Haplotype-resolved assembly of diploid and polyploid species and its applications

Inaugural-Dissertation

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

vorgelegt von

Rebecca Serra Mari
aus Merzig

Düsseldorf, April, 2025

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

3. Prof. Dr. Vikas Bansal

Tag der mündlichen Prüfung: 16.01.2026

Eidesstattliche Versicherung

Ich versichere an Eides Statt, dass die Dissertation von mir selbstständig und ohne unzulässige fremde Hilfe unter Beachtung der "Grundsätze zur Sicherung guter wissenschaftlicher Praxis an der Heinrich-Heine-Universität Düsseldorf" erstellt worden ist.

Düsseldorf, April 2025

Rebecca Serra Mari

Abstract

DNA, the molecular blueprint of life, is organized in the chromosomes of all eukaryotes. The DNA exists in multiple copies: humans are diploid and have two copies, other organisms such as most plants are polyploid and contain more than two genome copies. Reconstructing these exact copies, known as haplotypes, is referred to as phasing, or haplotype assembly. Modern sequencing technologies and algorithms have revolutionized genomic research by enabling highly accurate genome and haplotype assemblies, which facilitate many downstream analyses. Most progress has been made for diploids, while the field of polyploid genomics is lagging behind.

The first part of this dissertation investigates ancestral origins of genomic regions through shared haplotype blocks in high-quality phased genome assemblies. A Hidden Markov Model is presented and applied to a data set of assemblies from diverse ancestries, focusing on a Puerto Rican trio, to infer ancestry estimates.

The second part addresses a gap in polyploid phasing by introducing a novel algorithm based on read clustering and haplotype threading. It is applied to artificial polyploid datasets and the tetraploid potato genome, overcoming key challenges in polyploid phasing.

In the third part of this thesis, a method for the *de novo* assembly of a tetraploid potato genome is presented. It features a graph-based approach that uses long-read sequencing and a large progeny panel. It is demonstrated that the analysis of haplotype-specific k-mers in the progeny enables haplotype-resolved chromosome-scale assembly.

Finally, the fourth part explores the potential of assembly using a three-generation pedigree. In a shared assembly graph, child haplotypes are resolved via parent-specific k-mers. This approach identifies shared sequences and meiotic recombination breakpoints, demonstrating the utility of pangenome graphs for analysing genetic inheritance across generations.

Kurzfassung

Die DNA, der molekulare Bauplan des Lebens, ist in allen Eukaryoten in Chromosomen organisiert. Die DNA liegt in mehreren Kopien vor: Der Mensch ist diploid und besitzt zwei Kopien, andere Organismen wie die meisten Pflanzen sind polyploid und enthalten mehr als zwei Kopien des Genoms. Die Rekonstruktion dieser exakten Kopien, der so genannten Haplotypen, wird auch als Phasing oder Haplotyp-Assembly bezeichnet. Moderne Sequenziertechnologien und Algorithmen haben das Feld der Genomanalyse revolutioniert, indem sie hochpräzise Genom- und Haplotyp-Assemblies ermöglichen, die zahlreiche nachfolgende Analysen erleichtern. Die größten Fortschritte wurden bei diploiden Organismen erzielt, während die Forschung im polyploiden Bereich bislang nicht mit der Entwicklung Schritt halten konnte.

Der erste Teil dieser Arbeit untersucht den regionalen Ursprung von Genomregionen in neu rekonstruierten Haplotypen durch den Vergleich mit einer Kohorte aus verschiedenen Populationen. Es wird ein Hidden-Markov-Modell vorgestellt und auf einen Datensatz von Haplotypen von Personen verschiedener Abstammung angewendet–mit besonderem Fokus auf einem Trio aus Puerto Rico–um Schätzungen über die Abstammung anzustellen.

Der zweite Teil adressiert eine Lücke im polyploiden Phasing durch die Einführung eines neuen Algorithmus, der auf der Gruppierung von Reads und der Rekonstruktion von Haplotypen basiert. Dieser wird auf modellhafte polyploide Datensätze und das tetraploide Kartoffelgenom angewendet, wobei wesentliche Herausforderungen des polyploiden Phasings überwunden werden.

Im dritten Teil dieser Arbeit wird eine Methode zur Genomrekonstruktion eines tetraploiden Kartoffelgenoms vorgestellt. Es wird ein graphbasierter Ansatz entwickelt, der lange Reads und eine große Nachkommenpopulation verwendet. Dabei wird gezeigt, dass die Analyse der haplotyp-spezifischen k-mere in den Nachkommen eine Haplotyprekonstruktion auf Chromosomenebene ermöglicht.

Abschließend untersucht der vierte Teil das Potential des Haplotyp-Assembly anhand eines Drei-Generationen-Stammbaums. In einem gemeinsamen Assemblygraphen wer-

den die Haplotypen des Kindes durch elternspezifische k-mere aufgelöst. Dieser Ansatz identifiziert gemeinsame Sequenzen und meiotische Rekombination und demonstriert die Verwendung von Pangenomgraphen zur Analyse der genetischen Vererbung über mehrere Generationen hinweg.

Acknowledgements

First and foremost, I am deeply grateful to my supervisor, Tobias Marschall, for his support and guidance throughout my studies and during my PhD, and for his kind mentorship. I was fortunate to learn immensely from him over the years. I would also like to thank Gunnar Klau for agreeing to take on the role as second reviewer, and for his support and advice on two projects of this thesis, which was helpful and inspiring.

My heartfelt thanks go to all my peers, especially – but not restricted to – Ali Ghaffaari, Fawaz Dabbaghie, Haniyeh Barghi, Hufsah Ashraf, Hugo Magalhães, Jana Ebler, Kai Horny, Konstantinn Bonnet, Maryam Ghareghani, Mikko Rautiainen, Mir Henglin, Peter Ebert, Samarendra Pani and all past and present members of the Marschall Lab. It was a privilege to work in an environment with such kind and smart people, and I will cherish the time we spent together.
For proof-reading parts of this thesis and providing helpful feedback, I am grateful to Fawaz Dabbaghie, Hugo Magalhães, Jana Ebler, Kai Horny, Konstantinn Bonnet, Mir Henglin, Peter Ebert, Samarendra Pani, and Vithusan Suppiyar.

I would like to thank Sven Friedrich, André Feller and Jens Halbauer for being an immeasurable source of inspiration, comfort, and joy throughout the past years. Thank you.
Above all, I am thankful to my family for their continuous support, and to my friends. I am particularly grateful to Deborah Pitzius for being my life-long friend, and for many coworking hours. Finally, my deepest gratitude goes to Tim Goll, for all his love and support.

Contents

List of Figures

List of Tables

Introduction

The development, function and reproduction of a living organism is encoded in its DNA (deoxyribonucleic acid). The DNA molecule consists of two strands wound up in the shape of a double helix, where each strand consists of a sugar-phosphate backbone and a sequence of nucleotide bases – adenine (A), cytosine (C), guanine (G) and thymine (T). DNA sequencing is the process of making this nucleotide sequence accessible as readable data for analysis. Modern sequencing technologies allow the generation of reads deciphering tens of thousands of bases with 99.9% accuracy [220]. Other technologies output reads even hundreds of kilobases in length [48]. The process of reconstructing the full genome of an individual from sequencing reads is known as genome assembly. The first assembled human genome was presented in 2001 [96], but only recently, 20 years later, a fully complete sequence was published, filling the remaining gaps [146].

Any two humans share more than 99% of their DNA sequence [203]. The remaining variation is responsible for the diversity within our species; genomic variants influence phenotypic and behavioral traits and may be causal for multiple diseases [65, 110, 80]. This variation ranges from single-nucleotide polymorphisms, affecting one single base, to structural variants that can span tens of millions of bases [56].

Humans, animals and plants are eukaryotes and store their DNA in the nucleus of each cell, packaged into chromosomes. Humans are diploid and possess two copies of each chromosome - one inherited from each parent. Other organisms, particularly most plants, exhibit higher ploidy, such as the tetraploid potato, which carries four copies of each chromosome. The individual copies of the DNA are referred to as haplotypes and in particular, they provide a coherent representation of all variation that is jointly inherited from one parent. Reconstructing haplotypes, called phasing, is crucial for applications in population genetics and genomic medicine, or for breeding strategies in plants.

During meiosis, recombination between the paternal and maternal chromosome copies in the germ cell can lead to sequence exchange between the two. The chromosome copy which is passed on from that parent to the offspring will therefore contain a mixture of the two haplotypes. All humans share common ancestry and, consequently,

contain shared haplotype blocks. Analysis of these shared sequences enables studying population history and evolutionary developments [140, 132].

Accurate genome and haplotype assemblies are becoming increasingly accessible thanks to technological and algorithmic advancements. A number of these advancements are reflected in the structure of this thesis. First, Chapter 1 lays the groundwork for understanding the biological and computational key concepts underlying the work presented in this thesis.

Chapter 2 presents a method for local ancestry inference, developed as part of a project initiated by the Human Genome Structural Variation Consortium (HGSVC). The HGSVC utilized a pipeline for accurate phased assembly of 32 individuals based on single-cell strand sequencing and long reads. We make use of the shared haplotype blocks between individuals to investigate the population ancestry of the resulting haplotypes. We present a model that infers ancestry estimates for genomic regions and apply this to the 64 haplotype assemblies from diverse ancestries, with a focus on a Puerto Rican mother-father-child trio. This work was published as part of a *Science* publication [56].

In contrast to the rapid progress of diploid phasing methods, analogous advancements for the polyploid case are missing. We aim to address this gap by introducing a novel reference-based phasing approach tailored to polyploid genomes in Chapter 3. We examine polyploid phasing and its challenges and present `WhatsHap polyphase`, a two-step approach that clusters sequence reads based on similarities and then traces the haplotypes through these clusters. We phase artificial polyploid data and the genes of a tetraploid potato. This work was published in a *Genome Biology* article [184].

Inspired by the developments achieved in human genome assembly, especially the forthcoming of assembly methods that allow direct haplotype resolution, the focus shifts to *de novo* assembly in Chapter 4. The aim was to construct a haplotype-resolved assembly of the tetraploid potato genome. We developed a graph-based approach that employs a combination of accurate long reads of the cultivar Altus and sequencing data from a large offspring panel. Haplotype-specific k-mers are counted in the offspring samples and resulting k-mer count profiles are used for sorting reads by chromosome and haplotype. The methodology and the final assembly were published in *Genome Biology* [189].

Finally, Chapter 5 further explores the potential of graph-based assembly, inspired by the recent progress in the field of pangenomic representations. Using a combination of highly accurate and ultra-long sequencing data from a three-generational pedigree, we build a joint assembly graph and perform haplotype-resolved assembly of the child sample. The pangenome graph representation enables the identification of sequence frag-

ments that are transmitted across the three generations. This enables follow-up analysis of shared genomic tracts. In particular, we were able to detect meiotic recombination breakpoints.

Chapter 1

Background

1.1 DNA and genomes

The genetic information of a human individual is stored in the DNA (deoxyribonucleic acid). The DNA ensures the development and functioning of a living organism, and it contains the hereditary information that is passed on through generations. A DNA molecule consists of two long strands composed of a sugar-phosphate backbone and four types of bases: adenine (A), thymine (T), cytosine (C) and guanine (G). These bases pair in a specific manner – A with T and C with G – to form the double-stranded structure of the DNA, which is shaped in the characteristic double helix. The sequence of these *base pairs* encodes the genetic information.

The DNA is stored in the nucleus of every human cell. When stretched out, each cell would contain approximately 2 meters of DNA, while the nucleus itself is only 6 μm in size [6, 155]. Therefore, the DNA must be packaged very tightly, requiring several layers of efficient wrapping and coiling of the DNA strands (see Figure 1.1). To this end, the DNA molecules are wrapped around certain proteins (histones) and form DNA-protein complexes which are called *nucleosomes*. These complexes make up a structure called *chromatin*. The chromatin is then organized into *chromosomes*, of which humans typically contain 23 pairs: 22 autosomes (chromosomes 1 to 22), each present in two copies, and one pair of gonosomes (chromosome X and Y) which are relevant for sex development. Thus, a human genome typically contains 46 chromosomes. The human DNA consists of approximately 3.1 billion base pairs (Gigabase pairs, Gbp) [155]. The most accurate DNA assembly to date lists 3.055 Gbp for 22 autosomes and one X chromosome [146]. The Y chromosome has been shown to vary in size; among 43 assembled human Y chromosomes, a range between 45.2 and 84.9 Mbp was found (mean length of 57.6 Mbp) [79]. Another complete assembly of a Y chromosome contains 62.5 Mbp of sequence [171]. As

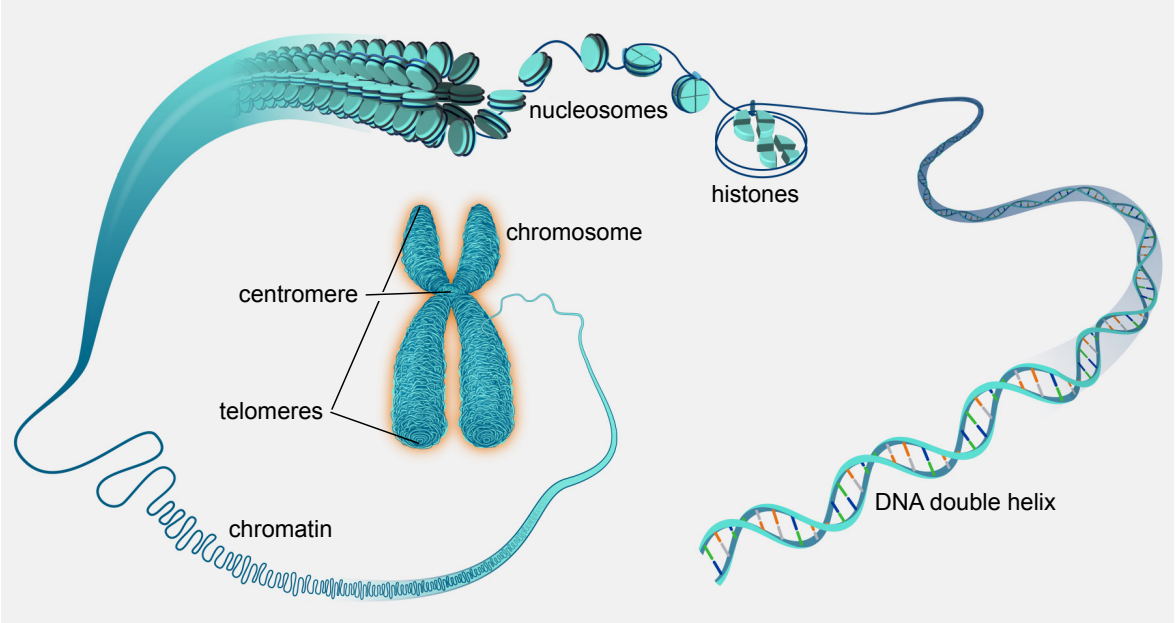


Figure 1.1: Organization of the DNA in the human cell. The DNA strands (dark blue) are wrapped around specific proteins (histones, cyan colour) to form protein-DNA-complexes called nucleosomes. The coiled structure of nucleosomes forms the chromatin, which in turn comprises the final chromosome as it is present in the cell nucleus. The chromosome center harbours the centromere, the ends contain the telomeres. Figure courtesy: National Human Genome Research Institute, published in the public domain. Text labels have been adapted from the original for this thesis.

humans contain two copies of the genome – one inherited from the mother, one from the father –, the total amount of base pairs in a human cell is more than 6 billion.

The center part of a human chromosome is called *centromere* and counts to the most complex regions of the genome due to its high diversity among individuals [10, 122]. The two ends are called *telomeres*. The sequences of the telomeres are commonly recognized by the characteristic sequence TTAGGG which is variably repeated numerous times [138, 92]. A chromosome typically possesses a shorter 'arm', termed p-arm, and a longer arm with the term q-arm. The human chromosomes 13, 14, 15, 21 and 22 are special in the sense that their short arms are extremely similar to one another, more than the other chromosomes [146]. These chromosomes are called *acrocentrics*. In particular, they harbour the human rDNA arrays, which are arrays of 45 kbp long tandem repeats (repeated nucleotide patterns directly adjacent to each other) that contain nearly identical sequence and represent a highly complex region in the human genome [146]. On average, a human genome contains 315 rDNA copies, but this number varies between individuals [146].

1.2 DNA sequencing: Reading the genetic code

At the beginning of all genetic analyses stands decoding the genetic information of the DNA. This is done by DNA sequencing, the process of determining the precise order of the nucleotides A, C, G and T within a DNA molecule.

The original, *first-generation* sequencing methods were developed in the mid-seventies by Frederick Sanger and, independently, by Maxam and Gilbert, but the latter was replaced in the course of further development of Sanger's method [73, 179]. In 1977, the Sanger method was published as the first sequencing method capable of sequencing up to 200-300 nucleotides [178]. This so-called *chain termination* method is based on polymerisation using the enzyme DNA polymerase to synthesise a complementary strand of DNA. Starting with a primer that binds near the region of interest, the enzyme extends the DNA strand by incorporating deoxynucleotide triphosphates (dNTPs), the building blocks of the DNA. Along with normal dNTPs, a small proportion of dideoxynucleotide triphosphates (ddNTPs) are included in the reaction.They are modified such that they lack the hydroxyl group at the 3' carbon, which prevents further strand extension. Occasionally, a ddNTP is incorporated into the strand, leading to the sequence chain being terminated. This results in DNA fragments of various lengths [178]. The fragments are separated by size using electrophoresis. Labelling the ddNTPs allows for identification of the terminal nucleotide of a fragment. Originally, this labelling was done via radiolabelling with radioactive phosphorus and then later got replaced by fluorescent dyes [136]. Identifying the fluorescent signal of the terminal nucleotides of each fragment reconstructs the original sequence. This underlying method is also called sequencing by synthesis (SBS) and remains the foundational concept that is used nowadays in modern sequencing platforms. It has served as the backbone for numerous advancements since then, such as the breakthrough introduced by next-generation sequencing methods.

1.2.1 Next-generation sequencing

While Sanger sequencing is a highly accurate method that laid the groundwork for sequencing and can still be of use today [46], the progress in genomic research of the last years would not have been possible without the development of far more efficient technologies which enable the sequencing of full genomes at substantially lower time scales and costs. These *next-generation* sequencing techniques are based on massively parallelising the concepts that Sanger had developed, where enzymatic synthesis and subsequent sequence detection are no longer two separate steps [136], but coupled into one step. One of the most prominent examples comes from Illumina. First, a library is created

by fragmenting the DNA first and then ligating adapters to the fragment ends. The fragments are loaded onto a flow cell, a glass slide containing oligonucleotides to which the fragments bind. In a process called bridge amplification, DNA polymerase synthesises a complementary strand to the bound fragment; this process is repeated multiple times and results in a cluster of multiple copies of the DNA fragment. Then, labelled dNTPs are incorporated into the synthesis, similar to the incorporation of the chain terminating ddNTPs in the Sanger method. The fluorescent label is detected and the corresponding nucleotide identified [197]. This is done with millions of DNA fragments in parallel, as opposed to just a single template fragment, and allows for a much higher throughput [12]. Such high-throughput methods are essential for large-scale projects such as whole genome sequencing.

1.2.2 Third-generation sequencing

The most recent sequencing methods are labelled as *third-generation*, which refers to those methods that do not rely on DNA amplification [73], but are able to sequence single molecules. This reduces possible artifacts [136] and allows for greater read lengths [136].

One of the most prominent approaches is the SMRT (single molecule, real time) method by PacBio (Pacific Biosciences) [60]. The applied principle remains the same: dNTPs are incorporated in the synthesis of a DNA strand. A circular DNA template is used, a so-called SMRTbell, consisting of a double-stranded DNA template with single-stranded adapters on the ends [123, 188]. An SMRT sequencing cell contains millions of zero-mode waveguides (ZMWs), which are wells to immobilize the single DNA molecules [60]. The SMRT cell then measures the light that is emitted during the incorporation of labelled nucleotides in real time.

The most accurate version of this method is circular consensus sequencing (CCS), in which the SMRT bell template fragment is traversed multiple times by the polymerase, so that several reads are produced for the same DNA fragment, which are then combined in a consensus read [123, 220]. This process yields highly accurate HiFi (high fidelity) reads, as the consensus step enables the eradication of sequencing errors in the produced subreads. In fact, HiFi reads have been shown to have a median accuracy of more than 99.9% [123].

Another approach is Nanopore Sequencing by ONT (Oxford Nanopore Technologies [48]), which is based on flow cells that contain protein nanopores set in an electrically-resistant polymer membrane, connected via electrodes to a channel in the sensor chip that is able to measure the electric current flowing through the nanopore. By loading

the DNA molecule onto the flow cell and passing the DNA through the pore, the current is disrupted in a specific way depending on which base is currently being translocated [221, 73, 99]. These current disruptions can be measured by the sensor in real time and decoded using basecalling algorithms to determine the DNA sequence [73]. In contrast to the PacBio technology, where read lengths are limited through the insert size of the SMRTbell template (10-30 kb long inserts are used [123]), the nanopores allow the passing of DNA molecules regardless of their length, as long as it is possible to extract high molecular weight (HMW) DNA [123]. Read lengths of 10-100 kb and also greater than 100 kb up to several Mb (ONT ultra-long reads) are possible, at the expense of lower correctness: Logsdon et al. report 87-98% accuracy [123].

1.2.3 Chromosome conformation capture with Hi-C and Pore-C

Chromosome conformation capture [49] allows the identification of long-range interactions between distant loci in the genome by examining how the chromosome is folded in three-dimensional space and which chromatin segments then are close to each other.

One of the most prevalent methods is Hi-C [120]. Here, cross-links between chromatin segments are created, the DNA segments are cut and ligated to the respective other cross-linked segment, after which the DNA is cut into fragments. Some of these fragments now contain the junction between the cross-linked chromatin segments. Since the restriction sites had been biotinylated, the relevant junction segments are now identifiable via the contained biotin and can be sequenced via massively parallel sequencing such as standard Illumina paired-end sequencing. This method enables the sequencing of DNA sites which are distant from each other in the DNA sequence, but in close proximity in the three-dimensional conformation of chromosomes in the nucleus. Contacts between DNA loci further apart than 20 kb can be created this way [120]. Thus, the created reads are still short – with all the benefits of short-read sequencing – while containing long-range information.

One shortcoming of Hi-C is that only pairwise contacts can be identified [53], but no higher-order contacts, since short Illumina reads are used. Pore-C, a technique that combines nanopore sequencing with chromosome conformation capture, is able to overcome this challenge. Here, DNA libraries are created via restriction and proximity ligation of cross-linked chromatin segments as described previously. Then, the ligated and restricted fragments are concatenated into Pore-C reads, so that multiple contacts can be represented. The resulting so-called concatemers are then sequenced via nanopore sequencing, allowing for long-range contact information from multiple loci [53, 40].

1.3 DNA variation: The parts that make a difference

Large parts of the DNA sequence in individuals of the same species are identical, meaning that all humans share the vast majority of their DNA. It is those small portions of DNA that differ which, together with environmental influences, define an individual's phenotypic traits. These differences are called *variants*. These can manifest in multiple types: Among the most common classes of variation are *single nucleotide polymorphisms* (SNPs), or single nucleotide variants (SNVs), where two DNA sequences differ in one single base. Other small variants are *insertions* and *deletions* (indels), which are fragments of inserted or deleted sequence that are typically below 50 bp in length [56]. Insertions and deletions longer than that are classified as *structural variants* (SVs), alongside with inversion polymorphisms, which are fragments of inverted sequence, and copy number variations (CNVs), where insertions and deletions lead to gains and losses of DNA segments that result in varying copy numbers of the affected segment in individuals [231].

On average, approximately 4 million SNPs are contained within a human genome [203], rendering them the most frequent type of variation, followed by indels with estimated numbers of around 1 million, taking up around 3.63 Mbp [203, 119]. Structural variants are less frequent – current estimates list 25,000-35,000 [203, 119] –, but they are larger in size and can take up lengths of several million base pairs [160, 56]. In total, the average length of all variation in the genome is estimated as 44 Mbp (23.2 Mbp of which are inversions), which corresponds to 0.76% of the full genome [203].

Such variations contribute to individual phenotypic differences, including physical traits and susceptibility to diseases [65, 110]. Analysing genetic variation is crucial for advancing studies in genomic medicine, as SNPs and other variants have been shown to be associated with many common types of diseases, including diabetes, cancer, coronary heart disease, or migraine [94, 11, 80, 86, 100].

Variants that arise in the germ cells (sperm and egg cells) can be inherited. These *germline variants*, when passed on to the child, will be present in all cells of the individual. In contrast, *somatic variants* occurring in somatic and therefore non-reproductive cells are not inheritable [228]. They arise in a single cell – for instance, as a result of DNA damage or DNA repair errors – and may accumulate by cell division. Both germline and somatic variation may contribute to diseases [228].

In order to detect variants in an individual (variant calling), its genome can be compared to a *reference sequence* of its species. This represents a genome assembly (Section 1.5) with very high accuracy, enabling it to serve as the 'blueprint' of this species' DNA. In aligning an individual's DNA sequence to the reference and comparing the corresponding

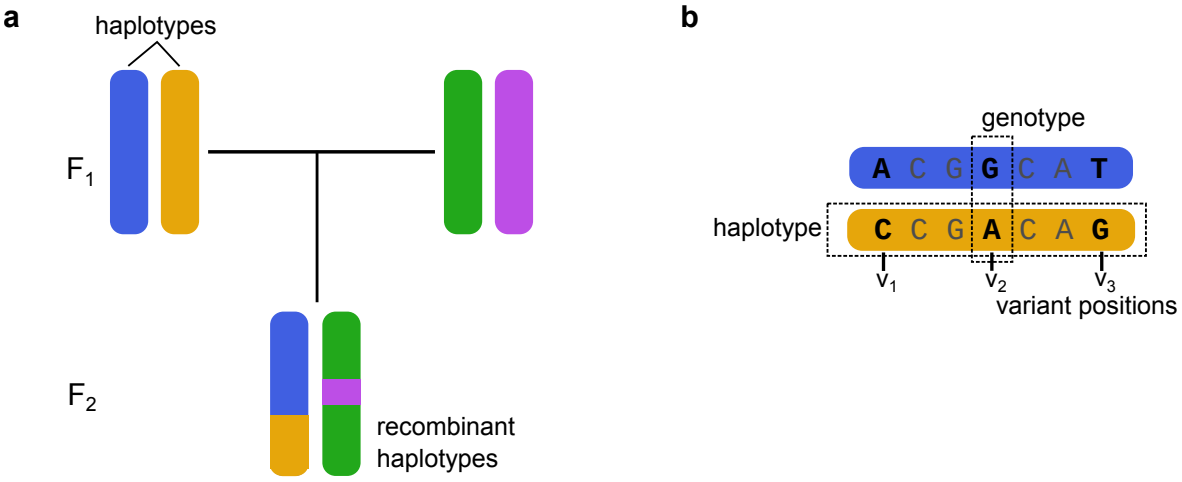


Figure 1.2: Haplotypes and recombination. a. Inheritance of paternal and maternal chromosomes. F_1 and F_2 denote the parent and the child generation, respectively. Haplotypes of one parent are given in blue/yellow, haplotypes of the second in green/purple. During meiosis, the two parental haplotypes cross-over and create recombined haplotypes in the child. **b.** Genotype and haplotype notation. Seven loci are shown, three of which are heterozygous. The genotype is denoted by the two alleles at a locus, while the haplotype marks the sequence of all alleles on the same homologous chromosome.

bases, genetic variation can be identified.

1.4 Haplotypes and inheritance

Humans are *diploid*, which means that each chromosome is present in two copies, one of which is inherited from the mother, one from the father. The two copies of a certain chromosome are called *homologous chromosomes* (homologs). The creation of reproductive cells, the gametes (egg and sperm cells in humans), happens in a process called meiosis. During this process, the pairs of homologous chromosomes align next to each other in the nucleus. Here, the arms of the two chromosomes in a pair may cross over and exchange genetic material, so that part of the paternal chromosome is transferred to the maternal one and vice versa (see Figure 1.2a). This process is called *recombination* and is one of the driving forces of genetic variability. The genomic positions where this cross-over happens are called *meiotic recombination breakpoints*. With increasing number of generations and, therefore, recombination events, the size of shared blocks of common ancestral DNA decreases. Due to recombination, human genomes are a mosaic of their ancestors' genomes. This principle represents an important cornerstone of the field of population genetics (see Chapter 2).

The two bases located at one position (locus) in the homologous chromosomes form

the two alleles of said locus. Together, these alleles shape the *genotype* at this position. A haplotype is then the collection of alleles on the same chromosome copy (see Figure 1.2b). In particular, alleles on the same haplotype are passed on jointly to the next generation. The process of determining which alleles lie on the same chromosome is called *phasing* or *haplotype assembly*. Knowledge about haplotypes is crucial for identifying variants that are associated with disease, where the co-occurrence of variants on the same chromosome copy can lead to mutual influence of effects [204].

Due to their importance, many methods to reconstruct haplotypes of individuals have been developed over the years. Most commonly, these are distinguished into three types: Reference-based (or read-based) phasing, statistical phasing and genetic phasing.

Reference-based phasing refers to determining haplotypes based on the sequence reads directly. First, reads are aligned to a reference sequence. Since one sequence read provides information about the biological proximity of genetic variants (as they are all on the same physical molecule), overlaps between reads can be employed to stretch this information over longer distances. This method relies on the availability of reads that span two or more heterozygous variants and therefore contain haplotype information. Hence, it profits from the increasing availability of high-quality reads and increased read lengths. Methods include HapCUT [15, 58] and WhatsHap [133].

Statistical phasing builds on genotype data present for many individuals from a population. It exploits the principle that due to shared ancestry, two individuals contain shared haplotype blocks. Statistical phasing approaches are typically based on maximum likelihood models that rely on haplotype frequencies at each variant site, so that rare variants cannot be reliably detected. Also, it has been shown that accuracy of these methods is dependent on the cohort size and decreases for small populations [51]. Common methods are SHAPEIT [51, 50] and Eagle [126].

Genetic phasing or trio-based phasing aims to infer haplotypes of a child sample based on information provided by the parental genomes. Applying basic Mendelian rules of inheritance, according to which exactly one allele is inherited from the mother and one from the father, haplotypes can be inferred for loci where the genotype of at least one parent is homozygous. As an example, if the genotype at one locus is C/G for the father, G/G for the mother, and C/G for the child, the 'C' allele can only have been inherited from the father and is thus assigned to the paternal haplotype. This process requires knowledge about the genotypes of the parents and does not work if all trio samples are heterozygous at a

locus. This principle has been implemented within WhatsHap [69] and is also common among modern genome assembly methods [35, 169, 107] (see Section 1.5).

Depending on the availability of data from diverse sequencing technologies, this separation into distinct types of methods nowadays is often softened, and several hybrid methods employing different aspects exist [50, 125, 82, 226]. While phasing in the diploid case is covered by many methods, it poses more challenges in the case of polyploid genomes, where there are more than two haplotypes (Section 1.7). In particular, haplotypes in diploid genomes are complements of each other (Figure 1.2b), while in polyploids, there can be multiple haplotypes with the same sequence. We target this problem in Chapter 3 and propose a method for polyploid phasing.

Modern methods for genome assembly also allow for direct resolution of haplotypes at the time of assembling the genome. This *de novo* haplotype assembly is considered in Section 1.5.

1.5 Genome assembly

Despite the advances in DNA sequencing technologies that allow for high-quality long-range information (Section 1.2), there is no single technology that is able to sequence the complete human genome from end to end. Therefore, after sequencing, the whole genome must be reconstructed from the reads. This process is called *de novo genome assembly*, one of the most fundamental tasks in the field of genomics and computational biology. This task is sometimes compared to solving a jigsaw puzzle [203] where sequencing reads are put together to form entire chromosomes. As many parts of the genome may be very similar or identical to each other, resolving these repeated sequences is particularly challenging.

The first working draft assemblies of the human genome published by Celera Genomics and the Human Genome Project [213, 96, 97] were the result of tremendous laboratory and costly efforts over multiple years. 20 years later, the remaining 200 Mbp of missing sequence from previous assemblies were filled and an assembly was published that covered the full genome from end to end (telomere-to-telomere) [146]. The previous reference suffered not only from missing and unplaced sequences, but also from assembly errors that led to biases in downstream genomic analyses such as imbalanced detection of insertions and deletions (Section 1.3) [3]. Assembling genomes of at least the same quality compared to the working drafts from years back is a routine step nowadays. This is the result of both technological and algorithmic advancements.

Typically, reconstructing the genome based on sequence reads is done by identifying

overlaps between reads and assembling overlapping parts to longer fragments. This process inherently benefits from long sequence reads and high *coverage*, which denotes the number of times the genome is covered by all sequencing reads. High coverage enables assembling the genome with as few gaps as possible. A major milestone for genome assembly methods was therefore the advent of sequencing technologies producing not only long reads, which were previously error-prone (such as ONT reads), but reads which are both long and accurate at the same time (in particular PacBio HiFi, see Section 1.2.2). The second major advancement was the evolution of ultra-long reads exceeding 100 kb in length.

Modern assemblers which can fully take advantage of HiFi and ultra-long reads are hifiasm [34] and Verkko [169]. They perform graph-based assembly, where accurate reads are first used to build an assembly graph. The methods slightly differ in the type of assembly graph they employ internally: In the case of hifiasm, an overlap graph is constructed; in the case of Verkko, a De Bruijn graph (DBG) is built. In the construction of an overlap graph, overlaps are computed for every pair of input reads. Each node is a sequence and an edge marks the overlap between two sequences. In a De Bruijn graph, for comparison, the reads are split into k-mers (fragments of length k), and each node is a distinct k-mer. An edge is drawn between two nodes when they share an overlap of (k-1) [37].

This principle is not new, as traditionally, many assembly methods internally used overlap graphs and yielded contigs as assembled fragments [39, 108, 106, 175]; many of which were based on the overlap-layout-consensus approach [63]. Usually, these graphs consist of linear node sequences that are separated by branching structures and tangles in the presence of heterozygosity and repeats. The linear structures can be extracted and output as contigs. Then, taking all reads making up this contig, a consensus sequence is built using the most common base at every contig position [63]. These contigs are output as primary assemblies. Sequence fragments representing alternative alleles may be output as alternate assemblies. Using additional data types that yield long-range information (such as optical maps [139]), these fragments can optionally be oriented along the genome and extended to scaffolds.

However, this strategy of assembling linear contigs with subsequent scaffolding nowadays is no longer needed. With current assemblers like Verkko and hifiasm that build assembly graphs with long linear components of several megabases in length [169], genome assembly can fully be performed in 'graph space'. As assembly graphs are constructed from the raw sequencing data, they are intended to contain the correct genomic path as well as information about heterozygosity between the reads. Via alignment of ultra-long reads, for instance, these genomic paths can be traced and output as assembled contigs

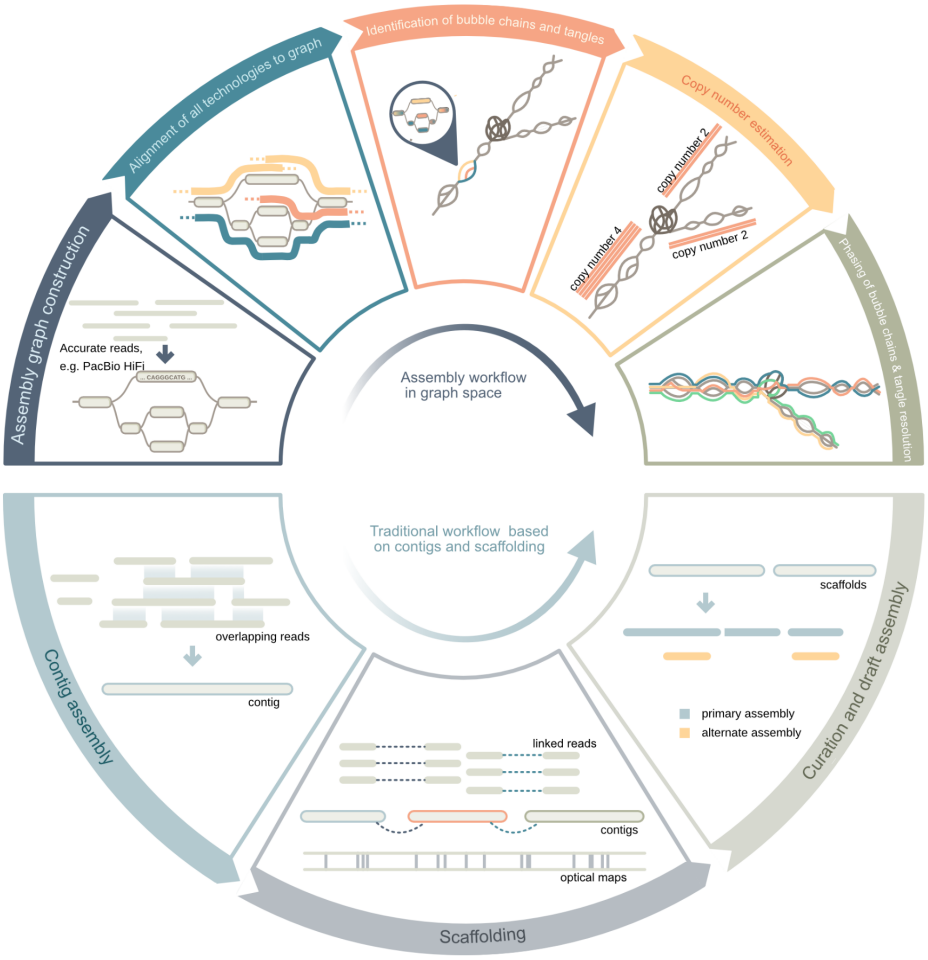


Figure 1.3: Overview over the assembly process. Top: Graph-based assembly. Bottom: Traditional assembly using overlap-layout-consensus approach.

[169]. These principles are illustrated in Figure 1.3.

1.5.1 Haplotype-resolved assembly

Until a few years ago, it was not possible for assemblers to directly distinguish between the homologous chromosomes during assembly [113]. For haplotype-resolved assembly, additional data types are needed to match heterozygous regions over long distances. These could be parental reads to perform trio binning [108] or Hi-C reads, both of which are supported by Verkko and hifiasm [169, 35, 36]. Another option is single-cell strand sequencing (Strand-seq) [177], which is used in haplotype assembly methods such as Graphasing [87] and the PGAS pipeline [158, 56]. Resulting phased assemblies from PGAS are highly accurate and suitable for follow-up analyses (see Chapter 2).

When these data types are available, high-quality assemblies of diploid genomes are

achieved routinely thanks to modern assembly methods. However, for polyploid genomes (Section 1.7), performance is still limited and methods are lacking. In Chapter 4, we therefore present a method for haplotype-resolved de novo assembly of tetraploid potato based on genetic data from an offspring panel. Chapter 5 covers an approach for haplotype-resolved assembly of diploid genomes in a parent-child trio. It is based on an assembly graph containing the three samples in a *pangenome*, a concept which is therefore introduced in the following (Section 1.6).

1.6 Pangenomes

Since the publication of the first sequenced human genome in 2001 [96, 213], reference sequences like this have been used as a baseline for all kinds of genomic analyses; they are treated as a blueprint of a species' genome (see Section 1.3). The initial release of the human reference has been continuously improved in quality and completeness [97]; for the past years, GRCh38 has been the standard reference sequence [182]. However, it also still contained nearly 200 Mbp of unresolved sequence and multiple errors, especially in more complex regions [203]. These remaining gaps could only be resolved recently, when the process of creating an ever-improved human reference sequence culminated in the release of a "telomere-to-telomere" (T2T) assembly of a human genome in 2022. This sequence, termed T2T-CHM13, was presented as the first fully complete assembly [146].

Although this allowed an unprecedented level of accuracy to be achieved, a single reference sequence is merely an approximation of the human genome, but cannot capture the differences between two individual genomes in the many complex and very variable regions of the genome [203]. *Reference bias* describes the lack of alternate alleles of variants from the reference, due to which subsequent read alignment often misses variants [119]. Read mapping and variant detection with the help of a single reference sequence therefore may be inaccurate. In particular, this concerns larger structural variants, which make up tens of megabases of sequence within human genomes [56] (Section 1.3).

Additionally, one single sequence can never capture the diversity of human genomes across different populations, which often show substantial differences, a limitation which is long known [118]. Studies showed later that humans from different populations contained considerable amounts of DNA that did not map to any part of the reference sequence [118, 190]. For example, a sequencing effort of 910 individuals of African descent revealed 296 Mbp of sequence not contained within GRCh38 [190].

Thus, the representation of the human genome offered by a reference sequence is inexact. To fully capture human diversity (and also the diversity within other species),

other reference structures are needed. Instead of the 'linear' representation of a single sequence, a more exact structure would contain a collection of references from multiple individuals of a species, called a *pangenome* [206].

The Human Pangenome Reference Consortium (HPRC) followed the idea to integrate multiple donor genomes, all in the quality of reference sequences, in a shared reference resource that deviates from the linear sequence used previously [218]. As proposed in a review by Taylor et al. [203], the three main components that define a pangenome are i) a set of haplotype-resolved assemblies in T2T or near-T2T quality, ii) an alignment of these haplotypes to each other, iii) functional annotations of each included sequence. There are multiple ways to represent such structures [203, 28, 144, 174], of which graph-based representations are the prevalent option [103, 71, 196, 70, 114, 149, 61].

To this end, the HPRC led an effort to create a pangenome representation of the human genome [119], presented in a graph structure. This pangenome reference consists of 47 genomes from donors of diverse ancestries from the continental populations of Africa, America, Asia and Europe and has shown to contain 119 Mb of polymorphic sequence not present in GRCh38 [119].

Many tools to construct pangenomes from sequences have been proposed, such as Minigraph [114], Minigraph-Cactus [91], and PGGB (Pangenome Graph Builder) [70], all of which were used in the construction of the current pangenome reference [119].

In comparison to linear references, pangenomes allow more exact read mapping and variant detection [203, 119]. Pangenomes also allow the representation of complex types of variation – including structural variants and tandem repeats [61] – which are too complex to be represented in one linear sequence [203]. This enables the analysis of genomic regions that have previously been inaccessible and the discovery of previously missed variants [196]; one reason for the rapid development of pangenome-based tools [57, 33, 59, 90, 195].

Apart from the HPRC, further pangenome projects are ongoing, some of which focus on specific populations, such as the Chinese Pangenome Consortium [68] or the draft Arab pangenome [145], while others refer to non-human species such as tomato [235], rice [163], and other agriculturally important plants [52, 203].

In Chapter 5, we build a pangenome graph containing the genomes of a mother-father-child trio in an effort to use the shared sequence present in the graph-based representation for haplotype-resolved assembly, as well as for assembly validation.

1.7 Polyploidy

While human genomes are *diploid*, meaning that each chromosome exists in two copies, for many species, this is not the case. *Polyploidy* describes the existence of more than two genome copies and is often seen in plant genomes. Most species of green plants (*Virdiplantae*) are polyploid [89]; according to Heslop-Harrison et al., "about half of all plants – both crops and species in their native habitats – are recent polyploids with chromosome sets from two or more ancestors" [89]. Other sources estimate that about 70% of all angiosperms (flowering plants) experienced polyploidization events in their evolutionary history [137], and this includes the agriculturally most important food and crop plants. The soybean (*Glycine max*), for instance, has been shown to have undergone polyploidization events in the past [74, 181]. It is an important protein source and additionally, like all legumes, important for agriculture, as they enrich the soil due to their ability to symbiose with microorganisms that fix nitrogen, which can then be made available in the soil via legume decomposition [181, 101] Further important polyploid plants [109] include crops such as tetraploid cotton (*Gossypium arboretum*) [111], hexaploid bread wheat (*Triticum aesthivum*) [208, 209], sugarcane (*Saccharum spp.*, with highly variable ploidy levels from 4x to 16x) [233, 16, 84], or tetraploid potato (*Solanum tuberosum*) [210], as well as economically important plants like tobacco and oil palms [192, 156] or ornamental plants [55, 170, 88].

Two forms of polyploidy are distinguished: *Allopolyploidy* and *autopolyploidy*, a notation that dates back to 1926 [102]. Autopolyploidy results from whole genome duplications within the same species, while allopolyploidy results from hybridizations of genomes from different species. Both intra- and interspecifically hybridized polyploids are often the result of meiotic events, such as unreduced gametes [134] leading to whole-genome duplication, and less likely due to hybridization followed by chromosome doubling [109, 135].

As a consequence, allopolyploids contain a complete set of homologous chromosomes for each subgenome in the nucleus, with differing divergence between these subgenomes [67, 135]. The involved species are often diverged from a common ancestor and the chromosomes between the diverged subgenomes are called *homeologous* to each other [89]. As an example, bread wheat (*Triticum aesthivum*) is an allohexaploid consisting of three subgenomes of three wild grass species [29]. Its genome can be described as AABBDD and involved two hybridizations between the three species. The first one happened between *Triticum urartu* (genome A) and an unknown relative of goatgrass (*Aegilops speltoides*, genome B) dating back to 0.5 to 0.8 million years ago [29]. The result-

ing hybrid, a tetraploid wheat (AABB) then hybridized with Tausch's goatgrass (*Aegilops tauschii*, genome D) 8,000-11,000 years ago, resulting in the common bread wheat [29].

In contrast, autopolyploids contain multiple sets of homologous chromosomes, such as the autotetraploid potato, where each chromosome is present four times (4n = 48).

Polyploid crop plants have been associated with many desirable traits such as adaptability to different stress factors such as harsh environments (high altitude, drought), increased growth rate and gigantism, which describes the increased size especially of harvested parts of crops [89]. As these are important traits for crop plants in particular [89, 217, 232, 62, 16], for some plants, breeding polyploid species is preferred over their diploid counterparts, such as for the kiwifruit [223] and for rice [227].

At the same time, the analysis of polyploid genomes creates additional challenges. For genome assembly (Section 1.5), especially haplotype-resolved assembly, polyploid genomes pose considerable challenges. As there are more than two haplotypes, the distinction between them can be difficult. This especially affects autopolyploids, whereas for allopolyploids, if the subgenomes are sufficiently distinct, the assembly problem is similar to assembling multiple separate diploid genomes.

Additionally, polyploidization comes with increased genomic complexity, often as a result of the hybridization processes [16]. Polyploid genomes tend to have higher heterozygosity, allele divergence, and increased frequency of meiotic recombination [152], each contributing to higher diversity [176]. Also, high repeat content can be another contributing factor to the difficulties around plant genome assembly, such as in the potato [200, 17, 189]. We challenge these problems in Chapters 3 and 4.

1.8 File formats

This section introduces the file formats that are used in analyses throughout this thesis. They are commonly used for storing raw sequence data, sequence alignment data and assemblies, both in linear and in graph space.

1.8.1 FASTA/FASTQ format

The sequencing data that are produced during DNA sequencing (Section 1.2) are typically stored in FASTA or FASTQ format. FASTA files were first described by Lipman and Pearson in 1988 [151] as part of a DNA and protein sequence alignment software package, and have remained the standard format for representing genomic sequences or also amino acid sequences ever since. Both FASTA and FASTQ are text-based files that consist of a

unique identifier for the sequence and the actual sequence data, all in text format. Thus, a DNA sequence in a FASTA file is encoded in two lines: The first line starts with a '>' symbol to indicate the identifier line, followed by the identifier itself, which can be any character string. The second line contains the actual sequence of bases, usually – in the case of DNA – composed of the letters 'A', 'C', 'G' and 'T' (Section 1.1) [151]. Other letters are possible as well; for instance, to represent unknown bases, 'N' is a common placeholder. An example for the FASTA format is given below:

```
>seq1
AACCAAGGTTGG
>seq2
GGTGGAACT
>seq3
GAGATATATATCCC
```

In comparison to that, the FASTQ format provides an extension to FASTA by allowing to store more information, including some about the sequencing run, and to assign a quality value to each nucleotide produced [42]. A sequence read in a FASTQ file consists of four lines: the first begins with a '@' symbol to denote the identifier line, while the second line contains the actual sequence of nucleotides. The third line merely consists of a separator symbol ('+') and the fourth line contains additional base call quality scores that give information about the quality of the base at that particular position. An example of the FASTQ format is shown below [1]:

```
@SIM:1:FCX:1:15:6329:1045 1:N:0:2
TGGACCT
+
<>;\#\#=>
```

1.8.2 SAM/BAM format

After alignment of sequence reads to a reference sequence, another set of reads, or an assembly (Section 1.5), the aligned data are stored in a sequence alignment map (SAM) or binary alignment map (BAM) file [115]. Each aligned sequence is represented by a tab-delimited line containing information about the alignment, such as the mapping position, the length of the mapped sequence, its quality and a 'CIGAR' string that encodes

[1] modified from https://help.basespace.illumina.com/files-used-by-basespace/fastq-files

```
S    n1    ACGTGACGCG
S    n2    CGCGTGCATGAGAC
S    n3    GCGACGTTTGAG
S    n4    AGACGTA
S    n5    CGCGTTACGT
L    n1    +    n2    +    4M
L    n1    +    n3    +    3M
L    n5    -    n2    +    4M
L    n5    -    n3    +    4M
L    n2    +    n4    +    4M
L    n3    +    n4    +    2M
```

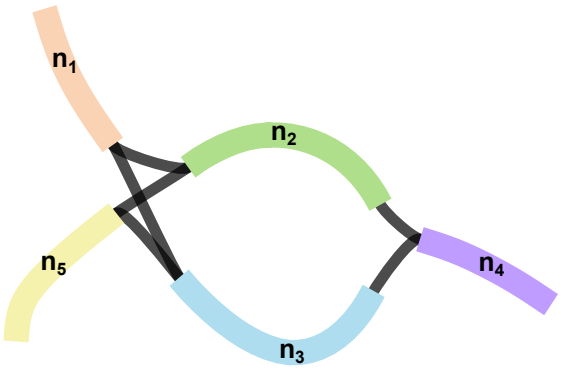


Figure 1.4: Example GFA file. Left: File structure of a graph consisting of five nodes $n_1, \ldots, n_5$ and six edges. Right: Visualisation of the corresponding sequence graph using Bandage [222].

the changes (matches, mismatches, deletions and insertions) that would have to be applied to the query in order to exactly match the reference [115]. Optionally, there may be a header containing metadata before the alignment lines. Similarly, BAM files are binary encoded, compressed versions of their SAM counterparts. Example:

```
@HD VN:1.6 SO:coordinate
@SQ SN:ref LN:45
r001 99 ref 7 26 4M2I4M1D3M = 37 39 TTAGTAGATACTG *
```

In this example, the header lines (starting with '@') denote that version 1.6 of the SAM specification was used, that the contained alignments are sorted by coordinate, i.e. first by reference name and second by first aligned position, and that the name of the reference sequence shall be given by 'ref' and the reference sequence is 45 nucleotides in length.

1.8.3 GFA format

The graphical fragment assembly (GFA) format describes sequence graphs, such as assembly graphs or pangenome graph representations (Section 1.5, Section 1.6) [207]. The main components are the actual sequences and the overlaps between them. The sequences represent the nodes in the graph and are denoted as *segments* in the GFA format. These are described in tab-separated lines starting with the identifier 'S' which also contain a unique segment name and the actual sequence [207]. Overlaps between two segments lead to an edge between the two corresponding nodes in the graph and are denoted as *links*. The link lines in GFA files, starting with 'L', contain the two segment names that are connected, their orientation, as well as the length of the overlap [207]. Apart from

segments and links, other types of records are possible to store within GFA files, but are omitted here. An example is given in Figure 1.4.

1.9 Evaluation statistics

To evaluate the quality and correctness of an assembly or phasing, several metrics can be computed. Here, we describe the phasing evaluation statistics which were used in Chapter 3 to evaluate polyploid haplotype phasing. For this, we introduce the Hamming rate and the switch error rate. For assemblies, a commonly used metric is the N50 value, which we also describe here.

The material from the following Section is taken from [184], the publication on which Chapter 3 is based.

Hamming rate For ploidy k, a set of ground truth haplotype sequences $h = \{h_1, ..., h_k\}$ and predicted haplotypes $h^* = \{h_1^*, ..., h_k^*\}$, we compute the number of Hamming errors HE as

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

where S_k represents the permutation group on $\{1, ..., k\}$ and $d_H()$ the Hamming distance between two sequences. The *Hamming rate (HR)* is then defined as the sum of Hamming errors divided by the total number of all phased variants. If subtracted from 1, the Hamming rate is equivalent to the *reconstruction rate* and the *correct phasing rate* presented in [141] and [224], respectively.

Switch error rate A well established evaluation metric for diploid phasing is the *switch error rate (SER)*, which can be adapted for the polyploid case. Instead of counting the number of incorrect alleles on each haplotype, the SER counts the minimum number of switches, i.e., how often the assignment between predicted and true haplotypes must be changed in order to reconstruct the true haplotypes from the predicted ones. The polyploid extension of the switch error was introduced as the *vector error rate* in [20].

Formally, for every position j let Π_j be the set of one-to-one mappings between 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 then defined as:

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

where m is the number of variants and $d_S(\pi_i,\pi_{i+1})$ the number of different mappings between π_i and π_{i+1}.

If the genotype of h^* is not equal to the genotype of h for every position, the set $\Pi_1 \times \ldots \times \Pi_m$ is empty and the vector error cannot be computed. Therefore, we compare only those positions for which the predicted genotype is correct and additionally report the fraction of *missing variants (MV)*, that is, either unphased or incorrectly genotyped variants.

Phasing tools may not phase the entire input region as one set of haplotypes. If the phasing between two consecutive variants is too uncertain (for read-based phasing: if not enough reads cover both variants, Section 1.4), the phasing might be split into *blocks*. For the evaluation in Chapter 3 [184], we applied the HR and SER on all reported phasing blocks separately and aggregated them. For this, the number of respective errors was summed up and divided by the total number of variants (HR) or by the total number of variants excluding the first variant in every block (SER).

N50 value The N50 is a commonly used metric to evaluate contiguity of assemblies. Frequently, genomes or haplotypes are not assembled in full chromosomes, but fragmented into several contigs. Given a set of contigs ordered by their lengths, the N50 value corresponds to the length of the shortest contig at 50% of the total assembly length. Let this length be n in an example assembly. Then, the N50 describes the value at which 50% of the assembly are present in contigs at least as long as n. Essentially, it presents a length weighted median of the contig lengths. In de novo assemblies, it is desirable to achieve N50s as large as possible, as this stands for contiguous assemblies composed of long contigs.

The N50 metric can also be applied in phasing evaluation. In the evaluation of the polyploid phasing method WHATSHAP POLYPHASE described in Chapter 3, it was used as a metric for contiguity of the phased blocks. The N50 block length is then the smallest block length needed to cover 50% of the considered genomic region when using only blocks of that size and larger.

1.10 Hidden Markov Models

The method presented in Chapter 2 that computes ancestry estimates of genome assemblies uses a Hidden Markov Model as underlying statistical model. This section gives a general introduction about these models, covering topics that are relevant for the work of Chapter 2.

A Hidden Markov Model (HMM) is a statistical model to represent systems that undergo transitions between a series of states, where the states themselves are not directly observable. It can be understood as the extension of a *Markov chain*, which is a mathematical model that describes a system which transitions between different states according to certain probabilistic rules [95, 236]. Formally, a Markov chain consists of the following components:

1. A finite or countable set of possible states $S = S_1, S_2, \ldots, S_N$
2. Transition probabilities which model the probability to transition from state S_i at time $t-1$ to state S_j at time t, given by $T_{i,j} = \mathbf{P}(X_t = S_j \mid X_{t-1} = S_i)$, where X_t describes the state of the system at time t
3. The initial state distribution which describes the probability to start in state S_i: $\pi_i = \mathbf{P}(X_1 = S_i)$

The key property of Markov chains is the so-called *Markov property*, which assumes that the probability of transitioning to a specific state depends only on the current state:

$$\mathbf{P}(X_{t+1} = S_j \mid X_t = S_i, X_{t-1} = S_k, \ldots, X_1 = S_l) = \mathbf{P}(X_{t+1} = S_j \mid X_t = S_i)$$

This implies that the state of the system at current time point t only depends on the state at time $t-1$, while all previous time points have no influence on the current state [95].

The Hidden Markov Model builds on the model of a Markov chain, with the difference that the states in a Markov chain are directly observable, while in the HMM, they are *hidden*. Each hidden state has a probability of emitting an observation sequence O. If the system is in state S_i, the probability of observing o is given by $E_i(o) = \mathbf{P}(O_t = o \mid X_t = S_i)$.

Thus, a Hidden Markov Model formally consists of a set of hidden states S, transition probabilities T and emission probabilities E, as well as an initial state distribution identical to the Markov chain [165, 19].

Typical goals of applying this model include inferring the most likely sequence of hidden states given a sequence of observed emissions, calculating the likelihood of a given observation sequence, or estimating the model parameters that best explain the observed

data. Application areas are diverse and include many other fields apart from genomics, such as natural language processing [8], finance [193] or neuroscience [25].

The most common algorithms to solve these common tasks are the **Viterbi algorithm**, which computes the likeliest state sequence $S_{1:T} = (S_1, \ldots, S_T)$ for a given emission sequence $o_{1:T}$, the **forward algorithm** to compute the likelihood of an observed sequence $o_{1:t}$ being produced by a given HMM, and the **forward-backward algorithm** [165], which will be described in the following. Details for these algorithms can be found in [19].

The algorithm we apply to the model in Chapter 2 is the forward-backward algorithm. It can be used to compute the probability of each state x at any time t, given the observed emission sequence $o_{1:t} = (o_1 \ldots, o_t)$. The algorithm consists of three steps. First, *forward* probabilities $\alpha(x)$ are computed, which give the likelihood of each state x to end in this state at time t, given the observations $o_{1:t}$. This procedure coincides with the forward algorithm.

$$\mathbf{P}(X_t = x, o_{1:t}) = \sum_{x'} \mathbf{P}(X_{t-1} = x', o_{1:t-1}) \mathbf{P}(X_t = x | X_{t-1} = x') \mathbf{P}(o_t | X_t = x)$$

In a second pass, we compute *backward* probabilities, which denote the probability to observe the remaining observations starting from t, $o_{t+1:T}$, given that we start in state x. These probabilities are given by:

$$\mathbf{P}(o_{t+1:T} | X_t = x) = \sum_{x'} \mathbf{P}(o_{t+2:T} | X_{t+1} = x') \mathbf{P}(X_{t+1} = x' | X_t = x) \mathbf{P}(o_{t+1} | X_{t+1} = x')$$

Finally, the results of both computations are combined to compute the *posterior* probabilities:

$$\mathbf{P}(X_t = x | o_{1:T}) = \frac{\mathbf{P}(o_{1:t}, X_t = x) \mathbf{P}(o_{t+1:T} | X_t = x)}{\mathbf{P}(o_1, \ldots, o_T)}$$

Given an observed sequence of emissions, this algorithm will result in the estimated probability to be in a certain state at any time point.

Chapter 2

Local ancestry inference

The work presented in this chapter was conducted as part of a project led by the Human Genome Structural Variation Consortium (HGSVC), aimed at generating haplotype-resolved genomes for a diverse set of human populations [56].

Accurate haplotype assemblies enable many downstream analyses, one of them being local ancestry analysis, which aims to determine the ancestral origin of specific segments of the genome.

My specific contribution to this project involved performing the local ancestry analysis using a Hidden Markov Model (HMM)-based method which was developed specifically for this purpose. This analysis aimed to provide an initial characterization of ancestry patterns across the diverse genome assemblies.

This section presents work conducted as part of a project by the Human Genome Structural Variant Consortium (HGSVC) which led to a publication in Science [56]. My contribution to this project was the implementation of a local ancestry inference method and its application to the assembled samples in order to detect ancestral populations of genomic regions. The analysis of the results has been done in collaboration with Ping-Hsun Hsieh, a co-author of this publication. Some material from [56] is re-used. See Section E.1 for the full list of author contributions and license information.

2.1 Background of local ancestry inference

Local ancestry inference (LAI) is the process of determining the ancestral origin of each position on the genome [219, 27, 18]. Every chromosome in an individual's genome is the product of genetic mixing across many generations, tracing back to matings between individuals from genetically diverse ancestral populations. Over time, these chromo-

somes have undergone frequent recombination events, resulting in each chromosome now being a mosaic of segments from different ancestral genomes. This is especially prevalent in individuals of more recently admixed populations, where the term *admixture* refers to populations emerging from migration effects between previously isolated populations [201]; one example being Admixed Americans such as Puerto Ricans, Peruvians or Cubans, for instance [140, 187].

Local ancestry inference helps understanding demographic history of populations [140, 132, 187]. For Admixed American populations, for instance, it could be shown that their genomes are composed of the three ancestral populations of Africans, Europeans and Native Americans [140, 132]. Analysing the lengths of the determined ancestry tracts allowed to date back admixture and to identify several likely pulses of migration [140]. Analysis of both recent and ancient admixture between populations thus helps shed light on evolutionary processes [132].

In addition to population genetics studies in the context of human demographic history, inferring local ancestry to genomic sites is relevant for the field of personalized medicine. Knowing the ancestral origin of genomic regions allows for ancestry-specific and ancestry-aware Genome Wide Association Studies (GWAS), which aim to find associations between genotypes and phenotypes, to detect variants associated with certain traits or diseases, for instance [211]. GWAS are reproducible only to a limited extent across different populations, especially across admixed ones [65, 187]. It has long been known that population stratification, which describes the presence of systematic differences in allele frequencies in different populations [66] may influence association studies [105]. Knowledge about admixture and population structure – which can be gained through LAI methods – is therefore crucial for determining the ability to transfer GWAS across populations [132].

For example, during the investigation of causes for African Americans having higher rates of kidney failure than European Americans, it was shown that two variants in the APOL1 gene on chromosome 22 (a gene which encodes apolipoprotein L-1) are associated with kidney disease [27, 72]. These variants were found only in chromosomes carried by African donors and absent in all chromosomes carried by European as well as East Asian donors (Japanese and Chinese, specifically) of the tested cohorts [72]. Such findings may then further catalyse the analysis of population-specific positive selection of certain variants [72, 187]. Notably, the inclusion of individuals from different ancestries into biobanks and reference panels allows for the detection of population-specific disease-associated variants and may additionally reduce disparities in personalized medicine.

LAI methods require the availability of genotyped and phased data sets to be used

as a reference panel. The most widely used has been the 1000 Genomes Project [205, 14], complemented by other projects that provide more diverse reference populations, such as the Human Genome Diversity Project (HGDP) [30, 21] and the Simons Genome Diversity Project [128]. The 'mosaic' structure of each individual genome being composed of blocks of shared ancestry is most commonly modelled via the Li-Stephens model [117], which allows to take linkage disequilibrium (LD) into account. Therefore, many LAI methods are based on adaptations of this standard model, with minor differences, and use a Hidden Markov Model as their underlying statistical model. These methods include HapMix [161], one of the first implementations which was only able to take two reference populations at a time into account, and LAMP-LD [18]. Other methods are based on extensions to the original HMM algorithm, such as RFMix [130], which infers ancestries using conditional random fields parametrized by random forests trained on reference panels. RFMix is widely used and referred to as the state-of-the-art.

2.2 A Hidden Markov Model for LAI

Here, we present a simple Hidden Markov Model-based workflow for ancestry inference on genomic segments that leads to a per-region chromosomal plot painting the different regions according to the ancestral population.

The underlying HMM is a version of the standard Li-and-Stephens inspired modelling of haplotypes based on reference panels, with easy-to-adjust parameters. An overview is shown in Figure 2.1. The input consists of the individual's haplotype (sequence s over variants v_i, $i \in \{1,2,\ldots,n\}$) given in a VCF file, as well as a panel of phased reference haplotypes in a multi-sample VCF file. For each haplotype in the reference panel, we additionally know the population the donor comes from. This data is input in a Hidden Markov Model which will be described in more detail later. Briefly, the reference samples mark the hidden states in the panel, where the haplotype sequence (0 or 1) are the emitted observations, and likelihoods of each reference at each variant position are computed based on the forward-backward algorithm. This can optionally be performed multiple times in a bootstrapping approach with differing sample subsets to reduce the computation time, depending on the size of the population panel. The output consists of a list of likelihoods per variant position, denoting the probability that the query can be described by the corresponding reference at this position. Lastly, these probabilities are summed up within each population. Based on the population with the highest probability, the corresponding genomic region is coloured accordingly, leading to a chromosome plot that is coloured by most likely ancestry per region.

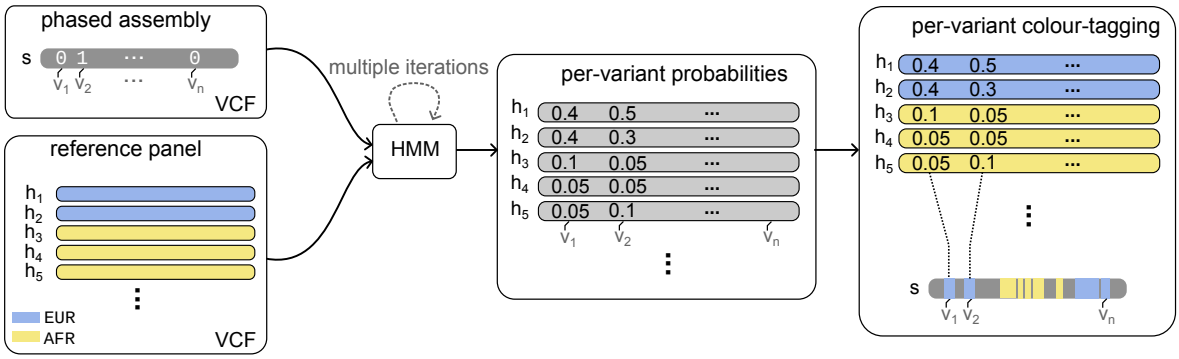


Figure 2.1: Overview of HMM-based LAI. A fully phased haplotype sequence s and reference haplotypes $h_1, \ldots, h_5$ over variants $v_1, \ldots, v_n$ are input into a Hidden Markov Model. The references are part of two different populations, European (EUR) and African (AFR). The HMM computes position-wise probabilities for each sequence in the reference panel, which are summed up per population (the blue values denoting the EUR samples, yellow denoting African samples). Last, the corresponding region in the genome is plotted in the colour code of the population which is the most likely ancestry of that region based on the computed probabilities.

2.2.1 Input: Query and reference panel

The population data can come from any high-quality large-scale sequencing effort; in our study, they are taken from the 1000 Genomes Project [14] as well as the Simons Genome Diversity Project [128]. The 1000 Genomes Project (1KG) provides data from 1,092 individuals from the five super-populations African (AFR), American (AMR), East Asian (EAS), European (EUR), and South Asian (SAS), which each contain numerous population groups. The population assignment is based on self-identification by the donors. The Simons Genome Diversity project further provides a resource over 300 human genomes across 142 diverse populations [128], in order to increase the genetic diversity of the available data sets and to include populations that have previously been missing. The input to the method is then a multi-sample VCF file containing phased variant calls of suitable reference sequences. The second input is another VCF file containing the haplotype-resolved assembly of the query.

The list of variants to be considered is restricted to heterozygous sites which are present in both the panel and the query, and which are phased.

2.2.2 Model definition

The computation of the most likely ancestral population for each genomic position is then performed by setting up a Hidden Markov Model that models the reference haplotypes from the panel as hidden state sequences and the query haplotype as the observed

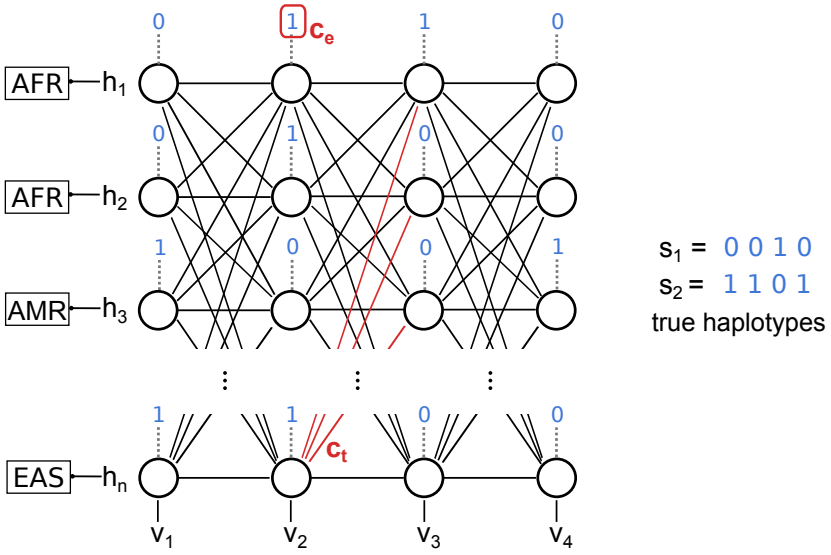


Figure 2.2: Hidden Markov Model for ancestry reconstruction. Hidden state space is given by the reference haplotypes h_1, h_2, …, h_n. Emissions are 0 or 1 and denote the alleles of the haplotype sequences. Each reference has a tag to a population identifier to denote that the underlying ancestries of the reference sequences are known and are part of the hidden states. The underlying true haplotypes of the query are given by s_1 and s_2. Costs for deviations between the allele emitted by a state and the true allele are denoted with c_e, here in reference h_1 at variant v_2. Costs for transitions between states from one position to another are denoted with c_t, here for instance for all transitions from h_n to all other states h_i ($i \neq n$) from variant position v_2 to v_3.

sequence and computing a standard forward-backward algorithm on this HMM. At each variant position, costs are inferred for a) discrepancies between query and reference haplotype and b) switches to a different reference haplotype happening between two variant positions. Thus, the probability will be high for reference haplotypes that largely coincide with the query, for blocks as long as possible. This results in a set of probabilities for each variant position, consisting of one probability per reference haplotype. The probabilities of references that belong to the same superpopulation are then added up to compute the likelihood for every superpopulation at each position.

Assuming the reference population contains n individuals, the hidden states are given by $2n$ haplotype sequences h_i. The states are here referenced by the sample name as well as a notifier for haplotypes 1 and 2. For each hidden state h_i and each variant position $j, j = 1,\ldots,m$, the observable state = $o_{i,j}$ denotes the allele that the reference haplotype (= the hidden state) possesses at this position. We assume a length of m bi-allelic variants, and alleles are encoded as 0 and 1, so that $o_{i,j} \in \{0,1\}$ for all $i = 1,\ldots,2n$, $j = 1,\ldots,m$. Thus, the observable state space is given by $O_{i,j} \in \{0,1\}^{2n\times m}$. The observed emission sequences in our approach are given by the two haplotypes of the assembly to be analysed. We define

them as s_1 and s_2; as the approach is run twice, on each of the haplotypes independently, for notation simplicity we will refer to each haplotype as s. Then, s is a sequence of alleles of length m, $s_i \in \{0,1\}, i = 1,\ldots,m$.

An example visualisation of the HMM is given in Figure 2.2, with references $h_1, \ldots, h_n$ over four variants.

The problem can then be summarised as follows: Given an HMM consisting of states $H = h_1,\ldots,h_{2n}$ over variants $v_1,\ldots,v_m$, transition probabilities T and emission probabilities E, as well as two observed sequences s_1 and s_2 ($s_1, s_2 \in \{0,1\}^m$). Compute the probability sets $\mathbf{P}(H_t = h \mid s_1)$ and $\mathbf{P}(H_t = h \mid s_2)$ for all states h and all time points t modelling the variant space.

We apply a forward-backward algorithm to fill a scoring matrix $S^{2n\times m}$, denoting the probabilities to be in state h at time t. The emission probability of state h_i at time t, $e(h_i,t)$ is the probability that the allele $s_1[t]$ (equivalent: $s_2[t]$) comes from reference sequence h_i. Thus, we apply the following cost function:

$$E(h_i,t) = \begin{cases} log(e^{-10}) & , s[t] \neq h_i[t] \\ log(1-e^{-10}) & , s[t] = h_i[t] \end{cases}$$

Similarly, we infer costs if we transition between two different reference haplotypes h_i at time t to h_j at $t+1$:

$$T(h_i,h_j) = \begin{cases} log(e^{-1}) & , i \neq j \\ log(1-e^{-1}) & , i = j \end{cases}$$

If we are given recombination rates, we can incorporate those into the computation of transition probabilities according to the Li-Stephens model [117].

Note that S is a negative cost matrix and logarithmic costs are used, so that during the computation of the forward-backward algorithm, log-likelihoods will be computed.

Applying the forward-backward algorithm, forward probabilities are then computed as follows:

$$\mathbf{P}(H_t = h, s_{1:t}) = \sum_j \mathbf{P}(H_{t-1} = h_j, s_{1:t-1})\mathbf{P}(H_t = h|H_{t-1} = h_j)\mathbf{P}(s_t|H_t = h)$$

The computation of backward probabilities works similarly:

$$\mathbf{P}(s_{t+1:T}|H_t = h) = \sum_j \mathbf{P}(s_{t+2:T}|H_{t+1} = h_j)\mathbf{P}(H_{t+1} = h_j|H_t = h)\mathbf{P}(s_{t+1}|H_{t+1} = h_j)$$

And lastly, these results are combined to compute the posterior probabilities:

$$\mathbf{P}(H_t = h|s_{1:T}) = \frac{\mathbf{P}(s_{1:t}, H_t = h)\mathbf{P}(s_{t+1:T}|H_t = h)}{\mathbf{P}(s_1, \ldots, s_T)}$$

After computing the posterior probabilities, we average over all probabilities belonging to haplotypes from the same population, resulting in an average likelihood per population per position. Then, the decision for the most likely ancestry estimate can be made based on the highest value or based on cut-offs, as done in the computation of ancestry estimates for the HGSVC project (see Section 2.4). To reduce the computation time when large panels are used, we added a bootstrapping option which allows random sampling of x samples from the panel y times, and averages the results over the y different runs.

Recombination

Our approach does not rely on an additional recombination map being included, but can incorporate it optionally. In case a recombination map file is given, which usually contains variant positions and their genetic distance in cM (centimorgan), we infer this information from the map. If a variant position v does exist in the query, but not in the map, its value is interpolated as follows: Given the positions of the two nearest neighbours (position-wise) p and s of the missing variant v. We denote with $cm(v)$ the genetic position in cM of variant v.

Then, $cm(s) - cm(p)/(s-p) * (v-p) + cm(p)$ gives an approximate genetic distance (in cM) for the interval between the current position and the previous neighbour. Adding this to the distance of the previous position, this gives an estimate for the value at the current position v. The differences $cm(i) - cm(i-1)$ are the final values in the recombination map at position i.

The genetic distance serves as an estimate for the recombination rate between two positions. According to the Li-Stephens model [117], we can then refine the transition probabilities. Given the estimated recombination rate c between two variants v_i and v_{i+1} and the effective population size N, the scaled recombination rate can be defined as $\rho_i = 4Nc$. Given ρ_i, the physical distance between v_i and v_{i+1} in base pairs d_i and the number of known haplotypes n, we define the Li-Stephens transition probabilities [117] as:

$$\mathbf{P}(H_{i+1} = h'|H_i = h) = \begin{cases} e^{-\frac{\rho_i}{n}d_i} + \left(1 - e^{-\frac{\rho_i}{n}d_i}\right)\left(\frac{1}{n}\right) & \text{if } h' = h \\ \left(1 - e^{-\frac{\rho_i}{n}d_i}\right)\left(\frac{1}{n}\right) & \text{if } h' \neq h \end{cases}$$

2.3 The HGSVC project

This section provides a summary of the full study published in [56]. Some material from this publication is re-used. My contribution consisted in the population genetic analysis of the assembled genomes and is described in Section 2.4.

The availability of long read sequencing and additional data types such as single-cell template strand sequencing (Strand-seq)[177] facilitated long-read based genome assemblies and even enabled the separation of the assemblies into distinct haplotypes [56]. Such fully haplotype-resolved human genome assemblies have considerably improved the discovery of all types of variants, in particular structural variants (SVs), and it has been shown that much more SVs could be detected based on long-read enabled phased assemblies [56].

The Human Genome Structural Variation Consortium (HGSVC) [56, 31] previously developed a workflow that combines long-read sequencing and Strand-seq to produce haplotype-phased diploid genome assemblies [158]. In particular, this PGAS pipeline does not require trio data for haplotype phasing. In a large-scale project launched by the HGSVC, this method was employed to create a resource of 70 phased human haplotype assemblies. This resource data set contained 35 individuals from all five 1000 Genomes Project superpopulations (AFR: n=11, AMR: n=5, EAS: n=7, EUR: n=7, SAS: n=5) [205]. 30 of the donors were sequenced with PacBio CLR data and 12 with PacBio HiFi , leading to an overlapping group being sequenced with both technologies.

The resulting assemblies showed high quality (QV > 40) and completeness (N50 > 25 Mb), as well as high phasing accuracy (median switch error rate of 0.12%). The accuracy and quality of the assemblies allowed for many follow-up analyses, focusing on the detection of novel SVs and their genotyping to provide a diverse genome panel. Variant calling was performed using a newly developed method PAV, which directly compares the two haplotype assemblies to a reference genome in order to detect variants [56]. PAV reported three different classes of variants: SNVs, indels (1–49 bp), and SVs (≥ 50 bp). The resulting variant call set contained 15.8 million SNVs, 2.3 million indels, 107,316 large insertions or deletions and 316 inversions. In contrast to that, the analysis of 2,504 short-read sequenced samples from the 1000 Genomes Project [14, 199] led to the discovery of merely 69,000 SVs, highlighting the great potential of long-read assemblies for novel variant detection, especially for structural variation.

2.4 Application: Local ancestry analysis in the context of the HGSVC project

This section reuses material from [56], focusing on work that was contributed by Ping-Hsun Hsieh and myself. My contribution was the application of the HMM-based ancestry inference to selected HGSVC samples and the comparison with RFMix results, which went into the creation of an ancestry plot for a Puerto Rican trio. The analysis of population-stratified variants was contributed by Ping-Hsun Hsieh.

The availability of accurate, high-quality haplotype-phased assemblies allows for many downstream analyses. As such, it provides an opportunity to explore the ancestry and population genetic properties of the genomes and SVs at multiple levels. We estimated ancestries of genomic blocks in the high-quality haplotype-resolved assemblies produced by the HGSVC [56] (see Section 2.3). We created ancestry-coloured plots for Chromosome 1 for all assemblies and put special focus on a trio with Puerto Rican origin.

To do so, we applied both the Hidden Markov Model described previously (see Section 2.2), which was specifically developed for this project, and RFMix [130].

For the reference panel to be used by the HMM and RFMix, we used published phased genomes from predetermined populations from the 1000 Genomes Project. To control for potential biases due to the inclusion of admixed individuals in the reference panel, we chose less admixed 1000 Genomes Project genomes, as determined by ADMIXTURE [7] and consistent with previous studies [132]. The final set contained data from 1203 individuals, including 472 from African populations: Luhya (LWK), Mende (MSL), Gambian Mandinka (GWD), Yoruba (YRI) and Esan (ESN); 381 individuals from European populations: Europeans from Utah (CEU), British (GBR), Finnish (FIN), Iberian (IBS) and Toscani (TSI); 145 from East Asia: Han Chinese (CHB), Southern Han Chinese (CHS) and Japanese (JPT); as well as 186 from South Asia: Telugu (ITU) and Tamil (STU). In addition, as part of our reference panel, we included 19 genomes from Native Americans from the Simons Genome Diversity Project [128] that show little European ancestry. These include individuals of the Quechuan population in Peru, Chane in Argentina, Karitiana and Surui in Brazil, Piapoco in Colombia, and Mayan, Mixe, Mixtec, Pima and Zapotec in Mexico.

Note that for the inference of local ancestry in Puerto Rican individuals, we also specifically explored and set the reference panel to be African, European, and Native American populations given the previously demonstrated population structure of this population, exhibiting a three-way admixture of the three aforementioned continental populations [132, 140].

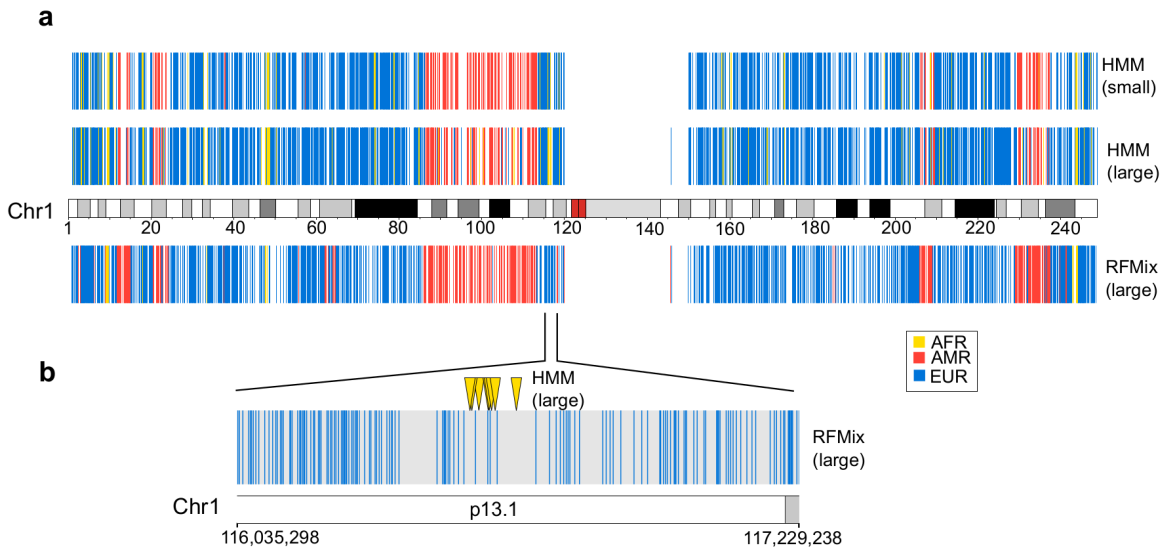


Figure 2.3: Example chromosome colouring. a. Chromosome 1 of sample HG00733 (PUR). One single haplotype is shown. Top: Results for the HMM-based approach run with the bootstrapping approach of 10 iterations and 30 sampled references from AFR and EUR, respectively, and all 19 AMR samples (Labelled as 'HMM (small)'). Below that ('HMM (large)'), results are shown for HMM using the larger panel of 419 samples (19 AMR, 200 EUR, 200 AFR). Bottom: Results for running RFMix on the panel of 419 samples. Genomic regions are coloured depending on the most likely ancestry estimate. **b.** Detailed region between positions 116,035,298 and 117,229,238, showing RFMix predictions (EUR-only region) and 8 variants assigned by the HMM to AFR ancestry with high confidence, i.e. probability > 0.9 (yellow markers).

As the runtime increases quadratically with increasing number of reference haplotypes in the panel, we opted for an approximative approach of subsampling the panel to ensure performance. To this end, we employed a bootstrapping approach, random sampling a fixed number of reference haplotypes from each superpopulation for multiple iterations and averaging over the results.

2.4.1 Results on HG00733

Results of the estimation are given in Figure 2.3. We applied both the HMM and RFMix to chromosome 1 of the phased assembly of HG00733, an individual from the Puerto Rican (PUR) population, as provided by the HGSVC. The variant call set of chromosome 1 contained 86,422 variants that were heterozygous, phased, and concordant with variant positions in the panel. For the bootstrapping run, we chose 10 iterations of sampling 79 individuals (30 from AFR and EUR each, and all 19 from AMR). To compare this to results given a larger panel, we added a second run on a panel containing 419 reference samples

(200 AFR, 200 EUR, and all 19 AMR samples).

For each variant, the computed likelihoods were averaged over all samples from the same population and the most likely population was output based on majority vote. Of the 86,422 variants, 2,577 had an assignment to one population with probability >= 0.9 (3,437 when using the panel with 419 samples). We denote such variants as *high-confidence* calls. Of those, 995 (1,251 for the larger panel) variants were assigned to one population with >= 0.99probability.

We ran RFMix [130] (v2.03-r0) using the parameters `-G 15 -e 5 -w 0.4 -n 5 -c 0.2 -s 0.2 -rf-minimum-snps=100 -reanalyze-reference` on the same panel of 419 samples. Note that we chose a node size of 5 to reduce bias resulting from unbalanced reference panel sizes, as done in previous studies [26, 132] and as suggested in the RFMix documentation [1].

Of the 1,251 estimates with highest confidence, 1,224 were assigned the same ancestry as with RFMix, leading to 97.84% concordant calls.

Results in Figure 2.3 show the high-confidence calls (> 0.9) for the HMM-based runs. For RFMix, which offers a variant-based and an interval-based output, the single variants were plotted. The results are largely coherent between the HMM and RFMix, although the HMM provides a more fine-grained resolution, especially regarding shorter intervals of predicted African ancestry, where the RFMix results seem smoothened over smaller regions. Figure 2.3 b shows an example region of a mismatch between the estimates of both tools. There are 8 AFR variant estimates that are not tagged as AFR by RFMix, but which are assigned to African ancestry with high-confidence by the HMM (probability > 0.9 on all variants; 0.9997, 0.9998, 0.9997, 0.9882, 0.9886, 0.9885, 0.9877, 0.9716, in that order). Looking at the allele frequencies in these populations, we find that the corresponding alleles for all these positions are very rare in the EUR population (containing 762 haplotypes from 381 EUR individuals), with AF=0.13% for 3 of them and AF=0.26% for the remaining 5 positions. In contrast, these alleles appear much more frequent in the AFR population (containing 942 haplotypes from 471 AFR individuals). There, allele frequencies range from 14.3% to 23.3%. Thus, the results from the HMM that infer AFR ancestry are much more consistent with the allele frequencies.

The results also show only minor differences in the high-confidence estimates of the larger reference panel and the bootstrapping results. In comparison, the computation using the large panel took 3:45 h to complete, while each iteration of a small panel takes 00:20 h. RFMix on the large panel ran for 1:08 hours.

[1] `https://gensoft.pasteur.fr/docs/RFMix/1.5.4/Manual.pdf`

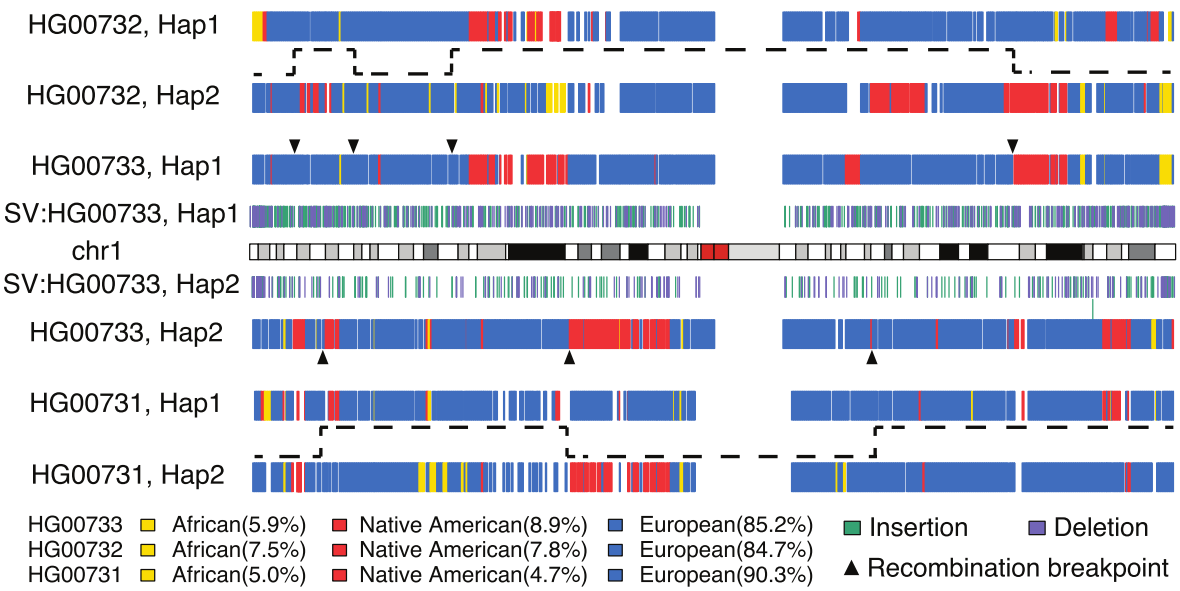


Figure 2.4: Ancestry-coloured chromosome 1 of a Puerto Rican trio. Inferred local ancestries for maternal (top, Hap1) and paternal (bottom, Hap2) haplotypes of HG00733 are compared with parental haplotypes (maternal, HG00732; paternal, HG00731). High-confidence calls of the HMM coinciding with RFMix estimates were used. Ancestral segments are coloured (African, yellow; Native American, red; and European, blue) and are consistent with the recent demographic history of the island. HG0733 SVs (≥ 50 bp; insertion, green; deletion, purple), inferred recombination breakpoints (triangles), and transmission of recombinant parental haplotypes (dashed lines) are shown. Figure taken from [56].

2.4.2 Results on PUR trio

Next, we performed the ancestry estimation for a Puerto Rican trio (samples HG00731, HG00732, HG00733). We assigned paternal and maternal haplotype ancestries accurately and distinguished recombination crossover events in the child haplotypes. Results are presented in Figure 2.4.

To avoid inaccurate ancestry calls, we removed SNVs in known gaps, segmental duplications, and heterochromatin, telomeric, and centromeric sequences (GRCh38) from this analysis. RFMix was again called with the following flags: `-G 15 -e 5 -w 0.4 -n 5 -c 0.2 -s 0.2 -rf-minimum-snps=100 -reanalyze-reference`.

We restricted the downstream analysis to a set of high-confidence ancestry markers where the HMM gave evidence for a single population with a probability >90%. To construct the final ancestry call sets for each individual haplotype, we first computed the concordant calls between the RFMix and HMM results. The per-variant ancestry resolution of the HMM approach allowed for re-calling the ancestry of a discordant region if sufficient evidence was given (>10 consecutive high-confidence markers).

Figure 2.4 shows the resulting ancestry inference of the Puerto Rican trio on chromosome 1, with HG00733 being the child and HG00732/HG00731 being the parents. We added Strand-seq inferred meiotic recombination breakpoints [31] for the two haplotypes of HG00733 (Figure 2.4, black triangles). By comparing the pattern of inferred ancestry with those breakpoints, we found that as expected, ancestry blocks on the paternal and maternal haplotypes of the child switch at the locations of the cross-over events in the parental chromosomes. The population-specific proportions for the three samples are consistent with previously reported ancestry proportions in PUR individuals [132]. The ancestry inference for the 32 HGSVC samples (omitting the children of the three trios) is shown in Supplementary Figure A.1.

Additional analysis of population stratification in the detected structural variants revealed 117 stratified SVs, which means they exhibit different allele frequencies in the different populations [66]. One example was a 2.8 kb insertion in the CLEC16A intron, a gene whose disruption is associated with type 1 diabetes [198]. This SV was rare in all populations (allele frequency <= 0.04) but the Admixed Americans (AF = 0.28), being the highest in Peruvians (AF = 0.39). Functional effect studies are missing, but notably, the rate of type 1 diabetes in Peruvians has been rated among the highest in the world [186]. This and other examples (details are skipped here and can be found in the manuscript [56]) are valuable starting points for further studies.

These results show that high quality phased assemblies allow for population-specific variant analysis and provide a valuable resource for follow-up studies in population genetics and evolutionary analysis.

2.5 Discussion and conclusions

We have developed a Hidden Markov Model that can be used for inferring ancestry estimates in genome assemblies. The method is based on the fact that common ancestry leads to shared haplotype blocks between individuals, where each genome is a mosaic structure of its ancestral genomes. This principle is exploited by using a reference panel of donors from diverse populations to compare the query assembly to. In the HMM, the references – including their population of origin – represent the hidden states which emit the alleles of the haplotypes. Computing a forward-backward algorithm on this model yields estimates for how common a certain haplotype context is in each population. We output position-wise vectors of probabilities for each population. This enables the comparison of estimations output by different methods and allows for potential corrections (Figure 2.3). When running both RFMix and the HMM on chromosome 1 of HG00733,

we found that the majority (97.84%) of the estimates output with highest confidence coincided with the RFMix estimates. As the computation time increases quadratically with the number of sequences in the panel, we implemented a bootstrapping approach for increased efficiency. It allows for subsampling smaller panels, reducing computation time (in our comparison, from 3:45 h to 00:20 h for each iteration, which can be run in parallel) while maintaining concordant results.

We applied this method to the HGSVC-led project where 35 human genomes were assembled in a haplotype-resolved manner, yielding high-quality assemblies from diverse populations that could be used to infer ancestries [56]. We computed ancestry tracts for all assembled haplotypes. The results on a Puerto Rican trio showed that the ancestry estimation was coherent with meiotic recombination breakpoints.

Despite working well for the underlying application purpose, the model is quite simple, as it is an implementation of the standard Li-Stephens model, which could be refined further. Both the transition and emission probabilities are currently oversimplified cost functions that aim to penalize switches between samples – not distinguishing between intra- and interpopulational switches – and deviations from the current allele. A different option would be to apply different transition probabilities for switching within and between populations. However, the HMM-based results showed high coherence to more complex methods such as RFMix.

More advanced methods like FLARE [27] model the emission probabilities by taking mutation rate, genotyping errors as well as gene conversion events into account, which may be a contributing factor to its higher accuracy compared to other methods, including RFMix. For distinguishing continental populations, this may not be as relevant, but it may become relevant in the distinction of subcontinental populations and, in general, populations which are less diverged, such as Japanese and Chinese, for instance [27]. On the other hand, simpler models like a basic HMM may be less accurate, but they work independently of having biological information about the underlying reference populations available, as they do not depend on biological parameters such as mutation rates, recombination rates, and genetic distance, and studies have shown [219] that these parameters can also influence the ancestry inference results.

One shortcoming of such a reference-panel based approach may be that regions of rare variants cannot be confidently assigned, as these methods require the alleles to be present in the panel. For the highest accuracy in calling ancestries, large panels of diverse populations are needed.

In the future, methods to infer local ancestry based on linear reference panels might become obsolete, as the field of pangenomics grows, and the first pangenome references

are published [119], aiming for a more diverse representation of the human genome as compared to linear references. Local ancestry inference methods could then be adapted to be used on pangenome graphs instead of reference panels, potentially in a similar manner as it is done in a recent LAI method [219], where a path is threaded through the population graph which fits the haplotype best.

Chapter 3

Haplotype threading for polyploid phasing

In contrast to diploid genomes, as within humans, genomes with more than two copies are substantially more challenging to assemble. Polyploid genomes–which plants often possess–can be highly diverse and structurally complex, often as a result of the hybridization processes leading to polyploidization (Section 1.7). Moreover, the simultaneous assembly of multiple haplotypes requires overcoming computational and algorithmic challenges. Therefore, methods for plant haplotype assembly have so far been missing, even though the availability of correctly phased haplotypes of plants is crucial for the design of advanced breeding strategies, for instance. This chapter presents WHATSHAP POLYPHASE, a polyploid phasing approach that consists of a read clustering and a subsequent cluster threading step. We elaborate on the problems of current polyploid phasing efforts and show that our method is able to overcome these challenges.

This chapter is based on a manuscript that was published in Genome Biology and on which I shared first authorship with Sven Schrinner and Jana Ebler [184]. My main contribution consisted in the design of the underlying haplotype threading algorithm and the implementation of its initial version. See Section E.2 for author contributions and license information.

3.1 Background

This section and its subsections include material from the publication [184]. Especially, some passages have been included directly and without modification from [184]. Further original material from [184] has been adapted for this thesis.

Polyploid genomes have more than two homologous sets of chromosomes (see Section 1.7). Polyploidy is common to many plant species, including important food crops like potato (*Solanum tuberosum*) and bread wheat (*Triticum aestivum*), as well as plants of economical and agricultural importance (Section 1.7). Resolving these genomes at the haplotype level is crucial for important applications: Evolutionary events, such as whole genome duplications, can be traced back and reveal the ancestry of polyploid organisms [225], contributing to uncover the evolutionary history of polyploid species. Moreover, knowledge of haplotypes is key for advanced breeding strategies and genome engineering, especially for enhancing yield quality in important crop species and adaptability to diverse environments [225, 214, 116]. These aspects are becoming increasingly urgent in the face of a growing world population and the challenges posed by climate change [229].

This chapter focuses on a method for phasing from long read information. For reference-based phasing in general, it is crucial that read alignments to the reference sequence span multiple variants, a requirement that short reads often fail to meet (see Section 1.4). This becomes particularly relevant in the polyploid context, where the phasing problem is complicated further by having to distinguish between not only two, but k haplotypes, with k being the ploidy. In the diploid case, the two haplotypes are complementary to each other, so that assembling one directly determines the second haplotype. In addition, long reads are better suited for haplotype assembly of polyploid plants, as their genomes exhibit many repeat elements [189, 109, 98] and underwent frequent structural variation events [216, 180]. Especially repetitive regions are difficult to resolve using short reads [123, 203].

There is a discrepancy between diploid phasing, where read-based methods are a routine step [150, 104], and polyploid phasing, which presents considerable challenges [104]. The underlying model which is most common to read-based phasing in the diploid context is the Minimum Error Correction (MEC) model. Its adaptation to polyploid genomes, however, exhibits substantial flaws, and alternative approaches are needed.

Sections 3.1.1 and 3.1.2 introduce the MEC model and existing phasing approaches, including their limitations in the polyploid setting. The methodology of WHATSHAP POLYPHASE is described in Section 3.2. Section 3.3 provides a summary of the results from applying WHATSHAP POLYPHASE to simulated tetraploid datasets and the genic regions of a potato sample.

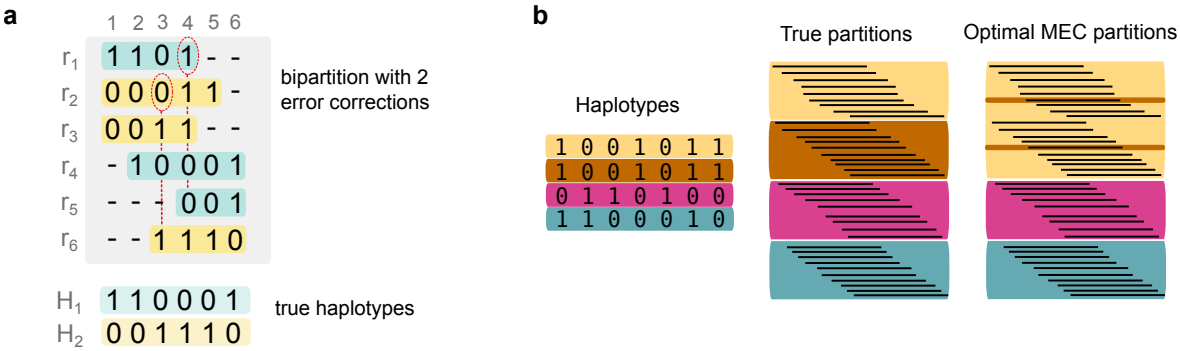


Figure 3.1: The MEC model. a. Diploid example. Six reads r_1 to r_6 are given over six variant positions. Uncovered alleles denoted with '-', alternative allele with 1, reference allele with 0. The reads can be sorted into two conflict-free bipartitions after flipping two bases (red circles in r_1 and r_2). The blue partition yields haplotype 110001, the yellow partition 001110. The MEC score corresponds to 2. **b.** MEC model in collapsed regions. Two haplotypes are identical (beige and brown). The MEC model is not able to distinguish between reads belonging to these haplotypes, so that one joint partition is created (right side) and the fourth partition is used to collect noisy reads. Figure **b** has been taken and adapted in colouring from [184], licensed under the CC BY 4.0 licence.

3.1.1 The MEC model

The Minimum Error Correction (MEC) model [121] is the most common and successful formalization for diploid haplotype assembly from sequence reads. Its aim is to create two sets of reads, each of which contains those reads that likely belong to the same haplotype. This is achieved by finding a bipartition that groups the reads into two sets that are conflict-free [150]. To create conflict-free partitions, it is necessary to apply *corrections* to the reads. The MEC model aims to minimize the number of these sequence corrections [150]. More formally, given a number n of sequencing reads, and m variant positions. Consider the read set $R = \{r_1, r_2, \ldots, r_n\}$, where each read r_i is a string of length m over the alphabet $\{0, 1, -\}$. We assume biallelic variants, so that $r_i[j] = 0$ if read r_i covers the reference allele at position j, and $r_i[j] = 1$ if the alternative allele is covered. If r_i does not cover variant j, we set $r_i[j] = -$.

Let $H = \{h_1, h_2, \ldots, h_k\}$ be a set of k haplotypes we aim to reconstruct, where $k = 2$ for diploid genomes. Each haplotype h_j is a string of length m over the alphabet $\{0, 1\}$.

The goal of the MEC model is to find k partitions $R'_1, \ldots, R'_k$ of R such that each partition provides an accurate assembly of a haplotype from H. The number of sequence corrections applied to reads in R in order to create the conflict-free partitions is to be minimized [121, 150, 185].

An example is illustrated in Figure 3.1a [185, 133]. A read set $R = r_1, \ldots, r_6$ is given over

six variant positions. The reads can be split into two bipartitions $\{r_1, r_4, r_5\}$ and $\{r_2, r_3, r_6\}$ that contain two errors. After two bases are flipped ($0 \rightarrow 1$ in r_2 and $1 \rightarrow 0$ in r_1), the read sets are conflict-free and yield the two haplotypes $H_1 = 110001$ and $H_2 = 001110$. Since two sequence corrections were necessary, the MEC score of this example equals 2.

It is evident that by design, this model is not suited to distinguish two (or more) haplotypes which share the same sequence – an optimal solution to the MEC problem would inevitably put the reads belonging to these haplotypes into the same partition. In polyploid genomes, it is not uncommon that multiple haplotypes are identical over some regions. Therefore, applying the MEC model to polyploid genomes is prone to errors based on incorrect read partitioning in regions of coinciding haplotypes. For an example in the tetraploid case, see Figure 3.1b. The two upper haplotypes (left panel) are identical. Instead of creating two separate read clusters for each of the haplotypes (middle panel), leading to a balanced set of partitions, the best solution to MEC would consist of one joint partition for both haplotypes, and optionally a small partition containing noisy reads (right panel). Consequently, this would lead to wrong haplotype assignments. MEC-based approaches are thus not immediately applicable to settings of polyploid phasing and, since MEC is an NP-hard problem [41], techniques required for optimal solutions such as dynamic programming techniques [150] quickly become infeasible for higher ploidies [24].

3.1.2 Related work

Over the past decade, several methods for polyploid phasing based on long reads have been developed. HapCompass [5, 4], introduced in 2013, is a theoretical framework based on spanning trees and the Minimum Weighted Edge Removal (MWER) criterion. A year later, HapTree [20], a maximum likelihood approach, was shown to outperform HapCompass in terms of accuracy and runtime [20, 142]. A simulation study from 2017 [142] evaluated HapCompass, HapTree, and a third tool named SDhaP [47], a semi-definite programming method based on an approximate MEC criterion. The study, using simulated data from a tetraploid potato genome, revealed that none of these methods were applicable for practical use due to both computational inefficiency that prohibits scaling to large genomic regions and qualitative insufficiencies like frequent failures and low accuracy. The authors concluded that there is 'clearly room for improvement in polyploid haplotyping algorithms' [142].

Since then, HPoP [224] was shown to outperform these previous approaches both in accuracy and runtime and became the state-of-the-art for a period of time. The under-

lying Polyploid Balanced Optimal Partition (PBOP) model aims to create k partitions of sequence reads to maximize similarities within partitions and differences between them. This can be seen as a polyploid generalization of MEC. When genotype information is present, these constraints are added to the model; this extension of HPoP is then referred to as H-PoPG.

More recent advancements, such as PolyHarsh [83] (Gibbs sampling method based on MEC), TriPoly [141] (trio-based inference), and SDA [32] (resolution of segmental duplications of higher ploidy), have not proven to be useful for whole-genome, single-individual haplotyping.

Overall, these methods lack practical usability, scalability, or accuracy for real-world polyploid phasing applications. Notably, no current method provides an accurate, efficient model designed specifically for polyploid phasing that can produce reliable blocks based on phasing certainty while being computationally feasible.

To bridge this gap, we developed WHATSHAP POLYPHASE, a method that is designed to properly address the challenges of polyploid phasing by offering an accurate model that departs from MEC. WHATSHAP POLYPHASE, like the core algorithm of WhatsHap [150, 133], is a read-based approach based on long sequencing reads (Section 1.4). In particular, WHATSHAP POLYPHASE is able to detect and phase regions where multiple haplotypes coincide by taking coverage into account via a newly established threading step. Additionally, our method is able to integrate information from input genotypes for accurate phasing results.

The following sections provide a description of the phasing algorithm with a focus on the 'haplotype threading' step including methodological details, and an evaluation of the approach on simulated and real datasets.

3.2 Phasing Model and Algorithm

This section and its subsections reuse material from [184]. Here, we describe the phasing algorithm, where we first provide a summary of the cluster editing part, which was mainly developed by Sven Schrinner, a co-author of this publication (shared first authorship). My main contribution was the development of the haplotype threading part, which is the focus of this section.

An overview of our method is given in Figure 3.2. Since WHATSHAP POLYPHASE is a reference-based phasing method, one essential input is a reference sequence, complemented by a BAM file containing read alignments to the reference, and a VCF file contain-

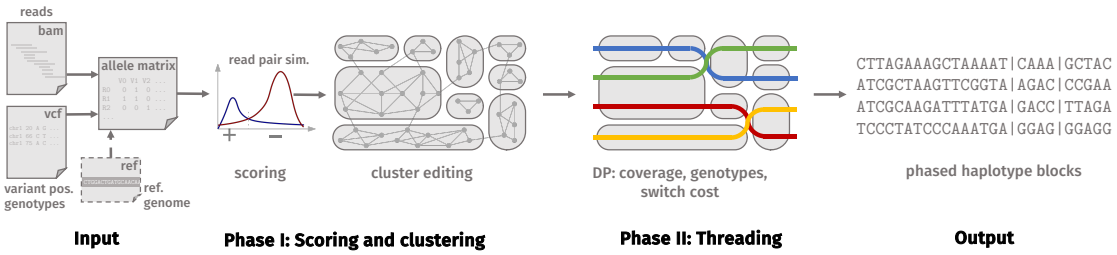


Figure 3.2: Overview of WHATSHAP POLYPHASE. The input allele matrix results from a given BAM and VCF file and an optional realignment step. Phase I: statistical scoring of each read pair classifies them into belonging to the same or to different haplotypes. The scores are used as weights for a graph over all reads, which is clustered by cluster editing (grey round shapes). Phase II threads k haplotypes (coloured lines) through the clusters (here $k = 4$) balancing coverage violations and switch costs while respecting the genotype information. This results in k phased haplotypes, subdivided into blocks (vertical lines). Figure taken from [184], licensed under the CC BY 4.0 licence.

ing variant calls of the query sample. The underlying approach of WHATSHAP POLYPHASE is made up of two stages. The first phase of the algorithm (Figure 3.2, second panel) aims to cluster reads that are likely to originate from the same haplotype. In short, this is done by computing a statistical similarity score for each pair of reads and constructing a graph where each read is a node and edges between the nodes are weighted by the similarity scores. The underlying clustering model, cluster editing [230], takes this weighted graph as input and finds the most cost-efficient way to transform it into a graph only consisting of disjoint cliques. A clique is then expected to contain reads from the same haplotype. The size of the graph makes it infeasible to compute an optimal solution to the problem, so we rely on an iterative heuristic to produce accurate clusters [23]. We deliberately make no assumptions on the ploidy at the clustering stage. In particular, reads of multiple haplotypes that are locally identical end up in the same cluster.

The second phase (Figure 3.2, third panel) consists of the actual haplotype assembly by threading k haplotypes through the set of clusters obtained in the first phase. The position-wise read coverage of a cluster influences whether it represents one or multiple haplotypes. In particular, this allows for multiple haplotypes covering the same cluster, which would be the expected behaviour in regions where the sequence is highly similar or identical across multiple haplotypes. During the threading step, we further ensure that cluster switches along a haplotype are minimized. If the genotype of the query sample is given, clusters are chosen for which the consensus genotype at a variant position fits the input genotype. When the phasing cannot be extended confidently across a variant pair, we create a new haplotype block (Figure 3.2, last panel) to minimize the number of

phasing errors within phased blocks.

3.2.1 Notation and objective

Here, we define the problem more formally and introduce the methodology behind our method. Let k be the ploidy of the query genome, containing n heterozygous variants. Let m be the number of sequence reads. We assume that all variants are biallelic and denote the reference allele with 0, the alternative with 1. Then, there are k true haplotypes of the sample, represented as sequences $H_0, \ldots, H_{k-1}$ of length n, where each $H_i[j]$ is the allele of the j-th variant, such that $H_i[j] \in \{0,1\}$ for all $j \in [0, \ldots, n-1]$. The set of reads R is represented by m sequences $r_0, \ldots, r_{m-1}$. Given that reads possess different lengths, we define each r_i over n variants and denote the alleles at position j with $r_i[j] \in \{0,1,-\}$, where '$-$' indicates that the read does not cover this position. Then, we aim to find k sequences $H_0^{'}, \ldots, H_{k-1}^{'}$ of length n which are close or identical to the true haplotypes $H_0, \ldots, H_{k-1}$.

3.2.2 Cluster editing

The following is a summary of the cluster editing part of the algorithm, which was developed by Sven Schrinner. It uses material modified from [184].

The first phase of the algorithm clusters reads likely originating from the same haplotype based on pairwise similarities, calculated using *Hamming rates*. The Hamming rate between two reads r and s is defined as the proportion of differing alleles (disagreements) to the total number of shared variants between the two reads (see Section 1.9). The expected Hamming rate between reads from the same haplotype, d_{same}, is primarily influenced by the sequencing error rate, as differences between reads from the same haplotype are likely due to sequencing errors alone. In contrast, the expected rate for reads from different haplotypes, d_{diff}, also reflects the inherent differences between different haplotypes. Two simplifying assumptions are used: all haplotypes are (i) equally frequent among the reads and (ii) equally distinct from each other. This leads to the follow-up assumption that on average, $\frac{1}{k}$ of all read pairs (where k is the ploidy) are from the same haplotype, allowing for the estimation of d_{same} as the lowest $\frac{1}{k}$ fraction of all Hamming rates.

Similarity scores for read pairs are then computed using the binomial probability density function, comparing observed disagreements with expected Hamming rates, with d_{same} and d_{diff} as success probabilities. Details are skipped here and can be found in [184]. A positive score suggests the two reads are from the same haplotype, while a negative score indicates different haplotypes. These scores guide the subsequent clustering

process.

Clustering is performed using the cluster editing model [230], which transforms the input graph (reads as nodes, similarities as edge weights) into disjoint cliques. These cliques indicate clusters of reads likely originating from the same haplotype. Practical issues, such as varying variant distances and regions of local haplotype similarities, can result in more clusters than the expected k haplotypes. The reduction of these clusters to k haplotypes is addressed in the subsequent step of the algorithm.

3.2.3 Haplotype threading

For the second phase of our algorithm, we developed a novel approach called *haplotype threading* which performs the actual phasing of a k-ploid genome to k haplotypes.

The input consists of a set C of clusters which is output by the cluster editing (Section 3.2.2), and a set of genotypes G. The output of this step is a sequence of k-tuples of clusters, representing the k haplotypes.

The previously computed read clusters C exhibit two properties: First, the number of clusters at a position $i \in \{0,\ldots,n-1\}$ can be larger than k, so that some clusters do not contribute to any computed haplotype. Second, the reads in a cluster $c \in C$ usually do not cover the whole chromosome, but only a part of the n variants, so that in order to obtain whole-chromosome haplotypes, these must be assembled from multiple clusters. This is done by *threading* a haplotype through the clusters, meaning that for every haplotype, a path through C is assembled by choosing one cluster $c \in C$ for each haplotype at every variant position i.

In a genome of ploidy k, we seek for k haplotypes and thus assemble all k sequences simultaneously by choosing k-tuples of clusters at each position. Duplicate clusters within tuples are allowed since reads from one cluster can belong to multiple true haplotypes: For regions with high local similarity between the true haplotypes, the corresponding reads are likely placed into one cluster by the cluster editing step.

In the threading process, we aim at achieving three objectives: (i) genotype concordance, (ii) read coverage concordance and (iii) haplotype contiguity.

Genotype concordance captures the agreement between the known target genotype and the chosen clusters. For the true haplotypes $H_0,\ldots,H_{k-1}$ of length n, the corresponding genotype can be described as the component-wise sum $G = H_0 + H_1 + \cdots + H_{k-1}$ and is denoted by $G = g_0, g_1, \ldots, g_{n-1}$, where $g_i \in 0,\ldots,k$. Furthermore, for each cluster c and each position i, we can compute the consensus $cons(c,i) \in \{0,1,-\}$ as the most frequent

allele among all reads in c at position i. Using this definition, we can compute a **consensus genotype** of a k-tuple $(c_0,\ldots,c_{k-1})$ at position i as $\sum_{j=0}^{k-1} cons(c_j,i)$.

For each position i, we then only take those cluster tuples into account whose consensus genotype at i is *concordant* with the target genotype, i.e., $\sum_{j=0}^{k-1} cons(c_j,i) = g_i$.

This reduces the search space of possible tuples and filters out non-promising combinations beforehand, which in turn increases efficiency and accuracy. In case there is no tuple with a concordant genotype at position i, we allow genotype deviations of 1; if this also fails, all possible tuples are considered.

Read coverage refers to the average number of reads within a cluster and affects the number of haplotypes that the cluster can represent.

Since in locally identical regions, multiple haplotypes can be threaded through the same cluster – which leads to multiple appearances of this cluster in the k-tuple – this number of haplotypes has to correspond to the coverage of the chosen cluster. The **relative coverage** of a cluster c at position i, $cov(c,i)$, describes the proportion of reads in c covering i to the number of reads in all clusters that cover i.

The **expected copy number** of a cluster c at i is the expected number of haplotypes that are threaded through c. We can compute it as $cn_{\text{exp}}(c,i) = \left\lceil k \times cov(c,i) - \frac{1}{2k} \right\rceil$.

The **true copy number** of c corresponds to the number of appearances of c in a chosen cluster tuple $(c_0,\ldots,c_{k-1})$. It is given by $cn_{\text{true}}((c_0,\ldots,c_{k-1}),c,i) = |\{i \mid i \in \{0,\ldots,k-1\}, c = c_i\}|$. Deviations of the true number of occurrences from the expected ones are penalized by a constant factor p_{cov} per cluster, so that a cluster tuple $(c_0,\ldots,c_{k-1})$ is evaluated by the cost function

$$\text{costs}_{\text{cov}}((c_0,\ldots,c_{k-1}),i) = \sum_{j=0}^{k-1} p_{\text{cov}}[[cn_{\text{exp}}(c_j,i) \neq cn_{\text{true}}((c_0,\ldots,c_{k-1}),c_j,i)]]$$

where $[[x \neq y]]$ returns 1 if $x \neq y$ and 0 otherwise.

Haplotype contiguity expects haplotypes to stay in the same cluster for as long as possible, so that switching of haplotypes between clusters is penalized. For two consecutive cluster tuples $(c_0,\ldots,c_{k-1})$ and $(c'_0,\ldots,c'_{k-1})$ at positions i and $i+1$, we denote the cost factor by p_{switch}, which results in the cost function

$$\text{costs}_{\text{switch}}((c_0,\ldots,c_{k-1}),(c'_0,\ldots,c'_{k-1})) = \sum_{i=0}^{k-1} p_{\text{switch}}[[c_i \neq c'_i]]$$

We developed a dynamic programming approach to rapidly find the optimal sequence

of tuples that minimizes all costs. We compute a two-dimensional matrix S with a column for every variant j from 0 to $n-1$ and a row for every possible genotype-conform tuple of clusters. Since the number of eligible cluster tuples can differ between variant positions, the columns of S do not necessarily have the same lengths. We denote the length of a column j with l_j. Using the cost functions defined above, $S[i,j]$ is then computed as

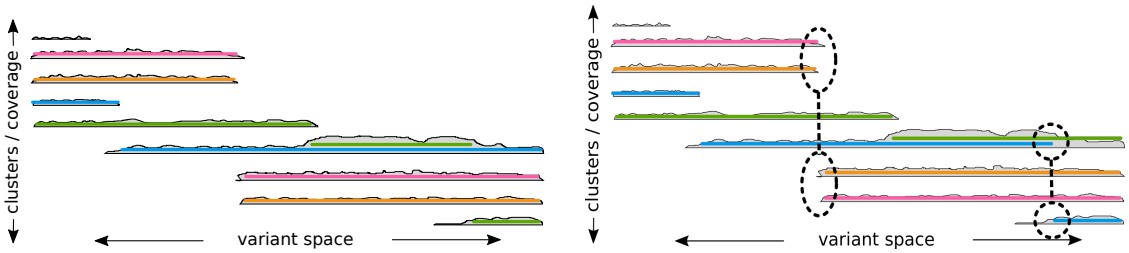
$$
\begin{aligned}
S[i,0] &= \text{costs}_{\text{cov}}(c_i, 0) \ , \\
S[i,j] &= \text{costs}_{\text{cov}}(c_i, j) + \\
&\qquad \min_{k \in \{0,\ldots,l_{j-1}-1\}} \left(S[k, j-1] + \text{costs}_{\text{switch}}(c_k, c_i) \right) \qquad \text{for } j > 0 \ ,
\end{aligned}
$$

where c_i denotes the cluster tuple in row i. The optimal threading score is then given by the minimum value in the last column. Starting at this position, we assemble the sequence of clusters with minimum costs via backtracing.

The threading process is illustrated in Figure 3.3a for $k = 4$. The clusters from the first step are drawn as grey shapes in a two-dimensional space, where the horizontal position refers to the variants covered by the reads inside a cluster and the height represents the relative coverage of a cluster at every position. The position on the y-axis has no numerical meaning and is just used for illustration purpose. Starting from the left, a 4-tuple of the five present clusters needs to be chosen. According to the coverage, the best choice is to thread one haplotype through each of the four clusters with highest coverage and to ignore the smallest one, as this is likely to contain noisy reads only. From thereon, the threads change clusters whenever a cluster ends or undergoes a drastic change in relative coverage. Note that this model allows for multiple sequences of clusters to have the same score. Therefore, when there is not enough evidence for combining clusters uniquely, we need to introduce a cut to the phasing output when unsure, see Figure 3.3b.

3.3 Results

This section contains joint work by all three co-authors. My contribution was to analyse and evaluate phasing results in collapsing regions and compare the results to non-collapsing regions. The Snakemake pipeline for the construction of the test data sets and the computation of the phasing statistics has been contributed by Jana Ebler. The pipeline has been extended to higher ploidies by Sven Schrinner. This section reuses material from [184] .



(a) Example of tetraploid threading **(b)** Critical switching positions

Figure 3.3: Visualisation of the threading. a. Clusters of reads are represented as grey shapes with their horizontal span indicating the covered variants and the height being the respective coverage. The $k = 4$ threads are shown as coloured lines passing through the clusters. Multiple threads can co-enter the same cluster if the coverage is suited. **b.** Alternative threading with the same score in our model. Two positions cause ambiguity and allow switches in the threading compared to (a). These are candidate cut positions to prevent switch errors in the final phasing. Figure taken from [184], licensed under the CC BY 4.0 licence.

We used common evaluation statistics that capture different properties of haplotype sequences to compare the solutions computed by both tools to ground truth haplotypes available for our data sets. For the definition of evaluation statistics, see Section 1.9.

3.3.1 Simulated polyploid data

We generated polyploid versions of human chromosome 1, specifically tetraploid, pentaploid, and hexaploid, by combining sequencing data from three individuals (NA19240, HG00514, HG00733), creating data sets at different coverages (40× and 80×). For all these samples, ground truth haplotypes were available [31]. Additionally, we created equivalent data sets using reads simulated by PBSIM [148]. Using these data sets, we then evaluated the phasing accuracy and computational performance of WHATSHAP POLYPHASE and compared it to the most viable competitor, H-POPG.

WHATSHAP POLYPHASE consistently achieved fewer switch errors compared to H-POPG, with at least 3 times lower SER on the varying data sets. For example, on the real tetraploid dataset at 40× coverage, WHATSHAP POLYPHASE produced a SER of 0.58% versus 2.01% for H-POPG. The full results are shown in Supplementary Table B.1.

The way that phasing methods introduce cuts to end the current haplotype block has a large effect on the phasing results. H-POPG cuts a haplotype only if two neighbouring variants are not connected by any read. This results in rather long blocks which also cover those regions that are phased with lower confidence and therefore prone to switch errors. These, however, impact the global phasing accuracy, and especially the Hamming rate.

WHATSHAP POLYPHASE by default defines blocks in a more sensitive way: new blocks are introduced whenever two variants cannot be connected with enough confidence, for instance, when one haplotype switches to another cluster during the threading step. However, we offer different settings for the block cut strategy, where varyingly sensitive criteria for block cuts can regulate the length and accuracy of the blocks.

With the most conservative method (default), blocks are shorter, represented by the lower N50 values, but more accurate, explaining the huge differences in Hamming rates (for instance, 1.48% for WHATSHAP POLYPHASE compared to 27.53% for H-POPG on the 40× tetraploid real data set). Across all configurations, the SER of WHATSHAP POLYPHASE falls below that of H-POPG (Supplementary Figure B.1).

For the most equivalent comparison, we applied the configuration closest to H-POPG, resulting in similar block lengths. After this adjustment, the N50 values and Hamming rates of both tools almost equalized, while WHATSHAP POLYPHASE still achieved 30%-40% lower SER. These results demonstrate that across all tested data sets and configurations, WHATSHAP POLYPHASE consistently produced more accurate phasing results, while also providing the flexibility to prioritize either block length or phasing confidence.

Locally identical haplotypes We have already stated (Section 3.1.1) that MEC-based approaches for polyploid phasing might experience struggles especially in regions where multiple haplotypes share the same sequence. Different configurations of locally similar haplotypes cannot properly be distinguished based on MEC scores. We evaluated the performance of WHATSHAP POLYPHASE in these specific regions and compared it to that of H-POPG.

First, we define a *collapsing region* as a region in the genome where two or more haplotypes share the same sequence over the length of at least 50 consecutive variants. Using the same data set previously described, we ran WHATSHAP POLYPHASE and H-POPG on the artificial tetraploid chromosome 1 of the simulated and real data sets with 40× and 80× coverage, respectively. Switch error rates were analysed separately in the collapsing regions and in the remaining (non-collapsing) regions, and in the full chromosome. Results are shown in Table 3.1.

In collapsing regions of the 40× real data set, H-POPG reached an SER of 2.02%, which corresponds to a 3.06-fold increase compared to the 0.66% SER of WHATSHAP POLYPHASE. In non-collapsing regions, this factor was only 1.19. A similar trend was visible for the 80× data set (factors 2.67 and 1.12, respectively), as well as for the equivalent simulated data. As with the previous experiments, the default block cut setting of WHATSHAP POLYPHASE further reduced the number of switch errors, and the difference between H-POPG and

Table 3.1: Comparison between the resulting switch error rates of WHATSHAP POLYPHASE (WH-PP) and H-POPG 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 chromosome 1 of the real (a) and simulated (b) tetraploid dataset on both 40× and 80× coverage. WH-PP denotes the default block cut strategy. For better comparability with H-POPG, a second setting with fewer block cuts is used (WH-PP*).

coverage	method	collapsing regions	non-collapsing regions	total
40×	WH-PP	0.29	0.69	0.60
	WH-PP*	0.66	1.81	1.65
	HPoP-G	2.02	2.16	2.02
80×	WH-PP	0.14	0.46	0.35
	WH-PP*	0.38	1.16	0.99
	HPoP-G	1.05	1.30	1.24

(a) real read data (tetraploid)

coverage	method	collapsing regions	non-collapsing regions	total
40×	WH-PP	0.18	0.45	0.43
	WH-PP*	0.45	1.29	1.19
	HPoP-G	2.01	1.63	1.68
80×	WH-PP	0.08	0.37	0.32
	WH-PP*	0.25	0.88	0.82
	HPoP-G	0.94	0.98	0.99

(b) simulated read data (tetraploid)

WHATSHAP POLYPHASE were substantially increased in collapsing regions as well. This shows that regions with locally identical haplotypes disproportionately affect the phasing qualities of H-POPG as a MEC-based model, while we were able to accurately phase haplotypes within these regions.

3.3.2 Potato data

The analysis presented here was mainly performed by Jana Ebler. This subsection provides a summary of her work and reuses material from [184].

We applied our method to real sequencing data from tetraploid potato (*Solanum tuberosum*), for which we generated both short Illumina reads and long Oxford Nanopore (ONT) reads. Read alignment to the potato reference genome published by the Potato Genome Sequencing Consortium (PGSC) [81] revealed unbalanced coverage, especially for the short reads (Supplementary Figure B.2a), which hinted at structural variations and re-

arrangements in the data. Consequently, we relied on the longer ONT reads for identifying SNPs, as these were better suited for variant calling in comparison to the short reads, particularly in the presence of structural variation. As the ONT reads expressed higher sequencing error rates, we first had to run an error correction pipeline to increase the quality of the reads [168]. Details are skipped here and can be found in [184]. After variant calling with the corrected reads and subsequent phasing, our method successfully phased 91% of the 36,274 genes that contained heterozygous variants, producing an average of 2.13 phased blocks per gene. A substantial portion of these genes, including many long ones, were fully phased (Supplementary Figure B.2b). For the FRIGIDA-like protein 5 isoform X2 gene (accession: XP_015169713), for instance, WHATSHAP POLYPHASE enabled haplotype-resolved assembly. Based on the phasing results, we separated the reads by haplotype and performed local assemblies for each, revealing an early STOP codon and a frameshift mutation in one haplotype (Supplementary Figure B.2c). These haplotype-specific differences were only visible through phasing, underscoring the importance of phasing in understanding gene architecture. Subsequent multiple sequence alignments of the phased haplotypes with the reference showed overall good alignment, with small differences that may be relevant as potential targets for functional follow-up studies (Supplementary Figure B.2d).

3.4 Discussion and conclusions

A significant challenge for reference-based phasing methods in general is their dependence on the quality of read alignments and subsequent variant calls. When there are substantial deviations from the reference genome, methods reliant on reference genomes are prone to errors.

First, this depends on the quality of the reads: alignments from reads that contain frequent sequencing errors will be of poor quality and cannot be trusted. However, accurate but short reads are also not useful. They do not bridge enough variant pairs to be useful for phasing, and struggle in regions with repeats and structural variation, as they often occur in polyploid genomes. For our phasing of the potato genome, we relied on long Oxford Nanopore reads, which were prone to sequencing errors and therefore aligned poorly to the reference. Without correcting the errors in a first step, which made the use of additional Illumina reads necessary, they would not have been usable. Since the time of publication of this project, there have been advancements in sequencing technologies, leading to increased read length and reduced sequencing errors. Especially the development of highly accurate long reads, such as PacBio Circular Consensus Sequencing (CCS),

or 'HiFi', reads [220], now alleviates this issue and enables the use of at the same time long and accurate sequencing reads.

Second, the quality of the reference genome is equally crucial for the success of reference-based phasing. At the time of publication, the lack of a high quality potato reference presented a severe limitation. More recent, accurate versions of the potato reference genome have since been published [154, 212], resolving previous gaps and improving completeness and accuracy.

Still, it is important to note that for a genome as variant-rich and heterogeneous as the potato and other polyploid plants, reads will still frequently deviate from the reference in multiple regions, even with a better reference and good quality reads. As our benchmark experiment on artificial polyploid data was created from human samples, it exhibited human-like heterogeneity, and there, our method was shown to work well, producing more accurate results than the competitor. In our potato experiment, however, we found that phasing the whole genome was not possible given the complexity of the genome and the insufficient alignments and variant calling, and we had to restrict to the gene regions. Nonetheless, we were able to phase the genes, providing a proof-of-concept of the method and enabling possible follow-up studies.

In consequence, one important finding of this study is the limited applicability of read-based methods for polyploid assembly in general. Recent advances have therefore shifted towards de novo assemblies [17, 200, 189], employing diverse sequencing technologies, including high-quality long reads (PacBio HiFi) and complementary data types such as single-cell sequencing, Hi-C data or pedigree information. Notably, the introduction of WhatsHap Polyphase Genetic [183], a genetic phasing method based on low-depth sequencing data from an offspring panel, has enabled the generation of longer haplotype blocks than read-based phasing alone, albeit producing more sparse haplotype blocks. These blocks can potentially be integrated with read-based phasing to enhance overall results [183].

However, WHATSHAP POLYPHASE is still valuable, as demonstrated by its recent application in evaluating one of such de novo assemblies, where it was run with the reference DMv6.1 [154] and ONT ultra-long reads. There, Bao et al. [17] used WhatsHap Polyphase to assess phasing accuracy in their de novo assembly of the C88 potato cultivar. A comparison of this local, read-based phasing over more than 3.4 million phased SNPs to the de novo assembled haplotypes showed high consistency across all haplotypes. In this case, two improvements since the time of the original publication were utilised: A more exact reference sequence and higher quality reads, which, after a lot of technological development, are less error-prone and substantially longer – two aspects that WHATSHAP

POLYPHASE benefits from, so that the method may be even better applicable nowadays than formerly.

To summarise, this application shows that WHATSHAP POLYPHASE can serve as a valuable addition for phasing of polyploid assemblies. However, for the creation of accurate haplotype-resolved assemblies of polyploids, many challenges are inherent to the general approach, so that it would be more advisable to not rely on reference-alignment-based methods, but to perform a de novo assembly with direct resolution of haplotypes instead. In the next chapter, we will introduce a method for the de novo assembly of the four haplotypes of the potato genome.

Chapter 4

Haplotype-resolved assembly of a tetraploid potato genome

In this chapter, we introduce a novel method for haplotype-resolved assembly of polyploid genomes and present an assembly of the autotetraploid potato cultivar Altus. Polyploid haplotype assembly poses a major computational challenge. Our method uses low-depth sequencing data from an offspring population to achieve chromosomal clustering and haplotype phasing on an assembly graph. We show that we are able to generate high-quality assemblies of individual chromosomes with haplotype-specific sequence resolution of whole chromosome arms.

This chapter is based on a publication in Genome Biology [189], of which I am the main author. All sections in this chapter reuse material from this publication which I contributed. Supplementary Section C.1 contains material that was contributed by Freya Ziegler, a co-author of this publication. See Section E.3 for author contributions and license information.

4.1 Background

This section reuses material from [189].

Polyploidy (see Section 1.7) is common in plant genomes and two forms are recognized. Allopolyploids arise from interspecific or intergeneric hybridization events, and the difference between subgenomes is usually sufficient to assemble them like diploids. This has been demonstrated for rapeseed, wheat and strawberry, among others [109]. In contrast, autopolyploids arise from genome duplications, and the presence of multiple sets

of the same homologous chromosomes means that haplotype-resolved sequence assemblies are much more challenging. One example is potato (*Solanum tuberosum*), most cultivars of which are autotetraploid [153]. Potato is a vital food crop in many developing countries [54], and the global production volume exceeds 300 million tons per year [22]. Because of this agronomic value, efforts to assemble potato genomes are of crucial importance.

The haplotype-resolved assembly of diploid genomes has been progressively refined, and accurate results are now possible as we have shown previously [56, 158]. In contrast, computational methods for polyploid haplotype assembly rarely lead to satisfying results, particularly for autotetraploids. Reference-based approaches for haplotype phasing in polyploid species align reads to an existing reference sequence but are often inaccurate [142]. Reference-based approaches in general have severe limitations, especially in the presence of structural variation [158]. For potato haplotype phasing, two reference genomes are currently used: the synthetic double monoploid potato clone DM1–3 516 R44 [154] and Solyntus, which is based on a diploid potato cultivar [212]. Reference-based algorithms for polyploid haplotype phasing include HapTree [20] and H-PoP [224]. Other methods target selected genomic regions to resolve haplotypes locally, for example using integer linear programming [194]. We previously developed WhatsHap polyphase, which was an improvement over contemporary methods but still relied on a reference genome [184] (see Chapter 3).

The de novo assembly of polyploid genomes without a reference is still an emerging strategy. Recently proposed workflows involve building a 'squashed' assembly with no or limited haplotype resolution at first, and using this as the basis for haplotype phasing. As even long-read sequencing is generally insufficient for long-range phasing, auxiliary data types are required. One example is single-cell pollen sequencing [234], which was recently used for comprehensive haplotype reconstruction in autotetraploid potato [200]. Another example is the recent publication of a potato assembly where a selfing population of 1034 samples was used [17].

Here, we propose an alternative method in which PacBio HiFi reads of the potato cultivar 'Altus' are combined with cost-effective low-coverage short-read sequences from multiple offspring samples. Accordingly, we generated PacBio HiFi reads (96× coverage) and created an initial assembly using hifiasm [35]. We assembled the individual haplotypes from the resulting assembly graph using sequencing data from 193 offspring of two potato cultivars (Altus and Colomba) at low coverage (~1.5× per haplotype) combined with a novel approach based on k-mers to identify the four haplotypes.

Approaches based on unique k-mers have facilitated the haplotype-resolved assembly

of diploid parent-offspring trios [107] and challenging regions of human chromosome 8, such as its centromere [124]. In the latter example, the authors created a library of singly unique nucleotide k-mers (SUNKs) to barcode long reads and assemble them into scaffolds especially in complex regions. Here, we make use of unique k-mers to phase the assembly graph of a parent genome from a polyploid offspring panel.

Our assembly mapped well to the latest version of the monoploid DM1–3 516 R44 reference (DMv6.1) and yielded haplotype-resolved assemblies of individual chromosomes with phased haplotype block lengths of up to 34 Mb, phased contig N50 values of up to 12 Mb, and a genome-wide phased contig N50 value of 7.5 Mb. Our approach also allows the detection and correction of assembly errors in the assembly graph as well as in previously published references.

4.2 Algorithm

Here, we describe the algorithmic workflow in more detail. Results of its application on the potato cultivar Altus are presented in Section 4.3.

This section and all subsections reuse material from [189] and are an extension of this publication.

4.2.1 Overall assembly strategy

A high-level overview of our workflow is shown in Figure 4.1. Starting with PacBio HiFi reads derived from the Altus genome (Figure 4.1a), we build an assembly graph using hifiasm, resulting in a partially haplotype-resolved graph with bubble-like structures representing the different haplotypes (Figure 4.1b). For each node in the assembly graph (so-called *unitig*), we detect unique k-mers (fragments of length k). (Figure 4.1b). We then estimate the dosage of each unitig, defined as the number of haplotypes to which each unitig contributes (Figure 4.1c). In the next step, we count the formerly detected unique k-mers in the Illumina reads for each of the 193 offspring samples (Figure 4.1d). Each unitig is thus represented by a k-mer count pattern consisting of 193 values. Nodes with similar count patterns, implying the inheritance of a node by the same subset of offspring samples, are therefore likely to be part of the same haplotype. We then make use of the k-mer count patterns to perform an initial clustering of the nodes into chromosomes (Figure 4.1e). The clustering procedure is followed by a step to determine the four haplotypes among nodes with dosage 1 (Figure 4.1f), and another step to add nodes with higher

dosages (Figure 4.1g). Ultimately, this yields a set of four haplotype clusters for each chromosome (Figure 4.1h). We complete the assembly by finding graph traversals through the clustered assembly graph and thereby assembling haplotype blocks (haplotigs).

4.2.2 Assembly graph construction

First, the Altus genome is sequenced using PacBio HiFi technology to produce highly accurate long reads. Also, Illumina short-read sequencing data is acquired that represents 193 offspring of the cross between Altus and another cultivar (Colomba).

The initial assembly is performed using hifiasm (v0.13) [35] taking the HiFi reads as input. This process generates an assembly graph that contains all the assembled, unprocessed (raw) unitigs, partially resolving the four haplotypes. Variation is represented by bubble structures in the graph, where a unitig branches into two or more other unitigs that eventually merge into another common node.

4.2.3 Dosage estimation

The *dosage* of a node (unitig) from the assembly graph describes the number of haplotypes that it represents. In the assembly of a diploid genome, a node can only be present either in one single haplotype or, representing a homozygous region, in both haplotypes. For a tetraploid genome, respectively, the dosage can take any value from the set {1, 2, 3, 4}. Knowing how many haplotypes pass through a contig of the initial hifiasm assembly graph is the first step in the assembly pipeline. To estimate the dosage for each contig, we analyse its average coverage of aligned reads. This is done by aligning all Altus HiFi reads to the graph contigs using minimap2 [112]. We filter out all alignments with a mapping quality below 60 (`samtools --min-MQ 60` [115]). Using the remaining alignments, we compute the sequencing depth at each base position.

Given that hifiasm graphs usually contain overlaps, we compute the intervals of non-overlapping sequences per node, which is the region of each node that is not part of any overlap with its neighbouring nodes, or the *unique sequence* of the contig. In the alignment, we then only consider positions mapping to this unique sequence to compute the sequence depth.

Given a contig i with n considered positions in the non-overlapping sequence, each position then has a read depth $c_i[j], j \in \{1,\dots,n\}$. The average coverage c_i of contig i is then computed as the average read depth over all positions: $c_i = \left(\sum_{j=1}^{n} c_i[j]\right)/n$.

We also compute the total average coverage m as the average over all contigs. Then,

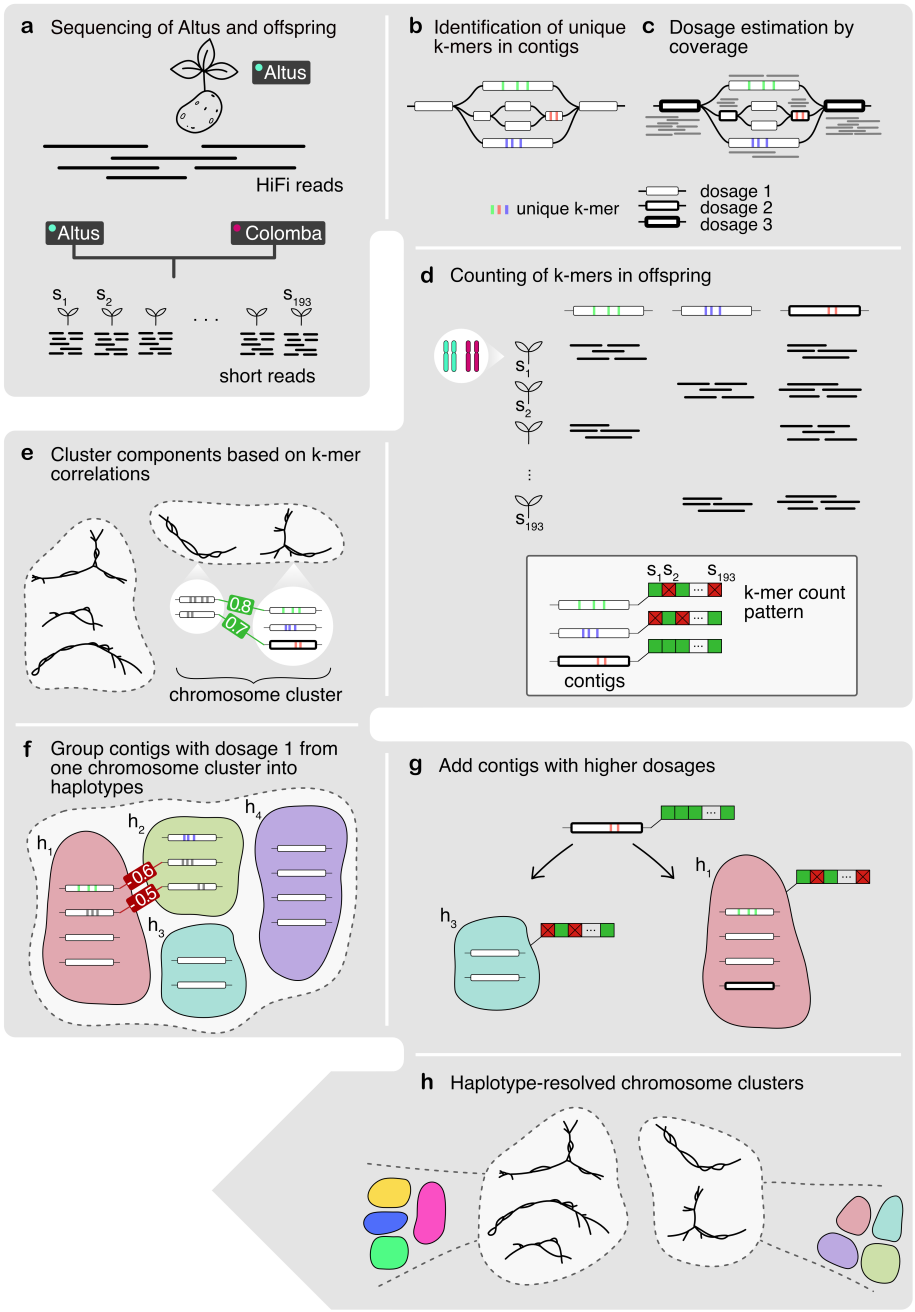


Figure 4.1: Overview of the workflow. **a.** The Altus genome is sequenced using PacBio HiFi technology, whereas the 193 genomes of the cross Altus × Colomba are sequenced on the Illumina platform. **b.** Hifiasm is used to assemble the Altus HiFi reads into an assembly graph. For each contig in the graph, unique k-mers are detected (denoted by the coloured bars). **c.** The HiFi reads are aligned to the contigs and the mapping depth is used to estimate dosages (1 to 4) for each contig. The different dosages are denoted by the thickness of the contig line (thicker outlines mean higher dosage). **d.** The unique k-mers are counted in the short reads of the offspring samples in order to compose a count pattern for each contig. **e.** For all nodes from the assembly graph components, the pairwise correlation of k-mer count patterns is computed and components are clustered to represent chromosomes. **f.** In each chromosome cluster, the nodes with estimated dosage 1 are first clustered into the four haplotypes, again based on pairwise correlations. **g.** The contigs with dosages > 1 are added to the clusters that contain most matching nodes in terms of k-mer count pattern correlations. **h.** This process results in chromosome clusters that contain subclusters for each haplotype. Taken from [189], licensed under the CC BY 4.0 licence.

the dosage d_i of contig i is estimated as follows:

$$d_i = \begin{cases} 1, & c_i \in \left]0.5m, 1.5m\right] \\ 2, & c_i \in \left]1.5m, 2.5m\right] \\ 3, & c_i \in \left]2.5m, 3.5m\right] \\ 4, & c_i \in \left]3.5m, 4.5m\right] \end{cases}$$

For $c_i > 4.5m$, we assign $d_i = 5$ to denote a repetitive contig.

4.2.4 K-mer counting

A *k-mer* denotes a fragment of length k. In the next step of our workflow, we count all possible k-mers (in our case k=71) within the unitigs. The k-mer counting process is based on the Jellyfish API [131]. From the full set of k-mers, we identify those appearing exactly once in the entire graph. Additionally, as the offspring samples are crosses between the two cultivars Altus and Colomba, we ensure the k-mers are unique for the Altus genome by disregarding those also found in the Colomba genome. Using Illumina sequencing data from Colomba, we construct a De Bruijn graph (DBG) of the Colomba genome using bcalm [38] and count the Altus k-mers in that DBG. K-mers that are present in the DBG – and thus, part of the Colomba genome – are excluded from the set. We report the resulting set of k-mers present only once in the entire assembly graph as well as only present in the Altus genome as *unique k-mers*.

Each unitig's set of unique k-mers is used as its identifier. Then, these k-mers are counted in each of the 193 offspring samples. To filter for extremely low k-mer counts, a unitig's k-mers are reported as *present* in an offspring only if at least 10% of the unitig's k-mer set are detected. Otherwise, they are reported as absent.

Based on the above, we can denote unitigs with unique k-mers as *phase-informative* and those without unique k-mers as *phase-uninformative*.

4.2.5 Correlation analysis for contig clustering

We expect the k-mer count patterns of two unitigs from the same haplotype to be similar. This is because when both unitigs originate from the same haplotype, the corresponding haplotype context is transmitted to the same subset of offspring samples. This assumption forms the basis of the correlation-based approach underlying our phasing method.

For this procedure, we select unitigs that are both phase-informative and have consistent dosage estimates based on both HiFi and ONT data. This ensures that the correlation

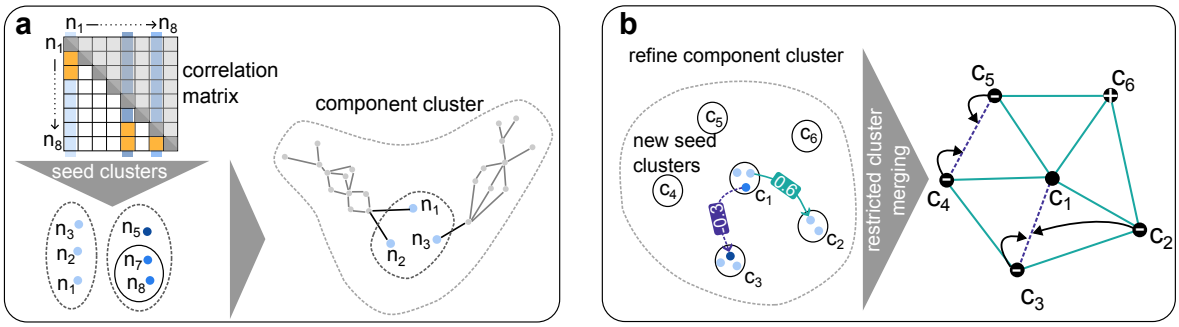


Figure 4.2: Concept of the clustering method. a: First phase, chromosome clustering. Each node is paired in a seed cluster with nodes of high pairwise correlation (marked in orange in the correlation matrix). Values are only marked in the lower diagonal matrix, as the correlation matrix is symmetrical. Seed clusters are marked in blue: n_1 pairs with n_2 and n_3, n_5 with n_7 and n_8, and n_7 with n_8. The latter two seed clusters can be merged due to the shared elements $\{n_7, n_8\}$. These clusters (n_1, n_2, n_3 in the example) are extended further by the remaining nodes of the components belonging to the cluster nodes. **b:** Second phase, cluster phasing. A component cluster resulting from the first clustering step gets further refined by building new seed clusters and merging them based on high positive or negative correlation. For this, positive edges (turquoise) and negative edges (purple) are created. Only c_6 can be merged with c_1, as all other clusters do not fulfil the requirements. Clusters are connected via an arrow with their corresponding contradicting edge.

clustering process starts with the most reliable dosage estimates. We compute Spearman correlation coefficients between the k-mer count patterns of the candidate unitigs and analyse the distribution of these correlations across the assembly graph.

The clustering of unitigs into haplotype-resolved chromosome clusters involves two main steps. First, we resolve the genome at the chromosome level. Given that chromosomes in the assembly graph may consist of several connected components along with additional singletons, it is essential to determine which graph components belong together. Second, we further divide each chromosomal cluster into four distinct clusters, one for each haplotype.

We make use of the previously computed k-mer counts in the progeny to cluster unitigs with similar k-mer count patterns. Our clustering procedure follows the idea that unitigs with highly similar patterns can be assigned to the same haplotype, as they are co-inherited to the same subset of offspring.

The similarity between the k-mer count patterns of two nodes is assessed using the Spearman correlation coefficient ρ. Nodes with highly positive correlations likely represent the same haplotype, whereas highly negative correlations suggest that the nodes belong to different haplotypes within the same chromosome. Only nodes from the same

Input: Correlation matrix **C**, threshold values θ_1, θ_2, component vector **co**
Output: Clustered nodes $\mathscr{C}$

$\mathscr{C} \leftarrow \emptyset$; // Initialize clusters
foreach *node* i **do**
  if $\max(\mathbf{C}[i,:]) > \theta_1$ **then**
    $c \leftarrow \{i\} \cup \{j : \mathbf{C}[i,j] > \theta_2\}$; // Form initial cluster
    $\mathscr{C} \leftarrow \mathscr{C} \cup \{c\}$; // Add cluster to set
  end
end
foreach *pair* $(c_1, c_2) \in \mathscr{C}$ **do**
  if $c_1 \cap c_2 \neq \emptyset$ **then**
    $c \leftarrow c_1 \cup c_2$; // Merge overlapping clusters
    $c' \leftarrow \bigcup \{\mathbf{co}[i] : i \in c\}$; // From nodes to components
    $\mathscr{C} \leftarrow (\mathscr{C} \setminus \{c_1, c_2\}) \cup \{c'\}$; // Update clusters
  end
end
return $\mathscr{C}$; // Return final clusters

Algorithm 1: Node clustering: Creating chromosomal clusters based on correlation.

chromosome should be highly correlated (positively or negatively), whereas for nodes lying on two separate chromosomes, the correlation coefficients should be low, as any similarities in their k-mer counts would occur by chance.

An overview of the clustering method is shown in Figure 4.2. In the first phase (Figure 4.2 a), which involves chromosomal clustering, we create initial 'seed' clusters. Each node i is paired with the set of nodes j with high pairwise correlation ($\rho_{i,j} > \theta$). These initial clusters, inherently containing many overlaps, are then merged based on shared elements. Finally, we identify the graph components associated with the nodes in these clusters and expand the clusters by incorporating the remaining nodes from components that have not yet been included. This process is detailed in Algorithm 1.

We denote the clusters yielded by the first step as *chromosomal clusters* or pseudochromosomes.

In the second step, we use the chromosomal clusters as input and determine the individual haplotypes for each chromosome. For this, we use the previously computed dosage estimation (Section 4.2.3) and begin by clustering unitigs with dosage 1. Being part of a single haplotype, these nodes can be assigned to a single cluster.

For the haplotype clustering, we follow an agglomerative method that, similar to the first clustering process, starts by building seed clusters between nodes with the highest correlations and then merges them into larger clusters, depending on certain restrictions.

For each node n with dosage 1, we initially create one cluster for n containing only

those unitigs with a high correlation to n ($\rho > \theta$) , producing a set of seed clusters. These clusters are subsequently merged according to the number of nodes they share. We assume that two nodes from different haplotypes will exhibit high negative correlation values and make use of this assumption to distinguish the different linkage groups. The concept is visualised in Figure 4.2 b and described in Algorithm 2.

Two clusters c_i and c_j are connected via a positive edge if there is one node pair (n_i, n_j), where $n_i \in c_i$ and $n_j \in c_j$, with high positive correlation ($\rho_{i,j} > \theta_1$). Conversely, they are connected via a negative edge if there is at least one node pair (n_i, n_j) with a high negative correlation ($\rho_{i,j} < \theta_2$). For two clusters c_i and c_j which share a positive edge, we say that c_i is a *positive neighbour* of c_j and vice versa. Then, two clusters c_i and c_j can be merged if three requirements are met:

1. c_i and c_j are connected via a positive edge
2. all positive neighbours of c_i are not a negative neighbour of c_j
3. all positive neighbours of c_j are not a negative neighbour of c_i

The latter two requirements ensure that no contradicting edge exists.

After the merging steps, the only nodes remaining outside of any clusters are those without high correlation values with other nodes ($\rho < \theta$). We include a remaining node n by assigning it to a cluster c_i if the three best hits (the nodes n_a, n_b and n_c with the highest correlation to n) all belong to c_i. If this is not the case, we are unable to assign n unambiguously to a single cluster and it is left unclustered.

Finally, we assign unitigs with higher dosages to the previously computed haplotype clusters. To cluster a node n with dosage x ($x \in \{2,3,4\}$), we compute the pairwise correlations between n and all nodes of all clusters c_1, c_2, c_3 and c_4, and add the node to the x clusters with the highest ratio of nodes that correlate positively with n.

While this rather conservative nature of our clustering procedure avoids misassemblies, it may prevent some unitigs from being added to any haplotype cluster; those remain unphased. Therefore, we employ a post-processing step in which we review the remaining unphased nodes and assign them to the cluster with the highest correlation.

4.2.6 Graph traversal and final haplotype assembly

The clustering method we described (Section 4.2.5) yields one cluster per chromosomal haplotype, each containing a set of unitigs. To enhance the completeness of the assembly, we then connect the clustered unitigs wherever possible by identifying graph traversals

Input: Correlation matrix **C**, Clusters $\mathscr{C}$, Thresholds θ_1, θ_2
Output: Haplotype clusters $\mathscr{S}$

foreach cluster $C \in \mathscr{C}$ **do**
  $\mathscr{S} \leftarrow \emptyset$;
  foreach node $i \in C$ **do**
    // Form seed clusters
    // Merge seed clusters with large intersection
    $\mathscr{S} \leftarrow \mathscr{S} \cup \{s\}$;
  end
  foreach $S_1, S_2 \in \mathscr{S}$ **do**
    if $\exists i \in S_1, j \in S_2 : \mathbf{C}[i,j] > \theta_1$ **then**
      Positive$[S_1] \leftarrow S_2$; // Create positive edges
    end
    if $\exists i \in S_1, j \in S_2 : \mathbf{C}[i,j] < \theta_2$ **then**
      Negative$[S_1] \leftarrow S_2$; // Create negative edges
    end
  end
  foreach $S_1, S_2 \in \mathscr{S}$ **do**
    if $S_2 \in$ Positive$[S_1]$
    and $\forall j \in$ Positive$[S_1] : j \notin$ Negative$[S_2]$
    and $\forall j \in$ Positive$[S_2] : j \notin$ Negative$[S_1]$ **then**
      $S \leftarrow S_1 \cup S_2$; // Merge clusters
    end
  end
  $\mathscr{S} \leftarrow \mathscr{S} \cup S$
end
return $\mathscr{S}$; // Return final clusters

Algorithm 2: Cluster phasing: Refining component clusters into haplotype clusters based on correlation.

within the assembly graph. This process results in contiguous blocks corresponding to the four haplotypes, which we refer to as haplotigs.

Starting with the unitig cluster for each chromosome, we reconstruct the ordering of unitigs across the chromosomes to identify all possible connections, with the aim of generating the most contiguous haplotypes. To do so, we rely on the graph topology in the assembly graph, inherently containing the order of the phased sequences. The principle of the graph traversal is visualised in Figure 4.3. We begin by implementing straightforward extensions (Figure 4.3, top panel): if a phased node has only one neighbouring node in either direction, that neighbour is considered to be part of the same haplotype, as that is the only possible path through the phased node. For simple bubble structures (four nodes, including source, sink and two branching nodes) where both the source and

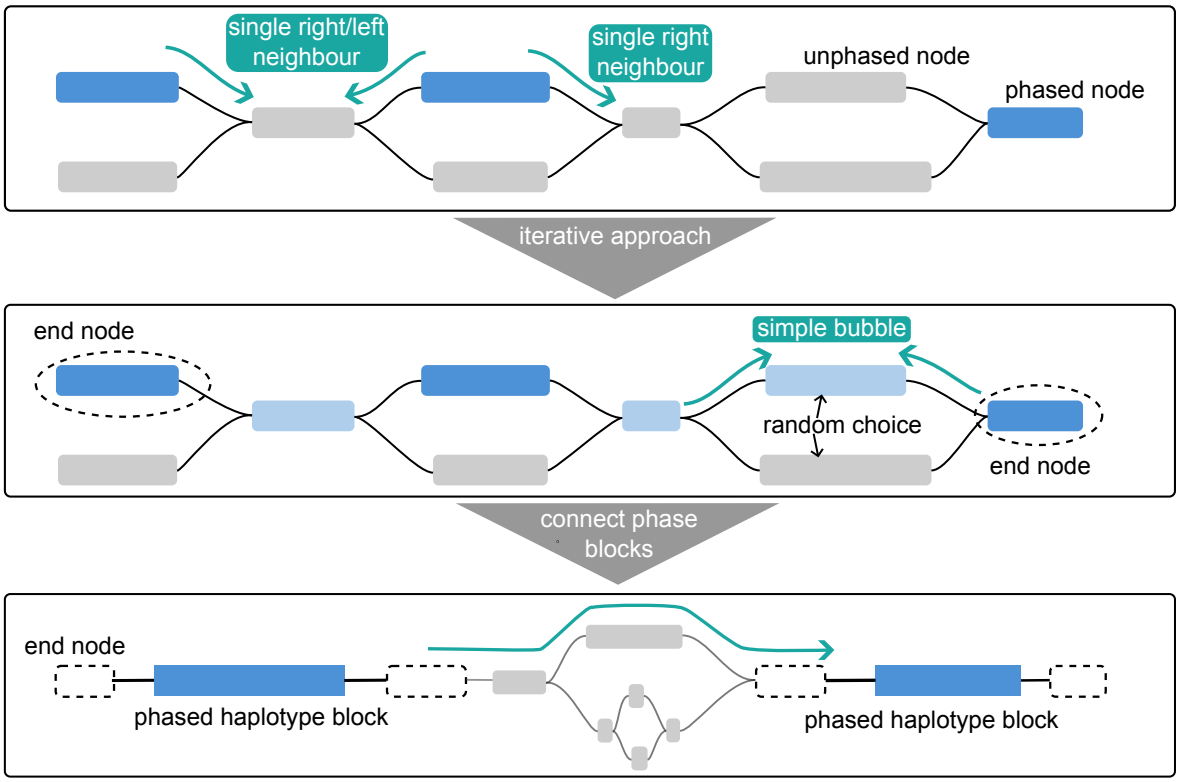


Figure 4.3: Overview of the graph traversal to bridge unphased nodes. Phased nodes from the same haplotype denoted in blue, unphased nodes in grey. **Top panel:** First, the obvious extensions to the phased paths are made. Unique neighbours to phased nodes are added to the same haplotype. **Mid panel:** Newly phased nodes marked in light blue. For simple bubbles where source and sink belong to one haplotype, the haplotype path must run up to one of the branching nodes. One of them is picked arbitrarily and the node sequence is replaced with placeholder characters for the haplotig sequence. **Bottom panel:** Two phased blocks are connected via a path through unphased nodes that connects the end nodes of the blocks (marked by dashed lines). This process results in the creation of a haplotig.

the sink node are phased, one of the two branching nodes is assumed to lie on the corresponding haplotype path (Figure 4.3, mid panel). If both branches are unphased, no information is available to pick the correct one, in which case a 'gap' is inserted by choosing one of the branching nodes arbitrarily (as no preferred option can be determined) and filling the corresponding sequence with placeholder characters ('N').

Next, we address the set of all phased nodes isolated from the rest of the graph. These form a set of linear block structures, representing haplotype blocks (Figure 4.3, bottom panel). For each block, we identify the two end nodes and reconstruct the node path (including the sequence) through the block. To reconstruct the order of these haplotype blocks, we aim to find connecting paths between the end nodes. We assume that the end

nodes of two different blocks can be connected if the only path connecting them leads through unphased nodes only. Based on this, we search for paths fulfilling this condition for every pair of block ends. For any block that can be connected uniquely to another, we concatenate the two block sequences and again insert placeholder characters over the length of the intermediate unphased fragment between them. Finally, we resolve any remaining overlaps between the extended node paths, resulting in the final output sequences.

4.3 Results

Here, we present the results of applying our k-mer-based method for haplotype-resolved assembly to the Altus genome. We present the results of each step of the workflow, including a comparison of our results to the existing reference sequence, and an evaluation of assembly quality.

This section and all subsections reuse material from [189].

4.3.1 Initial assembly graph

We produced HiFi sequencing data for the Altus genome with an average coverage of 24× per haplotype (73.7 Gb in total). Additionally, the Illumina short-read sequencing data for the 193 offspring of the cross Altus × Colomba consisted of 2 × 150 bp paired-end reads with an average coverage of 1.5× per haplotype.

The initial hifiasm graph consisted of 20,216 nodes (unitigs), 26,566 edges and contained 2798 Mb of sequence data. The N50 value of the unitigs was 1.34 Mb. The nodes of the unitig graph (Supplementary Figure C.1) within the 10 largest connected components covered 91–190 Mb each (1.27 Gb in total), 11 further components covered 45–66 Mb each (555.2 Mb in total) and a set of smaller components covered 20–32 Mb each (249.1 Mb in total). Additionally, 699 unitigs were not connected to any other node. In summary, the initial raw unitig graph provided a certain degree of haplotype resolution, indicated by the total amount of sequence data (3.8× the size of the DMv6.1 reference genome), but did not provide longer-range phasing at many loci, indicated by the substantial number of nodes shorter than 50 kb (Figure 4.4 a).

4.3.2 Dosage analysis

Figure 4.4 b shows the distribution of the average coverage. Here, we focus on long unitigs, i.e. those with a non-overlapping sequence of 100 kb or longer. Those covered ~80% of the total sequence in the graph. The same representation with the full unitig set can be seen in Supplementary Figure C.2. Three peaks were observed, representing approximate coverage values of 23, 46 and 69, consistent with dosages of 1, 2, and 3. A fourth peak (~92) was missing for the long nodes (Figure 4.4 b) and barely visible for all unitigs (Supplementary Figure C.2). This absence of unitigs covering all 4 haplotypes may indicate the existence of only a few homozygous regions and the complete absence of long homozygous stretches exceeding 100 kb.

For 6,212 unitigs, the sequence consisted solely of overlaps to both neighbouring nodes. Given the absence of a unique region, we therefore omitted these unitigs from the computation of coverage. These nodes accounted for 0.148 Gb in total, where the longest node was 42,105 bp.

Of the 8,290 nodes with a depth value above zero, 72.77% were labelled as dosage one, 15.01% as dosage two, 7.95% as dosage three, and 2.97% as dosage four. The remaining 1.3% of the contigs exceeded dosage four and are presumed to represent repetitive regions. In total, all nodes that could be assigned a dosage estimate comprised a sequence of 2.396 Gb.

These dosage estimates are a key step in our assembly process and are, due to repetitive sequence, difficult to validate using short read technologies. We therefore produced 162 Gb of long-read sequencing data from the Oxford Nanopore Technologies (ONT) platform (Supplementary Section C.1). We again estimated the dosage like described above. Comparing dosage estimates based on alignments of ONT reads to those based on HiFi reads, we found that for 6,982 nodes (84.2%), both estimates were equal. These nodes had a total length of 2.233 Gb, which is 93.22% of the formerly described 8,290 nodes. The joint distribution of ONT- and HiFi-based coverage estimates can be found in Supplementary Figure C.3, confirming the robustness of the dosage estimates, especially for nodes with sufficient amounts of unique sequence.

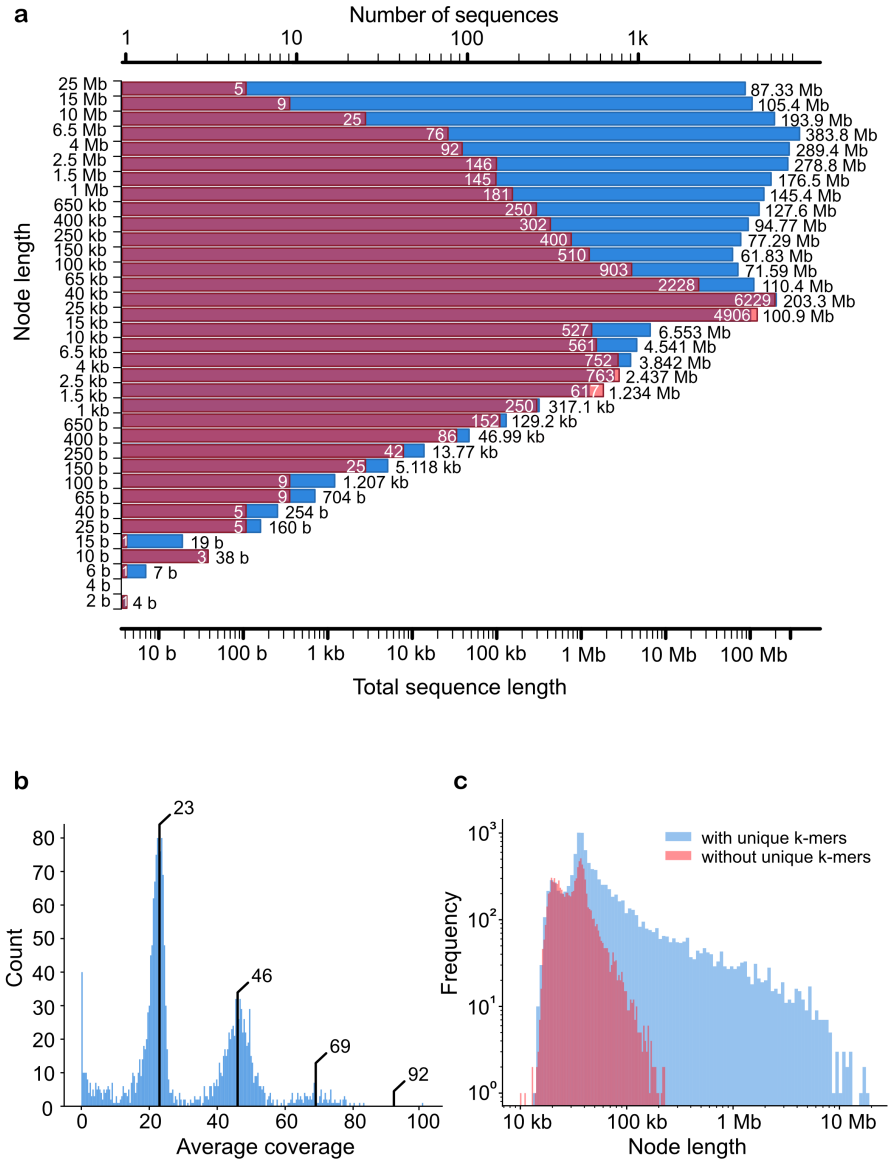


Figure 4.4: Initial assembly. a. Distribution of node lengths of the initial assembly graph. Red represents the count of each binned unitig length (the peak is 25–40 kb). Blue represents the aggregate length of a unitig bin, measured in base pairs. The two visible peaks show that the total sequence of unitigs between 25 and 40 kb is on par with the sequence taken up by those between 4.0 and 6.5 Mb. **b.** Dosage distribution of unitigs, excluding those with a unique sequence < 100 kb. The proportion of sequence that is covered by nodes at least 100 kb in length is 80%. The dosage peaks are marked by black bars (approximate coverage values of 23, 46 and 69). The peak for dosage 4 would be ~92. **c.** Length distribution of unitigs with unique k-mers compared to those without unique k-mers. Taken from [189], licensed under the CC BY 4.0 licence.

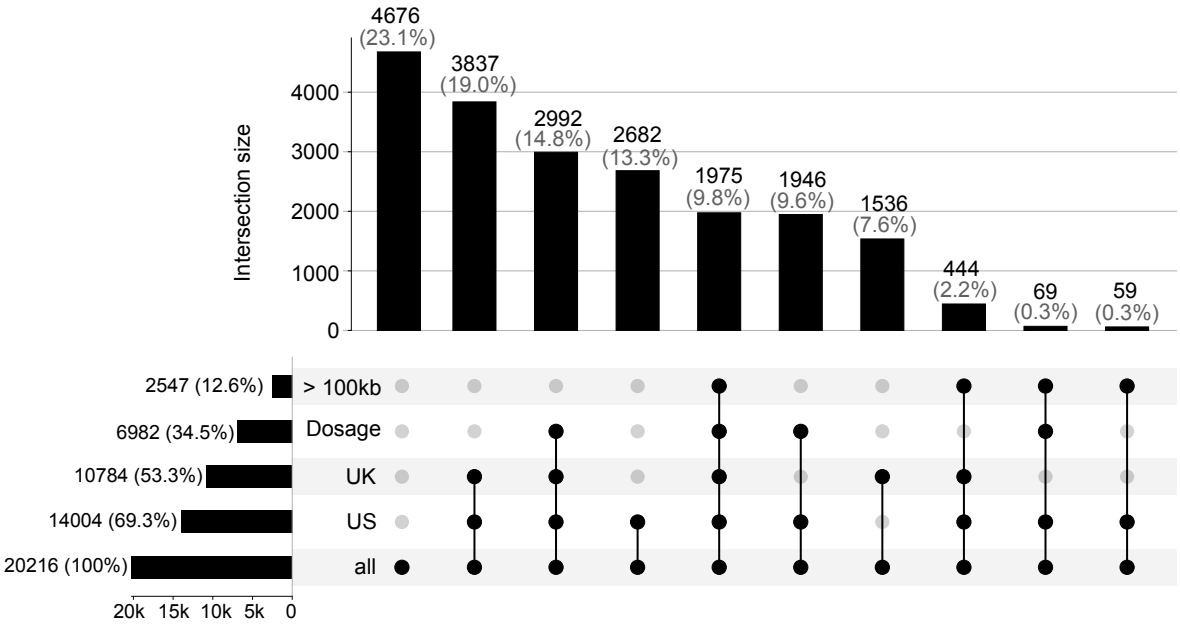


Figure 4.5: UpSet plot of the various node sets used throughout this study. 'Node' refers to the unitigs from the initial assembly graph. '>100 kb': nodes with a length of at least 100 kb. 'Dosage': Nodes for which we obtained dosage information, i.e. the dosage estimate of HiFi and ONT reads coincides. 'UK' (*unique k-mers*): Nodes that contain > 0 unique k-mers. 'US' (*unique sequence*): Nodes for which the unique – i.e. the non-overlapping – sequence is longer than 0 bp.

4.3.3 Analysis of k-mers

The analysis of node lengths for the sets of phase informative and uninformative nodes is shown in Figure 4.4 c. As anticipated, the uninformative unitigs were generally the shorter ones. Among the complete set of 20,216 nodes, we found that 10,784 (53.34%) were phase informative. The length of the sequence covered by informative nodes in relation to the sequence covered by all nodes was 88.15% (2.466 of 2.798 Gb). Specifically, the average node length in the set of phase informative nodes was 228.7 kb (N50 = 1.89 Mb) whereas the average for uninformative nodes was 35.1 kb (N50 = 37 kb). The longest unitig without a unique k-mer was 237 kb, compared to 19.11 Mb for the longest informative unitig. Thus, despite the relatively high number of phase uninformative nodes, most of the sequence (88.15%) was generally amenable to offspring-based phasing using our technique. Recall that 6,212 unitigs did not have a unique region due to overlaps, so that unique k-mers cannot be present in these nodes. The different node sets are visualised in Figure 4.5, showing the overlaps between nodes that are phase-informative, ultra-long (> 100 kb), those which contain a unique non-overlapping sequence, and those for which the dosage could be confidently estimated.

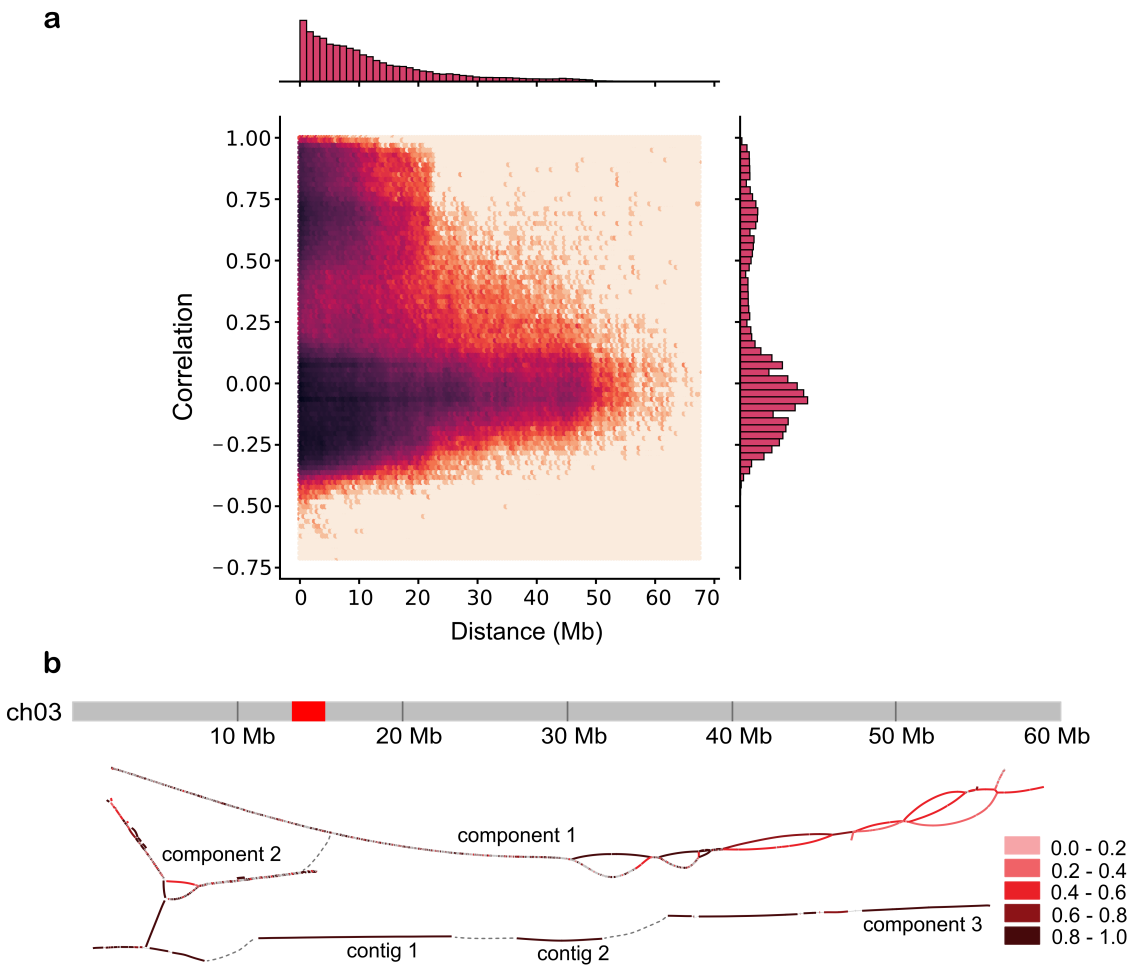


Figure 4.6: Correlation analysis. a. The correlation of all node pairs (nodes with dosage 1) in the 20 largest connected components as a function of the distance between nodes (in megabases). The 4,830 dosage-1 nodes of the largest components account for 947.78 Mb. After removing pairs which had no valid correlation (NaN), 701,582 pairs remained in the dataset for plotting. **b.** Reconstruction of the structure of chromosome 3 based on high correlation coefficients between nodes. Chromosome 3 is shown above, with the red block labelling the centromere as reported in the DMv6.1 annotation. The initial assembly consisted of three connected components and two additional contigs, which were manually placed at their approximate genomic location along the *x*-axis as determined by mapping the unitigs to DMv6.1 (the darker the colour of a contig, the higher the maximum correlation to any other contig beyond its component). Contig pairs with the highest correlation (here denoted by the darkest colour, representing a correlation coefficient of >= 0.8) could then be connected, revealing a more complete structure of the haplotype-resolved chromosome. The connected node pairs are marked by the dotted grey line. Taken from [189], licensed under the CC BY 4.0 licence.

4.3.4 Correlation clustering and graph traversal

Genomic loci that are at close distance – and thus, tightly linked – are likely to be transmitted together to progeny. Accordingly, we expect the offspring-based haplotype signal to get weaker with increasing distance due to the increased likelihood of recombination events separating loci. In line with this expectation, we observed a strong correlation for unitigs at distances < 10 Mb and decreasing correlation for greater distances (Figure 4.6 a). It is visible that negative correlation values, which indicate inversely related k-mer count patterns between the corresponding nodes, take up values of up to -0.5, and the distribution plot suggests a peak around the interval $[-0.2, -0.4]$. Thus, it seemed reasonable to choose $\theta = -0.3$ as a parameter in the cluster phasing process to set the threshold of high negative correlation between a node pair (corresponds to θ_2 in Algorithm 2). Accordingly, we chose parameters $\theta = 0.5$ and $\theta = 0.7$ as thresholds for high positive correlation needed to create seed clusters and to merge existing clusters; in line with the correlation distribution exhibiting a peak around $[0.5, 0.8]$, especially for closely located nodes.

Based on high correlation values, we were able to reconstruct areas from the graph that were unconnected in the initial assembly, such as broken bubble structures and unconnected fragments. A representative reconstruction of chromosome 3 is shown in Figure 4.6 b. In the initial assembly, this chromosome consisted of three connected components and two longer unconnected unitigs. By connecting unitig pairs with very high positive correlation coefficients ($\rho >= 0.8$), we were able to reconstruct the phased structure of the chromosome and to order the components and contigs accordingly.

Assembly results Applying our correlation-based clustering method (Section 4.2.5) to the unitigs of the assembly graph resulted in the generation of 48 clusters: 12 pseudo-chromosomes, each comprising four clusters of haplotagged unitigs.

The assembly results after clustering and subsequent creation of haplotigs via graph traversal (Section 4.2.6) are presented in Table 4.1. The longest haplotig per chromosome ranged from 11.50 Mb on chromosome 10 to 34.09 Mb on chromosome 7. The haplotig N50 value was ≤ 12 Mb and the total N50 value was 7.54 Mb.

We compared the assembled pseudo-chromosomes to the latest version of the monoploid reference, DMv6.1 [154]. Its length of 731.3 Mb is consistent with a k-mer based haploid genome size estimate from the HiFi data (730.6 Mb, Supplementary Figure C.4). To compute the N50 measures, we estimated the tetraploid genome size by using fourfold the length of DMv6.1 (2.925 Gb). The cumulative size comparison of each chromosome based on our assembled pseudo-chromosomes and DMv6.1 is provided in Table 4.1. The size of the individual phased chromosome was 3.5–4 times as large as the reference, and

Chromosome	Length of DMv6.1	N50	Longest haplotig	Sum of haplotigs
1	88.59	6.01	19.45	429.25
2	46.10	8.66	18.97	234.51
3	60.71	9.36	25.36	204.66
4	69.24	4.56	19.04	257.50
5	55.60	12.33	23.33	224.21
6	59.09	11.94	25.95	238.74
7	57.64	11.31	34.09	204.34
8	59.23	4.35	13.97	169.03
9	67.60	7.17	17.59	234.41
10	61.04	3.83	11.50	185.79
11	46.78	11.93	31.99	171.43
12	59.67	9.93	23.09	227.18
Total	731.29	7.54	34.01	2781.06

Table 4.1: Assembly statistics of the phased assembly. Results after clustering into haplotypes and constructing the final haplotigs. All lengths given in Mb. The phased length is defined as the sum of the unitig lengths contained in the four haplotypes for each chromosome. The N50 value is computed with four times the reference length as the underlying genome size. Lengths of the reference DMv6.1 given for comparison. Taken from [189].

the total phased length was ~3.8 times as large, consistent with structural variation and sequence loss on some of the haplotypes as previously observed for other cultivars [200].

Sizes of the resulting haplotype clusters alongside the size of the DMv6.1 chromosome are visualised in Figure 4.7 b. Some chromosomes, for instance 7 and 11, are rather complete, while others (e.g. chromosome 10) exhibit shorter lengths on all four clusters. For the latter, more nodes had to be excluded as they did not contain enough valid sequence to be reliably assigned. The light grey bars above each cluster size bar indicate how much sequence was contained in the first clustering step to form the chromosomal cluster, but had to be excluded from the second step. This value is not haplotype-specific, as the correct amount of sequence that each haplotype cluster is lacking cannot be specified further, so that we assigned equal portions of the unphased sequence from the chromosomal cluster to each haplotype for the purpose of the length estimate in Figure 4.7 b (light grey part split equally between haplotypes).

For chromosome 2, one haplotype appears larger than expected. To test whether this is due to an assembly misjoin to a different chromosome, we mapped the four haplotype clusters to the reference DMv6.1 (Supplementary Figure C.5). The alignment shows that the nodes from this haplotype map specifically to chromosome 2 in the reference, suggesting that it is unlikely to be a misjoin and more likely to be a result of duplicated

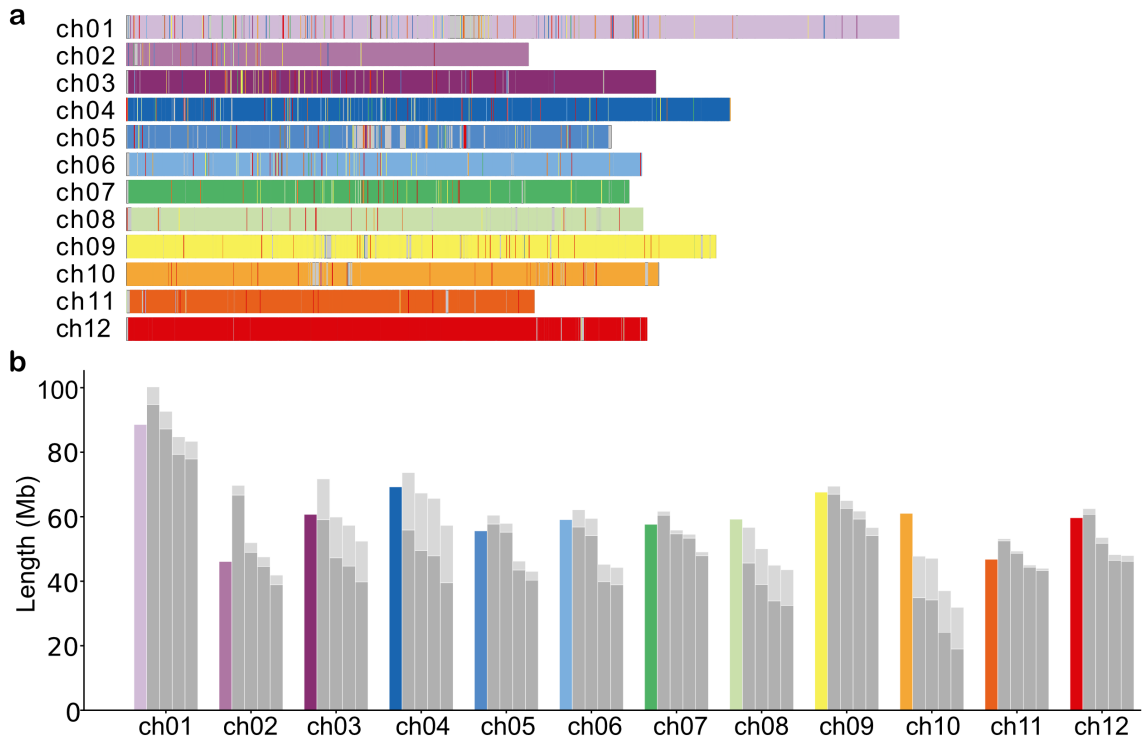


Figure 4.7: Clustering results. a. The unitigs of each chromosome cluster are mapped to the reference sequence DMv6.1, and the mapped interval is coloured accordingly. A different colour is used for each cluster. Ideally, one chromosome contains a single colour. **b.** Length comparison of the four haplotypes (grey bars) compared to the reference (coloured bars). The length is computed as the sum of the unitig lengths for all unitigs in a haplotype cluster. We have added the information on how much node sequence was included in the first clustering step to form the chromosomal cluster, but had to be excluded from the second step of phased clustering due to not being phase-informative (= lacking the unique k-mer information), marked by the light grey bars atop of each cluster size bar. This visualisation distributes the unphased sequence uniformly to each haplotype cluster. Note that the true assignment to haplotypes is, by definition, unknown for the unphased sequence. Taken from [189], licensed under the CC BY 4.0 licence.

sequence.

For a general comparison of the Altus assembly and DMv6.1, we mapped the resulting clusters to the reference using minimap2. The corresponding mapping intervals (Figure 4.7 a) indicated that all chromosomes in the assembly were nearly complete, and no large gaps were detected. In all chromosomes, the assembly consisted entirely of unitigs from one single cluster, supporting the robustness of our chromosome clustering process. For a more detailed visualisation showing mappings of all 48 chromosomal haplotype clusters, see Supplementary Figure C.6.

Downsampling experiment To investigate whether the same clustering performance could be achieved with fewer offspring samples, we repeated the clustering after downsampling the full set of 193 progeny samples to 50, 100, 120 and 150 samples, respectively. For the clustering procedure to work as expected, we found that at least 150 samples are necessary. With 50 and 100 samples, no significant clusters were created. With 120 samples, the clustering output 5 clusters instead of the expected 12 (one for each chromosome). Note that in this step, 12 is the expected number of clusters since the first step of our clustering method determines the chromosome clusters, of which we expect 12. It is only in the second step that these clusters are further refined to specific haplotype clusters. Since in our experiments with offspring sample numbers lower than 150, we noticed that even the 12 chromosomal clusters were not distinguished properly, we refrained from any further phasing steps. Only with increased sample size to 150 offspring and more, the k-mer counts got differentiated enough to distinguish between 12 chromosome clusters and output the 12 pseudo-chromosomes. The analyses presented in this chapter are all based on the full set of 193 offspring samples.

4.3.5 Comparison to earlier reference sequences reveal structural differences

We also utilised the correlation signal underlying the chromosome clustering method to detect structural differences between our assembly graphs and previous reference assemblies. Such differences can indicate assembly errors in either of the two assemblies, as well as structural differences in all or some haplotypes. When comparing the initial hifiasm assembly graph to the DMv6.1 reference, we detected two sets of nodes present on the same component of the graph that mapped to different chromosomes in DMv6.1 (Supplementary Figure C.6). For two unitig sets on separate chromosomes, we would expect to see little to no correlation between node pairs from the two sets. Indeed, for the two sets in question, the correlation distribution was very similar in shape to the correlation between one of the sets and a comparison set from a different chromosome (Supplementary Figure C.7 a). This probably indicates a false join in the hifiasm graph, which we corrected by manual curation. In this way, correlation analysis provides an opportunity to detect and correct residual assembly errors.

We then compared our assembly graphs to the diploid reference Solyntus [212] and found a number of larger structural differences (Supplementary Figure C.8). One example can be found in chromosome 8, where two regions are assembled from unitigs that belong to the same clusters as chromosome 7 and chromosome 1, respectively. To inves-

tigate whether this was a clustering artefact, an error in the Solyntus assembly, or a true structural difference, we mapped the connected components from the graph representing chromosome 8 individually to the Solyntus reference and identified one component that contained a large fragment of chromosome 1 but also the inserted region on chromosome 8 (Supplementary Figure C.9). We again compared the k-mer count correlations of all node pairs within the component, distinguishing between the sets of unitigs mapping to chromosomes 1 and 8. The former contained 563 nodes, covering 110.32 Mb, of which 315 featured unique k-mers and were thus suitable for the correlation computations (covered sequence = 102.48 Mb), whereas the latter contained 527 nodes, covering 74.6 Mb, of which 297 featured unique k-mers (covered sequence = 67.05 Mb). Again, we expect to see little or no correlation if two node sets originate from separate chromosomes. In this case, however, the distribution of correlations was consistent with the connections suggested by the assembly graph – contradicting the structure of the Solyntus reference (Supplementary Figure C.9 a). These results suggest there is either a large rearrangement that distinguishes between the Altus and Solyntus genomes, or a structural error in the Solyntus reference genome.

4.4 Structural analysis

This section reuses material from [189].

To analyse structural differences among the chromosomes in our assembly, we employed two distinct approaches for synteny analysis: SyntenyPlotteR [164] and SyRI [76]. Initially, we generated scaffolds from the assembled haplotigs using RagTag [9] and the DMv6.1 reference sequence, as the aforementioned tools require a singular chromosome-wide sequence as input for the synteny analysis. Subsequently, pairwise alignments were computed between the scaffolded haplotypes (h_1 to h_0, h_2 to h_1, h_3 to h_2, and h_0 to DMv6.1) using minimap2. The resulting synteny data were then input into SyntenyPlotteR. The synteny plot (Figure 4.8) provides an overview of the structural differences between the chromosomal haplotypes, including the reference. Overall, high synteny is observed among the haplotypes and also with the reference. Notably, there are no translocations of sequence fragments between different chromosomes, underscoring the robustness of the clustering process. For further analysis, we applied SyRI and plotsr [75] to the previously generated haplotype scaffolds in order to analyse structural differences for each individual chromosome. Synteny was computed between the scaffolded haplotypes and the reference DMv6.1 (Supplementary Figure C.10) as well as among the four haplotypes

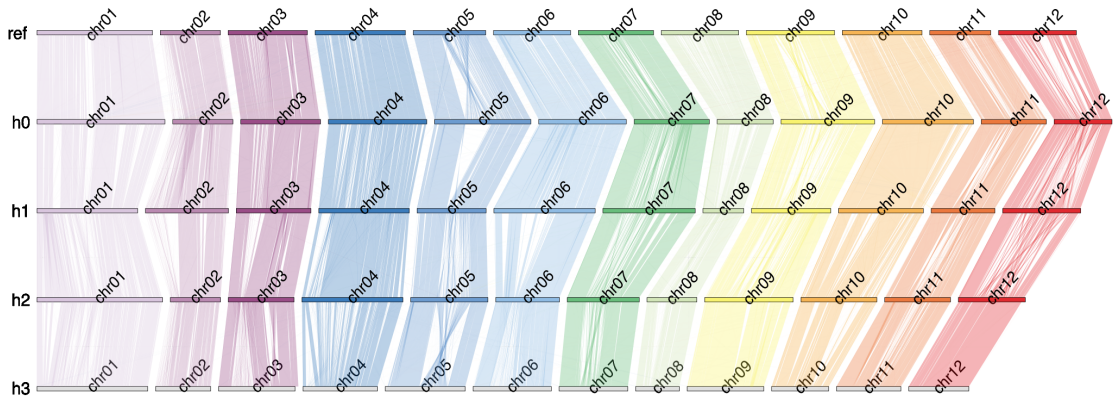


Figure 4.8: Synteny analysis. High-level overview of structural differences between chromosomal haplotypes. Top row: DMv6.1, rows 1-4: Haplotypes 0, 1, 2 and 3. Taken from [189], licensed under the CC BY 4.0 licence.

(Supplementary Figure C.11). These synteny patterns reveal extensive rearrangements, for instance on Chromosome 2, similar to those reported for other cultivars [200]. However, we caution that the outcomes of such analyses are dependent on alignment parameters, and a comprehensive investigation of rearrangement histories should be based on a larger number of different haplotypes.

4.5 Quality evaluation and comparison to other assemblies

This section reuses material from [189]. The generation of the Pore-C data is described in Supplementary Section C.1, reusing material from [189] which was contributed by Freya Ziegler, a co-author of this publication.

To evaluate our assembly, we performed further qualitative and quantitative analyses and compared the results to additional potato assemblies. To estimate the completeness of gene content, we computed BUSCO scores with BUSCO v5.4.7 [129] using the *solanales_odb10* database and obtained a completeness score of 96.8%. In comparison, this is similar to the completeness score listed for the DMv6.1 reference (97.9%), the assembly of C88 [17] (96.28%) and the assembly of Otava [200] (97.3%).

Furthermore, we annotated repeats using RepeatModeler2 [64] and RepeatMasker [2] and found 66.81% of repetitive content, the largest part of which was made up of long terminal repeats (LTRs). This is also in line with what was reported by Sun et al.[200] in the assembly of Otava (66% of repetitive content, LTRs being the most abundant group) and by Pham et al.[154] for the DMv6.1 assembly (66.8%).

For the qualitative assessment, we applied Merqury [172], a reference-free assembly evaluation method comparing k-mer databases of a read set and the assembly. Using Illumina data of Altus, we counted k-mers (k=21) using Meryl v1.4 [1, 172] and used the resulting k-mer database as input for Merqury, resulting in a QV value of 45.6677. This is similar to the QV value of C88 (46.6), while Sun et al. report a slightly better QV value of 51.7 for Otava [200]. We also used Merqury to assess assembly completeness at different stages of our pipeline: For the phased result, considering only the fully phase-resolved parts of the haplotigs, completeness was estimated at 89.48%. The result after the clustering – i.e., contigs put into a chromosome cluster, but possibly lacking accurate phasing due to insufficient k-mer information – is 97.93%. This value is similar to previously reported completeness scores of 97.3% for Otava and 99.05% for C88. The BUSCO scores per haplotype are 83.1%, 84%, 81.7% and 81.7%, respectively.

To assess the accuracy of phasing, we generated ~77 Gb of Pore-C data for Altus (Supplementary Section C.1). Subsequently, we aligned all Pore-C reads to the assembly graph nodes using minimap2 and filtered for a mapping quality of 50. We then computed the Pore-C 'coverage' of each node pair as the number of shared Pore-C reads spanning the pair. Our evaluation involves two criteria: Node pairs clustered to the same haplotype (*cis* pairs) should be supported by Pore-C reads (true positives), while node pairs from different haplotypes (*trans* pairs) should lack Pore-C support (true negatives).

Parameters that influence the ability of Pore-C data to assess a node pair in this manner include the distance between the two nodes in the graph, node length, and Pore-C 'coverage'. To quantify this, we conducted an analysis focusing on pairs of nodes longer than 1 Mb, applying various distance cut-offs (3 Mb, 5 Mb, 7 Mb, and 10 Mb) and Pore-C coverage cut-offs, ranging from 0 to 70 in steps of 1 and including a number of higher cut-offs (100 to 1200). The corresponding true positive and false positive rates were computed, and the ROC curves are presented in Figure 4.9. The areas under the curve (AUC) for the different curves are 0.96, 0.94, 0.93, and 0.93 for distance cut-offs of 3 Mb, 5 Mb, 7 Mb, and 10 Mb, respectively. These results suggest that Pore-C data can effectively support the correctness of phasing.

For a parameter setting with a distance cut-off of 5 Mb and a Pore-C coverage cut-off of 30, for instance, the achieved true positive rate is 95.22%, and the true negative rate is 92.56%. A total of 251 cis and 309 trans node pairs meet this condition. Even for node pairs further apart (up to 10 Mb), encompassing a larger number of pairs (502 cis, 706 trans), we maintain a sensitivity and specificity > 0.9, even though the Pore-C signal expectedly declines with increasing distance. In summary, the Pore-C data robustly validates the haplotype concordance of our assemblies.

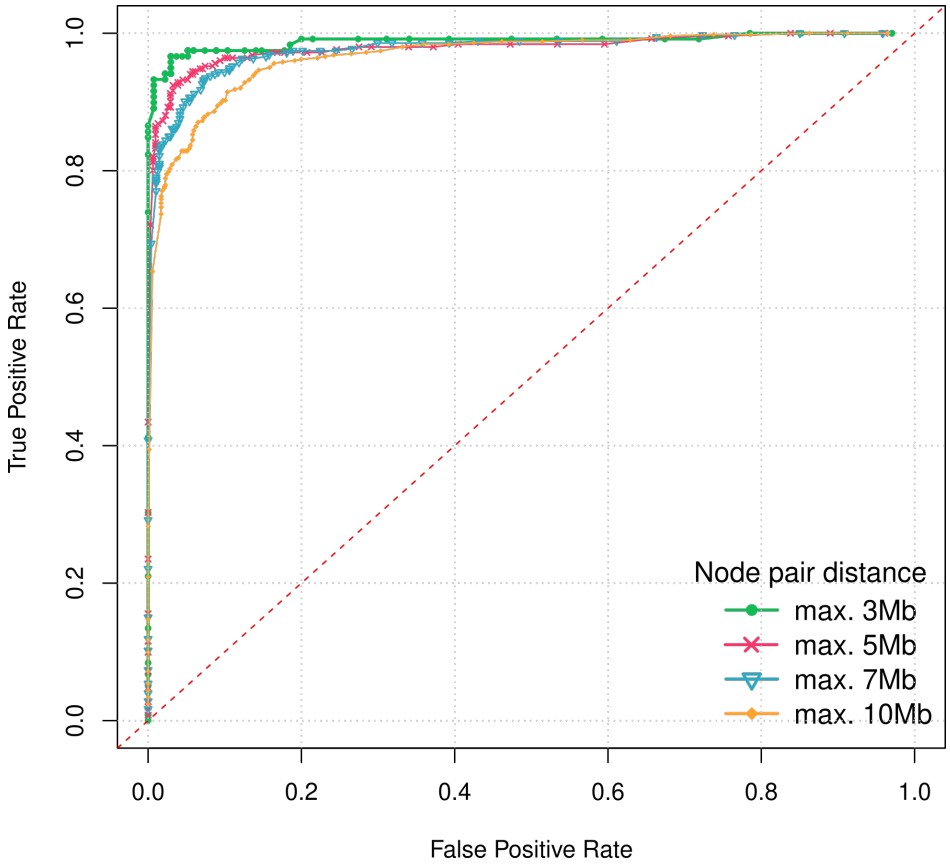


Figure 4.9: ROC curves for four different distance cut-offs. To compute the data points, all Pore-C coverage cut-offs between 0 and 70 (in steps of 1) were tested, as well as some higher cut-offs (ranging between 100 and 1200). Taken from [189], licensed under the CC BY 4.0 licence.

4.6 Discussion and conclusions

This section is an extension of material reused from [189].

We have developed a *de novo* assembly approach that uses accurate long reads and low-depth sequencing data from offspring samples to produce a phased assembly with haplotig lengths up to the length of chromosome arms. To achieve this, our method features multiple innovations. In particular, we designed a complete pipeline that uses haplotype-unique k-mers to chromosome sort and phase an assembly graph representing an autopolyploid genome. Importantly, this avoids intermediate steps that flatten the assemblies into contigs, instead resolving the haplotypes directly in the context of the graph topology, which might allow the unified integration of additional data types in the future.

The pseudo-chromosomes resulting from our assembly mapped well to the current monoploid reference genome, but we obtained ~3.8 times as much sequence data, which

indicates comprehensive haplotype resolution. By using low-pass offspring sequencing, our approach is immediately accessible in breeding and research settings where a population of offspring and standard sequencing facilities are available. An alternative route to assemblies of comparable quality is in the use of single-cell pollen sequencing [200]. Our method avoids the need for such data by computing and analysing correlation between the k-mer count patterns of different genomic loci across multiple progeny. This provides insights into how these loci are co-inherited, which in turn informs phasing and structural analysis of the genome. In parallel to our developments, Bao et al. [17] have published a similar phasing strategy to the one presented by us, but with the distinction that it uses a selfing population as offspring. We therefore view the two methods as complementary as they cater to different application settings.

Current limitations lie in the completeness of our assembly, which could still use improvement for some chromosomes. This is somewhat expected for two reasons: First, our method utilises unique k-mers, which poses a challenge in genomes as repeat-rich as the potato genome. Second, we use sequencing data from an offspring of the cross Altus × Colomba, so that those k-mers which are also present in the Colomba genome have to be discarded in order to make sure that only the signal from Altus is taken into account. Both aspects have an effect on the assembly completeness, as we found a number of unitigs to be unusable for phasing due to a lack of k-mers.

However, our method presents a solution for practical data encountered in breeding research where offspring genotyping data are typically determined at low coverage or reduced representation for cost constraints. There, the practice of cross-breeding distinct cultivars is the norm and extensive data from self-crosses are typically not available without additional experimental work. Here, we show that our method yields a reliable genome assembly tailored to such settings.

Algorithmically, there is potential for enhancing the accuracy of dosage estimation on the unitigs by employing a flow algorithm, which could go beyond the current estimation based solely on read coverage. More precise dosage estimates could, in turn, improve the clustering accuracy by having a more reliable set of nodes with dosage 1 to perform the main clustering, as well as more accurate information about nodes with higher dosages which are subsequently added to the clusters. Path threading could then potentially be optimized by considering all existing paths per component simultaneously.

Given that our method relies on the quality of the underlying assembly graph, we expect improvements in assembly methods to also enhance the effectiveness of our approach. The development of increasingly capable assembly methods is moving forward, both with improvements to existing methods like hifiasm [36], which we used in this

study, and with the development of newer methods, such as Verkko [169]. As these tools continue to develop, we expect that our graph-based framework for phased assembly will become increasingly robust and accurate.

We believe our presented assembly will be a valuable contribution to the potato genomics community. As the number of published assemblies of different potato cultivars slowly increases [202, 93], it becomes evident that graph-based pangenomics studies, like recently demonstrated for human genomes [119], will become possible for tetraploid potatoes in the near future.

Despite the rapid advances in phased plant genome assembly, haplotype-resolved chromosome-level assemblies remain challenging for complex autopolyploid genomes. The complete resolution of a haploid human genome foreshadows this development and highlights the methodological advantage of working directly on assembly graphs [146]. To resolve the most recalcitrant genomic loci, ultra-long Oxford Nanopore Technologies (ONT) reads have been aligned to assembly graphs constructed from PacBio HiFi reads [167]. We envision that our approach will be combined with such additional data types in future studies. This is currently hampered by difficulties in the preparation of ultra-long sequencing reads (> 100 kb) for plant genomes and the read length N50s of ONT reads produced in our study are currently smaller (N50 <= 33.43 kb). But we anticipate the technical challenges will be overcome in the next few years. In our present HiFi-based graphs, shorter unitigs tended to lack unique k-mers and 12% of the genome was part of such unitigs. Mapping additional sequencing data such as ultra-long ONT reads to the graphs could help to bridge the remaining gaps, allowing the inclusion of further graph nodes in the haplotype sequences and the completion of the assembly.

Chapter 5

Multigenerational graph-based assembly

In this chapter, we introduce an approach for constructing haplotype-resolved assemblies from a three-generation pedigree pangenome graph. Our method makes use of parental information in the joint assembly graph to resolve the haplotypes of the child's genome. We find that the joint assembly approach allows for identification of shared sequence that is transmitted across generations, enabling both validation of assemblies constructed from a single individual and inference of meiotic recombination breakpoints.

This chapter is based on an ongoing joint project with Monika Cechova and Karen Miga which is yet unpublished. All sections in this chapter use material from this project which I contributed. The preparation and curation of the single-sample assemblies was done by Monika Cechova. My focus was on creating the joint trio assembly, phasing of the child's genome and subsequent validation of the single-sample assemblies as well as analysis of meiotic recombination breakpoints.

5.1 Background

The Telomere-to-Telomere (T2T) consortium has made significant strides in creating fully haplotype-resolved and complete assemblies. After the publication of the first complete 'telomere-to-telomere', or T2T sequence of a human genome [146] (see Section 1.5), other advancements in the context of specific chromosomes or non-human genomes have followed [171, 127], and T2T-quality has become the new gold standard in assembly. In an ongoing joint project, we constructed assemblies of four family members of African-American descent, sequenced by the Washington University: PAN010/HG06803 (mother), PAN011/HG06804 (father), PAN027/HG06807 (daughter) and PAN028/HG06808 (granddaughter). Both accurate PacBio HiFi reads and long Oxford Nanopore reads are available

(see Table 5.1).

The concept of haplotype-resolved assembly using trio information is well established and has been implemented within multiple assembly applications [107, 35, 147, 69, 169]. The principle is based on Mendelian inheritance rules (see also Section 1.4): For example, if a child's genotype at one locus is heterozygous (0/1), the mother's is also heterozygous (0/1), and the father's is homozygous (1/1), the reference allele (0) in the child can be inferred to have been inherited from the mother. The principle of trio binning [107] involves partitioning sequence reads based on parental data before assembling the two partitioned read sets separately. However, this approach comes with the common drawback of genetic phasing where loci that are heterozygous in all family members cannot be phased [35].

A range of assembly software have included specific trio-based phasing modes to enable direct haplotype-resolved assembly in the presence of parental data, such as hifiasm [35] and Verkko [169]. In the trio binning process of hifiasm, first, k-mers are counted in parental Illumina reads and second, the unitigs in the hifiasm graph are labelled accordingly. Only those unitigs labelled with parent information are then used for the generation of a parental haplotype [35]. Similarly, Verkko contains a trio mode that also requires the computation of parental k-mer sets upfront and later integrates the haplotype information during the creation of haplotype paths through the assembly graph [169].

Here, we develop an alternative approach for the haplotype-resolved assembly of the child genome of a trio by inputting the pooled sequencing data of the trio into a single assembly method. Instead of relying on separate phasing steps, we aim at directly resolving haplotypes of the child genome within the joint assembly graph.

The creation of a first human pangenome reference through the Human Pangenome Reference Consortium (HPRC) [119] has given rise to numerous advancements in genomic research [215, 91, 78]. Among others, it has been shown that pangenomes yield a more comprehensive representation of human genetic variation [77]. In this chapter, we thereby create a small pangenome from three trio samples from the Washington University pedigree in an effort to leverage graph-based pangenomes for assembly. We have shown in Chapter 4 that graph-based k-mer approaches that make use of shared haplotype information can work well for haplotype-resolved assembly. In that case, a potato sample and several hundred offspring genomes were used to leverage inheritance information for phasing [189]. It is evident that the requirement for many offspring samples makes the exact approach not suitable for human genomes. However, the underlying concept remains relevant in scenarios involving human pedigrees and ultimately relates to the original concept of trio-based phasing.

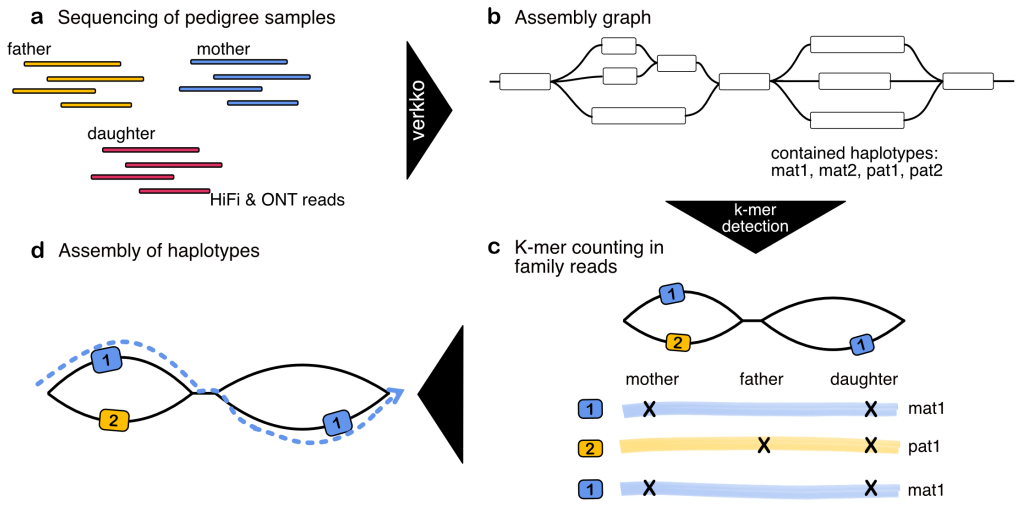


Figure 5.1: Overview of the workflow. a. HiFi and ONT reads of the mother, father and daughter are given. **b.** The sequencing data of the trio is collectively input into Verkko to build an assembly graph. **c.** Unique k-mers are detected in the nodes of the assembly graph. These node-specific k-mers are counted in the three individuals' genomes to assign haplotypes. **d.** The nodes that have been assigned to a haplotype are connected along the graph.

5.2 Assembly strategy

The general approach is similar to the method described in Chapter 4. A schematic of the method is shown in Figure 5.1. The first step involves constructing a joint assembly graph by applying Verkko (v1.3.1)[169] on pooled HiFi and ONT sequencing data from the trio comprising the father (PAN011), mother (PAN010), and daughter (PAN027) (Figure 5.1 a and b). This graph integrates data from three samples, effectively representing four haplotypes: the two maternal haplotypes from PAN010 and the two paternal haplotypes from PAN011. Since the haplotypes of PAN027 are mosaics derived from both parents, each sequence contributed by PAN027 should align with parental haplotypes, aside from potential de novo variants. We refer to the sequences of the graph nodes as *unitigs*.

Next, we detect unique k-mers in the graph (Figure 5.1 c). To do so, we identify fragments of length k (we use k=72) that appear exclusively on the sequence of a single graph node, allowing us to uniquely identify that node (= unitig). These k-mers are then searched and counted within the HiFi reads of all family members. Based on the resulting counts, the corresponding unitigs can be assigned to the subset of family members in

reads	PAN010	PAN021	PAN027	PAN028
HiFi	47×	44×	70×	44×
ONT	169×	180×	189×	191×

Table 5.1: Coverages of the HiFi and ONT reads for the four individuals. PAN010=mother, PAN011=father, PAN027=daughter, PAN028=granddaughter. 70×-73× of the ONT data refer to ultra-long reads (>100 kb).

whose genomes it is present. For instance, unitigs that appear exclusively in the child and mother are attributed to the maternal haplotype. Unitigs that are unique to the child may be further analysed as potential candidates for de novo variants.

As described previously (see Chapter 4), a threading step (Figure 5.1 d) is employed at the end, making use of the graph topology to link the phased unitigs and construct contiguous haplotypes.

5.3 Initial assembly

The genomes of PAN010, PAN011 and PAN027 (mother, father and daughter, respectively) were sequenced using PacBio HiFi technology as well as the ONT long read protocol. Coverages of the data are summarised in Table 5.1.

We pooled all HiFi and all ONT reads for the trio PAN010/PAN011/PAN027 and assembled them using Verkko v.1.3.1, yielding an initial assembly graph. A Bandage [222] visualisation of the graph is available in Figure 5.2.

The HiFi coverage in the graph as output by Verkko was 23.57× on average, with minimum and maximum values of 1.0× and 494.93×, respectively. The distribution of HiFi coverage on long unitigs (>100 kb) is shown in Figure 5.3 b. The mean coverage of ONT reads, as output by Verkko, was 74.72×, and values ranged from 0× to 3539.32×.

The initial graph consisted of 20,563 nodes, 25,219 edges and contained 8,689 Mb of sequence data, referring to homopolymer-compressed sequence. The uncompressed assembly as output by Verkko contained 12.497 Gb of sequence. The N50 value of the unitigs was 2.02 Mb, mean length was 0.42 Mb. The distribution of unitig lengths is visualised in Figure 5.3a.

The connected components contained between 362 and 1,534 unitigs, with the size of the components ranging from 189 Mb to 712.45 Mb. Mappings of the unitig sequences to CHM13 revealed that the smallest component corresponds to chr20 and the largest aligns with chr1, consistent with the chromosome sizes (Figure 5.2). Two exceptions are noted: The acrocentric chromosomes (chr13, chr14, chr15, chr21 and chr22) are all connected

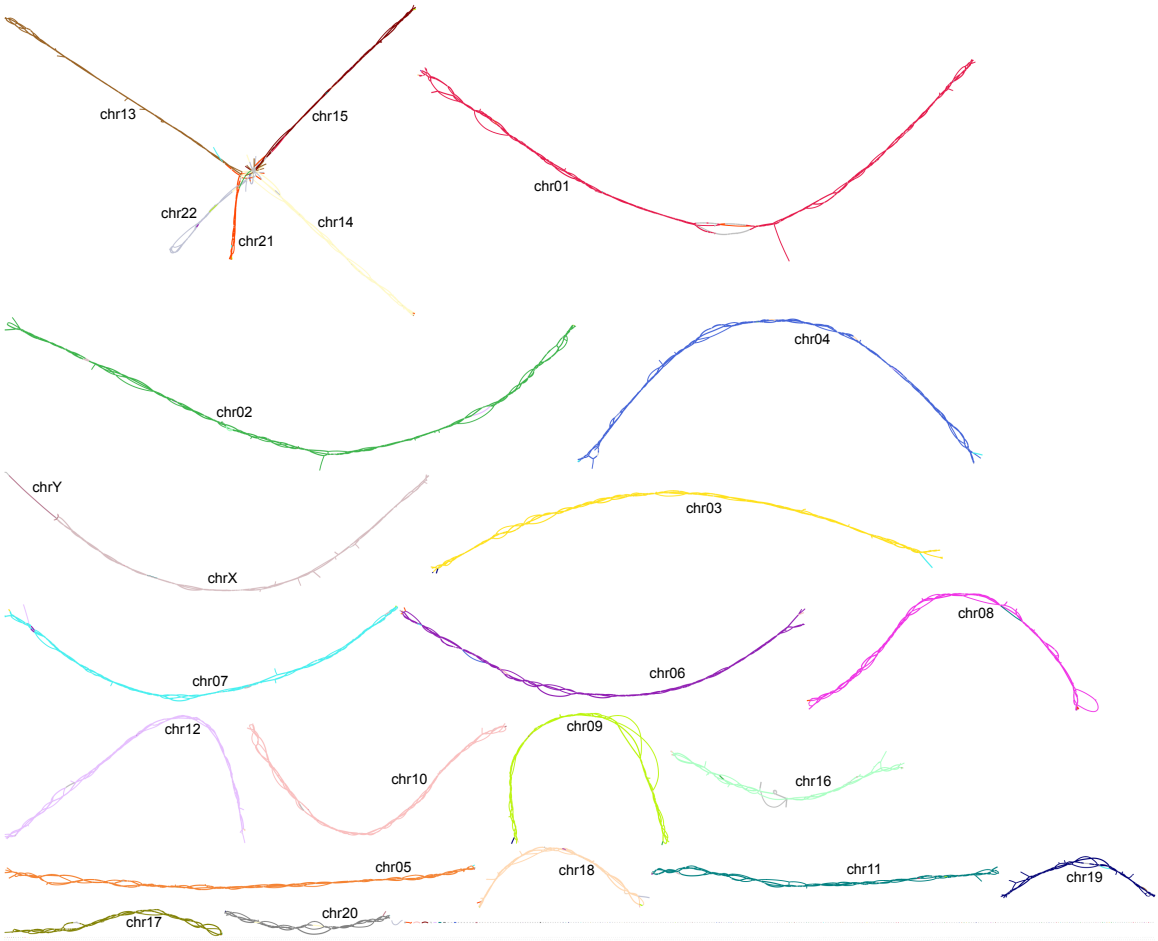


Figure 5.2: Assembly graph produced by Verkko. Bandage visualisation of Verkko graph built from HiFi and ONT reads of PAN011, PAN010 and PAN027. Unitigs are coloured and labelled based on their mapping to the reference CHM13.

within a single large component comprising 3,100 nodes and 1.171 Gb of sequence. This is expected (see Section 1.1) and in line with what was seen previously, such as the assembly graph of CHM13 [146]. Similarly, the X and Y chromosomes are clustered together in one component (630 nodes, 365.5 Mb). This is likely due to the pseudoautosomal regions (PARs), which are regions of homology between chromosomes X and Y [85]. These parts, located towards the ends of the chromosomes and spanning around 3.2 Mb [79], pair and recombine during meiosis and behave similar to homologous autosomes [85]. A total of 3,150 nodes (46 Mb of sequence) remained unconnected to any component, the majority (96%, 3,016 nodes) being shorter than 20 kb, with an average length of 14,603 bp.

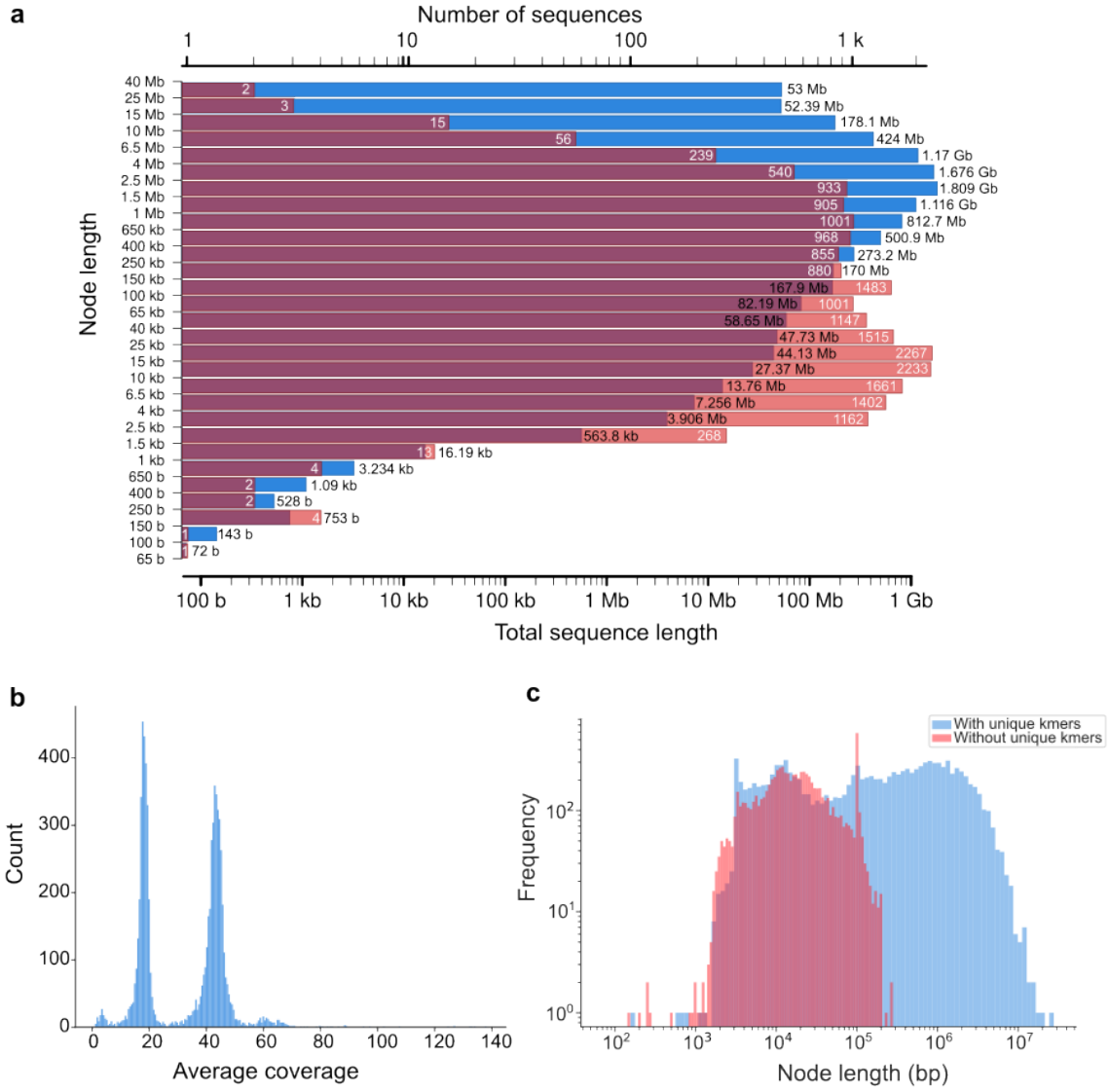


Figure 5.3: Initial assembly. a. Distribution of node lengths of the Verkko graph. **b.** HiFi coverage of nodes > 100 kb. **c.** Length distribution of unitigs with or without unique k-mers.

node set	no. of nodes	max length	min length	mean length	total length
all	20,563	27.63 Mb	0.000072 Mb	0.4 Mb	8688.88 Mb
k-mers	12,251	27.63 Mb	0.000072 Mb	0.69 Mb	8439.02 Mb
no k-mers	8,312	0.28 Mb	0.000143 Mb	0.03 Mb	249.86 Mb

Table 5.2: Statistics among different sets of nodes. First row: All nodes without restrictions. Second row: Phase-informative nodes, i.e. nodes with unique k-mers. Third row: Phase-uninformative nodes, i.e. without unique k-mers.

5.4 K-mer based assembly

Next, we counted all possible k-mers (k=72) within the graph unitigs and identified all unique k-mers. This allowed us to barcode the corresponding unitigs by their sets of unique k-mers. The k-mer counting process was performed using FastK [143]. 8,312 unitigs were left without any unique k-mers, accounting for 0.25 Gb of sequence in total. Figure 5.3 c shows the length distribution of unitigs with a set of unique k-mers versus those without. In accordance with the notation introduced in Chapter 4, we refer to the set of unitigs with unique k-mers as *phase-informative* and to all others as *phase-uninformative*. Basic statistics of the mentioned node sets are given in Table 5.2.

Notably, 97.12% of the entire sequence (8.44 Gb out of 8.69 Gb) were covered by phase-informative nodes. Within the set of phase-uninformative nodes, 3,750 lacked a unique region and consisted solely of overlaps from the neighbouring nodes. The average node length in this set was 33.53 kb, spanning a total sequence of 125.74 Mb.

Another 21.32% of the phase-uninformative set (1,772 unitigs in total) were part of the component spanning the acrocentric chromosomes; 923 of these nodes lacked a unique sequence altogether. Consequently, the acrocentric tangle remained a complex problem for assembly in this approach, as only 42.8% of the unitigs contained within were phase-informative.

Each unique graph k-mer identified in this step was then counted in the reads of all four pedigree members (mother, father, daughter, granddaughter). This allowed us to examine the co-occurrence of these k-mer markers in subsets of the family, providing information about which unitigs — barcoded by their set of k-mers — share sequence with which individuals. Based on this, we assigned the corresponding unitigs to the appropriate haplotype, for instance, to the paternal one if the markers of a unitig were found in the reads of the father and the daughter.

From this initial set of phased unitigs, we extended to longer spanning haplotypes by traversing the graph and connecting potential unphased and homozygous unitigs to the phased ones. We refer to the longer connected fragments as *haplotigs*. This resulted

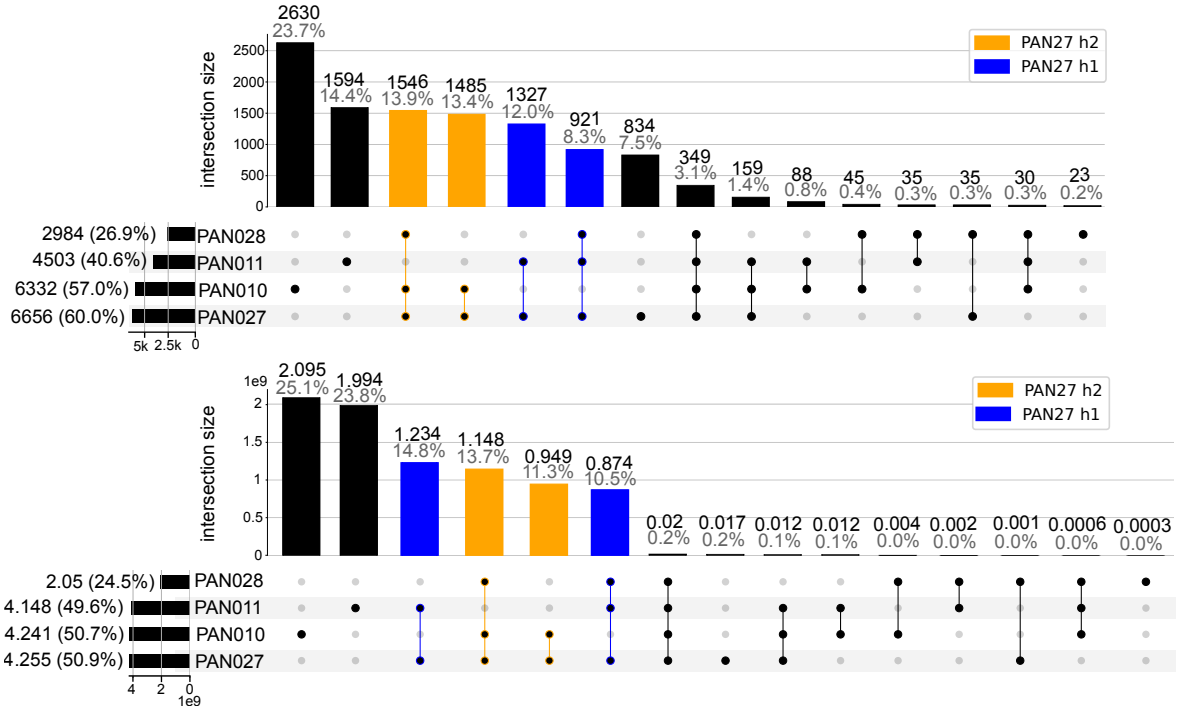


Figure 5.4: UpSet plot. Given are all possible combinations of sets combining PAN010 (mother), PAN011 (father), PAN027 (daughter) and PAN028 (granddaughter). **Top:** Number of shared nodes in the respective sample sets. 3031 nodes belong to the maternal haplotype of PAN027 (orange), 2248 nodes to the paternal haplotype (blue). **Bottom:** Length of the sequence in the sample sets. Maternal haplotype (orange) accounts for 2.097 Gb, paternal haplotype (blue) for 2.108 Gb of sequence. Note that the total length values include overlapping sequences between neighbouring nodes of the genome graph.

in two paths through each graph component which yielded the paternal and maternal haplotypes of PAN027 (daughter).

This last step of connecting nodes via path finding was performed on each graph component individually. For this, the graph was split into its components beforehand using `gfa_subgraphs` [43].

Figure 5.4 illustrates the distribution of different node sets following phasing through haplotagging of unitigs. The analysis revealed 3,031 nodes representing 2.097 Gb of maternal haplotype sequence and 2,248 nodes representing 2.108 Gb of paternal haplotype sequence. Both haplotypes accounted for approximately 25% of the total sequence. For both the paternal haplotype inherited by PAN011 and the maternal haplotype inherited by PAN010, the length of the second haplotype, which was not passed on to PAN027, was roughly equivalent. Consequently, each of these remaining haplotypes constituted about 25% of the total sequence.

5.5 Assembly results

We present the results for haplotype-resolved assemblies of PAN027. On average, a chromosome was assembled within 4 haplotigs, with a maximum of 6 for haplotype 1 of chromosome 2, and 10 haplotigs for haplotype 2 of chromosome 4. Chromosomes 22, 21, 20, 18, and 6, are assembled into a single haplotig for at least one haplotype. The N50 values of this assembly range from the length of chromosome arms to full chromosomes (see Table 5.3).

When comparing the lengths of our assemblies — computed as the sum of haplotig lengths for each assembled chromosome — to the lengths of CHM13 (Table 5.3), we observe a high degree of concordance. Notable exceptions are chromosome 9, where CHM13 contains a multi-megabase expansion [146], and chromosome 16.

The acrocentric chromosomes present the greatest challenge; due to the contained rDNA arrays, they are particularly hard to assemble [166]. Considering the complexity of these regions, a manual intervention was required to assemble at least parts of the chromosomes. To this end, we manually selected those graph nodes that were no part of the tangled region that encompasses all acrocentrics by their p-arms. By leaving out the single tangle containing the remaining parts of the chromosomes, including their p-arms, this enabled us to assemble at least the q-arms of the acrocentrics.

As a result, the lengths of these chromosomes are expectedly shorter than the reference. For example, chromosome 22 is 15 Mb and 9 Mb shorter in haplotypes 1 and 2, respectively, compared to the CHM13 reference, which lists the p-arm length as 12.8 Mb (15.7 Mb including the centromere). Chromosome 21 is 9 Mb and 10 Mb shorter, with a p-arm length of 11 Mb (11.3 Mb with the centromere). Chromosome 15 is shorter by 7 Mb and 16 Mb, and its p-arm length is recorded as 16.7 Mb (17.7 Mb including the centromere). Chromosome 14 is missing 5 Mb and 10 Mb, with a p-arm length of 10.1 Mb (12.7 Mb including the centromere), and chromosome 13 is 13 Mb and 90 Mb shorter (suggesting larger missing regions of the q-arm as well), while its p-arm consists of 15.5 Mb (17.5 Mb including the centromere). Similarly, chromosomes X and Y are connected as a single component in the graph (Figure 5.2), resulting in these chromosomes being assembled together. The resulting assembly length corresponds to the reference length of chromosome X.

chr	haplo	#haplotigs	length	N50	max	CHM13
1	1	5	243.28	210.64	210.64	248.39
	2	9	243.44	134.13	134.13	
2	1	6	242.78	89.08	112.11	242.70
	2	4	242.55	138.24	138.24	
3	1	4	202.64	181.82	181.82	201.11
	2	5	191.95	110.48	110.48	
4	1	3	192.69	185.17	185.17	193.57
	2	10	182.76	61.28	72.63	
5	1	3	182.27	157.48	157.48	182.05
	2	4	183.90	141.96	141.96	
6	1	1	172.01	172.01	172.01	172.13
	2	7	172.28	86.62	86.62	
7	1	2	161.25	127.21	127.21	160.57
	2	8	160.34	80.80	80.80	
8	1	3	146.20	44.36	65.68	146.26
	2	4	147.15	113.42	113.42	
9	1	2	136.26	116.68	116.68	150.62
	2	6	129.22	34.13	73.14	
10	1	2	134.53	134.50	134.50	134.76
	2	5	134.39	85.91	85.91	
11	1	3	134.85	134.72	134.72	135.13
	2	2	134.83	90.57	90.57	
12	1	4	133.48	92.97	92.97	133.32
	2	3	134.95	117.73	117.73	
13	1	2	99.89	72.30	72.30	113.57
	2	1	25.52	0.00	25.52	
14	1	3	96.75	81.44	81.44	101.16
	2	7	91.11	33.36	37.21	
15	1	2	92.15	67.16	67.16	99.75
	2	4	83.15	58.77	58.77	
16	1	2	90.48	42.89	47.59	96.33
	2	5	85.79	40.80	43.15	
17	1	3	83.77	82.51	82.51	84.28
	2	8	80.58	58.74	58.74	
18	1	1	80.49	80.49	80.49	80.54
	2	5	81.38	28.80	38.04	
19	1	2	63.56	48.10	48.10	61.71
	2	5	61.70	59.52	59.52	
20	1	1	66.51	66.51	66.51	66.21
	2	2	66.04	66.02	66.02	
21	1	1	36.62	36.62	36.62	45.09
	2	1	35.47	35.47	35.47	
22	1	2	36.39	12.87	23.52	51.32
	2	1	42.20	42.20	42.20	
XY	1	1	154.09	154.09	154.09	154.26 (X)
	2	8	152.87	46.60	50.62	62.46 (Y)

Table 5.3: Assembly statistics for PAN027. Showing the number of haplotigs, the length as the sum of all assembled haplotigs, the N50 and length of the longest haplotig, as well as the length of CHM13, for each chromosome. All lengths are given in Mb. CHM13v2.0 chromosome sizes as provided by `https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_009914755.1/`

5.6 Synteny analysis

This section refers to assemblies generated on an individual basis. These assemblies have been produced by Monika Cechova.

To compare our assemblies with the CHM13 reference genome, we identified syntenic regions. In order to do so, we first scaffolded the assembled haplotigs using RagTag [9], which involved aligning them to the CHM13 reference genome. The lengths of the scaffolded chromosomes, listed in Supplementary Table D.2, are mostly consistent with the pre-scaffolded lengths, although some smaller haplotigs that did not map to the reference were excluded.

Next, we utilised SyRI [76] to identify syntenic regions as well as regions of potential structural variation, including inversions, duplications, and translocations. The output was visualised using plotsR [75].

Results for chromosomes 10 and 16 are shown in Figure 5.5. Chromosome 10, among other chromosomes, showed high synteny to CHM13 (Figure 5.5 a). The only notable structural variation on the paternal chromosome is a 1.3 Mb inversion in the 10q11.22 region that starts at 47.39 Mb (47.22 Mb in CHM13); the same region on the maternal chromosome also harbours a 0.85 Mb inversion. In the maternal chromosome, the centromeric region (10p11.1 and 10q11.1) between 39.63 Mb and 41.66 Mb additionally show small translocated and duplicated regions. The inverted region lies within the same chromosome band q11.22 that has been shown to harbour numerous recurrent inversion polymorphisms [160]. In this particular region, six recurrent inversions in the mid length range (10 kb-100 kb and 100 kb-1 Mb) were detected previously [160].

In contrast, other chromosomes, in particular chromosomes 9 and 16, exhibit more structural variation, which is consistent with their previously observed size discrepancies and the known expansion on chr09 in CHM13. Paternal chromosome 16, shown in the top row of Figure 5.5 b, displays several duplicated and translocated regions. Some are located in the centromeric region (35.85 Mb - 37.83 Mb), where, due to its complexity, alignments are likely unreliable. Others are located in the 16q11.2 region, where also the maternal chromosome contains several translocations and duplications. The paternal haplotype harbours a 4.4 Mb inversion in the 16q22 region and a 0.61 Mb one in 16p13.11. The latter is another region known for multiple mid-size recurrent inversions (100 kb-1 Mb) [160].

We compared the assemblies produced by our pooled approach to those generated by assembling each sample individually [1]. For these individual assemblies, the same sequencing data was used, with Verkko (v1.4.1) handling the initial assembly. Phasing was achieved using Hi-C reads or, in the case of PAN027, alternatively using trio data. Additionally, the final assemblies went through manual curation. Moving forward, we will refer to this method as the single-sample assembly.

The lengths of the assembled chromosomes (see Supplementary Table D.1) largely

[1] https://github.com/biomonika/HPP/tree/main/T2T-Pedigree-project%20

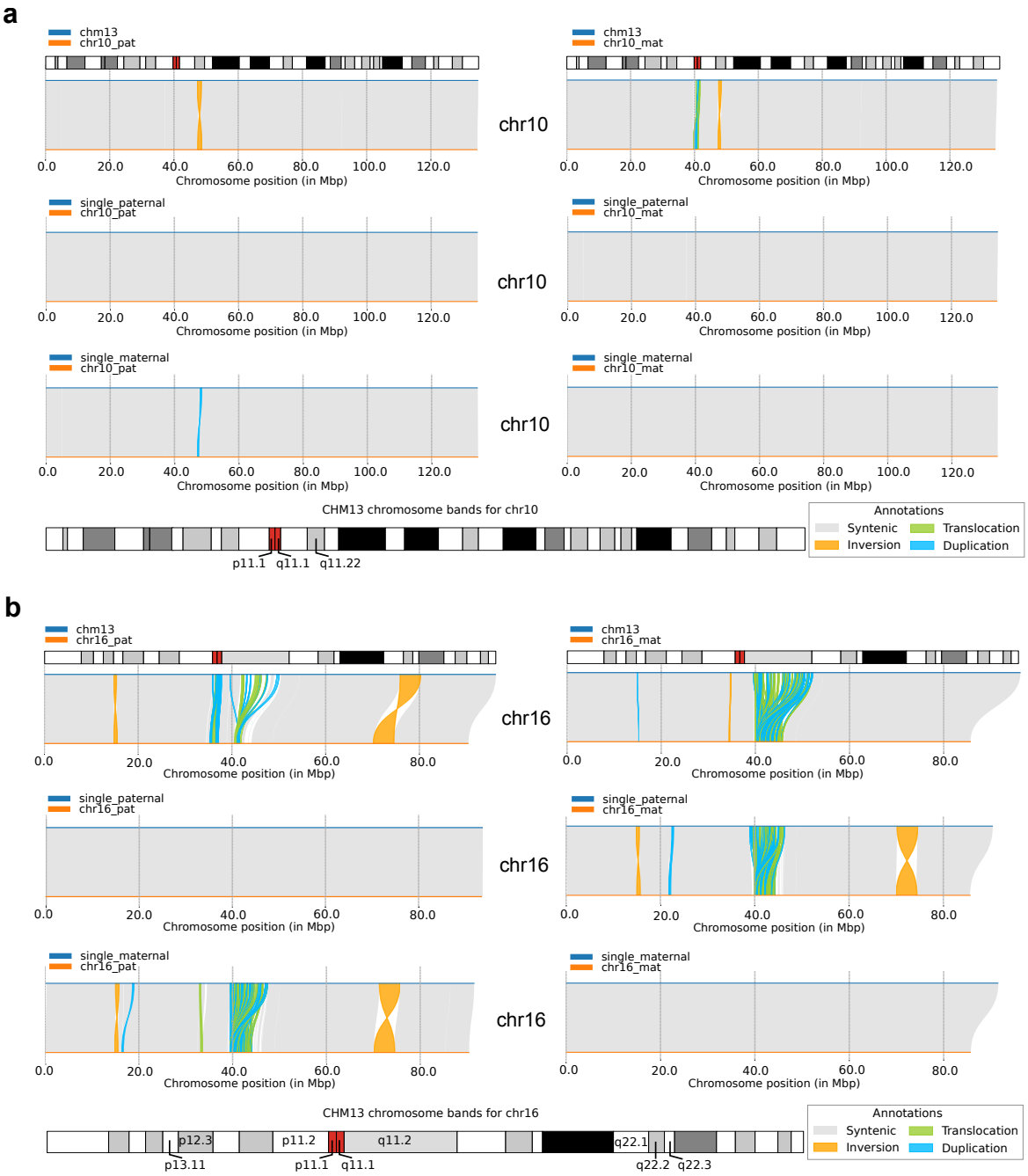


Figure 5.5: Synteny analysis for two chromosomes as computed by SyRI and plotsr. a: Chr10. **b:** Chr16. In both cases, the paternal haplotype is depicted on the left, the maternal one on the right. Top row: Comparison against CHM13 (reference: blue, assembly: orange). Middle row: Comparison against the paternal haplotype of the single-sample assemblies. Bottom row: Comparison against the maternal haplotype of the single-sample assemblies. Chromosome bands for CHM13 are shown in greater detail below; only bands with notable structural differences are labelled. The centromere is marked in red.

align with both the CHM13 reference and our graph-based assemblies.

In addition to the structural comparison of the PAN027 assemblies and CHM13, we also assessed synteny between our graph-based assemblies and the single-sample assemblies of PAN027. Most chromosomes display near-complete synteny, with only a few structural variants, primarily located around the centromere. Chromosome 16 exhibits similar variations as in comparison with CHM13 (Figure 5.5 b, middle and bottom rows). In contrast, chromosome 10 shows minimal variation (Figure 5.5 a, middle and bottom rows). Interestingly, the inversion in 10q11.22 that was detected in comparison to CHM13 is absent in this case; instead, both assemblies are highly syntenic. This might be indicative of an actual inversion in the genome; notably, it covers a region where four inversion polymorphisms were reported in [159], further supporting this possibility.

5.7 Annotation of conserved regions through transmission information

As the genomes of the trio are jointly assembled in one genome graph, whenever differences in the individuals' genomes are detected, this would show in the graph as a bifurcation. Consequently, graph unitigs that are shared between individuals from the pedigree will be identical across generations. Those unitigs that are tagged with samples from three generations – mother, daughter and granddaughter, or father, daughter and granddaughter – thus contain sequence that is present in the oldest generation and has been passed on to the next two generations. We refer to these as unitigs with two detected transmissions.

Similarly, unitigs tagged with samples from only two generations – where the sequence is present in the mother's and daughter's genomes (or father's and daughter's), but not in the granddaughter's genome – are termed as having one transmission.

We identified 2,467 unitigs with two transmissions, covering a total sequence length of 2.022 Gb. Of these, 1,546 nodes belong to the maternal haplotype (total length 1.148 Gb), indicating their presence in the mother, daughter, and granddaughter (PAN010, PAN027, and PAN028). Consequently, the remaining 921 nodes (total length 0.874 Gb) are part of the paternal haplotype.

Additionally, we identified 2,812 unitigs with one transmission, spanning a total of 2.183 Gb. Among these, 1,485 are maternal nodes (0.949 Gb), found in the mother and daughter, while 1,327 are paternal nodes (1.234 Gb), present in the father and daughter. All these node sets are illustrated in Figure 5.4.

Since unitigs with one or two transmissions are preserved across two or even three generations, they can be seen as 'confident' regions and thereby considered reliable indicators of assembly quality. Thus, to enhance quality control in single-sample assemblies, we mapped these unitigs to the assemblies and annotated the preserved regions accordingly.

We utilised minimap2 [112] to align unitigs with two transmissions, as well as those with one transmission, to the diploid assemblies. We focused on regions where unitigs mapped unambiguously and merged adjacent intervals if the gap between them was short (< 1 Mb).

Figure 5.6 shows the mapped regions within the full chromosomes.

Next, we identified discrepancies between the mapped unitigs and the assemblies by calling variants using PAV (v2.3.4) [56, 13]. For the paternal haplotype, we used the single-sample assembly of the paternal PAN027 haplotype as a reference and mapped the unitigs identified as paternal (joining the unitig sets with one and two transmissions). The called variants then highlight differences between the graph sequences and the single-sample assemblies. The same process was applied to the maternal haplotype. Variants are visualised in the additional tracks of Figure 5.6. The density of single nucleotide variants (SNVs) is shown in the track below each chromosome, the larger variants are denoted by a marker.

The SNVs appeared in clusters, evident through peaks in the SNV density. These clusters were often localised to specific chromosome regions, such as the centromere, as can be seen in chromosome 8 in the paternal haplotype or chromosome 1 in the maternal haplotype, for instance. Additionally, SNVs were found in areas not mapped by graph nodes with one or two transmissions, which are therefore not flagged as confident regions. Examples include the beginning of paternal chromosome 13 and the area around the centromere of chromosome 4. These regions showed variability in both haplotypes and could not be fully annotated.

In total, 6,747 SNVs were detected in the maternal haplotype, 10,300 in the paternal one. A few structural variants were detected by PAV. For the maternal haplotype, 54 deletions were found, affecting a total length of 252,641 bp. Among those, maternal chromosome 4 harboured a larger deletion of 156 kb. Furthermore, 65 insertions were detected (147,691 bp in total). For the paternal haplotype, 46 deletions affecting 100,836 bp as well as 51 insertions ranging over a total of 68,955 bp were detected.

In total, the unitigs preserved over three or two generations covered 2.097 Gb in the maternal haplotype, equating to 73.2% of the single assembly (2.865 Gb). Of those, 99.98% (2.0966 Gb) could be verified as there were no differences found. For the paternal haplo-

type, the total length of preserved unitigs was 2.108 Gb, 73.4% of the provided assembly (2.871 Gb). Nearly all of that could be verified in this way, as less than 0.01% of the preserved regions was part of the detected variation.

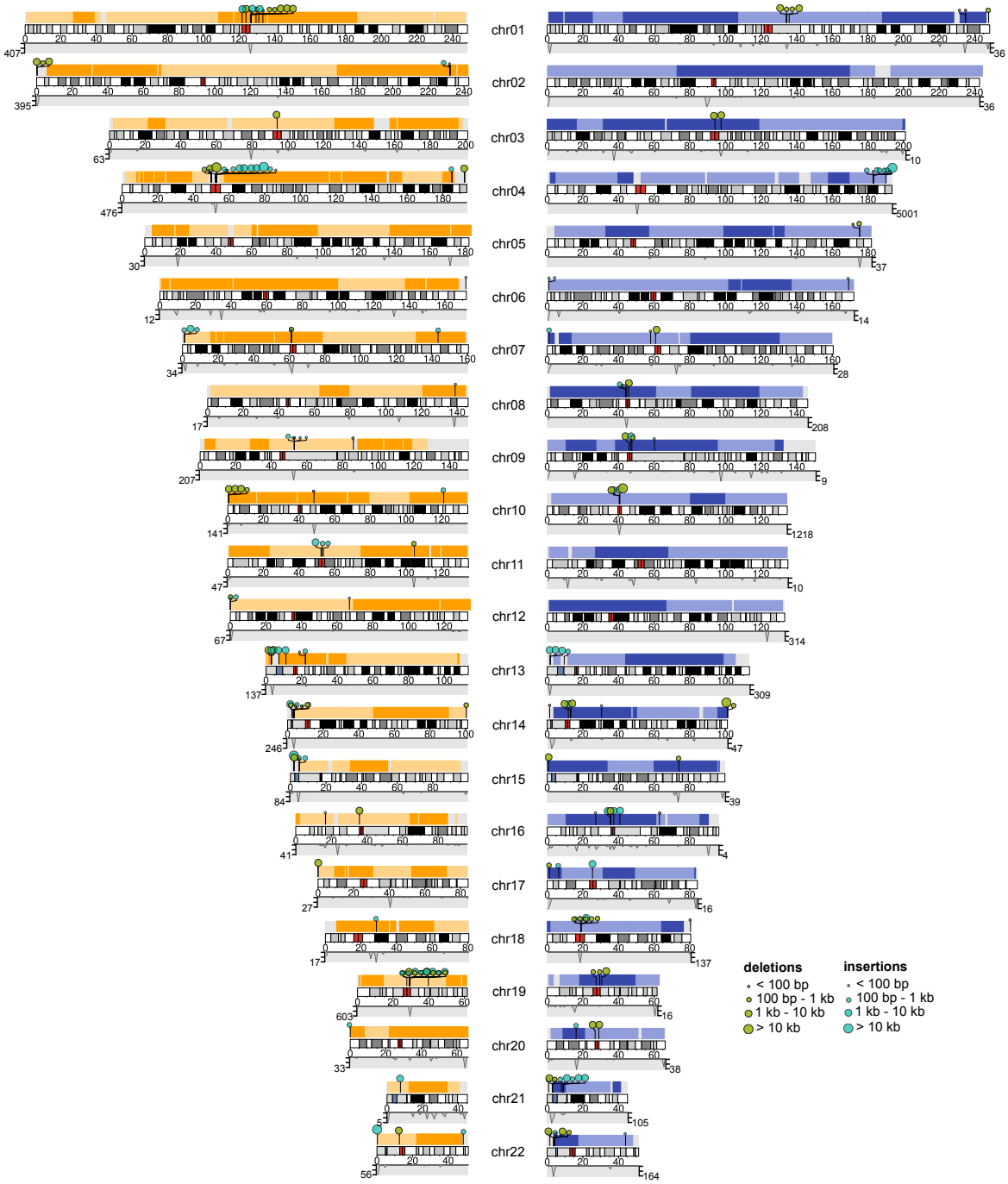


Figure 5.6: Ideograms of annotated regions in single assemblies. Shown are those regions in the single assemblies (left: maternal haplotype, right: paternal haplotype) where nodes with two and with one transmission are mapped. Orange: mapping of maternal haplotype unitigs (dark colour: present in PAN010/PAN027/PAN028, light colour: PAN010/PAN027), blue: mapping of paternal haplotype unitigs (dark: PAN011/PAN027/PAN028, light: PAN011/PAN027). The circular markers denote structural variants, the track below the ideograms shows the SNV distribution in 100 kb bins (all variants as detected by PAV). The scale next to the SNV track ranges from 0 to the maximum binned SNV count.

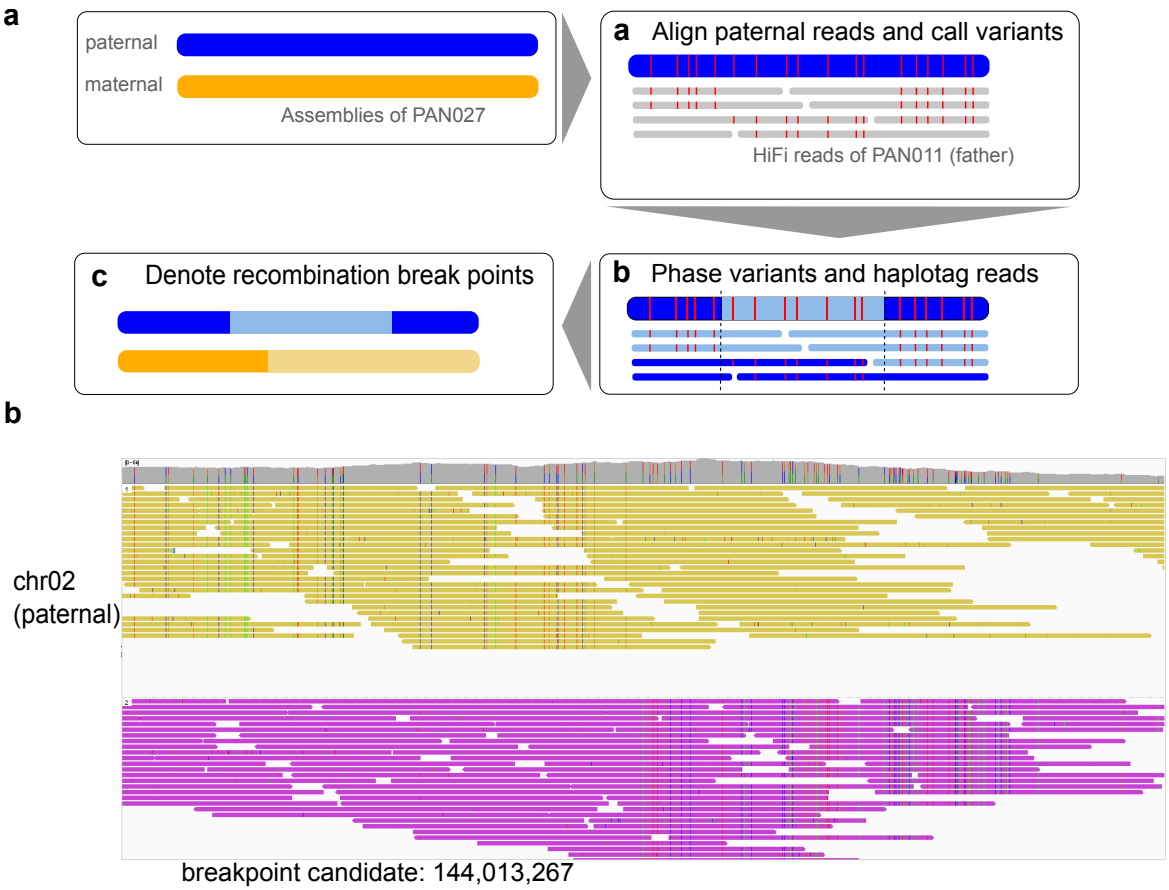


Figure 5.7: Recombination analysis. a: Overview of the procedure for generating potential meiotic recombination breakpoints. Starting from the assemblies, the HiFi reads of the corresponding parent are aligned, variants are called and phased. The light and dark shades of blue in panel b denote the maternal and paternal reads of PAN011. **b**: Screenshot from IGV [173] of one potential breakpoint position on paternal chr02 around 144 Mb. HiFi read alignments are coloured by haplotype.

5.8 Computation of recombination breakpoints

We analysed the single-sample assemblies with respect to meiotic recombination breakpoints. In order to infer breakpoints, we used two methods.

The first one is illustrated in Figure 5.7. At first, the paternal HiFi reads, which contain sequence from both haplotypes, are aligned to the paternal haplotype assembly of PAN027 using minimap2 (v.2.26) [112]. Next, variant calling is performed using DeepVariant (v1.5.0) [157]. We anticipate observing variation when a read aligns to a region belonging to the father's paternal haplotype but originates from his maternal haplotype, or vice versa. Additionally, we expect homozygosity in alignments from the father's maternal or paternal haplotype to regions in the assembly that correspond to the same haplotype. Between two of such regions, we expect a phase switch. The analysis is repeated with the

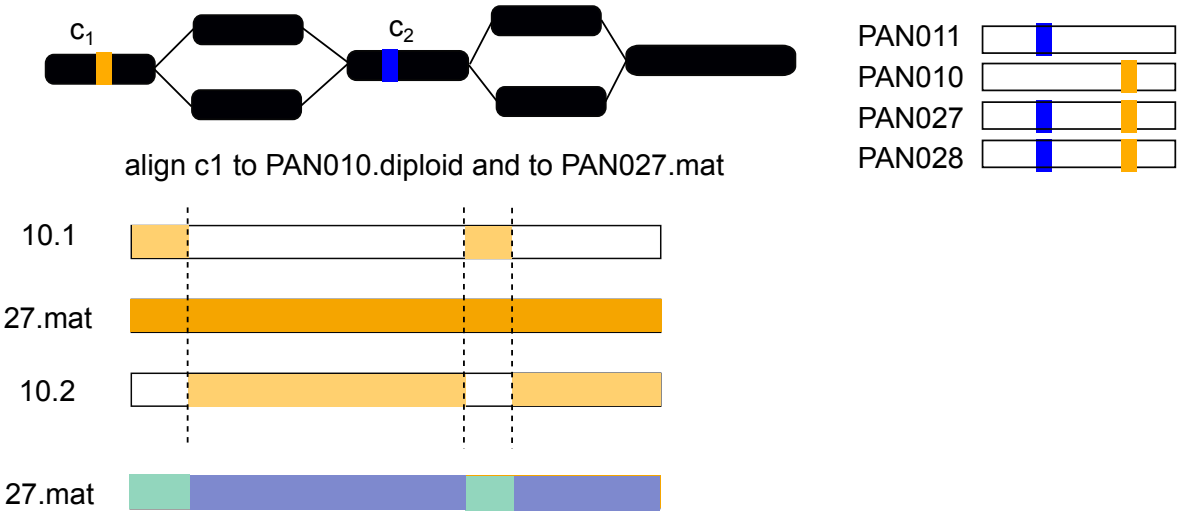


Figure 5.8: Alternative method of detecting recombination breakpoints. Haplotype-specific nodes are detected: c_1 is part of the maternal haplotype, c_2 of the paternal haplotype. c_1 is aligned to the assembly of the mother and to the maternal daughter haplotype. Switches in the two maternal haplotypes indicate recombination candidates.

mother's HiFi reads and the maternal haplotype accordingly.

To pinpoint the phase switch locations, the resulting callsets are phased with WhatsHap[150], additionally using the ONT reads.

The variant calling yielded 4,829,307 heterozygous variants in the paternal call set, of which 3,354,553 remained after subsequent filtering using `bcftools`[45] to keep only phased variants. The maternal call set included 4,721,957 heterozygous variants , with 3,316,209 phased. Further filtering was applied to the phased variant sets using `vcftools` [44], where less confident variant calls were removed based on quality (`--minQ 30`) and depth (`--max_meanDP 65`). The final call sets contained 3,004,396 heterozygous phased SNVs in the paternal case, and 2,960,084 in the maternal one.

From the phased VCF, potential recombination breakpoints were identified by inspecting phase changes within phased blocks, as output by WhatsHap. We flagged a variant site as a potential switch position if a phase change occurred within a phaseset, but excluded it if another switch appeared within 100 kb to prevent including 0|1|0 flip errors within a short distance in regions with a low number of variants.

For a more conservative approach, we apply a second method to detect recombination, based on the shared unitigs identified previously. The principle is shown in Figure 5.8. We align the nodes that are transmitted from one parent to the child to the diploid assembly of the corresponding parent and to the child's haplotype of said parent. Then, we scan the mapped regions to check which of the two haplotypes of the parent the nodes map to. Recombination breakpoints are given by switches between haplotype 1 and hap-

lotype 2 in the parent. Results of both methods are shown in Figure 5.9, where the variant calling-based approach yields more breakpoint candidates than the one based on the shared nodes, as expected. In the maternal haplotype, 46 candidates were found based on shared node mappings, 45 of which (97.83%) were also supported by the variant calling. In the paternal haplotype, 23 candidates were detected, all of which were support by the variant calling-based approach. The latter approach tends to overestimate breakpoint positions, as a breakpoint is inserted merely based on phase switches in the phasing. In the maternal haplotype, 32 additional candidate positions were estimated which lacked support by the shared nodes; in the paternal haplotype, 44 additional candidates were found. In both cases, these numbers are enriched by local 'clusters' where multiple consecutive variants within a short interval are selected as potential switch locations, meaning that there is no clear separation between two phase sets, but rather multiple consecutive short phase sets. One example are the five breakpoint candidates on maternal Chromosome 16 in the pericentromeric region, for example (Figure 5.9). Other estimates are clearly unreliable, as they are shortly followed by the end of a phase set, and thus cannot give any long-range phasing information. Examples are all candidates detected in paternal Chromosomes 16 or 22 (Figure 5.9).

Figure 5.7 b shows an example of a recombination position. Candidates computed by the two methods, in particular those that were detected using one method, but not the other, can be manually inspected via IGV [173] and potentially validated. Reads were first haplotagged by `WhatsHap haplotag`, using the read alignments and the phased call set. Colouring and grouping them by haplotype shows that before the switch position, all reads from one haplotype exhibit a lot of variation to the reference (which is the paternal haplotype assembly), while no variants are detected in the reads from the second haplotype. On the other side of the position in question, this is reversed.

For the final curation of the most likely set of breakpoint positions, we included all breakpoints consistent with the two methods. For the remaining variant-calling based candidates, we examined each position individually in IGV as described above. By examining the variation in the two haplotagged read sets around the location, four additional breakpoints in both haplotypes each could be added; an example for a recovered location on paternal chromosome 6 is shown in Supplementary Figure D.1 alongside another example of a candidate that could not be validated.

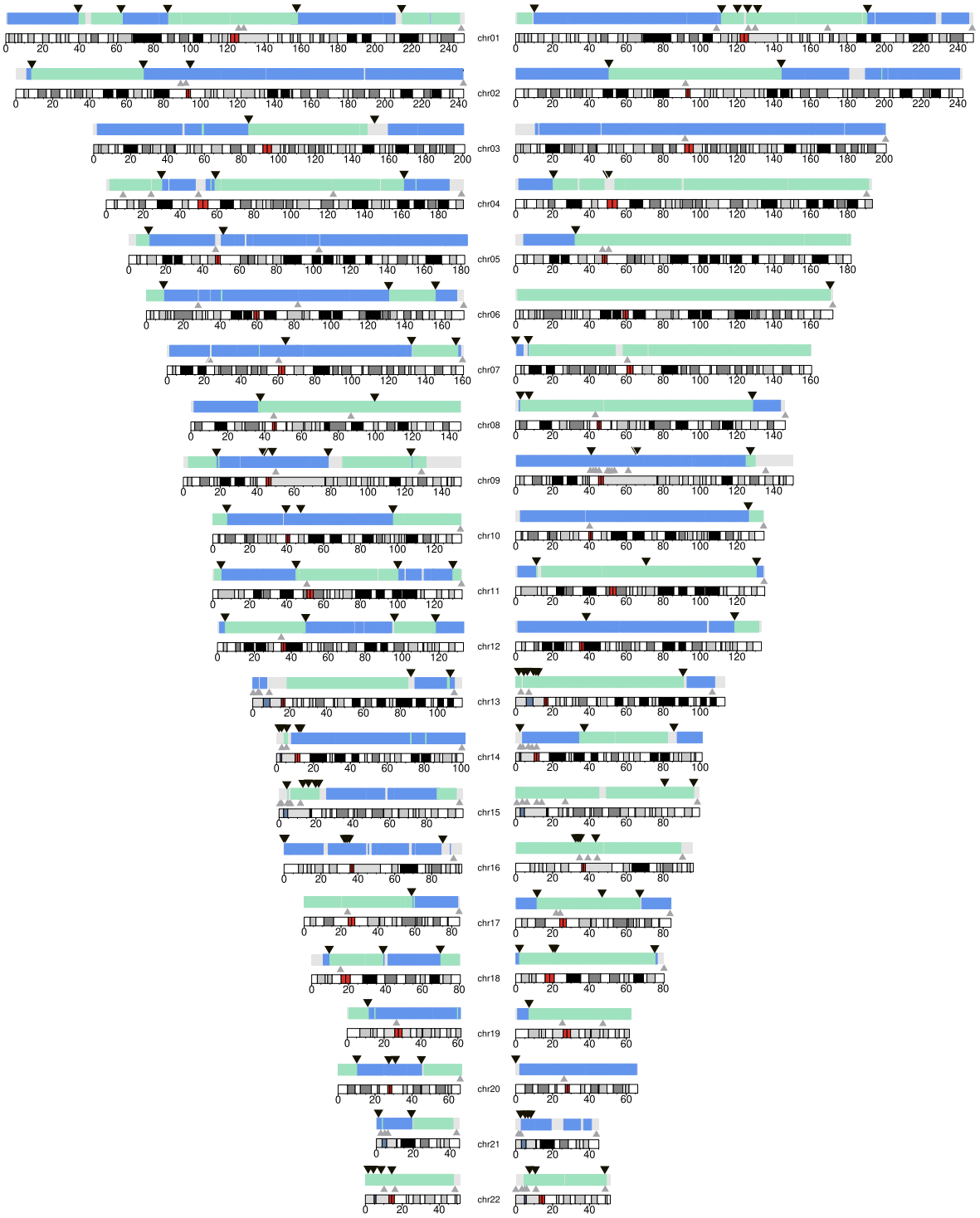


Figure 5.9: Meiotic recombination breakpoints of the PAN027 single-sample assembly. Breakpoints have been detected using alignments of the blocks shared across generations. Left: maternal haplotype, right: paternal haplotype. Colour change between blue and green regions denotes a phase switch. The black triangles denote breakpoint positions as estimated by the variant-based method. The grey triangles below indicate the end positions of a phase set as output by WhatsHap.

5.9 Discussion and conclusions

In this chapter, we introduced an novel approach for haplotype-resolved assembly using data from a three-generational pedigree. By pooling data from three members of a family trio (father, mother, and daughter) and leveraging Verkko, we constructed an assembly graph representing a pangenome. This method offers an alternative to the traditional approach of assembling genomes on an individual basis.

Using the pangenome graph and shared k-mers between the child and parent samples, we successfully assembled the child's genome to haplotype resolution. All autosomes were assembled, with a length comparison to CHM13 indicating that most chromosomes were assembled with high completeness. However, the acrocentric chromosomes posed significant challenges. These chromosomes became tangled within a single component of the assembly, making it difficult to fully resolve them. As a result, only the long arm of these chromosomes is predominantly present in the final assembly.

The assemblies demonstrated low fragmentation; several chromosomes were assembled as single haplotigs, and after scaffolding against CHM13, the final assemblies contained at most five haplotigs per chromosome. This method presents a promising alternative to individual assembly by directly utilising the genetic information shared within the pangenome graph. By tagging nodes in the graph based on their shared sequence content with the four pedigree members (including the granddaughter), we could identify unitigs present across multiple generations. These unitigs, transmitted either once or twice, highlight high-confidence regions in assemblies derived from single samples.

We used these unitigs to detect variants, such as SNVs and structural variations, by comparing them against the single-sample assemblies. This lays the groundwork for more detailed future analyses. In particular, further investigation into structural variants, especially the longer ones – such as the 156 kb deletion on chromosome 4 – would be valuable. On the other hand, we showed that large parts (> 70%) of the single assemblies can be validated by comparison to these preserved regions, taking advantage of the shared unitigs being identical across generations.

Future work might involve the analysis of de novo variants, as the multi-generation approach may help in distinguishing between germline and somatic variation. In a trio that involves only two generations, de novo variants in the child could either be somatic or the result of recent germline mutations in the parents. Adding a third generation enables the detection of these variants in the grandchild, where presence of a variant may be indicative of a germline mutation, and absence would more likely indicate a somatic variant.

Additionally, we explored potential meiotic recombination breakpoints, a crucial step towards generating haplotype-resolved assemblies from parental haplotypes. We applied two alternative methods, one based on variant calling from read alignments, the other based on the shared nodes. All of the breakpoints detected by the latter method – the more conservative one – were validated by the first method, suggesting these are confident candidates.

In summary, our results demonstrate that incorporating family data into a joint assembly graph for haplotype assembly can complement standard assembly methods. The single-sample assemblies used for comparison were of very high quality, also due to the high coverage across all sequencing data types. In cases of lower coverage, the pangenome approach may offer a more robust alternative, potentially mitigating the loss of information. Future research could explore this by systematically reducing coverage to determine the point at which standard assembly methods begin to fail, and whether the pooled approach offers a distinct advantage.

Conclusion

The past decade has seen remarkable advancements the field of genome and haplotype assembly. Not only have sequencing technologies continuously improved length and quality of sequencing reads, but also, substantial methodological progress has been achieved. As such, methods have emerged that successfully integrate multiple types of reads and allow for haplotype resolution during the assembly process. They are able to combine accurate reads with even the longest reads available to date, and with auxiliary data such as Strand-seq, Hi-C or trio data. Additionally, there has been a paradigm shift from linear to graph-based assembly, and from single reference sequences to pangenomes. Thanks to this progress, breakthroughs like the completion of the human genome and the first pangenome reference have been achieved.

The accessibility of high-quality haplotype-resolved assemblies allows for multiple downstream analyses. In Chapter 2, we presented a method for inferring local ancestry in phased assemblies. We applied a Hidden Markov model to a set of phased assemblies resulting from the PGAS pipeline developed as part of the HGSVC. The model utilized a reference panel consisting of samples from multiple populations from the 1000 Genomes Project. It output estimates of the most likely ancestral population at each variant position of the assembled samples. Chromosomal regions were colored by population in the resulting plots. We computed ancestries for each of the 64 assembled haplotypes and focused on a mother-father-child trio of Puerto Rican ancestry. Here, ancestry coloring coincided with known meiotic recombination breakpoints.

Challenges are the scalability to very large panels, as computation time depends quadratically on the number of samples in the panel. A bootstrapping approach allows for subsampling and parallelization. Additionally, reference-dependent approaches inherently fail to represent rare variants. As the trend goes towards pangenomics and graph representations instead of single linear sequences, future panel-based methods for local ancestry inference might switch to graph space as well. In that case, a pangenome reference would be used instead of the reference panel. This project demonstrated how accurately phased assemblies help not only detect new variation, but also allow for the

analysis of this variation in population-specific contexts.

While advancements in phasing have enabled routine haplotype assemblies of diploid species, the field of polyploid genomics is lagging behind. Polyploid genomes – with their high heterozygosity, repeat content and frequent structural variation – pose substantial challenges. Chapter 3 introduced `WhatsHap polyphase`, a polyploid phasing method based on read clustering and subsequent haplotype assembly by threading paths through the clusters. Resulting phased blocks had at least 3 times lower switch error rates compared to a competitor method. Our method yielded accurate results especially in regions where multiple haplotypes coincided, a phenomenon that is common in polyploid genomes. We obtained results on artificial tetraploid data as well as proof-of-concept results on the potato genes. Phasing of the full potato genome was not possible as the methodology at the time suffered from two main limitations: Insufficient quality of the potato reference sequence and error-prone long reads that had to be corrected in a separate step. These limitations are expected to become less relevant in the foreseeable future and `WhatsHap polyphase` has already proven useful for evaluating recent de novo assemblies.

As the potato genome is both variant-rich and highly heterogeneous, alignment-based methods in general are suboptimal. To bridge the gap of poor reference-quality potato assemblies, we built a de novo assembly, presented in Chapter 4. Enabled by the development of graph-based assembly methods and long-read sequencing, as well as inspired by advancements made for human genomes, we developed a kmer-based method that assembled the genome of the tetraploid potato cultivar Altus. The approach involved the construction of an assembly graph and subsequent haplotype-resolved assembly using nearly 200 offspring samples. Haplotypes were determined by detecting co-inherited loci through analysis of count patterns of loci-specific k-mers in the offspring. Assembled contigs reached up to the size of chromosome arms. However, with an average N50 value of 7.5, this assembly still did not reach the quality and completeness attained for human genomes and could not be considered 'telomere-to-telomere' yet.

One remaining challenge is the dependency on the existence of unique k-mers, which contradicts with the repetitive nature of the potato genome. Also, ultra-long reads, which have considerably improved assemblies of human genomes and, among others, enabled the telomere-to-telomere reference, are so far lacking. When these become available, we expect to resolve the remaining unphased regions on the genome graphs to also achieve complete polyploid references in the near future.

Similar to this k-mer based approach, Chapter 5 presented a method for assembly based on a pangenome approach on a threegenerational pedigree. We constructed an

assembly graph representing a mother-father-child trio. Shared sequence-specific k-mers between child and parents allowed resolving the child's haplotypes. Several chromosomes were assembled in a single contig. The representation of this trio in a pangenome graph allowed direct access to sequences shared across generations. This allowed for detecting meiotic recombination breakpoints and will further allow follow-up analyses of genomic tracts transmitted between multiple generations.

Taken together, the methods and results presented in this thesis contribute to advancing genome analysis at multiple levels: from applied analyses in phased assemblies, to improved phasing in polyploid genomes, to de novo assembly strategies in both human and plant contexts. In doing so, they also reflect the broader developments in the field – a shift from phasing as a separate step towards fully haplotype-resolved assemblies, enabled by advances in sequencing technology and new methodological frameworks. This shift is accompanied by a move from linear representations to graph-based, pangenomic frameworks. While human genomics has seen rapid progress, polyploid genomes have long remained an open challenge. This work contributes to closing that gap and to enabling haplotype-resolved assemblies in diploid, but also in complex polyploid genomes – offering methods that may support future developments as the field continues to evolve.

Bibliography

[1] Meryl. `https://github.com/marbl/meryl` [Accessed: (15.04.2025)].

[2] Repeatmasker. `http://www.repeatmasker.org` [Accessed: (15.04.2025)].

[3] S. Aganezov, S. M. Yan, D. C. Soto, M. Kirsche, S. Zarate, P. Avdeyev, D. J. Taylor, K. Shafin, A. Shumate, C. Xiao et al. A complete reference genome improves analysis of human genetic variation. *Science*, 376(6588):eabl3533, 2022.

[4] 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.

[5] D. Aguiar and S. Istrail. Haplotype assembly in polyploid genomes and identical by descent shared tracts. *Bioinformatics*, 29(13):i352–60, 2013.

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

[7] D. H. Alexander, J. Novembre, and K. Lange. Fast model-based estimation of ancestry in unrelated individuals. *Genome Research*, 19(9):1655–1664, 2009.

[8] T. Almutiri and F. Nadeem. Markov Models Applications in Natural Language Processing: A Survey. *International Journal of Information Technology and Computer Science*, 14:1–16, 2022.

[9] M. Alonge, L. Lebeigle, M. Kirsche, K. Jenike, S. Ou, S. Aganezov, X. Wang, Z. B. Lippman, M. C. Schatz, and S. Soyk. Automated assembly scaffolding using RagTag elevates a new tomato system for high-throughput genome editing. *Genome Biology*, 23(1):258, 2022.

[10] N. Altemose, G. A. Logsdon, A. V. Bzikadze, P. Sidhwani, S. A. Langley, G. V. Caldas, S. J. Hoyt, L. Uralsky, F. D. Ryabov, C. J. Shew et al. Complete genomic and epigenetic maps of human centromeres. *Science*, 376(6588):eabl4178, 2022.

[11] L. T. Amundadottir, P. Sulem, J. Gudmundsson, A. Helgason, A. Baker, B. A. Agnarsson, A. Sigurdsson, K. R. Benediktsdottir, J.-B. Cazier, J. Sainz et al. A common variant associated with prostate cancer in European and African populations. *Nature Genetics*, 38(6):652–658, 2006.

[12] An introduction to Next-Generation Sequencing Technology. https://www.illumina.com/content/dam/illumina-marketing/documents/products/illumina_sequencing_introduction.pdf [Accessed: (15.04.2025)].

[13] P. Audano. PAV: An assembly-based approach for discovering structural variants, indels, and point mutations in long-read phased genomes. In *ASHG Annual Meeting*, 2021-10-20.

[14] A. Auton, G. R. Abecasis, D. M. Altshuler, R. M. Durbin, D. R. Bentley, A. Chakravarti, A. G. Clark, P. Donnelly, E. E. Eichler, P. Flicek et al. A global reference for human genetic variation. *Nature*, 526(7571):68–74, 2015.

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

[16] Y. Bao, Q. Zhang, J. Huang, S. Zhang, W. Yao, Z. Yu, Z. Deng, J. Yu, W. Kong, X. Yu et al. A chromosomal-scale genome assembly of modern cultivated hybrid sugarcane provides insights into origination and evolution. *Nature Communications*, 15(1):3041, 2024.

[17] Z. Bao, C. Li, G. Li, P. Wang, Z. Peng, L. Cheng, H. Li, Z. Zhang, Y. Li, W. Huang et al. Genome architecture and tetrasomic inheritance of autotetraploid potato. *Molecular Plant*, 15(7):1211–1226, 2022.

[18] Y. Baran, B. Pasaniuc, S. Sankararaman, D. G. Torgerson, C. Gignoux, C. Eng, W. Rodriguez-Cintron, R. Chapela, J. G. Ford, P. C. Avila et al. Fast and accurate inference of local ancestry in Latino populations. *Bioinformatics*, 28(10):1359–1367, 2012.

[19] D. Barber. *Bayesian Reasoning and Machine Learning*. Cambridge University Press, Cambridge, 2012.

[20] 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.

[21] A. Bergström, S. A. McCarthy, R. Hui, M. A. Almarri, Q. Ayub, P. Danecek, Y. Chen, S. Felkel, P. Hallast, J. Kamm et al. Insights into human genetic variation and population history from 929 diverse genomes. *Science*, 367(6484):eaay5012, 2020.

[22] P. R. J. Birch, G. Bryan, B. Fenton, E. M. Gilroy, I. Hein, J. T. Jones, A. Prashar, M. A. Taylor, L. Torrance, and I. K. Toth. Crops that feed the world 8: Potato: are the trends of increased global production sustainable? *Food Security*, 4(4):477–508, 2012.

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

[24] 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. PMID: 27280382.

[25] J. P. Borst and J. R. Anderson. The discovery of processing stages: Analyzing eeg data with hidden semi-markov models. *NeuroImage*, 108:60–73, 2015.

[26] S. R. Browning, K. Grinde, A. Plantinga, S. M. Gogarten, A. M. Stilp, R. C. Kaplan, M. L. Avilés-Santa, B. L. Browning, and C. C. Laurie. Local ancestry inference in a large US-based Hispanic/Latino study: Hispanic community health study/study of latinos (HCHS/SOL). *G3 (Bethesda)*, 6(6):1525–1534, 2016.

[27] S. R. Browning, R. K. Waples, and B. L. Browning. Fast, accurate local ancestry inference with flare. *The American Journal of Human Genetics*, 110(2):326–335, 2023.

[28] T. Büchler, J. Olbrich, and E. Ohlebusch. Efficient short read mapping to a pangenome that is represented by a graph of ed strings. *Bioinformatics*, 39(5):btad320, 2023.

[29] E. Cavalet-Giorsa, A. González-Muñoz, N. Athiyannan, S. Holden, A. Salhi, C. Gardener, J. Quiroz-Chávez, S. M. Rustamova, A. F. Elkot, M. Patpour et al. Origin and evolution of the bread wheat D genome. *Nature*, 633(8031):848–855, 2024.

[30] L. L. Cavalli-Sforza. The human genome diversity project: past, present and future. *Nature Reviews Genetics*, 6(4):333–340, 2005.

[31] M. J. Chaisson, A. D. Sanders, X. Zhao, A. Malhotra, D. Porubsky, T. Rausch, E. J. Gardner, O. Rodriguez, L. Guo, R. L. Collins et al. Multi-platform discovery of haplotype-resolved structural variation in human genomes. *Nature Communications*, 10(1):1784, 2019.

[32] M. J. P. Chaisson, S. Mukherjee, S. Kannan, and E. E. Eichler. Resolving multicopy duplications de novo using polyploid phasing. *Research in Computational Molecular Biology*, 10229:117–133, 2017.

[33] 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):291, 2019.

[34] H. Cheng, M. Asri, J. Lucas, S. Koren, and H. Li. Scalable telomere-to-telomere assembly for diploid and polyploid genomes with double graph. *Nature Methods*, 21(6):967–970, 2024. Epub 2024 May 10.

[35] 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.

[36] H. Cheng, E. D. Jarvis, O. Fedrigo, K.-P. Koepfli, L. Urban, N. J. Gemmell, and H. Li. Haplotype-resolved assembly of diploid genomes without parental data. *Nature Biotechnology*, 40(9):1332–1335, 2022.

[37] R. Chikhi, A. Limasset, S. Jackman, J. T. Simpson, and P. Medvedev. On the Representation of De Bruijn Graphs. *Journal of Computational Biology*, 22(5):336–352, 2015. PMID: 25629448.

[38] R. Chikhi, A. Limasset, and P. Medvedev. Compacting de Bruijn graphs from sequencing data quickly and in low memory. *Bioinformatics*, 32(12):i201–i208, 2016.

[39] C.-S. Chin, P. Peluso, F. J. Sedlazeck, M. Nattestad, G. T. Concepcion, A. Clum, C. Dunn, R. O'Malley, R. Figueroa-Balderas, A. Morales-Cruz et al. Phased diploid genome assembly with single-molecule real-time sequencing. *Nature Methods*, 13(12):1050–1054, 2016.

[40] Chromatin conformation. `https://nanoporetech.com/applications/investigations/chromatin-conformation` [Accessed: (15.04.2025)].

[41] R. Cilibrasi, L. van Iersel, S. Kelk, and J. Tromp. On the complexity of several haplotyping problems. In R. Casadio and G. Myers, editors, *Algorithms in Bioinformatics*, pages 128–139, Berlin, Heidelberg, 2005. Springer Berlin Heidelberg.

[42] P. J. A. Cock, C. J. Fields, N. Goto, M. L. Heuer, and P. M. Rice. The sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Research*, 38(6):1767–1771, 2009.

[43] F. Dabbaghieh. gfa_subgraphs. `https://github.com/fawaz-dabbaghieh/gfa_subgraphs/tree/main` [Accessed: (15.04.2025)].

[44] P. Danecek, A. Auton, G. Abecasis, C. A. Albers, E. Banks, M. A. DePristo, R. E. Handsaker, G. Lunter, G. T. Marth, S. T. Sherry et al. The variant call format and VCFtools. *Bioinformatics*, 27(15):2156–2158, 2011.

[45] P. Danecek, J. K. Bonfield, J. Liddle, J. Marshall, V. Ohan, M. O. Pollard, A. Whitwham, T. Keane, S. A. McCarthy, R. M. Davies et al. Twelve years of SAMtools and BCFtools. *Gigascience*, 10(2), 2021.

[46] R. S. Daniels, R. Harvey, B. Ermetal, Z. Xiang, M. Galiano, L. Adams, and J. W. McCauley. A sanger sequencing protocol for sars-cov-2 s-gene. *Influenza and Other Respiratory Viruses*, 15(6):707–710, 2021.

[47] S. Das and H. Vikalo. SDhaP: haplotype assembly for diploids and polyploids via semi-definite programming. *BMC Genomics*, 16:260, 2015.

[48] D. Deamer, M. Akeson, and D. Branton. Three decades of nanopore sequencing. *Nature Biotechnology*, 34, 2016.

[49] J. Dekker, K. Rippe, M. Dekker, and N. Kleckner. Capturing chromosome conformation. *Science*, 295(5558):1306–1311, 2002.

[50] O. Delaneau, B. Howie, A. J. Cox, J.-F. Zagury, and J. Marchini. Haplotype Estimation Using Sequencing Reads. *The American Journal of Human Genetics*, 93(4):687–696, 2013.

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

[52] R. Della Coletta, Y. Qiu, S. Ou, M. B. Hufford, and C. N. Hirsch. How the pan-genome is changing crop genomics and improvement. *Genome Biology*, 22(1):3, 2021.

[53] A. S. Deshpande, N. Ulahannan, M. Pendleton, X. Dai, L. Ly, J. M. Behr, S. Schwenk, W. Liao, M. A. Augello, C. Tyer et al. Identifying synergistic high-order 3D chromatin conformations from genome-scale nanopore concatemer sequencing. *Nature Biotechnology*, 40:1488–1499, 2022.

[54] A. Devaux, P. Kromann, and O. Ortiz. Potatoes for sustainable global food security. *Potato Research*, 57(3-4):185–199, 2014.

[55] G. J. Dowrick. The chromosomes of Chrysanthemum. *Heredity*, 7(1):59–72, 1953.

[56] 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.

[57] 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.

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

[59] 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):1654–1660, 2017.

[60] J. Eid, A. Fehr, J. Gray, K. Luong, J. Lyle, G. Otto, P. Peluso, D. Rank, P. Baybayan, B. Bettman et al. Real-time DNA sequencing from single polymerase molecules. *Science*, 323(5910):133–138, 2009.

[61] 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(Volume 21, 2020):139–162, 2020.

[62] T. L. Elliott, A. M. Muasya, and P. Bureš. Complex patterns of ploidy in a holocentric plant clade (Schoenus, Cyperaceae) in the Cape biodiversity hotspot. *Annals of Botany*, 131(1):143–156, 2022.

[63] E. Espinosa, R. Bautista, I. Fernandez, R. Larrosa, E. L. Zapata, and O. Plata. Comparing assembly strategies for third-generation sequencing technologies across different genomes. *Genomics*, 115(5):110700, 2023.

[64] J. M. Flynn, R. Hubley, C. Goubert, J. Rosen, A. G. Clark, C. Feschotte, and A. F. Smit. RepeatModeler2 for automated genomic discovery of transposable element families. *Proceedings of the National Academy of Sciences*, 117(17):9451–9457, 2020.

[65] K. A. Frazer, S. S. Murray, N. J. Schork, and E. J. Topol. Human genetic variation and its contribution to complex traits. *Nature Reviews Genetics*, 10(4):241–251, 2009.

[66] M. L. Freedman, D. Reich, K. L. Penney, G. J. McDonald, A. A. Mignault, N. Patterson, S. B. Gabriel, E. J. Topol, J. W. Smoller, C. N. Pato et al. Assessing the impact of population stratification on genetic association studies. *Nature Genetics*, 36(4):388–393, 2004.

[67] R. T. Gaeta and J. Chris Pires. Homoeologous recombination in allopolyploids: the polyploid ratchet. *New Phytologist*, 186(1):18–28, 2010.

[68] Y. Gao, X. Yang, H. Chen, X. Tan, Z. Yang, L. Deng, B. Wang, S. Kong, S. Li, Y. Cui et al. A pangenome reference of 36 chinese populations. *Nature*, 619(7968):112–121, 2023.

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

[70] E. Garrison, A. Guarracino, S. Heumos, F. Villani, Z. Bao, L. Tattini, J. Hagmann, S. Vorbrugg, S. Marco-Sola, C. Kubica et al. Building pangenome graphs. *Nature Methods*, 21(11):2008–2012, 2024.

[71] 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.

[72] G. Genovese, D. J. Friedman, M. D. Ross, L. Lecordier, P. Uzureau, B. I. Freedman, D. W. Bowden, C. D. Langefeld, T. K. Oleksyk, A. L. U. Knob et al. Association of Trypanolytic ApoL1 Variants with Kidney Disease in African Americans. *Science*, 329(5993):841–845, 2010.

[73] A. M. Giani, G. R. Gallo, L. Gianfranceschi, and G. Formenti. Long walk to genomics: History and current approaches to genome sequencing and assembly. *Computational and Structural Biotechnology Journal*, 18:9–19, 2020.

[74] N. Gill, S. Findley, J. G. Walling, C. Hans, J. Ma, J. Doyle, G. Stacey, and S. A. Jackson. Molecular and Chromosomal Evidence for Allopolyploidy in Soybean . *Plant Physiology*, 151(3):1167–1174, 2009.

[75] M. Goel and K. Schneeberger. plotsr: visualizing structural similarities and rearrangements between multiple genomes. *Bioinformatics*, 38(10):2922–2926, 2022.

[76] M. Goel, H. Sun, W.-B. Jiao, and K. Schneeberger. SyRI: finding genomic rearrangements and local sequence differences from whole-genome assemblies. *Genome Biology*, 20(1):277, 2019.

[77] C. Groza, C. Schwendinger-Schreck, W. A. Cheung, E. G. Farrow, I. Thiffault, J. Lake, W. B. Rizzo, G. Evrony, T. Curran, G. Bourque et al. Pangenome graphs improve the analysis of structural variants in rare genetic diseases. *Nature Communications*, 15(1):657, 2024.

[78] A. Guarracino, S. Buonaiuto, L. G. de Lima, T. Potapova, A. Rhie, S. Koren, B. Rubinstein, C. Fischer, H. J. Abel, L. L. Antonacci-Fulton et al. Recombination between heterologous human acrocentric chromosomes. *Nature*, 617(7960):335–343, 2023.

[79] P. Hallast, P. Ebert, M. Loftus, F. Yilmaz, P. A. Audano, G. A. Logsdon, M. J. Bonder, W. Zhou, W. Höps, K. Kim et al. Assembly of 43 human y chromosomes reveals extensive complexity and variation. *Nature*, 621(7978):355–364, 2023.

[80] M.-R. Han, W. Zheng, Q. Cai, Y.-T. Gao, Y. Zheng, M. K. Bolla, K. Michailidou, J. Dennis, Q. Wang, A. M. Dunning et al. Evaluating genetic variants associated with breast cancer risk in high and moderate-penetrance genes in Asians. *Carcinogenesis*, 38(5):511–518, 2017.

[81] 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 et al. 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.

[82] D. He and E. Eskin. Hap-seqx: expedite algorithm for haplotype phasing with imputation using sequence data. *Gene*, 518(1):2–6, 2013.

[83] D. He, S. Saha, R. Finkers, and L. Parida. Efficient algorithms for polyploid haplotype phasing. *BMC Genomics*, 19(Suppl 2):110, 2018.

[84] A. L. Healey, O. Garsmeur, J. T. Lovell, S. Shengquiang, A. Sreedasyam, J. Jenkins, C. B. Plott, N. Piperidis, N. Pompidor, V. Llaca et al. The complex polyploid genome architecture of sugarcane. *Nature*, 628(8009):804–810, 2024.

[85] A. Helena Mangs and B. J. Morris. The human pseudoautosomal region (PAR): Origin, function and future. *Curr Genomics*, 8(2):129–136, 2007.

[86] A. Helgadottir, G. Thorleifsson, K. P. Magnusson, S. Grétarsdottir, V. Steinthorsdottir, A. Manolescu, G. T. Jones, G. J. E. Rinkel, J. D. Blankensteijn, A. Ronkainen et al. The same sequence variant on 9p21 associates with myocardial infarction, abdominal aortic aneurysm and intracranial aneurysm. *Nature Genetics*, 40(2):217–224, 2008.

[87] M. Henglin, M. Ghareghani, W. T. Harvey, D. Porubsky, S. Koren, E. E. Eichler, P. Ebert, and T. Marschall. Graphasing: phasing diploid genome assembly graphs with single-cell strand sequencing. *Genome Biology*, 25(1):265, 2024.

[88] V. Herklotz and C. M. Ritz. Multiple and asymmetrical origin of polyploid dog rose hybrids (Rosa L. sect. Caninae (DC.) Ser.) involving unreduced gametes. *Annals of Botany*, 120(2):209–220, 2016.

[89] J. S. P. Heslop-Harrison, T. Schwarzacher, and Q. Liu. Polyploidy: its consequences and enabling role in plant diversification and evolution. *Annals of Botany*, 131(1):1–10, 2022.

[90] 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):35, 2020.

[91] G. Hickey, J. Monlong, J. Ebler, A. M. Novak, J. M. Eizenga, Y. Gao, H. J. Abel, L. L. Antonacci-Fulton, M. Asri, G. Baid et al. Pangenome graph construction from genome alignments with Minigraph-Cactus. *Nature Biotechnology*, 42(4):663–673, 2024.

[92] A. M. Hinchie, S. L. Sanford, K. E. Loughridge, R. M. Sutton, A. H. Parikh, A. A. Gil Silva, D. I. Sullivan, P. Chun-On, M. R. Morrell, J. F. McDyer et al. A persistent variant telomere sequence in a human pedigree. *Nature Communications*, 15(1):4681, 2024.

[93] G. Hoopes, X. Meng, J. P. Hamilton, S. R. Achakkagari, F. de Alves Freitas Guesdes, M. E. Bolger, J. J. Coombs, D. Esselink, N. R. Kaiser, L. Kodde et al. Phased, chromosome-scale genome assemblies of tetraploid potato reveal a complex genome, transcriptome, and predicted proteome landscape underpinning genetic diversity. *Molecular Plant*, 15(3):520–536, 2022.

[94] A. Huerta-Chagoya, P. Schroeder, R. Mandla, J. Li, L. Morris, M. Vora, A. Alkanaq, D. Nagy, L. Szczerbinski, J. G. S. Madsen et al. Rare variant analyses in 51,256 type 2 diabetes cases and 370,487 controls reveal the pathogenicity spectrum of monogenic diabetes genes. *Nature Genetics*, 56(11):2370–2379, 2024.

[95] O. Häggström. *Finite Markov Chains and Algorithmic Applications*, volume 52 of *London Mathematical Society Student Texts*. Cambridge University Press, Cambridge, 2002.

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

[97] International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature*, 431(7011):931–945, 2004.

[98] International Wheat Genome Sequencing Consortium (IWGSC). Shifting the limits in wheat research and breeding using a fully annotated reference genome. *Science*, 361(6403), 2018.

[99] M. Jain, H. E. Olsen, B. Paten, and M. Akeson. The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. *Genome Biology*, 17, 2016.

[100] S. Kaur, A. Ali, U. Ahmad, Y. Siahbalaei, A. K. Pandey, and B. Singh. Role of single nucleotide polymorphisms (SNPs) in common migraine. *The Egyptian Journal of Neurology, Psychiatry and Neurosurgery*, 55(1):47, 2019.

[101] E. Kebede. Contribution, utilization, and improvement of legumes-driven biological nitrogen fixation in agricultural systems. *Frontiers in Sustainable Food Systems*, 5, 2021.

[102] H. Kihara and T. Ono. Chromosomenzahlen und systematische Gruppierung der Rumex-Arten. *Zeitschrift für Zellforschung und Mikroskopische Anatomie*, 4(3):475–481, 1926.

[103] 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.

[104] G. W. Klau and T. Marschall. A guided tour to computational haplotyping. In *Unveiling Dynamics and Complexity*, volume 10307 of *Lecture Notes in Computer Science*, pages 50–63, Cham, 2017. Springer.

[105] W. C. Knowler, R. C. Williams, D. J. Pettitt, and A. G. Steinberg. Gm3;5,13,14 and type 2 diabetes mellitus: an association in American Indians with genetic admixture. *The American Journal of Human Genetics*, 43(4):520–526, 1988.

[106] M. Kolmogorov, J. Yuan, Y. Lin, and P. A. Pevzner. Assembly of long, error-prone reads using repeat graphs. *Nature Biotechnology*, 37(5):540–546, 2019.

[107] S. Koren, A. Rhie, B. P. Walenz, A. T. Dilthey, D. M. Bickhart, S. B. Kingan, S. Hiendleder, J. L. Williams, T. P. L. Smith, and A. M. Phillippy. De novo assembly of haplotype-resolved genomes with trio binning. *Nature Biotechnology*, 2018.

[108] S. Koren, B. P. Walenz, K. Berlin, J. R. Miller, N. H. Bergman, and A. M. Phillippy. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Research*, 27(5):722–736, 2017.

[109] M. Kyriakidou, H. H. Tai, N. L. Anglin, D. Ellis, and M. V. Strömvik. Current strategies of polyploid plant genome sequence assembly. *Frontiers in Plant Science*, 9:1660, 2018.

[110] M. J. Landrum, J. M. Lee, M. Benson, G. R. Brown, C. Chao, S. Chitipiralla, B. Gu, J. Hart, D. Hoffman, W. Jang et al. Clinvar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Research*, 46(D1):D1062–D1067, 2018.

[111] F. Li, G. Fan, K. Wang, F. Sun, Y. Yuan, G. Song, Q. Li, Z. Ma, C. Lu, C. Zou et al. Genome sequence of the cultivated cotton gossypium arboreum. *Nature Genetics*, 46(6):567–572, 2014.

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

[113] H. Li and R. Durbin. Genome assembly in the telomere-to-telomere era. *Nature Reviews Genetics*, 25(9):658–670, 2024.

[114] H. Li, X. Feng, and C. Chu. The design and construction of reference pangenome graphs with minigraph. *Genome Biology*, 21(1):265, 2020.

[115] 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.

[116] K.-T. Li, M. Moulin, N. Mangel, M. Albersen, N. M. Verhoeven-Duif, Q. Ma, P. Zhang, T. B. Fitzpatrick, W. Gruissem, and H. Vanderschuren. Increased bioavailable vitamin B6 in field-grown transgenic cassava for dietary sufficiency. *Nature Biotechnology*, 33:1029–1032, 2015.

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

[118] R. Li, Y. Li, H. Zheng, R. Luo, H. Zhu, Q. Li, W. Qian, Y. Ren, G. Tian, J. Li et al. Building the sequence map of the human pan-genome. *Nature Biotechnology*, 28(1):57–63, 2010.

[119] 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. *Nature*, 617(7960):312–324, 2023.

[120] E. Lieberman-Aiden, N. L. van Berkum, L. Williams, M. Imakaev, T. Ragoczy, A. Telling, I. Amit, B. R. Lajoie, P. J. Sabo, M. Dorschner et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, 326:289–293, 2009.

[121] R. Lippert, R. Schwartz, G. Lancia, and S. Istrail. Algorithmic strategies for the single nucleotide polymorphism haplotype assembly problem. *Briefings in Bioinformatics*, 3(1):23–31, 2002.

[122] G. A. Logsdon, A. N. Rozanski, F. Ryabov, T. Potapova, V. A. Shepelev, C. R. Catacchio, D. Porubsky, Y. Mao, D. Yoo, M. Rautiainen et al. The variation and evolution of complete human centromeres. *Nature*, 629(8010):136–145, 2024.

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

[124] 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.

[125] 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.

[126] P.-R. Loh, P. F. Palamara, and A. L. Price. Fast and accurate long-range phasing in a UK Biobank cohort. *Nature Genetics*, 48(7):811–816, 2016.

[127] K. D. Makova, B. D. Pickett, R. S. Harris, G. A. Hartley, M. Cechova, K. Pal, S. Nurk, D. Yoo, Q. Li, P. Hebbar et al. The complete sequence and comparative analysis of ape sex chromosomes. *Nature*, 630(8016):401–411, 2024.

[128] S. Mallick, H. Li, M. Lipson, I. Mathieson, M. Gymrek, F. Racimo, M. Zhao, N. Chennagiri, S. Nordenfelt, A. Tandon et al. The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature*, 538(7624):201–206, 2016.

[129] M. Manni, M. R. Berkeley, M. Seppey, F. A. Simão, and E. M. Zdobnov. BUSCO update: Novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. *Molecular Biology and Evolution*, 38(10):4647–4654, 2021.

[130] B. K. Maples, S. Gravel, E. E. Kenny, and C. D. Bustamante. RFMix: a discriminative modeling approach for rapid and robust local-ancestry inference. *The American Journal of Human Genetics*, 93(2):278–288, 2013.

[131] 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.

[132] A. R. Martin, C. R. Gignoux, R. K. Walters, G. L. Wojcik, B. M. Neale, S. Gravel, M. J. Daly, C. D. Bustamante, and E. E. Kenny. Human Demographic History Impacts Genetic Risk Prediction across Diverse Populations. *The American Journal of Human Genetics*, 100(4):635–649, 2017.

[133] M. Martin, M. Patterson, S. Garg, S. O Fischer, N. Pisanti, G. W. Klau, A. Schönhuth, and T. Marschall. WhatsHap: fast and accurate read-based phasing. *bioRxiv*, 2016.

[134] A. S. Mason and J. C. Pires. Unreduced gametes: meiotic mishap or evolutionary mechanism? *Trends in Genetics*, 31(1):5–10, 2015.

[135] A. S. Mason and J. F. Wendel. Homoeologous Exchanges, Segmental Allopolyploidy, and Polyploid Genome Evolution. *Frontiers in Genetics*, 11, 2020.

[136] W. R. McCombie, J. D. McPherson, and E. R. Mardis. Next-Generation Sequencing Technologies. *Cold Spring Harbor Perspectives in Medicine*, 2019.

[137] L. A. Meyers and D. A. Levin. ON THE ABUNDANCE OF POLYPLOIDS IN FLOWERING PLANTS. *Evolution*, 60(6):1198–1206, 2006.

[138] J. Meyne, R. L. Ratliff, and R. K. Moyzis. Conservation of the human telomere sequence (ttaggg)n among vertebrates. *Proceedings of the National Academy of Sciences*, 86(18):7049–7053, 1989.

[139] S. Moore, J. McGowan-Jordan, A. C. Smith, K. Rack, U. Koehler, M. Stevens-Kroef, H. Barseghyan, R. Kanagal-Shamanna, R. Hastings, and O. behalf of the ISCN Standing Committee. Genome mapping nomenclature. *Cytogenetic and Genome Research*, 163(5-6):236–246, 2024.

[140] A. Moreno-Estrada, S. Gravel, F. Zakharia, J. L. McCauley, J. K. Byrnes, C. R. Gignoux, P. A. Ortiz-Tello, R. J. Martínez, D. J. Hedges, R. W. Morris et al. Reconstructing the Population Genetic History of the Caribbean. *PLOS Genetics*, 9(11):1–19, 2013.

[141] 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.

[142] 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, 2017.

[143] G. Myers. FastK: A K-mer counter (for HQ assembly data sets). `https://github.com/thegenemyers/FASTK` [Accessed: (15.04.2025)].

[144] V. Mäkinen, B. Cazaux, M. Equi, T. Norri, and A. I. Tomescu. Linear time construction of indexable founder block graphs, 2020.

[145] N. Nassir, M. A. Almarri, M. Kumail, N. Mohamed, B. Balan, S. Hanif, M. AlObathani, B. Jamalalail, H. Elsokary, D. Kondaramage et al. A draft arab pangenome reference. *bioRxiv*, 2024.

[146] 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.

[147] S. Nurk, B. P. Walenz, A. Rhie, M. R. Vollger, G. A. Logsdon, R. Grothe, K. H. Miga, E. E. Eichler, A. M. Phillippy, and S. Koren. HiCanu: accurate assembly of segmental duplications, satellites, and allelic variants from high-fidelity long reads. *Genome Research*, 30(9):1291–1305, 2020.

[148] Y. Ono, K. Asai, and M. Hamada. PBSIM: PacBio reads simulator—toward accurate genome assembly. *Bioinformatics*, 29(1):119–121, 2012.

[149] B. Paten, A. M. Novak, J. M. Eizenga, and E. Garrison. Genome graphs and the evolution of genome inference. *Genome Research*, 27(5):665–676, 2017. © 2017 Paten et al.; Published by Cold Spring Harbor Laboratory Press.

[150] 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.

[151] W. R. Pearson and D. J. Lipman. Improved tools for biological sequence comparison. *Proceedings of the National Academy of Sciences*, 85(8):2444–2448, 1988.

[152] A. Pecinka, W. Fang, M. Rehmsmeier, A. A. Levy, and O. M. Scheid. Polyploidization increases meiotic recombination frequency in Arabidopsis. *BMC Biology*, 9(1):24, 2011.

[153] M. Petek, M. Zagorščak, Ž. Ramšak, S. Sanders, Š. Tomaž, E. Tseng, M. Zouine, A. Coll, and K. Gruden. Cultivar-specific transcriptome and pan-transcriptome reconstruction of tetraploid potato. *Scientific Data*, 7(1):249, 2020.

[154] G. M. Pham, J. P. Hamilton, J. C. Wood, J. T. Burke, H. Zhao, B. Vaillancourt, S. Ou, J. Jiang, and C. R. Buell. Construction of a chromosome-scale long-read reference genome assembly for potato. *Gigascience*, 9(9), 2020.

[155] A. Piovesan, M. C. Pelleri, F. Antonaros, P. Strippoli, M. Caracausi, and L. Vitale. On the length, weight and GC content of the human genome. *BMC Research Notes*, 12(1):106, 2019.

[156] V. Pomiès, N. Turnbull, S. Le Squin, I. Syahputra, E. Suryana, T. Durand-Gasselin, B. Cochard, and F. Bakry. Occurrence of triploids in oil palm and their origin. *Annals of Botany*, 131(1):17–32, 2022.

[157] 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.

[158] 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.

[159] D. Porubsky, W. T. Harvey, A. N. Rozanski, J. Ebler, W. Höps, H. Ashraf, P. Hasenfeld, B. Paten, A. D. Sanders, T. Marschall et al. Inversion polymorphism in a complete human genome assembly. *Genome Biology*, 24(1):100, 2023.

[160] D. Porubsky, W. Höps, H. Ashraf, P. Hsieh, B. Rodriguez-Martin, F. Yilmaz, J. Ebler, P. Hallast, F. A. Maria Maggiolini, W. T. Harvey et al. Recurrent inversion polymorphisms in humans associate with genetic instability and genomic disorders. *Cell*, 185(11):1986–2005.e26, 2022.

[161] A. L. Price, A. Tandon, N. Patterson, K. C. Barnes, N. Rafaels, I. Ruczinski, T. H. Beaty, R. Mathias, D. Reich, and S. Myers. Sensitive detection of chromosomal segments of distinct ancestry in admixed populations. *PLOS Genetics*, 5(6):1–18, 2009.

[162] B. Pucker. Plant DNA extraction and preparation for ONT sequencing, 2020.

[163] P. Qin, H. Lu, H. Du, H. Wang, W. Chen, Z. Chen, Q. He, S. Ou, H. Zhang, X. Li et al. Pangenome analysis of 33 genetically diverse rice accessions reveals hidden genomic variations. *Cell*, 184(13):3542–3558.e16, 2021.

[164] S. Quigley, J. Damas, D. M. Larkin, and M. Farré. syntenyPlotteR: a user-friendly R package to visualize genome synteny, ideal for both experienced and novice bioinformaticians. *Bioinformatics Advances*.

[165] L. Rabiner and B. Juang. An introduction to hidden markov models. *IEEE ASSP Magazine*, 3(1):4–16, 1986.

[166] M. Rautiainen. Ribotin: automated assembly and phasing of rdna morphs. *Bioinformatics*, 40(3):btae124, 2024.

[167] M. Rautiainen and T. Marschall. GraphAligner: rapid and versatile sequence-to-graph alignment. *Genome Biology*, 21(1):253, 2020.

[168] M. Rautiainen, V. Mäkinen, and T. Marschall. Bit-parallel sequence-to-graph alignment. *Bioinformatics*, 35(19):3599–3607, 2019.

[169] M. Rautiainen, S. Nurk, B. P. Walenz, G. A. Logsdon, D. Porubsky, A. Rhie, E. E. Eichler, A. M. Phillippy, and S. Koren. Telomere-to-telomere assembly of diploid chromosomes with verkko. *Nature Biotechnology*, 41(10):1474–1482, 2023.

[170] O. Raymond, J. Gouzy, J. Just, H. Badouin, M. Verdenaud, A. Lemainque, P. Vergne, S. Moja, N. Choisne, C. Pont et al. The Rosa genome provides new insights into the domestication of modern roses. *Nature Genetics*, 50(6):772–777, 2018.

[171] A. Rhie, S. Nurk, M. Cechova, S. J. Hoyt, D. J. Taylor, N. Altemose, P. W. Hook, S. Koren, M. Rautiainen, I. A. Alexandrov et al. The complete sequence of a human Y chromosome. *Nature*, 621(7978):344–354, 2023.

[172] A. Rhie, B. P. Walenz, S. Koren, and A. M. Phillippy. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biology*, 21(1):245, 2020.

[173] 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.

[174] M. Rossi, M. Oliva, B. Langmead, T. Gagie, and C. Boucher. Moni: A pangenomic index for finding maximal exact matches. *Journal of Computational Biology*, 29(2):169–187, 2022. PMID: 35041495.

[175] J. Ruan and H. Li. Fast and accurate long-read assembly with wtdbg2. *Nature Methods*, 17(2):155–158, 2020.

[176] A. Salman-Minkov, N. Sabath, and I. Mayrose. Whole-genome duplication as a key factor in crop domestication. *Nature Plants*, 2(8):16115, 2016.

[177] A. D. Sanders, E. Falconer, M. Hills, D. C. J. Spierings, and P. M. Lansdorp. Single-cell template strand sequencing by Strand-seq enables the characterization of individual homologs. *Nature Protocols*, 12(6):1151–1176, 2017.

[178] 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.

[179] E. E. Schadt, S. Turner, and A. Kasarskis. A window into third-generation sequencing. *Human Molecular Genetics*, 19(R2):R227–R240, 2010.

[180] S.-V. Schiessl, E. Katche, E. Ihien, H. S. Chawla, and A. S. Mason. The role of genomic structural variation in the genetic improvement of polyploid crops. *The Crop Journal*, 7(2):127–140, 2019.

[181] J. Schmutz, S. B. Cannon, J. Schlueter, J. Ma, T. Mitros, W. Nelson, D. L. Hyten, Q. Song, J. J. Thelen, J. Cheng et al. Genome sequence of the palaeopolyploid soybean. *Nature*, 463(7278):178–183, 2010.

[182] V. A. Schneider, T. Graves-Lindsay, K. Howe, N. Bouk, H.-C. Chen, P. A. Kitts, T. D. Murphy, K. D. Pruitt, F. Thibaud-Nissen, D. Albracht et al. Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. *Genome Research*, 27(5):849–864, 2017.

[183] S. Schrinner, R. Serra Mari, R. Finkers, P. Arens, B. Usadel, T. Marschall, and G. W. Klau. Genetic polyploid phasing from low-depth progeny samples. *iScience*, 25(6):104461, 2022.

[184] 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):252, 2020.

[185] R. Schwartz. Theory and algorithms for the haplotype assembly problem. *Communications in Information and Systems*, 10(1):23–38, 2010.

[186] S. N. Seclen, M. E. Rosas, A. J. Arias, and C. A. Medina. Elevated incidence rates of diabetes in Peru: report from PERUDIAB, a national urban population-based longitudinal study. *BMJ Open Diabetes Research & Care*, 5(1):e000401, 2017.

[187] R. Secolin, A. Mas-Sandoval, L. R. Arauna, F. R. Torres, T. K. de Araujo, M. L. Santos, C. S. Rocha, B. S. Carvalho, F. Cendes, I. Lopes-Cendes et al. Distribution of local ancestry and evidence of adaptation in admixed populations. *Scientific Reports*, 9(1):13900, 2019.

[188] Sequencing 101: from DNA to discovery — the steps of SMRT sequencing. `https://www.pacb.com/blog/steps-of-smrt-sequencing/` [Accessed: (15.04.2025)].

[189] R. Serra Mari, S. Schrinner, R. Finkers, F. M. R. Ziegler, 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. *Genome Biology*, 25(1):26, 2024.

[190] R. M. Sherman, J. Forman, V. Antonescu, D. Puiu, M. Daya, N. Rafaels, M. P. Boorgula, S. Chavan, C. Vergara, V. E. Ortega et al. Assembly of a pan-genome from deep sequencing of 910 humans of african descent. *Nature Genetics*, 51(1):30–35, 2019.

[191] C. Siadjeu, B. Pucker, P. Viehöver, D. C. Albach, and B. Weisshaar. High contiguity de novo genome sequence assembly of trifoliate yam (dioscorea dumetorum) using long read sequencing. *Genes*, 11(3), 2020.

[192] N. Sierro, M. Auberson, R. Dulize, and N. V. Ivanov. Chromosome-level genome assemblies of Nicotiana tabacum, Nicotiana sylvestris, and Nicotiana tomentosiformis. *Scientific Data*, 11(1):135, 2024.

[193] I. R. Sipos, A. Ceffer, and J. Levendovszky. Parallel Optimization of Sparse Portfolios with AR-HMMs. *Computational Economics*, 49(4):563–578, 2017.

[194] E. Siragusa, N. Haiminen, R. Finkers, R. Visser, and L. Parida. Haplotype assembly of autotetraploid potato using integer linear programing. *Bioinformatics*, 35(21):4534, 2019.

[195] J. Sirén, P. Eskandar, M. T. Ungaro, G. Hickey, J. M. Eizenga, A. M. Novak, X. Chang, P.-C. Chang, M. Kolmogorov, A. Carroll et al. Personalized pangenome references. *Nature Methods*, 21(11):2017–2023, 2024.

[196] 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.

[197] 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.

[198] S. A. Soleimanpour, A. Gupta, M. Bakay, A. M. Ferrari, D. N. Groff, J. Fadista, L. A. Spruce, J. A. Kushner, L. Groop, S. H. Seeholzer et al. The diabetes susceptibility gene Clec16a regulates mitophagy. *Cell*, 157(7):1577–1590, 2014.

[199] 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.

[200] 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*, 2022.

[201] T. Tan and E. G. Atkinson. Strategies for the genomic analysis of admixed populations. *Annual Review of Biomedical Data Science*, 6(Volume 6, 2023):105–127, 2023.

[202] D. Tang, Y. Jia, J. Zhang, H. Li, L. Cheng, P. Wang, Z. Bao, Z. Liu, S. Feng, X. Zhu et al. Genome evolution and diversity of wild and cultivated potatoes. *Nature*, 606(7914):535–541, 2022.

[203] D. J. Taylor, J. M. Eizenga, Q. Li, A. Das, K. M. Jenike, E. E. Kenny, K. H. Miga, J. Monlong, R. C. McCoy, B. Paten et al. Beyond the human genome project: The age of complete human genome sequences and pangenome references. *Annual Review of Genomics and Human Genetics*, 25(Volume 25, 2024):77–104, 2024.

[204] 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.

[205] The 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491(7422):56–65, 2012.

[206] The Computational Pan-Genomics Consortium . Computational pan-genomics: status, promises and challenges. *Briefings in Bioinformatics*, 19(1):118–135, 2016.

[207] The GFA Format Specification. `(http://gfa-spec.github.io/GFA-spec/GFA1.html)` [Accessed: 15.04.2024].

[208] The International Wheat Genome Sequencing Consortium (IWGSC). A chromosome-based draft sequence of the hexaploid bread wheat (*Triticum aestivum*) genome. *Science*, 345(6194):1251788, 2014.

[209] The International Wheat Genome Sequencing Consortium (IWGSC). Shifting the limits in wheat research and breeding using a fully annotated reference genome. *Science*, 361(6403):eaar7191, 2018.

[210] The Potato Genome Sequencing Consortium. Genome sequence and analysis of the tuber crop potato. *Nature*, 475(7355):189–195, 2011.

[211] E. Uffelmann, Q. Q. Huang, N. S. Munung, J. de Vries, Y. Okada, A. R. Martin, H. C. Martin, T. Lappalainen, and D. Posthuma. Genome-wide association studies. *Nature Reviews Methods Primers*, 1(1):59, 2021.

[212] N. van Lieshout, A. van der Burgt, M. E. de Vries, M. Ter Maat, D. Eickholt, D. Esselink, M. P. W. van Kaauwen, L. P. Kodde, R. G. F. Visser, P. Lindhout et al. Solyntus, the New Highly Contiguous Reference Genome for Potato (*Solanum Tuberosum*). *G3*, 10(10):3489–3495, 2020.

[213] 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. Erratum in: Science 2001 Jun 5;292(5523):1838.

[214] R. G. F. Visser, C. W. B. Bachem, T. Borm, J. de Boer, H. J. van Eck, R. Finkers, G. van der Linden, C. A. Maliepaard, J G A M, R. Voorrips et al. Possibilities and challenges of the potato genome sequence. *Potato Research*, 57(3-4):327–330, 2014.

[215] M. R. Vollger, P. C. Dishuck, W. T. Harvey, W. S. DeWitt, X. Guitart, M. E. Goldberg, A. N. Rozanski, J. Lucas, M. Asri, H. J. Abel et al. Increased mutation and gene conversion within human segmental duplications. *Nature*, 617(7960):325–334, 2023.

[216] S. Walkowiak, L. Gao, C. Monat, G. Haberer, M. T. Kassa, J. Brinton, R. H. Ramirez-Gonzalez, M. C. Kolodziej, E. Delorean, D. Thambugala et al. Multiple wheat genomes reveal global variation in modern breeding. *Nature*, 588(7837):277–283, 2020.

[217] G. Wang, N. Zhou, Q. Chen, Y. Yang, Y. Yang, and Y. Duan. Gradual genome size evolution and polyploidy in Allium from the Qinghai–Tibetan Plateau. *Annals of Botany*, 131(1):109–122, 2021.

[218] T. Wang, L. Antonacci-Fulton, K. Howe, H. A. Lawson, J. K. Lucas, A. M. Phillippy, A. B. Popejoy, M. Asri, C. Carson, M. J. P. Chaisson et al. The human pangenome project: a global resource to map genomic diversity. *Nature*, 604(7906):437–446, 2022.

[219] Y. Wei, D. Zhi, and S. Zhang. Fast and accurate local ancestry inference with Recomb-Mix. *bioRxiv*, 2024.

[220] 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.

[221] Flow cells and nanopores. `https://nanoporetech.com/platform/technology/flow-cells-and-nanopores` [Accessed: (15.04.25)].

[222] R. R. Wick, M. B. Schultz, J. Zobel, and K. E. Holt. Bandage: interactive visualization of de novo genome assemblies. *Bioinformatics*, 31(20):3350–3352, 2015.

[223] J.-H. Wu, A. R. Ferguson, B. G. Murray, Y. Jia, P. M. Datson, and J. Zhang. Induced polyploidy dramatically increases the size and alters the shape of fruit in Actinidia chinensis. *Annals of Botany*, 109(1):169–179, 2011.

[224] 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.

[225] J. Yang, M.-H. Moeinzadeh, H. Kuhl, J. Helmuth, P. Xiao, S. Haas, G. Liu, J. Zheng, Z. Sun, W. Fan et al. Haplotype-resolved sweet potato genome traces back its hexaploidization history. *Nature Plants*, 2017.

[226] W. Y. Yang, F. Hormozdiari, Z. Wang, D. He, B. Pasaniuc, and E. Eskin. Leveraging reads that span multiple single nucleotide polymorphisms for haplotype inference from sequencing data. *Bioinformatics*, 29(18):2245–2252, 2013.

[227] H. Yu, T. Lin, X. Meng, H. Du, J. Zhang, G. Liu, M. Chen, Y. Jing, L. Kou, X. Li et al. A route to de novo domestication of wild allotetraploid rice. *Cell*, 184(5):1156–1170.e14, 2021.

[228] Z. Yu, T. H. H. Coorens, M. M. Uddin, K. G. Ardlie, N. Lennon, and P. Natarajan. Genetic variation across and within individuals. *Nature Reviews Genetics*, 25(8):548–562, 2024.

[229] Y. Yuan, B. L. Ton, W. J. W. Thomas, J. Batley, and D. Edwards. Supporting crop plant resilience during climate change. *Crop Science*, 63(4):1816–1828, 2023.

[230] C. Zahn. Approximating symmetric relations by equivalence relations. *Journal of the Society for Industrial & Applied Mathematics*, 12, 1964.

[231] M. Zarrei, J. R. MacDonald, D. Merico, and S. W. Scherer. A copy number variation map of the human genome. *Nature Reviews Genetics*, 16(3):172–183, 2015.

[232] L. Zhang, S. Wu, X. Chang, X. Wang, Y. Zhao, Y. Xia, R. N. Trigiano, Y. Jiao, and F. Chen. The ancient wave of polyploidization events in flowering plants and their facilitated adaptation to environmental stress. *Plant, Cell & Environment*, 43(12):2847–2856, 2020.

[233] Q. Zhang, Y. Qi, H. Pan, H. Tang, G. Wang, X. Hua, Y. Wang, L. Lin, Z. Li, Y. Li et al. Genomic insights into the recent chromosome reduction of autopolyploid sugarcane Saccharum spontaneum. *Nature Genetics*, 54(6):885–896, 2022.

[234] W. Zhang, C. Luo, F. Scossa, Q. Zhang, B. Usadel, A. R. Fernie, H. Mei, and W. Wen. A phased genome based on single sperm sequencing reveals crossover pattern and complex relatedness in tea plants. *Plant Journal*, 105(1):197–208, 2021.

[235] Y. Zhou, Z. Zhang, Z. Bao, H. Li, Y. Lyu, Y. Zan, Y. Wu, L. Cheng, Y. Fang, K. Wu et al. Graph pangenome captures missing heritability and empowers tomato breeding. *Nature*, 606(7914):527–534, 2022.

[236] E. Çinlar. *Probability and Stochastics*. Graduate Texts in Mathematics. Springer New York, NY, 1 edition, 2011.

Appendix A

Local ancestry inference

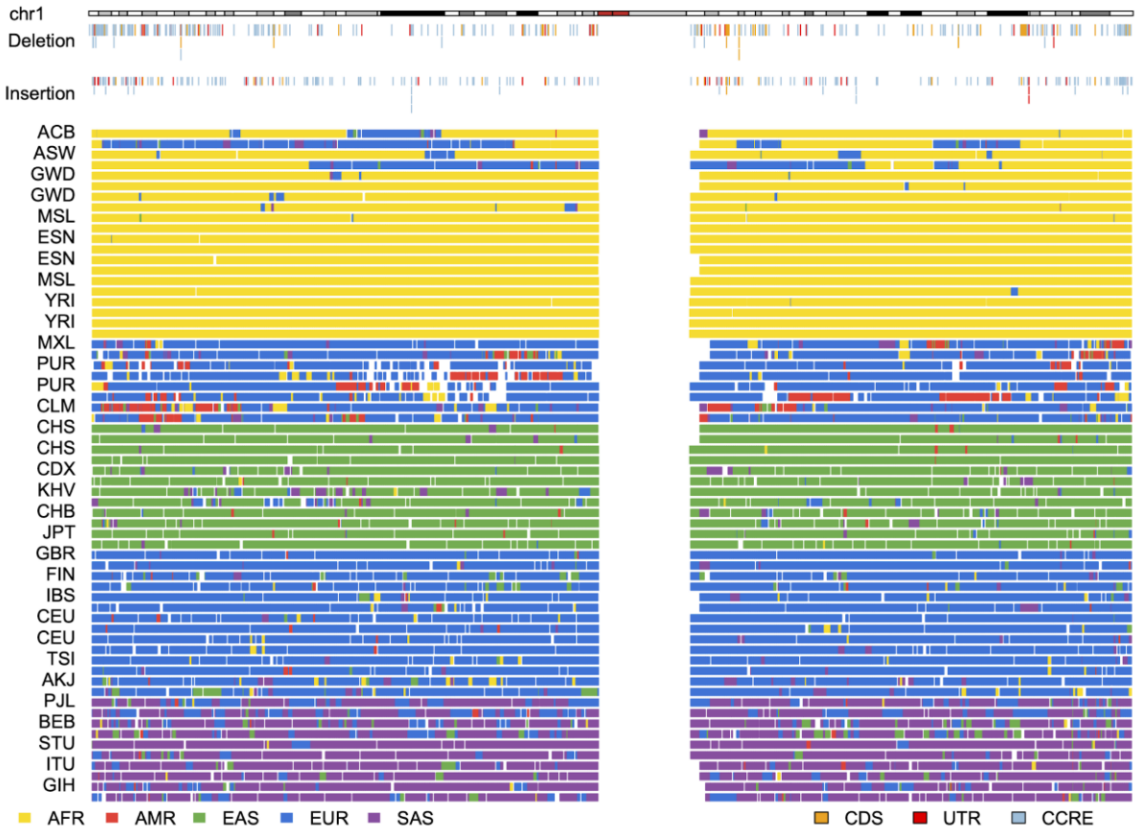


Figure A.1: Inferred local ancestry blocks across chromosome 1 for haplotype-phased assemblies. Ancestry calls are determined using RFMix and our Hidden Markov Model-based approach. White spaces are discordant calls between the two approaches, known gaps, centromeres, and/or segmental duplications. Deletion and insertion SVs that overlap with coding sequences (CDS), untranslated regions (UTR), and putative promoter sequences (CCRE) are annotated above. Figure taken from [56] (Supplementary Information).

Appendix B

Haplotype threading for polyploid phasing

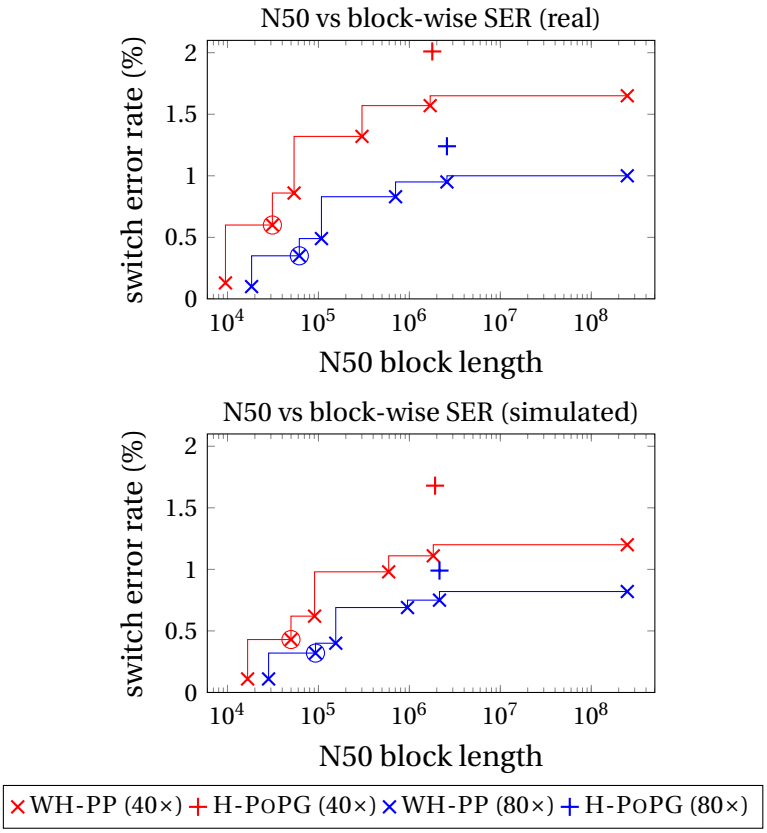


Figure B.1: N50 and SER. N50 block lengths and the respective block-wise switch error rates for different block cut strategies of WHATSHAP POLYPHASE (default strategy marked by a circle) on the real tetraploid read dataset (top) and the simulated tetraploid dataset (bottom) with 40× and 80× coverage. Figure taken from [184], licensed under the CC BY 4.0 licence.

Table B.1: 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. The total length of the chromosome is 249 Mb. Table taken from [184].

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

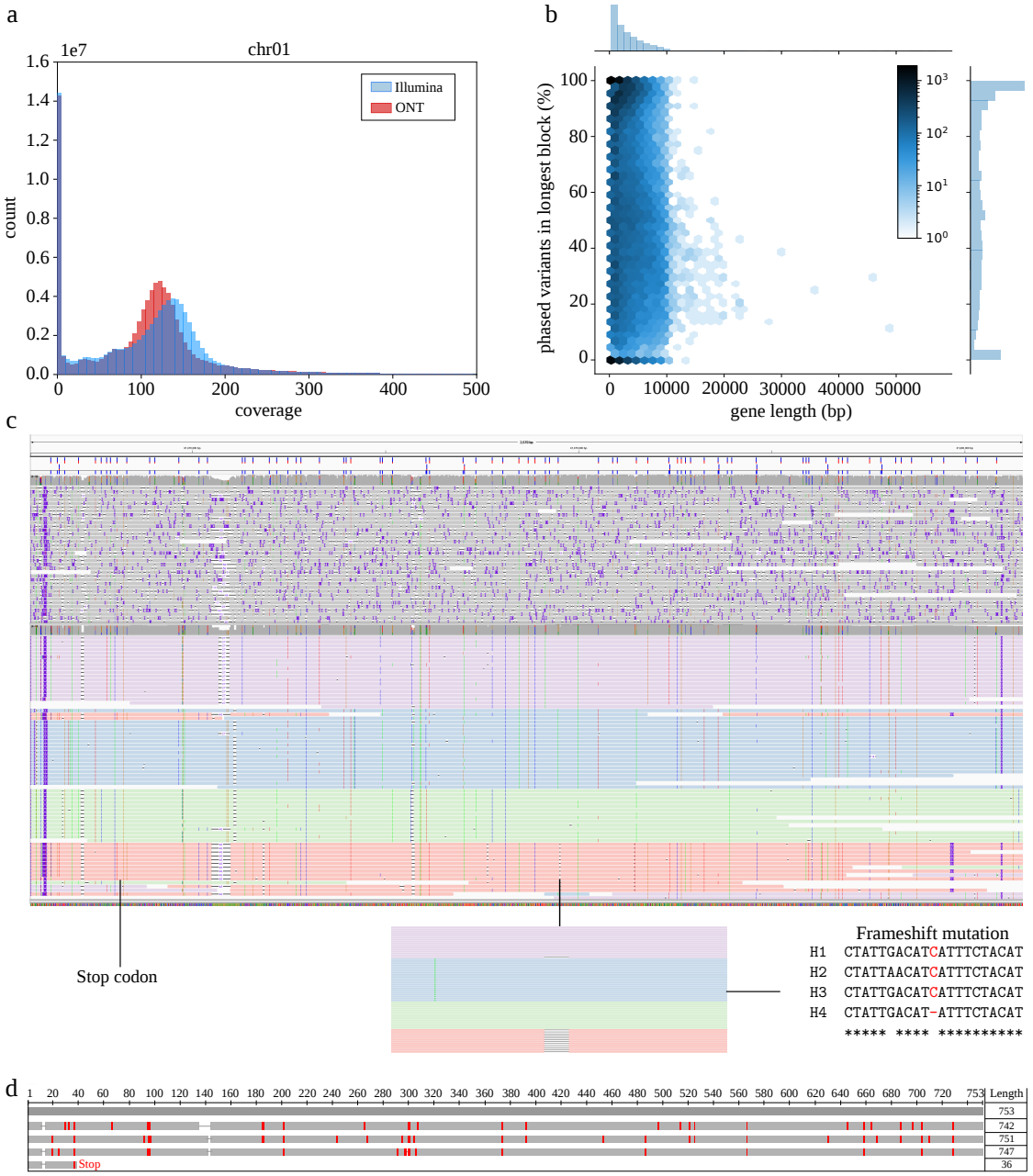


Figure B.2: Phasing of potato genome. **(a)** Per-base coverage distribution of Illumina and ONT MinION alignments on Chr01. **(b)** Fraction of phased variants in relation to gene length. The x-axis shows the gene length and the y-axis the percentage of phased variants in the longest block. Axis histograms and hexagons illustrate the distribution of data points. **(c)** IGV [173] 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 [184], licensed under the CC BY 4.0 licence.

Appendix C

Haplotype-resolved assembly of tetraploid potato

C.1 Data production

The following section is taken from [189] and was contributed by Freya Ziegler, a co-author of this publication.

Plants were grown under greenhouse conditions at Düsseldorf university. After three weeks, young leaves from a single plant were harvested and immediately frozen in liquid nitrogen. Next, DNA was extracted from 1 g of frozen leaf material as previously described [162, 191]. The DNA was size-selected using the Circulomics Short-Read Eliminator XL Kit (Circulomics Cat# SKU SS-100-111-01). DNA quality was assayed on a 1% agarose gel and using a NanoDrop Spectrophotometer (Thermo Fisher Scientific, USA). Sequencing libraries were prepared using the Oxford Nanopore Technologies (ONT) library preparation and sequencing kit SQK-LSK114, following standard protocols suggested by the manufacturer (Oxford Nanopore, UK). In brief, genomic DNA fragments were repaired and 3'-adenylated using the NEBNext FFPE DNA Repair Mix and the NEBNext Ultra II End Repair/ A-Tailing Module (New England Biolabs, USA). Sequencing adapters provided by ONT were then ligated using NEBNext Quick Ligation Module (NEB). After purifying the product with AMPure XP beads (Beckmann Coulter, CA, USA), libraries were loaded onto primed 10.4.1 Spot-On Flow Cells and sequenced using a PromethION sequencer (Oxford Nanopore Technologies, Oxford, UK) for 72-hours. Basecalling was performed using Oxford Nanopore guppy software (v6.3.8) with "super" accuracy models resulting in 162 Gb of ONT reads passing the quality filter.

Restriction enzyme Pore-C libraries (RE-Pore-C) were prepared following the ONT

Plant RE-Pore-C protocol (Oxford Nanopore Technologies), employing DpnII as the restriction enzyme [53]. After an overnight incubation, the enzyme was heat-denatured to facilitate the ligation of adjacent DNA clusters. Subsequent processes included protein degradation and decrosslinking, liberating chimeric Pore-C double-stranded DNA polymers. DNA quality and concentration were monitored using 1% agarose gel electrophoresis, a NanoDrop spectrophotometer (Thermo Fisher Scientific), and the Qubit DNA Assay Kit with a Qubit fluorimeter (Thermo Fisher Scientific). Genomic DNA fragments underwent repair, end repair, and A-tailing via NEBNext FFPE DNA Repair Mix (New England BioLabs Inc) and the NEBNext Ultra II End Repair/A-Tailing Module (New England BioLabs Inc). Afterwards, adapters were ligated and clean up was performed (ONT ligation sequencing DNA V14 (SQK-LSK114) protocol). Resulting libraries were sequenced on R10.4.1 PromethION flow cells, with a runtime set to 100 hours in accurate speed mode (260 base pairs per second). Flowcells were flushed and reloaded after 24, 48, and 72 hours. In total, seven Runs were performed and basecalling was done using guppy v6.4.8 (ONT).

C.2 Supplementary Figures

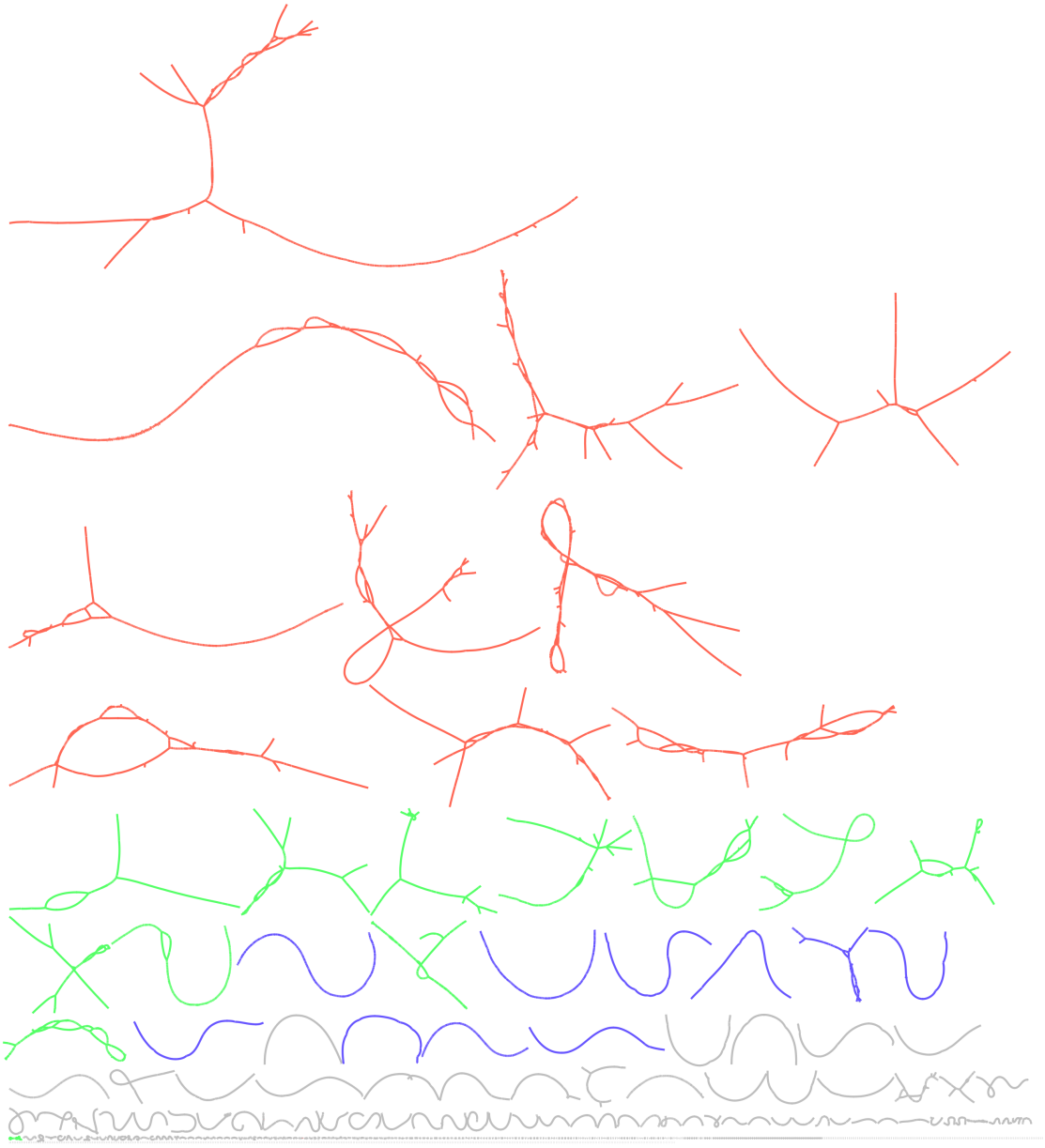


Figure C.1: Bandage visualisation of the hifiasm raw unitig graph. The different size categories are indicated by the colouring: red represents the largest components (91–190 Mb), green the second largest (45–66 Mb), and blue the third largest (20–32 Mb). Figure taken from [189] (Supplementary Material), licensed under the CC BY 4.0 licence.

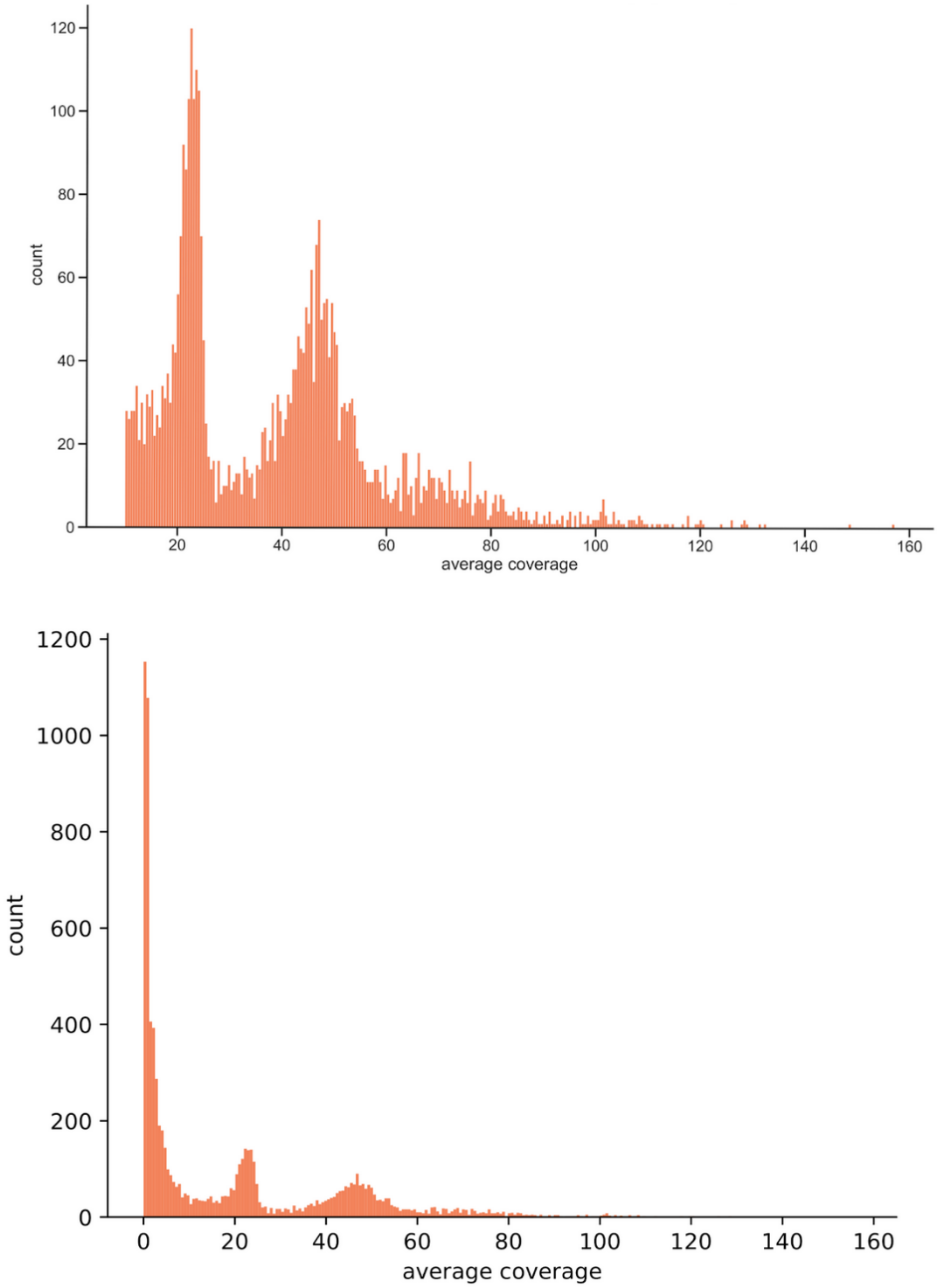


Figure C.2: Dosage distribution of unitigs. Top: Unitigs with coverage < 10 are filtered out for better visualisation. Bottom: All unitigs are shown. Figure taken from [189] (Supplementary Material), licensed under the CC BY 4.0 licence.

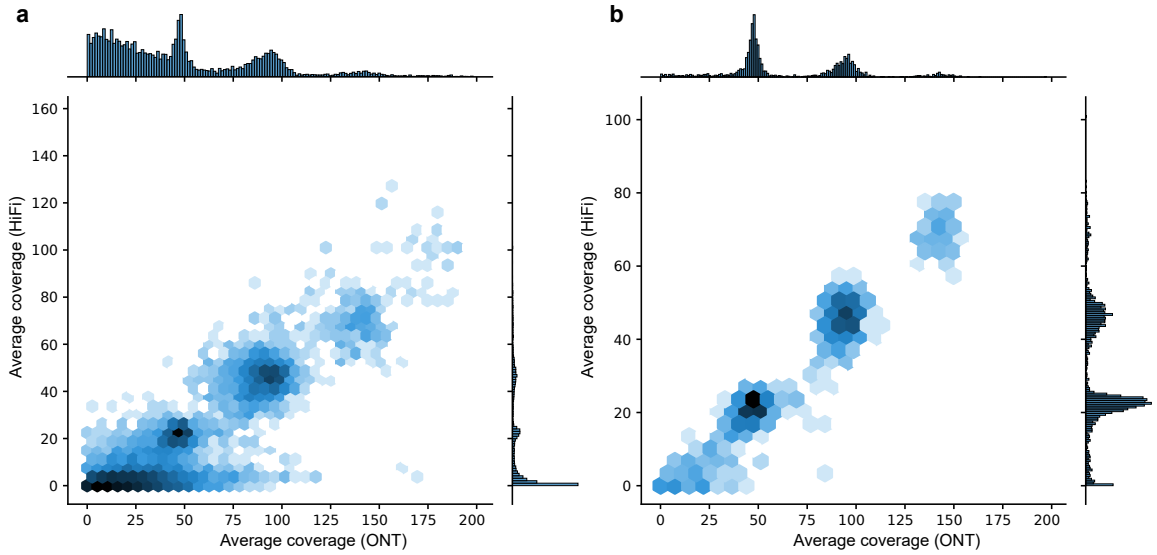


Figure C.3: ONT and HiFi coverage. Joint coverage of ONT reads as well as HiFi reads, mapped to the hifiasm assembly graph of Altus. **a.** All nodes without any filtering applied, **b.** Filtered to only show coverage of nodes with a unique sequence length of at least 100 kb. Figure taken from [189] (Supplementary Material), licensed under the CC BY 4.0 licence.

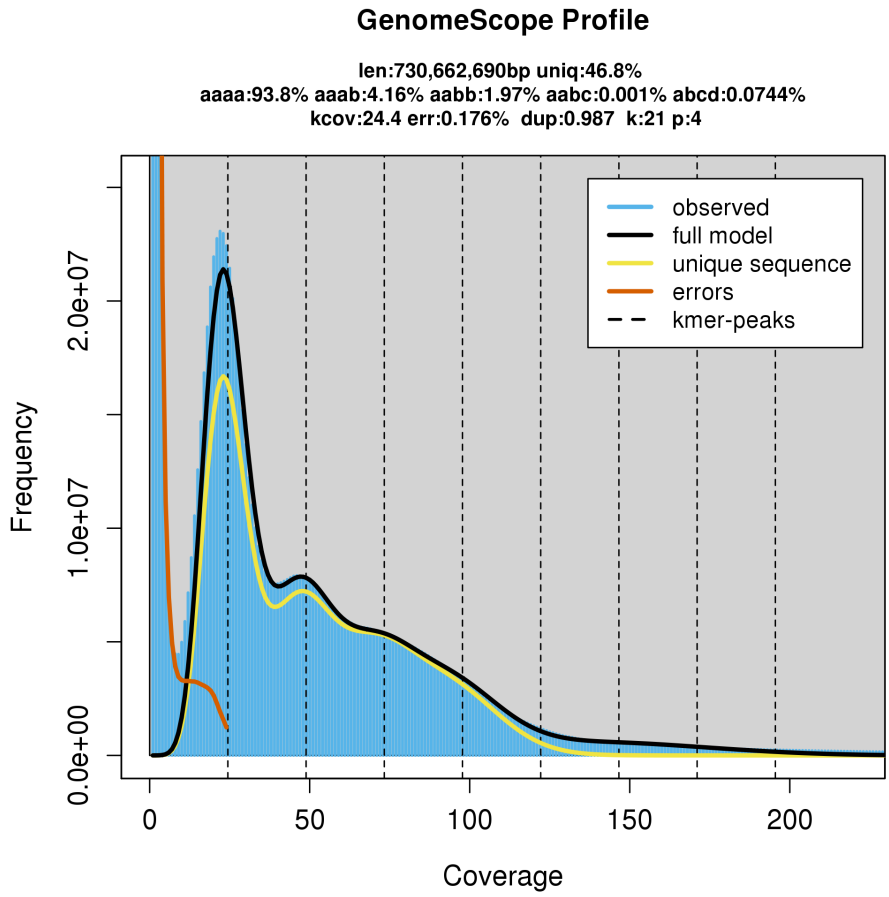


Figure C.4: Genome size estimation using GenomeScope. Figure taken from [189] (Supplementary Material), licensed under the CC BY 4.0 licence.

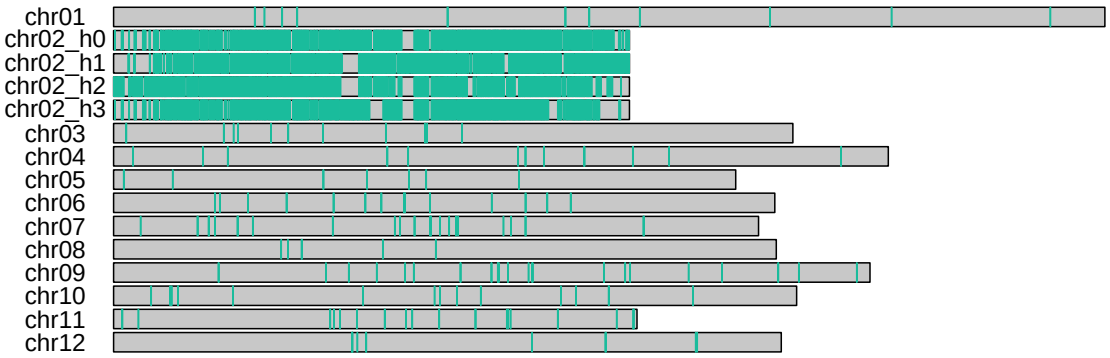


Figure C.5: Mapping chr02 unitigs to DMv6.1. The unitigs from all four haplotype clusters of Chromosome 2 are mapped to the DMv6.1 reference. The colouring indicates mapped regions. Figure taken from [189] (Supplementary Material), licensed under the CC BY 4.0 licence.

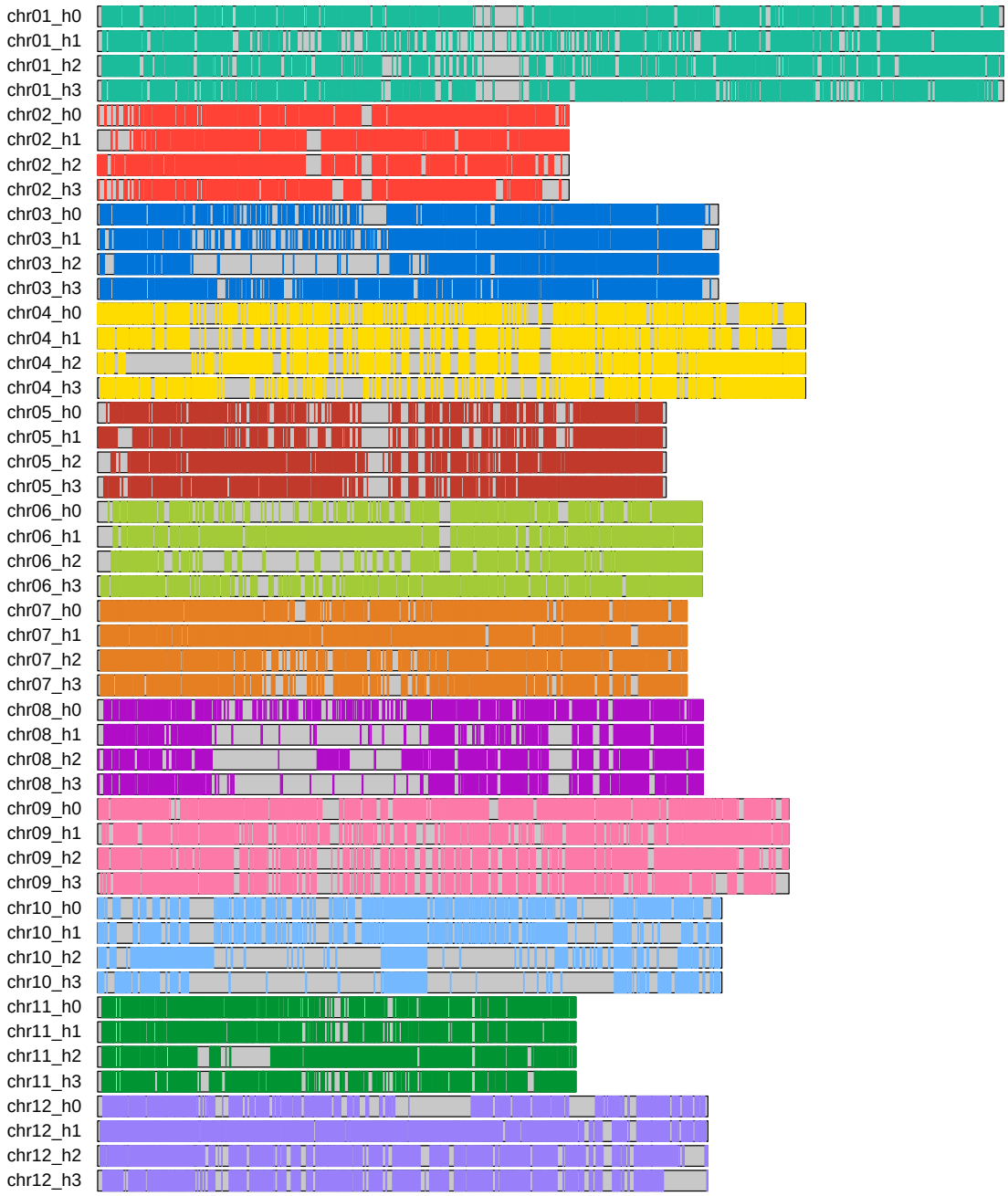


Figure C.6: Haplotype-resolved version of main Figure 4a. All unitigs from the 48 haplotype clusters are mapped to the reference DMv6.1. Figure taken from [189] (Supplementary Material), licensed under the CC BY 4.0 licence.

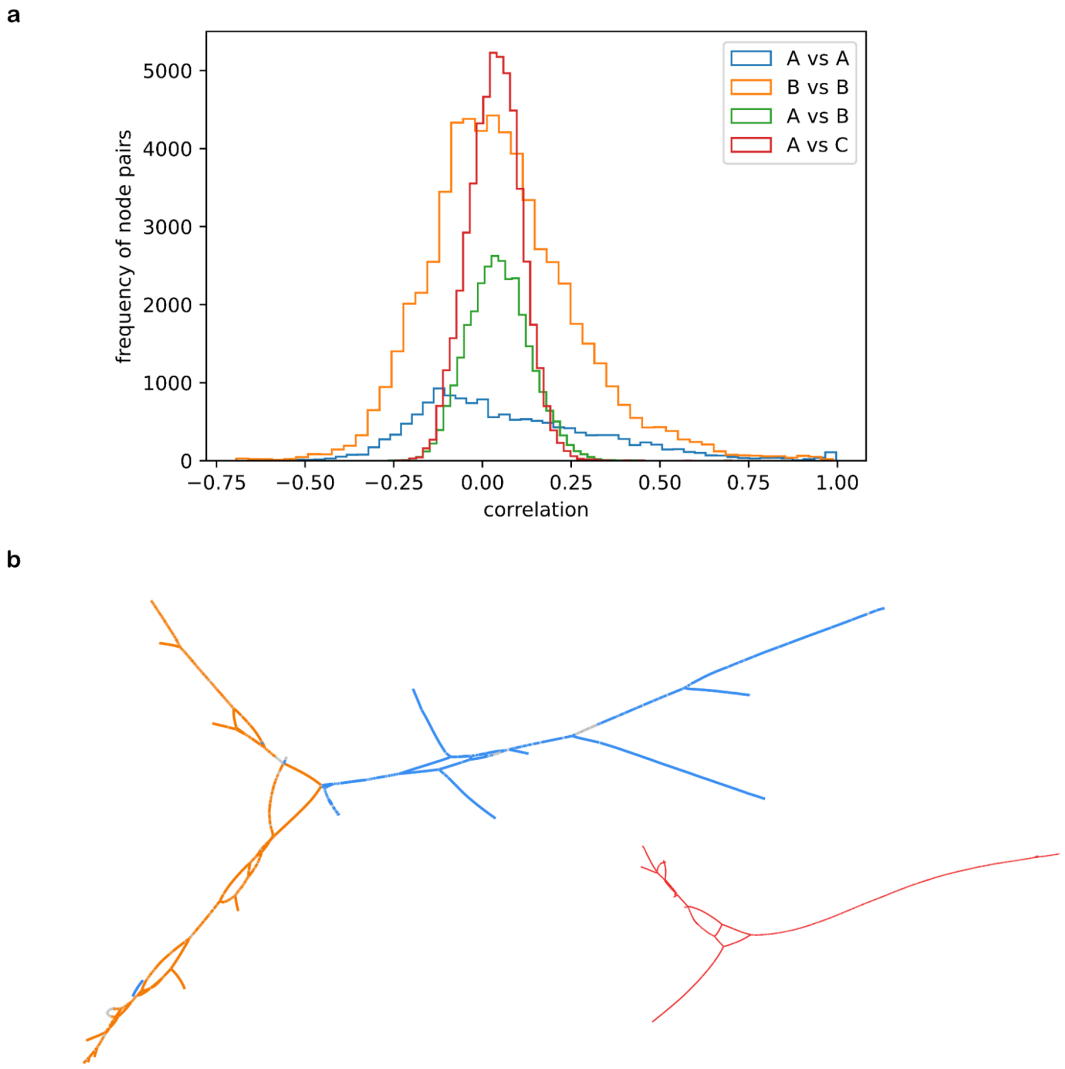


Figure C.7: Analysis of assembly errors in the hifiasm graph between Chromosomes 10 and 12 of DMv6.1. a. Distribution of the pairwise correlation of nodes contained within set A (blue), within set B (orange), between sets A and B (green), and between A and an arbitrarily chosen different component C (red). **b.** Left: The component of the assembly graph that contains node sets A (unitigs that map to Chromosome 12 in the DMv6.1 sequence, coloured in blue) and B (unitigs that map to Chromosome 10 in the DMv6.1 sequence, coloured in orange). Right: The component used for comparison, labelled C (unitigs are coloured in red). Figure taken from [189] (Supplementary Material), licensed under the CC BY 4.0 licence.

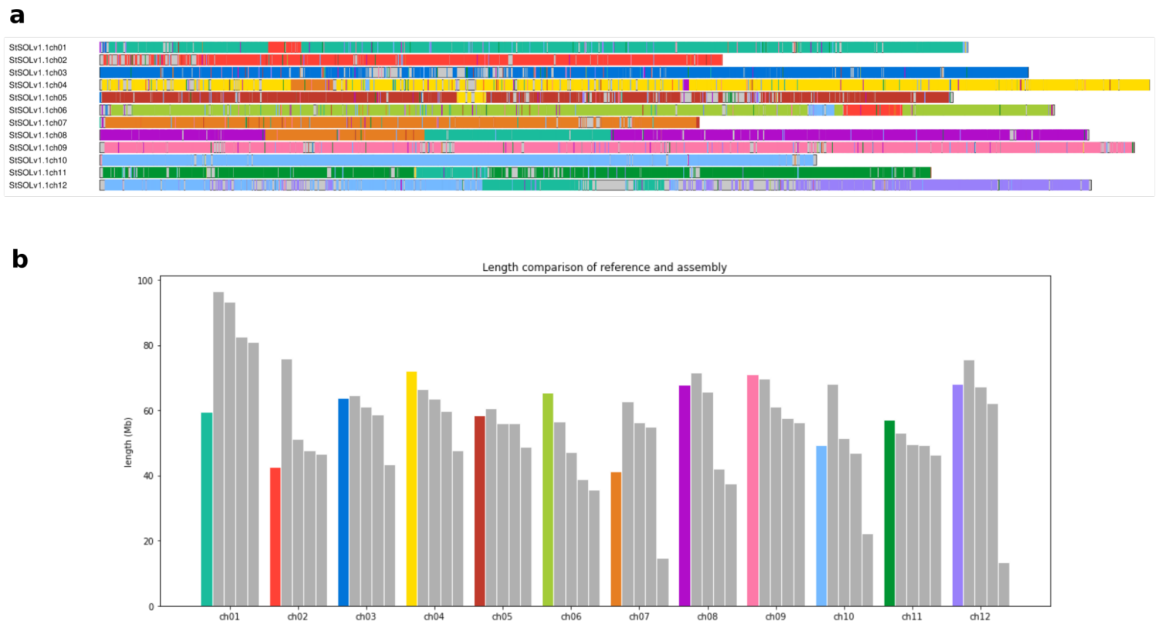


Figure C.8: Mapping to Solyntus. The clustered nodes are mapped to the Solyntus v1.1 reference sequence. Figure taken from [189] (Supplementary Material), licensed under the CC BY 4.0 licence.

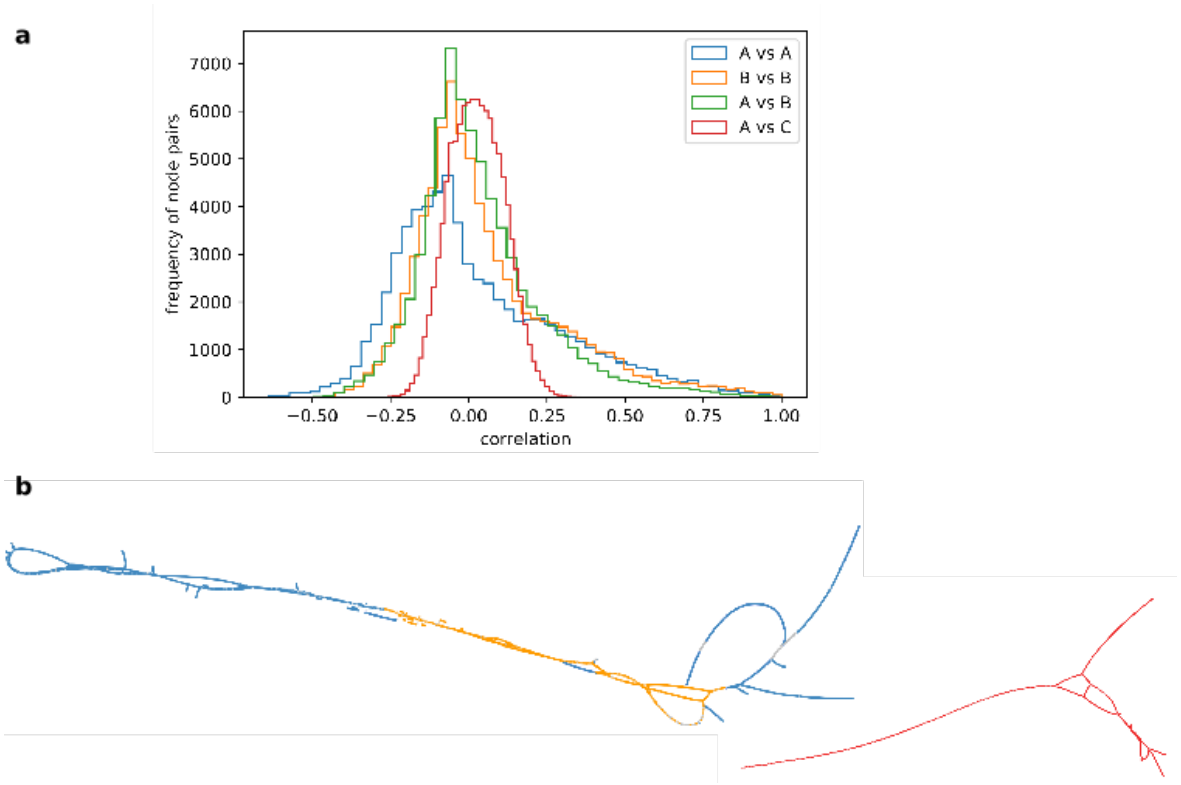


Figure C.9: Analysis of assembly inserts between Chromosomes 1 and 8 in the Solyntus v1.1 reference sequence. a. Distribution of the pairwise correlation of nodes contained within set A (blue), within set B (orange), between sets A and B (green), and between A and an arbitrarily chosen different component C (red). **b.** Left: The component of the assembly graph that contains node sets A (unitigs that map to Chromosome 1, coloured in blue) and B (unitigs that map to chromosome 8, coloured in orange). Right: The component used for comparison, labelled C (unitigs are coloured in red). Figure taken from [189] (Supplementary Material), licensed under the CC BY 4.0 licence.

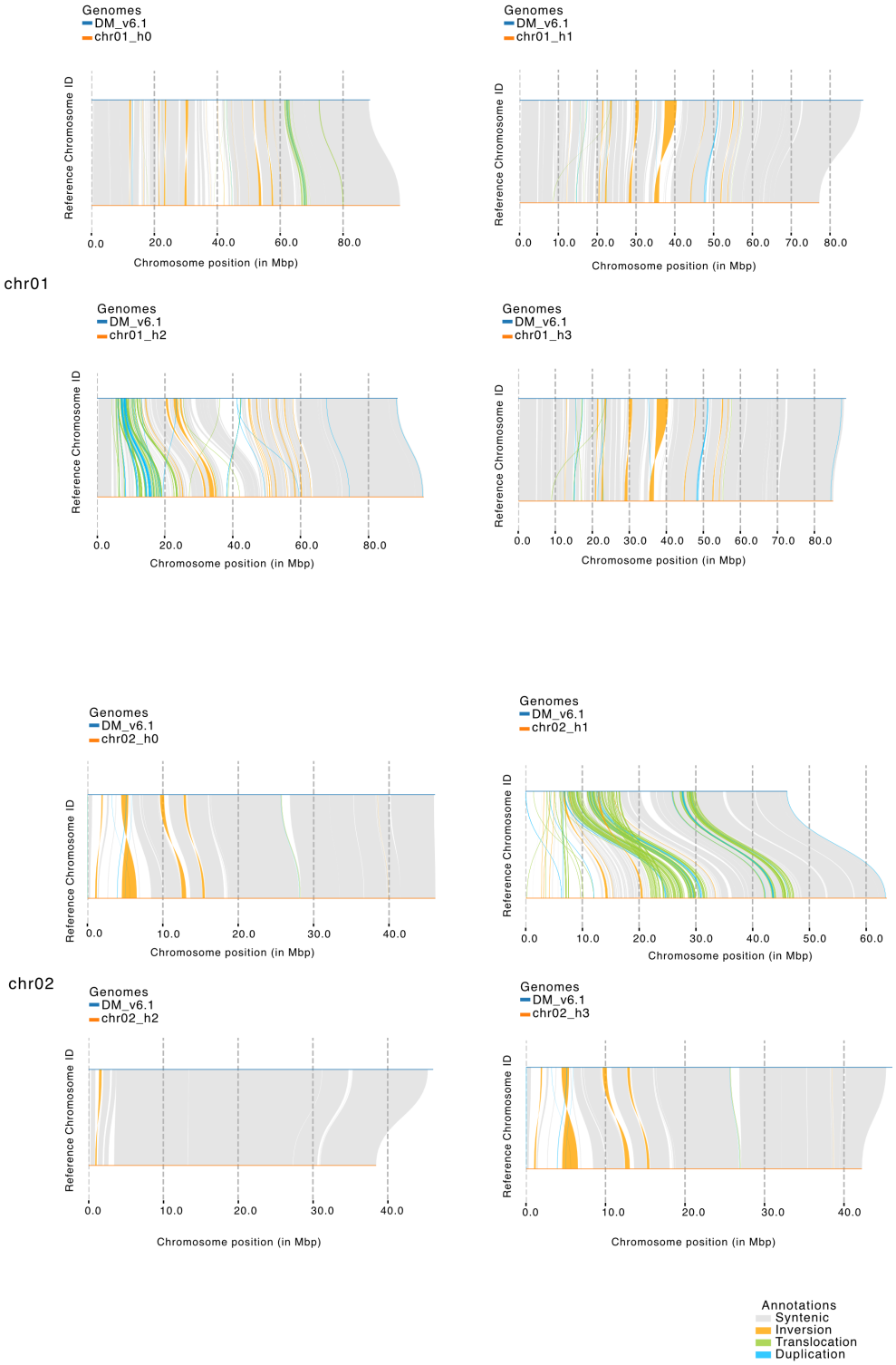


Figure C.10: Synteny analysis. Comparison between all haplotypes of chr01 (top) and chr02 (bottom) and the reference sequence DMv6.1. Figure taken from [189] (Supplementary Material), licensed under the CC BY 4.0 licence.

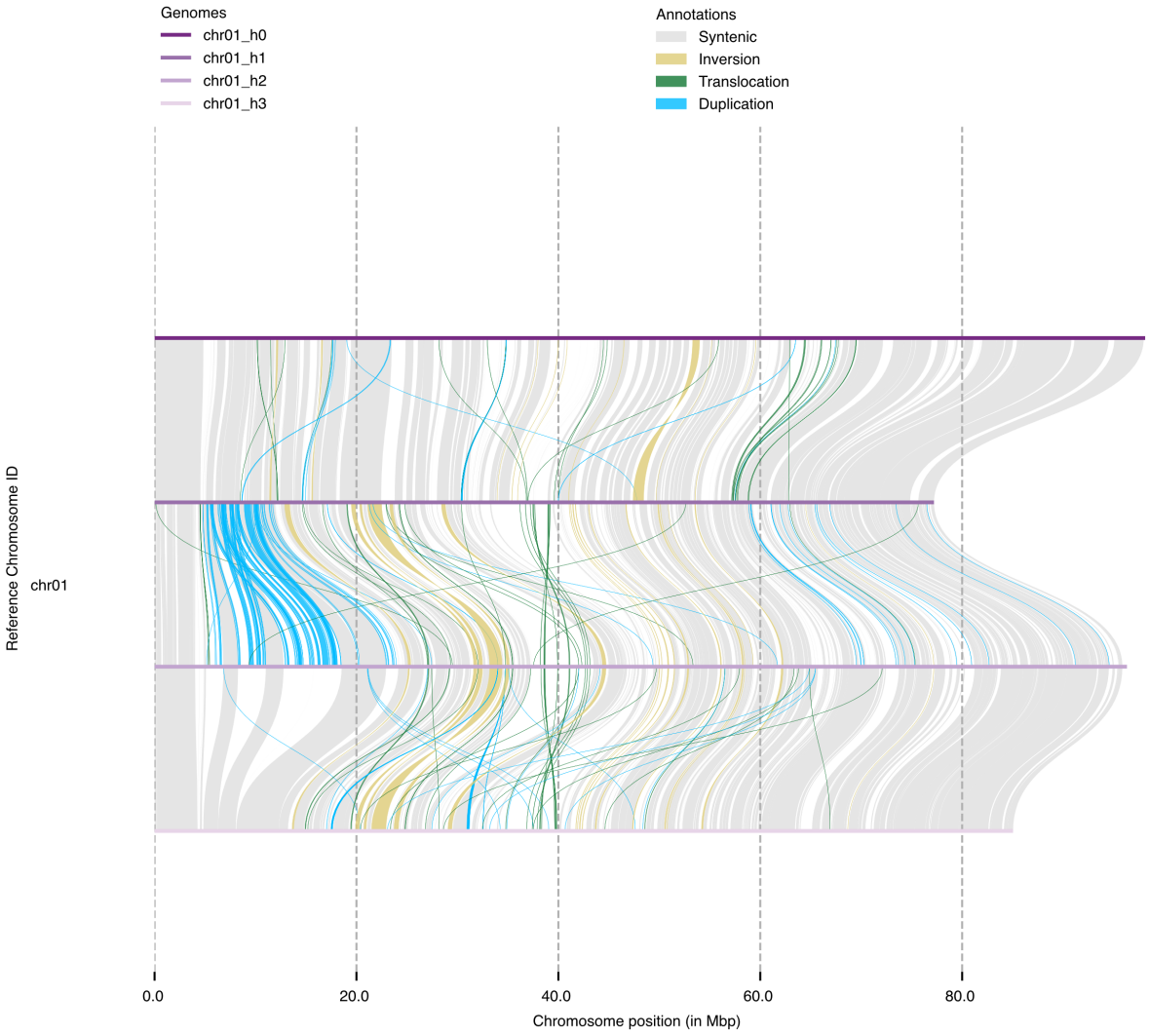


Figure C.11: Synteny analysis of chr01. Figure taken from [189] (Supplementary Material), licensed under the CC BY 4.0 licence.

Appendix D

Multigenerational graph-based assembly

chromosome	paternal haplotype	maternal haplotype
1	247.77	247.02
2	244.35	242.51
3	202.56	201.41
4	190.45	192.39
5	182.15	183.84
6	172.02	171.87
7	161.23	160.16
8	146.19	147.17
9	136.21	129.07
10	134.51	134.38
11	134.72	134.80
12	133.49	134.86
13	106.74	109.44
14	103.03	100.41
15	98.82	98.03
16	90.51	91.59
17	83.72	84.21
18	80.49	81.16
19	63.52	62.75
20	66.52	66.23
21	43.90	43.39
22	48.38	48.47

Table D.1: Lengths of single-sample assemblies from `https://github.com/biomonika/HPP/tree/main/T2T-Pedigree-project%20` [Accessed: 15.04.2025]

chromosome	paternal haplotype	maternal haplotype
1	243.25	243.42
2	242.74	242.55
3	202.64	191.84
4	192.69	182.58
5	182.27	183.88
6	172.01	172.26
7	161.25	160.22
8	146.20	147.08
9	136.26	129.20
10	134.50	134.30
11	134.72	134.83
12	133.48	134.88
13	99.89	25.52
14	96.72	90.55
15	92.15	83.15
16	90.48	85.77
17	83.77	80.43
18	80.49	81.29
19	63.56	61.49
20	66.51	66.02
21	36.62	35.47
22	36.39	42.20

Table D.2: Lengths of our graph-based assemblies after scaffolding using RagTag.

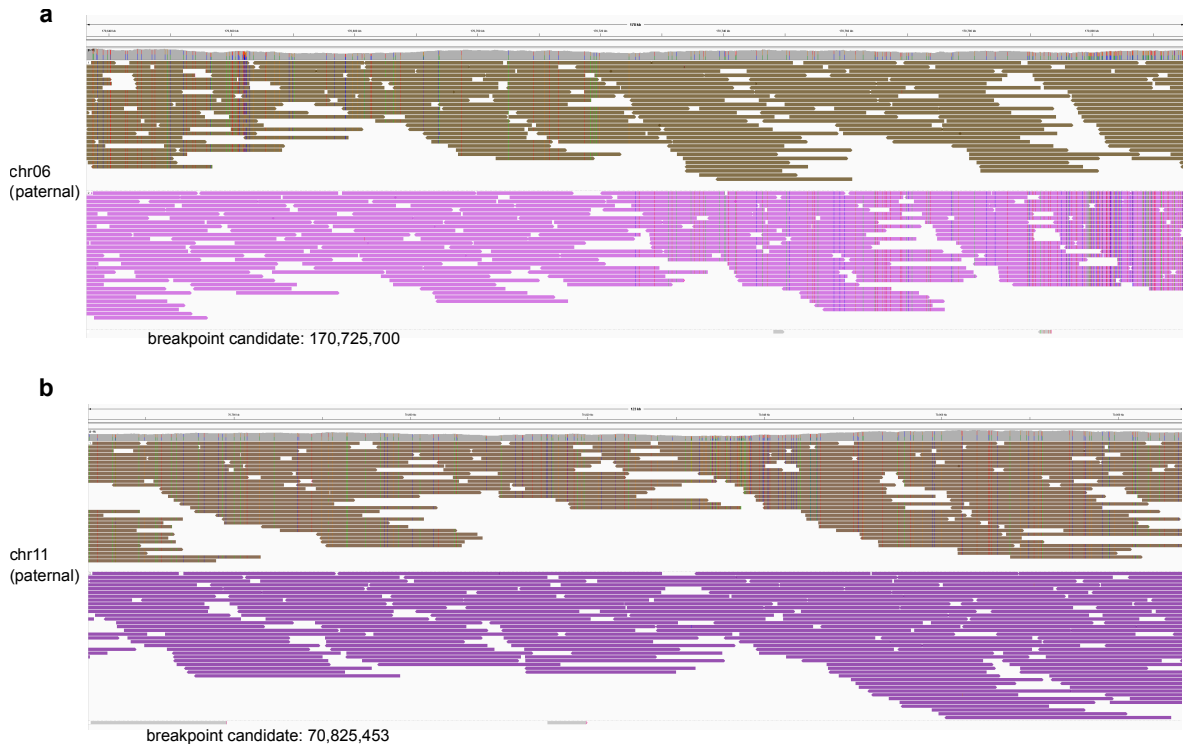


Figure D.1: IGV analysis of recombination breakpoint candidates. The single-sample assembly of the paternal haplotype of PAN027 serves as the reference, HiFi reads of PAN011 are aligned to it and coloured and grouped according to their haplotype assignment. **a.** Around the candidate position on paternal chromosome 6, the read set that shows variation changes. This phase switch indicate a likely recombination breakpoint. The visible area covers 178 kb. **b.** Around the candidate position on paternal chromosome 11, no switch in variation is visible between the two haplotagged read sets, signalling the absence of a phase switch. The alignment does not support the recombination candidate. The visible area covers 123 kb.

Appendix E

Publication details and copyright information

The following presents a list of the publications that Chapters 2, 3 and 4 are based on. These publications are:

- Ebert, Audano, Zhu, Rodriguez-Martin et al.: Haplotype-resolved diverse human genomes and integrated analysis of structural variation. *Science*, 372(6573), 2021.
- Schrinner, Serra Mari, Ebler et al.: Haplotype threading: accurate polyploid phasing from long reads. *Genome Biology*, 21(1), 2020.
- Serra Mari et al.: Haplotype-resolved assembly of a tetraploid potato genome using long reads and low-depth offspring data. *Genome Biology*, 25(1):26, 2024.

Author contributions and licence details are given in the following.

E.1 Haplotype-resolved diverse human genomes and integrated analysis of structural variation

The manuscript "Haplotype-resolved diverse human genomes and integrated analysis of structural variation" was published in *Science* [56]. Chapter 2 is based on this publication and reuses some material from [56].

E.1.1 Licence

The article was published under the terms of the Science Journals Default License and is available under `https://doi.org/10.1126/science.abf7117`. According to the

Reprints and Permissions stated online (`https://www.science.org/content/page/reprints-and-permissions`), the following holds: "If you are the author of an article that was published in a Science journal, you retain the rights to use your paper and its contents as permitted under the License to Publish that you agreed to during the submission process." The License to Publish states:

> "[All authors] retain the non-exclusive right to use the Work in the following ways after publication of the Work by AAAS, without further permission: [...] Reprint the Final Published Version in print format for inclusion in a thesis or dissertation that the Author writes. [...]"

Permission for re-use in print and electronic has been granted by AAAS (Licence number 5917560872454).

E.1.2 Authors

The author list, as given in the manuscript [56], is as follows (* indicates joint first authorship):

> 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

E.1.3 Author contributions

The author contributions, as stated in the manuscript [56], are as follows:

"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."

E.2 Haplotype threading: accurate polyploid phasing from long reads

The manuscript "Haplotype threading: accurate polyploid phasing from long reads" was published in *Genome Biology* [184]. On this publication, I share first authorship with Sven Schrinner and Jana Ebler. Material of this manuscript has been re-used for Chapter 3.

E.2.1 Licence

The article was published under the CC-BY 4.0 licence and is available under `https://doi.org/10.1186/s13059-020-02158-1`. According to the online version of the article, the following holds:

E.2.2 Authors

The author list, as given in the manuscript [184], is as follows (* denotes joint first authorship):

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

E.2.3 Author contributions

The author contributions, as stated in the manuscript [184], are as follows:

"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.3 Haplotype-resolved assembly of a tetraploid potato genome using long reads and low-depth offspring data

The manuscript "Haplotype-resolved assembly of a tetraploid potato genome using long reads and low-depth offspring data" was published in *Genome Biology* [189]. Material of this manuscript has been re-used for Chapter 4.

E.3.1 Licence

The article was published under the CC-BY 4.0 licence and is available under `https://doi.org/10.1186/s13059-023-03160-z`. According to the online version of the article, the following holds:

> "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 licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence 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 licence, visit http://creativecommons.org/licenses/by/4.0/. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated in a credit line to the data."

E.3.2 Authors

The author list, as given in the manuscript [189], is as follows:

Rebecca Serra Mari, Sven Schrinner, Richard Finkers, Freya Maria Rosemarie Ziegler, Paul Arens, Maximilian H.-W. Schmidt, Björn Usadel, Gunnar W. Klau, Tobias Marschall

E.3.3 Author contributions

The author contributions, as stated in the manuscript [189], are as follows:

> "**R.S.M.** and T.M. designed the assembly method, with input from S.S., R.F., B.U., and G.W.K. **R.S.M.** implemented the assembly method, performed the assemblies, and created the figures. **R.S.M.**, S.S., R.F., G.W.K., B.U., and T.M. discussed and interpreted results. **R.S.M.**, R.F., B.U., and T.M. wrote the manuscript, with input from S.S. and G.W.K. P.A. provided offspring sequencing data. M.HW.S. produced HiFi data and ran an initial assembly. F.M.R.Z. produced ONT data and wrote text describing data production. All authors read and approved the final manuscript."