [68, 194]. Our SNP calls contained around twice as many transitions as transversions (Figure 3.3b) resulting in ti:tv ratios between 2.01 and 2.02 for all samples. We obtained het:hom ratios between 1.37 and 2.20 for all our 11 callset samples. These numbers are in line with respective results for African (AFR), American (AMR), Asian (EAS) and European (EUR) individuals reported in previous studies [194, 195]. Furthermore, our callset contains comparable numbers of insertions and deletions (Figure 3.3c), except for the expected enrichment for insertion alleles for SVs [24]. We show detailed counts of distinct variant alleles for all types in Figure 3.3d (first row) and for the individual samples in Table B.2. We distinguish small variants (1–19 bp), midsize variants (20–49 bp) and large variants ($\geq$ 50 bp).

3.4 Constructing a pangenome reference

This section re-uses material presented in [49].

We created an acyclic and directed pangenome graph representing our variant callset (Figure 3.4). Variants produce bubbles in the graph with branches that define the corresponding alleles. The input haplotypes can be represented as paths through the resulting pangenome. When constructing the graph, we represent sets of variants overlapping across haplotypes as a single bubble, with potentially multiple branches reflecting all the allele sequences observed in the haplotypes in the respective genomic region (Figure 3.4). The total number of bubbles in the resulting graph is presented in the last row of Figure 3.3d. We represent the pangenome in terms of a fully phased, multisample VCF file that contains one entry for each bubble in the graph (Figure 3.4b). At each site, the number of branches of the bubble is limited by the number of input haplotype sequences and the genotypes of each sample define two paths through this graph, corresponding to the respective haplotypes. We keep track of which individual input variants contribute to each bubble in the graph, so that we can convert our pangenome graph representation back to the set of input variants. In this way, we can translate genotypes computed by a genotyper for all these bubbles to a genotype for each individual callset variant.

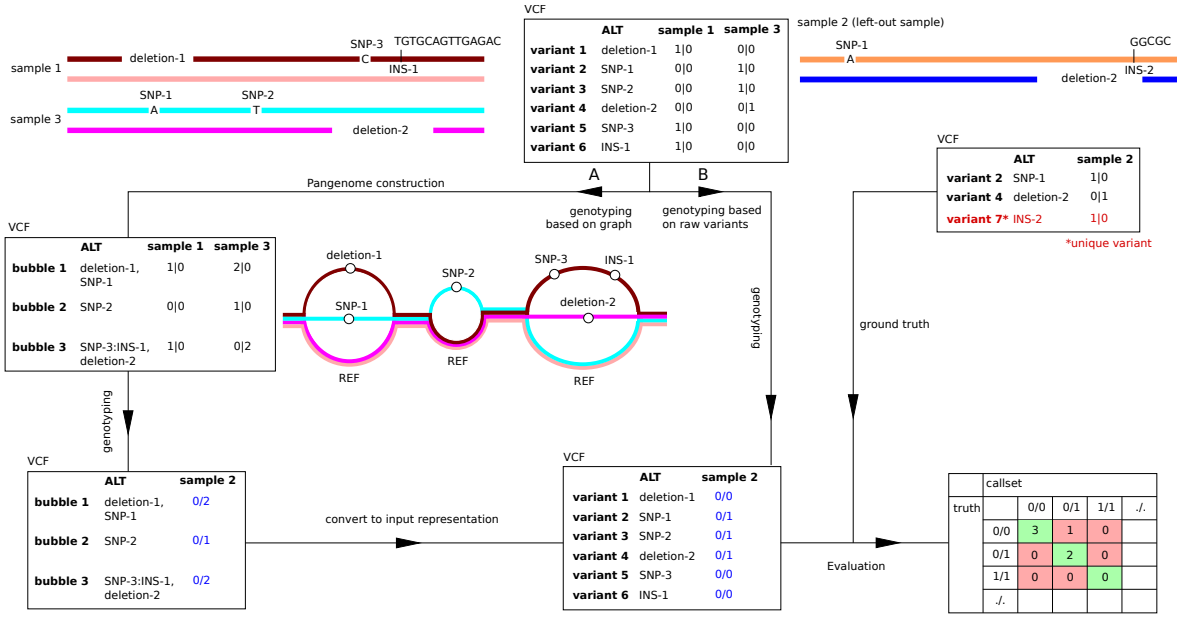


Figure 3.5: Overview of leave one out experiment. We illustrate the leave-one-out experiment using three samples. Variants are called in all samples based on haplotype-resolved assemblies. For evaluation, we construct a biallelic callset containing all variants detected in samples 1 and 3, and a biallelic truth set containing all variants called in the left out sample (sample 2). The former set of variants is used for genotyping, the latter for evaluation. When running PanGenie, BayesTyper and Platypus, we first convert the variant calls into a pangenome graph representation (stored as multiallelic VCF) and genotyped the corresponding bubbles (A). We keep track of which bubbles consist of which variant alleles so that genotypes can later be converted back to the original variant representation. For the other tools tested (GATK, Platypus, GraphTyper, Giraffe), we directly used the callset variants as input, without creating the graph (B). The genotypes predicted by each tool are then compared to the variants detected in the left out sample for evaluation. Variants unique to the left out sample cannot be genotyped correctly by any re-genotyping approach (marked in red). We exclude such variants when computing weighted genotype concordances and adjusted precision/recall/F-score metrics. Figure taken from [49].

3.5 Comparison to existing genotyping methods

This section and all its subsections re-use material presented in [49].

We conducted a 'leave-one-out experiment' (Figure 3.5) to mimic a realistic scenario in which we genotyped variants detected from haplotype-resolved assemblies of a set of known samples in a new, unknown sample. We used Illumina reads from the Genome in a Bottle (GIAB) consortium [207] and 1000 Genomes Project high-coverage data [21]. We collected variants that we called from the assemblies across all but one sample (see Section 3.3) and used them as input for genotyping the left-out sample (we refer to this set as *known variants* in the following). We used the set of variants called from the assemblies of the left-out sample for evaluation (*evaluation variants*). We ran this experiment twice, leaving out samples NA12878 and NA24385, respectively. In addition to running PanGenie, we ran

BayesTyper [175] (k-mer-based), Platypus [158], GATK HaplotypeCaller [39], GraphTyper [53], Paragraph [26] and Giraffe [176] (all mapping-based) to re-genotype the same set of variants (Figure 3.5). We ran our experiments on coverage levels 30×, 20×, 10× and 5×. Not all tools are designed to handle all types of variants. Therefore, we ran GATK only on SNPs, small and midsize variants and Paragraph only on midsize and large variants. GraphTyper and Giraffe were run on large variants only.

As input for PanGenie (commit 1f3d2d2), BayesTyper (v.v1.5) and Paragraph (v.2a), we constructed a pangenome graph representation based on the known variants in the same way as described in Section 3.4. We kept track of which variant alleles each resulting bubble consists of, so that genotypes derived for all bubbles can later be converted back to the original variant representation. For the other genotypers tested (GATK 4.1.3.0, Platypus 0.8.1, GraphTyper 2.7.1 and Giraffe v.1.30.0), we directly used the set of known variants as input, without generating the graph representation first, because we observed that these tools could better handle variants represented in this way. As a result of running all genotypers, we had one VCF file per tool containing genotypes for all our known variants. We used the evaluation variants to evaluate the genotype predictions of all tools. Figure 3.5 provides an illustration of the leave-one-out experiment.

Note that re-genotyping a set of known variants in a new sample is different from variant detection. Variants unique to the new sample that have not been seen in the callset samples can not be genotyped because re-genotyping methods genotype only variants that they are given as input. We provide the number of unique variants present in each panel sample in Table B.2. To analyze genotyping performance, we introduce the weighted genotype concordance (wGC) which puts equal emphasis on the ability to detect all three possible genotypes (Figure 3.5.1). As an alternative view on the performance of the individual methods, we offer precision, recall and F-score, all in an unadjusted version and an adjusted version that does not penalize methods for 'missing' variants that are undetectable because they are not in the input set (see Section 3.5.1 for a detailed description on all evaluation metrics). We consider two configurations for PanGenie: 'high-gq' filtering, where we use only genotypes reported with high quality scores and treat all other variants as not genotyped, and 'all', where we consider all reported genotypes regardless of their quality.

3.5.1 Evaluation metrics

Weighted genotype concordance

In the biallelic representation, each genotyped variant is either absent from the truth set (0/0, in case it is not present in the left out sample), heterozygous (0/1) or homozygous (1/1). We construct a confusion matrix counting all cases (Figure 3.6). The counts on the diagonal (labeled T_0/0, T_0/1, T_1/1) correspond to correctly genotyped variants. All others are errors. For all three genotypes, we compute the concordances by counting the number of correct predictions and divide it by the total number of variants in that category:

genotype concordance

	callset				
truth		0/0	0/1	1/1	./.
	0/0	T_0/0	F_0/0	F_0/0	
	0/1	F_0/1	T_0/1	F_0/1	
	1/1	F_1/1	F_1/1	T_1/1	
	./.				

precision

	callset				
truth		0/0	0/1	1/1	./.
	0/0		FP	FP	
	0/1		TP	FP	
	1/1		FP	TP	
	./.		FP	FP	

recall

	callset				
truth		0/0	0/1	1/1	./.
	0/0				
	0/1	FN	TP	FN	FN
	1/1	FN	FN	TP	FN
	./.				

Figure 3.6: PanGenie evaluation metrics. Metrics used to evaluate genotyping results and how they define errors. Figure taken from [49].

$$\text{conc}(0/0) = \frac{\text{T\_0/0}}{\text{T\_0/0} + \text{F\_0/0}} \quad \text{conc}(0/1) = \frac{\text{T\_0/1}}{\text{T\_0/1} + \text{F\_0/1}} \quad \text{conc}(1/1) = \frac{\text{T\_1/1}}{\text{T\_1/1} + \text{F\_1/1}}$$

Since we genotype all variants detected across multiple samples (including many rare alleles) in our evaluation sample, the majority of variants will be absent in the evaluation sample. That is, the number of variants whose true genotype is 0/0, will be higher compared to the ones with genotypes 0/1 or 1/1. To adjust for unqeual numbers of 0/0, 0/1 and 1/1 genotypes in our ground truth, we compute the weighted genotype concordance as:

$$\text{weighted genotype concordance} = \frac{\text{conc}(0/0) + \text{conc}(0/1) + \text{conc}(1/1)}{3}$$

As mentioned previously, we exclude all variants unique to the evaluation sample when computing the weighted genotype concordance (see Table B.2), since these variants are not part of the set of input variants given to all genotypers and thus will not be considered for genotyping (as all tools re-genotype variants and do not detect them).

Fraction of genotyped variants

Many of the re-genotyping tools we consider can report genotypes "./." for input variants that they are not able to genotype. For each tool, we compute the fraction of input variants that were reported with such an "untyped" genotype.

(Adjusted) Precision/Recall/F-score

We use `RTG vcfeval` [29] in order to compute precision and recall for our genotype predictions. We compute two versions of precision, recall and F-scores: taking all variants into account (including those that are unique to the evaluation sample, hence missing from the input set and undetectable by re-genotyping, see Table B.2), and an *adjusted* version, where we remove all variants unique to the evaluation sample from the truth set. Therefore, the unadjusted version combines the effects of variants missing from the input set to be genotyped and the performance of the genotyping method, while the adjusted version aims to only measure the performance of the method (and does not penalize variants absent in the

input set). True positives, false positives and false negatives are defined as shown in Figure 3.6 [29] and precision, recall and F-score are defined as:

$$\text{precision} = \frac{TP}{TP + FP} \qquad \text{recall} = \frac{TP}{TP + FN} \qquad \text{F-score} = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

Note on Precision/Recall/F-score metrics

We offer precision/recall/F-score metrics to facilitate comparison to other studies, including on methods for variant calling. However, these metrics come with the following caveats when evaluating re-genotyping experiments and should hence be interpreted accordingly: the more samples we use to generate the set of known variants to genotype in a new sample, the larger the amount of rare variants and thus the larger the fraction of variants whose true genotype of the new sample is 0/0. That is, we can make our set of input variants (almost) arbitrarily large by adding variants absent from the new sample. The possibility of adding noise when including a large number of rare alleles when constructing pangenome representations is a known effect and an important consideration [147]. As a consequence, the number of false positive calls increases with the increase in the number of tested variants, while the number of true positive calls is limited by the actual number of variants present in the new sample, reducing the precision. An example is shown in Figure B.1a. This also explains why the precision we see for all genotypers in our evaluation is sometimes small compared to the genotype concordance (Figure B.1b).

3.5.2 Evaluation regions

We stratified our analyses by considering variants outside and inside short-tandem repeats (STRs) and variable-number tandem repeats (VNTRs) [93]. We annotated variants according to their repeat status and observed that between 68% and 72% of midsize (20–49 bp) and large variants ($\geq$ 50 bp) are repeat associated, respectively (Table B.3).

In addition, we classified the genome into ‘complex’ and ‘biallelic’ regions based on the bubble structure of our pangenome graph: all variants located inside of complex bubbles, that is, bubbles with more than two branches, fell into the first category, and the remaining regions into the second. Consider Figure 3.4 for an example: the first and third bubbles are complex, thus all variants contained inside these bubbles fall into the category ‘complex’. The second bubble is biallelic and therefore the corresponding SNP variant is considered ‘biallelic’. For our leave-one-out experiment for sample NA12878, we show the number of variants falling into the different categories in Table B.3. It can be observed that most complex bubbles are located inside STR/VNTR regions (Table B.3). In addition, more than half of all midsize and large variants are located in these repetitive regions.

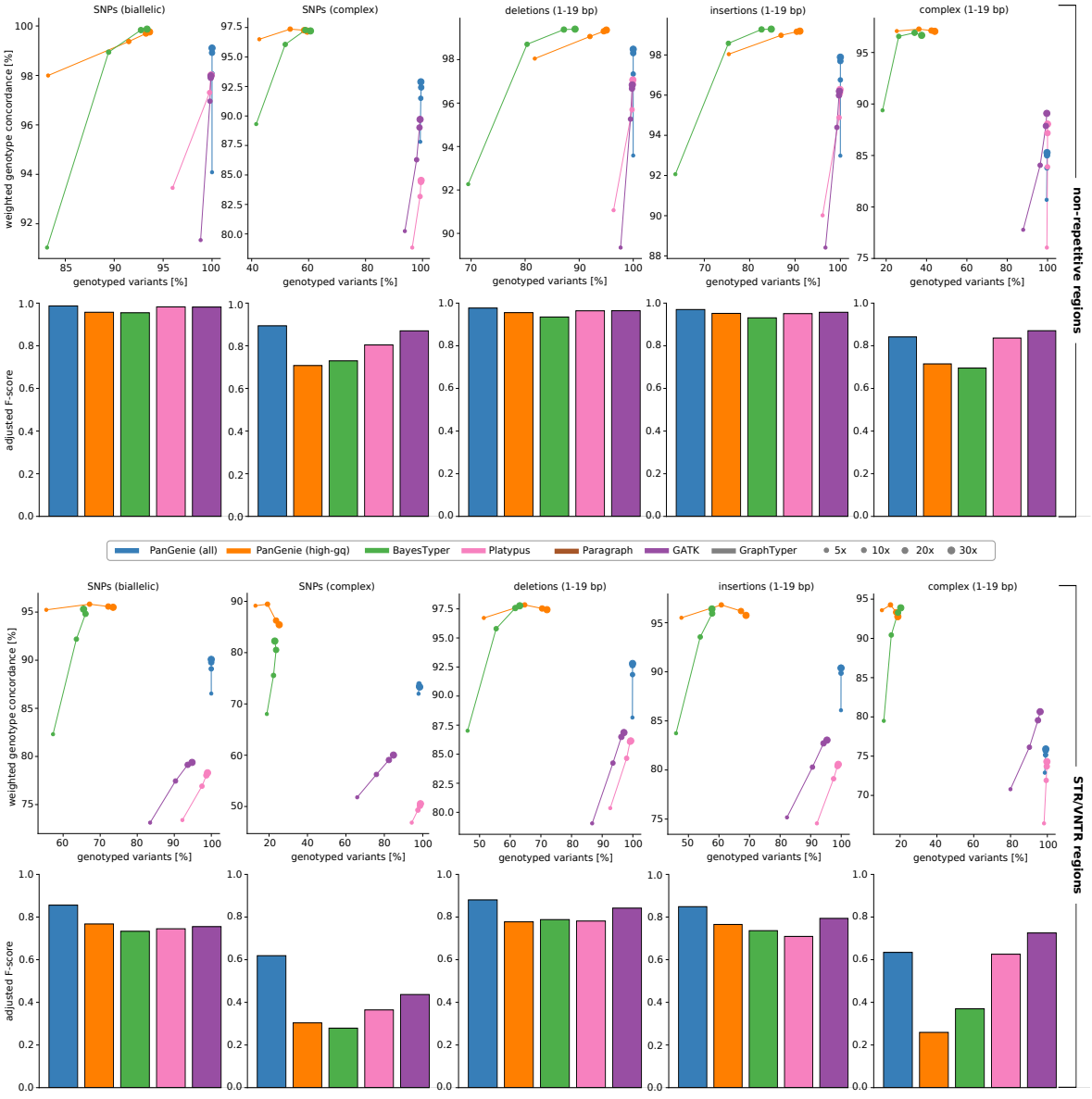


Figure 3.7: Results of leave-one-out experiment (SNPs and small variants). The wGC at different coverages for sample NA12878 and F-scores for coverage 30× in nonrepetitive (top) and STR/VNTR regions (bottom). We ran PanGenie, BayesTyper, Paragraph, Platypus, GATK and GraphTyper to re-genotype all callset variants. Besides not applying any filter on the reported genotype qualities ('all'), we additionally report genotyping statistics for PanGenie when using 'high-gq' filtering (genotype quality ≥ 200). Biallelic SNPs, insertions and deletions include all respective variants in biallelic regions of the genome, whereas complex contains all variant alleles falling into regions with complex bubbles in the pangenome graph representation. Figure taken from [49].

3.5.3 Results

Genotyping results for NA12878 (Figure 3.7, Figure 3.8, Figures B.2-B.3) and NA24385 (Figures B.5–B.10) were similar, showcasing consistency of results across samples. In the

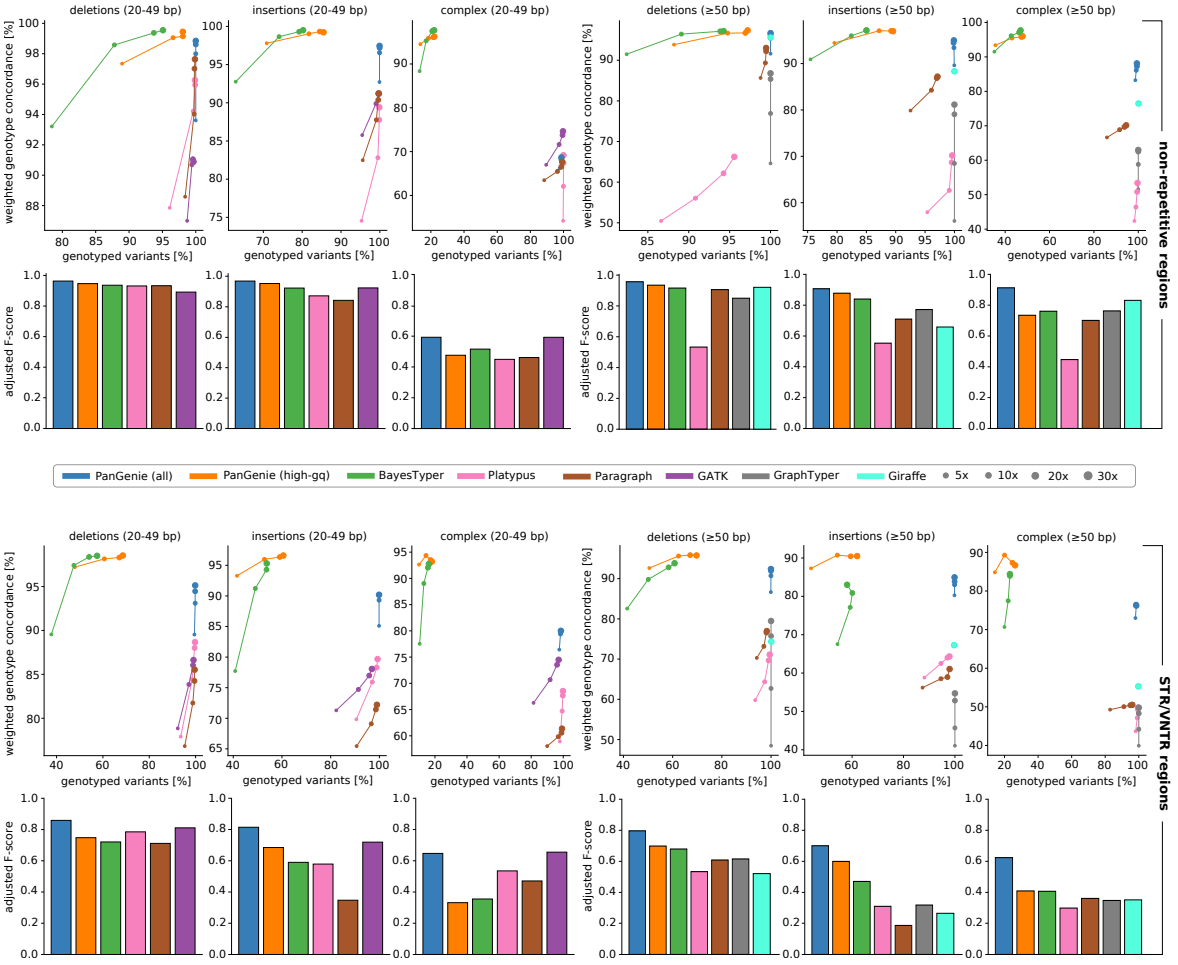


Figure 3.8: Results of leave-one-out experiment (midsize and large variants). The wGC at different coverages for sample NA12878 and F-scores for coverage 30× in nonrepetitive (top) and STR/VNTR regions (bottom). We ran PanGenie, BayesTyper, Paragraph, Platypus, GATK, GraphTyper and Giraffe to re-genotype all callset variants. Besides not applying any filter on the reported genotype qualities ('all'), we additionally report genotyping statistics for PanGenie when using 'high-gq' filtering (genotype quality $\geq$ 200). Insertions and deletions include all respective variants in biallelic regions of the genome, whereas complex contains all variant alleles (insertions or deletions) falling into regions with complex bubbles in the pangenome graph representation. Figure taken from [49].

following, the results for sample NA12878 are discussed.

For biallelic SNPs in nonrepetitive regions, all methods reached very good levels of genotype concordance and F-scores (Figure 3.7), with all F-scores $>$ 0.95 at coverage 30×. For biallelic SNPs in repetitive regions, PanGenie still achieved an F-score of 0.85, whereas the second-best tool GATK reached only 0.75 (Figure 3.7). In repetitive regions, BayesTyper had the largest fraction of untyped SNPs of all tools, resulting in lowest recall of 0.6 for biallelic SNPs and 0.17 for SNPs inside of complex bubbles (Figure B.3).

For small insertions and deletions, PanGenie ('all') outperformed the mapping-based approaches, in particular in STR/VNTR regions (wGC of 90.4% for insertions and 92.8% for

coverage	method	NA12878				NA24385			
		time total	time genotyping	memory total	memory genotyping	time total	time genotyping	memory total	memory genotyping
5	PanGenie	21:06:10	19:42:05	84.8	36.4	31:44:24	29:30:54	84.6	36.2
	BayesTyper	27:23:15	26:22:21	39.3	39.3	36:31:30	35:20:37	39.2	39.2
	Platypus	18:12:42	1:20:10	18.2	0.2	20:39:51	1:31:42	8.7	0.1
	GATK[1]	34:41:06	17:24:26	18.2	0.4	35:24:17	15:53:15	8.7	0.4
	Paragraph[2]	39:49:37	22:57:04	18.2	10.1	40:51:58	21:43:48	11.1	11.1
	GraphTyper[3]	22:06:44	5:14:12	18.2	0.2	23:12:02	4:03:52	8.7	0.2
10	PanGenie	21:36:59	19:27:31	84.8	36.4	33:07:31	29:29:26	84.7	36.2
	BayesTyper	38:42:03	37:20:08	40.7	40.7	36:05:15	34:16:52	40.7	40.7
	Platypus	35:20:29	1:42:35	18.6	0.4	42:57:08	1:57:21	8.8	0.3
	GATK[1]	59:42:39	25:21:58	18.6	0.4	67:21:06	25:36:00	8.8	0.5
	Paragraph[2]	66:02:14	32:24:20	18.6	13.2	86:19:41	45:19:54	12.2	12.2
	GraphTyper[3]	42:52:25	9:14:31	18.6	0.3	49:30:28	8:30:41	8.8	0.2
20	PanGenie	23:46:08	19:39:33	84.8	36.4	24:24:09	19:41:24	84.7	36.3
	BayesTyper	32:03:53	29:59:40	41.0	41.0	44:49:37	41:59:38	41.1	41.1
	Platypus	68:38:45	2:11:46	28.4	0.7	81:28:44	2:42:48	8.8	0.5
	GATK[1]	107:04:36	39:18:12	28.4	0.5	120:51:45	40:43:18	8.8	0.8
	ParaGraph[2]	137:18:30	70:51:31	28.4	14.3	139:56:03	61:10:07	12.9	12.9
	GraphTyper[3]	84:34:29	18:07:30	28.4	0.5	92:58:00	14:12:04	8.8	0.3
30	PanGenie	24:58:54	19:31:51	84.8	36.4	26:48:22	19:41:23	84.7	36.3
	Bayestyper	32:24:13	29:34:54	41.1	41.1	48:30:38	44:34:30	44.4	44.4
	Platypus	99:12:01	1:59:29	39.1	1.0	123:09:20	3:02:53	8.8	0.9
	GATK[1]	143:57:46	44:54:12	39.1	0.5	176:26:20	54:21:41	8.8	0.9
	Paragraph[2]	210:28:50	113:16:17	39.1	14.7	256:00:10	135:53:43	13.3	13.3
	GraphTyper[3]	123:03:06	25:50:33	39.1	0.7	141:57:38	21:51:11	8.8	0.5
	Giraffe[3]	3043:47:18	11:10:38	188.7	45.2	-	-	-	-

[1] GATK was run on SNPs, small and midsize variants only.
[2] Paragraph was run on midsize and large variants only.
[3] GraphTyper and Giraffe were run on large variants only.

Table 3.1: Runtime and memory usage of different genotypers. Runtime (in CPU hhh:mm:ss) and peak memory usage (in GB) of the different genotyping methods at different coverages. For all methods, we show the total resources needed for producing genotypes from raw, unaligned sequencing reads ("total"), as well as the resources needed only for the genotyping step ("genotyping"). Thus, for Platypus, GATK, Paragraph and GraphTyper, the latter excludes the time needed to generate alignments against the reference genome. For Giraffe, it excludes the time for graph construction with vg, indexing and alignment. For k-mer-based approaches (PanGenie and BayesTyper), it excludes the k-mer counting step. All tools were run on an HPC-cluster predominantly consisting of Intel E5-2697v2 (2$\times$ 12 cores and 128 GB of RAM) and Intel Xeon Gold 6136 (2$\times$ 12 cores and 192 GB of RAM) nodes. Table taken from [49].

deletions; Figure 3.7), where the best mapping-based tools (GATK) achieved a wGC of 83% and 86.9% for biallelic insertions and deletions, respectively, at coverage 30$\times$. BayesTyper and PanGenie using 'high-gq' filtering achieved the highest wGCs, both $>$ 99% for non-repetitve and $>$ 97% for repetitive regions (Figure 3.7). For both tools, these good wGCs came at the expense of relatively few genotyped variants, with PanGenie being able to genotype slightly more. We also evaluated our results for SNPs, small and midsize variants using the GIAB high-confidence small variant callset [208] as a ground truth (Figure B.4).

Performance differences were largest for midsize and large variants (Figure 3.8). PanGenie clearly outperformed the mapping-based approaches, especially in repeat regions. Here, PanGenie ('all') reached wGCs for large SVs of 85%, 92% and 76% for biallelic insertions,

biallelic deletions and variants in complex multiallelic regions, respectively, at coverage 30×. This is in contrast with the performance of the best mapping-based tool, achieving only 64%, 79% and 51%, respectively. BayesTyper reached high wGCs, but left 42%, 39% and 77% of these variants untyped, respectively. Using 'high-gq' filtering, PanGenie could reach concordances similar or superior to BayesTyper, while still being able to type much larger fractions of variants (Figure 3.8). PanGenie's genotyping performance for large SVs in repetitive regions is underscored also by the F-score (Figure 3.8): for large biallelic insertions, for example, PanGenie ('all') showed an F-score of 0.7 whereas all other tools reach F-scores < 0.5. We additionally used the SVs contained in the syndip benchmark set [109] to evaluate genotyping performance. Although the absolute results tended to be slightly worse for all tools, PanGenie again produced the most accurate genotype predictions and outperformed the other tools (Figure B.11).

3.5.4 Resources

Comparison of runtimes and memory usages

The runtime and peak memory usage of all genotypers is presented in Table 3.1. For all methods, we measured the resources needed to produce genotypes given the raw, unaligned sequencing reads ('total') as well as the resources needed specifically for genotyping ('genotyping'). For the mapping-based approaches (Platypus, GATK, Paragraph, GraphTyper and Giraffe) the latter excludes the resources needed for aligning the sequencing reads, for the k-mer-based approaches (PanGenie and BayesTyper) it excludes the resources needed for counting k-mers. Note that not all tools are able to genotype all considered variant types. We ran GATK only on SNPs, small and midsize variants. Paragraph was only run on midsize and large variants and GraphTyper only on large variants. We ran Giraffe only for sample NA12878 at coverage 30× and only on large variants, as we observed a very high runtime for its graph alignment step. All tools were run on an HPC-cluster predominantly consisting of Intel E5-2697v2 (2×12 cores and 128 GB of RAM) and Intel Xeon Gold 6136 (2×12 cores and 192 GB of RAM) nodes.

Asymptotic runtime of PanGenie

PanGenie is based on a Hidden Markov Model which, for each variant position, defines one state for each pair of haplotypes in the input panel. Given m variants to be genotyped and n panel haplotypes (which equals twice the number of samples), there will be $O(n^2 \cdot m)$ states. Applying the Forward-Backward algorithm to the HMM corresponds to a runtime quartic in the number of haplotypes, but linear in the number of variants since only states corresponding to the previous/next variant position need to be considered in order to compute Forward/Backward probabilities, respectively. Therefore, the total runtime is $O(n^4 \cdot m)$. If the number of panel haplotypes grows, the algorithm will get slow. We have therefore developed a subsampling strategy to speed up the algorithm when applied to larger pan-

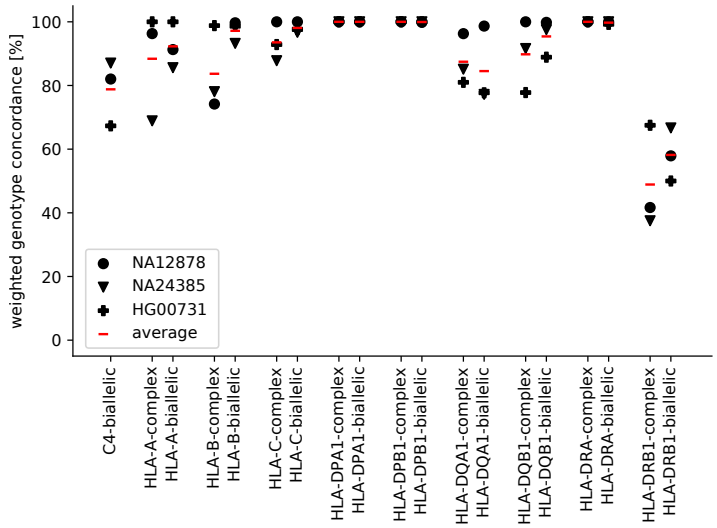


Figure 3.9: HLA genotyping. Weighted genotype concordances for samples NA12878, NA24385 and HG00731 resulting from a "leave-one-out" experiment for HLA genes, as well as the average weighted genotype concordance across all three samples (red). For each gene, we separately computed concordances for the simpler, "biallelic" regions, as well as the more difficult "complex" regions. Figure taken from [49].

els. It will be presented in Chapter 4. For all experiments in this chapter however, we ran PanGenie without subsampling using the full HMM.

3.6 Accuracy in the Major Histocompatibility Complex

This section re-uses material presented in [49]. The evaluation of the assemblies was performed by co-authors of this publication.

To evaluate the accuracy of all 14 haplotype-resolved assemblies in the human leukocyte antigen (HLA) region, we used HLA*ASM [42] to determine assembly HLA types. HLA*ASM successfully processed 27 of 28 input assemblies and identified perfect (edit distance 0) HLA-G group matches [159] for all classic HLA loci (HLA-A, -B, -C, -DQA1, -DQB1 and -DRB1) in all processed input assemblies with one exception (HLA-DRB1 in NA19238), which was resolved by manual curation with minimap2 [106]. To verify the accuracy of the assembly HLA types, we integrated publicly available HLA genotype data for samples from the 1000 Genomes Project [4, 41, 65] for HLA-A, -B, -C, -DQB1 and -DRB1, intersected these with the assembly-implied HLA types, and found perfect agreement in all evaluated cases (9 samples and 85 individual genotype comparisons).

We analyzed our genotyping performance inside of the MHC region. In order to generate the callset described in Sections 3.3 and 3.4, we had used a reference genome that contained sequences of alternative HLA haplotypes in addition to the sequence of chromosome 6 (on which MHC is located). Therefore, many assembly contigs aligned to these

additional haplotypes instead of the respective region on chromosome 6 during variant calling. As a result, the MHC region was not covered well by our callset. We therefore used the same pipeline to generate a second version of our variant calls and pangenome graph using a reference genome that contains only chromosomes 1-22, chromosome X and chromosome Y. We evaluated PanGenie's genotyping performance based on a 'leave-one-out' experiment for samples HG00731, NA12878 and NA24385. Analogously to what we described in Section 3.5 and Figure 3.5, we repeatedly constructed callsets and pangenome graphs excluding the respective samples and evaluated genotypes by comparing to the variants detected in the left out sample. As mentioned previously, we restricted our evaluation to all variants that are genotypable and excluded such that are unique to the left out sample.

We present weighted genotype concordances that we obtained for the HLA genes in Figure 3.9. We separately evaluated variants (all types) located in biallelic regions of the genome and such located in regions with complex bubbles in the pangenome graph. Our callset did not fully cover the C4 genes (C4A and C4B) since the region was not completely covered by contig alignments in most haplotypes (including one of the haplotypes of NA12878) possibly due to the presence of large structural variants in this region. Thus, the evaluation for these genes only corresponds to the parts that were accessible for variant calling.

3.7 Genotyping larger cohorts

This section re-uses material presented in [49].

The low runtime of PanGenie makes it well suited to genotype larger cohorts. As an example use case, we applied it to a set of 300 samples consisting of 100 randomly selected trios from the 1000 Genomes Project using high-coverage data [21]. We used our pangenome graph representation containing all 11 assembly samples as an input for PanGenie, genotyped all bubbles and later converted the resulting genotypes back to obtain genotypes for the individual callset variants. Similar to the approaches that we use in Chapter 4 to analyze HGSVC and HPRC variants across the whole 1000 Genomes cohort, we employed Mendelian consistency of the genotyped trios and the genotype quality reported by PanGenie to compute an integrated score for genotyping reliability of each variant. To this end, we defined different filters based on the predicted genotypes that we list below. One metric used for defining filters is the Mendelian consistency. We computed the Mendelian consistency for each variant by counting the number of trios for which the predicted genotypes are consistent with Mendelian laws. We considered only trios with at least two different genotypes, that is, we excluded a trio if all three genotypes were 0/0, 0/1 or 1/1. This resulted in a more strict definition of Mendelian consistency. In addition to genotyping all 300 trio samples, we also genotyped all 11 panel samples using the full input panel. Genotyping samples that are also in the panel helped us to find cases where panel haplotypes and reads disagreed and thus

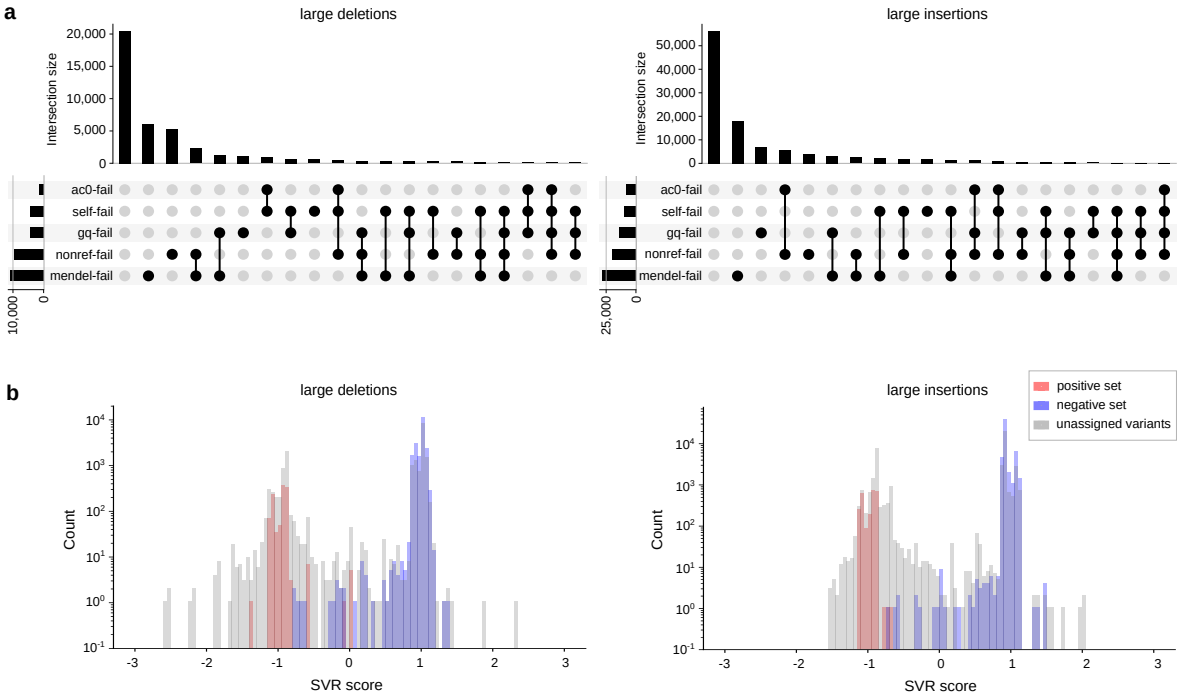


Figure 3.10: Cohort genotype filtering. a Shown are all combinations of filters that we applied to our genotyped variant callset and the respective number of variants in each subset. The black dots indicate that the respective filter failed. **b** Distributions of SVR scores predicted for the positive set (blue), negative set (red) and unassigned variants (grey). Figures taken from [49].

was another useful filter criterion. We defined filters as follows:

- **ac0-fail:** a variant fails this filter if it was genotyped with an allele frequency of 0.0 across all samples.
- **mendel-fail:** a variant fails this filter if the fraction of Mendelian consistent trios was $< 90\%$.
- **gq-fail:** a variant fails this filter if it was genotyped with a genotype quality < 200 in more than 5 samples.
- **self-fail:** a variant fails this filter if the genotype concordance across all panel samples was $< 90\%$.
- **non-ref-fail:** a variant fails this filter if it was genotyped as 0/0 across all panel samples.

For all combinations of filters, we show the number of large deletions and large insertions in each category in Figure 3.10a. To define a strict, high-quality set of variants, we selected all alleles that passed all five filters (Table 3.2). In addition to defining a strict set, we constructed a 'lenient' set for our SV calls (≥ 50 bp) using a machine-learning approach based

variant type	unfiltered set	strict set	lenient set
SNP	12,095,177	11,234,462	
small INS	1,922,163	1,198,663	
small DEL	1,811,123	1,202,791	
midsize INS	110,882	57,699	
midsize DEL	80,027	40,752	
large INS	108,929	56,290	84,836
large DEL	41,499	20,490	34,290

Table 3.2: Number of variants before/after filtering. Number of variants in the unfiltered, strict and lenient sets. The lenient set was only computed for SVs (≥ 50 bp), therefore strict and lenient sets for all other variants are identical. Table taken from [49].

on support vector regression. We used the strict set as a positive set and defined a negative set consisting of all variants that were typed with an allele frequency (AF) $>$ 0.0 and failed at least three filters. For large insertions, the negative set contained 2,611 variants, and for large deletions 1,125. The model then predicted scores between -1 (worst) and 1 (best) for all variants that were in neither the positive nor the negative set. The prediction used 33 features collected from the predicted genotypes based on self-genotyping accuracies of the panel samples, allele frequencies computed across the panel samples and all genotyped samples, genotype quality and Mendelian consistency. We show the distribution of scores for our variant calls in Figure 3.10b. The lenient set was then constructed by adding all variants with a score $>$ -0.5 to our strict SV set. The resulting set of variants contained 78% and 83% of all insertion SVs and deletion SVs, respectively (Table 3.2). To confirm that the lenient set still offers very good genotyping performance, we analyzed allele frequencies and heterozygosities observed from the predicted genotypes for all variants in the lenient set and observed a relationship close to what is expected from the Hardy–Weinberg equilibrium (HWE; Figure 3.11a). When testing for HWE, 90.7% of SV alleles inside of repeats, and 90.9% outside of repeats, showed no significant deviation. Furthermore, observed allele frequencies across all 200 unrelated samples were in excellent agreement with coarse-grained AF estimates obtained from the 22 haplotype assemblies of our 11 input samples (Figure 3.11b). Note that neither of these two measures, HWE and agreement in estimated AFs, has been used when defining the lenient set and therefore serves as independent evidence for PanGenie's performance. PanGenie on average only took about 30 single-core CPU hours per sample.

Our callset contains 209 of 250 medically relevant SVs reported by GIAB [192]. We observed that 174 medically relevant SVs were contained in our lenient set, of which 119 were part of our strictly filtered set. We show the score distribution for these variants as well as AFs and heterozygosities observed across all 200 unrelated samples for the lenient set in Figure B.12.

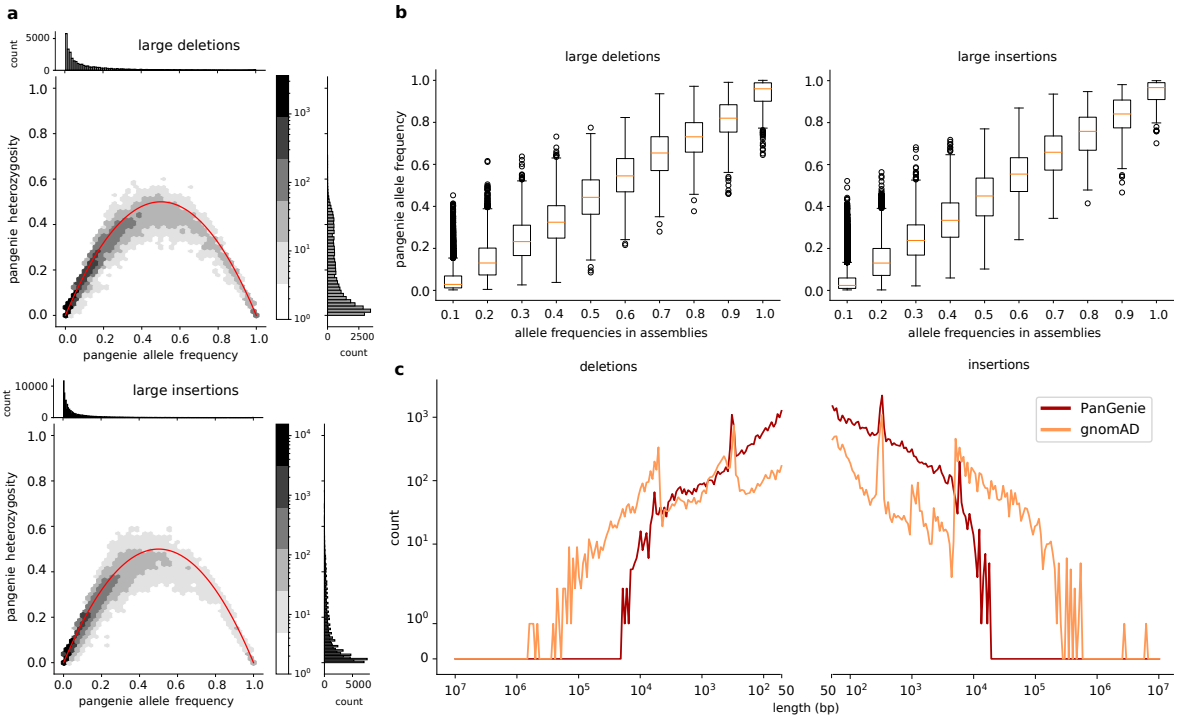


Figure 3.11: Genotyping large cohorts. **a** The hexbin plots show the relationship between AFs and heterozygosities of the PanGenie genotypes for all 200 unrelated samples from the 1000 Genomes Project. The barplots show the one-dimensional distributions of both features (top: AF, right: heterozygosity). All large insertions ($\geq$ 50 bp, n = 84,836) and deletions ($\geq$ 50 bp, n = 34,290) contained in our lenient set were taken into account. **b** Comparison of AFs computed from the PanGenie genotypes for 200 samples and the corresponding AFs observed in the 11 assembly samples from which variants were called. As in a), we consider all large insertions ($\geq$ 50 bp, n = 84,836) and deletions ($\geq$ 50 bp, n = 34,290) contained in our lenient set. In the boxplots, lower and upper limits of the box represent the lower and upper quartiles (Q1 and Q3); the median is marked in yellow. Lower and upper whiskers are defined as Q1-1.5 (Q3-Q1) and Q3 + 1.5 (Q3-Q1), respectively, and outliers are marked by dots. **c** Length distribution of the number of common insertions and deletions (AF $\geq$ 5%) contained in the PanGenie lenient callset and gnomAD. Figure taken from [49].

3.8 Comparison to gnomAD

This section re-uses material presented in [49].

We compared the 119,126 SV alleles genotypable by PanGenie (lenient set) with the SVs that are part of the Genome Aggregation Database (gnomAD) [32]; gnomAD contains SVs collected across 14,891 genomes from different populations [32]. Since gnomAD calls were generated relative to reference genome version GRCh37, we used UCSC liftOver (`https://genome.ucsc.edu/cgi-bin/hgLiftOver`) to convert their coordinates to GRCh38. Requiring a reciprocal overlap of at least 50% or a start, end and variant length deviation of $<$ 200 bp, we found that both callsets had 34,468 variants in common, whereas 84,658 (71%) of our SV alleles were not contained in gnomAD. This finding is consistent with previous observations

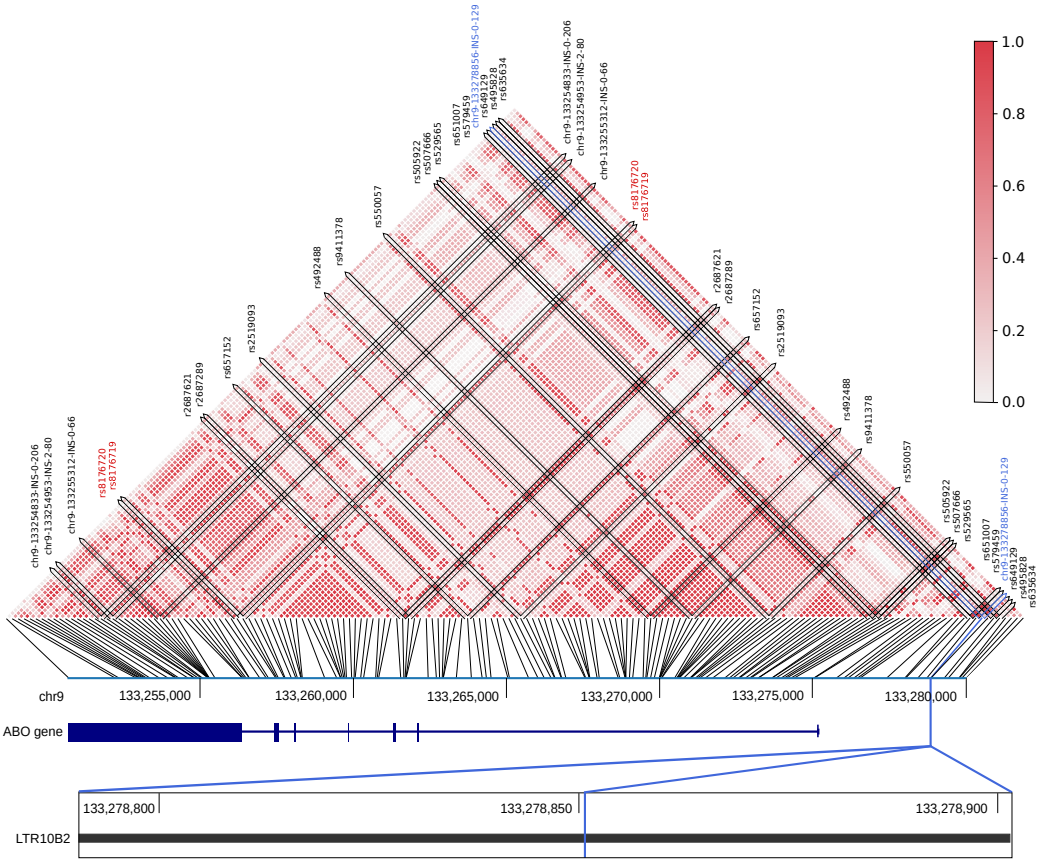


Figure 3.12: LD analysis. We calculated the LD for GWAS variants and SVs that were part of our assembly-based callset. We detected an insertion (marked in blue) close to the ABO gene which was in LD with six GWAS SNPs. The plots show all callset variants in this region; GWAS variants are annotated with their name. Variants colored in red correspond to blood-type markers. Figure taken from [49].

that short-read-based SV detection misses most SVs [206]. Of those 84,658 SVs, around 80% were located in STR/VNTR regions. Furthermore, 43% of these 84,658 variants were common variants with AF $\geq$ 0.05 across all genotyped samples. The length distribution of common insertions and deletions (Figure 3.11c) demonstrates the ability of PanGenie to genotype variants in regions inaccessible by callers based on short-read data alone, and shows its particular impact when genotyping insertions and shorter deletions.

3.9 LD analysis

This section re-uses material presented in [49]. The analysis on nonhuman primates was performed by co-authors of this publication.

Based on the genotypes obtained across all 200 unrelated samples (Section 3.7), we performed an LD analysis. We used gatk4 [39] (v.4.1.9.0) to annotate the calls with variant IDs

from dbSNP (build 154) [174]. We selected all SNPs from our callset that were contained at least five times in the NHGRI-EBI GWAS (genome-wide association studies) catalog [20] and used plink [148] (v.190b618) to determine SVs that are in LD with the GWAS variants ($r^2 \geq 0.8$) within a window of 1 Mb.

For 147 of 3,404 disease-associated SNPs from NHGRI-EBI, we found nearby structural variants that were in LD ($r^2 \geq 0.8$). An insertion of length 129 bp located at position 133,278,856 on chromosome 9, close to the ABO gene, looked particularly interesting (Figure 3.12). It is in LD with six GWAS variants (rs2519093, rs495828, rs507666, rs579459, rs635634 and rs651007) which are related to low-density lipoprotein-cholesterol levels [20]. Of note, neither the GWAS SNPs nor the insertions are in LD with blood-type markers present in our callset (rs8176747 [156], rs8176746 [124], rs8176743 [156], rs8176742 [185], rs8176741 [185], rs8176740 [185], rs7853989 [185], rs1053878 [185], rs8176720 [185] and rs8176719 [124]). The insertion is located in a long tandem repeat (LTR10B2 for ERV1 endogenous retrovirus). Analysis of the insertion sequence revealed that it contains three exact copies of a 43 bp sequence ("TAACGCAGTTTCTGTTTCTGTGTCCTTCCCCTATTGGCTGGGG"; Figure B.13), which appears with copy number 1 in the reference genome. We thus concluded that this insertion is a repeat expansion, leading to four copies of this repeated subsequence. A comparison with nonhuman primate genomes [99, 115] showed that the 43-mer occurs as two copies in gorilla (*Gorilla gorilla*), but is a single copy in chimpanzee (*Pan troglodytes*), bonobo (*Pan paniscus*) and the Sumatran orangutan (*Pongo abelii*). This suggests independent expansion events or incomplete lineage sorting in humans and gorillas.

Another interesting association was an intronic insertion of length 322 bp located at position 28,264,365 on chromosome 12, inside the CCDC91 gene close to a regulatory element reported by ENCODE [55] (Figure B.14). It was in LD with two GWAS variants (rs10843151 and rs11049566), which are both linked to body fat [20]. One of these SNPs, rs10843151, is in perfect LD with many other variants in this region, which suggests that it is probably embedded in the same haplotype block. Such perfect LD provides further evidence that PanGenie is accurately genotyping new insertions within short-read sequencing data.

3.10 Discussion and conclusions

This section re-uses material presented in [49].

We presented an algorithm, PanGenie, that can leverage the long-range haplotype information inherent to a panel of assembled haplotypes in combination with read k-mer counts for genotyping an uncharacterized sample. Although we generated such pangenome reference panels from haplotype-resolved assemblies for the present study, generating these pangenomes was not the main focus of this thesis and PanGenie is not restricted to panels created in this way. In fact, it can be applied to any acyclic genome graph with fully phased

path information.

Traditionally, longer variants are especially difficult to genotype based on short reads only, because such variants are often located in repetitive or duplicated regions of the genome, leading to the difficulty of unambiguously aligning the reads. Approaches based on k-mers additionally lack connectivity information contained in the reads because they do not use the order of k-mers stemming from the same read or read pair. PanGenie overcomes these limitations of short reads because it incorporates long-range haplotype information inherent to the pangenome reference panel that it uses. In comparison to BayesTyper, a graph-based genotyper relying on k-mers, PanGenie genotypes a large fraction of variants not typable by the former. For SVs and indels, PanGenie clearly outperforms mapping-based approaches, which require alignments of reads to a reference genome. Compared with Paragraph, a graph-based method relying on such read alignments, PanGenie produces better genotyping results while additionally providing the ability to jointly genotype SNPs, indels and SVs. Our approach was faster than the other methods, especially when comparing with the mapping-based approaches. The fast runtime makes PanGenie well suited for genotyping larger cohorts, providing the basis for population genetic analysis. In the present study, we have presented an application to a cohort of 300 samples that suggests that SVs in LD with disease-associated SNPs may functionally underlie these associations.

We have hence presented a method that is both fast and leverages a haplotype-resolved pangenome reference to enable genotyping of otherwise inaccessible variants. Although we have tested it only on human data so far, PanGenie can be applied to any diploid genome once corresponding panels of high-quality phased assemblies become available for other species. Our method offers a powerful approach for genotyping and association studies, on ever-larger cohorts, for all variant types – including those currently understudied due to technical limitations.

3.10.1 Limitations and future directions

Although PanGenie improves results over other methods in repetitive regions of the genome, genotyping within these remains challenging. While biallelic variants are less problematic, more complex cases such as segmental duplications, α-satellite repeats or acrocentric DNA are hard to access because of the lack of unique k-mers, but also because such regions are still difficult to assemble. Once a panel of telomere-to-telomere assemblies becomes available, future experiments can clarify which additional loci are amenable to genotyping with PanGenie.

Our model assumes that the unknown haplotypes of the sample to be genotyped are mosaics of the given panel haplotypes. Therefore, currently it cannot be used to genotype rare variants that are present only in the sample, but in none of the other haplotypes. We believe that there are exciting opportunities to develop methods to discover variation that our approach has not captured because it was not present in the reference panel. For example, one could either filter the reads for as yet “unexplained” k-mers and use those for the discovery

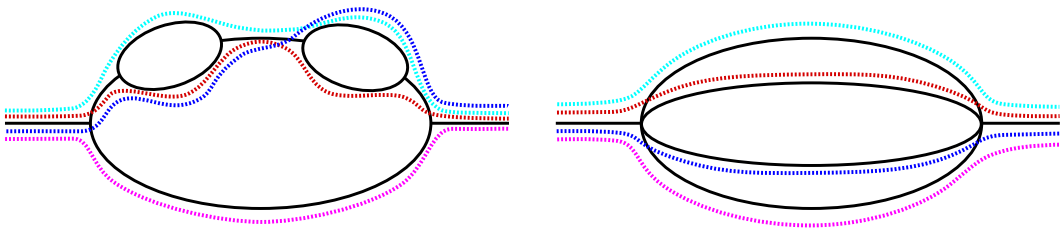


Figure 3.13: Nested variants representation. A bubble structure with nested variants is shown on the left, as well as four haplotype paths. On the right side, this bubble is represented in terms of the merged, multiallelic VCF file format used by PanGenie. Each distinct path through the bubble will be interpreted as an allele, even though paths are partially identical. The latter representation does not allow recombination between the two nested variants.

of rare variants, or utilize PanGenie's output as a personalized pangenome reference graph to map reads to.

Currently, PanGenie computes recombination probabilities only based on the distance between two adjacent variant positions. In order to increase genotyping accuracy, recombination probabilities could be adjusted to account for recombination hotspots in the genome, or to regions with low recombination rates.

The runtime of our method depends on the number of input haplotypes, because we defined a hidden state for each possible pair of haplotypes that can be selected for each bubble. Therefore, additional engineering would be required to use much larger panels, which could be approached similarly to how statistical phasing packages prune the solution space and/or proceed iteratively [19, 37, 79]. Such techniques could also pave the way toward a version of PanGenie for polyploid genomes, which would be prohibitively slow when implemented without such additional optimization. Another option could be to compute founder sequences [134, 190] from the input haplotypes prior to genotyping in order to reduce the runtime for large panels. This smaller set of representative haplotype sequences could then be used as input to PanGenie instead of the full set of haplotypes. Similar ideas have been used recently to successfully reduce the complexity of accurate, pangenome-based variant detection [135].

PanGenie determines k-mers unique to a bubble region for genotyping (Section 3.2.1). These k-mers are allowed to occur in one or more alleles inside of the bubble, but nowhere else in the pangenome. In the current implementation of PanGenie, the number of unique k-mers selected for larger bubbles is restricted to 300 to reduce memory usage and runtime. Which k-mers are selected is completely arbitrary. Here, more sophisticated strategies to choosing the most informative k-mers could be implemented. For example, one could try to select k-mers such that each allele is equally covered by k-mers. Also, one could prefer those k-mers that distinguish alleles by selecting the ones that are not only specific to a bubble region but also occur only in exactly one of the allele paths inside of the bubble.

Currently, the pangenome graph provided as input to PanGenie is stored in VCF format. Each bubble in the graph is represented in terms of a variant record and the alternative alleles specify all distinct paths covered by the input haplotypes in this genomic region (Figure 3.13, right). For large bubbles, there might be a different path for each haplotype, although haplotypes might be partly identical. However, this VCF-based representation does not account for nested variation that might be present in a bubble and thus does not consider the actual graph structure underlying such a region (Figure 3.13, left). This might especially be problematic as the number of haplotypes increases. The more haplotypes there are, the more alleles might be overlapping across haplotypes leading to larger bubbles that tend to have a distinct allele path for each single haplotype in the resulting VCF. Genotyping such bubbles is challenging, as PanGenie needs to decide between a large number of possible alternative alleles. Therefore, leveraging a more sophisticated pangenome graph structure that encodes variation in terms of nested bubbles and is independent of a reference genome might be more beneficial. It might be helpful to genotype nested variants separately instead of considering the whole bubble at once, allowing recombination between nested variants inside of a bubble. Recursive strategies for genotyping nested variation within larger bubbles could be useful here, e.g. as used by gramtools [105], however, this might require a different problem formulation which is not based on a Hidden Markov Model. An existing format that is able to encode nested variation is the GFA format. It encodes a graph by defining its nodes (“segments”) and edges between them (“links”), along with sequence information. However, such graphs are more difficult to handle since files are often large, and also because bubble detection algorithms are required in order to localize bubbles in these graphs prior to genotyping.

Chapter 4

Application: Genotyping Large Cohorts

The recent advances in accurate long read sequencing (Section 2.2) as well as the development of efficient methods for chromosome-scale assembly (Section 1.8) have enabled the generation of highly accurate, haplotype-resolved assemblies of human samples on larger scales [24, 59, 145]. Such assemblies enable the construction of highly accurate pangenome references, which provide insights into complex genomic regions that were previously difficult to analyze. Alternative allele sequences are missing from the current reference genome. Furthermore, many SVs are located in repetitive sequence context. As a consequence, over two-thirds of SVs have been missed by short-read based callers [24, 199, 206]. Methods based on pangenome graphs enable fast, short-read based genotyping of variants previously inaccessible (Chapter 3). This allows to include such SVs in population-based downstream analyses, such as genome-wide association studies.

Recently, two consortia have produced haplotype-resolved assemblies for multiple human samples. The Human Genome Structural Variation Consortium (HGSVC) [24, 46] generated phased assemblies for 35 human samples, focusing on variant calling and the analysis of structural variations across these genomes. More recently, the Human Pangenome Reference Consortium (HPRC) [90, 113] produced such assemblies for 47 human samples, with the goal of providing an alternative to the linear reference genome. Both consortia deliver pangenome representations that can serve as input for pangenome-based genotyping methods like PanGenie, which was presented in Chapter 3. The fast runtime of PanGenie enables accurate genotyping of thousands of samples based on short sequencing reads using such pangenome references. This chapter presents the results of applying PanGenie to the datasets of the HGSVC and HPRC consortia. It demonstrates that PanGenie works well in practice and that it can be used to produce high quality genotypes across large populations. The main focus is on structural variation, as many SVs have been challenging to detect and analyze in previous studies. It is further demonstrated how SNP genotypes across large populations can be used in order to detect potential carriers of rare inversions in human samples. The ex-

periments presented here demonstrate the added value from pangenome-based genotyping being able to access variation previously inaccessible by short reads.

4.1 HGSVC project

This section presents the work of the Human Genome Structural Variation Consortium (HGSVC) and focuses on my contributions to this project. Results were published as part of a Science publication [46]. Sections 4.1.3, 4.1.4 and 4.1.5 re-use material from this paper. Section 4.1.2 re-uses material from [49]. See Sections E.4 and E.5 for information on author contributions and publication details.

4.1.1 Introduction

The Human Genome Structural Variation Consortium (HGSVC) [24, 46] constructed highly accurate, haplotype-resolved assemblies of human samples using the PGAS pipeline [145] based on long-read PacBio sequencing data (CLR and CCS) as well as Strand-seq data. Unlike other methods [98], PGAS does not depend on parent-child trio data [46, 145] (see also Section 1.8). Haplotype-resolved assemblies for 35 human samples (70 haplotypes) were produced in this way, which included 32 unrelated individuals. These assemblies were used in order to call genetic variants, including SNPs, indels and SVs across these samples. The focus here was especially on SVs, as these new assemblies enable the discovery of novel SVs previously inaccessible by short-read or long-read based data. Variant calling was performed using a new method PAV, developed by a co-author, which detects variants by comparing the haplotype sequences of the assembly samples to the linear reference genome, producing a callset with phased genotypes for all assembly samples. As a part of this project, I genotyped the variants detected by PAV across a diverse cohort consisting of 3,202 human samples [2, 21] using PanGenie, enabling estimation of allele frequencies as well as the discovery of associations between genotypes and gene expression, splicing and candidate disease loci [46].

4.1.2 Speeding up PanGenie for larger panels

This section re-uses material presented in [49].

The asymptotic runtime of the Forward-Backward algorithm implemented in PanGenie is $O(n^4 \cdot m)$, where n is the number of reference haplotypes and m the number of genotyped variants (see Section 3.5.4 for details). Thus, the algorithm might get slow once the number of haplotypes in the panel gets larger. Therefore, we have implemented a subsampling approach to efficiently genotype larger panels, such as the ones produced by the HGSVC and HPRC projects. Given n input haplotypes, the idea is to subsample l sets of haplotypes each of size k from the full set. The Forward-Backward algorithm is then run separately on each of these subsets in order to produce genotype predictions for the variants. In the end, genotype likelihoods produced based on each subset for a variant position are combined by iteratively adding up likelihoods and normalizing them. PanGenie automatically switches to this subsampling approach if the number of input haplotypes in the panel exceeds 30. It

was used for the HGSVC experiments presented in the following sections as well as for the HPRC experiments presented in Section 4.3.

Asympotic runtime

Let us assume we split the set of n input haplotypes in l subsets each of a fixed size k. When genotyping m variants, PanGenie's genotyping step is now run separately on each of the l sets in time $O(k^4 \cdot m)$ (Section 3.5.4). This will result in a total runtime linear in the number of subsets, i.e. $O(l \cdot k^4 \cdot m)$.

4.1.3 Variant calling from haplotype-resolved assemblies

This section provides an overview of the variant calling pipeline that was developed and run by Peter Audano, a co-author of [46] and re-uses some material from this publication.

Variants were identified from the haplotype-resolved assemblies with the Phased Assembly Variant (PAV) caller. For each haplotype, the assembly contigs were aligned to the GRCh38 reference genome using minimap2 [106]. Multi-mapping issues were resolved by trimming alignments based on a dynamic programming approach. Variants contained within alignments were detected from the CIGAR string of the mapped contigs. Structural variants can cause breaks in the alignments. In such cases, alignment breakpoints were analyzed in order to identify insertion or deletion events. In order to combine variant calls generated for each haplotype by PAV, a three-step merging approach was applied to decide whether similar alleles observed across different haplotypes represented the same variant or not. First, variant alleles matching exactly were intersected. Second, variant calls with intersecting reference coordinates were combined if their reciprocal overlap (RO) was at least 50%, i.e. the interval covered by a variant overlapped at least 50% of the interval covered by another variant and vice versa. For insertions, the end position was computed as the sum of start position and length. Finally, variants within 200 bp and 50% overlap by size were merged, since smaller variants are often missed by the RO criterion. SNP variants were only combined if their position and alternative base were exactly the same. The output of PAV is a fully-phased, multisample VCF file containing all variant alleles after merging. It was used in subsequent steps in order to generate an acyclic and directed pangenome representation of these calls, which was used as input to PanGenie.

4.1.4 Genotyping SVs across a cohort of 3,202 individuals

This section re-uses material presented in [46].

The variants detected by PAV were genotyped across all 3,202 samples from the 1000 Genomes Project [2, 21] in order to demonstrate the utility of PanGenie for large cohorts. For each sample, we provided PanGenie with short-read sequencing reads (in FASTQ for-

	PAV calls	Minus $\geq$ 20% missing	Minus chrY + chr_*	Merged set (PanGenie input)
SNPs	15.810.489	15.610.830	15.609.695	15.488.649
Deletions (1-49 bp)	649.065	642.348	642.341	636.570
Insertions (1-49 bp)	405.981	400.527	400.513	397.095
Deletions ($\geq$ 50 bp)	41.393	37.327	37.169	35.862
Insertions ($\geq$ 50 bp)	66.197	61.788	61.604	60.283

Table 4.1: Number of HGSVC variants. The first column provides the number of input variants (from PAV callsets). The second and third columns show the numbers of variants obtained after removing positions with at least 20% missing alleles in the panel and those located outside of chromosomes 1-22 or chromosome X. The last column provides the numbers of variants contained in the final graph used as input for PanGenie. Table taken from [46].

mat) as well as a multisample VCF file containing phased variant calls from the 64 assemblies of all unrelated samples (i.e. excluding six child haplotypes). This input VCF file was derived from the merged PAV callsets produced for SNPs, indels and SVs. At first, we removed all positions for which more than 20% of the panel haplotypes carried a missing allele. Furthermore, we kept only variants located on chromosomes 1-22 and chromosome X for genotyping. We then created a multiallelic VCF representation in which overlapping variants were combined into multiallelic bubbles using the same approach as explained in Section 3.4. For each bubble, we kept track of the individual variants it was composed of, which allowed us to later translate genotypes computed for bubbles to genotypes for the underlying PAV variants. Some variants in the input PAV callsets were overlapping on the same haplotype (e.g. a SNP inside of a deletion). We removed such conflicts by setting the corresponding alleles to missing (“.”). In this way, the VCF encodes an acyclic and directed pangenome graph representing the panel genomes. Table 4.1 provides an overview of the number of PAV variants obtained after the different pangenome graph construction steps. We genotyped all bubbles contained in the pangenome graph across all samples from the 1000 Genomes Project using PanGenie. We converted the genotypes computed for the bubbles back to genotypes for the PAV variants from which the graph was constructed.

We started with a pilot set consisting of 300 individuals selected from the 3,202 samples of the 1000 Genomes cohort. This subset was constructed by randomly choosing 20 trios from each of the five superpopulations (AFR, AMR, EAS, EUR, SAS). We then ran PanGenie in order to genotype all 15.5 million SNPs, 1.03 million indels, and 96.1 thousand SVs across these 300 samples. For comparison, we additionally ran state-of-the-art method Paragraph [26] to derive genotypes for all SVs. As a quality control measure, we computed allele frequencies across all 200 unrelated samples from the PanGenie and Paragraph genotypes and compared them to the allele frequencies derived from the PAV calls for all 64 unrelated assembly haplotypes. For Paragraph, we observed allele frequency correlations (Pearson correlation) of 0.61 and 0.54 for SV deletions and insertions. For PanGenie, these values were 0.85 and 0.86, respectively. Plots comparing the Paragraph and PanGenie allele

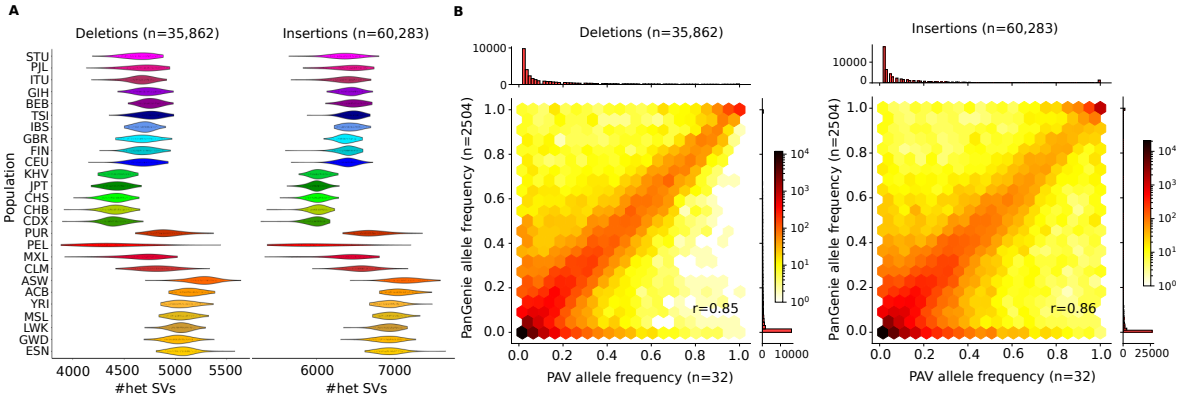


Figure 4.1: Unfiltered HGSVC genotypes. A Shown are the number of SVs typed as heterozygous by PanGenie for different populations. The plots are based on the unfiltered callset containing all 96,145 SVs (35,862 deletions and 60,283 insertions) **B** Concordance of AF estimates from the assembly-based PAV discovery callset and AF estimates from genotyping unrelated Illumina genomes (n = 2,504) with PanGenie (unfiltered genotype set of 96,145 SVs). Marginal histograms are in linear scale. Figures taken from [46].

frequencies to the ones for PAV indicate that Paragraph tends to genotype variants as heterozygous and especially struggles with genotyping insertions (Figure C.1). We concluded that PanGenie seems to be more suitable for genotyping our SV calls and proceeded with genotyping all SNPs, indels, and SVs across the 3,202 samples using PanGenie. We determined the number of heterozygous SVs for each population from these genotypes (Figure 4.1A) and observed higher numbers for the African populations, reflecting their increased genetic diversity. We also computed allele frequency correlations from the allele frequencies derived from the PanGenie genotypes for all 2,504 unrelated samples and the PAV calls. We observed correlations of 0.98, 0.95, and 0.85 for SNPs, indels and SVs, respectively (Figure C.2B, Figure C.3B, Figure 4.1B). However, these numbers indicate that there were variants for which PanGenie and PAV allele frequencies differ significantly. In order to filter out such potentially wrongly genotyped calls, we defined a strict subset of variants based on statistics we computed from the genotypes of the 3,202 samples. Similarly to what we described in our PanGenie publication [49] (Section 3.7), we defined the five filters listed below:

- **ac0_fail**: a variant fails this filter if it was genotyped as absent (genotype 0/0 or ./.) by PanGenie in all 3,202 samples (i.e. the allele frequency was zero).
- **mendel_fail**: there are 602 trios among the 3,202 genotyped samples. For each variant, we counted the number of trios with Mendelian-consistent genotypes. As in [49] (Section 3.7), we only took trios with at least two different genotypes into consideration, meaning we skipped trios in which all samples were typed as 0/0, 0/1 or 1/1, respectively. A variant fails this filter if the Mendelian consistency is below 90%.
- **gq_fail**: a variant fails this filter if there were at least 200 genotype predictions with genotype quality < 200.

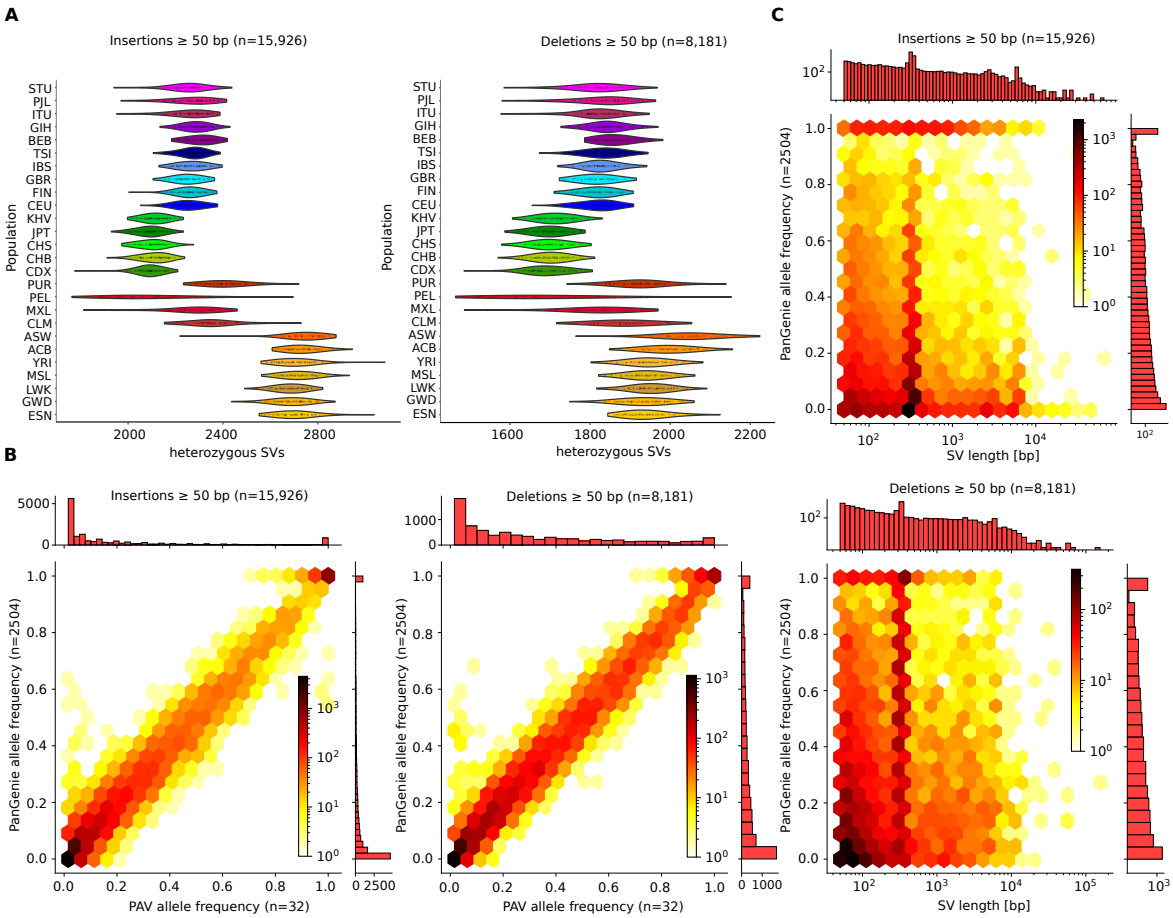


Figure 4.2: Strictly filtered HGSVC genotypes. **A** Distribution of heterozygous SV counts per diploid genome broken down by population, based on PanGenie genotypes passing strict filters. **B** Concordance of AF estimates from the assembly-based PAV discovery callset and AF estimates from genotyping unrelated Illumina genomes (n = 2,504) with PanGenie (strict genotype set of 24,107 SVs). Marginal histograms are in linear scale. **C** PanGenie allele frequencies were computed based on the HGSVC genotypes of all 2,504 unrelated samples. Only SVs contained in the strict set (n = 24,107) are considered. Figures taken from [46].

- **nonref_fail**: a variant fails this filter if all panel samples were genotyped as homozygous reference.
- **loo_fail**: in addition to genotyping the 3,202 samples, we conducted a leave-one-out experiment, in which we repeatedly took out one of the panel samples from the input and used PanGenie to genotype it based on the remaining samples in the panel. We then compared the predicted genotypes to the ground-truth genotypes of the left-out sample. This enabled us to compute the genotype concordance across all panel samples at each variant position. This filter fails if the genotype concordance of the panel samples is below 80%.

We applied these filters to all SNPs, indels and SVs genotyped by PanGenie. For SVs, 16,343 out of 60,238 insertions (27%) failed the “ac0_fail” filter and were genotyped with an allele

frequency of zero across all 3,202 samples (Figure C.4A, Figure C.5A). For deletions, 8,948 out of 35,862 (25%) were genotyped as homozygous reference in all samples. About 57% of these variants are rare and were carried by only a single haplotype in the input panel. Such variants are, in particular, difficult to genotype by a panel-based approach like PanGenie, especially if the k-mer counts show no strong indication for the presence of an allele in a sample. To obtain a filtered callset, we removed all variants for which at least one of the five filters failed. This led to a rather stringent, but high-quality set of genotypes that served as a basis for further analysis. Our filtered set contains 12,283,650 SNPs (79%), 705,893 indels (68%), and 24,107 SVs (25%). We provide callset statistics for this strict set in Figure 4.2. Figure 4.2A shows the number of heterozygous SVs for different populations, demonstrating expected patterns of diversity [2]. Figure 4.2B shows the allele frequencies obtained for SVs across the PanGenie genotypes, as well as the corresponding allele frequencies in the input panel haplotypes. The allele frequencies for PanGenie were computed based on all 2,504 unrelated individuals from the full panel of the 3,202 samples. For both variant types, insertions and deletions, the allele frequencies matched well with very few outliers, indicating that the genotypes are of good quality. Note that allele frequencies were not directly used for filtering. We obtained an allele frequency correlation of 0.99 (0.98 for deletions, 0.99 for insertions). Likewise, allele frequency correlations for SNPs and indels in this filtered set were both equal to 0.99 (Figure C.2, Figure C.3). We also investigated the relationship between variant length and allele frequencies across the PanGenie genotypes. The peaks (Figure 4.2C) show a clear tendency of Alu insertions towards lower allele frequencies, while an opposite trend is observed for Alu deletions. We suspect that this behavior is caused by Alu insertions present in the reference genome. Figures C.6 and C.7 show the relationship between F_{ST} values of the five superpopulations (AFR, AMR, EAS, EUR, SAS) and the length of the SVs. F_{ST} is a measure used to analyze population structure [72]. It calculates the proportion of genetic variation among subpopulations relative to the overall genetic variation [123]. Here, for each superpopulation, F_{ST} was computed between individuals belonging to the respective superpopulation and the union of the remaining populations. We observed higher F_{ST} values for African and East Asian populations indicating higher degrees of differentiation among these populations.

For SVs, our filtered set contained only 25% of all input variants. Besides defining a strict set, we used support vector regression in order to generate a larger, more lenient set of SVs. This approach differs from the one introduced in Section 3.7 in the set of features that are used for the regression. Here, we included additional features computed from a leave-one-out experiment performed for all panel samples and concordances with k-mer-based presence/absence genotyping contributed by Tobias Rausch, a co-author of [46]. Our model is designed to assign scores close to -1 to poorly genotyped SVs and scores around 1 to those passing all of our filters. For training, we used our strict SV set as “true positives”. The “true negatives” were defined as all variants genotyped with an allele frequency larger than zero (“ac0_fail” did not fail) that failed at least three of the remaining filters. We predicted

callset	SV Insertions			SV Deletions		
	number	correlation	genotype concordance [%]	number	correlation	genotype concordance [%]
unfiltered set	60.283	0,86	85,1	35.862	0,85	86
lenient, cutoff -0.5	31.680	0,95	89,8	18.660	0,95	91,5
lenient, cutoff 0.0	27.565	0,97	92,1	16.065	0,97	93,3
lenient, cutoff 0.5	23.128	0,98	94,3	13.384	0,98	94,7
strict set	15.926	0,99	96,9	8.181	0,98	96,1

Table 4.2: Callset statistics for the HGSVC lenient set. The table lists the number of variants ($\geq$ 50 bp) contained in the unfiltered, strict and lenient SV callsets. The machine-learning-based process to define the lenient set can be tuned by setting different cutoffs separating high- from low-quality SVs. Columns named "correlation" state Pearson correlation between PanGenie (2,504 unrelated samples) and PAV genotype allele frequencies. Columns named "genotype concordance" state average genotype concordance for all assembly haplotypes estimated via leave-one-out experiments. Table taken from [46].

regression scores for all yet unlabeled SVs (with an allele frequency > 0). We then created more lenient callsets by adding those variants to the strict set for which the scores were above a certain threshold. We investigated three different cutoffs: -0.5, 0.0, and 0.5. Table 4.2 contains corresponding callset statistics. As expected, numbers improved as we increased the cutoff used to define a lenient callset. While allele frequency correlations were 0.86 and 0.85 for insertions and deletions in the unfiltered set, respectively, they reached levels between 0.95 and 0.98 when applying the different cutoffs. Likewise, average genotype concordances of the PanGenie genotypes with the PAV calls for the assembly samples increased, reaching levels above 94% for cutoff 0.5.

Based on these statistics we chose threshold -0.5 as our final value to define the lenient set. It contained 31,680 SV insertion alleles and 18,660 deletion alleles (Table 4.2) and still reached allele frequency correlations of 0.95 for both SV types, indicating that the PanGenie allele frequencies were consistent with the PAV results (Figure 4.3a). Average genotype concordances were 89.8% for insertions and 91.5% for deletions. Furthermore, we compared allele frequencies and heterozygosities and found that variants largely behaved as expected by Hardy-Weinberg equilibrium (Figure C.2B, Figure C.3B, Figure C.4B, Figure C.5B).

4.1.5 Added value from graph-based genotyping into short-read WGS data

This section provides a summary on evaluation results produced by co-authors presented in [46] and re-uses material from this publication.

A comparison of the PanGenie genotypes for the 1000 Genomes samples to state-of-the-art short-read SV discovery sets showed that 59.9% and 42.5% of SV alleles contained in the lenient and strict sets, respectively, were not detectable from short reads. Figure 4.3b shows the length distribution of common SVs (allele frequency > 5 %) contained in the PAV calls, the PanGenie lenient set, as well as three short-read based callsets [3, 32, 46]. Results

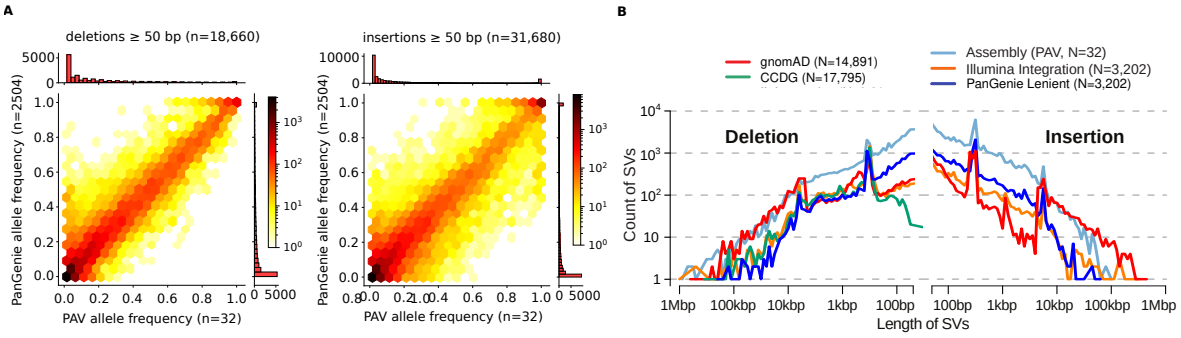


Figure 4.3: Leniently filtered HGSVC genotypes. **A** For PanGenie, allele frequencies were computed based on the genotypes of all 2,504 unrelated samples. The PAV allele frequencies were computed based on all 64 assemblies. Only SVs ($\geq$ 50 bp) contained in our lenient callset (cutoff -0.5, n = 50,340) were considered. **B** Length distribution of common SV sites (AF $>$ 5%) represented in assembly-based callsets, including variants genotyped by using PanGenie and all common variants from population-scale studies from the Genome Aggregation Database (gnomAD-SV) and CCDG (insertions from CCDG omitted because of lack of data). Figures taken from [46].

demonstrate that short-read based callsets miss large fractions of common SVs, which is consistent with the results shown in Chapter 3 as well as with results of previous studies [206]. The assembly-based callset PAV and the PanGenie genotypes contained increased numbers of deletions below 250 bp and insertions under 1 kbp, which were missed by short-read based callers. The ability to genotype variants across a cohort that were previously inaccessible to purely short-read based callers also enabled including such variants in genome-wide Quantitative trait locus (QTL) analyses. QTL analysis aims at identifying associations between genotypes and phenotypes (molecular and clinical), with the goal to explain the roles of genetic variation in diseases. A QTL analysis based on the PanGenie strict set showed that 48% of the most likely causal expression QTL SVs ("lead eQTLs") were previously inaccessible by purely short-read based callers. These results demonstrate the added value from pangenome-based genotyping with PanGenie: it is able to reliably genotype a large portion of (common) structural variants that purely short-read based callers are not able to access, enabling the inclusion of such variants in downstream analyses.

4.1.6 Discussion

The pangenome-based genotyping method PanGenie was applied to a pangenome representing 32 human samples (64 haplotypes), created from high quality haplotype-resolved genome assemblies. In this way, variants were re-genotyped across a large and diverse cohort of 3,202 samples from different superpopulations and filtered sets of genotypes were constructed. Evaluation based on different population-based statistics, like the Mendelian consistency or comparison of allele frequencies, showed that our genotypes are of high quality. We showed that 50,340 structural variants can be reliably genotyped by PanGenie across the populations, and 59.9% of these variants were previously inaccessible by

linear alignment based, short-read discovery callsets, demonstrating the added value of a pangenome-based framework. Our genotyping method took only around 30 CPU hours per sample for genotyping, unlike other approaches, that require time-consuming read alignments to a reference genome and are thus much slower. Furthermore, results demonstrate that SV genotypes produced by PanGenie enable QTL analysis for the discovery of disease associated variants. Our analyses revealed 31.9% of SV-eQTLs and 48% of lead SV-eQTLs that were previously not accessible by short-read approaches.

However, the study also reveals some limitations of PanGenie. Especially genotyping rare variants with low allele frequencies across the assembly samples was challenging, since PanGenie tended to genotype them as absent, resulting in allele frequencies of zero across all 3,202 samples. As a consequence, 26.3% of all SV alleles needed to be excluded from any filtered set, since they could not be re-typed in any sample and therefore did not provide any information for downstream analyses. Possible reasons why PanGenie tends to genotype such variants as absent could be the lack of unique k-mers for all or some of the variant alleles at a variant locus. The absence of unique k-mers in a region or unbalanced k-mer distributions across the alleles can lead to wrong conclusions about the genotypes, because in such cases, genotypes are imputed from the panel haplotypes. Therefore, if a variant is rare in the panel, there is a bias towards genotyping it as absent. Another factor that could have affected PanGenie's genotyping performance for rare and common variants is that during variant calling with PAV, similar overlapping SV alleles were merged into a single allele. In some cases, this might have led to "overmerging" alleles, i.e. merging sequences not actually representing the same variant allele. Since PanGenie is a k-mer based method, small differences in allele sequences between the panel and the sample to genotype might lead to wrong genotypes because the observed k-mer spectrum is different from the expected one.

4.2 Identifying rare inversions

This section demonstrates how the HGSVC genotypes produced by PanGenie for the 1000 Genomes cohort can be used in order to detect rare inversions. The material presented has been published as part of a Cell publication [146]. Sections 4.2.1 and 4.2.2 re-use material presented in this publication. See Section E.6 for information on author contributions and publication details.

4.2.1 Introduction

This section re-uses material presented in [146].

Compared to insertions and deletions, inversions are a relatively rare class of structural variants in humans [3, 46, 181]. In case of an inversion, a chromosomal segment is inverted (Section 1.3). Unlike other classes of SVs, inversions remain challenging to discover and analyze [3, 32, 46, 70, 80, 94, 97, 155, 167, 181]. This is because inversions are often flanked by segmental duplications that are too long to be spanned by sequencing reads. Furthermore, balanced inversions do not lead to any gains or losses of DNA, which makes their detection more difficult. Inversions can suppress recombination and have been shown to be associated with diseases, such as Haemophilia A or the Hunter syndrome [16, 102, 180]. In our study [146], we characterized the full spectrum of inversions across 41 human samples based on Strand-seq data, the haplotype-resolved assemblies generated by the HGSVC project [46] (Section 4.1), and Bionano optical mapping data. We detected 729 inversions with lengths ranging from 50 bp to several mega base pairs. As a part of this study, I developed an approach to find potential carriers of rare inversions detected across the 41 samples in a large cohort of 3,202 samples from the 1000 Genomes project. This approach uses the SNP genotypes computed by PanGenie across the 3,202 samples in the HGSVC project described in detail in Section 4.1. The idea is to find samples that share a high fraction of rare SNPs with the samples carrying the inversion in the respective genomic region.

4.2.2 Identifying potential inversion carriers using PanGenie

This section re-uses material presented in [146].

The genotypes computed for the HGSVC SNPs across all 3,202 genomes (see Section 4.1) can be used in order to detect potential carriers of rare inversions. Here, we used them in order to identify potential carriers of two inversions of interest both of which were absent from the other Strand-seq discovery samples. The first one is a 23 Mbp pericentromeric inversion on chromosome 2 detected in NA19650 with its distal breakpoint lying near ($\sim$ 2 Mbp) the ancestral chromosome 2 fusion point specific to humans. The second inversion is a 5 Mbp inversion located on chromosome 15 detected in HG02492. It overlaps the well-known Prader-Willi/Angelman syndrome (PWAS) type II critical region [30] and has been postulated to predispose to disease [63].

We considered SNP alleles present in the inversion haplotypes of these two samples as follows. We used the HGSVC SNP genotypes [46] (Section 4.1.4) available for all 3,202 samples from the 1000 Genomes Project [21] to determine which of these SNPs are rare (allele frequency ≤ 0.01 across unrelated samples). In HG02492 we identified 103, and for NA19650 we detected 333 such rare SNPs within the respective inverted segments. We then counted these rare alleles in the genotypes of all 3,202 samples. Those samples that shared a high number of rare SNPs with the respective inversion haplotype were considered potential carriers of the inversion. For the inversion on chromosome 15, we identified four samples sharing a high number of rare SNP alleles with HG02492 (Figure 4.4A): HG02491 (102/103 alleles in common; 99%; mother of HG02492), HG02784 (101/103; 98%), HG02725 (74/103; 72%), and HG03639 (74/103; 72%). The shared alleles are evenly distributed along the inverted sequence (Figure 4.4B). All these samples are part of the Punjabi South Asian population. For the pericentromeric inversion on chromosome 2, sample NA19648 (mother of NA19650) shared 330/333 (99%) rare alleles with the respective haplotype-resolved segment in NA19650 (Figure 4.5A,B).

Fluorescence *In Situ* Hybridization (FISH) validation experiments were performed in order to verify the results of our approach. For the inversion on chromosome 15, all five samples predicted as carriers could be verified (Figure 4.4C) which suggests a founder inversion event in the Punjabi population in association with this haplotype. Likewise, both carriers of the inversion on chromosome 2 could be verified by FISH (Figure 4.5C).

4.2.3 Discussion

In this section, an approach was presented that can detect carriers of rare inversions in a large cohort for which SNP genotypes are available. It was demonstrated to successfully detect carriers of two rare inversions of interest, which could be experimentally verified using FISH. However, there are some limitations. The method presented here is mainly suited for detecting longer inversion events, since it requires rare alleles to be present in the inverted region. Here, we applied the approach to inversions of 5 Mbp and 23 Mbp in length. Furthermore, it requires phased SNP genotypes to be available for the detection sample, as well as SNP genotypes for a cohort.

In conclusion, the experiments show that rare inversions from a set of discovery samples can be accurately inferred in large whole-genome sequencing cohorts based on SNP genotypes. This enables to study associations between inversion polymorphisms and diseases, which were previously not accessible to investigation [146].

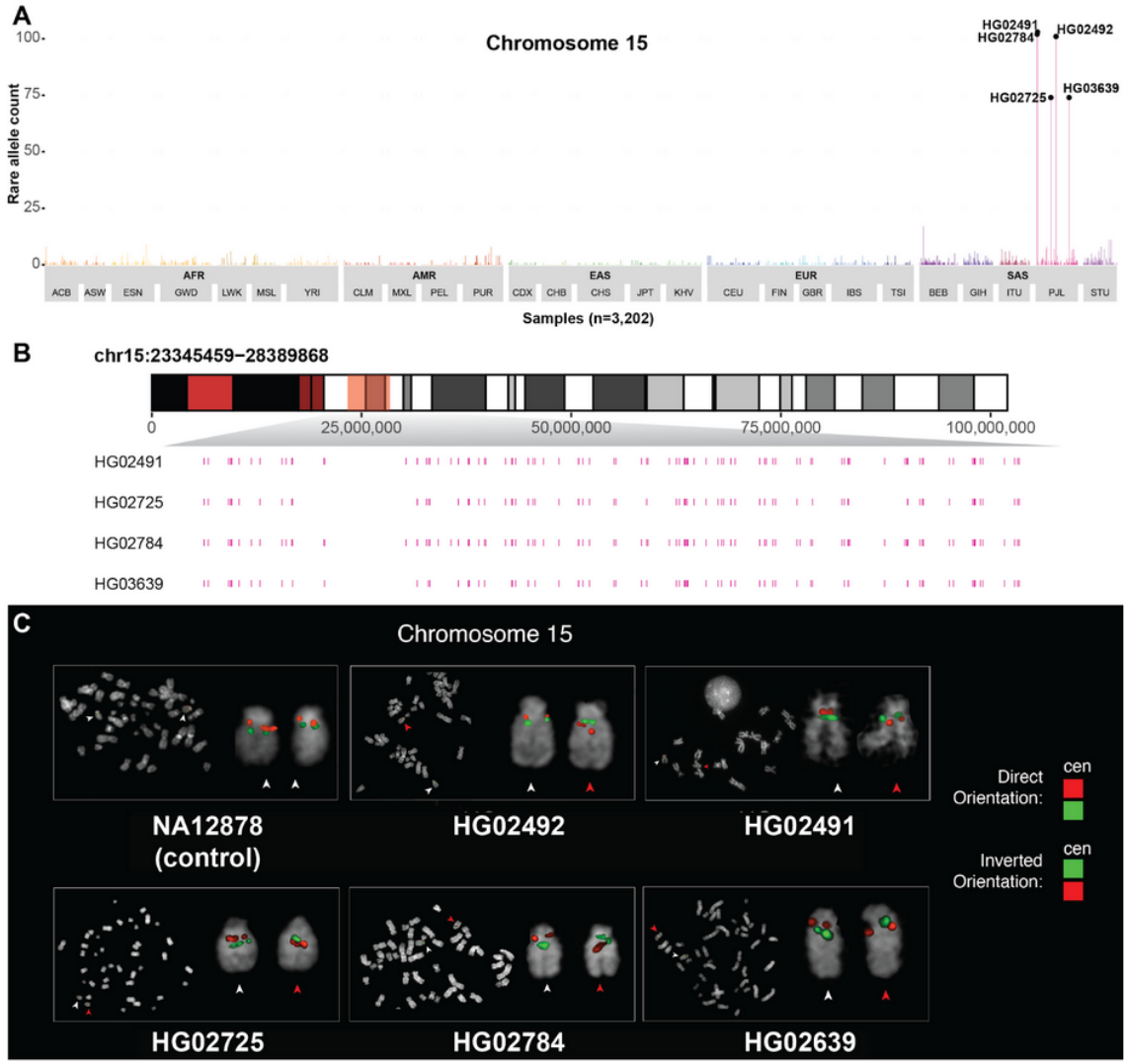


Figure 4.4: Rare inversion on chromosome 15. **A** Barplot showing the number of rare alleles shared between an index individual (HG02492) and all 1000 Genomes samples (n = 3,202) stratified by superpopulation (AFR, AMR, EAS, EUR and SAS) and population (same abbreviations used as in [2]). Individuals with the largest number of shared rare alleles are highlighted by a black dot and a sample specific identifier. **B** Distribution of detected shared rare alleles along the inverted region. The inverted region is highlighted by a transparent red rectangle. Note that rare alleles for the chromosome 15 specific inversion are evenly distributed along the whole inversion for all predicted inversion carriers. **C** FISH validation. Both chromosome 15 homologs have a direct orientation in the control individual (NA12878) while HG02491, HG03639, HG02725 and HG02784 are all carriers of the inversion in heterozygous state. White arrowheads indicate chromosomes in direct orientation while red arrowheads indicate chromosomes with the inversion (ABC8-41788900G7 in red mapping at chr15:23751929-23796236; RP11-640H21 in green mapping at chr15:27894428-28091240). Figure taken from [146].

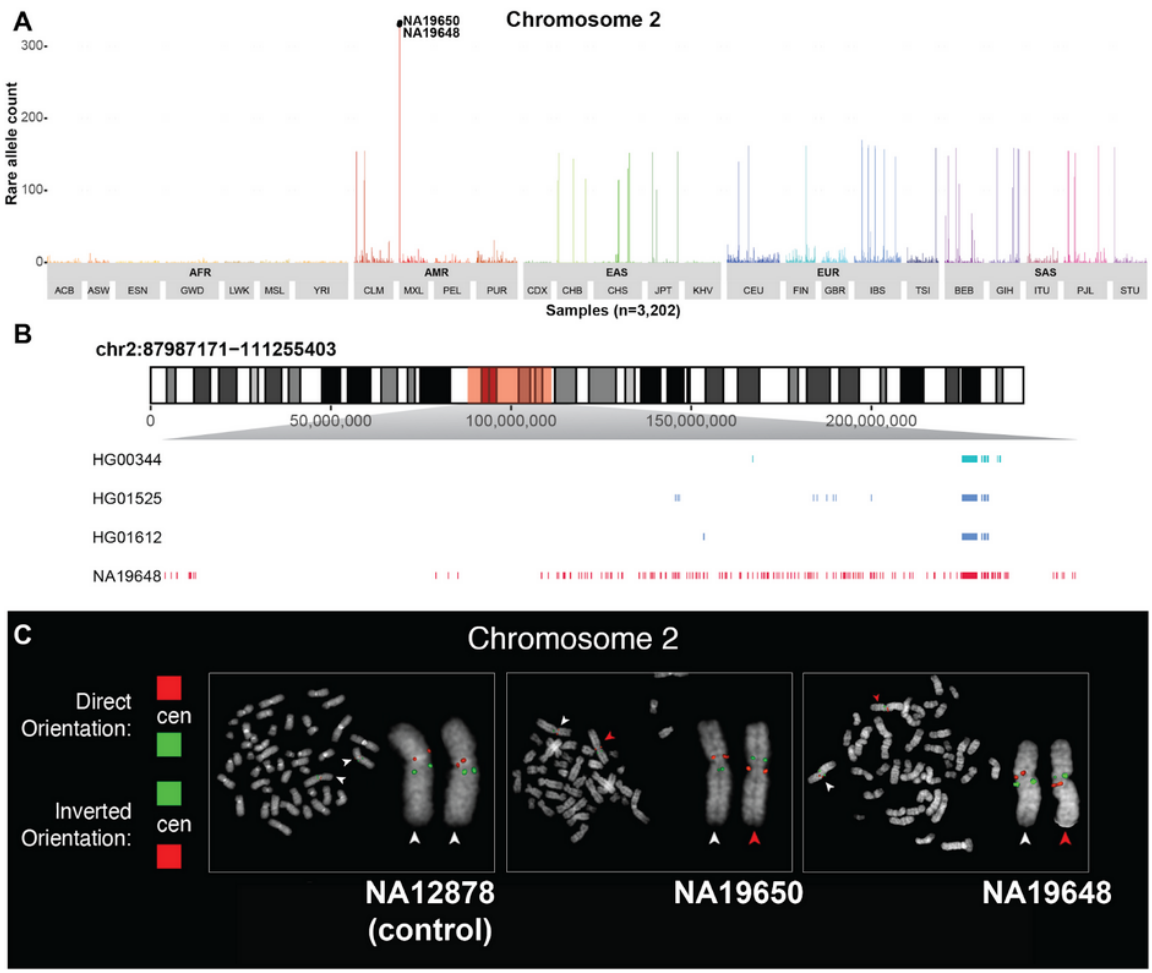


Figure 4.5: Rare inversion on chromosome 2. **A** The barplot shows the number of rare alleles shared between NA19650 and all 1000 Genomes samples (n = 3,202) stratified by superpopulation (AFR, AMR, EAS, EUR, SAS) and population (same abbreviations used as in [2]). Individuals with the largest number of shared rare alleles are highlighted by a black dot. **B** A distribution of detected shared rare alleles along the inverted region on chromosome 2. The inverted region is highlighted by a transparent red rectangle. Rare alleles for the chromosome 2 specific inversion are evenly distributed along the whole inverted region only in a single sample (NA19648). **C** FISH results of a ~ 23.2 Mb inversion on chromosome 2 (chr2:88064758-111283969) are shown. Both chromosome 2 homologs have a direct orientation in the control individual (NA12878) while NA19650 and NA19648 individuals are inverted in heterozygous state. White arrowheads indicate chromosomes in direct orientation while red arrowheads indicate chromosomes with the inversion (ABC8-2121940H19 in red mapping at chr2:88223569-88269173; WI2-1849B17 in green mapping at chr2:110712025-110745244). Figure taken from [146].

4.3 HPRC project

This section presents the work of the Human Pangenome Reference Consortium (HPRC). As a part of this project, I used the pangenome graphs produced from haplotype-resolved assemblies of 44 samples to genotype variants across 3,202 samples using PanGenie. This work is currently under revision and available as a preprint [113]. Sections 4.3.2, 4.3.3 and 4.3.4 re-use material from this preprint. See Section E.7 for information on author contributions and publication details.

4.3.1 Introduction

The Human Pangenome Reference Consortium (HPRC) [90, 113] generated haplotype-resolved assemblies for 47 human samples based on PacBio CCS data and parental Illumina short reads using Trio-Hifiasm [27]. The goal was to construct a pangenome graph capturing the genetic diversity across all 44 unrelated samples (excluding three children) that can replace the linear reference genome. Using a graph instead of a linear reference helps overcoming reference biases arising from missing alternative sequences [187].

The main difference between the pangenome references generated by the HGSVC and HPRC is the way they were constructed and represented. In the HGSVC project, assemblies were individually aligned to the linear reference genome in order to identify variants, and variant calls were later merged across samples to obtain a joint, reference-based callset across all samples that can be interpreted as a pangenome structure. In the HPRC project, a sequence graph is constructed based on a multiple sequence alignment of the assemblies. Unlike the reference-based representation used by the HGSVC, the HPRC graph allows representing variation nested in sequences absent from the reference. The linear reference is only used in order to assign reference coordinates to the variants present in the assembly which allows representing variants in VCF format. Variants are detected from the pangenome by studying the bubble structures in the graph and by comparing traversals of haplotypes and the reference genome through bubble regions.

As a part of the HPRC project, I genotyped the variation obtained from the newly constructed pangenome graph across all 3,202 samples from the 1000 Genomes project [2, 21] using PanGenie, demonstrating that this new reference structure allows genotyping structural variants not typable based on the HGSVC data.

4.3.2 Pangenome construction and variant calling

This section presents a summary on the pangenome construction methods presented in [113] developed by co-authors and re-uses some material from this preprint.

Several versions of the pangenome were constructed based on different construction methods. The version underlying the genotyping experiments described in this chapter is based on Minigraph and the Cactus genome aligner [9, 76, 110]. First, Minigraph [110] was used in order to iteratively construct a graph starting from the GRCh38 reference genome and

progressively adding 88 haplotype-resolved assemblies constructed from 44 human samples. After this step, the graph contained only structural variants $\geq$ 50 bp. In order to include SNPs and small variants, it was extended with a base-level alignment of homology relationships between assemblies by using Cactus [9]. Non-reference sequences of at least 100 kbp that were either identified as being satellite, which could not be assigned to a reference chromosome or which appeared unaligned to the remaining assemblies, were removed from the graph. The result is a sequence graph in which nodes correspond to DNA segments. The graph represents a multiple sequence alignment of the 88 haplotypes and each haplotype is represented as a path through this graph.

Bubbles in the graph were detected using the tool `vg deconstruct` [142]. It decomposes the pangenome graph into sets of nested subgraphs, so-called snarls, each of which corresponds to a collection of genetic variants. Briefly, this is done by converting the graph into a biedged graph, in which nodes are represented as "black edges" and edges as "grey edges" [142]. Pairs of black edges are determined that need to be cut in order to disconnect the graph such as to create separated components [142]. These snarls can be nested. Nesting relationships can be represented by arranging the snarls in a tree and annotating them based on their level (LV) in this tree. Top-level snarls are annotated by LV = 0, and snarls contained in others are annotated by LV > 0. Snarls are output in VCF format. Figure 4.6a (left) provides an example of a bubble region in the tree. On the right, the corresponding biedged graph is shown. The top-level snarl represents the whole bubble and can be generated by cutting edges a and e. The two nested variants are represented in terms of two LV = 1 snarls, resulting from cutting edges b and c, and c and d, respectively. The snarls are represented in terms of the three VCF records shown below, listing their respective allele sequences in the graph.

4.3.3 Genotyping SVs across a cohort of 3,202 individuals

This section and all subsections re-use material presented in [113].

We used the VCF file created based on the snarl traversal of the Minigraph-Cactus (MC) graph as a basis for genotyping. We used vcfbub (`https://github.com/pangenome/vcfbub`, version 0.1.0) with parameters `-l 0` and `-r 100000` in order to filter the VCF. It removes all non-top-level bubbles from the VCF (LV > 0) unless they are nested inside a top-level bubble with a reference length exceeding 100 kbp, i.e. top-level bubbles longer than that are replaced by their child nodes in the snarl tree. The VCF also contained the haplotypes for all 44 assembly samples, representing paths in the pangenome graph. We additionally removed all records for which more than 20% of all 88 haplotypes carried a missing allele ("."), which can happen in case of gaps in the assemblies. This resulted in a set of 22,133,782 bubbles. In a next step, we used PanGenie (version 1.0.0) to genotype these bubbles across all 3,202 samples from the 1000 Genomes project based on high coverage Illumina reads [21]. We observed a per-sample runtime of 53 single-core CPU hours, as well as a memory

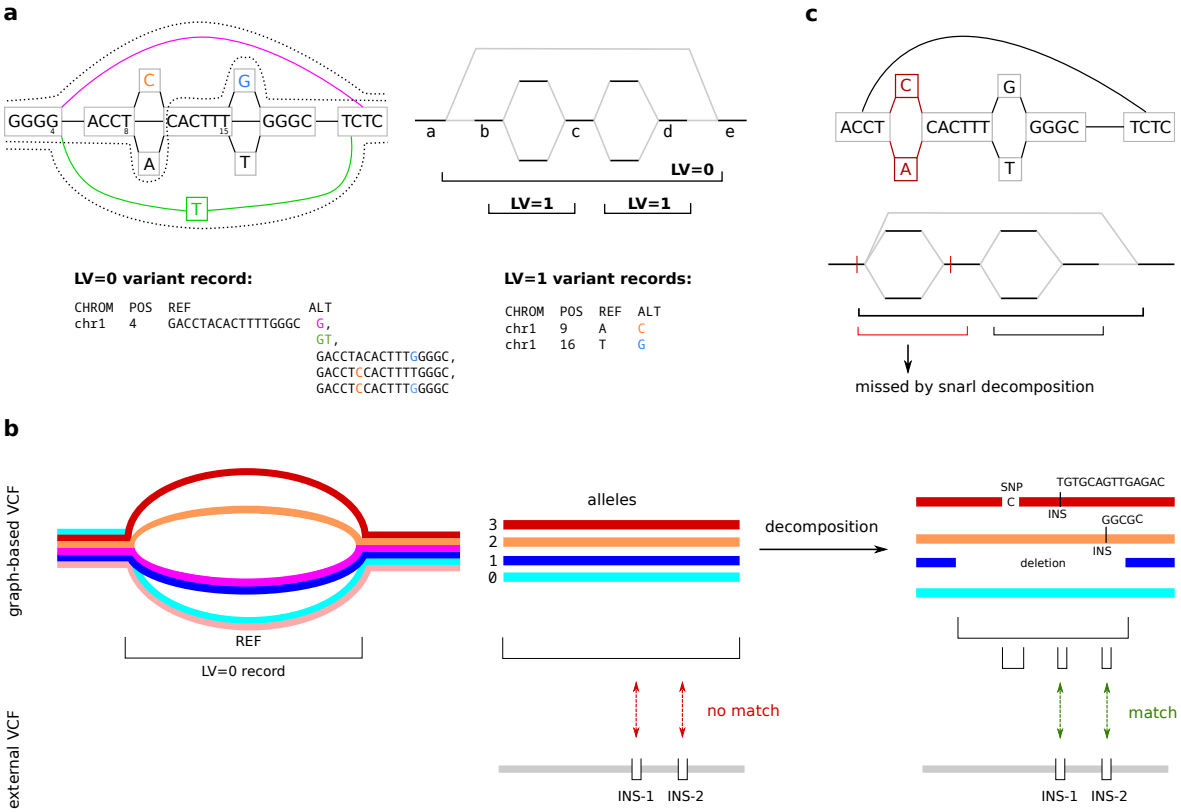


Figure 4.6: Decomposition of graph bubbles. **a** On the left, a bubble in the pangenome graph is shown which contains nested variation. Dotted lines represent haplotype paths in the graph. On the right, the corresponding biedged representation of the graph is shown. Snarls are represented in VCF format as shown below. **b** Shown is a multiallelic bubble contained in the snarl-based VCF (LV = 0 record). Using the coordinates of the whole bubble when comparing to external callsets leads to errors, since the insertions carried by the second and third haplotypes are not detected. The decomposition aims at identifying which individual variant alleles each haplotype carries inside of the bubble and enables proper comparison to external callsets. **c** Example for which the snarl decomposition fails to detect nested variation. The leftmost nested SNP is not represented as a separate snarl, since it shares its leftmost black edge with the top-level snarl. Panel b) taken from [113].

usage of 153 GB.

Decomposition of variants

Genotyping results in genotypes for all top-level bubbles across all 1000 Genomes samples. While biallelic bubbles can be easily classified representing SNPs, indels or SVs, this becomes more difficult for multiallelic bubbles contained in the VCF. Especially larger multiallelic bubbles can contain a high number of nested variant alleles overlapping across haplotypes, represented as a single bubble in the graph. Since we considered only top-level bubbles for genotyping, any information on nested variants is no longer present in the VCF representation of our genotyped bubbles. This is especially problematic when comparing the genotypes computed for the whole bubble to external callsets, as coordinates of the top-level bubble

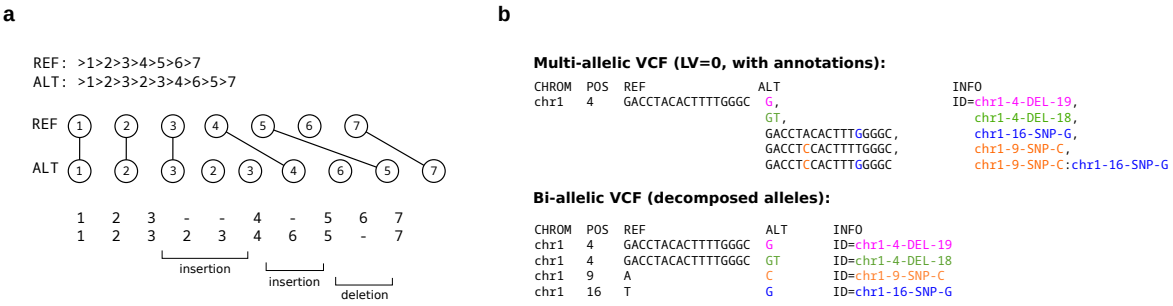
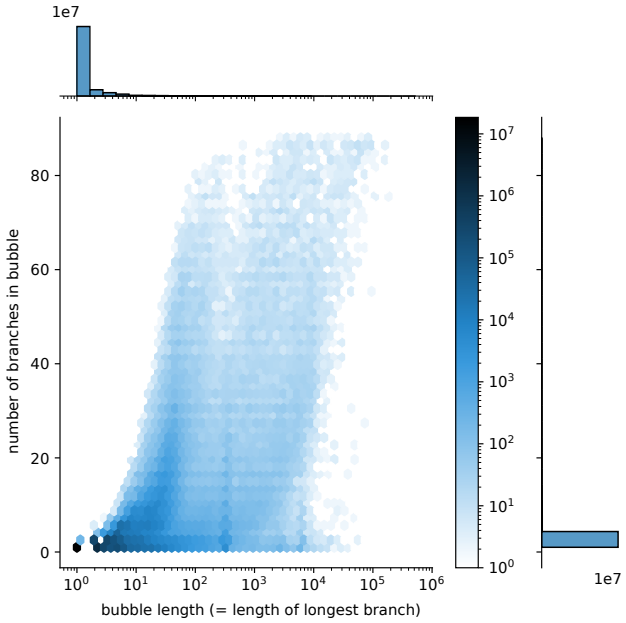


Figure 4.7: Traversal-based decomposition. **a** The idea of the decomposition approach is to compare the sequence of visited node IDs (node traversal) of each alternative allele (that is covered by at least one of the haplotypes present in the graph) to the traversal of the reference allele. Each node in the reference traversal is matched to its leftmost occurrence in the alternative traversal (if existent), resulting in an alignment of the traversals. The nested alleles can then be determined from the insertions, deletions and mismatches in the alignment. In this example, the alternative allele can be decomposed in two insertions and one deletion. **b** Two VCF files are produced. The multiallelic VCF contains the same records as the input VCF, just with annotations for all alternative alleles added to the INFO field. Each ALT allele is annotated by a sequence of IDs encoding the nested alleles, separated by ":". The second VCF is a biallelic one, containing a separate record for each nested variant ID, i.e. it contains all alleles after decomposition. Figure taken from [113].

do not necessarily represent the exact coordinates of individual variant alleles carried by a sample in this region (Figure 4.6b).

The information on nested variants initially provided by the snarl tree could help to tackle this problem. However, the snarl-based decomposition of bubbles sometimes misses nested variants in cases where a nested bubble shares a black edge with a higher level snarl. See Figure 4.6c for an example: the leftmost nested bubble will not be detected since cutting at the positions marked in red will not disconnect the corresponding subgraph from the rest of the graph. We have therefore introduced an alternative decomposition approach which aims at detecting all variant alleles nested inside of multiallelic top-level records. The idea is to detect variants from the sequences of visited node IDs (node traversals) of the reference and alternative alleles of all top-level bubbles. Given the node traversals of a reference and alternative path through a bubble, our approach is to match each reference node to its leftmost occurrence in the alternative traversal, resulting in an alignment of the node traversals (Figure 4.7a). Nested alleles can then be determined based on insertions, deletions and mismatches in this alignment. Since the node traversals of the alternative alleles can visit the same node more than once (which is not the case for the reference alleles of the MC graph), this approach is not guaranteed to reconstruct the optimal sequence alignment underlying the nodes in these repeated regions. As an output, the decomposition process generates two VCF files. The first one is a multiallelic VCF which contains exactly the same variant records as the input VCF, just that annotations for all alternative alleles of a record were added to the ID tag in the INFO field. For each alternative allele, the ID tag contains IDs encoding all nested variants it is composed of, separated by a colon. The second VCF

a



b

allele type	# alleles in biallelic regions	# alleles in multiallelic regions
SNPs	18,392,518	1,801,599
indels	2,250,931	4,597,184
SV INS	12,192	242,420
SV DEL	4,232	52,969
SV Other	1,296	100,700

Figure 4.8: HPRC allele statistics. **a** Each of the 88 haplotypes contained in the graph defines a path through each bubble. The plot shows the number of different paths covered by the haplotypes in a bubble as a function of the length of the bubble. Here, the length of a bubble is defined by the sequence of the longest such path. **b** Number of variant alleles located inside of biallelic and multiallelic regions of the graph. Biallelic regions include all bubbles with only two alternative paths, multiallelic regions include all bubbles in which haplotypes cover more than two alternative paths through the bubble. Figure taken from [113].

is biallelic and contains a separate record for each nested variant ID defining reference and alternative allele of the respective variant (Figure 4.7b). Both VCFs are different representations of the same genomic variation, i.e. before and after decomposition. We applied this decomposition method to the MC-based VCF file, used the multiallelic output VCF as input for PanGenie to genotype bubbles, and used the biallelic VCF as well as the IDs in order to translate PanGenie's genotypes for top-level bubbles to genotypes for all individual nested variant alleles. All downstream analyses of the genotypes are based on this biallelic representation (i.e. after decomposition). While the majority of short bubbles (< 10 bp) are biallelic, especially large bubbles (> 1000 bp) tend to be multiallelic. Sometimes each of the 88 haplotypes contained in the graph covers a different path through such a bubble (Figure 4.8a), leading to a VCF record with 88 alternative alleles listed. We determined the number of variant alleles located inside of biallelic and multiallelic bubbles in the pangenome after decomposition. As expected, the majority of SV alleles is located inside of the more complex, multiallelic regions of the pangenome (Figure 4.8b).

Genotyping evaluation based on assembly samples

Similarly to what we presented in [49] (Section 3.5), we performed a "leave-one-out" experiment in order to evaluate PanGenie's genotyping performance for the callset samples. For this purpose, we repeatedly removed one of the panel samples from the MC VCF and genotyped it using only the remaining samples as an input panel for PanGenie. We later used the genotypes of the left-out sample as ground truth for evaluation. We repeated this experiment for five of the callset samples (HG00438, HG00733, HG02717, NA20129 and HG03453) using 1000 Genomes high coverage Illumina reads [21]. PanGenie is a re-genotyping method. Therefore, like any other re-typer, it can only genotype variants contained in the input panel VCF, that is, it is unable to detect variants unique to the genotyped sample. For this reason, we excluded truth set variants (after decomposition) that were only contained in the left-out sample for evaluation. We used the weighted genotype concordance [49] (Section 3.5.1) as a metric to evaluate the genotyping performance. Figure 4.9 shows the results stratified by different regions. Figure 4.9A shows concordances in biallelic and multiallelic regions of the MC VCF. The biallelic regions include only bubbles with two branches. The multiallelic regions include all bubbles in which haplotypes cover more than two different paths. Figure 4.9B shows the same results stratified by genomic regions defined by GIAB [137] (obtained from locations specified in Section C.2):

- **low-mappability:** difficult to map regions as well as segmental duplications
- **repeats:** short tandem repeats and repeat annotations from UCSC Genome Browser [93]
- **other-difficult:** union of regions that are difficult to access, including MHC and KIR regions or regions poorly assembled in the reference genome
- **easy:** regions outside of other difficult regions such as tandem repeats, homopolymers, difficult to map regions, segmental duplications and high/low GC content

Here and in the following, we consider results for SNPs, indels (1-49 bp), SV deletions, SV insertions and "other" SV alleles, defined as follows: SV deletions include all alleles for which length(REF) $\geq$ 50 bp and length(ALT) = 1, SV insertions include all alleles for which length(REF) = 1 and length(ALT) $\geq$ 50 bp. All other alleles with a length $\geq$ 50 bp are included in "others".

Overall, weighted genotype concordances were high for all variant types. Especially variant alleles in biallelic regions of the graph were very well genotypable. Alleles inside of multiallelic bubbles were more difficult to genotype correctly since PanGenie needed to decide between several possible alternative paths, while there were only two such paths for biallelic regions (Figure 4.9a). Furthermore, genotyping accuracy depended on the genomic context (Figure 4.9b). Regions with low mappability, repetitive regions and other difficult regions were harder to genotype than regions classified as "easy" by GIAB.

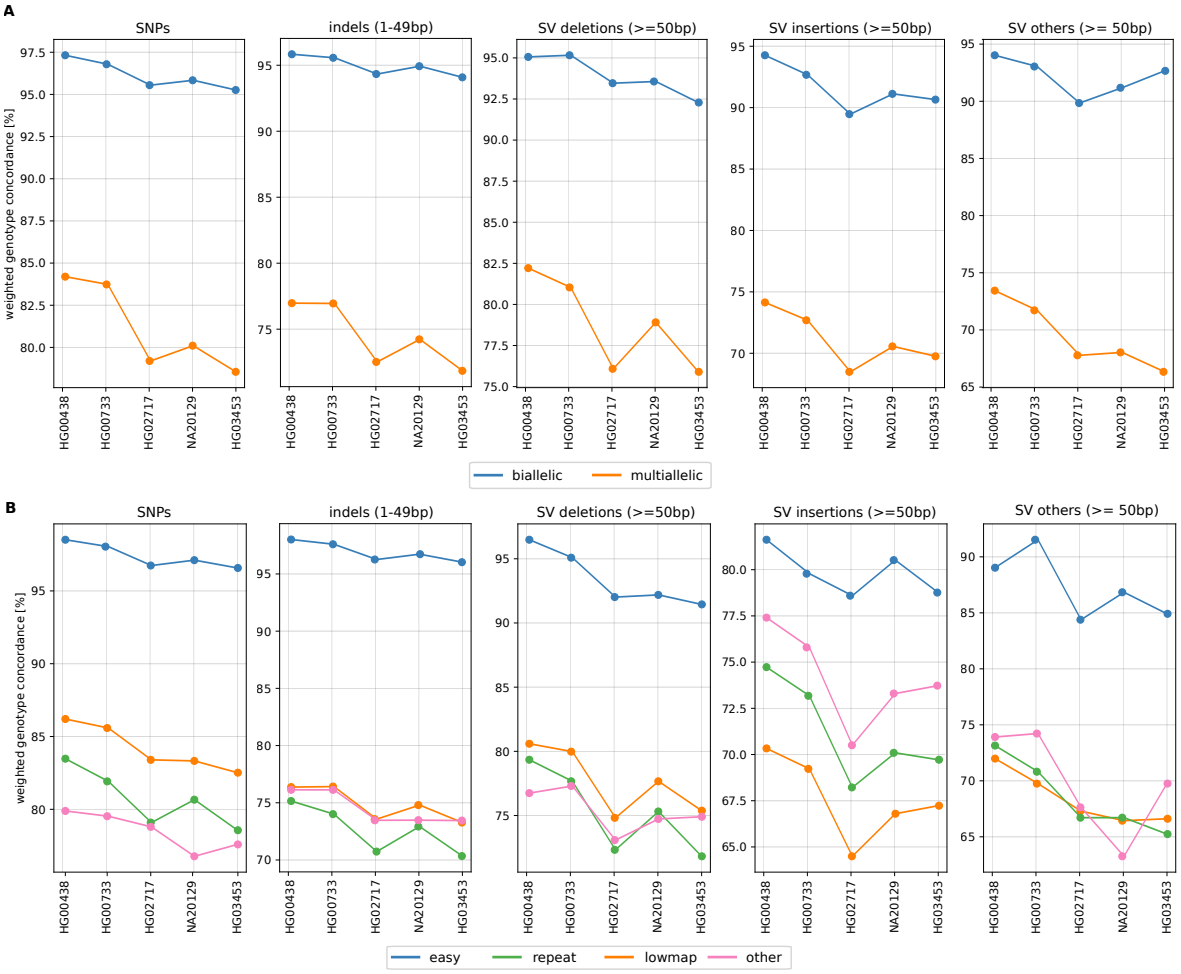


Figure 4.9: HPRC leave-one-out experiment. A leave-one-out experiment was conducted by repeatedly removing one of the assembly-samples from the panel VCF and genotyping it based on the remaining samples. Plots show the resulting weighted genotype concordances for different variant allele classes. **a** Weighted genotype concordances are stratified by graph complexity: biallelic regions of the MC graph include only bubbles with two branches, and multiallelic regions include all bubbles with > 2 different alternative paths defined by the 88 haplotypes. **b** Results of the same experiment stratified by different genomic regions defined by GIAB. Figure taken from [113].

Creating a high quality subset

We generated genotypes for all 3,202 samples from the 1000 Genomes Project with PanGenie and defined a high quality subset of SV alleles that we can reliably genotype. For this purpose, we applied a machine learning approach similar to what we have presented previously [46, 49] (Sections 3.7 and 4.1.4). We used five filters in order to define positive and negative subsets of variants:

- **ac0_fail**: this filter fails if a variant was genotyped as absent (0/0 or ./.) across the cohort samples (AF $= 0.0$).
- **mendel_fail**: this filter fails if the Mendelian consistency for a variant across trios was

variant type	# alleles in unfiltered set	# alleles in positive set	# alleles in final set
SNPs	20,194,117	15,069,514	
indels	6,848,115	2,399,842	
SV INS	254,612	32,431	84,755
SV DEL	57,201	13,356	28,433
SV Other	101,996	8,334	26,489

Table 4.3: Number of variant alleles in HPRC callsets. Shown are the numbers of variant alleles for SNPs, indels and SVs in the unfiltered set, the positive set and the final, filtered set. Indels include all variant alleles between 1-49 bp in length, SVs include all variant alleles $\geq$ 50 bp. SV insertions/deletions contain all clean insertions/deletions, all other SVs are defined as "others". Table taken from [113].

less than 80%. As before [46, 49] (Sections 3.7 and 4.1.4), we used a strict definition of Mendelian consistency which excluded all trios with only 0/0, only 0/1 and only 1/1 genotypes.

- **gq_fail**: a variant fails this filter if the number of genotypes with quality $>$ 200 was less than 50.
- **self_fail**: a variant fails this filter if the genotyping accuracy of a variant allele across the panel samples was less than 90%.
- **nonref_fail**: a variant fails this filter if not a single non-0/0 genotype was genotyped correctly across all panel samples.

The positive set included all variant alleles that passed all five filters. The negative set contained all variant alleles that passed the ac0_fail filter but failed at least three of the other filters. We show the number of SV alleles failing the different filters in Figures C.8, C.9 and C.10 (top panels). We trained a support vector regression (SVR) approach [49] (Section 3.7) based on 33 features including allele frequencies, Mendelian consistencies and the number of alternative alleles transmitted from parents to children. We applied this method to all remaining variant alleles genotyped with an AF $>$ 0, resulting in a score between -1 (bad) and 1 (good) for each. We finally defined a filtered set of variants which included the positive set, as well as all variant alleles with a score of $\geq$ -0.5.

We show the number of variant alleles contained in the unfiltered set, the positive set as well as the filtered set in Table 4.3. Since our focus was on SVs and since 65% of all SNPs and indels were already contained in the positive set, we applied our machine learning approach only to SVs. We found that 50%, 33% and 26% of all deletion, insertion and "other" alleles, respectively, were contained in the final, filtered set of variants. Note that these numbers take all distinct SV alleles contained in the callsets into account. Especially for insertions and "other" SVs, many of these alleles were very similar, with sometimes only a single base pair differing. Therefore, it is likely that many of these actually represent the same events. Our genotyping and filtering approach helps to remove such redundant alleles.

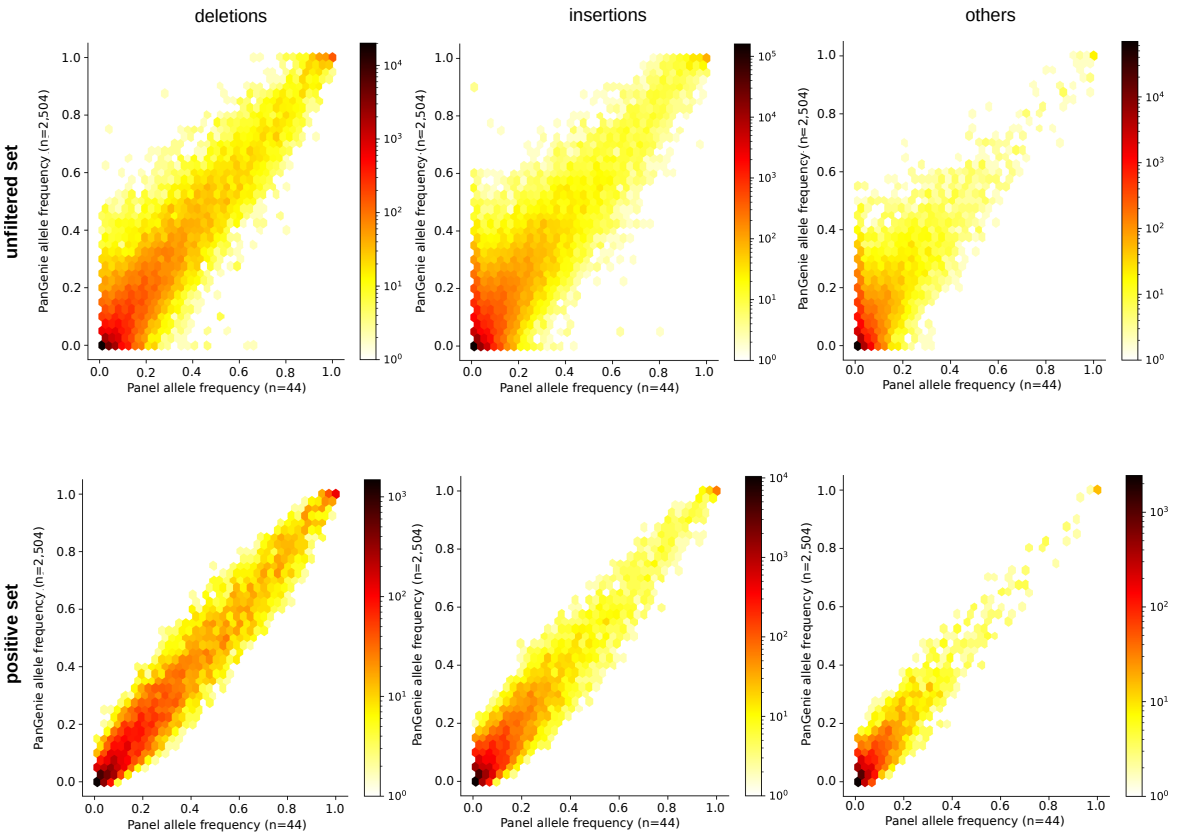


Figure 4.10: HPRC unfiltered and positive sets. Comparison of allele frequencies observed from the PanGenie genotypes for all 2,504 unrelated 1000 Genomes samples and the allele frequencies observed across the 44 assembly samples in the MC graph. Results are shown for the unfiltered (57,201 deletions, 254,612 insertions, 101,996 "other" alleles) and positive sets (13,356 deletions, 32,431 insertions, 8,334 "other" alleles). Figures taken from [113].

In order to evaluate the quality of the PanGenie genotypes, we compared the allele frequencies observed for the SV alleles across all 2,504 unrelated 1000 Genomes samples to their allele frequencies observed across the 44 assembly samples in the MC callset. We observed that the allele frequencies between both sets matched well, resulting in correlations (Pearson) of 0.93, 0.87 and 0.81 for deletions, insertions and "other" alleles contained in the unfiltered set (Figure 4.10, top). For the positive set, we observed allele frequency correlations of 0.97, 0.96 and 0.95, respectively (Figure 4.10, bottom). For our final, filtered set, these correlations were 0.96, 0.93 and 0.90, for deletions, insertions and "other" alleles, respectively (Figure 4.11a), indicating high-quality genotypes. We also analyzed the heterozygosity of the PanGenie genotypes across all 2,504 unrelated 1000 Genomes samples contained in the filtered set and found that the results are consistent with the Hardy-Weinberg equilibrium (Figures C.8, C.9 and C.10, lower panel).

A direct comparison of HGSVC and HPRC genotyping results is difficult for several rea-

sons. For HGSVC, variant calling included a merging step in which similar alleles were combined into a single variant record, while in the HPRC set, similar alleles were always kept separate. Differences in variant representation are also reflected in the different SV categories: while the HGSVC callset represents variants in terms of insertions and deletions, the HPRC set sometimes represents variants in terms of "other" variant alleles, that could neither be classified as insertions nor deletions. While we used a similar machine learning approach to filter our sets, our definition of filters and the set of features used was not exactly identical in order to account for these differences. We therefore compared both sets based on the number of SVs per sample in the filtered sets.

Number of SVs per sample

To quantify our ability to detect additional SVs, we compared our filtered set of genotypes to the HGSVC PanGenie genotypes (v2.0 "lenient" set [46], Section 4.1.4) and Illumina-based 1kGP SV genotypes [21]. A direct comparison of the three callsets was difficult. The HGSVC and HPRC callsets are based on variant calls produced from haplotype-resolved assemblies of 32 and 44 samples, respectively [46] (Section 4.1). For each callset, variants were re-genotyped across all 3,202 samples from the 1000 Genomes Project. Note that the callset samples for HPRC and HGSVC are disjoint. Since re-genotyping cannot discover novel variants, both callsets will miss variants carried by 3,202 samples that were not seen in the assembly samples. In contrast, the 1kGP callset contains short-read based variant calls produced for each of the 3,202 samples from the 1000 Genomes Project. As mentioned before, another difference between the HGSVC and HPRC callsets is that in the HGSVC callset, highly similar alleles were merged into a single record to correct for representation differences across different samples or haplotypes. The HPRC callset however, kept all these alleles separately even if there was only a single basepair difference between them. To make the callsets better comparable, we merged clusters of highly similar alleles in the HPRC filtered set prior to comparisons with other callsets. This was done with truvari ([56], version v3.1.0) using the command: `truvari collapse -r 500 -p 0.95 -P 0.95 -s 50 -S 100000`. In order to be able to properly compare the callsets despite their differences, we counted the number of SV alleles present in each sample (heterozygous or homozygous) in each callset and plotted the corresponding distributions stratified by genome annotations from GIAB (same as above, Figure 4.11b) as well as using repeat annotations directly derived from the graph (Figure C.11). We also generated the same plot including only common SV alleles with an AF $>$ 5% across all 3,202 samples (Figure C.12). These plots show that the HPRC and HGSVC callsets were able to access more structural variants (HPRC: 18,483 SVs/sample, HGSVC: 12,997 SVs/sample) across the genome than the short-read-based 1kGP callset (9,596 SVs/sample), especially deletions $<$ 300 bp and insertions (Figure 4.11c). This confirms that short-read based SV discovery relative to a linear reference genome misses a large portion of SVs located in regions inaccessible by short-read alignments, which has been reported previously by several studies [24, 46, 206].

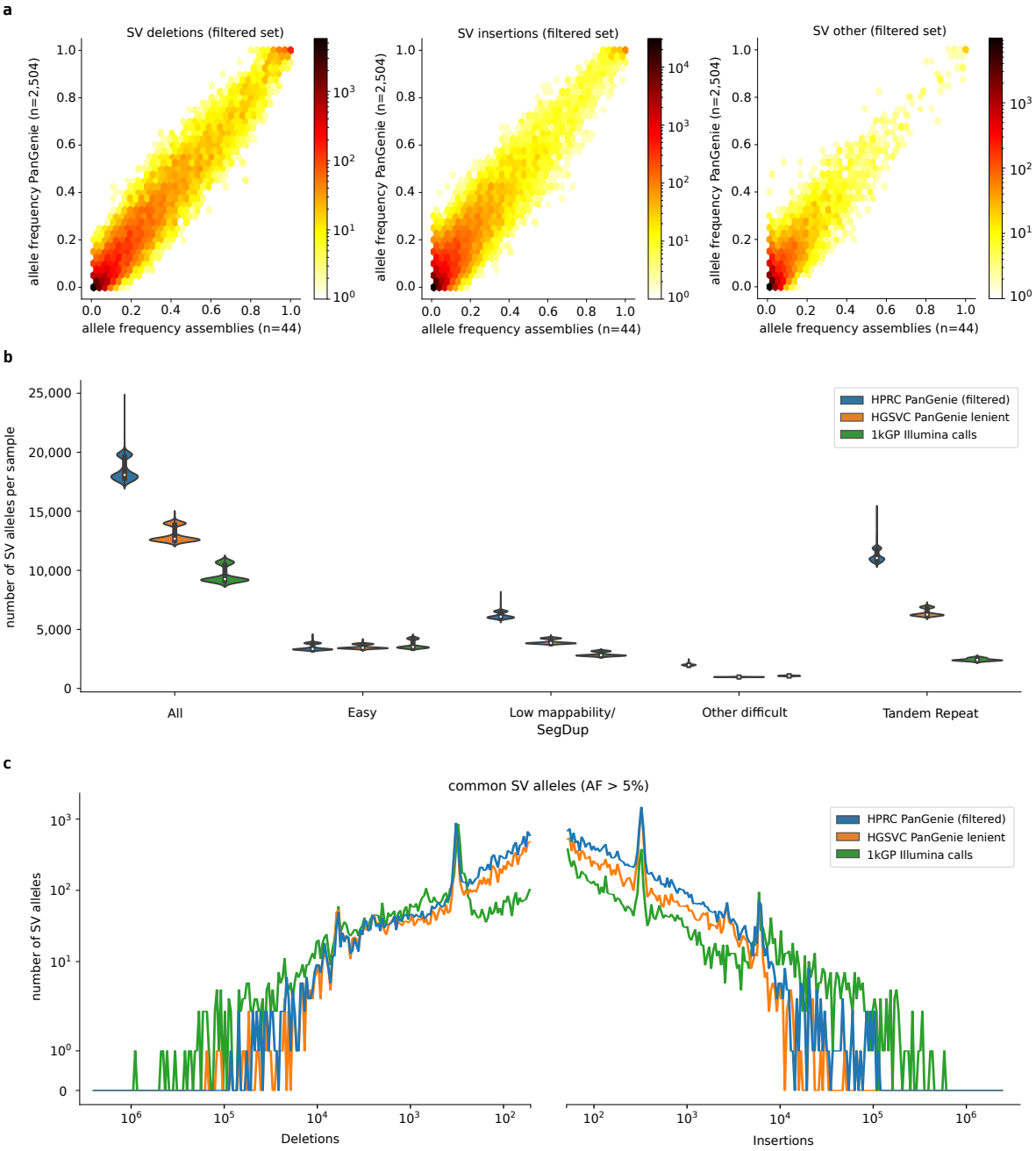


Figure 4.11: HPRC filtered set. a Comparison of allele frequencies observed from the PanGenie genotypes for all 2,504 unrelated 1000 Genomes samples and the allele frequencies observed across the 44 assembly samples in the MC graph. The PanGenie genotypes include all variants contained in the filtered set (28,433 deletions, 84,755 insertions, 32,431 "other" alleles). **b** Shown are the number of SVs present (genotype 0/1 or 1/1) in each of the 3,202 samples from the 1000 Genomes Project in the filtered HPRC genotypes (PanGenie), the HGSVC lenient set and the 1kGP Illumina calls in GIAB regions. **c** Shown is the length distribution of SV insertions and SV deletions contained in the filtered HPRC genotypes (PanGenie), the HGSVC lenient set and the 1kGP Illumina calls. Only variants with an allele frequency $> 5\%$ across the 3,202 samples are considered. Figure taken from [113].

Expectedly, the number of SVs per sample within "easy" genomic regions was consistent

across all three callsets, while especially in low mappability and tandem repeat regions, the use of our pangenome reference led to pronounced gains (Figure 4.11b), including for common variants (Figure 4.11c, Figure C.11).

In order to evaluate the novel SVs in our filtered HPRC callset, we re-visited the leave-one-out experiment we had performed previously on the unfiltered set of variants (see above). We restricted the evaluation to the subset of variants that are (a) in our filtered set but not in the 1kGP Illumina calls ("novel"), (b) in our filtered set as well as in the 1kGP Illumina callset ("known"), and to (c) all variants in our filtered set. In order to find matches between our set and the Illumina calls, we used a criterion based on reciprocal overlap of at least 50%. Results are shown in Figure 4.12. We generated two versions of this figure: the first one (top panel) ignores variants that were only contained in the left-out sample and thus not typable by any re-genotyping method, and the second one includes these variants (bottom panel). In general, genotype concordances of all filtered variants (brown, dark purple) were slightly higher compared to the concordances we observed for the unfiltered set (Figure 4.12). Furthermore, concordances of the known variants were highest. This is expected, since these variants tend to be in regions easier to access by short reads. Concordances for novel variants were slightly worse. This is also expected, since these variants tend to be located in more complex genomic regions that are generally harder to access. However, even for these variants, concordances were still high, indicating that the PanGenie genotypes for these variants are indeed of high quality.

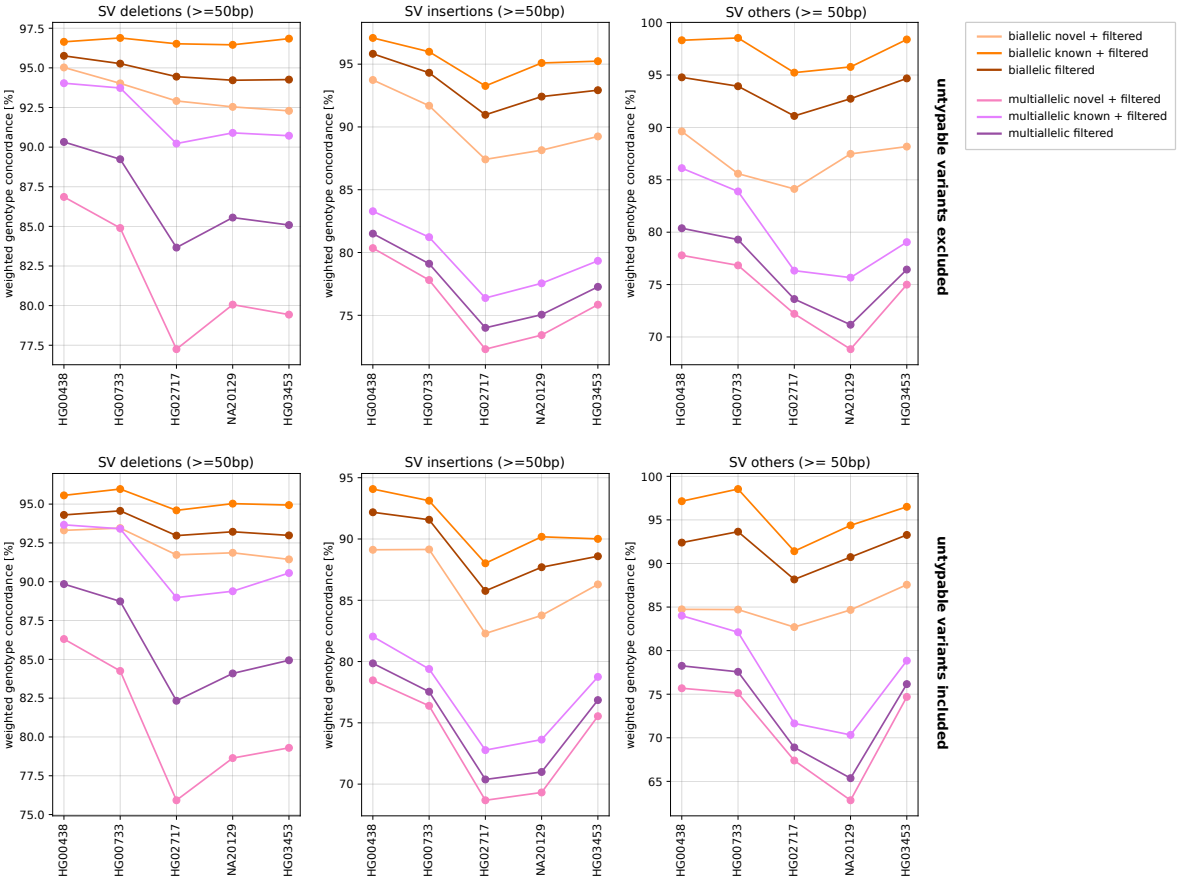


Figure 4.12: HPRC novel and known variants. A leave-one-out experiment was conducted by repeatedly removing one of the assembly samples from the panel VCF and genotyping it based on the remaining samples. Plots show the resulting weighted genotype concordances for variants in our filtered PanGenie set. The novel variants include only SVs not contained in the 1kGP Illumina set, the known variants include only variants contained in these Illumina calls. Weighted genotype concordances are stratified by graph complexity: biallelic regions of the MC graph include only bubbles with two branches, and multiallelic regions include all bubbles with > 2 different alternative paths defined by the 88 haplotypes. The top panel excludes variants that are unique to the left-out sample and thus not typable by any re-genotyping method while the bottom panel includes them. Figure taken from [113].

Evaluation based on medically relevant SVs

In addition to all 1000 Genomes samples we also genotyped sample HG002/NA24385 based on Illumina reads from [207]. We used the GIAB CMRG benchmark containing medically relevant SVs [192], downloaded from: `https://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/giab/release/AshkenazimTrio/HG002_NA24385_son/CMRG_v1.00/GRCh38/StructuralVariant/`) for evaluation. Like for the 1000 Genomes samples, we used the MC-based VCF (see above) containing variant bubbles and haplotypes of 44 assembly samples as an input panel for PanGenie. We extracted all variant alleles with a length ≥ 50 bp from our genotyped VCF (biallelic version, after decomposition). We con-

unadjusted			adjusted (without untypables)		
precision	recall	F-score	precision	recall	F-score
0.74	0.75	0.75	0.74	0.81	0.78

Table 4.4: Medically relevant SV benchmarking. Shown are results of comparing PanGenie genotypes computed for HG002 to the GIAB CMRG benchmark. The left part of the table shows results for all SV alleles, the right part excludes SVs from the truth set that are unique to HG002 and thus untypable because they are absent from the input panel used for PanGenie. Table taken from [113].

verted the ground truth VCF into a biallelic representation using `bcftools norm -m -any` and kept all alleles with length $\geq$ 50 bp. We used truvari ([56], version v3.1.0) with parameters `--multimatch --includebed <medically-relevant-sv-bed> -r 2000 --no-ref a -C 2000 --passonly` in order to compare our genotype predictions to the medically relevant SVs. Results are shown in Table 4.4 (left). Since PanGenie is a re-typing method, it can only genotype variants provided in the input and thus cannot detect novel alleles. Since HG002 is not among the panel samples, the input VCF misses variants unique to HG002. Thus, these unique variants cannot be genotyped by PanGenie and were counted as false negatives during evaluation. Therefore, we computed an "adjusted" version of the recall which excludes SV alleles unique to HG002 (i.e. alleles not in the graph) from the truth set for evaluation. In order to identify which SV alleles were unique, we compared each of the 44 panel samples to the ground truth VCF using truvari in order to identify the false negatives for each sample. Then we computed the intersection of false negative calls across all samples. The resulting set then contains all variant alleles unique to the HG002 ground truth set. We found 15 such unique SV alleles among the GIAB CMRG variants. We removed these alleles from the ground truth set and recomputed precision/recall statistics for our genotypes. Adjusted precision/recall values are shown in Table 4.4 (right).

4.3.4 Discussion

This section re-uses some material presented in [113].

The HPRC generated haplotype-resolved assemblies of 47 human genomes, the largest set of fully phased genome assemblies that is currently available. Compared to earlier efforts, including the HGSVC (Section 4.1), the assemblies are of better quality, reaching an average median base-level accuracy that is an order of magnitude higher than for the HGSVC assemblies and an N50 that is nearly twice as large. A subset of 44 samples (88 haplotypes) was used in order to construct a pangenome graph, representing all genetic variation present in the haplotypes, with the goal to provide a better alternative to the linear reference genome.

We have demonstrated how such a pangenome graph can be used in order to accurately genotype genetic variants, especially structural variants, based on short-reads. The approach presented here differs from the work presented in Section 4.1 in the way the panel used for genotyping was created. For HGSVC, variants were called based on alignments of the

assemblies to the linear reference genome and variant calls generated in this way were later combined into a pangenome graph by merging overlapping variant sites into bubbles. In contrast, the HPRC took a different approach by constructing a pangenome graph in a reference-free manner, based on multiple sequence alignments of the assemblies. Variants were later detected from the bubbles, requiring a decomposition step to identify variation nested inside of bubbles.

PanGenie produced high quality, short-read based genotypes for the 3,202 samples from the 1000 Genomes project based on this pangenome reference, and, as illustrated also in previous experiments (Sections 3, 4.1), enabled fast genotyping of SVs otherwise missed by linear reference based, short-read callers. An evaluation based on medically relevant SVs showed high genotyping accuracies even in difficult-to-access regions.

A direct comparison to the HGSVC callset was difficult since the underlying callset samples were disjoint and also because of the different representation of SVs. While similar SV alleles were merged into a single variant allele in the HGSVC callset, the HPRC callset kept separate records even for highly similar SVs. Therefore, we compared the number of SVs carried by each sample after genotyping in order to correct for these differences, revealing higher numbers for the HPRC callset. These results suggest that the improvements are likely a combination of both increased numbers of individuals in the pangenome and improved genome assemblies that retain better sequence-level resolution of SV haplotypes and thus, the pangenome delivers substantially better SV calling than earlier approaches. This enables the inclusion of tens of thousands of additional SV alleles into further downstream analyses.

However, the experiments also reveal some limitations. The pangenome graph contains many very large, complex bubbles containing nested variants. Genotyping such multiallelic regions is generally more difficult due to the high number of possible alternative alleles. The “leave-one-out” experiment showed that the performance of PanGenie in such regions was worse compared to the more easy, biallelic regions. In addition, there is a need for more sophisticated methods to evaluate genotyping performance in such regions, especially because many of the allelic sequences are often very similar and current methods do not take this factor into account.

Similar to what was observed from the HGSVC experiments (Section 4.1), rare variants were difficult to genotype as they tended to be genotyped as absent across the 1000 Genomes cohort. Due to not merging similar SV alleles into a single record, and the higher number of samples present in the graph (44 instead of 32), this problem is even more prominent here, since the proportion of rare alleles is larger. About 48% of all SV alleles were genotyped with an allele frequency of zero across the cohort and could thus not be included in any downstream analysis. However, this number includes variants with missing genotype predictions (“./.”). In some cases, PanGenie failed to provide a genotype prediction for the carrier samples and typed them as “./.”, while the other samples were correctly typed as 0/0, leading to an allele frequency of zero for the respective variant. Only around 16% variants were genotyped as “0/0” in each of the 1000 Genomes samples. Additionally, unlike

for HGSVC, similar SV alleles were not combined into a single allele in the HPRC callset in order to correct for small representation differences or variant calling errors in some of the samples. This can lead to multiple, slightly different SV alleles present in the callset that actually represent the same SV event. Therefore, a fraction of those alleles that were typed as absent across all samples might also be such redundant alleles correctly filtered out by PanGenie.

In summary, our experiments demonstrate how PanGenie can be applied to a pangenome reference in order to efficiently genotype genetic variants, especially SVs, across large cohorts. The set of high quality genotypes produced gave access to more structual variants than previous approaches, including the HGSVC.

4.4 Conclusions

In this chapter, PanGenie was applied to pangenome references recently constructed from haplotype-resolved assemblies by the HGSVC and HPRC consortia. It was demonstrated that PanGenie enabled accurate and fast genotyping across a large cohort of 3,202 human samples based on short reads when using the two pangenome graphs containing 32 and 44 human samples. By leveraging the information provided by the known haplotypes, PanGenie was able to genotype a large fraction of structural variants that were previously inaccessible by short-read based callers. This demonstrates that short-read based genotyping of SVs benefits from the additional sequence information inherent in the pangenome that is missing from the linear reference. Being able to genotype previously inaccessible SVs across the cohort samples enables the inclusion of such variants into downstream analyses studying their association with disease. Here, we demonstrated that our SV genotypes could be used for Quantitative trait locus analysis, revealing SV-eQTLs that were previously not detectable by short-read approaches. Furthermore, we showed how SNP genotypes produced by PanGenie across the cohort samples enabled the detection of carriers of rare inversions. All predicted carriers could be verified experimentally by FISH.

There are still some limitations and possibilities to improve the current genotyping model. We presented a subsampling strategy implemented in PanGenie which speeds up the runtime as the panel gets larger (Section 4.1.2). The idea was to randomly divide the panel into multiple sets and then run genotyping separately on all of them. Probabilities computed for each subset are then combined iteratively to obtain the final genotype likelihoods. This iterative approach might not be ideal, as it does not put equal weight on the results of each subset when combining likelihoods. Here, alternative strategies could be considered that sum up probabilities first and then normalize the result. Another factor that should be taken into account is that for variants with many alternative alleles, not every subset might cover all these alleles. Thus, not every subset can contribute likelihoods to each possible genotype at the locus. This might lead to biases when computing the final likelihoods. Which strategy works best when combining likelihoods from different subsets is not clear and still needs to be determined in future experiments.

In the near future, highly accurate haplotype-resolved assemblies will become available for larger sets of samples. Thus, instead of containing 64 or 88 haplotypes, future pangenome references will contain hundreds of haplotypes, resulting in larger, more complex graphs. This will impose challenges for pangenome-based genotyping methods like PanGenie: as the number of samples increases, the bubbles in the graph will get more complex, leading to a graph with fewer but larger bubble structures caused by overlapping variants across haplotypes. Genotyping such bubbles will become more difficult due to the higher number of possible alternative alleles the genotyper has to choose between, especially since PanGenie does not genotype nested variants separately (Section 3.10). Furthermore, higher numbers of haplotypes in the graphs lead to a higher runtime and memory usage of PanGenie. While

PanGenie was still very fast on the currently largest graph containing 88 haplotypes (53 single-core CPU hours per sample, 153 GB RAM memory per sample), runtime and memory usage will increase linearly as the number of haplotypes gets larger, leading to problems for graphs containing several hundreds of samples. Currently, storing variant information, keeping track of which paths cover which alleles and which unique k-mers occur in which allele sequence, are the components of the code that require the most memory (Figure C.13). The computation of genotype likelihoods based on the HMM is the most time-consuming step (Figure C.13). How to adapt the model to larger panels is still an open problem. As mentioned earlier (Section 3.10), ideas used in statistical phasing could be explored, as well as reducing the panel sizes based on founder sequences (Section 3.10). Another idea is to subsample smaller sets of haplotypes from the input panel, use the Viterbi algorithm to compute the best two haplotype sequences of each subset and compute the final likelihoods based on the resulting set of haplotypes using the Forward-Backward algorithm. Furthermore, a more efficient way of computing the Forward and Backward probabilities in the current model could be implemented. Below, we present an idea developed very recently that helps to reduce the asymptotic runtime of the Forward-Backward algorithm from $O(n^4 \cdot m)$ to $O(n^2 \cdot m)$ and, combined with the ideas mentioned above, could be extremely helpful in making PanGenie applicable to larger panels.

As more haplotypes are available, it would be interesting to analyze how the genotyping accuracy changes as more haplotypes are added to the graph. On the one hand, adding more samples gives access to more variation that can be re-genotyped in a sample but on the other hand, the accuracy might also become worse once the graph gets too complex. Therefore, it would be helpful to investigate how the genotyping accuracy behaves once the panel gets larger and whether it drops once the graph gets too complex.

4.4.1 Computing PanGenie's Forward-Backward algorithm more efficiently

The idea described in this subsection was suggested by Mikko Rautiainen (personal communication, December 3, 2022).

The asymptotic runtime of PanGenie (Section 3.5.4) can be drastically reduced by introducing helper variables to simplify the computation of Forward- and Backward probabilities (Section 3.2). The idea is based on the fact that the recombination probabilities are independent of the particular haplotypes recombining between columns and only depend on the number of recombination events (none, one or two).

For a bubble position v, we can define the following variables which can be computed in time quadratic in the number of haplotypes, $O(n^2)$:

$$\alpha_v(H_{v,i,*}) = \sum_t \alpha_v(H_{v,i,t}) \quad \forall i$$
$$\alpha_v(H_{v,*,j}) = \sum_s \alpha_v(H_{v,s,j}) \quad \forall j$$
$$\alpha_v(H_{v,*,*}) = \sum_{s,t} \alpha_v(H_{v,s,t})$$

$$\beta_v(H_{v,i,*}) = \sum_t \beta_v(H_{v,i,t}) \cdot P(\mathcal{O}_v|H_{v,i,t}) \quad \forall i$$
$$\beta_v(H_{v,*,j}) = \sum_s \beta_v(H_{v,s,j}) \cdot P(\mathcal{O}_v|H_{v,s,j}) \quad \forall j$$
$$\beta_v(H_{v,*,*}) = \sum_{s,t} \beta_v(H_{v,s,t}) \cdot P(\mathcal{O}_v|H_{v,s,t})$$

Then, the Forward probabilities can be computed in constant time in the following way:

$$\begin{aligned}
\alpha_v(H_{v,i,j}) &= \sum_{s,t} \alpha_{v-1}(H_{v-1,s,t}) \cdot P(H_{v,i,j}|H_{v-1,s,t}) \cdot P(\mathcal{O}_v|H_{v,i,j}) \\
\alpha_v(H_{v,i,j}) &= P(\mathcal{O}_v|H_{v,i,j}) \cdot \sum_{s,t} \alpha_{v-1}(H_{v-1,s,t}) \cdot P(H_{v,i,j}|H_{v-1,s,t}) \\
\alpha_v(H_{v,i,j}) &= P(\mathcal{O}_v|H_{v,i,j}) \cdot \Big[q_r^2 \cdot \alpha_{v-1}(H_{v-1,i,j}) \\
&\quad + p_r q_r \cdot \Big(\sum_t \alpha_{v-1}(H_{v-1,i,t}) - \alpha_{v-1}(H_{v-1,i,j}) \Big) \\
&\quad + q_r p_r \cdot \Big(\sum_s \alpha_{v-1}(H_{v-1,s,j}) - \alpha_{v-1}(H_{v-1,i,j}) \Big) \\
&\quad + p_r^2 \cdot \Big(\sum_{s,t} \alpha_{v-1}(H_{v-1,s,t}) - \sum_s \alpha_{v-1}(H_{v-1,s,j}) - \sum_t \alpha_{v-1}(H_{v-1,i,t}) \\
&\quad + \alpha_{v-1}(H_{v-1,i,j}) \Big) \Big] \\
\alpha_v(H_{v,i,j}) &= P(\mathcal{O}_v|H_{v,i,j}) \cdot \Big[q_r^2 \cdot \alpha_{v-1}(H_{v-1,i,j}) \\
&\quad + p_r q_r \cdot \Big(\alpha_{v-1}(H_{v-1,i,*}) - \alpha_{v-1}(H_{v-1,i,j}) \Big) \\
&\quad + q_r p_r \cdot \Big(\alpha_{v-1}(H_{v-1,*,j}) - \alpha_{v-1}(H_{v-1,i,j}) \Big) \\
&\quad + p_r^2 \cdot \Big(\alpha_{v-1}(H_{v-1,*,*}) - \alpha_{v-1}(H_{v-1,i,*}) - \alpha_{v-1}(H_{v-1,*,j}) + \alpha_{v-1}(H_{v-1,i,j}) \Big) \Big]
\end{aligned}$$

Using the same trick, the Backward probabilities can be computed in constant time as:

$$\begin{aligned}\beta_v(H_{v,i,j}) = q_r^2 \cdot \beta_{v+1}(H_{v+1,i,j}) \cdot P(\mathcal{O}_{v+1}|H_{v+1,i,j}) \\ + p_r q_r \cdot \Big(\beta_{v+1}(H_{v+1,i,*}) - \beta_{v+1}(H_{v+1,i,j}) \cdot P(\mathcal{O}_{v+1}|H_{v+1,i,j})\Big) \\ + q_r p_r \cdot \Big(\beta_{v+1}(H_{v+1,*,j}) - \beta_{v+1}(H_{v+1,i,j}) \cdot P(\mathcal{O}_{v+1}|H_{v+1,i,j})\Big) \\ + p_r^2 \cdot \Big(\beta_{v+1}(H_{v+1,*,*}) - \beta_{v+1}(H_{v+1,i,*}) - \beta_{v+1}(H_{v+1,*,j}) \\ + \beta_{v+1}(H_{v+1,i,j}) \cdot P(\mathcal{O}_{v+1}|H_{v+1,i,j})\Big)\end{aligned}$$

This reduces the runtime of the Forward-Backward algorithm from $O(n^4 \cdot m)$ to $O(n^2 \cdot m)$.

Summary

In Chapter 2 of this thesis, alignment-based phasing methods were discussed and applied to different datasets in order to reconstruct haplotypes of diploid and polyploid genomes. Our new algorithm for polyploid phasing, `whatshap polyphase`, was specifically designed to handle regions of locally identical haplotypes, for which competing tools show lower performances, as they aim to separate haplotypes based on their dissimilarities. We phased a tetraploid potato genome and demonstrated how to generate haplotype-resolved assemblies of genes. We further demonstrated how accurate, PacBio CCS reads improved phasing performances over other long-read based methods and that they enable phasing without relying on short reads for variant calling. While both of these methods provided high quality phasing of SNPs, the experiments showed that problems arise in regions that are highly variable or contain structural variation, hindering them to provide haplotype predictions on the scale of a whole chromosome. Accurate long sequencing reads also provide the basis for reference-free *de novo* reconstruction of haplotypes which are able to cover complex regions. Besides enabling to study the haplotype sequences of single individuals in detail, such haplotype assemblies can be used to construct pangenome graphs aiming to provide a complete overview of the variation present in different samples of the same species. Such graphs can be modeled as sequence graphs. Nodes describe sequences and edges connect sequences such that each haplotype is represented as a path through this graph. Besides many other applications, pangenome graphs are beneficial for genotyping of genetic variants, especially of structural variants. They provide information that enables fast genotyping of variants based on short sequencing data which does not depend on time-consuming alignments of reads to a reference genome. In Chapter 3, we introduced PanGenie, a pangenome-based genotyping method that uses k-mer counts from short-read data of a new sample in combination with a pangenome graph containing haplotypes of a set of known samples in order to genotype the new sample. The k-mer information indicates which alleles are carried by a sample, while the haplotype paths are used to impute genotypes in regions for which no unique k-mers can be identified. We showed that PanGenie outperformed competing methods in terms of speed and genotyping accuracy, especially for structural variants in complex, repetitive regions of the genome, providing insights into regions previously not accessible by short-read based methods. Application of PanGenie to the 3,202 samples of the 1000 Genomes project based on pangenome structures produced by the HGSVC and HPRC consortia (Chapter 4) demonstrated that it was fast, produced high quality genotypes in practice

and allowed genotyping structural variants that other short-read based methods were not able to access. This allowed the inclusion of such variants into Quantitative trait locus analyses enabling the identification of SVs that are associated with diseases. Furthermore, SNP genotypes produced by PanGenie across the cohort samples enabled the detection of carriers of rare inversions. However, as pointed out in Chapters 3 and 4, several challenges arise as future pangenome graphs will be constructed from hundreds of haplotype sequences. It is still unclear how to adapt the genotyping model to efficiently work on such graphs. Founder sets or strategies used by statistical phasing tools might provide useful techniques. In general, it still needs to be determined what the "ideal" size of a pangenome is in terms of the number of haplotypes present in the graph. Adding more samples helps to include rare variants thus increasing accuracy but also the complexity of the graph, which is challenging for many downstream analyses, such as genotyping or read alignment.

Bibliography

[1] 1000 Genomes Project Consortium. A map of human genome variation from population scale sequencing. *Nature*, 467(7319):1061, 2010.

[2] 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature*, 526(7571):68, 2015.

[3] H. J. Abel, D. E. Larson, A. A. Regier, C. Chiang, I. Das, K. L. Kanchi, R. M. Layer, B. M. Neale, W. J. Salerno, C. Reeves, et al. Mapping and characterization of structural variation in 17,795 human genomes. *Nature*, 583(7814):83–89, 2020.

[4] L. Abi-Rached, P. Gouret, J.-H. Yeh, J. Di Cristofaro, P. Pontarotti, C. Picard, and J. Paganini. Immune diversity sheds light on missing variation in worldwide genetic diversity panels. *PLoS One*, 13(10):e0206512, 2018.

[5] D. Aguiar and S. Istrail. HapCompass: a fast cycle basis algorithm for accurate haplotype assembly of sequence data. *Journal of Computational Biology*, 19(6):577–590, 2012.

[6] D. Aguiar and S. Istrail. Haplotype assembly in polyploid genomes and identical by descent shared tracts. *Bioinformatics*, 29(13):i352–i360, 2013.

[7] M. U. Ahsan, Q. Liu, L. Fang, and K. Wang. NanoCaller for accurate detection of SNPs and indels in difficult-to-map regions from long-read sequencing by haplotype-aware deep neural networks. *Genome Biology*, 22(1):1–33, 2021.

[8] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter. *Molecular Biology of the Cell. 4th edition*. Garland Science, 2002.

[9] J. Armstrong, G. Hickey, M. Diekhans, I. T. Fiddes, A. M. Novak, A. Deran, Q. Fang, D. Xie, S. Feng, J. Stiller, et al. Progressive cactus is a multiple-genome aligner for the thousand-genome era. *Nature*, 587(7833):246–251, 2020.

[10] V. Bansal and V. Bafna. HapCUT: an efficient and accurate algorithm for the haplotype assembly problem. *Bioinformatics*, 24(16):i153–i159, 2008.

[11] D. Barber. *Bayesian reasoning and machine learning*. Cambridge University Press, 2012.

[12] E. Berger, D. Yorukoglu, J. Peng, and B. Berger. HapTree: a novel bayesian framework for single individual polyplotyping using NGS data. *PLoS Computational Biology*, 10(3):e1003502, 2014.

[13] D. Beyter, H. Ingimundardottir, A. Oddsson, H. P. Eggertsson, E. Bjornsson, H. Jonsson, B. A. Atlason, S. Kristmundsdottir, S. Mehringer, M. T. Hardarson, et al. Long-read sequencing of 3,622 icelanders provides insight into the role of structural variants in human diseases and other traits. *Nature Genetics*, 53(6):779–786, 2021.

[14] D. N. Bharadwaj. Polyploidy in crop improvement and evolution. In *Plant Biology and Biotechnology*, pages 619–638. Springer, 2015.

[15] S. Böcker, S. Briesemeister, and G. W. Klau. Exact algorithms for cluster editing: Evaluation and experiments. *Algorithmica*, 60(2):316–334, 2011.

[16] M.-L. Bondeson, N. Dahl, H. Malmgren, W. J. Kleijer, T. Tönnesen, B.-M. Carlberg, and U. Pettersson. Inversion of the IDS gene resulting from recombination with IDS-related sequences in a common cause of the hunter syndrome. *Human Molecular Genetics*, 4(4):615–621, 1995.

[17] P. Bonizzoni, R. Dondi, G. W. Klau, Y. Pirola, N. Pisanti, and S. Zaccaria. On the minimum error correction problem for haplotype assembly in diploid and polyploid genomes. *Journal of Computational Biology*, 23(9):718–736, 2016.

[18] B. L. Browning and S. R. Browning. Genotype imputation with millions of reference samples. *The American Journal of Human Genetics*, 98(1):116–126, 2016.

[19] B. L. Browning, Y. Zhou, and S. R. Browning. A one-penny imputed genome from next-generation reference panels. *The American Journal of Human Genetics*, 103(3):338–348, 2018.

[20] A. Buniello, J. A. L. MacArthur, M. Cerezo, L. W. Harris, J. Hayhurst, C. Malangone, A. McMahon, J. Morales, E. Mountjoy, E. Sollis, et al. The NHGRI-EBI GWAS catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Research*, 47(D1):D1005–D1012, 2019.

[21] M. Byrska-Bishop, U. S. Evani, X. Zhao, A. O. Basile, H. J. Abel, A. A. Regier, A. Corvelo, W. E. Clarke, R. Musunuri, K. Nagulapalli, et al. High-coverage whole-genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. *Cell*, 185(18):3426–3440, 2022.

[22] C. J. Castro and T. F. F. Ng. U50: a new metric for measuring assembly output based on non-overlapping, target-specific contigs. *Journal of Computational Biology*, 24(11):1071–1080, 2017.

[23] M. J. Chaisson, S. Mukherjee, S. Kannan, and E. E. Eichler. Resolving multicopy duplications de novo using polyploid phasing. In *International Conference on Research in Computational Molecular Biology*, pages 117–133. Springer, 2017.

[24] M. J. Chaisson, A. D. Sanders, X. Zhao, A. Malhotra, D. Porubsky, T. Rausch, E. J. Gardner, O. L. Rodriguez, L. Guo, R. L. Collins, et al. Multi-platform discovery of haplotype-resolved structural variation in human genomes. *Nature Communications*, 10(1):1–16, 2019.

[25] D. I. Chasman, M. Schürks, V. Anttila, B. de Vries, U. Schminke, L. J. Launer, G. M. Terwindt, A. M. van den Maagdenberg, K. Fendrich, H. Völzke, et al. Genome-wide association study reveals three susceptibility loci for common migraine in the general population. *Nature Genetics*, 43(7):695–698, 2011.

[26] S. Chen, P. Krusche, E. Dolzhenko, R. M. Sherman, R. Petrovski, F. Schlesinger, M. Kirsche, D. R. Bentley, M. C. Schatz, F. J. Sedlazeck, et al. Paragraph: a graph-based structural variant genotyper for short-read sequence data. *Genome Biology*, 20(1):1–13, 2019.

[27] H. Cheng, G. T. Concepcion, X. Feng, H. Zhang, and H. Li. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nature Methods*, 18(2):170–175, 2021.

[28] C. Chiang, R. M. Layer, G. G. Faust, M. R. Lindberg, D. B. Rose, E. P. Garrison, G. T. Marth, A. R. Quinlan, and I. M. Hall. SpeedSeq: ultra-fast personal genome analysis and interpretation. *Nature Methods*, 12(10):966, 2015.

[29] J. G. Cleary, R. Braithwaite, K. Gaastra, B. S. Hilbush, S. Inglis, S. A. Irvine, A. Jackson, R. Littin, M. Rathod, D. Ware, et al. Comparing variant call files for performance benchmarking of next-generation sequencing variant calling pipelines. *bioRxiv*, page 023754, 2015.

[30] B. P. Coe, K. Witherspoon, J. A. Rosenfeld, B. W. Van Bon, A. T. Vulto-van Silfhout, P. Bosco, K. L. Friend, C. Baker, S. Buono, L. E. Vissers, et al. Refining analyses of copy number variation identifies specific genes associated with developmental delay. *Nature Genetics*, 46(10):1063–1071, 2014.

[31] F. S. Collins, M. L. Drumm, J. L. Cole, W. K. Lockwood, G. F. Vande Woude, and M. C. Iannuzzi. Construction of a general human chromosome jumping library, with application to cystic fibrosis. *Science*, 235(4792):1046–1049, 1987.

[32] R. L. Collins, H. Brand, K. J. Karczewski, X. Zhao, J. Alföldi, L. C. Francioli, A. V. Khera, C. Lowther, L. D. Gauthier, H. Wang, et al. A structural variation reference for medical and population genetics. *Nature*, 581(7809):444–451, 2020.

[33] N. Craddock, M. E. Hurles, N. Cardin, R. D. Pearson, V. Plagnol, S. Robson, D. Vukcevic, C. Barnes, D. F. Conrad, E. Giannoulatou, et al. Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls. *Nature*, 464(7289):713, 2010.

[34] S. Das and H. Vikalo. SDhaP: haplotype assembly for diploids and polyploids via semi-definite programming. *BMC Genomics*, 16(1):1–16, 2015.

[35] S. Das, L. Forer, S. Schönherr, C. Sidore, A. E. Locke, A. Kwong, S. I. Vrieze, E. Y. Chew, S. Levy, M. McGue, et al. Next-generation genotype imputation service and methods. *Nature Genetics*, 48(10):1284–1287, 2016.

[36] J. De Grouchy. Chromosome phylogenies of man, great apes, and old world monkeys. *Genetica*, 73(1):37–52, 1987.

[37] O. Delaneau, J. Marchini, and J.-F. Zagury. A linear complexity phasing method for thousands of genomes. *Nature Methods*, 9(2):179–181, 2012.

[38] O. Delaneau, J.-F. Zagury, M. R. Robinson, J. L. Marchini, and E. T. Dermitzakis. Accurate, scalable and integrative haplotype estimation. *Nature Communications*, 10(1):1–10, 2019.

[39] M. A. DePristo, E. Banks, R. Poplin, K. V. Garimella, J. R. Maguire, C. Hartl, A. A. Philippakis, G. Del Angel, M. A. Rivas, M. Hanna, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics*, 43(5):491, 2011.

[40] A. Dilthey, C. Cox, Z. Iqbal, M. R. Nelson, and G. McVean. Improved genome inference in the MHC using a population reference graph. *Nature Genetics*, 47(6):682, 2015.

[41] A. T. Dilthey, P.-A. Gourraud, A. J. Mentzer, N. Cereb, Z. Iqbal, and G. McVean. High-Accuracy HLA type inference from Whole-Genome sequencing data using population reference graphs. *PLoS Computational Biology*, 12(10):e1005151, 2016.

[42] A. T. Dilthey, A. J. Mentzer, R. Carapito, C. Cutland, N. Cereb, S. A. Madhi, A. Rhie, S. Koren, S. Bahram, G. McVean, and A. M. Phillippy. HLA*LA-HLA typing from linearly projected graph alignments. *Bioinformatics*, 35(21):4394–4396, 2019.

[43] J. Ding, A. Bashashati, A. Roth, A. Oloumi, K. Tse, T. Zeng, G. Haffari, M. Hirst, M. A. Marra, A. Condon, et al. Feature-based classifiers for somatic mutation detection in tumour–normal paired sequencing data. *Bioinformatics*, 28(2):167–175, 2012.

[44] D. D. Dolle, Z. Liu, M. Cotten, J. T. Simpson, Z. Iqbal, R. Durbin, S. A. McCarthy, and T. M. Keane. Using reference-free compressed data structures to analyze sequencing reads from thousands of human genomes. *Genome Research*, 27(2):300–309, 2017.

[45] J. Dubcovsky, M.-C. Luo, G.-Y. Zhong, R. Bransteitter, A. Desai, A. Kilian, A. Kleinhofs, and J. Dvořák. Genetic map of diploid wheat, Triticum monococcum L., and its comparison with maps of Hordeum vulgare L. *Genetics*, 143(2):983–999, 1996.

[46] P. Ebert, P. A. Audano, Q. Zhu, B. Rodriguez-Martin, D. Porubsky, M. J. Bonder, A. Sulovari, J. Ebler, W. Zhou, R. Serra Mari, et al. Haplotype-resolved diverse human genomes and integrated analysis of structural variation. *Science*, 372(6537), 2021.

[47] J. Ebler, A. Schönhuth, and T. Marschall. Genotyping inversions and tandem duplications. *Bioinformatics*, 33(24):4015–4023, 2017.

[48] J. Ebler, M. Haukness, T. Pesout, T. Marschall, and B. Paten. Haplotype-aware diplotyping from noisy long reads. *Genome Biology*, 20(1):1–16, 2019.

[49] J. Ebler, P. Ebert, W. E. Clarke, T. Rausch, P. A. Audano, T. Houwaart, Y. Mao, J. O. Korbel, E. E. Eichler, M. C. Zody, et al. Pangenome-based genome inference allows efficient and accurate genotyping across a wide spectrum of variant classes. *Nature Genetics*, 54(4):518–525, 2022.

[50] P. Edge and V. Bansal. Longshot enables accurate variant calling in diploid genomes from single-molecule long read sequencing. *Nature Communications*, 10(1):1–10, 2019.

[51] P. Edge, V. Bafna, and V. Bansal. HapCUT2: robust and accurate haplotype assembly for diverse sequencing technologies. *Genome Research*, 27(5):801–812, 2017.

[52] H. P. Eggertsson, H. Jonsson, S. Kristmundsdottir, E. Hjartarson, B. Kehr, G. Masson, F. Zink, K. E. Hjorleifsson, A. Jonasdottir, A. Jonasdottir, et al. Graphtyper enables population-scale genotyping using pangenome graphs. *Nature Genetics*, 49(11):1654, 2017.

[53] H. P. Eggertsson, S. Kristmundsdottir, D. Beyter, H. Jonsson, A. Skuladottir, M. T. Hardarson, D. F. Gudbjartsson, K. Stefansson, B. V. Halldorsson, and P. Melsted. GraphTyper2 enables population-scale genotyping of structural variation using pangenome graphs. *Nature Communications*, 10(1):1–8, 2019.

[54] J. M. Eizenga, A. M. Novak, J. A. Sibbesen, S. Heumos, A. Ghaffaari, G. Hickey, X. Chang, J. D. Seaman, R. Rounthwaite, J. Ebler, et al. Pangenome graphs. *Annual Review of Genomics and Human Genetics*, 21:139–162, 2020.

[55] ENCODE Project Consortium et al. The ENCODE (ENCyclopedia of DNA elements) project. *Science*, 306(5696):636–640, 2004.

[56] A. C. English, V. K. Menon, R. Gibbs, G. A. Metcalf, and F. J. Sedlazeck. Truvari: Refined structural variant comparison preserves allelic diversity. *bioRxiv*, 2022.

[57] J. Ernst and M. Kellis. ChromHMM: automating chromatin-state discovery and characterization. *Nature Methods*, 9(3):215–216, 2012.

[58] S. Garg, M. Martin, and T. Marschall. Read-based phasing of related individuals. *Bioinformatics*, 32(12):i234–i242, 2016.

[59] S. Garg, A. Fungtammasan, A. Carroll, M. Chou, A. Schmitt, X. Zhou, S. Mac, P. Peluso, E. Hatas, J. Ghurye, et al. Chromosome-scale, haplotype-resolved assembly of human genomes. *Nature Biotechnology*, 39(3):309–312, 2021.

[60] E. Garrison and G. Marth. Haplotype-based variant detection from short-read sequencing. *arXiv preprint arXiv:1207.3907*, 2012.

[61] E. Garrison, J. Sirén, A. M. Novak, G. Hickey, J. M. Eizenga, E. T. Dawson, W. Jones, S. Garg, C. Markello, M. F. Lin, et al. Variation graph toolkit improves read mapping by representing genetic variation in the reference. *Nature Biotechnology*, 36(9):875–879, 2018.

[62] I. Giegling, A. M. Hartmann, H.-J. Möller, and D. Rujescu. Anger-and aggression-related traits are associated with polymorphisms in the 5-HT-2A gene. *Journal of Affective Disorders*, 96(1-2):75–81, 2006.

[63] G. Gimelli, M. A. Pujana, M. G. Patricelli, S. Russo, D. Giardino, L. Larizza, J. Cheung, L. Armengol, A. Schinzel, X. Estivill, et al. Genomic inversions of human chromosome 15q11–q13 in mothers of angelman syndrome patients with class II (BP2/3) deletions. *Human Molecular Genetics*, 12(8):849–858, 2003.

[64] G. Gonnella, N. Niehus, and S. Kurtz. GfaViz: flexible and interactive visualization of gfa sequence graphs. *Bioinformatics*, 35(16):2853–2855, 2019.

[65] P.-A. Gourraud, P. Khankhanian, N. Cereb, S. Y. Yang, M. Feolo, M. Maiers, J. D. Rioux, S. Hauser, and J. Oksenberg. HLA diversity in the 1000 genomes dataset. *PLoS One*, 9(7): e97282, 2014.

[66] R. Goya, M. G. Sun, R. D. Morin, G. Leung, G. Ha, K. C. Wiegand, J. Senz, A. Crisan, M. A. Marra, M. Hirst, et al. SNVMix: predicting single nucleotide variants from next-generation sequencing of tumors. *Bioinformatics*, 26(6):730–736, 2010.

[67] C. Grasso and C. Lee. Combining partial order alignment and progressive multiple sequence alignment increases alignment speed and scalability to very large alignment problems. *Bioinformatics*, 20(10):1546–1556, 2004.

[68] Y. Guo, F. Ye, Q. Sheng, T. Clark, and D. C. Samuels. Three-stage quality control strategies for DNA re-sequencing data. *Briefings in Bioinformatics*, 15(6):879–889, 2013.

[69] M. B. Hamilton. *Population genetics*. John Wiley & Sons, 2021.

[70] R. E. Handsaker, V. Van Doren, J. R. Berman, G. Genovese, S. Kashin, L. M. Boettger, and S. A. McCarroll. Large multiallelic copy number variations in humans. *Nature Genetics*, 47(3): 296–303, 2015.

[71] M. A. Hardigan, E. Crisovan, J. P. Hamilton, J. Kim, P. Laimbeer, C. P. Leisner, N. C. Manrique-Carpintero, L. Newton, G. M. Pham, B. Vaillancourt, X. Yang, Z. Zeng, D. S. Douches, J. Jiang, R. E. Veilleux, and C. R. Buell. Genome reduction uncovers a large dispensable genome and adaptive role for copy number variation in asexually propagated Solanum tuberosum. *The Plant Cell*, 28(2):388–405, 2016.

[72] D. L. Hartl and A. G. Clark. *Principles of population genetics, fourth edition*. Sinauer associates Sunderland, 2007.

[73] D. He, S. Saha, R. Finkers, and L. Parida. Efficient algorithms for polyploid haplotype phasing. *BMC Genomics*, 19(2):171–180, 2018.

[74] D. Heller and M. Vingron. SVIM: structural variant identification using mapped long reads. *Bioinformatics*, 35(17):2907–2915, 2019.

[75] G. Hickey, D. Heller, J. Monlong, J. A. Sibbesen, J. Sirén, J. Eizenga, E. T. Dawson, E. Garrison, A. M. Novak, and B. Paten. Genotyping structural variants in pangenome graphs using the vg toolkit. *Genome Biology*, 21(1):1–17, 2020.

[76] G. Hickey, J. Monlong, A. Novak, J. M. Eizenga, H. Li, B. Paten, H. P. R. Consortium, et al. Pangenome graph construction from genome alignment with minigraph-cactus. *bioRxiv*, 2022.

[77] G. Holley and P. Melsted. Bifrost: highly parallel construction and indexing of colored and compacted de Bruijn graphs. *Genome Biology*, 21(1):1–20, 2020.

[78] B. Howie, J. Marchini, and M. Stephens. Genotype imputation with thousands of genomes. *G3: Genes, Genomes, Genetics*, 1(6):457–470, 2011.

[79] B. N. Howie, P. Donnelly, and J. Marchini. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genetics*, 5(6): e1000529, 2009.

[80] A. J. Iafrate, L. Feuk, M. N. Rivera, M. L. Listewnik, P. K. Donahoe, Y. Qi, S. W. Scherer, and C. Lee. Detection of large-scale variation in the human genome. *Nature Genetics*, 36(9): 949–951, 2004.

[81] Illumina Inc. An introduction to Next-Generation Sequencing Technology, 2017. URL `https://www.illumina.com/content/dam/illumina-marketing/documents/products/illumina_sequencing_introduction.pdf`. visited on 2022-04-27.

[82] Illumina Inc. FASTQ files explained, 2022. URL `https://support.illumina.com/bulletins/2016/04/fastq-files-explained.html`. visited on 2022-04-26.

[83] Illumina Inc. MiniSeq Specifications, 2022. URL https://www.illumina.com/systems/sequencing-platforms/miniseq/specifications.html. visited on 2022-04-29.

[84] Illumina Inc. MiSeq Specifications, 2022. URL https://www.illumina.com/systems/sequencing-platforms/miseq/specifications.html. visited on 2022-04-29.

[85] Illumina Inc. NextSeq Series Specifications, 2022. URL https://www.illumina.com/systems/sequencing-platforms/nextseq/specifications.html. visited on 2022-04-29.

[86] Illumina Inc. NovaSeq 6000 System Specifications, 2022. URL https://www.illumina.com/systems/sequencing-platforms/novaseq/specifications.html. visited on 2022-04-29.

[87] Illumina Inc. Illumina sequencing platforms, 2022. URL https://www.illumina.com/systems/sequencing-platforms.html. visited on 2022-04-29.

[88] International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, 2001.

[89] Z. Iqbal, M. Caccamo, I. Turner, P. Flicek, and G. McVean. De novo assembly and genotyping of variants using colored de Bruijn graphs. *Nature Genetics*, 44(2):226, 2012.

[90] E. D. Jarvis, G. Formenti, A. Rhie, A. Guarracino, C. Yang, J. Wood, A. Tracey, F. Thibaud-Nissen, M. R. Vollger, D. Porubsky, et al. Semi-automated assembly of high-quality diploid human reference genomes. *Nature*, 611(7936):519–531, 2022.

[91] T. Jiang, Y. Liu, Y. Jiang, J. Li, Y. Gao, Z. Cui, Y. Liu, B. Liu, and Y. Wang. Long-read-based human genomic structural variation detection with cuteSV. *Genome Biology*, 21(1):1–24, 2020.

[92] B. H. Juang and L. R. Rabiner. Hidden markov models for speech recognition. *Technometrics*, 33(3):251–272, 1991.

[93] D. Karolchik, A. S. Hinrichs, T. S. Furey, K. M. Roskin, C. W. Sugnet, D. Haussler, and W. J. Kent. The UCSC Table Browser data retrieval tool. *Nucleic Acids Research*, 32(suppl_1):D493–D496, 2004.

[94] J. M. Kidd, G. M. Cooper, W. F. Donahue, H. S. Hayden, N. Sampas, T. Graves, N. Hansen, B. Teague, C. Alkan, F. Antonacci, et al. Mapping and sequencing of structural variation from eight human genomes. *Nature*, 453(7191):56–64, 2008.

[95] D. Kim, J. M. Paggi, C. Park, C. Bennett, and S. L. Salzberg. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nature Biotechnology*, 37(8):907–915, 2019.

[96] D. C. Koboldt, Q. Zhang, D. E. Larson, D. Shen, M. D. McLellan, L. Lin, C. A. Miller, E. R. Mardis, L. Ding, and R. K. Wilson. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Research*, 22(3):568–576, 2012.

[97] J. O. Korbel, A. E. Urban, J. P. Affourtit, B. Godwin, F. Grubert, J. F. Simons, P. M. Kim, D. Palejev, N. J. Carriero, L. Du, et al. Paired-end mapping reveals extensive structural variation in the human genome. *Science*, 318(5849):420–426, 2007.

[98] S. Koren, A. Rhie, B. P. Walenz, A. T. Dilthey, D. M. Bickhart, S. B. Kingan, S. Hiendleder, J. L. Williams, T. P. Smith, and A. M. Phillippy. De novo assembly of haplotype-resolved genomes with trio binning. *Nature Biotechnology*, 36(12):1174–1182, 2018.

[99] Z. N. Kronenberg, I. T. Fiddes, D. Gordon, S. Murali, S. Cantsilieris, O. S. Meyerson, J. G. Underwood, B. J. Nelson, M. J. Chaisson, M. L. Dougherty, et al. High-resolution comparative analysis of great ape genomes. *Science*, 360(6393), 2018.

[100] Z. N. Kronenberg, A. Rhie, S. Koren, G. T. Concepcion, P. Peluso, K. M. Munson, D. Porubsky, K. Kuhn, K. A. Mueller, W. Y. Low, et al. Extended haplotype-phasing of long-read de novo genome assemblies using Hi-C. *Nature Communications*, 12(1):1–10, 2021.

[101] M. Kyriakidou, S. R. Achakkagari, J. H. Gálvez López, X. Zhu, C. Y. Tang, H. H. Tai, N. L. Anglin, D. Ellis, and M. V. Strömvik. Structural genome analysis in cultivated potato taxa. *Theoretical and Applied Genetics*, 133(3):951–966, 2020.

[102] D. Lakich, H. H. Kazazian, S. E. Antonarakis, and J. Gitschier. Inversions disrupting the factor VIII gene are a common cause of severe haemophilia A. *Nature Genetics*, 5(3):236–241, 1993.

[103] L. Lecompte, P. Peterlongo, D. Lavenier, and C. Lemaitre. SVJedi: genotyping structural variations with long reads. *Bioinformatics*, 36(17):4568–4575, 2020.

[104] C. Lee, C. Grasso, and M. F. Sharlow. Multiple sequence alignment using partial order graphs. *Bioinformatics*, 18(3):452–464, 2002.

[105] B. Letcher, M. Hunt, and Z. Iqbal. Gramtools enables multiscale variation analysis with genome graphs. *Genome Biology*, 22(1):1–27, 2021.

[106] H. Li. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, 34(18):3094–3100, 2018.

[107] H. Li. SNP vs SNV, 2021. URL `http://lh3.github.io/2021/03/15/snp-vs-snv`. visited on 2022-10-20.

[108] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, and R. Durbin. The sequence alignment/map format and SAMtools. *Bioinformatics*, 25(16):2078–2079, 2009.

[109] H. Li, J. M. Bloom, Y. Farjoun, M. Fleharty, L. Gauthier, B. Neale, and D. MacArthur. A synthetic-diploid benchmark for accurate variant-calling evaluation. *Nature Methods*, 15(8): 595–597, 2018.

[110] H. Li, X. Feng, and C. Chu. The design and construction of reference pangenome graphs with minigraph. *Genome Biology*, 21(1):1–19, 2020.

[111] N. Li and M. Stephens. Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics*, 165(4):2213–2233, 2003.

[112] Y. Li, N. D. Roberts, J. A. Wala, O. Shapira, S. E. Schumacher, K. Kumar, E. Khurana, S. Waszak, J. O. Korbel, J. E. Haber, et al. Patterns of somatic structural variation in human cancer genomes. *Nature*, 578(7793):112–121, 2020.

[113] W.-W. Liao, M. Asri, J. Ebler, D. Doerr, M. Haukness, G. Hickey, S. Lu, J. K. Lucas, J. Monlong, H. J. Abel, et al. A draft human pangenome reference. *bioRxiv*, 2022.

[114] G. A. Logsdon, M. R. Vollger, and E. E. Eichler. Long-read human genome sequencing and its applications. *Nature Reviews Genetics*, 21(10):597–614, 2020.

[115] G. A. Logsdon, M. R. Vollger, P. Hsieh, Y. Mao, M. A. Liskovykh, S. Koren, S. Nurk, L. Mercuri, P. C. Dishuck, A. Rhie, et al. The structure, function and evolution of a complete human chromosome 8. *Nature*, 593(7857):101–107, 2021.

[116] P.-R. Loh, P. Danecek, P. F. Palamara, C. Fuchsberger, Y. A. Reshef, H. K. Finucane, S. Schoenherr, L. Forer, S. McCarthy, G. R. Abecasis, et al. Reference-based phasing using the Haplotype Reference Consortium panel. *Nature Genetics*, 48(11):1443–1448, 2016.

[117] R. Luo, C.-L. Wong, Y.-S. Wong, C.-I. Tang, C.-M. Liu, C.-M. Leung, and T.-W. Lam. Exploring the limit of using a deep neural network on pileup data for germline variant calling. *Nature Machine Intelligence*, 2(4):220–227, 2020.

[118] D. Malhotra, S. McCarthy, J. J. Michaelson, V. Vacic, K. E. Burdick, S. Yoon, S. Cichon, A. Corvin, S. Gary, E. S. Gershon, et al. High frequencies of de novo CNVs in bipolar disorder and schizophrenia. *Neuron*, 72(6):951–963, 2011.

[119] G. Marçais and C. Kingsford. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics*, 27(6):764–770, 2011.

[120] S. Marcus, H. Lee, and M. C. Schatz. SplitMEM: a graphical algorithm for pan-genome analysis with suffix skips. *Bioinformatics*, 30(24):3476–3483, 2014.

[121] A. M. Maxam and W. Gilbert. Sequencing end-labeled DNA with base-specific chemical cleavages. *Methods in Enzymology*, 65(1):499–560, 1980.

[122] W. R. McCombie, J. D. McPherson, and E. R. Mardis. Next-generation sequencing technologies. *Cold Spring Harbor Perspectives in Medicine*, 9(11):a036798, 2019.

[123] P. G. Meirmans and P. W. Hedrick. Assessing population structure: Fst and related measures. *Molecular Ecology Resources*, 11(1):5–18, 2011.

[124] D. Melzer, J. R. Perry, D. Hernandez, A.-M. Corsi, K. Stevens, I. Rafferty, F. Lauretani, A. Murray, J. R. Gibbs, G. Paolisso, et al. A genome-wide association study identifies protein quantitative trait loci (pQTLs). *PLoS Genetics*, 4(5):e1000072, 2008.

[125] A. Menelaou and J. Marchini. Genotype calling and phasing using next-generation sequencing reads and a haplotype scaffold. *Bioinformatics*, 29(1):84–91, 2013.

[126] I. Minkin and P. Medvedev. Scalable multiple whole-genome alignment and locally collinear block construction with SibeliaZ. *Nature Communications*, 11(1):1–11, 2020.

[127] I. Minkin, S. Pham, and P. Medvedev. TwoPaCo: an efficient algorithm to build the compacted de Bruijn graph from many complete genomes. *Bioinformatics*, 33(24):4024–4032, 2017.

[128] B. Mor, S. Garhwal, and A. Kumar. A systematic review of hidden markov models and their applications. *Archives of Computational Methods in Engineering*, 28(3):1429–1448, 2021.

[129] E. Motazedi, D. de Ridder, R. Finkers, S. Baldwin, S. Thomson, K. Monaghan, and C. Maliepaard. TriPoly: haplotype estimation for polyploids using sequencing data of related individuals. *Bioinformatics*, 34(22):3864–3872, 2018.

[130] E. Motazedi, R. Finkers, C. Maliepaard, and D. de Ridder. Exploiting next-generation sequencing to solve the haplotyping puzzle in polyploids: a simulation study. *Briefings in Bioinformatics*, 19(3):387–403, 2018.

[131] National Center for Biotechnology Information, U.S. National Library of Medicine. BLAST topics. URL `www.ncbi.nlm.nih.gov/BLAST/fasta.shtml`. visited on 2022-04-26.

[132] S. B. Needleman and C. D. Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, 48(3):443–453, 1970.

[133] M. S. Nicoloso, H. Sun, R. Spizzo, H. Kim, P. Wickramasinghe, M. Shimizu, S. E. Wojcik, J. Ferdin, T. Kunej, L. Xiao, et al. Single-nucleotide polymorphisms inside microRNA target sites influence tumor susceptibility. *Cancer Research*, 70(7):2789–2798, 2010.

[134] T. Norri, B. Cazaux, D. Kosolobov, and V. Mäkinen. Linear time minimum segmentation enables scalable founder reconstruction. *Algorithms for Molecular Biology*, 14(1):1–15, 2019.

[135] T. Norri, B. Cazaux, S. Dönges, D. Valenzuela, and V. Mäkinen. Founder reconstruction enables scalable and seamless pangenomic analysis. *Bioinformatics*, 37(24):4611–4619, 2021.

[136] S. Nurk, S. Koren, A. Rhie, M. Rautiainen, A. V. Bzikadze, A. Mikheenko, M. R. Vollger, N. Altemose, L. Uralsky, A. Gershman, et al. The complete sequence of a human genome. *Science*, 376(6588):44–53, 2022.

[137] N. D. Olson, J. Wagner, J. McDaniel, S. H. Stephens, S. T. Westreich, A. G. Prasanna, E. Johanson, E. Boja, E. J. Maier, O. Serang, et al. PrecisionFDA Truth Challenge V2: Calling variants from short and long reads in difficult-to-map regions. *Cell Genomics*, 2(5):100129, 2022.

[138] Y. Ono, K. Asai, and M. Hamada. PBSIM: Pacbio reads simulator-toward accurate genome assembly. *Bioinformatics*, 29(1):119–121, 2013.

[139] Oxford Nanopore Technologies. How does nanopore DNA/RNA sequencing work?, 2017. URL `https://nanoporetech.com/how-it-works`. visited on 2022-04-28.

[140] T. S. Painter. A comparison of the chromosomes of the rat and mouse with reference to the question of chromosome homology in mammals. *Genetics*, 13(2):180, 1928.

[141] J. S. Papadopoulos and R. Agarwala. COBALT: constraint-based alignment tool for multiple protein sequences. *Bioinformatics*, 23(9):1073–1079, 2007.

[142] B. Paten, J. M. Eizenga, Y. M. Rosen, A. M. Novak, E. Garrison, and G. Hickey. Superbubbles, ultrabubbles, and cacti. *Journal of Computational Biology*, 25(7):649–663, 2018.

[143] M. Patterson, T. Marschall, N. Pisanti, L. Van Iersel, L. Stougie, G. W. Klau, and A. Schönhuth. WhatsHap: weighted haplotype assembly for future-generation sequencing reads. *Journal of Computational Biology*, 22(6):498–509, 2015.

[144] R. Poplin, P.-C. Chang, D. Alexander, S. Schwartz, T. Colthurst, A. Ku, D. Newburger, J. Dijamco, N. Nguyen, P. T. Afshar, et al. A universal SNP and small-indel variant caller using deep neural networks. *Nature Biotechnology*, 36(10):983–987, 2018.

[145] D. Porubsky, P. Ebert, P. A. Audano, M. R. Vollger, W. T. Harvey, P. Marijon, J. Ebler, K. M. Munson, M. Sorensen, A. Sulovari, et al. Fully phased human genome assembly without parental data using single-cell strand sequencing and long reads. *Nature Biotechnology*, 39(3): 302–308, 2021.

[146] D. Porubsky, W. Höps, H. Ashraf, P. Hsieh, B. Rodriguez-Martin, F. Yilmaz, J. Ebler, P. Hallast, F. A. M. Maggiolini, W. T. Harvey, et al. Recurrent inversion polymorphisms in humans associate with genetic instability and genomic disorders. *Cell*, 185(11):1986–2005, 2022.

[147] J. Pritt, N.-C. Chen, and B. Langmead. FORGe: prioritizing variants for graph genomes. *Genome Biology*, 19(1):1–16, 2018.

[148] S. Purcell, B. Neale, K. Todd-Brown, L. Thomas, M. A. Ferreira, D. Bender, J. Maller, P. Sklar, P. I. De Bakker, M. J. Daly, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics*, 81(3):559–575, 2007.

[149] L. R. Rabiner and B. H. Juang. An introduction to hidden Markov models. *IEEE ASSP Magazine*, pages 4–15, 1986.

[150] L. R. Rabiner, S. E. Levinson, and M. M. Sondhi. On the application of vector quantization and hidden markov models to speaker-independent, isolated word recognition. *Bell System Technical Journal*, 62(4):1075–1105, 1983.

[151] G. Rakocevic, V. Semenyuk, W.-P. Lee, J. Spencer, J. Browning, I. J. Johnson, V. Arsenijevic, J. Nadj, K. Ghose, M. C. Suciu, et al. Fast and accurate genomic analyses using genome graphs. *Nature Genetics*, 51(2):354–362, 2019.

[152] T. Rausch, T. Zichner, A. Schlattl, A. M. Stütz, V. Benes, and J. O. Korbel. DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics*, 28(18): i333–i339, 2012.

[153] M. Rautiainen and T. Marschall. GraphAligner: rapid and versatile sequence-to-graph alignment. *Genome Biology*, 21(1):1–28, 2020.

[154] M. Rautiainen, S. Nurk, B. P. Walenz, G. A. Logsdon, D. Porubsky, A. Rhie, E. E. Eichler, A. M. Phillippy, and S. Koren. Verkko: telomere-to-telomere assembly of diploid chromosomes. *bioRxiv*, 2022.

[155] R. Redon, S. Ishikawa, K. R. Fitch, L. Feuk, G. H. Perry, T. D. Andrews, H. Fiegler, M. H. Shapero, A. R. Carson, W. Chen, et al. Global variation in copy number in the human genome. *Nature*, 444(7118):444–454, 2006.

[156] M. E. Reid and G. A. Denomme. DNA-based methods in the immunohematology reference laboratory. *Transfusion and Apheresis Science*, 44(1):65–72, 2011.

[157] A. Rhoads and K. F. Au. PacBio sequencing and its applications. *Genomics, Proteomics & Bioinformatics*, 13(5):278–289, 2015.

[158] A. Rimmer, H. Phan, I. Mathieson, Z. Iqbal, S. R. Twigg, A. O. Wilkie, G. McVean, G. Lunter, W. Consortium, et al. Integrating mapping-, assembly-and haplotype-based approaches for calling variants in clinical sequencing applications. *Nature Genetics*, 46(8):912, 2014.

[159] J. Robinson, K. Mistry, H. McWilliam, R. Lopez, and S. G. E. Marsh. IPD–the immuno polymorphism database. *Nucleic Acids Research*, 38(Database issue):D863–9, 2010.

[160] J. T. Robinson, H. Thorvaldsdóttir, W. Winckler, M. Guttman, E. S. Lander, G. Getz, and J. P. Mesirov. Integrative genomics viewer. *Nature Biotechnology*, 29(1):24, 2011.

[161] J. Ruan and H. Li. Fast and accurate long-read assembly with wtdbg2. *Nature Methods*, 2019.

[162] Samtools. The Variant Call Format (VCF) Version 4.3 Specification, 2021. URL `https://samtools.github.io/hts-specs/VCFv4.3.pdf`. visited on 2022-04-25.

[163] S. J. Sanders, A. G. Ercan-Sencicek, V. Hus, R. Luo, M. T. Murtha, D. Moreno-De-Luca, S. H. Chu, M. P. Moreau, A. R. Gupta, S. A. Thomson, et al. Multiple recurrent de novo CNVs, including duplications of the 7q11.23 Williams syndrome region, are strongly associated with autism. *Neuron*, 70(5):863–885, 2011.

[164] F. Sanger and A. R. Coulson. A rapid method for determining sequences in DNA by primed synthesis with dna polymerase. *Journal of Molecular Biology*, 94(3):441–448, 1975.

[165] F. Sanger, S. Nicklen, and A. R. Coulson. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences*, 74(12):5463–5467, 1977.

[166] S. D. Schrinner, R. Serra Mari, J. Ebler, M. Rautiainen, L. Seillier, J. J. Reimer, B. Usadel, T. Marschall, and G. W. Klau. Haplotype threading: accurate polyploid phasing from long reads. *Genome Biology*, 21(1):1–22, 2020.

[167] J. Sebat, B. Lakshmi, J. Troge, J. Alexander, J. Young, P. Lundin, S. Manér, H. Massa, M. Walker, M. Chi, et al. Large-scale copy number polymorphism in the human genome. *Science*, 305(5683):525–528, 2004.

[168] J. Sebat, B. Lakshmi, D. Malhotra, J. Troge, C. Lese-Martin, T. Walsh, B. Yamrom, S. Yoon, A. Krasnitz, J. Kendall, et al. Strong association of de novo copy number mutations with autism. *Science*, 316(5823):445–449, 2007.

[169] F. J. Sedlazeck, P. Rescheneder, M. Smolka, H. Fang, M. Nattestad, A. Von Haeseler, and M. C. Schatz. Accurate detection of complex structural variations using single-molecule sequencing. *Nature Methods*, 15(6):461–468, 2018.

[170] R. Serra Mari, S. Schrinner, R. Finkers, P. Arens, M. H.-W. Schmidt, B. Usadel, G. W. Klau, and T. Marschall. Haplotype-resolved assembly of a tetraploid potato genome using long reads and low-depth offspring data. *bioRxiv*, 2022.

[171] K. Shafin, T. Pesout, P.-C. Chang, M. Nattestad, A. Kolesnikov, S. Goel, G. Baid, M. Kolmogorov, J. M. Eizenga, K. H. Miga, et al. Haplotype-aware variant calling with PEPPER-Margin-DeepVariant enables high accuracy in nanopore long-reads. *Nature Methods*, 18(11): 1322–1332, 2021.

[172] A. Shajii, D. Yorukoglu, Y. William Yu, and B. Berger. Fast genotyping of known SNPs through approximate k-mer matching. *Bioinformatics*, 32(17):i538–i544, 2016.

[173] J. Shendure, S. Balasubramanian, G. M. Church, W. Gilbert, J. Rogers, J. A. Schloss, and R. H. Waterston. DNA sequencing at 40: past, present and future. *Nature*, 550(7676):345–353, 2017.

[174] S. T. Sherry, M.-H. Ward, M. Kholodov, J. Baker, L. Phan, E. M. Smigielski, and K. Sirotkin. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Research*, 29(1):308–311, 2001.

[175] J. A. Sibbesen, L. Maretty, and A. Krogh. Accurate genotyping across variant classes and lengths using variant graphs. *Nature Genetics*, 50(7):1054, 2018.

[176] J. Sirén, J. Monlong, X. Chang, A. M. Novak, J. M. Eizenga, C. Markello, J. A. Sibbesen, G. Hickey, P.-C. Chang, A. Carroll, et al. Pangenomics enables genotyping of known structural variants in 5202 diverse genomes. *Science*, 374(6574):abg8871, 2021.

[177] B. E. Slatko, A. F. Gardner, and F. M. Ausubel. Overview of next-generation sequencing technologies. *Current Protocols in Molecular Biology*, 122(1):e59, 2018.

[178] T. F. Smith, M. S. Waterman, et al. Identification of common molecular subsequences. *Journal of Molecular Biology*, 147(1):195–197, 1981.

[179] M. W. Snyder, A. Adey, J. O. Kitzman, and J. Shendure. Haplotype-resolved genome sequencing: experimental methods and applications. *Nature Reviews Genetics*, 16(6):344–358, 2015.

[180] A. H. Sturtevant. Genetic factors affecting the strength of linkage in Drosophila. *Proceedings of the National Academy of Sciences of the United States of America*, 3(9):555, 1917.

[181] P. H. Sudmant, T. Rausch, E. J. Gardner, R. E. Handsaker, A. Abyzov, J. Huddleston, Y. Zhang, K. Ye, G. Jun, M. Hsi-Yang Fritz, et al. An integrated map of structural variation in 2,504 human genomes. *Nature*, 526(7571):75–81, 2015.

[182] C. Sun and P. Medvedev. Toward fast and accurate SNP genotyping from whole genome sequencing data for bedside diagnostics. *Bioinformatics*, 35(3):415–420, 2019.

[183] H. Sun, W.-B. Jiao, K. Krause, J. A. Campoy, M. Goel, K. Folz-Donahue, C. Kukat, B. Huettel, and K. Schneeberger. Chromosome-scale and haplotype-resolved genome assembly of a tetraploid potato cultivar. *Nature Genetics*, 54(3):342–348, 2022.

[184] T2T Consortium. CHM13 Cell Line. URL `https://sites.google.com/ucsc.edu/t2tworkinggroup/chm13-cell-line?authuser=0`. visited on 2022-10-20.

[185] J. L. Taylor-Cousar, M. A. Zariwala, L. H. Burch, R. G. Pace, M. L. Drumm, H. Calloway, H. Fan, B. W. Weston, F. A. Wright, M. R. Knowles, et al. Histo-blood group gene polymorphisms as potential genetic modifiers of infection and cystic fibrosis lung disease severity. *PloS One*, 4 (1):e4270, 2009.

[186] R. Tewhey, V. Bansal, A. Torkamani, E. J. Topol, and N. J. Schork. The importance of phase information for human genomics. *Nature Reviews Genetics*, 12(3):215–223, 2011.

[187] The Computational Pan-Genomics Consortium. Computational pan-genomics: status, promises and challenges. *Briefings in Bioinformatics*, 19(1):118–135, 2018.

[188] The Genome of the Netherlands Consortium. Whole-genome sequence variation, population structure and demographic history of the dutch population. *Nature Genetics*, 46(8):818–825, 2014.

[189] I. Turner, K. V. Garimella, Z. Iqbal, and G. McVean. Integrating long-range connectivity information into de Bruijn graphs. *Bioinformatics*, 34(15):2556–2565, 2018.

[190] E. Ukkonen. Finding founder sequences from a set of recombinants. In *International Workshop on Algorithms in Bioinformatics*, pages 277–286. Springer, 2002.

[191] J. C. Venter, M. D. Adams, E. W. Myers, P. W. Li, R. J. Mural, G. G. Sutton, H. O. Smith, M. Yandell, C. A. Evans, R. A. Holt, et al. The sequence of the human genome. *Science*, 291 (5507):1304–1351, 2001.

[192] J. Wagner, N. D. Olson, L. Harris, J. McDaniel, H. Cheng, A. Fungtammasan, Y.-C. Hwang, R. Gupta, A. M. Wenger, W. J. Rowell, et al. Curated variation benchmarks for challenging medically relevant autosomal genes. *Nature Biotechnology*, pages 1–9, 2022.

[193] T. Walsh, J. M. McClellan, S. E. McCarthy, A. M. Addington, S. B. Pierce, G. M. Cooper, A. S. Nord, M. Kusenda, D. Malhotra, A. Bhandari, et al. Rare structural variants disrupt multiple genes in neurodevelopmental pathways in schizophrenia. *Science*, 320(5875):539–543, 2008.

[194] J. Wang, L. Raskin, D. C. Samuels, Y. Shyr, and Y. Guo. Genome measures used for quality control are dependent on gene function and ancestry. *Bioinformatics*, 31(3):318–323, 2014.

[195] J. Wang, D. C. Samuels, Y. Shyr, and Y. Guo. Population structure analysis on 2504 individuals across 26 ancestries using bioinformatics approaches. *BMC Bioinformatics*, 16(15):P19, 2015.

[196] K. Watanabe. Potato genetics, genomics, and applications. *Breeding Science*, 65(1):53–68, 2015.

[197] J. D. Watson and F. H. Crick. Molecular structure of nucleic acids: a structure for deoxyribose nucleic acid. *Nature*, 171(4356):737–738, 1953.

[198] J. Weischenfeldt, O. Symmons, F. Spitz, and J. O. Korbel. Phenotypic impact of genomic structural variation: insights from and for human disease. *Nature Reviews Genetics*, 14(2):125, 2013.

[199] A. M. Wenger, P. Peluso, W. J. Rowell, P.-C. Chang, R. J. Hall, G. T. Concepcion, J. Ebler, A. Fungtammasan, A. Kolesnikov, N. D. Olson, et al. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nature Biotechnology*, 37(10):1155–1162, 2019.

[200] D. L. Wheeler, D. M. Church, S. Federhen, A. E. Lash, T. L. Madden, J. U. Pontius, G. D. Schuler, L. M. Schriml, E. Sequeira, T. A. Tatusova, et al. Database resources of the National Center for Biotechnology. *Nucleic Acids Research*, 31(1):28–33, 2003.

[201] N. M. Williams, I. Zaharieva, A. Martin, K. Langley, K. Mantripragada, R. Fossdal, H. Stefansson, K. Stefansson, P. Magnusson, O. O. Gudmundsson, et al. Rare chromosomal deletions and duplications in attention-deficit hyperactivity disorder: a genome-wide analysis. *The Lancet*, 376(9750):1401–1408, 2010.

[202] M. Xie, Q. Wu, J. Wang, and T. Jiang. H-PoP and H-PoPG: heuristic partitioning algorithms for single individual haplotyping of polyploids. *Bioinformatics*, 32(24):3735–3744, 2016.

[203] C. Zahn. Approximating symmetric relations by equivalence relations. *Journal of the Society for Industrial & Applied Mathematics*, 12, 1964. doi: 10.1137/0112071.

[204] X. Zhang, R. Wu, Y. Wang, J. Yu, and H. Tang. Unzipping haplotypes in diploid and polyploid genomes. *Computational and Structural Biotechnology Journal*, 18:66–72, 2020.

[205] X. Zhao, A. M. Weber, and R. E. Mills. A recurrence-based approach for validating structural variation using long-read sequencing technology. *Gigascience*, 6(8):gix061, 2017.

[206] X. Zhao, R. L. Collins, W.-P. Lee, A. M. Weber, Y. Jun, Q. Zhu, B. Weisburd, Y. Huang, P. A. Audano, H. Wang, et al. Expectations and blind spots for structural variation detection from long-read assemblies and short-read genome sequencing technologies. *The American Journal of Human Genetics*, 108(5):919–928, 2021.

[207] J. M. Zook, D. Catoe, J. McDaniel, L. Vang, N. Spies, A. Sidow, Z. Weng, Y. Liu, C. E. Mason, N. Alexander, et al. Extensive sequencing of seven human genomes to characterize benchmark reference materials. *Scientific Data*, 3(1):1–26, 2016.

[208] J. M. Zook, J. McDaniel, N. D. Olson, J. Wagner, H. Parikh, H. Heaton, S. A. Irvine, L. Trigg, R. Truty, C. Y. McLean, et al. An open resource for accurately benchmarking small variant and reference calls. *Nature Biotechnology*, 37(5):561–566, 2019.

Appendix A

Appendices: Application and advances in haplotype phasing

A.1 Accurate polyploid phasing from long reads

coverage	method	SER (%)	HR (%)	N50 (bp)	runtime (s)	memory (GB)
40×	WH-PP	0.58	1.48	29529	3333	1.41
	WH-PP*	1.39	28.72	1692352	3433	1.42
	H-PoPG	2.01	27.53	1785293	2230	9.97
80×	WH-PP	0.31	1.43	54434	12694	2.52
	WH-PP*	0.74	28.27	2587104	13042	2.89
	H-PoPG	1.24	27.66	2587104	4368	9.99

(a) real tetraploid read data

coverage	method	SER (%)	HR (%)	N50 (bp)	runtime (s)	memory (GB)
40×	WH-PP	0.42	1.74	48815	1960	1.10
	WH-PP*	1.00	26.57	1830943	2004	1.17
	H-PoPG	1.67	26.37	1917094	1414	9.96
80×	WH-PP	0.29	2.51	86227	5738	1.78
	WH-PP*	0.68	25.23	2142893	5865	2.04
	H-PoPG	0.98	25.65	2142893	2843	9.97

(b) simulated tetraploid read data

coverage	method	SER (%)	HR (%)	N50 (bp)	runtime (s)	memory (GB)
40×	WH-PP	0.86	1.57	22625	2331	1.05
	WH-PP*	2.01	25.34	1361459	2377	1.07
	H-PoPG	3.50	24.78	1453040	2357	9.97
80×	WH-PP	0.47	1.18	33438	5031	1.69
	WH-PP*	1.33	23.64	1701753	5118	1.87
	H-PoPG	2.24	24.76	1748404	4849	9.96

(c) simulated pentaploid read data

coverage	method	SER (%)	HR (%)	N50 (bp)	runtime (s)	memory (GB)
40×	WH-PP	1.12	1.82	16785	25841	1.30
	WH-PP*	2.35	27.03	3877456	25860	1.79
	H-PoPG	3.85	26.75	4490129	5450	9.96
80×	WH-PP	0.48	0.97	26711	10331	1.98
	WH-PP*	1.34	25.63	4540968	10827	2.63
	H-PoPG	2.37	25.93	4721421	11563	10.89

(d) simulated hexaploid read data

Table A.1: Phasing evaluation on artificial polyploid human. Comparison of `whatshap polyphase` and H-PoPG on tetraploid real (a) and simulated (b) datasets, pentaploid simulated dataset (c) and hexaploid simulated dataset (d). Performances are based on the switch error rate (SER), block-wise Hamming rate (HR) and N50 for the block size. For better comparability with H-PoPG, a second setting (WH-PP*) with less block-cuts was used in addition to the default block cut strategy (WH-PP). The total length of the chromosome is 249 Mb. Table taken from [166].

coverage	method	collapsing regions	non-collapsing regions	total
40×	WH-PP*	0.66	1.81	1.65
	H-PoPG	2.02	2.16	2.02
	$\mathrm{SER}(\frac{H-PoPG}{WH-PP*})$	3.06	1.19	1.22
80×	WH-PP*	0.38	1.16	0.99
	H-PoPG	1.05	1.30	1.24
	$\mathrm{SER}(\frac{H-PoPG}{WH-PP*})$	2.76	1.12	1.25

(a) real read data

coverage	method	collapsing regions	non-collapsing regions	total
40×	WH-PP*	0.45	1.29	1.19
	H-PoPG	2.01	1.63	1.68
	$\mathrm{SER}(\frac{H-PoPG}{WH-PP*})$	4.47	1.62	1.41
80×	WH-PP*	0.25	0.88	0.82
	H-PoPG	0.94	0.98	0.99
	$\mathrm{SER}(\frac{H-PoPG}{WH-PP*})$	3.76	1.11	1.21

(b) simulated read data

Table A.2: Phasing evaluation in/outside collapsing regions. Comparison of resulting switch error rates of H-PoPG and `whatshap polyphase` using block lengths that are comparable to H-PoPG (WH-PP*) on collapsing regions over at least 50 variants as compared to non-collapsing regions and the average throughout the genome. Results (switch error rates in %) are presented for chromsome 1 of the real (a) and the simulated (b) dataset, testing 40× and 80× coverage. The third row marks the quotient between the switch error rate of H-PoPG and that of `whatshap polyphase` to highlight by which magnitude the results differ. Table taken from [166].

Appendix B

Appendices: PanGenie: Pangenome-based genome inference

type	variants (unfiltered)	variants (callable regions)	variants (mendelian consistent)	bubbles in pangenome graph
SNP	13,628,117	12,560,841	12,095,177	11,556,580
small insertion	2,229,474	2,163,433	1,922,163	810,298
small deletion	2,026,998	1,961,042	1,811,123	819,445
small complex	0	0	0	597,044
midsize insertion	123,304	120,505	110,882	20,300
midsize deletion	87,263	85,114	80,027	12,720
midsize complex	0	0	0	87,392
large insertion	135,150	123,990	108,929	18,325
large deletion	48,724	45,419	41,499	4,397
large complex	0	0	0	52,272

Table B.1: Variant calling statistics. Numbers of variants obtained at different stages of variant calling/pangenome graph construction. The first column corresponds to the number of raw variant calls made across all individual haplotypes. The second column contains the number of variants within the callable regions, that is, after removing sites with more than 20% of missing (“./.”) genotypes. The third column shows the number of variants left after removing sites with Mendelian inconsistencies and corresponds to our final variant callset. The last column presents the number of bubbles in the graph after constructing a pangenome from all variants in the previous column. Columns 1-3 contain only variant alleles that can be classified as SNPs, insertions and deletions. In the graph however, overlapping variant alleles are combined into multiallelic bubbles. All such bubbles with more than two branches are defined as “complex”. Table taken from [49].

		SNP	small INS	small DEL	midsize INS	midsize DEL	large INS	large DEL
HG00512	total	3,724,332	444,561	442,280	15,635	14,685	14,742	8,623
	unique	251,925	48,811	42,229	5,531	2,831	5,830	1,412
HG00513	total	3,767,798	447,734	444,429	16,025	15,019	15,098	8,900
	unique	258,979	46,955	39,971	5,758	2,978	6,191	1,540
HG00731	total	3,792,925	446,416	448,667	15,630	15,002	15,006	8,717
	unique	237,125	43,607	38,016	5,299	2,585	5,645	1,336
HG00732	total	3,850,476	552,952	469,341	16,943	15,401	15,882	9,082
	unique	258,515	122,268	57,321	6,157	2,961	6,554	1,552
HG02818	total	4,604,971	559,626	566,604	20,166	18,904	17,837	10,740
	unique	605,439	100,220	97,157	8,881	5,455	8,820	2,710
HG03125	total	4,631,416	576,887	580,655	20,290	19,106	18,132	10,759
	unique	608,299	113,155	108,740	9,021	5,574	9,074	2,726
HG03486	total	4,679,604	582,370	575,677	20,421	19,430	18,376	11,027
	unique	670,228	118,933	106,370	9,146	5,935	9,364	2,891
NA12878	total	3,775,211	445,739	448,293	16,029	15,027	15,374	8,777
	unique	247,345	46,445	40,769	5,739	2,816	6,001	1,444
NA19238	total	4,629,589	562,928	584,771	20,099	19,081	18,321	10,870
	unique	606,621	102,902	108,613	8,803	5,496	9,066	2,771
NA19239	total	4,573,111	551,810	575,465	19,611	18,642	17,721	10,664
	unique	589,863	98,383	106,349	8,327	5,466	8,636	2,620
NA24385	total	3,761,904	459,907	447,880	16,144	15,046	15,023	8,769
	unique	247,111	56,090	43,103	5,710	2,792	5,921	1,370
total	total	12,095,177	1,922,163	1,811,123	110,882	80,027	108,929	41,499

Table B.2: Variants in pangenome graph. Total number of variants detected across all assembly samples ("total"), as well as the number of variants unique to a sample, that is, variants seen only in the respective sample and in none of the other samples ("unique"). Table taken from [49].

		non-repetitive regions	**repeat regions**	**repeat regions [%]**
SNPs	biallelic	10,736,632	527,498	4.7 %
	complex	179,476	368,385	67.2 %
small	biallelic INS	682,987	115,702	14.5%
	biallelic DEL	696,243	123,055	15.0 %
	complex INS+DEL	1,238,489	817,458	39.7 %
midsize	biallelic INS	9,313	10,997	54.2 %
	biallelic DEL	5,651	7,909	58.3 %
	complex INS + DEL	43,105	104,734	70.8 %
large	biallelic INS	7,537	10,757	58.8 %
	biallelic DEL	2,212	2,397	52.0 %
	complex INS + DEL	29,277	89,297	75.3 %

Table B.3: Number of variants in repetitive and non-repetitive regions. Shown are the numbers of variants located inside and outside of STR/VNTR regions for sample NA12878. "biallelic" corresponds to all genomic regions outside of complex bubbles (= bubbles with more than two branches) in our pangenome graph. "complex" corresponds to all callset variants that are located inside of complex bubbles. Table taken from [49].

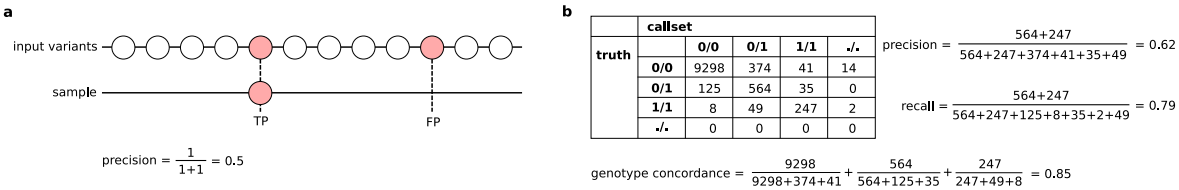


Figure B.1: Evaluation example. a Shown is a set of variants (represented by circles) used as input for genotyping and a sample carrying only one of these variants (i.e. the true genotype for all others is 0/0). Both variants colored in red were genotyped as non-0/0, resulting in a genotyping error for the second one which is counted as a false positive call. This example illustrates why precision/recall measures are not ideal for a re-genotyping scenario: when variants are called across a large set of samples and then genotyped in another, new sample, there will be a large fraction of variants absent in the new sample. Thus, even if the false positive rate is low, the number of false positives increases as the number of input variants increases, while the number of true positives is limited. In this example, only a single genotyping error decreases the precision to 0.5. **b** Confusion matrix for PanGenie and the resulting values for precision, recall and weighted genotype concordance (matrix for large insertions in repetitive regions). While the false positive rate is low (only 415 of 9713 variants absent from the truth set are genotyped as non-0/0), it is relatively high compared to the number of true positives which explains why the precision is so low compared to the weighted genotype concordance. The same effect applies to all re-genotyping methods. Figure taken from [49].

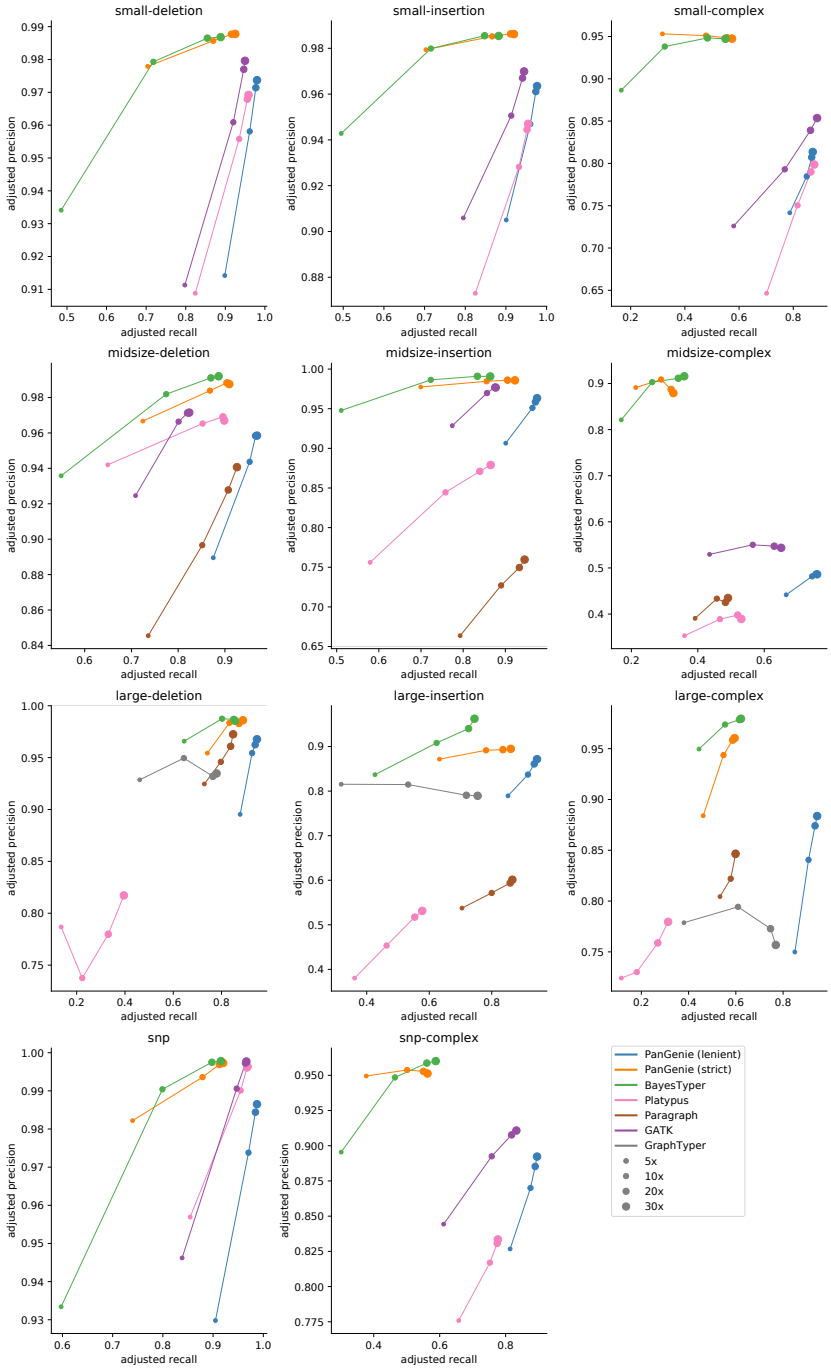


Figure B.2: Adjusted precision/recall for NA12878 (non-repetitive regions). Adjusted precision/recall at different coverages for sample NA12878. We ran PanGenie, BayesTyper, Paragraph, Platypus, GATK, GraphTyper and Giraffe in order to re-genotype all callset variants. Besides not applying any filter on the reported genotype qualities ("all"), we additionally report genotyping statistics for PanGenie when using "high-gq" filtering (genotype quality ≥ 200). SNPs, insertions and deletions include all respective variants in biallelic regions of the genome, while complex contains all variant alleles falling into regions with complex bubbles in the pangenome graph representation. Figure taken from [49].

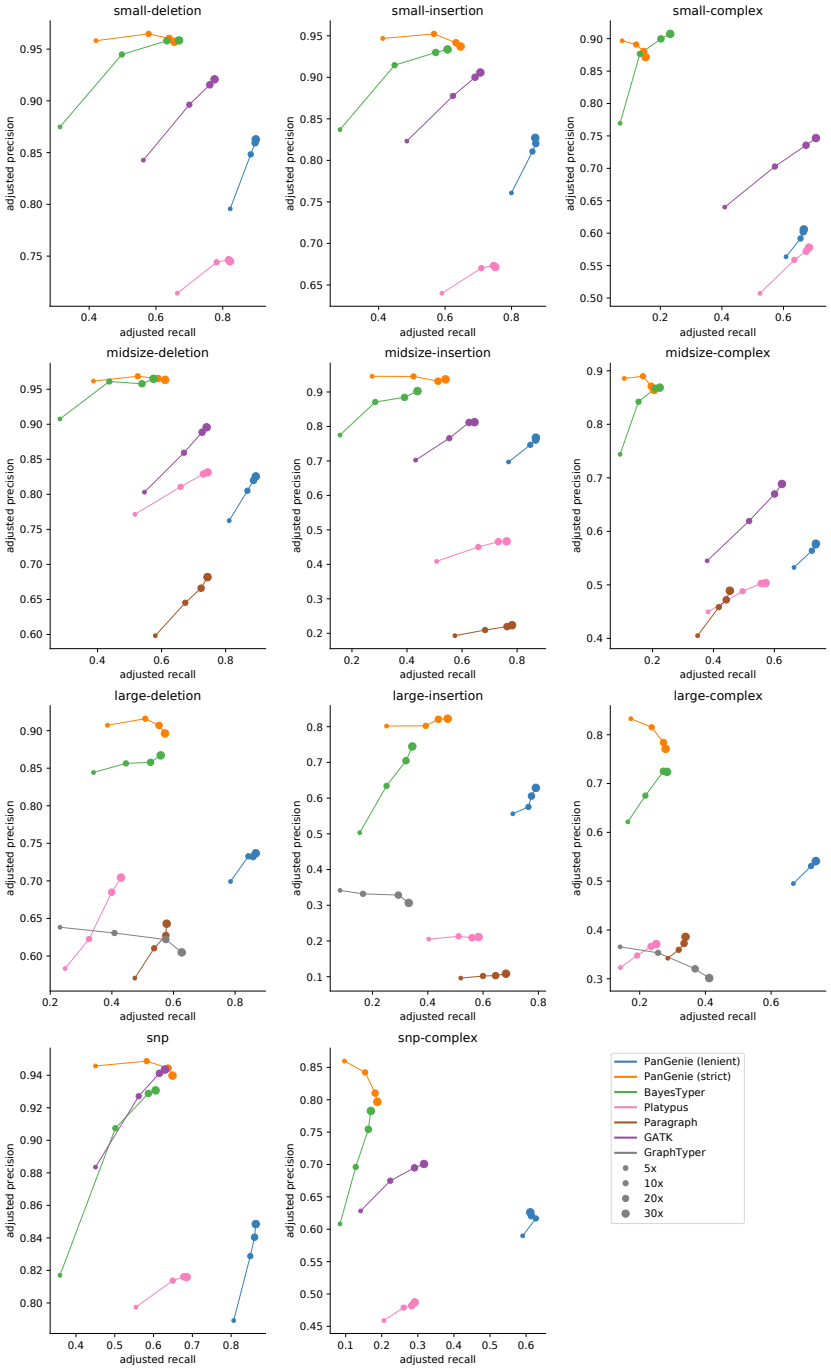


Figure B.3: Adjusted precision/recall for NA12878 (STR/VNTR regions). Adjusted precision/recall at different coverages for sample NA12878. We ran PanGenie, BayesTyper, Paragraph, Platypus, GATK, GraphTyper and Giraffe in order to re-genotype all callset variants. Besides not applying any filter on the reported genotype qualities ("all"), we additionally report genotyping statistics for PanGenie when using "high-gq" filtering (genotype quality ≥ 200). SNPs, insertions and deletions include all respective variants in biallelic regions of the genome, while complex contains all variant alleles falling into regions with complex bubbles in the pangenome graph representation. Figure taken from [49].

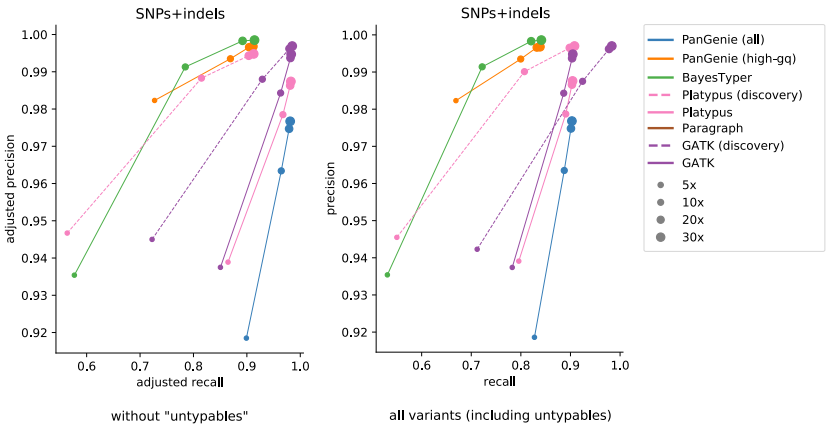


Figure B.4: Comparison to GIAB small variants for NA12878. The GIAB small variants benchmark set [208] was used as ground truth for evaluating the results of our "leave-one-out" experiment for SNPs and indels (< 50bp). We computed the adjusted precision and recall (left), as well as the un-adjusted versions (right) including variants unique to NA12878 and thus not genotypable by a re-genotyping approach. GATK and Platypus were additionally run in detection mode. Figure taken from [49].

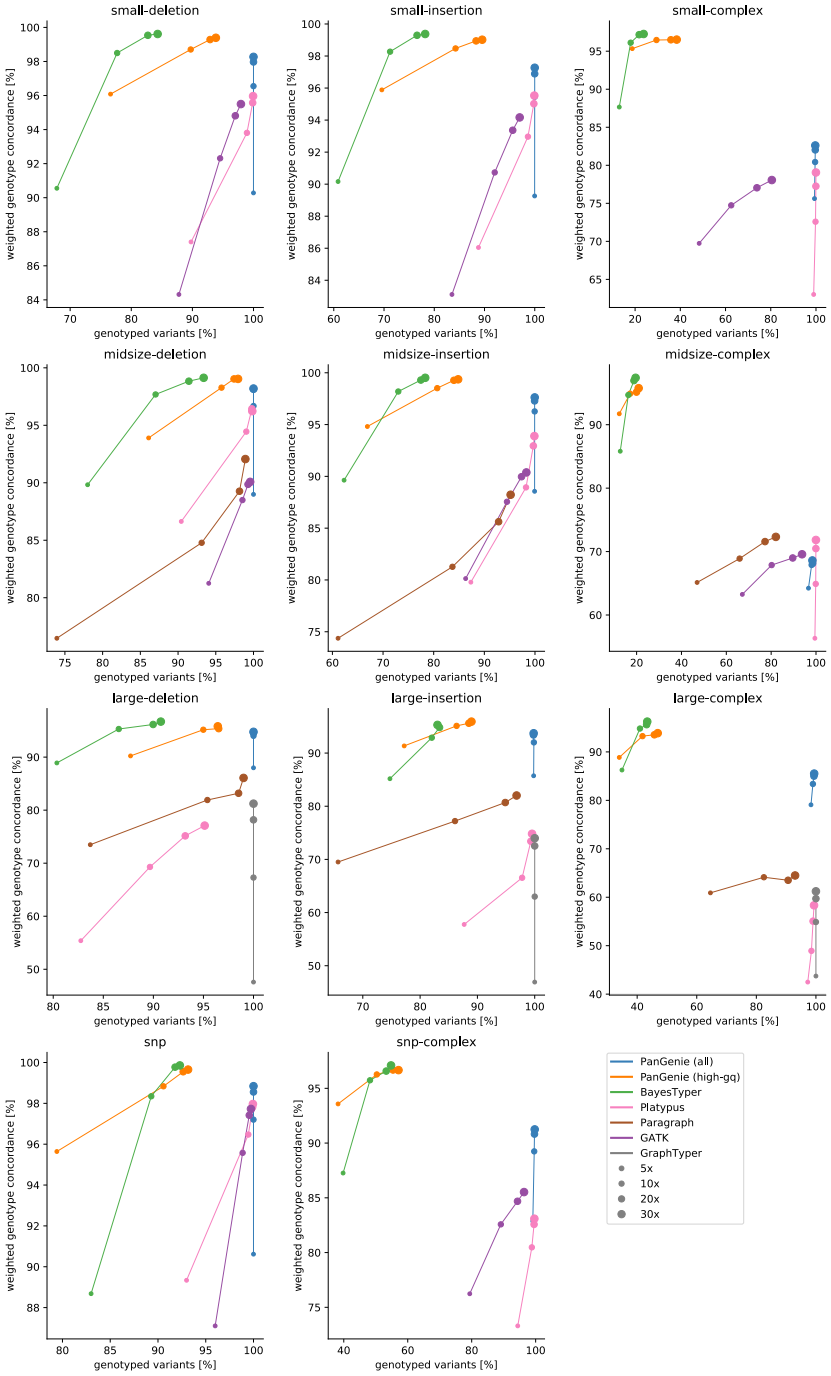


Figure B.5: Weighted genotype concordance for NA24385 (non-repetitive regions). Weighted genotype concordance at different coverages for sample NA24385. We ran PanGenie, BayesTyper, Paragraph, Platypus, GATK, GraphTyper and Giraffe in order to re-genotype all callset variants. Besides not applying any filter on the reported genotype qualities ("all"), we additionally report genotyping statistics for PanGenie when using "high-gq" filtering (genotype quality ≥ 200). SNPs, insertions and deletions include all respective variants in biallelic regions of the genome, while *complex* contains all variant alleles falling into regions with complex bubbles in the pangenome graph representation. Figure taken from [49].

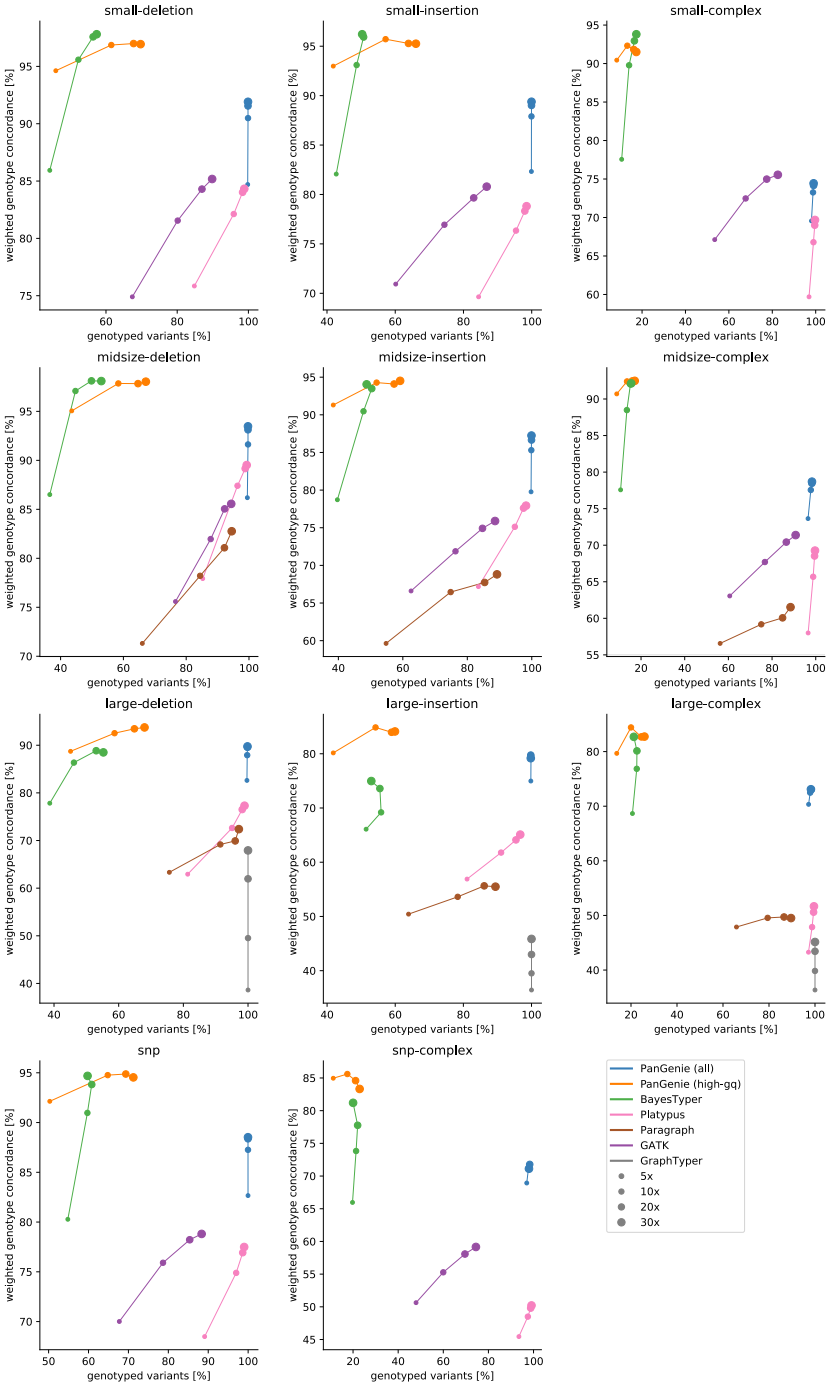


Figure B.6: Weighted genotype concordance for NA24385 (STR/VNTR regions). Weighted genotype concordance at different coverages for sample NA24385. We ran PanGenie, BayesTyper, Paragraph, Platypus, GATK, GraphTyper and Giraffe in order to regenotype all callset variants. Besides not applying any filter on the reported genotype qualities ("all"), we additionally report genotyping statistics for PanGenie when using "high-gq" filtering (genotype quality ≥ 200). SNPs, insertions and deletions include all respective variants in biallelic regions of the genome, while *complex* contains all variant alleles falling into regions with complex bubbles in the pangenome graph representation. Figure taken from [49].

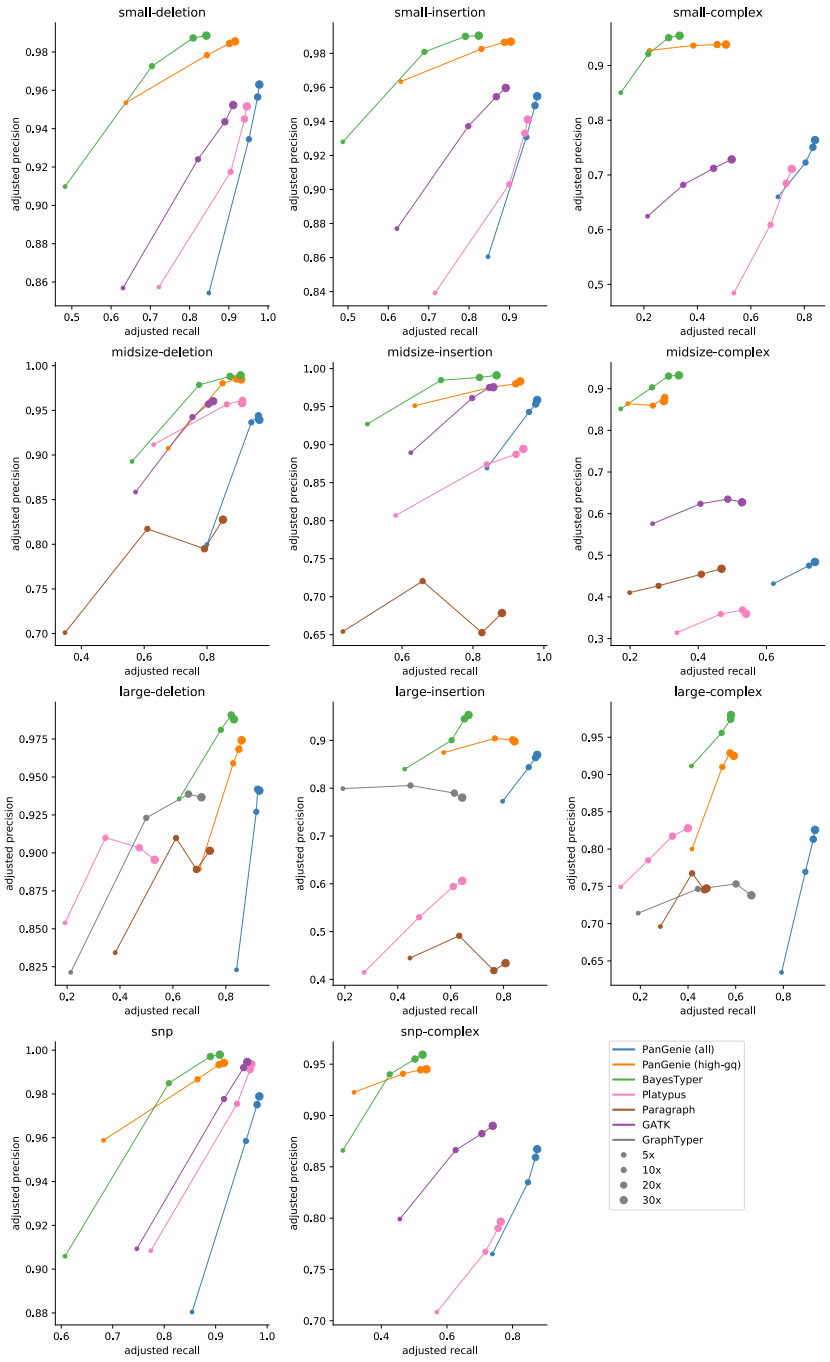


Figure B.7: Adjusted precision/recall for NA24385 (non-repetitive regions). Adjusted precision/recall at different coverages for sample NA24385. We ran PanGenie, BayesTyper, Paragraph, Platypus, GATK, GraphTyper and Giraffe in order to re-genotype all callset variants. Besides not applying any filter on the reported genotype qualities ("all"), we additionally report genotyping statistics for PanGenie when using "high-gq" filtering (genotype quality ≥ 200). SNPs, insertions and deletions include all respective variants in biallelic regions of the genome, while *complex* contains all variant alleles falling into regions with complex bubbles in the pangenome graph representation. Figure taken from [49].

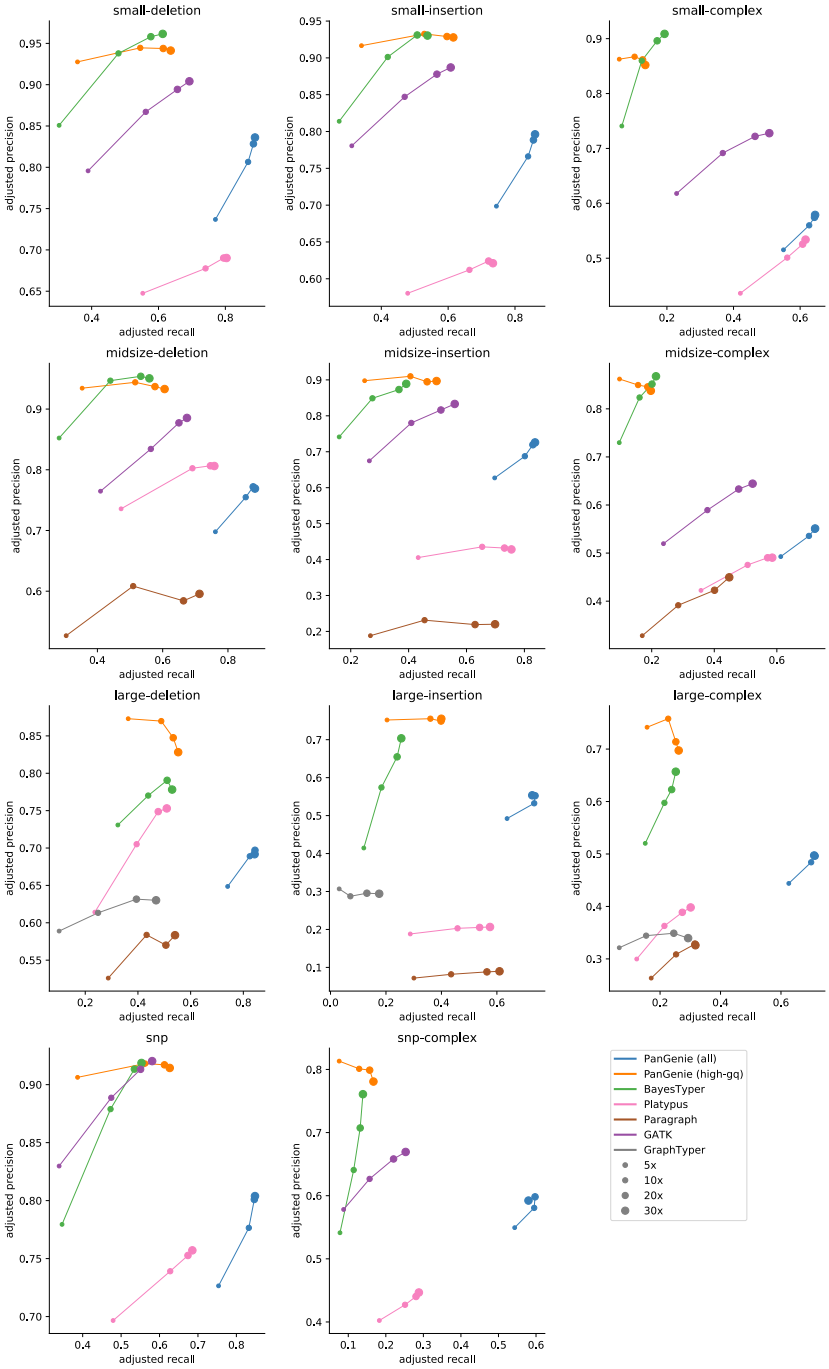


Figure B.8: Adjusted precision/recall for NA24385 (STR/VNTR regions). Adjusted precision/recall at different coverages for sample NA24385. We ran PanGenie, BayesTyper, Paragraph, Platypus, GATK, GraphTyper and Giraffe in order to re-genotype all callset variants. Besides not applying any filter on the reported genotype qualities ("all"), we additionally report genotyping statistics for PanGenie when using "high-gq" filtering (genotype quality ≥ 200). SNPs, insertions and deletions include all respective variants in biallelic regions of the genome, while *complex* contains all variant alleles falling into regions with complex bubbles in the pangenome graph representation. Figure taken from [49].

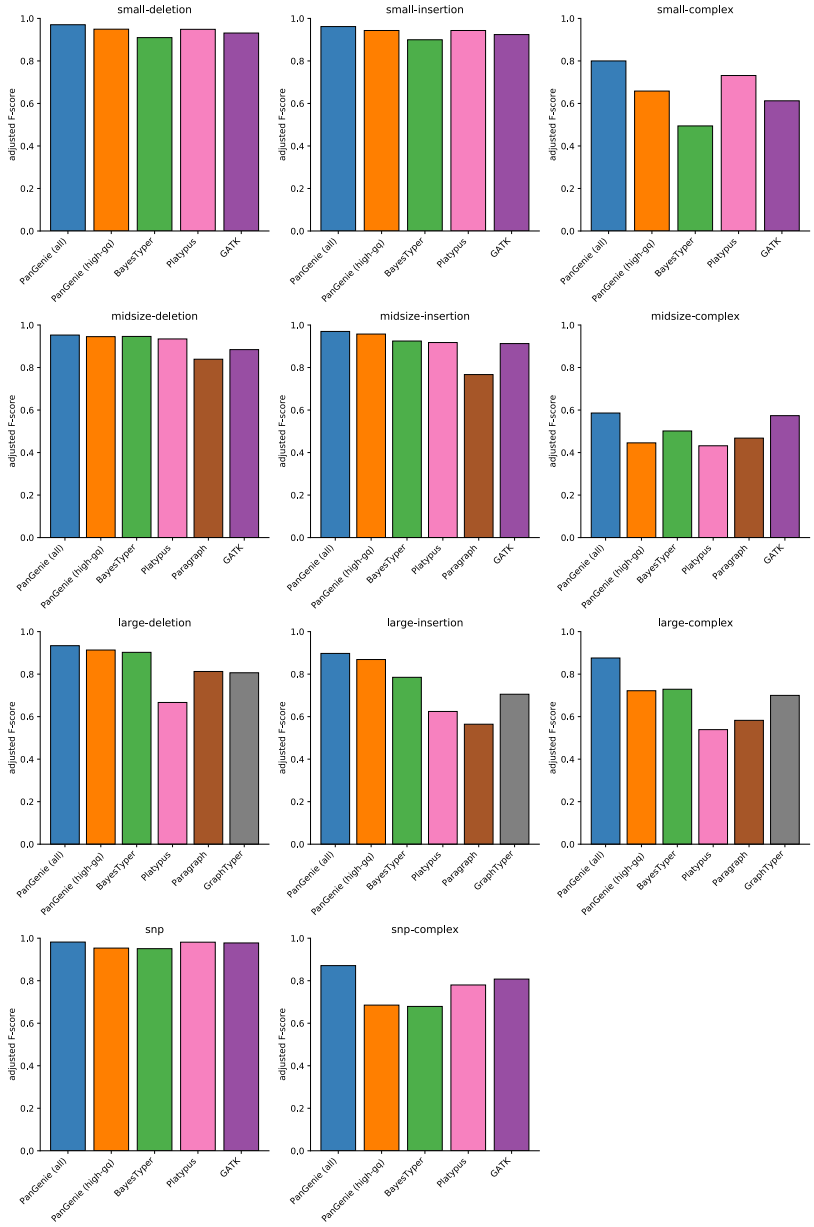


Figure B.9: Adjusted F-score for NA24385 (non-repetitive regions). Adjusted F-score at coverage $30\times$ for sample NA24385. We ran PanGenie, BayesTyper, Paragraph, Platypus, GATK, GraphTyper and Giraffe in order to re-genotype all callset variants. Besides not applying any filter on the reported genotype qualities ("all"), we additionally report genotyping statistics for PanGenie when using "high-gq" filtering (genotype quality ≥ 200). SNPs, insertions and deletions include all respective variants in biallelic regions of the genome, while *complex* contains all variant alleles falling into regions with complex bubbles in the pangenome graph representation. Figure taken from [49].

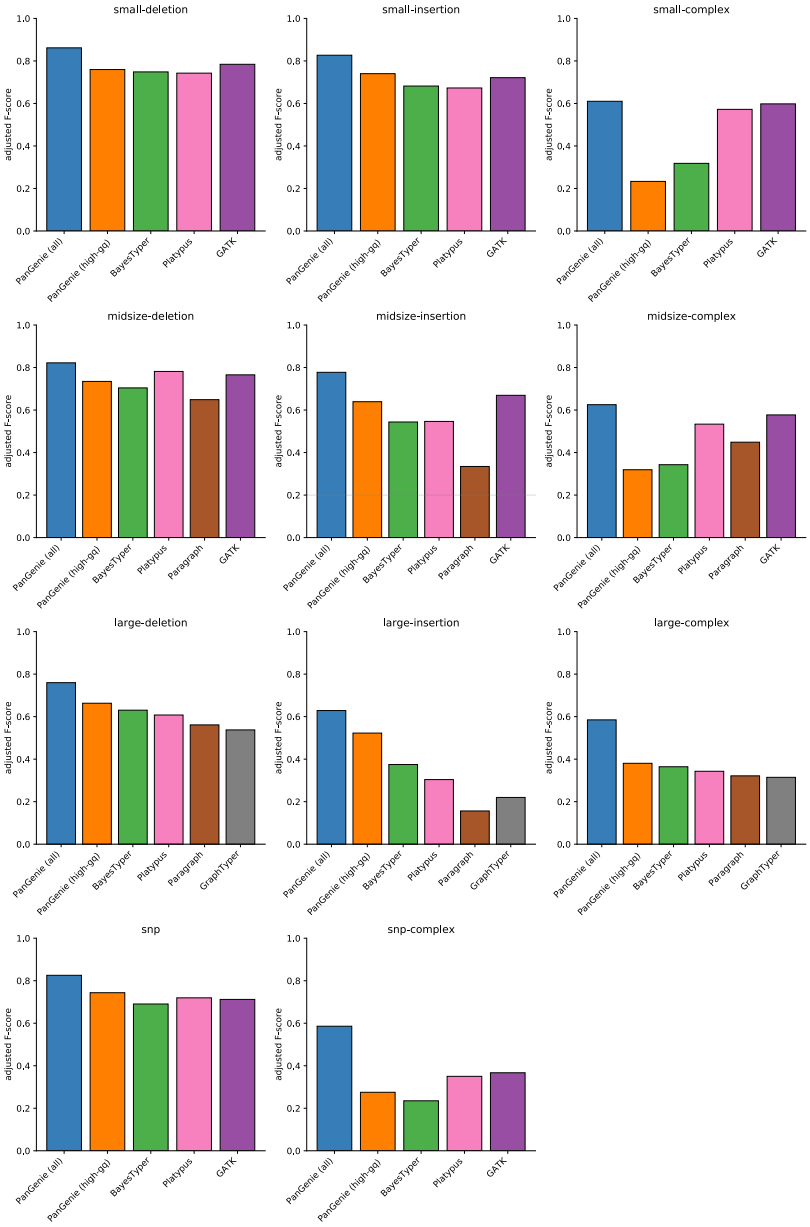


Figure B.10: Adjusted F-score for NA24385 (STR/VNTR regions). Adjusted F-score at coverage $30\times$ for sample NA24385. We ran PanGenie, BayesTyper, Paragraph, Platypus, GATK, GraphTyper and Giraffe in order to re-genotype all callset variants. Besides not applying any filter on the reported genotype qualities ("all"), we additionally report genotyping statistics for PanGenie when using "high-gq" filtering (genotype quality ≥ 200). SNPs, insertions and deletions include all respective variants in biallelic regions of the genome, while *complex* contains all variant alleles falling into regions with complex bubbles in the pangenome graph representation. Figure taken from [49].

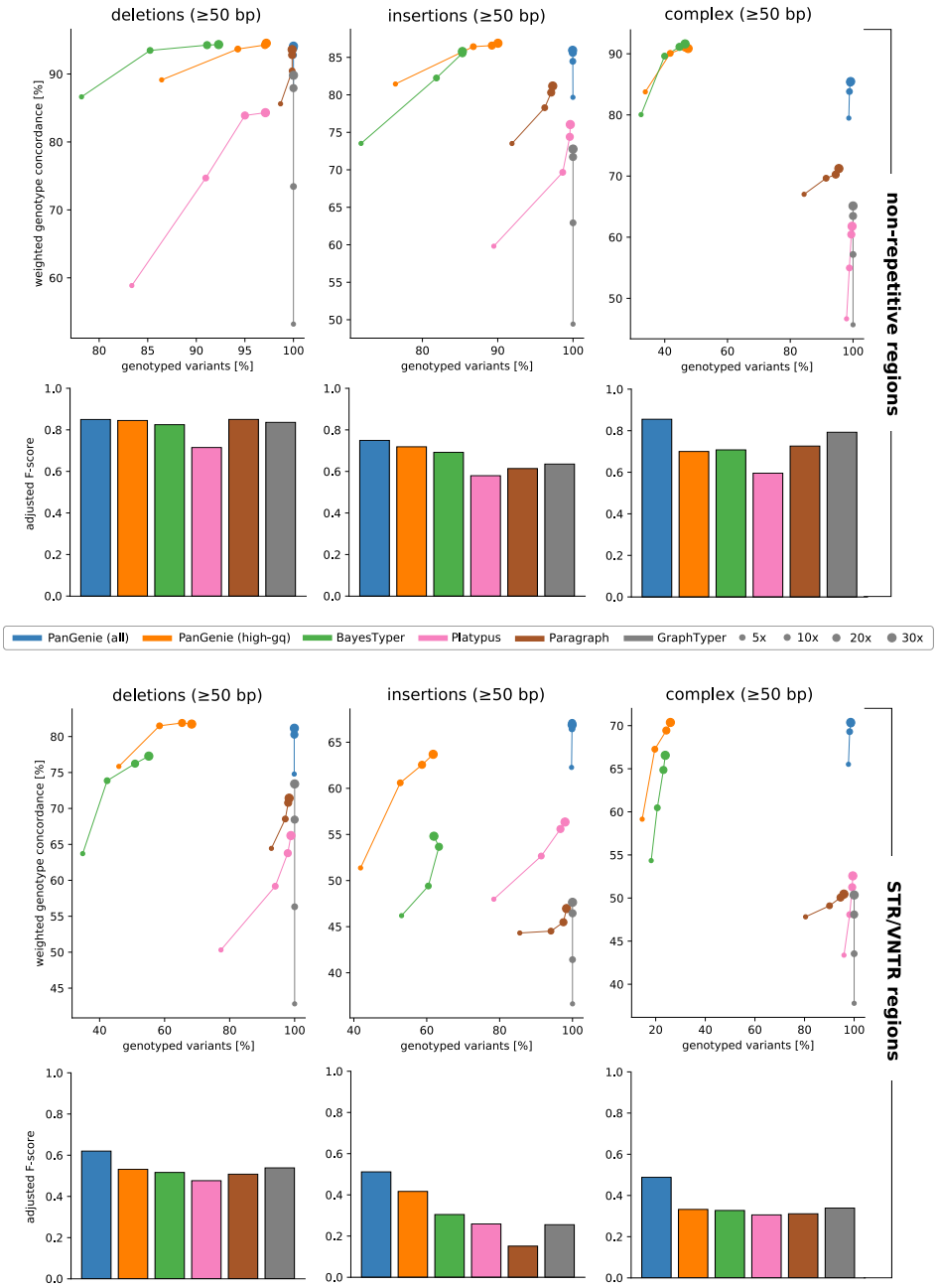


Figure B.11: Comparison to syndip benchmark SVs. SVs contained in the syndip benchmark set were used as ground truth for evaluation. We computed the weighted genotype concordance and the adjusted precision and recall metrics to evaluate our results. Figure taken from [49].

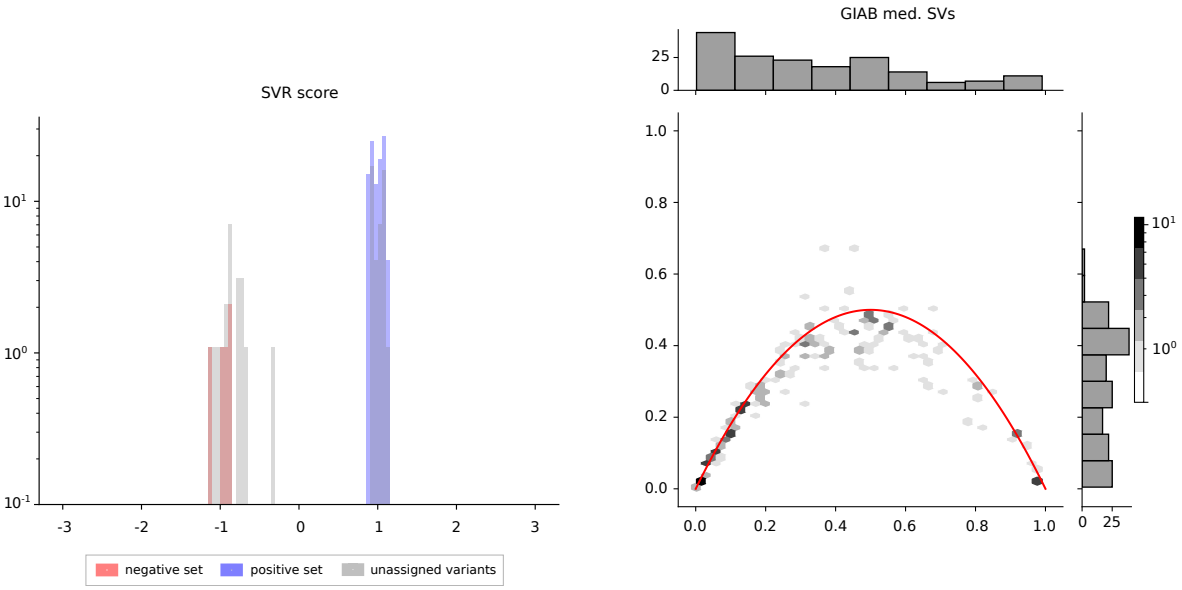


Figure B.12: GIAB medically relevant SVs in our lenient set. Distribution of SVR scores for all 209 GIAB medically relevant genes that are part of our variant callset (left), as well as heterozygosities and allele frequencies observed across all 200 unrelated trio samples in our lenient set (right). Figure taken from [49].

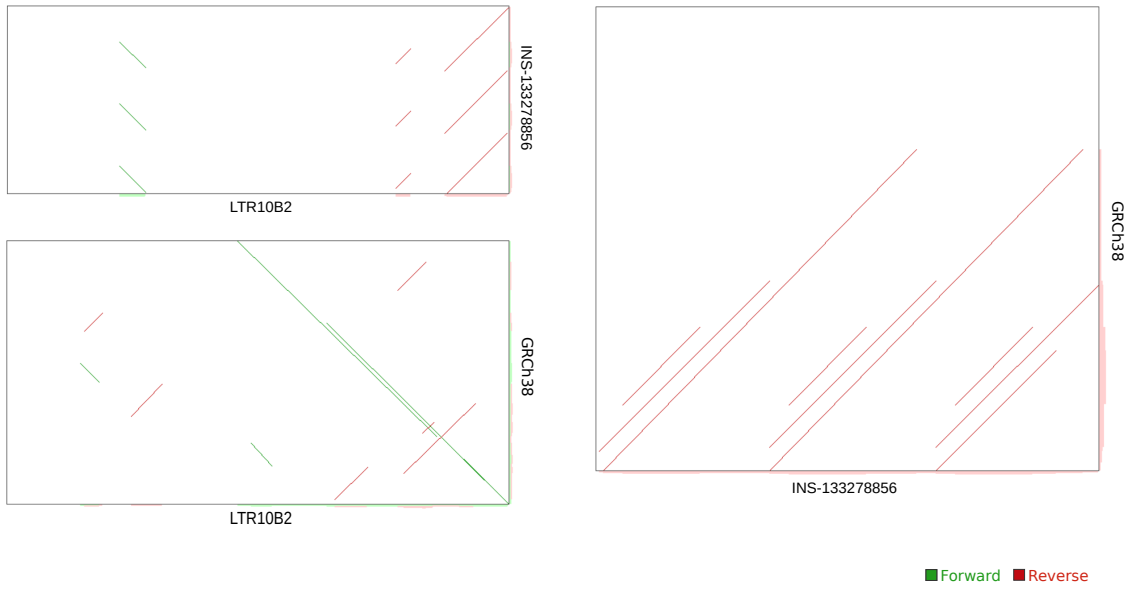


Figure B.13: LD analysis for ABO. Pairwise dot plots of the insertion sequence, LTR10B2 consensus sequence and the reference sequence of this region (GRCh38). Figure taken from [49].

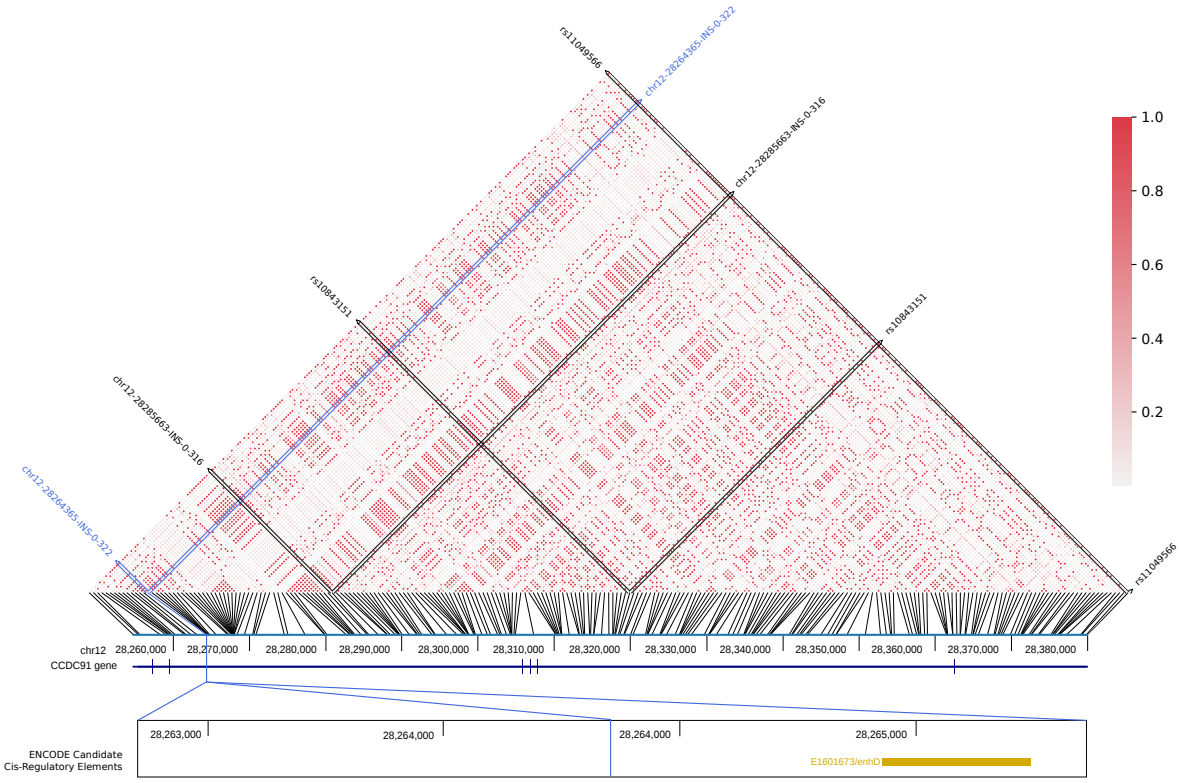


Figure B.14: LD analysis for CCDC91. The plot shows an insertion (marked in blue) in the CCDC91 gene that is in linkage disequilibrium with two GWAS SNPs (rs10843151 and rs11049566). The plot shows all callset variants with AF≥ 0.05 in this region, GWAS variants are annotated with their name. Figure taken from [49].

Appendix C

Appendices: Application: genotyping large cohorts

C.1 HGSVC project

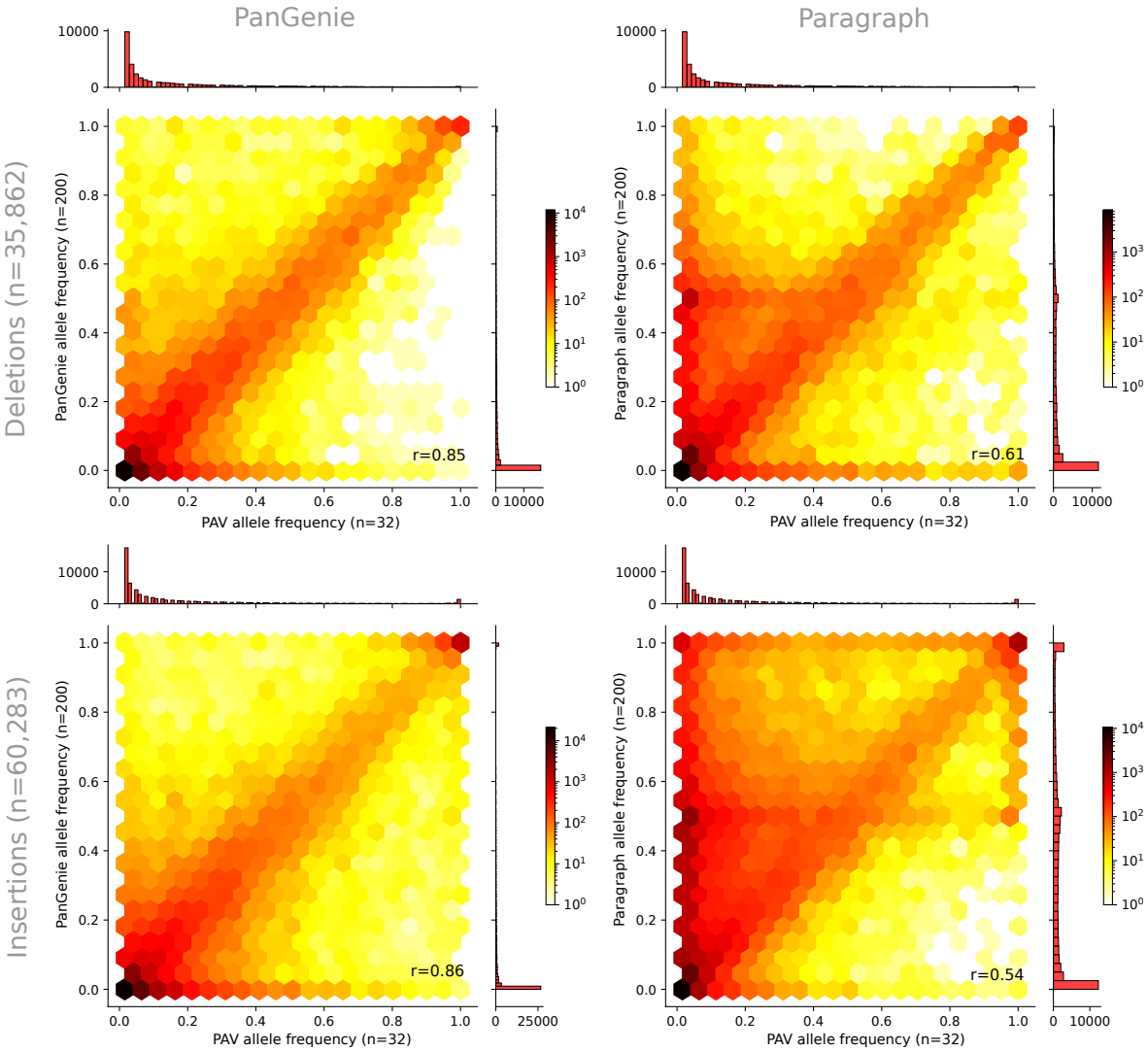


Figure C.1: Comparison PanGenie and Paragraph on HGSVC calls. Paragraph and PanGenie were run on a subset of 100 trios (300 samples) in order to derive genotypes for all SVs (n = 96,145). Allele frequencies were computed based on the genotypes of both methods for all 200 unrelated samples. PAV allele frequencies were computed based on all 64 assembly haplotypes. Figure taken from [46].

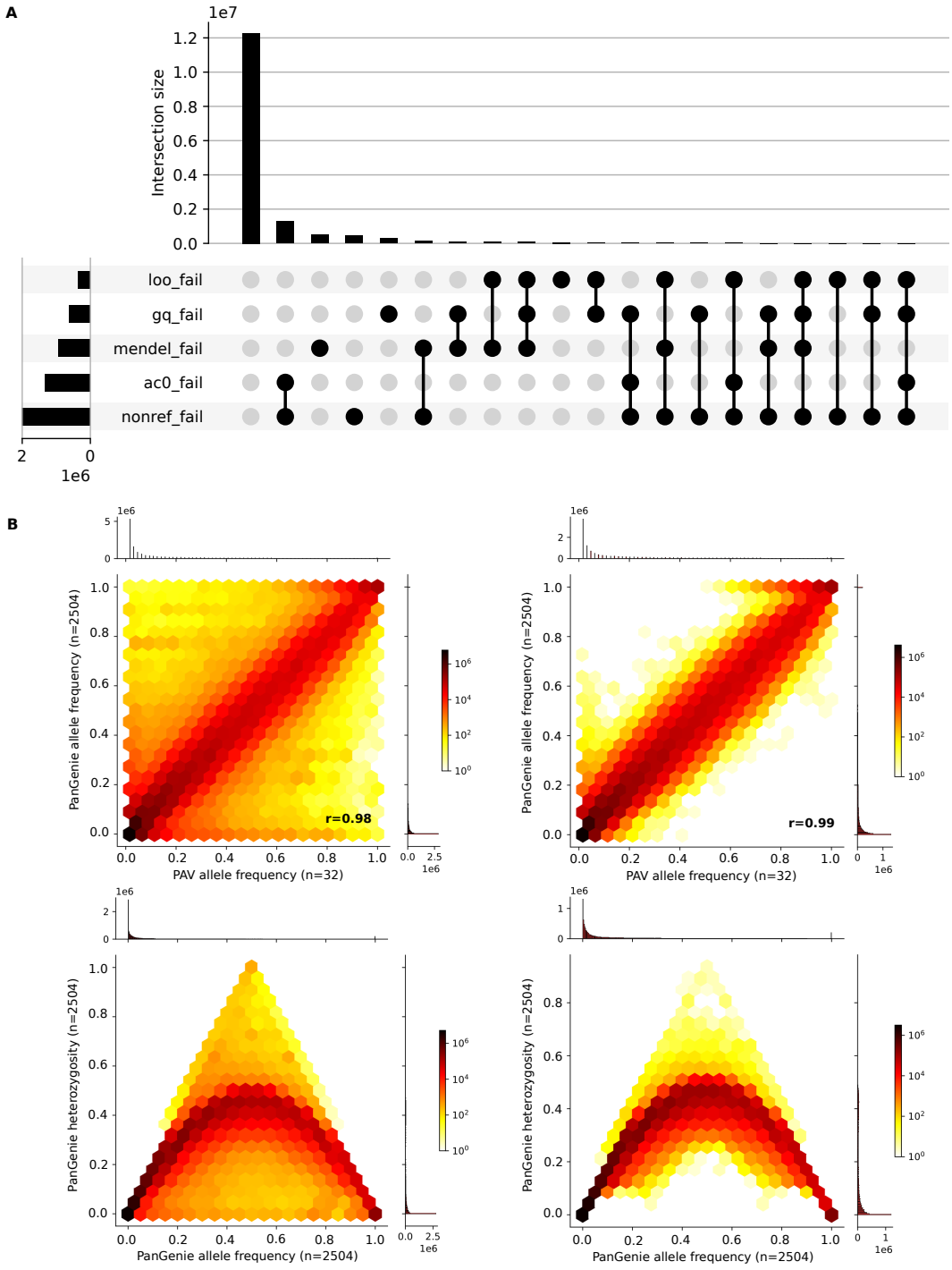


Figure C.2: HGSVC genotyping results for SNVs. **A** Shown are number of SNVs failing different filters for the set of all genotyped SNVs (n = 15,488,649). **B** The two plots on top show the allele frequencies observed for the PanGenie genotypes across all unrelated samples for all SNVs (left, n = 15,488,649) and those contained in the strict set (right, n = 12,283,650). The two plots below compare the heterozygosity of the PanGenie genotypes to the allele frequencies for all SNVs in the unfiltered (left) and strict sets (right).

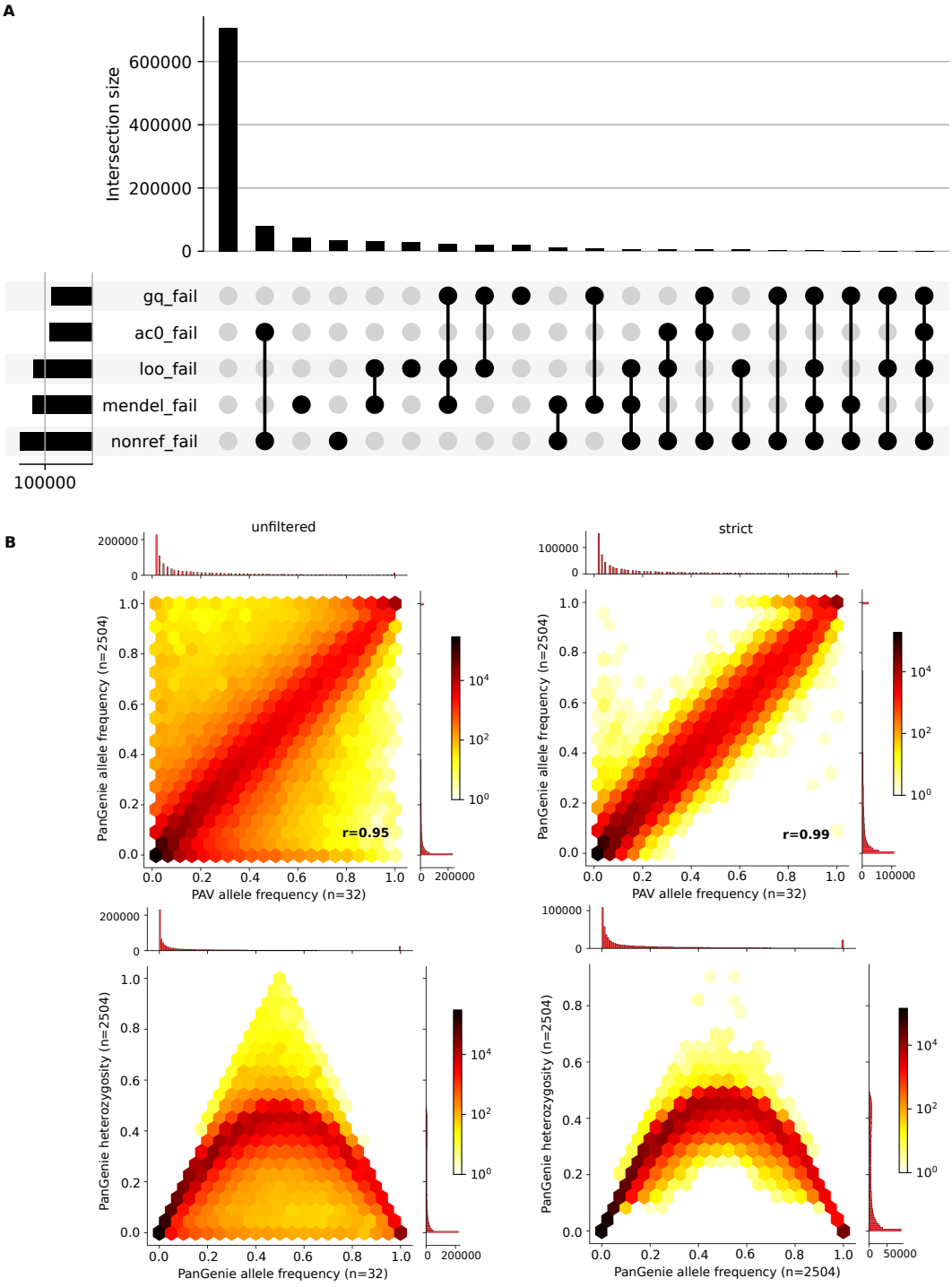


Figure C.3: HGSVC genotyping results for indels. **A** Shown are number of indels failing different filters for the set of all genotyped indels (1-49bp, n = 1,033,665). **B** The two plots on top show the allele frequencies observed for the PanGenie genotypes across all unrelated samples for all indels (left, n = 1,033,665) and those contained in the strict set (right, n = 705,893). The two plots below compare the heterozygosity of the PanGenie genotypes to the allele frequencies for all indels in the unfiltered (left) and strict sets (right).

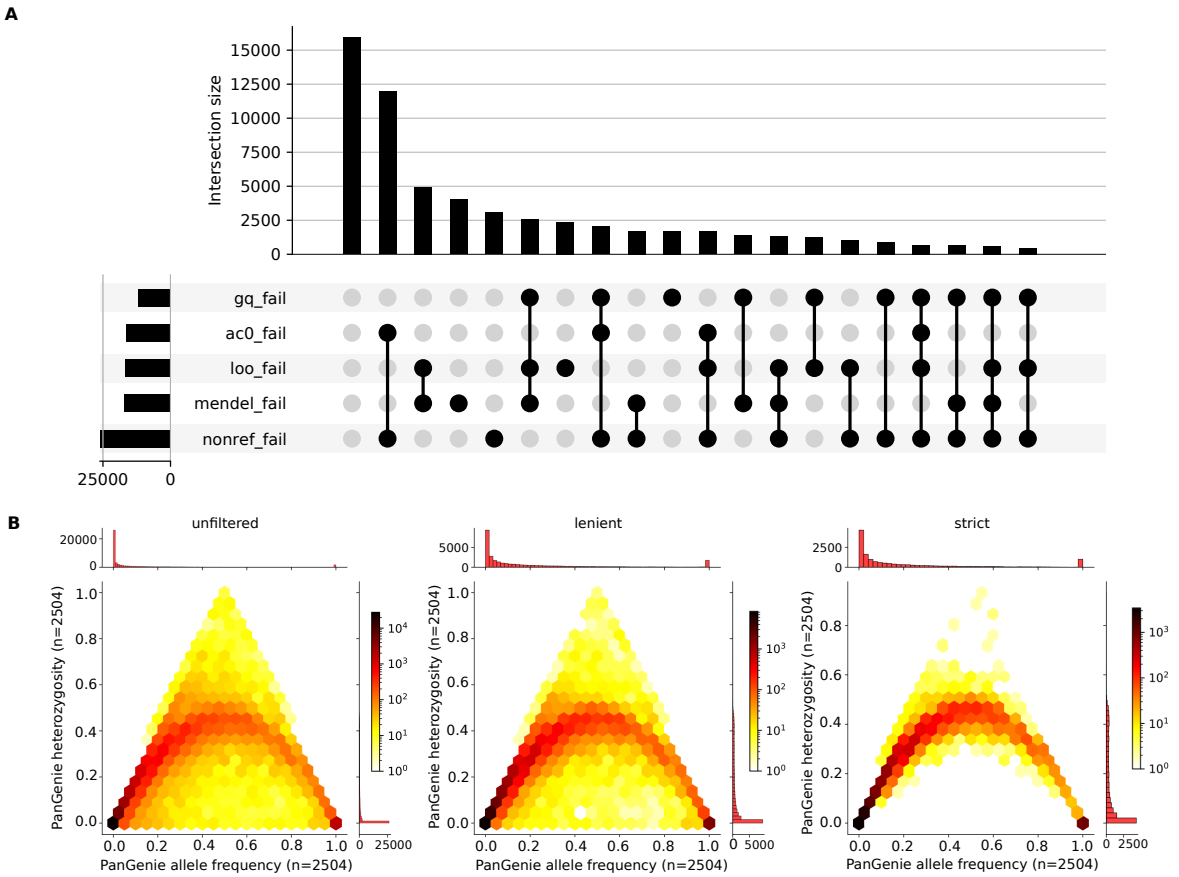


Figure C.4: Filtering HGSVC SV insertions. **A** The upset plot shows the number of variants failing the different filters for the set of all genotyped SV insertions (n = 60,283). **B** The plots show the heterozygosity of the PanGenie genotypes for all unrelated samples on the y-axis and the PanGenie allele frequencies on the x-axis. The leftmost plot corresponds to the unfiltered set containing 60,283 SV insertions, the middle one corresponds to the lenient set containing n = 31,680 insertions and the plot on the right corresponds to the strict set containing n = 15,826 SV insertions. Figures taken from [46].

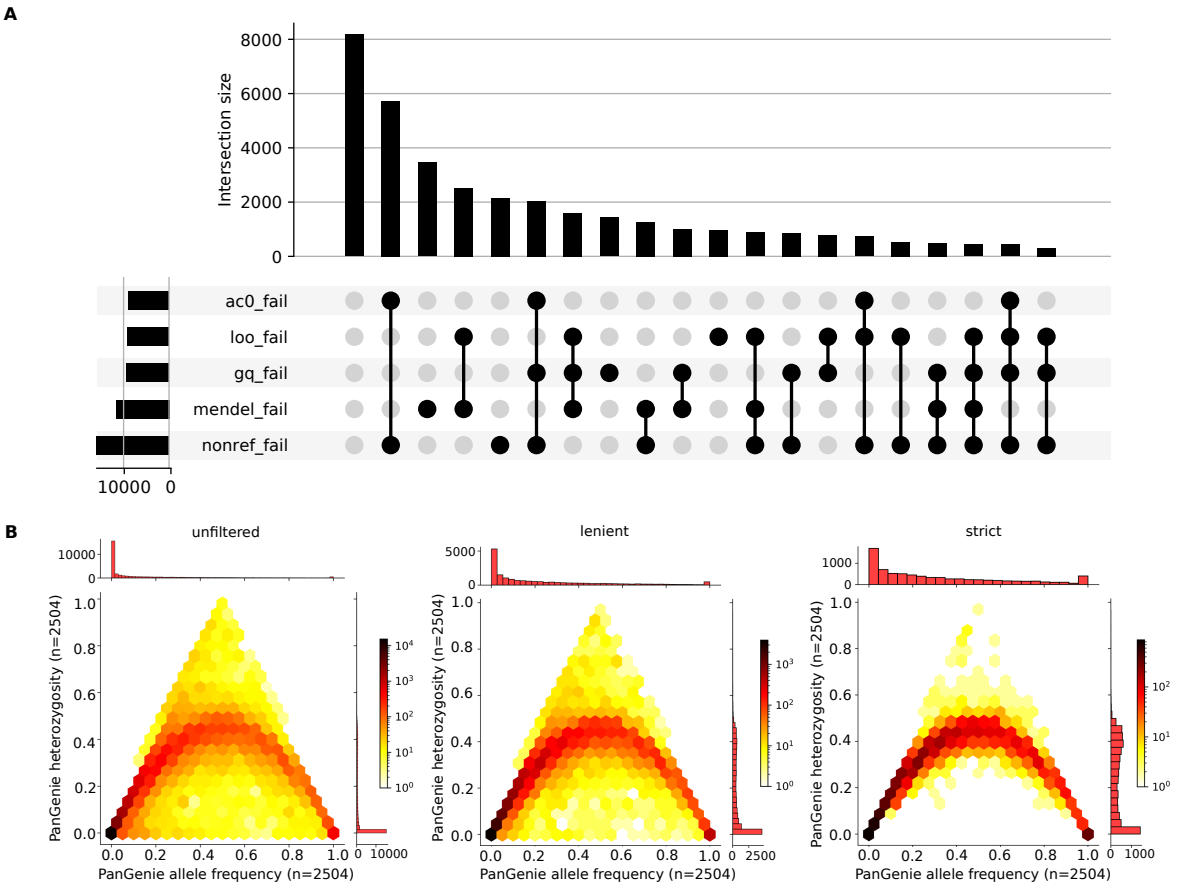


Figure C.5: Filtering HGSVC SV deletions. A The upset plot shows the number of variants failing the different filters for the set of all genotyped SV deletions (n = 35,862). **B** The plots show the heterozygosity of the PanGenie genotypes for all unrelated samples on the y-axis and the PanGenie allele frequencies on the x-axis. The leftmost plot corresponds to the unfiltered set containing 35,862 SV deletions, the middle one corresponds to the lenient set containing n = 18,660 deletions and the plot on the right corresponds to the strict set containing n = 8,181 SV deletions. Figures taken from [46].

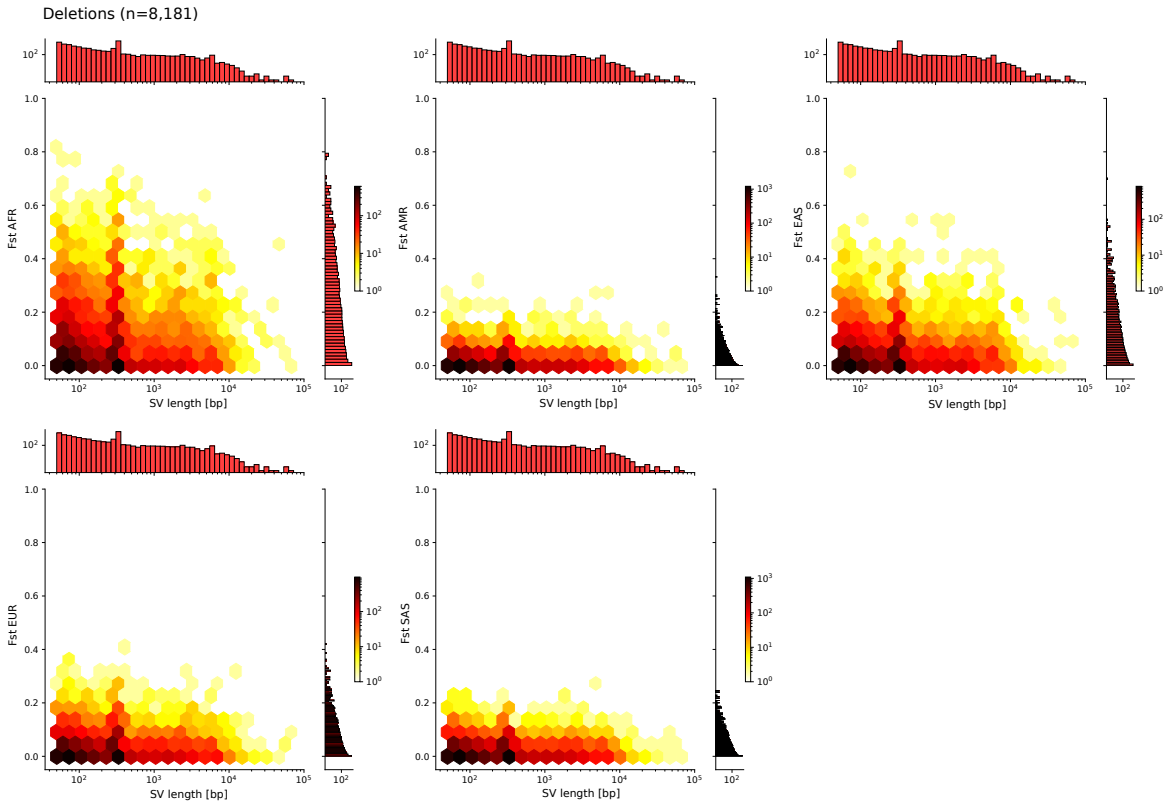


Figure C.6: F_{ST} versus SV length for all superpopulations (deletions). For each superpopulation, F_{ST} values were computed by comparing to the union of the remaining populations. The plots are based on the strictly filtered PanGenie calls containing 8,181 deletions. Figure taken from [46].

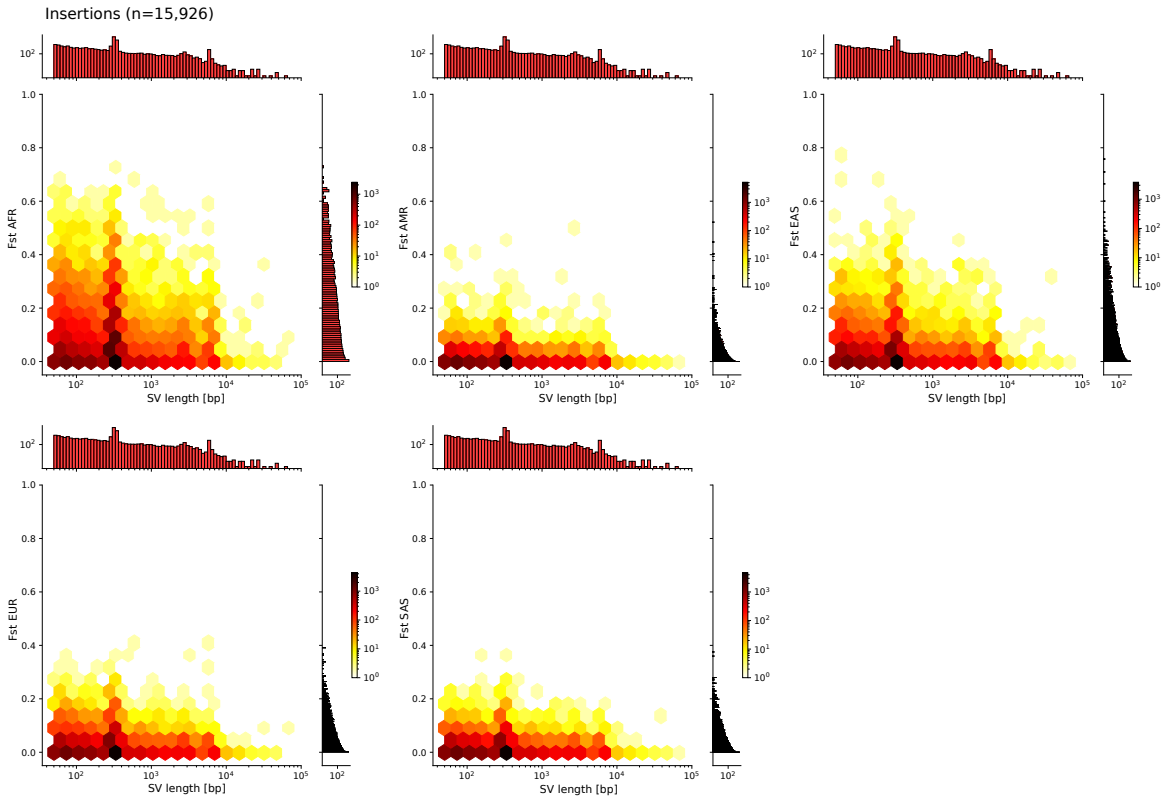


Figure C.7: F_{ST} versus SV length for all superpopulations (insertions). For each superpopulation, F_{ST} values were computed by comparing to the union of the remaining populations. The plots are based on the strictly filtered PanGenie calls containing 15,926 insertions. Figure taken from [46].

C.2 HPRC project

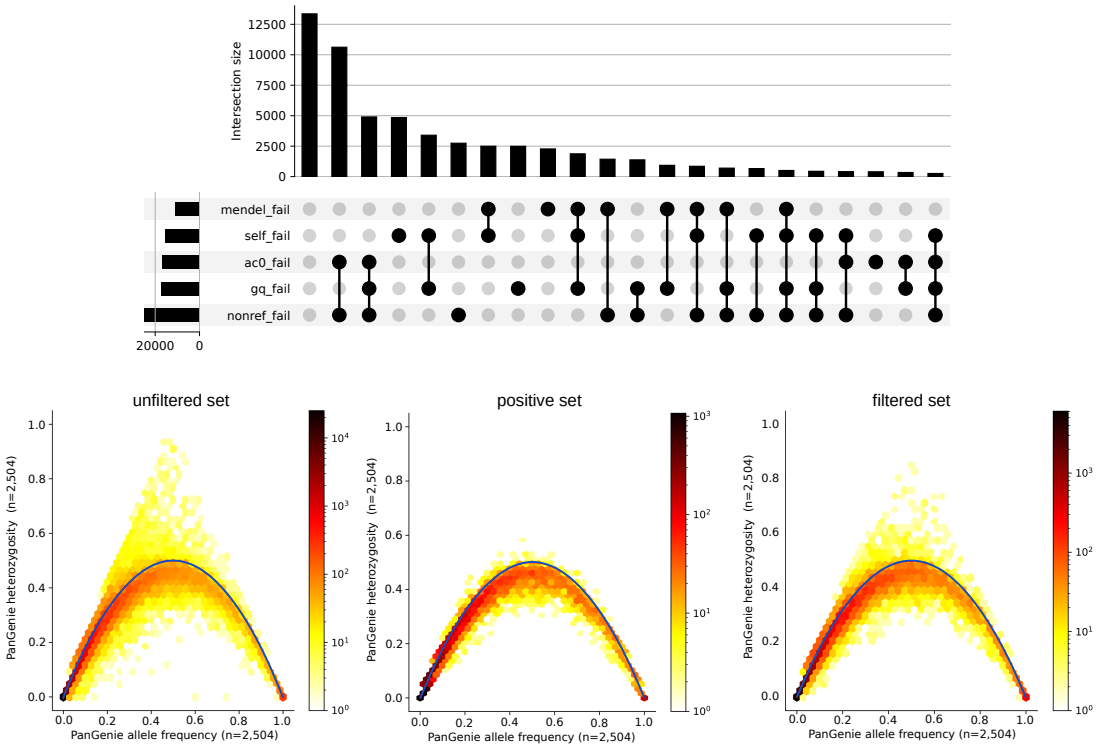


Figure C.8: HPRC callset statistics SV deletions. Top: Shown are the number of deletion alleles failing different filters computed based on the genotypes of the 1000 genomes genotypes for all 57,201 SV deletion alleles. **Bottom:** Shown are callset statistics for all SV deletion alleles ($\geq$ 50 bp) in the unfiltered set (left, n = 57,201), the positive set (middle, n = 13,356) and the final filtered set (right, n = 28,433). The plots compare the heterozygousity across the PanGenie genotypes for all 2,504 unrelated samples to the PanGenie allele frequencies. The blue line indicates the expected relationship based on Hardy-Weinberg equilibrium. Bottom panel figures are taken from [113].

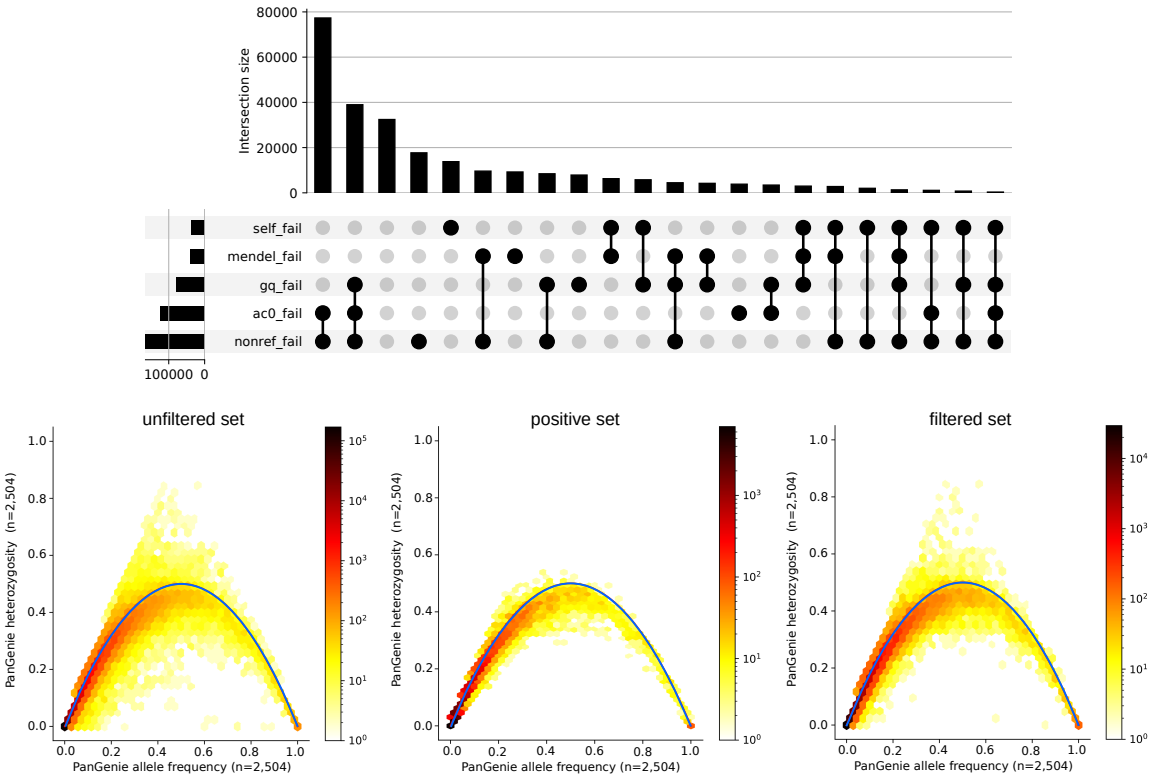


Figure C.9: HPRC callset statistics SV insertions. Top: Shown are the number of insertion alleles failing different filters computed based on the genotypes of the 1000 genomes genotypes for all 254,612 SV insertion alleles. **Bottom:** Shown are callset statistics for all SV insertion alleles ($\geq$ 50 bp) in the unfiltered set (left, n = 254,612), the positive set (middle, n = 32,431) and the final filtered set (right, n = 84,755). The plots compare the heterozygosity across the PanGenie genotypes for all 2,504 unrelated samples to the PanGenie allele frequencies. The blue line indicates the expected relationship based on Hardy-Weinberg equilibrium Bottom panel figures are taken from [113].

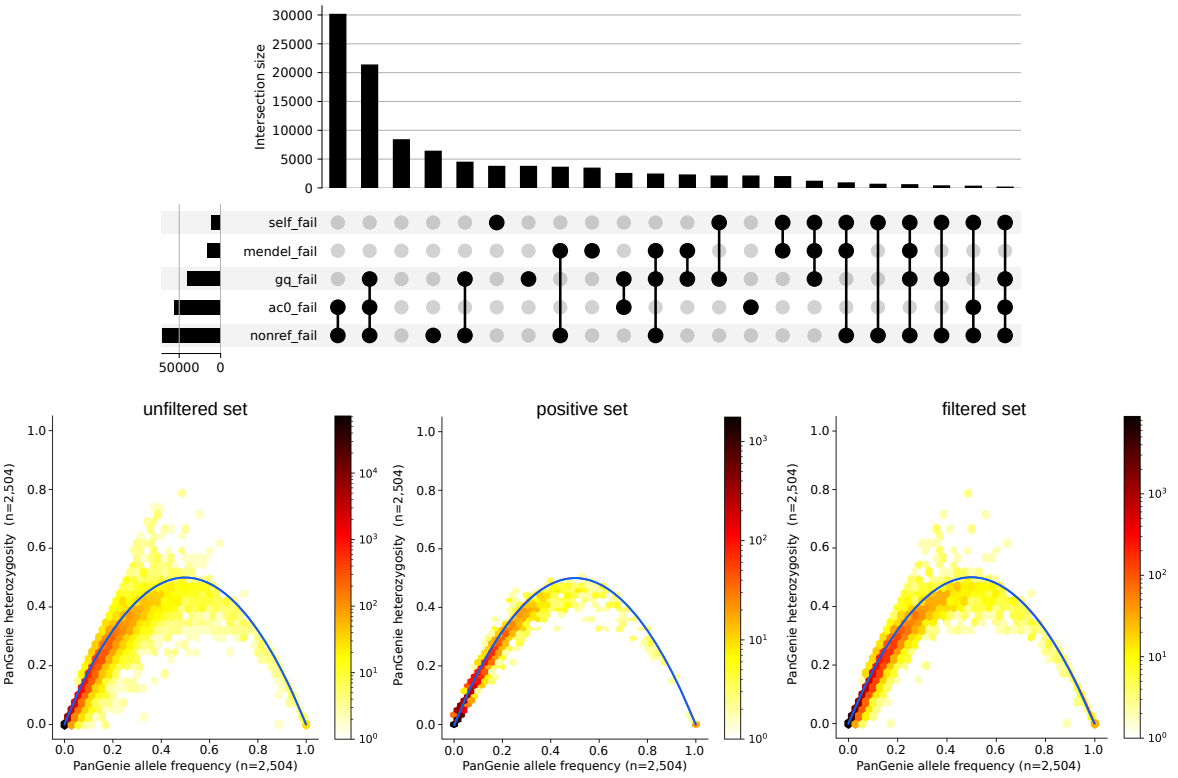


Figure C.10: HPRC callset statistics SV others. Top: Shown are the number of other SV alleles failing different filters computed based on the genotypes of the 1000 genomes genotypes for all 101,996 SV alleles that are neither insertions nor deletions. **Bottom:** Shown are callset statistics for all SV alleles that are neither a clean insertion nor a clean deletion ($\geq$ 50 bp) in the unfiltered set (left, n = 101,996), the positive set (middle, n = 8,334) and the final filtered set (right, n = 32,431). The plots compare the heterozygosity across the PanGenie genotypes for all 2,504 unrelated samples to the PanGenie allele frequencies. The blue line indicates the expected relationship based on Hardy-Weinberg equilibrium. Bottom panel figures are taken from [113].

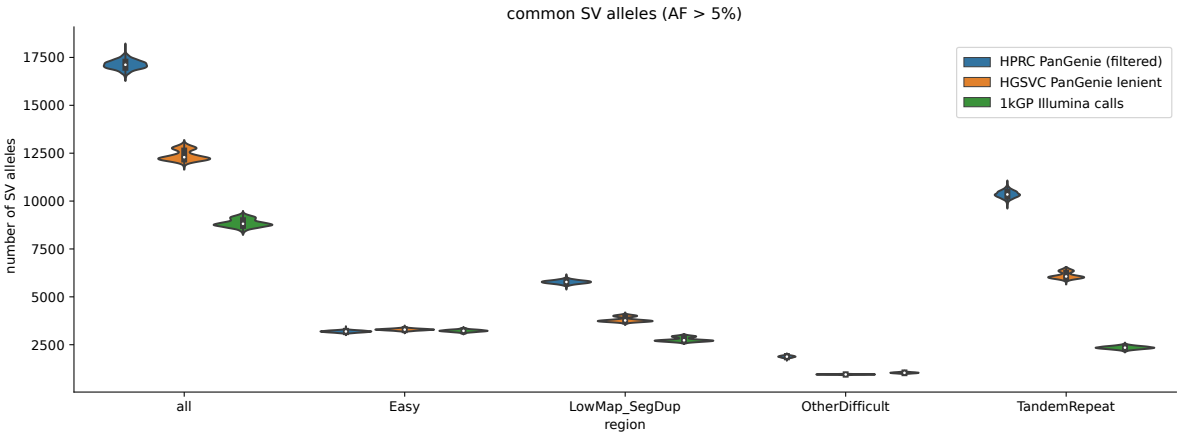


Figure C.11: HPRC common SVs. Shown are the number of SVs present (genotype 0/1 or 1/1) in each of the 3,202 1000 Genomes Project samples in the filtered HPRC genotypes (PanGenie), the HGSVC lenient set and the 1kGP Illumina calls in GIAB regions. Only common alleles with an allele frequency above 5% across all samples are taken into account. Figure taken from [113].

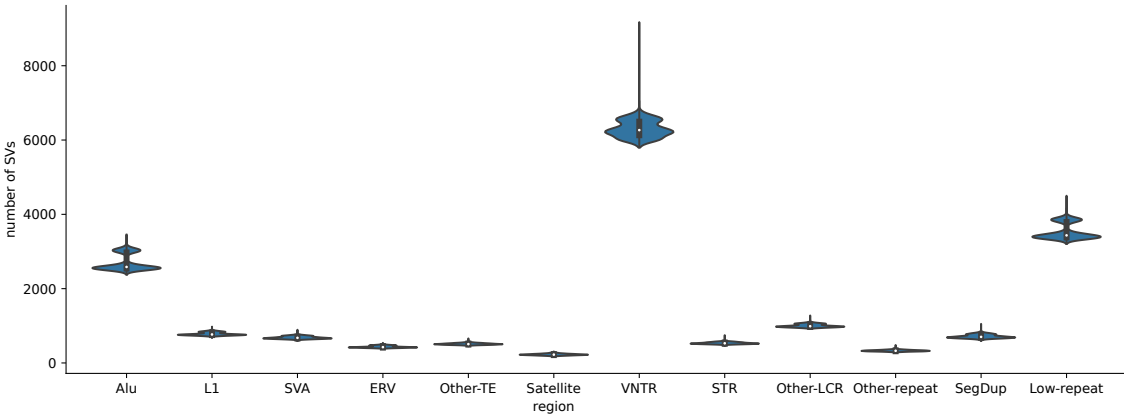


Figure C.12: HPRC SVs in repeat regions. Shown are the number of SVs present (genotype 0/1 or 1/1) in each of the 3,202 1000 Genomes Project samples in the filtered HPRC genotypes (PanGenie) in different repeat categories. Repeat annotations are based on the cactus-minigraph graph. Figure taken from [113].

Stratification files from GIAB

- easy: `https://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/giab/release/genome-stratifications/v3.0/GRCh38/union/GRCh38_notinalldifficultregions.bed.gz`
- low-mappability: `https://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/giab/release/genome-stratifications/v3.0/GRCh38/union/GRCh38_alllowmapandsegdupregions.bed.gz`
- repeats: `https://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/giab/release/genome-stratifications/v3.0/GRCh38/LowComplexity/GRCh38_AllTandemRepeats_gt100bp_slop5.bed.gz`
- other-difficult: `https://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/giab/release/genome-stratifications/v3.0/GRCh38/OtherDifficult/GRCh38_allOtherDifficultregions.bed.gz`

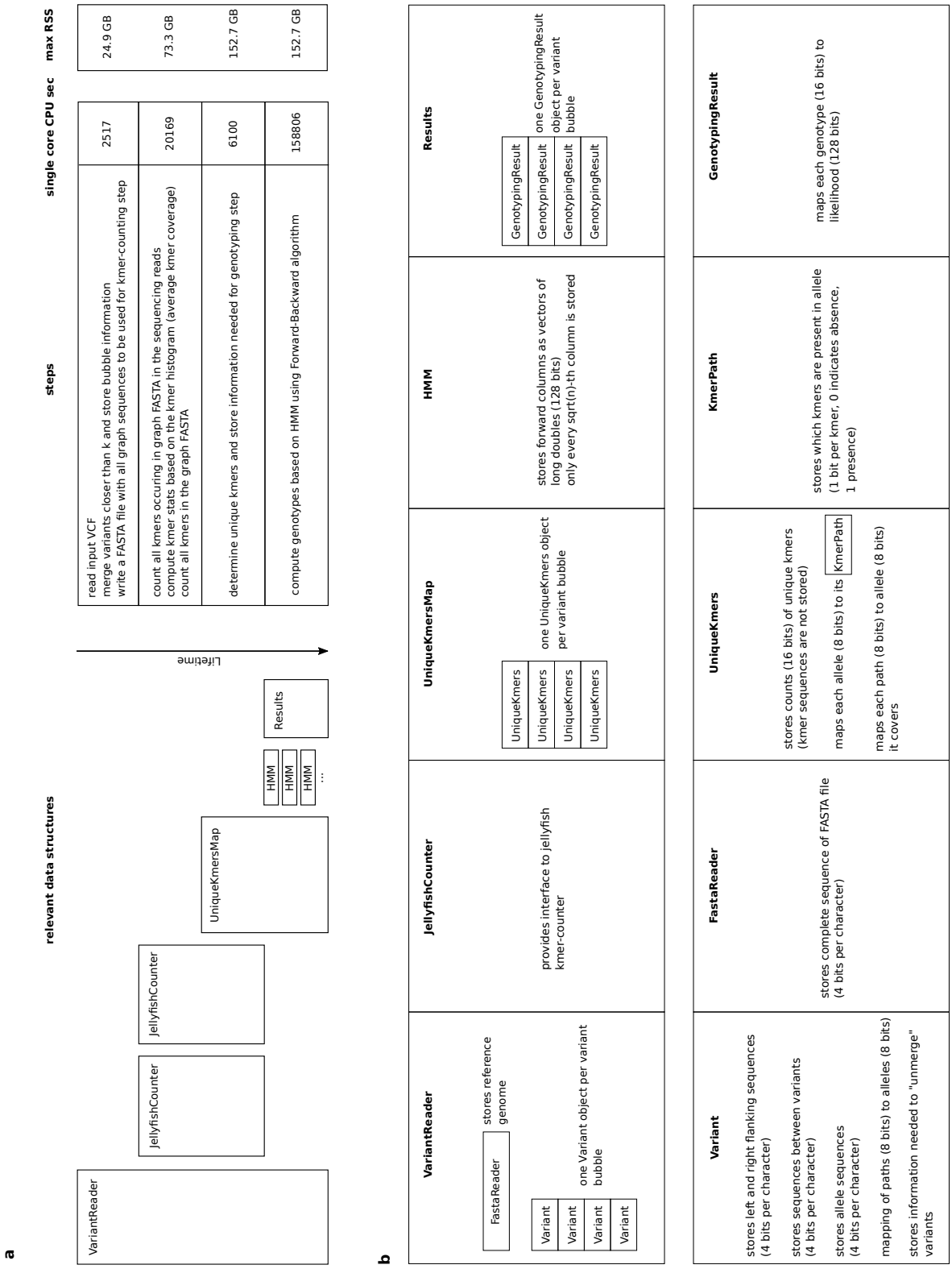


Figure C.13: Runtime and memory usage of PanGenie a Shown are the runtime (single-core CPU seconds) and maximum resident set size (RSS) of PanGenie when run on the HPRC graph containing 88 haplotypes. On the left, the lifetime of relevant data objects created during execution of the code are shown. **b** Implementation details of the most relevant classes of PanGenie.

Appendix D

Code Availability

The implementation of `whatshap polyphase` is available as open source code under the MIT licence at: `https://github.com/whatshap/whatshap`. The scripts and pipelines for re-producing the results presented in Section 2.1 are available at: `https://github.com/eblerjana/whatshap-polyphase-experiments`. The scripts and pipelines used to produce the results presented in Section 2.2 are available at: `https://github.com/PacificBiosciences/hg002-ccs/tree/master/phasing`.

The implementation of PanGenie is available as open source code under the MIT licence at: `https://github.com/eblerjana/pangenie`. The pipelines for re-producing the results presented in Chapter 3 are available at: `https://bitbucket.org/jana_ebler/genotyping-experiments/src/master/`.

The pipelines for re-producing the results presented in Sections 4.1, 4.2 and 4.3 are available at: `https://bitbucket.org/jana_ebler/genotyping-experiments/src/hgsvc-paper/`, `https://bitbucket.org/jana_ebler/rare-inversions/src/master/` and `https://bitbucket.org/jana_ebler/hprc-experiments/src/master/genotyping-experiments/`, respectively.

Appendix E

Published articles underlying this thesis

E.1 Haplotype threading: accurate polyploid phasing from long reads

The manuscript "Haplotype threading: accurate polyploid phasing from long reads" [166] was published in *Genome Biology*. Author information, author contributions, license and copyright information are listed in the subsections below.

E.1.1 Authors

Sven D. Schrinner*, Rebecca Serra Mari*, Jana Ebler*, Mikko Rautiainen, Lancelot Seillier, Julia J. Reimer, Björn Usadel, Tobias Marschall, Gunnar W. Klau.

* joint first authors

E.1.2 Contributions

Author contributions as stated in the manuscript [166]:

> "SDS, RSM, JE, GWK, and TM developed the algorithmic concepts and designed the study. RSM designed the haplotype threading algorithm and implemented a prototype. SDS designed and implemented the cluster editing algorithm, designed the block cut strategies, and optimized the threading implementation. JE performed the evaluation and analyzed the potato dataset. MR ran the error correction on the potato reads. LS, JJR, and BU performed potato sequencing, and BU helped with the interpretation of phasing results. SDS, RSM, and JE integrated all software components into WhatsHap and tested the workflow. SDS, RSM, JE, GWK, and TM wrote the paper. All authors read and approved the final manuscript."

E.1.3 Licence and copyright information

E.2 Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome

The manuscript "Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome" [199] was published in *Nature Biotechnology*. Author information, author contributions, licence and copyright information are listed in the subsections below.

E.2.1 Authors

Aaron M. Wenger*, Paul Peluso*, William J. Rowell, Pi-Chuan Chang, Richard J. Hall, Gregory T. Concepcion, Jana Ebler, Arkarachai Fungtammasan, Alexey Kolesnikov, Nathan D. Olson, Armin Töpfer, Michael Alonge, Medhat Mahmoud, Yufeng Qian, Chen-Shan Chin, Adam M. Phillippy, Michael C. Schatz, Gene Myers, Mark A. DePristo, Jue Ruan, Tobias Marschall, Fritz J. Sedlazeck, Justin M. Zook, Heng Li, Sergey Koren, Andrew Carroll, David R. Rank, Michael W. Hunkapiller

* joint first authors

E.2.2 Contributions

Author contributions as stated in the manuscript [199]:

> "A.M.W., D.R.R., M.W.H. and P.P. designed the study. D.R.R. and P.P. developed the sample preparation protocol and performed sample preparation. D.R.R., P.P. and Y.Q. performed sequencing. A.C., A.K., C-S.C., M.A.D. and P.C. adapted the algorithms and implementation of DeepVariant. A.C., A.F., A.K., A.M.P., A.M.W., A.T., C-S.C., D.R.R., F.J.S., G.M., G.T.C., H.L., J.E., J.M.Z., J.R., M.A., M.A.D., M.C.S., M.M., N.D.O., P.C., P.P., R.J.H., S.K., T.M. and W.J.R. performed analysis. A.C., A.M.P., C-S.C., D.R.R., F.J.S., J.M.Z., M.A.D., M.C.S. and M.W.H. supervised analysis. A.C., A.M.W., D.R.R., G.M., J.M.Z., P.P., R.J.H., S.K. and W.J.R. wrote the manuscript. See Supplementary Note for more detailed author contributions. All authors reviewed and approved the final manuscript."

and:

> "CCS Library Preparation and Sequencing: D.R.R., P.P., Y.Q.
> Quality Evaluation of CCS Reads: A.M.W., G.M., R.J.H.
> Increased Mappability of CCS Reads: R.J.H.
> Small Variant Detection in CCS Reads: A.C., A.K., C-S.C., F.J.S., J.M.Z., M.A.D., N.D.O., P.C., W.J.R.
> Phasing Small Variants: J.E., T.M., W.J.R.
> Improving Small Variant Detection with Haplotype Phasing: A.C., A.K., M.A.D., P.C., W.J.R.
> Structural Variant Detection in CCS Reads: A.M.W., A.T., F.J.S., H.L., M.C.S., M.A., M.M.
> De Novo Assembly of CCS Reads: A.F., A.M.P., A.M.W., D.R.R., J.R., G.T.C., S.K.
> Coverage Requirements for Variant Calling and De Novo Assembly: A.C., A.K., A.M.W., G.T.C., J.E., T.M., W.J.R.
> Revising and Expanding Genome in a Bottle Benchmarks: A.M.W., J.M.Z., N.D.O."

E.2.3 Licence and copyright information

According to the link provided in the "Rights and permissions" section of the online version of the manuscript at `https://doi.org/10.1038/s41587-019-0217-9`, the following holds:

> "Ownership of copyright in original research articles remains with the Author, and provided that, when reproducing the contribution or extracts from it or from the Supplementary Information, the Author acknowledges first and reference publication in the Journal, the Author retains the following non-exclusive rights: To reproduce the contribution in whole or in part in any printed volume (book or thesis) of which they are the author(s). The author and any academic institution,

> where they work, at the time may reproduce the contribution for the purpose of course teaching. To reuse figures or tables created by the Author and contained in the Contribution in oral presentations and other works created by them. To post a copy of the contribution as accepted for publication after peer review (in locked Word processing file, of a PDF version thereof) on the Author's own web site, or the Author's institutional repository, or the Author's funding body's archive, six months after publication of the printed or online edition of the Journal, provided that they also link to the contribution on the publisher's website. Authors wishing to use the published version of their article for promotional use or on a web site must request in the normal way."

E.3 Fully phased human genome assembly without parental data using single-cell strand sequencing and long reads

The manuscript "Fully phased human genome assembly without parental data using single-cell strand sequencing and long reads" [145] was published in *Nature Biotechnology*. Author information, author contributions, licence and copyright information are listed in the subsections below.

E.3.1 Authors

David Porubsky*, Peter Ebert*, Peter A. Audano, Mitchell R. Vollger, William T. Harvey, Pierre Marijon, Jana Ebler, Katherine M. Munson, Melanie Sorensen, Arvis Sulovari, Marina Haukness, Maryam Ghareghani, Human Genome Structural Variation Consortium, Peter M. Lansdorp, Benedict Paten, Scott E. Devine, Ashley D. Sanders, Charles Lee, Mark J. P. Chaisson, Jan O. Korbel, Evan E. Eichler, Tobias Marschall

* joint first authors

E.3.2 Contributions

Author contributions as stated in the manuscript [145]:

> "D.P., P.E., E.E.E. and T.M. designed the study. P.E. and D.P. implemented the assembly workflow. P.A.A., M.J.P.C. and A.S. performed structural variant analysis. M.R.V. and D.P. analyzed assemblies for universal breaks, segmental duplications and collapses. An earlier HiFi dataset was provided by S.E.D. and used during method development. M.H. and B.P. compared assemblies to trio-binned Shasta assemblies. HGSVC members engaged in fruitful discussions, led by C.L., at the biannual consortium meetings. P.M. performed assembly graph analyses. J.E. produced multi-sample callsets for comparative assembly analysis. W.T.H.

> performed variant calling for phasing and processed Hi-C data. K.M.M. generated HiFi PacBio data. M.S. sequenced BAC clones for validation. D.P., P.A.A., M.R.V. and T.M. prepared the main display items. D.P., P.E., P.A.A., M.R.V., E.E.E. and T.M. wrote the manuscript, with input from A.D.S., M.G., P.M.L. and J.O.K."

I wrote the pipeline to generate a multi-sample callset from haplotype-resolved assemblies and applied it to call variants from the assemblies generated for five samples.

E.3.3 Licence and copyright information

The manuscript was published under a Creative Commons Licence as stated in section "Rights and permissions" in the online version: `https://www.nature.com/articles/s41587-020-0719-5`:

> "This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit `http://creativecommons.org/licenses/by/4.0/`."

E.4 Pangenome-based genome inference allows efficient and accurate genotyping across a wide spectrum of variant classes

The manuscript "Pangenome-based genome inference allows efficient and accurate genotyping across a wide spectrum of variant classes" [49] was published in *Nature Genetics*. Author information, author contributions, licence and copyright information are listed in the subsections below.

E.4.1 Authors

Jana Ebler, Peter Ebert, Wayne E. Clarke, Tobias Rausch, Peter A. Audano, Torsten Houwaart, Yafei Mao, Jan Korbel, Evan E. Eichler, Michael C. Zody, Alexander T. Dilthey, Tobias Marschall

E.4.2 Contributions

Author contributions as stated in the manuscript:

> "J.E. and T.M. developed the algorithms and designed the study. J.E. implemented PanGenie and the pangenome graph construction. P.E. generated the assemblies. T.H. and A.T.D provided ideas for graph construction and preliminary versions of the graph. J.E., T.M. designed the experiments and J.O.K., P.A.A., E.E.E., M.C.Z., T.R., A.T.D., W.E.C. and T.H. contributed ideas and suggestions. J.E. performed all experiments. W.E.C. helped with the Paragraph evaluation. Y.M., P.A.A. and E.E.E. provided the results for non-human primates discussed in the L.D. analysis. T.H and A.T.D. evaluated the assemblies in the HLA region. J.E. and T.M. wrote a draft of the paper and all authors contributed edits and comments. All authors approved the final manuscript."

E.4.3 Licence and copyright information

The manuscript was published under a Creative Commons Licence as stated in section "Rights and permissions" in the online version: `https://doi.org/10.1038/s41588-022-01043-w`:

> "This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit `http://creativecommons.org/licenses/by/4.0/`."

E.5 Haplotype-resolved diverse human genomes and integrated analysis of structural variation

The manuscript "Haplotype-resolved diverse human genomes and integrated analysis of structural variation" [46] was published in *Science*. Author information, author contributions, licence and copyright information are listed in the subsections below.

E.5.1 Authors

Peter Ebert*, Peter A. Audano*, Qihui Zhu*, Bernardo Rodriguez-Martin*, David Porubsky, Marc Jan Bonder, Arvis Sulovari, Jana Ebler, Weichen Zhou, Rebecca Serra Mari, Feyza Yilmaz, Xuefang Zhao, PingHsun Hsieh, Joyce Lee, Sushant Kumar, Jiadong Lin, Tobias Rausch, Yu Chen, Jingwen Ren, Martin Santamarina, Wolfram Höps, Hufsah Ashraf, Nelson T. Chuang, Xiaofei Yang, Katherine M. Munson, Alexandra P. Lewis, Susan Fairley, Luke J. Tallon, Wayne E. Clarke, Anna O. Basile, Marta Byrska-Bishop, André Corvelo, Uday S. Evani, Tsung-Yu Lu, Mark J. P. Chaisson, Junjie Chen, Chong Li, Harrison Brand, Aaron M. Wenger, Maryam Ghareghani, William T. Harvey, Benjamin Raeder, Patrick Hasenfeld, Allison A. Regier, Haley J. Abel, Ira M. Hall, Paul Flicek, Oliver Stegle, Mark B. Gerstein, Jose M. C. Tubio, Zepeng Mu, Yang I. Li, Xinghua Shi, Alex R. Hastie, Kai Ye, Zechen Chong, Ashley D. Sanders, Michael C. Zody, Michael E. Talkowski, Ryan E. Mills, Scott E. Devine, Charles Lee, Jan O. Korbel, Tobias Marschall, Evan E. Eichler

* joint first authors

E.5.2 Contributions

Author contributions as stated in the manuscript [46]:

> “PacBio production sequencing: K.M.M., A.P.L., Q.Z., L.J.T., and S.E.D. Strand-seq production: A.D.S., B.R., P.H., and J.O.K. Phased genome assembly: P.E., P.A.A., D.P., Q.Z., F.Y., W.T.H., and T.M. Assembly analysis: P.E. Assembly-based variant calling: P.A.A. Variant QC, merging, and annotation: P.A.A., T.R., M.J.P.C., J.R., T.L., Z.C., Y.C., K.Y., J.L., X.Y., and J.O.K. Assembly scaffolding: F.Y., D.P., and P.E. Additional long-read callsets: P.A.A., Y.C., Z.C., W.T.H., J.R., and A.M.W. Short-read SV calling and merging: X.Z., Q.Z., H.J.A., H.B., N.T.C., W.E.C., A.C., U.S.E., S.E.D., I.M.H., W.T.H., A.A.R., M.C.Z., and M.E.T. Bionano Genomics SV discovery and analysis: F.Y., J.L., and A.R.H. Strand-seq inversion detection and genotyping: D.P., W.T.H., H.A., M.G., T.M., A.D.S., and J.O.K. MEI discovery and integration: B.R.-M., W.Z., M.S., N.T.C., J.M.C.T., J.O.K., R.E.M., and S.E.D. Variant hotspot analysis: D.P. and E.E.E. Breakpoint analysis: S.K., J.L., X.Y., M.G., K.Y., and J.O.K. PanGenie genotyping: J.E. and T.M. Illumina genotype analysis: J.E., X.Z., W.E.C., P.E., T.R., P.A.A., H.B., J.O.K., M.E.T., M.C.Z., and T.M. RNA-seq and QTL analysis: M.J.B., A.S., Z.M., J.C., C.L., M.B.-B., A.O.B., O.S., Y.I.L., X.S., M.C.Z., and J.O.K. Ancestry and population genetic analyses: PH.H., R.S.M., P.A.A., T.M., and E.E.E. Data archiving: S.F., P.A.A., K.M.M., and P.F. Organization of supplementary materials: Q.Z. and C.L. Display items: P.A.A., P.E., J.E., A.R.H., PH.H., R.S.M., T.M., D.P., T.R., B.R.-M., M.S., F.Y., X.Z., and W.Z. Manuscript writing: P.A.A., P.E., B.R.-M., A.S., D.P.,

> PH.H., Q.Z., F.Y., A.R.H., J.L., M.E.T., M.J.B., X.S., S.E.D., J.O.K., T.M., and E.E.E. HGSVC Co-chairs: C.L., J.O.K., and E.E.E."

I contributed all PanGenie experiments to this work which included the preparation of the input files needed for PanGenie, running all genotyping experiments and evaluating the genotyping results.

E.5.3 License and copyright information

According to the AAAS Rights & Permissions office the following holds:

> "After publication of a manuscript in an AAAS journal, the author may reprint their full manuscript or portions of the manuscript in a thesis or dissertation written by the author as part of a course of study at an educational institution in print & electronic formats. Credit must be given to the first appearance of the material in the appropriate issue of the AAAS journal."

E.6 Recurrent inversion polymorphisms in humans associate with genetic instability and genomic disorders

The manuscript "Recurrent inversion polymorphisms in humans associate with genetic instability and genomic disorders" [146] was published in *Cell*. Author information, author contributions, licence and copyright information are listed in the subsections below.

E.6.1 Authors

David Porubsky*, Wolfram Höps*, Hufsah Ashraf*, PingHsun Hsieh, Bernardo Rodriguez-Martin, Feyza Yilmaz, Jana Ebler, Pille Hallast, Flavia Angela Maria Maggiolini, William T. Harvey, Barbara Henning, Peter A. Audano, David S. Gordon, Peter Ebert, Patrick Hasenfeld, Eva Benito, Qihui Zhu, Charles Lee, Francesca Antonacci, Matthias Steinrücken, Christine R. Beck, Ashley D. Sanders, Tobias Marschall, Evan E. Eichler, Jan O. Korbel

* joint first authors

E.6.2 Contributions

Author contributions as stated in the manuscript [146]:

> "Conceptualization, D.P., A.D.S., T.M., E.E.E., and J.O.K.; methodology & software, D.P., W.H., H.A., P. Hsieh, B.R.-M., and M.S.; formal analysis, D.P., W.H., H.A., P. Hsieh, B.R.-M., F.Y., J.E., and P. Hallast; investigation, D.P., W.H., H.A., P. Hsieh, B.R.-M., A.D.S., M.S., and C.R.B.; resources, HGSVC, Q.Z., C.L., P. Hasenfeld, A.D.S., T.M., E.E.E., and J.O.K.; computational support, W.T.H.,

> P.A.A., B.H., and D.S.G.; validation, F.A.M.M., P.E., E.B., and F.A.; writing, D.P., W.H., H.A., B.R.-M., T.M., E.E.E., and J.O.K., with input from all authors."

In this work, I analyzed PanGenie genotypes of rare variants in order to find candidate samples that carry inversions.

E.6.3 License and copyright information

This paper was published under the Creative Commons CC-BY-NC licence as stated in the online version: `https://doi.org/10.1016/j.cell.2022.04.017`

E.7 A Draft Human Pangenome Reference

The manuscript "A Draft Human Pangenome Reference" [113] is publicly available as a *bioArxiv* preprint. It is currently under revision. Author information, author contributions, licence and copyright information are listed in the subsections below.

E.7.1 Authors

Wen-Wei Liao*, Mobin Asri*, Jana Ebler*, Daniel Doerr, Marina Haukness, Glenn Hickey, Shuangjia Lu, Julian K. Lucas, Jean Monlong, Haley J. Abel, Silvia Buonaiuto, Xian H. Chang, Haoyu Cheng, Justin Chu, Vincenza Colonna, Jordan M. Eizenga, Xiaowen Feng, Christian Fischer, Robert S. Fulton, Shilpa Garg, Cristian Groza, Andrea Guarracino, William T Harvey, Simon Heumos, Kerstin Howe, Miten Jain, Tsung-Yu Lu, Charles Markello, Fergal J. Martin, Matthew W. Mitchell, Katherine M. Munson, Moses Njagi Mwaniki, Adam M. Novak, Hugh E. Olsen, Trevor Pesout, David Porubsky, Pjotr Prins, Jonas A. Sibbesen, Chad Tomlinson, Flavia Villani, Mitchell R. Vollger, Human Pangenome Reference Consortium, Guillaume Bourque, Mark JP Chaisson, Paul Flicek, Adam M. Phillippy, Justin M. Zook, Evan E. Eichler, David Haussler, Erich D. Jarvis, Karen H. Miga, Ting Wang, Erik Garrison, Tobias Marschall, Ira Hall, Heng Li, Benedict Paten

* joint first authors

E.7.2 Contributions

Author contributions as stated in the manuscript [113]:

> "**Paper writing** Wen-Wei Liao, Mobin Asri, Jana Ebler, Daniel Doerr, Marina Haukness, Glenn Hickey, Shuangjia Lu, Jean Monlong, Shilpa Garg, Erik Garrison, Tsung-Yu Lu, Matthew W. Mitchell, Adam M. Novak, Trevor Pesout, Jonas

A. Sibbesen, Mitchell R. Vollger, Guillaume Bourque, Karen H. Miga, Tobias Marschall, Ira Hall, Benedict Paten, Robert S. Fulton, Richard E. Green, Leanne Haggerty, Hugh E. Olsen, Fergal J. Martin

...

Pangenome applications: structural variants Jana Ebler, Glenn Hickey, Haley J. Abel, William T Harvey, Pjotr Prins, Erik Garrison, Evan E. Eichler, Tobias Marschall, Hanlee P. Ji, Hugo Magalhães"

I ran PanGenie to genotype the 1000 Genomes samples and evaluated the results for structural variants.

E.7.3 License and copyright information