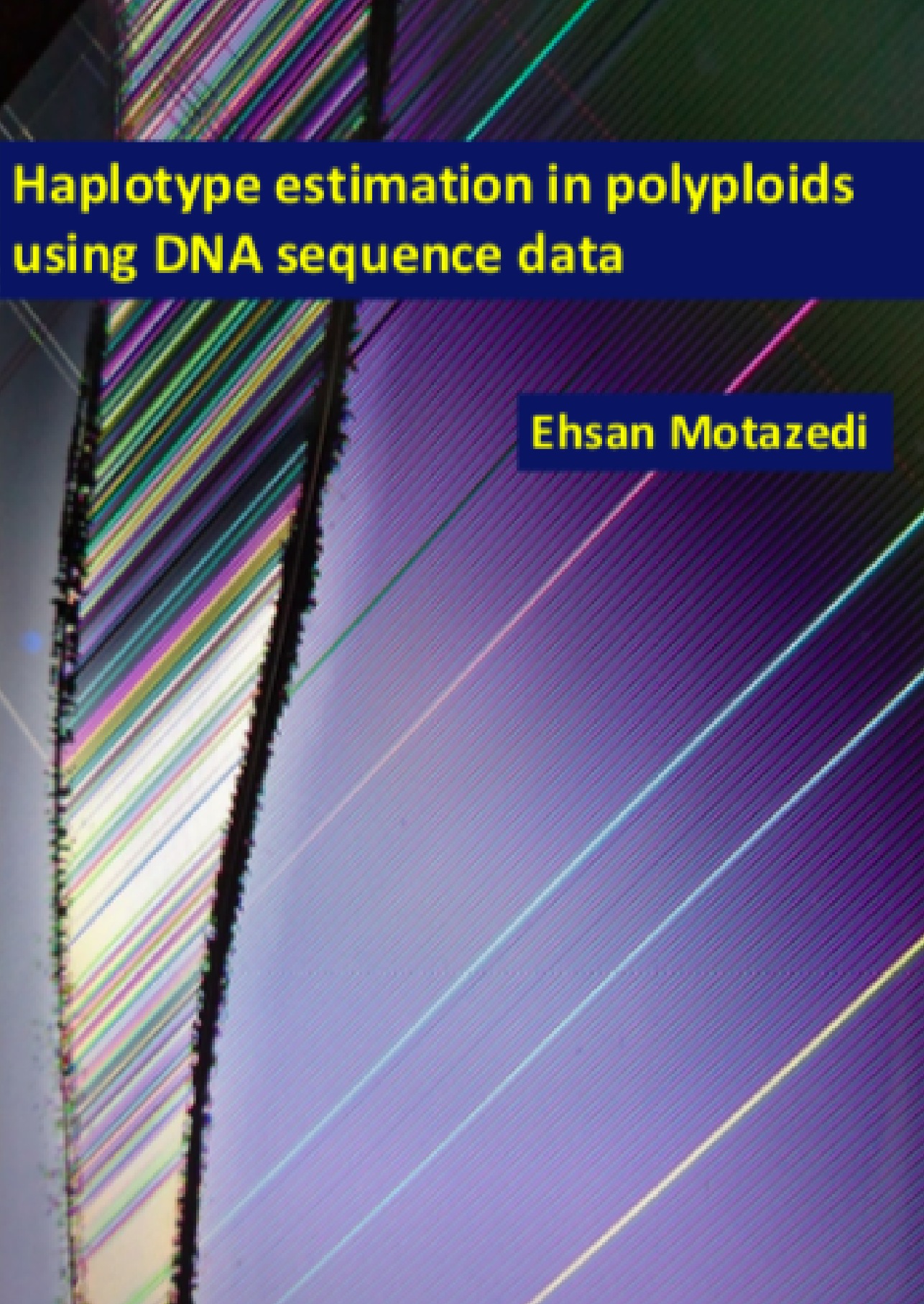
Haplotype estimation in polyploids using DNA sequence data

Ehsan Motazedi



Propositions

1. Sequence-based haplotyping in polyploid populations should make use of inheritance information.
   (this thesis)

2. The average read length influences the quality of sequence-based haplotyping more than sequencing depth.
   (this thesis)

3. The high expectations in genome editing are a result of a lack of knowledge about non-additive effects.

4. There is a positive correlation between the modesty of the first presentation of a new scientific concept and its long-term impact.

5. Obligatory inclusion of one PhD student in the thesis committee will bring more challenge and rigor to the PhD defense.

6. It is easier to find oriental food in the Netherlands than traditional Dutch cuisine.

Propositions belonging to the thesis, entitled

'Haplotype estimation in polyploids using DNA sequence data'.

Ehsan Motazedi
Wageningen, 7 November 2019.

Haplotype estimation in polyploids using DNA sequence data

Ehsan Motazedi

Thesis committee

Promotor
Prof. Dr D. de Ridder
Professor of Bioinformatics
Wageningen University & Research

Co-promotors
Dr C.A. Maliepaard
Associate Professor, Plant Breeding
Wageningen University & Research

Dr H.J. Finkers
Researcher, Plant Breeding
Wageningen University & Research

Other members
Prof. Dr F.A. van Eeuwijk, Wageningen University & Research
Prof. Dr T. Marschall, Max-Planck-Institut für Informatik, Saarbrücken, Germany
Dr J. Endelman, University of Wisconsin-Madison, USA
Dr A. van der Burgt, Solynta N.V., Wageningen

The work presented in this thesis was conducted under the auspices of the Graduate School Experimental Plant Sciences

Haplotype estimation in polyploids using DNA sequence data

Ehsan Motazedi

Thesis

submitted in fulfilment of the requirements for the degree of doctor
at Wageningen University
by the authority of the Rector Magnificus
Prof. Dr A.P.J. Mol,
in the presence of the
Thesis Committee appointed by the Academic Board
to be defended in public
on Thursday 7 November 2019
at 11 a.m. in the Aula.

Ehsan Motazedi
Haplotype estimation in polyploids using DNA sequence data, 153 pages.

PhD thesis, Wageningen University, Wageningen, the Netherlands (2019)
With references, with summary in English and in Dutch

ISBN 978-94-6395-121-0
DOI 10.18174/500092

To all farmers,
who make sense out of the selected genes.

Contents

1

Introduction

The horse saw the camel,
Uttered with laughter hoarse:
“Such a monstrous freak of a horse!”
The camel rejoined:
“You, a horse?
Not nearly, for sure!
An immature camel, that is what you are.”
Only God knew, omniscient indeed
That they were mammals
Of different breed

Vladimir Mayakovsky (1893-1930)

Sexual reproduction is the main mechanism for maintaining biodiversity in higher organisms, including plants, as an offspring receives at random half of the genes of each parent. Many organisms are diploids, that is they contain two copies of each chromosome representing their maternal and paternal ancestry. Polyploidization is a natural or artificial process that leads to an increase in the number of chromosomal copies to $k > 2$, which is a major force in plant diversification and has yielded many important and widely used agronomic species such as wheat, potato, ryegrass, alfalfa, cotton and ornamental flowers. However, the same phenomenon dramatically adds to the genetic and genomic complexity and has caused a lag in the analysis of these important plants. The advent of modern molecular technologies and powerful computational resources has changed this situation in recent years and led to the development of novel methods for the genetic and genomic analysis of polyploid crops [15]. In particular, the increasing prevalence of high-throughput DNA sequencing technologies proves to be an asset for deciphering polyploid genomes. The challenge is to deal with the computational complexity of analysing these data and the associated noise, as well as in combining the genomic measurements with other sources of information, most notably genetic data.

This thesis focuses on the estimation of polyploid haplotypes from DNA sequence and genetic data. Below, I first clearly define haplotypes and explain their importance (Section 1.1). Polyploidy and its consequences are discussed in Section 1.2, with an emphasis on the relevance for plant breeding. The problem of haplotype estimation from DNA sequence reads is introduced in Section 1.3 and several estimation algorithms are introduced for diploids and polyploids. This chapter is concluded by providing an outline of the remaining chapters (Section 1.4).

1.1. Background

The nuclear DNA in eukaryotes is stored and packed in the form of chromatin, i.e. a set of often linear chromosomes wrapped around histones. In diploid organisms, the chromosomes can be divided into pairs that consist of one maternal and one paternal chromosome, hosting the same set of genes at the same loci. The chromosomes in each pair are called homologous chromosomes, which have very similar nucleotide sequences with nonetheless potentially important allelic or structural variations. The number of chromosomes in a somatic cell is denoted by $2n$ and the number of chromosomes in a diploid cell denoted by $2x$, where x is the basic or *haploid* chromosome number equal to the number of homologous groups. The nucleotide sequence of each chromosome is called its *haplotype*.

To quantify the inheritance patterns of a phenotype it is important to take into account that the genes (and DNA segments in general) located on the same chromosome tend to be inherited together unless recombination events occur, the frequencies of which depend on the distance between the genes (or the DNA segments) and their positions on the chromosome. This fact is used to construct genetic linkage maps and to identify genomic markers associated with a phenotype and located within proximity of the unknown actual causative loci, by investigating the co-segregation of the genetic markers with the phenotype [33, 74]. With the advent of single nucleotide polymorphism (SNP) markers, which can be determined by high-throughput assays such as SNP arrays [43, 72], high resolution genetic maps have been constructed that landmark

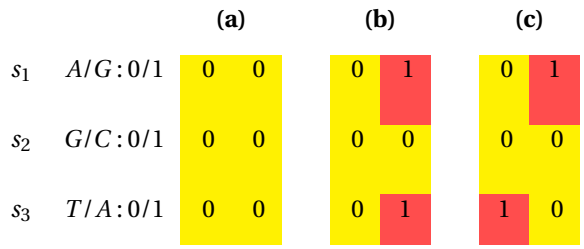


Figure 1.1: **(a)** Two normal haplotypes of a gene containing 3 bi-allelic SNPs with the variant nucleotides coded by 0/1. **(b)** Two mutations (red) on the same haplotype render one copy of the gene dysfunctional. **(c)** One deleterious mutation on each copy of the gene renders both copies dysfunctional.

the whole genome and can narrow associated genomic regions down to a few centiMorgans. The power of genetic studies is, however, highly dependent on the informativeness of the employed genetic markers as well as the strength of their linkage to the causative loci.

While haplotypes can be defined as the nucleotide sequence of each chromosome at a genomic locus, a more general definition is as the set of ordered genomic variants over each chromosome that are transmitted together from a parent to the offspring. The application of the latter definition is not limited to sequencing experiments, as genotyping experiments can use the same definition in which context haplotypes correspond to ordered genotype alleles over each chromosome. Also, in statistical genetics and breeding applications the interest often lies in the variant genomic positions, which are represented by a set of genomic markers, rather than by the whole DNA sequence. The problem of determining the order of marker alleles on each chromosome is called *haplotyping* or *phasing*.

Knowing the haplotypes of the parents and the offspring allows the unambiguous following of meiotic transmission through generations, which is a great advantage in studying trait inheritance [60, 79]. Haplotypes cover the genetic diversity in a population and can be used as powerful multi-allelic markers in genetic studies [6, 36, 53]. In addition, phasing information is often helpful in estimating missing SNP alleles using the other alleles in a haplotype, an approach known as genotype imputation [16, 49].

Haplotypes are also of direct biological importance, as both gene expression and protein function can be affected by a single variant allele being in cis or trans with other alleles [86] (Figure 1.1). Besides, haplotypes can determine epigenetic features of the genome, such as DNA methylation [11, 32]. In plant and animal breeding, haplotypes are invaluable tools to assist selecting the best varieties and races, or parents for crossing. Molecular selection, compared to the traditional selection approaches, provides a means to select desirable individuals very early in the breeding cycle, hence reducing the time and the cost needed to end up with efficient animal breeds or plant cultivars [20, 45, 61]. In short, haplotypes are essential units of inheritance and therefore of prime interest in genetic analysis.

In case the homologous haplotypes are the same in an individual (except maybe at a single marker position) with the same allele on each homologous chromosome, i.e. all of the SNPs (except maybe just one) are *homozygous*, knowing the SNP genotypes imme-

diately leads to the determination of the haplotypes. In this situation, the phasing problem is reduced to finding markers that are in tight linkage. To find such markers, one has to look into the inheritance of distinct alleles at a set of segregating loci, in a population of homozygous mosaic lines derived from crosses between homozygous founders that each contain one of the non-identical alleles at each locus. This approach been successfully applied for quantitative trait locus (QTL) analysis and diversity profiling of recombinant inbred lines (RIL) [66, 98], advanced intercross recombinant inbred lines (AI-RIL) [8] and multi-parent advanced generation intercross (MAGIC) lines [41, 68], as well as in homozygosity mapping [78].

However, if the two alleles are different at more than one SNP site, i.e. at least two SNPs are *heterozygous*, various haplotypes may be deduced from the same set of genotypes for an individual and hence the determination of the haplotypes will not be trivial. This situation is often encountered with outcrossing plants, e.g. those that are not self-compatible such as turnip (*Brassica rapa*), or vegetatively propagated plants such as potato (*Solanum tuberosum*) and banana (*Musa acuminata*) which are propagated by tubers and by division, respectively. In this situation, the phasing should usually be indirectly estimated by statistical methods based on observations and assumptions about haplotype diversity, segregation pattern and, if sequence data is available, sequencing information. I will return to this problem in Section 1.3.

1.2. Polyploidy and its consequences

The replication of DNA, the migration of the chromatids to opposite poles and the disjunction of homologous chromosomes (during anaphase I) or sister chromatids (during anaphase II) are not error free. While errors in the former process yield novel structural and allelic variants in the genome, errors in the latter can result in the formation of gametes with more than one copy of some or all of the chromosomes. In case such a gamete manages to get fused with another gamete, the result will be a zygote with more than 2 copies of some or all of the chromosomes. This cellular state is known as partial or complete *polyploidy*, respectively. Polyploidy can also occur by a failure of cell division after mitotic doubling during early cleavage of a fertilised egg as well as by polyspermy [69].

In humans, even partial polyploidy is usually lethal to the embryo, with a few exceptions such as trisomy 21, i.e. the possession of an extra copy of (part of) chromosome 21, also known as Down's syndrome. While this trisomy also enhances embryonic mortality, surviving individuals often suffer from congenital malformations and mental retardation and have a reduced life expectancy. Sex chromosome aneuploidies are usually less detrimental compared to autosomal anomalies and occur in at least 1 in 400 births [50], with well-known examples being monosomy X (45, X) or Turner syndrome [51], trisomy X (47, XXX) [85] and Klinefelter syndrome (with the most widespread karyotype being 47, XXY) [28]. Complete polyploidy has indeed very rarely been reported in live-born infants [81].

In general, polyploidy is considered a rare event within the animal kingdom [31], the reason of which remains a subject of debate. While in contrast to humans, many lower animals are equally viable with diploid and polyploid chromosome numbers (a well-known example of which is *Drosophila melanogaster* [62]), polyploidy seems to in-

terfere with the animal sex-differentiating process which is often based on the diploid mechanisms such as the determination of sex in "XX/XY" and "WZ/ZZ" forms [63]. As a result of this, triploid individuals ($2n = 3x$), which are considered usual intermediate bridges towards higher ploidy levels, are unlikely to express the sex-determining genes in the proper ratios and therefore become infertile or can be fertile as only male or as only female, i.e. with no genetically compatible mate [63]. Nevertheless, important examples of viable polyploidy are observed among animals, such as the Salmonidae family considered to be tetraploid ($2n = 4x$) [7], and several tetraploid species of African clawed frog (*Xenopus* spp.) [40].

Within the plant kingdom, polyploidy is very frequent and presumably a central feature in plant diversification and speciation [93], facilitated by self-compatibility and asexual propagation. The incidence of polyploid species within flowering plant genera (the angiosperms) has been estimated to be 30 to 80% [57], while polyploidy is considered rare in gymnosperms [38]. Wood *et al.* [93] report that 15% and 31% of speciation events in angiosperms and in ferns, respectively, are accompanied by an increase in the ploidy level.

In plants, polyploidization can also be initiated at the sporophytic stage because of mitotic irregularities in apical meristems [22], yielding polyploid chimeras with up to 70% polyploid cells in some species [27]. Some studies have even suggested that polyploidy is present in a majority of the somatic cells comprising the body of *Arabidopsis thaliana* ($2n = 2x = 10$), which is an important model plant [12]. This somatic polyploidy can also be, and has been, introduced artificially by radiative or chemical treatments in some species in order to help study polyploidy. Several sustainable polyploid cultivars in the *Poaceae* grass family have also been obtained by this so-called neopolyploidization process [18, 64]. It is worth mentioning that somatic polyploidy occurs also in animal tissues and is a distinguishing feature of anaplastic cancer cells [75, 96].

Among polyploids, it is important to recognise two distinct modes of chromosomal segregation: segregation through preferential bivalent pairing between specific chromosomes or *disomic* segregation versus segregation through the formation of multivalent chromosome crossings or non-preferential bivalent crossings of randomly pairing chromosomes, known as *polysomic* inheritance. The disomic mode is often observed when genomes of rather distant species have come together in a polyploid hybrid, a very well known example of which is bread wheat (*Triticum aestivum*), a hexaploid ($2n = 6x = 42$) composed of three homoeologous sets, i.e. A, B and D subgenomes each with 2 copies (AABBDD) [35, 56]. The polysomic mode is in contrast observed often in polyploids that consist of multiple copies of the same genome, with the well known example of cultivated potato (*S. tuberosum*), which is a tetraploid ($2n = 4x = 48$) with 4 copies of the same genome (AAAA) [59, 71].

In case the polyploidization has occurred due to duplication/multiplication of the same original species genome, the resulting polyploid is called an *autopolyploid*. In case two (or more) different genomes came together, most likely as a result of accidental or intended crossing between different species, the polyploid species is called an *allopolyploid*[1]. While disomic segregation and preferential bivalents (between the homologous chromosomes originating from the same species) are usually expected for allopolyploids, even in a well-known allopolyploid such as wheat loci have been de-

tected whose deletion enhances meiotic homoeologous pairing and hence non-disomic behaviour [77]. In contrast, polysomic inheritance patterns are usually expected for autopolyploids, such as potato. Segmental allopolyploids are formed by hybridisation of parental genomes with heterogeneous divergence, leading to a meiotic situation in which certain chromosomes or segments of chromosomes pair randomly as homologs, while others pair separately as homoeologs [80]. An example of this class is cultivated chrysanthemum (*Chrysanthemum*× *morifolium* Ramat.), for which evidence has been found of both disomic segregation [70] and polysomic segregation [70, 88] at different crosses and loci.

Sybenga (1996) presents segmental allopolyploidy as a transient state in the evolution of neopolyploids, and occasionally as a stable state in which some sets of chromosomes are well differentiated and behave as in allopolyploids while the others behave as in autopolyploids [83]. Bourke *et al.* [14] suggest that pairing affinities may vary along chromosome arms and demonstrate segmental allopolyploidy in a cross of tetraploid cut rose (*Rosa hybrida*). As we will see in Section 1.3, the segregation mode should be taken into account for the correct phasing estimation of heterozygous markers in polyploids.

1.2.1. Consequences of polyploidy in plants

The duplication of genes, which is the direct consequence of polyploidization, exposes a recent polyploid species to a period of instability after which a successful polyploid emerges adapted to survive and to compete with its diploid relatives [21]. During this adaptation process, some of the duplicated genes are either eliminated or silenced while others might change their function. This *diploidization* process may finally result in novel diploid species [65, 99], contributing to evolution and biodiversity [87]. The surviving duplicated genes can also offer extra plasticity to individual organisms by masking recessive deleterious mutations and by diversifying gene function and gene regulation [21, 89] that eventually yields novel phenotypes [2].

Besides the mentioned evolutionary advantages, polyploidization has major importance for plant breeding. The increased heterozygosity and the buffering of deleterious alleles, as well as the increment in plant organs such as tubercles, flowers and seeds (the so-called "gigas" effects) have traditionally led to thriving land races and in the modern era to increasingly improved cultivars [76]. The repression of meiotic division in polyploids with an odd number of chromosomes, such as banana ($2n = 3x = 33$), results in sterility and hence seedless fruits that have consumption appeal.

Table 1.1 lists several polyploid crops and their commercial use, among which staple crops such as potato (*S. tuberosum*) and wheat (*T. aestivum* and *T. turgidum*), as well as crops of industrial value such as cotton (*Gossypium hirsutum* and *G. barbadense*) and tobacco (*Nicotiana tabacum*). According to 2006 FAO statistics, 58.9% of the total area under cultivation (corresponding to 47.7% of the agricultural mass production) in the European union and the United Kingdom is occupied by polyploids, underlining the

[1] This nomenclature dates back to Kihara and Ono [39]: "Unter *Polyploidie* müssen wir heute zwei verschiedene Erscheinungen unterscheiden, nämlich die *Autopolyploidie* und die *Allopolyploidie*. Unter Autopolyploidie versteht man die Verdoppelung desselben Chromosomensatzes; unter Allopolyploidie die durch das Zusammenkommen verschiedener Chromosomensätze auf dem Wege der Bastardierung erfolgte Chromosomenvermehrung."

Table 1.1: Examples of commercially important polyploid crops and their classification.

Crop name	ploidy level and chromosome number	Dominant form	Commercial interest	Origin
Bread wheat (*Triticum aestivum* L.)	$6x = 42$	Allopolyploidy	Grain	Natural
Durum wheat (*Triticum turgidum* L.)	$4x = 28$	Allopolyploidy	Grain	Natural
Potato (*Solanum tuberosum* L.)	$4x = 48$	Autopolyploidy	Tuber	Natural
Leek (*Allium ampeloprasum* L.)	$4x = 32$	Autopolyploidy	Vegetable	Natural
Coffea (*Coffea arabica* L.)	$4x = 44$	Allopolyploidy	Beverage	Natural
Peanut (*Arachis hypogaea* L.)	$4x = 40$	Allopolyploidy	Nuts and Oil	Natural
Banana (*Musa acuminata* Colla)	$3x = 33$	Autopolyploidy	Fruit	Natural
Kiwi fruit (*Actinidia chinensis* Planch., *A. deliciosa* A.Chev.)	$4x = 116$, $6x = 174$	Autopolyploidy	Fruit	Natural
Sweet potato (*Ipomea batatas* (L.) Lam.)	$6x = 90$	Segmental Allopolyploidy	Tubercle	Natural
Alfalfa (*Medicago sativa* subsp. *sativa* L.)	$4x = 42$	Autopolyploidy	Forage	Natural
Rapeseed (*Brassica napus* L.)	$4x = 38$	Allopolyploidy	Oil	Natural
Tobacco (*Nicotiana tabacum* L.)	$8x = 56$	Allopolyploidy	Industrial	Natural
Strawberry (*Fragaria* × *ananassa* Duchesne)	$8x = 56$	Allopolyploidy	Fruit	Natural
Rye (*Secale cereale* L.)	$4x = 28$	Autopolyploidy	Grain and Forage	Synthetic
Ryegrass (*Lolium perenne* L.)	$4x = 28$	Autopolyploidy	Forage	Synthetic
Alstroemeria (*Alstroemeria* × *hybrida* L.)	$3x = 24$, $4x = 32$	Autopolyploidy	Ornamental	Natural
Sugarcane (*Saccharum officinarum* L.)	$8x = 80$	Allopolyploidy	Industrial	Natural
Cut rose, Garden rose (*Rosa* × *hybrida* L.)	$4x = 28$	Segmental Allopolyploidy	Ornamental	Natural
Cotton (*Gossypium hirsutum* L., *G. barbadense* L.)	$4x = 52$	Allopolyploidy	Industrial	Natural

importance of polyploids in present day agriculture.

1.3. The problem of haplotype estimation from sequence data

Ideally, the phasing of SNP markers will be directly known if one can separate chromosomes before genotyping or sequencing. However, current approaches that allow for this are expensive and labor intensive, hence low-throughput [26, 34, 55]. Therefore, haplotype estimation algorithms are often used to indirectly obtain the phasing from the unphased high-throughput SNP array or sequence-based genotypes [17].

Phasing methods that only consider genotypes usually aim to maximise the likelihood of the observed genotypes in a population using a likelihood function that relates an *a priori* set of compatible haplotypes to the observed population allele frequencies [1, 19, 25, 82]. These methods can usually handle only a limited number of markers and work best with large populations. Besides, they are mostly designed for diploid populations. Only recently methods have been developed for polyploids, such as TetraOrigin [101] (targeting recombinant markers in tetraploid bi-parental F1-populations with a known linkage map) and HaplotypeR (unpublished, Voorrips *et al.*), which are bound by similar limitations to a greater extent compared to the diploid algorithms.

With the advent of shotgun and next generation sequencing technologies, individual phasing has become a possibility by using the sequence reads of a single individual. Current high-throughput sequencing technologies shatter the DNA into small pieces ranging in length from a few hundred base pairs, e.g. with Illumina sequencing-by-synthesis technology, to several kilo base pairs, e.g. with nanopore sequencers. Thus, one obtains a large number of sequence reads each originating from a random position along the target DNA, whose average length and depth of coverage depend on the employed library preparation and sequencing approaches. To retrieve the haplotypes, it is therefore necessary to first align the obtained reads according to their genomic coordinates, either

by *de novo* assembly [48, 54] or by mapping the reads to a reference sequence [46, 47], and to detect genomic variations and determine the genotypes by comparing the bases called by the overlapping reads at the same genomic position [24, 29].

After determining the coordinates of the reads and detecting the alleles within the reads, the reads can be assigned to the k haplotypes of an individual (with k the ploidy level) using the fact that the reads of the same haplotype must contain the same alleles at their overlapping sites. Considering only bi-allelic SNPs and discarding the homozygous SNP sites of an individual (as all the haplotypes contain the same allele at these sites), one has to deal with two complementary haplotypes for a diploid, i.e. each haplotype can be derived from the other by converting the reference/alternate alleles of the other haplotype to alternate/reference alleles. This drastically simplifies the problem and theoretically allows perfect retrieval of the haplotypes. In polyploids, however, the k haplotypes of an individual need not to be complementary, as a haplotype can have a dosage up to $k-1$ (discarding the dosage k which corresponds to a completely homozygous region). In reality, one has also to deal with sequencing errors, the rate of which varies from around 0.1-2% of the called bases for Illumina to around 10-30% for Oxford Nanopore Technology (ONT). The read alignment, variant calling and genotyping steps are of course influenced by the sequencing errors, as well as by the sequencing depth, read length and the complexity of the target DNA. These errors can result in the false detection of more distinct haplotypes than actually present and hamper the naive phasing explained above. Therefore, it is necessary to *estimate* the phasing using optimisation approaches that deal with ploidy levels $k > 2$, as well as with sequencing errors.

In the remainder of this section, I introduce several read-based phasing approaches which have been so far developed, for both diploids (Section 1.3.1) and polyploids (Section 1.3.2), using the notations and concepts introduced for the diploids also in the polyploid case.

1.3.1. Haplotype estimation for diploids

As only those DNA fragments that include at least two heterozygous SNP sites contain phasing information, homozygous SNP sites and the fragments containing just a single SNP can be discarded in the outset from the data. Assuming to have m aligned DNA fragments over a region including n heterozygous SNP sites, the sequence information can be stored in the SNP-fragment matrix $M_{m\times n}$, in which each row represents a fragment and each column represents a heterozygous site. The elements of each fragment could be labeled by the alphabet $0/1/2/3$ and $-$, in which 0 and $1/2/3$ represent the wild and mutant alleles, respectively, and '$-$' represents the uncalled bases or *holes*.

In case a paired-end or mate pair sequencing library has been used, each fragment should consist of two sequence reads split by a gap corresponding to consecutive holes (Figure 1.2). Recently, 10X Genomics has released its Chromium system that piggybacks on Illumina technology to generate long fragments (with a typical average length of 50 kb) composed of many short sequence reads with holes in between. It is also possible to have holes within a single sequence read as a result of discarding low quality base calls (Figure 1.2).

With the above definition, two fragments f_i and f_j are said to be in conflict at position s if:

$$f_i(s) \neq f_j(s),\ f_i(s) \neq -,\ f_j(s) \neq - \tag{1.1}$$

which means that both f_i and f_j are non-missing, but contain different alleles at position s. Accordingly, two fragments are *conflict-free* if they are not in conflict at any overlapping position. The phasing algorithm is thus based on dividing the rows of M into $k = 2$ conflict-free groups, so that the haplotypes can be determined by the consensus of the reads within each group. With no error in the reads and only bi-allelic SNPs allowed (reducing the alphabet to $\{0, 1, -\}$), a trivial clustering algorithm could accomplish the task in $O(mn)$ time (provided that the sequencing coverage is enough so that at least one fragment covers each of the $n-1$ adjacent SNP pairs) and M is called *feasible*. In presence of errors, however, the aforementioned partition would not be possible and therefore M would be *infeasible*. An example of a feasible and an infeasible SNP-fragment matrix is given in Figure 1.2.

A practical phasing algorithm must therefore (minimally) manipulate the fragments in order to make M feasible. As intuitively inspired by Figure 1.2, to make M feasible one may think of: (1) discarding fragments (corresponding to omitting rows in M), (2) discarding SNPs (corresponding to omitting columns in M), or (3) flipping alleles within the

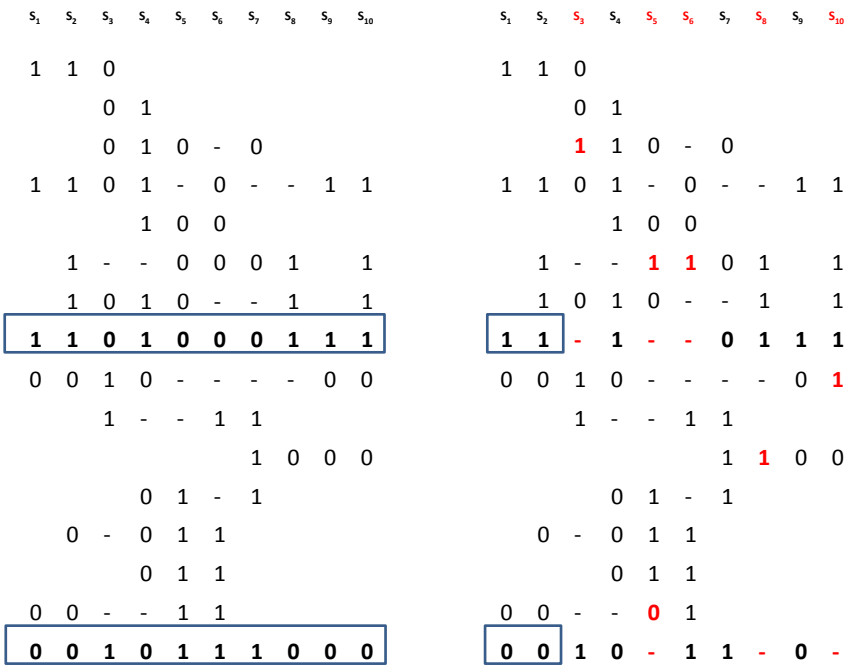


Figure 1.2: A feasible (left) and an infeasible SNP-fragment matrix (right). Sequencing errors are denoted by red letters. Consensus haplotypes are given in boxes. For the infeasible matrix, only the first two SNPs have been phased.

reads (corresponding to converting a presumably wrong letter in M to another letter). By considering a cost function for any of the mentioned manipulations, haplotype reconstruction is translated into an exponentially bounded optimisation problem of making M feasible at minimum cost.

This formulation has been formally treated in Lancia *et al.* [44], Rizzi *et al.* [73] and Lippert *et al.* [52] by building fragment conflict graphs or SNP conflict graphs corresponding to M and then applying graph theoretic algorithms to obtain perfect graphs at a minimum pruning cost. Specifically, they discuss Minimum Fragment Removal (MFR), Minimum SNP removal (MSR) and Minimum Error Correction or Minimum Letter Flip (MEC or MLF) approaches, corresponding to the three types of manipulation discussed above.

Lancia *et al.* [44] prove that the MFR and MSR are polynomially solvable for *gapless* fragments, i.e. fragments containing no holes, with only 0 and 1 alleles corresponding to bi-allelic SNPs, but are NP-hard in general and do not have a polynomial solution. Nonetheless, the exact algorithms they suggest are bounded by too high-degree polynomials even for the usually unrealistic gapless case, and are therefore by no means practical [73]. Approximate algorithms are introduced by Rizzi *et al.* [73], which are polynomial in terms of m and n, but exponential in terms of the maximum number of holes making them impractical for paired end and mate pair technologies. Moreover, ignoring a whole fragment or SNP could often cause too much information loss, unless one deals with a very low quality fragment or SNP. However, such fragments and SNPs are usually discarded by preprocessing of the sequence reads or during the alignment and variant calling steps. Wang *et al.* [90] discuss the more applicable MEC criterion, and prove that the exact algorithm is NP-hard, even for the gapless case. Zhao *et al.* [100] investigate the weighted version of MLF (WMLF) where a weight matrix is assigned to M by, for instance, weighing each fragment letter by its associated base calling quality (Phred score) and prove the NP-hardness of the exact solution.

Iterative hill-climbing and heuristic procedures have therefore been the premise of practical phasing methods. Examples are found in complete weighted MLF (CWMLF), in which fragment and SNP removal operations are applied in addition to weighted letter flip to improve the clustering of the fragments [100], genetic algorithm (GA) based MEC [90], MEC models with two distance functions [92], MEC using an iterative max-cut algorithm on the SNP graph with weighted edges that reflect the difference between the number of reads consistent and inconsistent with the current phasing between the vertices of the edge [9], particle swarm optimisation (PSO) [94] and self organising map (SOM) [95] algorithms for the MEC model and a greedy fragment clustering algorithm with refinement based on a fuzzy conflict graph (FastHap) [58]. Wang *et al.* [91] suggest a Markov Chain (MC) model for assembling haplotypes, which follows a dynamic programming (DP) approach to extend d-meric haplotypes by choosing the most probable extension whose probability is calculated from the SNP-matrix and the d-letter overlap between consecutive d-mers. The sequencing errors are here implicitly dealt with by preferring the most frequent, i.e. compatible, extensions. A more general graphical approach is discussed in Kuleshov [42], applying the max-sum message passing algorithm to the two-dimensional Bayesian network that relates the reads to putative haplotypes. Bansal *et al.* [10] suggest a Metropolis algorithm called HASH to obtain the empirical

posterior distribution of the haplotypes from the SNP-matrix and the base calling qualities. As the exhaustive Monte Carlo update over the set of all possible haplotypes is computationally prohibitive, they perform a local search by modifying the haplotypes at the most conflicting SNP sites at each iteration, followed by the Metropolis accept-reject rule. The set of these conflicting SNP sites is chosen by applying a min-cut partitioning algorithm to the SNP graph (weighted in a modified similar approach as in Bansal and Bafna [9]) and is updated regularly during the run of the Monte Carlo algorithm.

I end this section by discussing in some detail two diploid phasing algorithms, HapCompass [3] and SDhaP [23], that have been generalised to the polyploid case (Section 1.3.2). Aguiar and Istrail [3] introduce a Minimum Edge Removal (MER) algorithm, based on resolving conflicting cycles in a special SNP-graph, called the Compass graph, by removing a set of edges with a total weight closest to zero. Their algorithm would correspond to discarding parts of the erroneous fragments to make M feasible. The Compass graph $G = (V, E, W)$ has the heterozygous sites as nodes, $V = \{s_1, s_2, ..., s_n\}$, and each edge e_{ij} is weighted by the number of fragments $\mathscr{F}$ that cover both s_i and s_j and suggest a $\begin{smallmatrix} & s_i & s_j \\ h_1 & 1 & 1 \\ h_2 & 0 & 0 \end{smallmatrix}$ phasing, minus the number of fragments that suggest a $\begin{smallmatrix} & s_i & s_j \\ h_1 & 1 & 0 \\ h_2 & 0 & 1 \end{smallmatrix}$ phasing:

$$w_{ij} = \sum_{f \in \mathscr{F}} \mathscr{C}(f, i, j)$$
$$\mathscr{C}(f, i, j) = \begin{cases} 1 & \big(f(i) = 1 \wedge f(j) = 1\big) \vee \big(f(i) = 0 \wedge f(j) = 0\big) \\ -1 & \big(f(i) = 1 \wedge f(j) = 0\big) \vee \big(f(i) = 0 \wedge f(j) = 1\big) \end{cases} \tag{1.2}$$

Obviously, zero weighted edges are indecisive about the phasing between the two sites, and are therefore omitted from the graph. Every path in the Compass graph corresponds to a phasing for its end nodes, and a Compass graph is called *happy* if all paths between s_i and s_j suggest the same phase. Each pair of paths between s_i and s_j forms a simple cycle in the Compass graph, and it is easy to show that a cycle is conflicting, i.e. indicates conflicting phasings between s_i and s_j, if and only if the number of its negative edges, i.e. e_{ij} with $w_{ij} < 0$ as calculated by Equation 1.2, is odd. The HapCompass algorithm removes a minimum weight set of edges through an iterative local search in the cycle basis of the Compass graph in order to make the graph happy, and the maximum spanning tree of the resulting happy Compass graph gives the desired phasing for the n SNPs.

Das and Vikalo [23] consider the MEC optimisation problem as an NP-complete quadratic integer programming problem, which can be reformulated as a standard semidefinite problem which is approximately solvable in polynomial time using Goemans-Williamson algorithm [30]. They implement an efficient version of this algorithm, called SDhaP, by factorising the sparse SNP-fragment matrix and refining the final solution by greedy flipping of the haplotype alleles to further reduce the MEC, if possible.

1.3.2. Haplotype estimation for polyploids

The estimation of haplotypes from sequence data is much more complicated for polyploids than for diploids. Due to the nature of polyploid genomes, even the read alignment and genotype calling steps are challenging compared to the diploid case. Tang *et al.* [84] implemented a naive haplotype extension method based on minimum overlapping and clustering requirements in order to improve the quality of SNP calling from

de novo assembled contigs, which approach was extended by Nijveen *et al.* [67] to accommodate short NGS fragments. However, their haplotyping scheme is heuristic and inefficient, in the sense that it discards many fragments and is sensitive to the order of the input fragments. The most well-known optimisation based polyploid algorithms developed so far include polyploid HapCompass [4, 5], HapTree [13], polyploid SDhaP [23] and H-PoP [97] which I present here.

Aguiar and Istrail [4] extend their graphical haplotype estimation approach for diploids [3] by introducing a modified version of the Compass graph (Section 1.3.1) that has k nodes for each SNP site s_i in a k-ploid, corresponding to the k alleles. To represent the k haplotypes, k edges are drawn between each SNP pair's nodes (s_i, s_j) weighted according to a maximum likelihood model for the phasing between s_i and s_j conditional on the sequence fragments that cover s_i and s_j. A global minimum weighted edge removal (MWER) algorithm is applied to detect and eliminate edges with conflicting phasing information with the aid of auxiliary chain graphs. Each chain graph takes a set of SNPs that make a cycle in the Compass graph and detects phasing conflicts within the cycle, if present. The MWER algorithm then tries to resolve the conflicts by eliminating a number of edges with minimum total weight corresponding to the phasings the least likely conditional on the sequence fragments. At the end, HapCompass reports the most likely haplotypes over the complete set of n SNPs from the conflict-free Compass graph, by finding k disjoint maximum spanning trees through an efficient greedy algorithm [5]. It is worth noting that the current version of HapCompass can handle multi-allelic SNPs for polyploids, although the original algorithm is presented for bi-allelic markers. Besides, HapCompass has been popular compared to other methods, as it accepts input in the conventional alignment (BAM) and variant calling (VCF) formats and it has a user friendly command line interface [37].

The HapTree algorithm developed by Berger *et al.* [13] extends the phasing SNP by SNP from s_1 to s_n, keeping only the most likely phasings at each extension step up to s_i when proceeding to include s_{i+1} in the phasing. An ordered tree is used to represent the extensions, in which the nodes at level i correspond to the phasing extensions that include $s_1, s_2, \ldots, s_i$ and the degree of each node equals the number of extensions possible for its associated phasing. To obtain the phasing of n SNPs in polynomial time, HapTree performs a greedy *branching and pruning* step at each extension level by removing child vertices whose extension probabilities fall below a preset threshold $0 \leqslant \rho \leqslant 1$ (branching) as well as children whose relative extension probability with respect to the maximum extension probability at the current level is below a preset threshold $0 \leqslant \kappa \leqslant 1$ (pruning). The current version of HapTree handles only bi-allelic SNPs, although extensions to multi-allelic markers are also discussed by Berger *et al.* [13].

The polyploid version of SDhaP [23] starts with random phasings (considered points in the metric phasing space), and finds the local optimum of MEC by implementing a gradient-descent method. To this purpose and in order to overcome the singularities caused by the countable phasing space, the phasing space is reformulated by representing each phasing as a k-simplex in a connected n-dimensional space subject to semidefinite constraints. To obtain the MEC phasing, a greedy best solution is first obtained by relaxing the semidefinite constraints. The projection of each sequence fragment on the k-simplex representing this preliminary phasing assigns it to one of its vertices (based

on minimum distance), and the haplotype corresponding to each vertex is determined by the consensus of its associated sequence fragments. The estimated haplotypes are subsequently refined by greedy flipping of their alleles to further reduce the MEC, if possible. Also, SDhaP makes use of the sparseness of the SNP-fragment matrix to provide a fast and efficient implementation in polynomial time.

The last method that I discuss here, H-PoP developed by Xie *et al.* [97], tries to partition the sequence fragments into k groups, so that the similarity is maximised between the fragments within each group while the difference is maximised between the fragments that are assigned to different groups. Xie *et al.* [97] introduce a heuristic dynamic-programming algorithm called Polyploid Balanced Optimal Partition (PBOP) to obtain the k groups. The haplotype of each group is afterwards determined by consensus.

1.4. Outline of this thesis

The research described in this thesis is motivated by several research questions focusing on sequence-based haplotype estimation of polyploid crops. As various methods have been proposed to estimate the haplotypes from noisy sequence data, the first challenge is to develop a uniform framework to evaluate and compare these methods in different situations. Considering the single individual nature of most of the so far proposed methods, another main challenge is to develop algorithms for sequence-based phasing in a population.

In **Chapter 2**, we develop a simulation pipeline for polyploid genomes with given heterozygosity rated and dosage distributions, from which sequencing data can be generated using the available technology specific sequence read-simulators. We use this pipeline to evaluate several state-of-the-art single individual haplotyping (SIH) algorithms by comparing their estimates to the original haplotypes. We show that the conditional log-likelihood is a better score for polyploid haplotyping compared to the MEC, at the cost of computational complexity, and that all of the SIH methods suffer performance issues at ploidy levels higher than four ($k > 4$), as well as at low sequencing depths and with short-length DNA fragments.

In **Chapter 3**, we introduce *TriPoly*, a novel haplotyping approach for polyploid trios. We compare TriPoly to SIH methods and demonstrate its better performance with both short-read and long-read sequencing, especially at low sequencing depths. For this comparison, we extend the simulation pipeline of Chapter 2 to generate offspring from simulated parental genomes, taking a simplified model of meiotic recombination and chromosome segregation into account. In **Chapter 4**, we introduce a family-based approach, called *PopPoly*, that specifically targets moderate to large F1-families and short-read sequence data. Through simulations, we show that PopPoly outperforms SIH methods and TriPoly provided that the population size is sufficient (more than 5 offspring, say). Besides, PopPoly improves the genotypes obtained from sequence data by conventional tools (such as FreeBayes [29]) that do not consider pedigree information. We also apply both TriPoly and PopPoly to different F1-populations of tetraploid potato.

In **Chapter 5**, we address the question of extracting the partial phasing information in the reads to obtain haplotype marker scores for genetic association analysis. We introduce a latent Poisson model for the read count of each haplotype containing a few SNP markers, the rate of which we estimate by the expectation-maximisation (EM) method.

This approach has been implemented in the command line tool *AcroPoly*. Finally, the thesis is concluded by **Chapter 6** with a discussion of the main findings and the current haplotyping perspective for polyploids.

References

[1] Abecasis, G. R., Cherny, S. S., Cookson, W. O., and Cardon, L. R. (2002). Merlin—rapid analysis of dense genetic maps using sparse gene flow trees. *Nature Genetics*, **30**(1), 97.

[2] Adams, K. L. and Wendel, J. F. (2005). Polyploidy and genome evolution in plants. *Current Opinion in Plant Biology*, **8**(2), 135–141.

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

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

[5] Aguiar, D., Wong, W. S., and Istrail, S. (2014). Tumor haplotype assembly algorithms for cancer genomics. In *Pacific Symposium on Biocomputing*, page 3. World Scientific.

[6] Akey, J., Jin, L., and Xiong, M. (2001). Haplotypes vs single marker linkage disequilibrium tests: what do we gain? *European Journal of Human Genetics*, **9**(4), 291.

[7] Allendorf, F. W. and Thorgaard, G. H. (1984). Tetraploidy and the evolution of salmonid fishes. In *Evolutionary Genetics of Fishes*, pages 1–53. Springer.

[8] Balasubramanian, S., Schwartz, C., Singh, A., Warthmann, N., Kim, M. C., Maloof, J. N., Loudet, O., Trainer, G. T., Dabi, T., Borevitz, J. O., *et al.* (2009). QTL mapping in new Arabidopsis thaliana advanced intercross-recombinant inbred lines. *PloS One*, **4**(2), e4318.

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

[10] Bansal, V., Halpern, A. L., Axelrod, N., and Bafna, V. (2008). An mcmc algorithm for haplotype assembly from whole-genome sequence data. *Genome research*, **18**(8), 1336–1346.

[11] Bell, C. G., Finer, S., Lindgren, C. M., Wilson, G. A., Rakyan, V. K., Teschendorff, A. E., Akan, P., Stupka, E., Down, T. A., Prokopenko, I., *et al.* (2010). Integrated genetic and epigenetic analysis identifies haplotype-specific methylation in the FTO type 2 diabetes and obesity susceptibility locus. *PloS One*, **5**(11), e14040.

[12] Bennett, M. D. (2004). Perspectives on polyploidy in plants–ancient and neo. *Biological Journal of the Linnean Society*, **82**(4), 411–423.

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

[14] Bourke, P. M., Arens, P., Voorrips, R. E., Esselink, G. D., Koning-Boucoiran, C. F., van't Westende, W. P., Santos Leonardo, T., Wissink, P., Zheng, C., Van Geest, G., *et al.* (2017). Partial preferential chromosome pairing is genotype dependent in tetraploid rose. *The Plant Journal*, **90**(2), 330–343.

[15] Bourke, P. M., Voorrips, R. E., Visser, R. G., and Maliepaard, C. (2018). Tools for genetic studies in experimental populations of polyploids. *Frontiers in Plant Science*, **9**.

[16] Browning, B. L. and Browning, S. R. (2009). A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *The American Journal of Human Genetics*, **84**(2), 210–223.

[17] Browning, S. R. and Browning, B. L. (2011). Haplotype phasing: existing methods and new developments. *Nature Reviews Genetics*, **12**(10), 703–714.

[18] Chae, W. B., Hong, S. J., Gifford, J. M., Rayburn, A. L., Widholm, J. M., and Juvik, J. A. (2013). Synthetic polyploid production of Miscanthus sacchariflorus, Miscanthus sinensis, and Miscanthus x giganteus. *Gcb Bioenergy*, **5**(3), 338–350.

[19] Clark, V. J., Metheny, N., Dean, M., and Peterson, R. J. (2001). Statistical estimation and pedigree analysis of CCR2-CCR5 haplotypes. *Human Genetics*, **108**(6), 484–493.

[20] Collard, B. C. and Mackill, D. J. (2008). Marker-assisted selection: an approach for precision plant breeding in the twenty-first century. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **363**(1491), 557–572.

[21] Comai, L. (2005). The advantages and disadvantages of being polyploid. *Nature Reviews Genetics*, **6**(11), 836.

[22] D'Amato, F. (1964). Endopolyploidy as a factor in plant tissue development. *Caryologia*, **17**(1), 41–52.

[23] Das, S. and Vikalo, H. (2015). SDhaP: haplotype assembly for diploids and polyploids via semi-definite programming. *BMC Genomics*, **16**(1), 260.

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

[25] Excoffier, L. and Slatkin, M. (1995). Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. *Molecular Biology and Evolution*, **12**(5), 921–927.

[26] Fan, H. C., Wang, J., Potanina, A., and Quake, S. R. (2011). Whole-genome molecular haplotyping of single cells. *Nature Biotechnology*, **29**(1), 51.

[27] Frisch, B. and Nagl, W. (1979). Patterns of endopolyploidy and 2C nuclear DNA content (Feulgen) in Scilla (Liliaceae). *Plant Systematics and Evolution*, **131**(3-4), 261–276.

[28] Frühmesser, A. and Kotzot, D. (2011). Chromosomal variants in Klinefelter syndrome. *Sexual Development*, **5**(3), 109–123.

[29] Garrison, E. and Marth, G. (2012). Haplotype-based variant detection from short-read sequencing. *arXiv preprint arXiv:1207.3907 [q-bio.GN]*.

[30] Goemans, M. X. and Williamson, D. P. (1995). Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming. *Journal of the ACM (JACM)*, **42**(6), 1115–1145.

[31] Gregory, T. R. and Mable, B. K. (2005). Polyploidy in animals. In *The Evolution of the Genome*, pages 427–517. Elsevier.

[32] Guo, S., Diep, D., Plongthongkum, N., Fung, H.-L., Zhang, K., and Zhang, K. (2017).

Identification of methylation haplotype blocks aids in deconvolution of heterogeneous tissue samples and tumor tissuei-of-origin mapping from plasma DNA. *Nature Genetics*, **49**(4), 635.

[33] Hackett, C. A., McLean, K., and Bryan, G. J. (2013). Linkage analysis and QTL mapping using SNP dosage data in a tetraploid potato mapping population. *PLoS One*, **8**(5), e63939.

[34] Huang, M., Tu, J., and Lu, Z. (2017). Recent advances in experimental whole genome haplotyping methods. *International Journal of Molecular Sciences*, **18**(9), 1944.

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

[36] Jansen, R. C., Jannink, J.-L., and Beavis, W. D. (2003). Mapping quantitative trait loci in plant breeding populations. *Crop Science*, **43**(3), 829–834.

[37] Kamneva, O. K., Syring, J., Liston, A., and Rosenberg, N. A. (2017). Evaluating allopolyploid origins in strawberries (Fragaria) using haplotypes generated from target capture sequencing. *BMC Evolutionary Biology*, **17**(1), 180.

[38] Khoshoo, T. (1959). Polyploidy in gymnosperms. *Evolution*, **13**(1), 24–39.

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

[40] Kobel, H. R. and Du Pasquier, L. (1986). Genetics of polyploid Xenopus. *Trends in Genetics*, **2**, 310–315.

[41] Kover, P. X., Valdar, W., Trakalo, J., Scarcelli, N., Ehrenreich, I. M., Purugganan, M. D., Durrant, C., and Mott, R. (2009). A multiparent advanced generation inter-cross to fine-map quantitative traits in Arabidopsis thaliana. *PLoS Genetics*, **5**(7), e1000551.

[42] Kuleshov, V. (2014). Probabilistic single-individual haplotyping. *Bioinformatics*, **30**(17), i379–i385.

[43] LaFramboise, T. (2009). Single nucleotide polymorphism arrays: a decade of biological, computational and technological advances. *Nucleic Acids Research*, **37**(13), 4181–4193.

[44] Lancia, G., Bafna, V., Istrail, S., Lippert, R., and Schwartz, R. (2001). SNPs problems, complexity, and algorithms. In *Algorithms—ESA 2001*, pages 182–193. Springer.

[45] Lande, R. and Thompson, R. (1990). Efficiency of marker-assisted selection in the improvement of quantitative traits. *Genetics*, **124**(3), 743–756.

[46] Langmead, B. and Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, **9**(4), 357.

[47] Li, H. and Durbin, R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*, **25**(14), 1754–1760.

[48] Li, R., Zhu, H., Ruan, J., Qian, W., Fang, X., Shi, Z., Li, Y., Li, S., Shan, G., Kristiansen, K., *et al.* (2009). De novo assembly of human genomes with massively parallel short read sequencing. *Genome Research*.

[49] Li, Y., Willer, C. J., Ding, J., Scheet, P., and Abecasis, G. R. (2010). MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genetic Epidemiology*, **34**(8), 816–834.

[50] Linden, M. C., Bender, B. G., and Robinson, A. (1995). Sex chromosome tetrasomy and pentasomy. *Pediatrics*, **96**(4), 672–682.

[51] Lippe, B. (1991). Turner syndrome. *Endocrinology and Metabolism Clinics of North America*, **20**(1), 121–152.

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

[53] Lu, Y., Zhang, S., Shah, T., Xie, C., Hao, Z., Li, X., Farkhari, M., Ribaut, J.-M., Cao, M., Rong, T., *et al.* (2010). Joint linkage–linkage disequilibrium mapping is a powerful approach to detecting quantitative trait loci underlying drought tolerance in maize. *Proceedings of the National Academy of Sciences*, **107**(45), 19585–19590.

[54] Luo, R., Liu, B., Xie, Y., Li, Z., Huang, W., Yuan, J., He, G., Chen, Y., Pan, Q., Liu, Y., *et al.* (2012). SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *Gigascience*, **1**(1), 18.

[55] Ma, L., Xiao, Y., Huang, H., Wang, Q., Rao, W., Feng, Y., Zhang, K., and Song, Q. (2010). Direct determination of molecular haplotypes by chromosome microdissection. *Nature Methods*, **7**(4), 299.

[56] Marcussen, T., Sandve, S. R., Heier, L., Spannagl, M., Pfeifer, M., Jakobsen, K. S., Wulff, B. B., Steuernagel, B., Mayer, K. F., Olsen, O.-A., *et al.* (2014). Ancient hybridizations among the ancestral genomes of bread wheat. *Science*, **345**(6194), 1250092.

[57] Masterson, J. (1994). Stomatal size in fossil plants: evidence for polyploidy in majority of angiosperms. *Science*, **264**(5157), 421–424.

[58] Mazrouee, S. and Wang, W. (2014). Fasthap: fast and accurate single individual haplotype reconstruction using fuzzy conflict graphs. *Bioinformatics*, **30**(17), i371–i378.

[59] Mendoza, H. and Haynes, F. (1974). Genetic basis of heterosis for yield in the autotetraploid potato. *Theoretical and Applied Genetics*, **45**(1), 21–25.

[60] Meuwissen, T. H. and Goddard, M. E. (2001). Prediction of identity by descent probabilities from marker-haplotypes. *Genetics Selection Evolution*, **33**(6), 605.

[61] Mohan, M., Nair, S., Bhagwat, A., Krishna, T., Yano, M., Bhatia, C., and Sasaki, T. (1997). Genome mapping, molecular markers and marker-assisted selection in crop plants. *Molecular Breeding*, **3**(2), 87–103.

[62] Morgan, L. (1925). Polyploidy in Drosophila melanogaster with two attached X chromosomes. *Genetics*, **10**(2), 148–178.

[63] Muller, H. (1925). Why polyploidy is rarer in animals than in plants. *The American Naturalist*, **59**(663), 346–353.

[64] Myers, W. (1939). Colchicine induced tetraploidy in perennial ryegrass: Lolium perenne L. *Journal of Heredity*, **30**(11), 499–504.

[65] Naganowska, B., Wolko, B., Śliwińska, E., and Kaczmarek, Z. (2003). Nuclear DNA content variation and species relationships in the genus Lupinus (Fabaceae). *Annals of Botany*, **92**(3), 349–355.

[66] N'Diaye, A., Haile, J. K., Cory, A. T., Clarke, F. R., Clarke, J. M., Knox, R. E., and Pozniak, C. J. (2017). Single marker and haplotype-based association analysis of semolina and pasta colour in elite durum wheat breeding lines using a high-density consensus

map. *PloS One*, **12**(1), e0170941.

[67] Nijveen, H., van Kaauwen, M., Esselink, D. G., Hoegen, B., and Vosman, B. (2013). QualitySNPng: a user-friendly SNP detection and visualization tool. *Nucleic Acids Research*, page gkt333.

[68] Ogawa, D., Nonoue, Y., Tsunematsu, H., Kanno, N., Yamamoto, T., and Yonemaru, J.-i. (2018). Discovery of QTL alleles for grain shape in the Japan-MAGIC rice population using haplotype information. *G3: Genes, Genomes, Genetics*, **8**(11), 3559–3565.

[69] Otto, S. P. and Whitton, J. (2000). Polyploid incidence and evolution. *Annual Review of Genetics*, **34**(1), 401–437.

[70] Park, S. K., Arens, P., Esselink, D., Lim, J. H., and Shin, H. K. (2015). Analysis of inheritance mode in chrysanthemum using EST-derived SSR markers. *Scientia Horticulturae*, **192**, 80–88.

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

[72] Rickert, A. M., Kim, J. H., Meyer, S., Nagel, A., Ballvora, A., Oefner, P. J., and Gebhardt, C. (2003). First-generation snp/indel markers tagging loci for pathogen resistance in the potato genome. *Plant Biotechnology Journal*, **1**(6), 399–410.

[73] Rizzi, R., Bafna, V., Istrail, S., and Lancia, G. (2002). Practical algorithms and fixed-parameter tractability for the single individual SNP haplotyping problem. In *Algorithms in Bioinformatics*, pages 29–43. Springer.

[74] Rosyara, U. R., De Jong, W. S., Douches, D. S., and Endelman, J. B. (2016). Software for genome-wide association studies in autopolyploids and its application to potato. *The Plant Genome*, **9**(2).

[75] Sandberg, A. A., Ishihara, T., Moore, G. E., and Pickren, J. W. (1963). Unusually high polyploidy in a human cancer. *Cancer*, **16**(10), 1246–1254.

[76] Sattler, M. C., Carvalho, C. R., and Clarindo, W. R. (2016). The polyploidy and its key role in plant breeding. *Planta*, **243**(2), 281–296.

[77] Sears, E. R. (1977). Genetics society of canada award of excellence lecture an induced mutant with homoeologous pairing in common wheat. *Canadian Journal of Genetics and Cytology*, **19**(4), 585–593.

[78] Seelow, D., Schuelke, M., Hildebrandt, F., and Nürnberg, P. (2009). HomozygosityMapper—an interactive approach to homozygosity mapping. *Nucleic Acids Research*, **37**(suppl_2), W593–W599.

[79] Simko, I., Haynes, K. G., Ewing, E. E., Costanzo, S., Christ, B. J., and Jones, R. W. (2004). Mapping genes for resistance to Verticillium alboi-atrum in tetraploid and diploid potato populations using haplotype association tests and genetic linkage analysis. *Molecular Genetics and Genomics*, **271**(5), 522–531.

[80] Spoelhof, J. P., Soltis, P. S., and Soltis, D. E. (2017). Pure polyploidy: closing the gaps in autopolyploid research. *Journal of Systematics and Evolution*, **55**(4), 340–352.

[81] Stefanova, I., Jenderny, J., Kaminsky, E., Mannhardt, A., Meinecke, P., Grozdanova, L., and Gillessen-Kaesbach, G. (2010). Mosaic and complete tetraploidy in live-born infants: two new patients and review of the literature. *Clinical Dysmorphology*, **19**(3), 123–127.

[82] Stephens, M., Smith, N. J., and Donnelly, P. (2001). A new statistical method for

haplotype reconstruction from population data. *The American Journal of Human Genetics*, **68**(4), 978–989.

[83] Sybenga, J. (1996). Chromosome pairing affinity and quadrivalent formation in polyploids: do segmental allopolyploids exist? *Genome*, **39**(6), 1176–1184.

[84] Tang, J., Vosman, B., Voorrips, R. E., van der Linden, C. G., and Leunissen, J. A. (2006). QualitySNP: a pipeline for detecting single nucleotide polymorphisms and insertions/deletions in EST data from diploid and polyploid species. *BMC Bioinformatics*, **7**(1), 438.

[85] Tartaglia, N. R., Howell, S., Sutherland, A., Wilson, R., and Wilson, L. (2010). A review of trisomy X (47, XXX). *Orphanet Journal of Rare Diseases*, **5**(1), 8.

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

[87] Van de Peer, Y., Mizrachi, E., and Marchal, K. (2017). The evolutionary significance of polyploidy. *Nature Reviews Genetics*, **18**(7), 411.

[88] Van Geest, G., Voorrips, R. E., Esselink, D., Post, A., Visser, R. G., and Arens, P. (2017). Conclusive evidence for hexasomic inheritance in chrysanthemum based on analysis of a 183k SNP array. *BMC Genomics*, **18**(1), 585.

[89] Wang, J., Tian, L., Lee, H.-S., Wei, N., Jiang, H., Watson, B., Madlung, A., Osborn, T., Doerge, R., Comai, L., *et al.* (2005a). Genome-wide non-additive gene regulation in Arabidopsis allotetraploids. *Genetics*.

[90] Wang, R.-S., Wu, L.-Y., Li, Z.-P., and Zhang, X.-S. (2005b). Haplotype reconstruction from SNP fragments by minimum error correction. *Bioinformatics*, **21**(10), 2456–2462.

[91] Wang, R.-S., Wu, L.-Y., Zhang, X.-S., and Chen, L. (2006). A markov chain model for haplotype assembly from snp fragments. *Genome Informatics*, **17**(2), 162–171.

[92] Wang, Y., Feng, E., and Wang, R. (2007). A clustering algorithm based on two distance functions for mec model. *Computational biology and chemistry*, **31**(2), 148–150.

[93] Wood, T. E., Takebayashi, N., Barker, M. S., Mayrose, I., Greenspoon, P. B., and Rieseberg, L. H. (2009). The frequency of polyploid speciation in vascular plants. *Proceedings of the national Academy of sciences*, **106**(33), 13875–13879.

[94] Wu, J., Wang, J., *et al.* (2009a). A practical algorithm based on particle swarm optimization for haplotype reconstruction. *Applied Mathematics and Computation*, **208**(2), 363–372.

[95] Wu, L.-Y., Li, Z., Wang, R.-S., Zhang, X.-S., and Chen, L. (2009b). Self-organizing map approaches for the haplotype assembly problem. *Mathematics and Computers in Simulation*, **79**(10), 3026–3037.

[96] Wurster-Hill, D. H. and Maurer, L. H. (1978). Cytogenetic diagnosis of cancer: Abnormalities of chromosomes and polyploid levels in the bone marrow of patients with small cell anaplastic carcinoma of the lung. *Journal of the National Cancer Institute*, **61**(4), 1065–1075.

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

[98] Yonemaru, J.-i., Ebana, K., and Yano, M. (2014). HapRice, an SNP haplotype

database and a web tool for rice. *Plant and Cell Physiology*, **55**(1), e9–e9.

[99] Zhang, Y., Xu, G.-h., Guo, X.-y., and Fan, L.-j. (2005). Two ancient rounds of polyploidy in rice genome. *Journal of Zhejiang University. Science. B*, **6**(2), 87.

[100] Zhao, Y.-Y., Wu, L.-Y., Zhang, J.-H., Wang, R.-S., and Zhang, X.-S. (2005). Haplotype assembly from aligned weighted SNP fragments. *Computational Biology and Chemistry*, **29**(4), 281–287.

[101] Zheng, C., Voorrips, R. E., Jansen, J., Hackett, C. A., Ho, J., and Bink, M. C. (2016). Probabilistic multilocus haplotype reconstruction in outcrossing tetraploids. *Genetics*, **203**(1), 119–131.

2

Exploiting Next Generation Sequencing to solve the haplotyping puzzle in polyploids: a simulation study

This chapter has been published with minor modifications in: Ehsan Motazedi, Richard Finkers, Chris Maliepaard, Dick de Ridder, **Exploiting next-generation sequencing to solve the haplotyping puzzle in polyploids: a simulation study**, Briefings in Bioinformatics, Volume 19, Issue 3, May 2018, Pages 387-403

Abstract

Haplotypes are the units of inheritance in an organism, and many genetic analyses depend on their precise determination. Methods for haplotyping single individuals use the phasing information available in Next Generation Sequencing reads, by matching overlapping SNPs while penalizing post hoc nucleotide corrections made. Haplotyping diploids is relatively easy, but the complexity of the problem increases drastically for polyploid genomes, which are found in both model organisms and in economically relevant plant and animal species. While a number of tools are available for haplotyping polyploids, the effects of the genomic makeup and the sequencing strategy followed on the accuracy of these methods have hitherto not been thoroughly evaluated.

We developed the simulation pipeline *haplosim* to evaluate the performance of three haplotype estimation algorithms for polyploids: HapCompass, HapTree and SDhaP, in settings varying in sequencing approach, ploidy levels and genomic diversity, using tetraploid potato as the model. Our results show that sequencing depth is the major determinant of haplotype estimation quality, that 1 *kb* PacBio CCS reads and Illumina reads with large insert-sizes are competitive, and that all methods fail to produce good haplotypes when ploidy levels increase. Comparing the three methods, HapTree produces the most accurate estimates, but also consumes the most resources. There is clearly room for improvement in polyploid haplotyping algorithms.

2.1. Introduction

The advent of sequencing technology has had tremendous impact in genomics and genetics over the last years. The human (*Homo sapiens*) genome, published in 2003 [15], formed the basis for large-scale efforts to catalogue sequence variation found between individual genomes, in particular single nucleotide polymorphisms (SNPs) [49]. Subsequently, the international HapMap project [21] sequenced a large number of individuals to discover haplotype blocks, i.e. genomic regions containing co-segregating SNPs. These haplotype blocks and their so-called haplotype tagging SNP markers, heterozygous SNPs whose alleles predict the presence of a certain haplotype, formed the basis for the development of high-density SNP arrays [18, 34], capable of determining rare or less-frequent genotypes in a population, which were used in a large number of studies relating SNPs to phenotypes such as ecological traits, diseases and disorders [52].

Following the success of the human genome project, many animal and plant species were sequenced and genotyped, notably mouse ear cress (*Arabidopsis thaliana* [30], fruit fly (*Drosophila*) [14], chicken (*Gallus gallus)* [27], pig (*Sus scrofa*) [24], potato (*Solanum tuberosum*) [20], tomato (*S. lycopersicum*) [16] and hot pepper (*Capsicum annuum*) [31] (the last three being plant genera within the *Solanacea* family). This has impacted not only fundamental genetics research, but has also revolutionized the fields of animal and plant breeding, by relating thousands of previously unknown genomic variants to physiological, morphological and economically important traits such as yield per generation and disease resistance [23, 41].

The introduction of high-throughput, relatively cheap and reliable next-generation sequencing (NGS) technologies made it possible to determine most of the variants directly within a single genome rather than using a pre-defined set of marker variants as proxies for the other variants [8]. Efficient tools have been developed to call variants based on NGS data, e.g. FreeBayes [19] and GATK [42], and also to link variants on the same homologous chromosome, so-called haplotype phasing. However, while phasing of nearby variants occurring within the average NGS read length is relatively straightforward, long-range haplotyping using NGS data remains a challenge. Nevertheless haplotyping is important in many areas: in fundamental biology, to improve our understanding of genome structure, recombination and evolution [12]; in medicine, to obtain a full picture of the genetic variation in a population potentially linked to diseases and traits [22] and to investigate the effect of compound heterozygosity [11]; and in animal and plant breeding, to move from phenotype-based to genotype-based crossing and selection of individuals [25, 40]. Moreover, the knowledge of haplotypes can help reveal the linkage disequilibrium pattern in a population and hence increase the power and coverage of genetic analysis by allowing the imputation of a large number of alleles using a limited set of genotyped loci [26, 28] .

In the diploid case, haplotyping algorithms aim to divide the aligned reads into two complementary sets, each covering a specific region of, say, n heterozygous sites, so that the nucleotides are the same at the overlapping sites of the reads within each set, but different between the sets. The algorithmic challenge then is to take the occurrence of sequencing and variant calling errors into account [35, 38, 48]. Minimum Error Correction (MEC), the most prevalent approach, uses single base-flips for reads that conflict with both of the read sets, presumably due to sequencing or variant calling errors, to

assign them to one of these. The aim is to find a configuration that requires a minimal number of such base flips [48, 54]. This strategy is the basis of several diploid haplotyping algorithms [4, 5, 33, 55–58, 60]. Recently, diploid aware genome assemblers based on long reads, produced for instance by PacBio and Nanopore technologies [43], have also been developed. These algorithms use the overlap-layout-consensus approach to construct the assembly graph and obtain the primary contigs from the raw reads, and try to resolve the haplotypes by calling heterozygous SNPs within the contigs. These SNPs are used to separate the long reads into two groups, the so called "haplotigs", from which consensus sequences are obtained to determine the phasing [13].

For polyploid genomes, the problem could be formulated as dividing the reads into $k > 2$ groups, but the generalization from the diploid case is not straightforward. Specifically, polyploids can be classified as allopolyploids, autopolyploids or mixture types, i.e. so-called segmental allopolyploids. In the allopolyploid case, the constituent subgenomes of the polyploid are derived from adequately distant diploid ancestors that do not usually recombine with each other, a situation observed among several species in the plant kingdom such as tetraploid and hexaploid wheat [45]. Under specific circumstances, one may treat the sub-genome haplotypes as separate diploids and an ad hoc phasing solution could be still achievable using the algorithms developed for diploids. As an example, this strategy has been successfully applied to pasta wheat *Triticum turgidum* [32], which is a self-fertilizing allotetraploid for which the ancestral diploid genomes are also known, to determine the variation on each sub-genome by dividing the transcriptome reads into two groups using HapCut [4]. In the case of (partial) autopolyploids, however, recombination is observed between homologues belonging to different subgenomes and unlike for the diploid case, knowledge of one haplotype does not automatically determine the phasing of the others. Besides, some haplotypes may be (locally) identical and thus several configurations could have the same MEC score. Moreover, the computational complexity of haplotype reconstruction increases rapidly with an increase in ploidy [2, 17]. The diploid approaches are therefore in general not applicable to polyploids.

Still, haplotype assembly for polyploids is highly relevant, as many interesting organisms have polyploid genomes and haplotyping will help unravel the range of the complex recombinations allowed by such genomes. Within the animal kingdom, triploidy and tetraploidy are observed in treefrog (*Xenopus laevis*) [53] and zebrafish (*Danio rerio*) [59], both important model organisms in evolutionary biology. Moreover, many economically important crops and ornamentals are polyploid, including tetraploid alfalfa (*Medicago sativa*), triploid banana (*Musa acuminata* × *M. balbisiana*), tetraploid leek (*Allium ampeloprasum*), tetraploid potato (*S. tuberosum*), tetraploid hard wheat (*T. durum*), hexaploid bread wheat (*T. aestivum*), tetraploid, hexaploid and octoploid strawberry species including *Faragaria moupinesis* (k=4), *F. moschata* (k=6), *F.* × *ananassa* (k=8) and several hybrid cotton (*Gossypium*, tetraploid or hexaploid) and rose (*Rosa*, tetraploid) species.

Here we review three state-of-the-art haplotyping algorithms for polyploids: HapCompass [2, 3], HapTree [7] and SDhaP [17], and evaluate their accuracy through extensive simulations of random genomes and NGS reads. Using the highly heterozygous tetraploid potato (*S. tuberosum*) as a model, we generated random genomes using a real-

istic stochastic model with parameters SNP density and distribution of SNP dosages, i.e. the number of alternative alleles at each SNP site, derived from a recent genomic study of potato [51]. In addition, we simulated genomes at higher levels of ploidy with the same SNP density, as well as tetraploid genomes with different SNP densities and haplotype dosages, in order to investigate the effects of genome characteristics on the estimation. Moreover, we considered various sequencing depths, paired-end insert-sizes and sequencing technologies to quantify the impact of these parameters on the haplotyping. We provide guidelines to apply the haplotyping methods in practice, and show the characteristics of each method in various situations.

The pipeline used is available as software package *haplosim*, which allows simulation for various sequencing approaches, genomic characteristics and variation models.

2.2. Material and Methods

While several studies have used experimental data, e.g. the human haplotype panels and sequence reads, to evaluate the efficiency of diploid haplotyping algorithms [1], experimentally obtained haplotypes are often not available for polyploids at a scale enabling insightful statistical comparison. Therefore, the evaluation of polyploid haplotyping algorithms has been based on artificial data sets [2, 3, 7, 17]. Here we also rely on simulation to evaluate the performance of these methods. Compared to the previous studies, our approach has the advantage of simulating all parts of a practical haplotyping pipeline, encompassing the careful simulation of genomes and sequence reads based on real data and the application of standard software for read alignment and genotype calling. In contrast, previous studies relied on the direct simulation of SNP-fragment matrix using simplifying assumptions. An additional advantage of our simulation approach is that it allows to investigate the effects of SNP-density, similarity between homologues, ploidy level, sequencing depth, sequencing technology and DNA library size on the quality of haplotype estimates, which is usually not feasible using real data.

We developed a multi-stage pipeline, *haplosim*, to simulate polyploid individuals, with various genomic characteristics, that are sequenced *in silico*. After individual SNP detection and dosage estimation, the haplotypes are estimated, separately for each individual, by the available algorithms: HapCompass [2, 3] , HapTree [7] and SDhaP [17] in the next steps. In the last step of the pipeline., the estimated haplotypes are compared to the originally simulated haplotypes using quantitative measures (Figure 2.1).

For the first step (Figure 2.1-A), polyploid genomes are produced from a reference DNA sequence by introducing heterozygous regions containing bi-allelic SNPs, using the command-line tool *haplogenerator* that we developed to this purpose. Next, NGS reads are simulated for each produced individual using ART [29] and PBSIM [44] for Illumina and PacBio, respectively, and the reads are mapped back to their reference genome using *bwa-mem* [36] (with the settings recommended in its manual for Illumina and PacBio reads). The alignments are pre-processed to generate BAM files and remove duplicates by samtools [37] and Picardtools [9], after which SNPs are called using FreeBayes [19] (Figure 2.1-B). The processed alignments, the reference and the VCF files are used in the haplotyping step by HapCompass [2, 3], HapTree [7] and SDhaP [17] (Figure 2.1-C) and the obtained haplotype estimates are compared to the original haplotypes by the command line tool *hapcompare* that we developed using several measures of estimation

quality (Figure 2.1-D). These steps are explained in detail below.

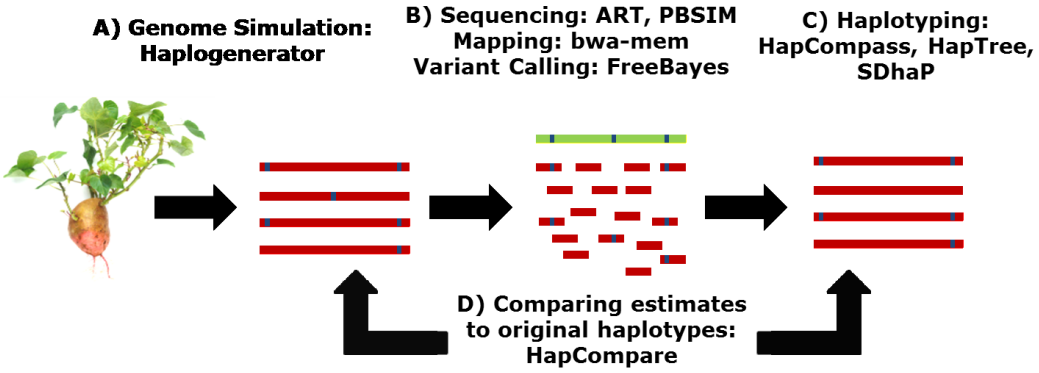


Figure 2.1: *Haplosim* pipeline to generate, estimate and evaluate haplotypes. Random genomes and haplotypes are produced by *Haplogenerator*, from which NGS reads are simulated and mapped backed to the reference. SNPs are called at the next step and the haplotypes estimated by HapTree, HapCompass and SDhaP. The estimates are compared to the original haplotypes by *hapcompare*.

2.2.1. Polyploid haplotyping software

Currently, three optimization based haplotyping algorithms are available for polyploids that make use of the sequence reads of a single sample: HapCompass [2, 3], HapTree [7] and SDhaP [17] (Table 2.1). Among these three algorithms, HapTree and SDhaP have separate software releases for diploids and polyploids, which may have major performance differences not discussed here. We explain each method assuming a genomic region containing l heterozygous SNP sites s_1, s_2,...,s_l, a ploidy level $k > 2$ and an NGS dataset containing m (paired-end) reads. We define a "fragment" as the sequence of the determined alleles at the heterozygous sites within a (paired-end) read. For simplicity, we focus on the most prevalent type of SNPs, bi-allelic SNPs, for which the alleles can be represented by '0' (the reference) and '1' (the alternative).

a) HapCompass: Aguiar and Istrail (2013) extended their graphical haplotype estimation approach for diploids [2], by constructing the polyploid *compass graph*, which has k nodes for each variant site s_i in a k-ploid corresponding to the k alleles at that site [3]. To each SNP pair (s_i, s_j) covered by at least one of the m fragments, the phas-

Table 2.1: Summary of the polyploid haplotype estimation algorithms using sequence reads

Algorithm	Principle	Release Version	Separate diploid release
HapCompass (Aguiar and Istrail 2013)	Graph based using Weighted Minimum Edge Reduction criterion	HapCompass v0.8.2	No
HapTree (Berger et al. 2014)	Using Bayesian probability tree with Maximum Relative Likelihood criterion	HapTree_v0.1	Yes
SDhaP (Das and Vikalo 2015)	Minimum Error Correction criterion with Semi-Definite approximation	SDhaP_poly	Yes

ing with the largest likelihood is assigned by a polyploid likelihood model, conditional on the covering fragments and assuming a fixed base calling error rate. k edges are accordingly added to the compass graph between the nodes at s_i and s_j sites, representing the k homologues covering s_i and s_j and weighted by their likelihoods. A global *minimum weighted edge removal* (*MWER*) algorithm is applied to detect and eliminate edges with conflicting phasing information from the compass graph with the aid of auxiliary *chain graphs*. Each chain graph takes a set of variants that make a cycle in the compass graph and detects phasing conflicts within the cycle, if present. The MWER algorithm then tries to resolve the conflicts by eliminating a number of edges with minimum total weight corresponding to the least likely phasings. Finally, the most likely haplotypes are found over the full set of SNPs from the conflict-free compass graph by finding k disjoint maximum spanning trees through an efficient greedy algorithm, corresponding to the k most likely homologues covering the l SNP sites.

b) HapTree: The HapTree algorithm, developed by Berger et al. (2014) [7], builds a tree representing a subset of likely phasing solutions for l heterozygous sites and tries to find the most probable path from s_1 to s_l in the tree as the best phasing, by calculating the probability of each phasing conditional on the m fragments, assuming a fixed base calling error rate. Nevertheless, as an exhaustive search over the full tree of all solutions would be computationally prohibitive, the tree is built and extended site by site with branching and pruning to greedily eliminate the (relatively) low probability paths from the final tree. In doing so, HapTree calculates the relative probabilities of the haplotypes at each extension using the relative probabilities of the haplotype at the previous extension that survived the branching and pruning, taking the error model into account.

c) SDhaP: The third algorithm, polyploid SDhaP designed by Das and Vikalo (2015), is a semi-definite programming approach that aims to find an approximate MEC solution by a greedy searching of the space of all possible phasings from s_1 to s_l [17]. The algorithm starts with random initial haplotypes, and tries to find the MEC solution by making changes to these initial haplotypes according to a gradient-descent method. To this end, the MEC problem is reformulated as a semi-definite optimization task and preliminary solutions are obtained in polynomial time by exploiting the sparseness of overlaps between fragments for efficient implementation. These preliminary estimates are subsequently refined by greedy flipping of the alleles in the estimated homologues to further reduce the MEC, if possible. By this flipping, SDhaP allows making changes to the dosages of the alternative alleles estimated during variant calling, which could sometimes lower the error correction score. Therefore, the dosages of corresponding SNPs in the SDhaP estimates and original haplotypes could differ, in contrast to the estimates produced by HapTree and HapCompass.

2.2.2. Simulation of polyploid genomes and NGS reads

a) Haplotype generation

We developed the command line tool *haplogenerator* for generating artificial genomes and their haplotypes with desired characteristics. Specifying an indexed fasta file as input, *haplogenerator* applies random insertions, deletions and mutations to the input sequence according to a chosen stochastic model to produce modified fasta files for each of the k-genomes of a k-ploid individual. A separate haplotype file is also made con-

taining the phasing of the generated variants. In the haplotype file, the reference and alternative alleles are numerically coded, assigning 0 to the nucleotide present in the input reference and the following integers to the alternative alleles. The coordinate of each variant on its contig is also specified within the haplotype file.

The random indel and mutation sites are scattered across the input genome according to a selected built-in stochastic model for the distance between consecutive variations, or alternatively by sampling with replacement from a given empirical distribution for this distance. At each position i, possible alternative alleles are generated: for indels, an inserted or deleted nucleotide; for mutations, nucleotides other than the reference (or just one nucleotide for obtaining bi-allelic SNPs). Next, a dosage d_i is assigned to the alternative allele based on the ploidy s_i, according to user-specified probabilities for d_i=1 to $d_i = k$, and d_i out of k homologues are selected randomly to get their allele at s_i changed to an alternative. To account for linkage disequilibrium, we imposed an additional step after this dosage assignment to reassign the alternative alleles at each s_{i+1}, $i = 1, 2, ..., n-1$, to the homologues containing the alternative alleles at s_i (as much as the numbers of alternative alleles, i.e. the dosages, at s_i and s_{i+1} allow), with a reassignment decision made independently for each site from s_2 to s_n with an arbitrary probability set to 0.4.

b) Simulation of NGS reads

We used the technology specific simulator ART [29] to generate paired-end reads from Illumina MiSeq and HiSeq 2500 technologies [43, 46], and PBSIM [44] to simulate Circular Consensus Sequencing (CCS) and Continuous Long Reads (CLR) from Pacific BioScience [39, 46]. The average length of single Illumina reads was set to the maximum allowed by ART ($125bp$ and $250bp$ for HiSeq 2500 and MiSeq, respectively). For PBSIM, the average read length was set to $1kb$ and $5kb$ with CCS, and to $10kb$ with CLR. Setting these averages, both softwares generated reads with random lengths following the built-in distributions derived from empirical data for each technology.

Each homologue was "sequenced" separately, and the reads were combined to simulate the output of real sequencing apparatus. Average sequencing depths were specified for each homologue to obtain the desired average total depth equal to k times the per homologue depth, with k being the ploidy. Both ART and PBSIM consider a discrete uniform distribution with the user-set mean for the depth at each position, and hence the standard deviation of the total depth was dependent on the average depth per homologue, c, and equal to $\sqrt{\frac{kc(c+2)}{12}}$ for a k-ploid.

The choice of the sequencing strategies in our study was based on the efficiency and performance of the available techniques [46], as well as their practical convenience. In particular, we did not consider single-ended reads of Illumina as preliminary assessment showed that they produce low quality estimates with a large number of gaps in the solution.

c) Simulation of polyploid datasets

In order to simulate realistic polyploid genomes, we chose tetraploid potato (*S. tuberosum)* as the model organism, due to the availability of a reference genome [20] as well as NGS data containing genomic variation of 83 diverse cultivars [51]. The sequence of

chromosome 5 from PGSC v4.03 DM draft genome [20] was used as the template sequence for haplogenerator. We selected random contiguous regions with $20kb$ length from this template to be used as references for simulating genomes. In selecting the references, we rejected $20kb$ regions of the template that contained more than 20% undetermined nucleotides, and omitted these undetermined sites (denoted by 'N' in PGSC v4.03 DM sequence) before introducing mutations in the accepted regions. As the length of many genes falls below $20kb$, choosing references with this size allows us to evaluate haplotype estimation for amplicon sequences covering a complete gene. Random bi-allelic SNPs were introduced in each reference to produce synthetic tetraploid genomes according to the built-in *lognormal* model of haplogenerator, with the mean and the standard deviation of the log-distance between the SNPs being set to 3.0349 and 1.293, respectively, corresponding to an expected SNP frequency of 1 per $21bp$ with a standard deviation of $27bp$. The distribution of the dosages, d_i, was similarly set equal to that from [51], with percentages equal to 50%, 23%, 14% and 13% of simplex ($d_i = 1$), duplex, ($d_i = 2$), triplex ($d_i = 3$) and quadruplex ($d_i = 4$) SNPs.

In order to investigate the effect of library preparation, we considered various insert-sizes for paired-end Illumina reads, namely end-to-end insert-sizes of 235, 300, 400, 500, 600 and $800bp$ with HiSeq 2500 and 400, 450, 500, 600 and $800bp$ with MiSeq. For evaluation of the effect of sequencing depth on haplotyping, 2×, 5×, 8×, 10×, 20×, 22×, 25×, 28×, 30× and 35× average coverages were considered per homologue for each of these insert-sizes.

To investigate the effects of genome characteristics, the ploidy level, the dosage of different homologues and the SNP density on the quality of haplotype estimation, additional genomes were generated in a similar way by haplogenerator. Considering the same proportion of simplex and duplex SNPs, i.e. SNPs with dosages equal to 1 and 2, respectively, as in [51] and considering equal proportions for the dosages higher than 2, we simulated genomes with $3n$, $4n$, $6n$, $8n$, $10n$ and $12n$ ploidy levels to investigate the effect of the ploidy, and simulated modified tetraploid genomes that contained only two distinct homologues with simplex and triplex dosages to investigate the effect of similarity between the homologues on haplotype estimation. While these scenarios assume a SNP-density model valid for *S. tuberosum*, they still can show the pattern by which ploidy level and similarity between the homologues influence the quality of haplotyping.

Finally, tetraploid genomes with SNP densities lower than that of the highly heterozygous *S. tuberosum* [51] were simulated to observe the effect of SNP density, with average frequencies of 1 per $22bp$ to 1 per $110bp$.

In total, 250 individuals were simulated for each of the above mentioned scenarios by choosing 25 random references from the template and generating 10 genomes with randomly distributed bi-allelic SNPs for each selected reference (Figure 2.2).

2.2.3. Evaluation of the estimated haplotypes

As several types of error occurring in different steps of the haplotyping pipeline could cause differences between the actual haplotypes and their estimates, we needed several measures of consistency to be able to capture all of them, as summarized in Table 2.2. These errors include the absence or wrong dosage of original SNPs in the estimates, presence of spurious SNPs, discontinuity of the estimated haplotypes, i.e. presence of gaps

between estimated haplotype blocks, and finally wrong extension of homologues leading to incorrect phasing. We also included an extra measure, the *failure rate* (*FR*) for each algorithm, regardless of the quality of haplotype estimation, as it could happen that the haplotyping tools failed to produce any estimate for some of the individuals.

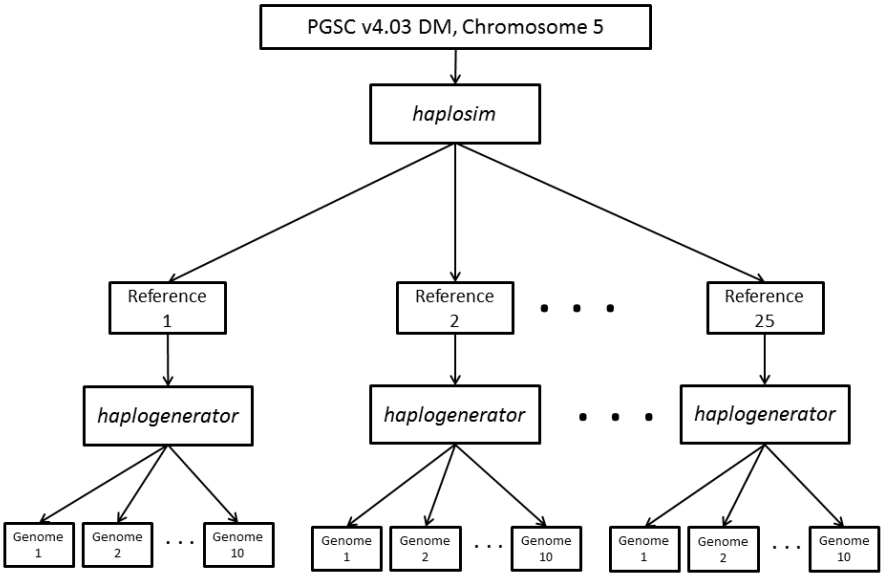


Figure 2.2: The schematic diagram of simulation for each considered scenario. 10 references of length 20kb are chosen from the draft sequence chromosome 5 [20], from each 10 polyploid genomes are simulated containing bi-allelic SNPs randomly distributed according to the lognormal distance model, to obtain datasets of size 250.

Table 2.2: Summary of the measures used to asses the quality of haplotyping

Measure	Description	Limitation
PAR	The average accuracy of phasing for any possible SNP pair	SNPs present in both original haplotypes and estimates, no account of gaps
VER*	Number of wrong extensions of homologues with (parts of) other homologues	SNPs present in both original haplotypes and estimates with same dosages, no account of gaps
SMR	Proportion of original SNPs missing in the estimates	No account of phasing, false positive SNPs in estimates and dosages
IDR	Proportion of SNPs with an incorrect dosage in estimated haplotypes	SNPs present in both original haplotypes and estimates, no account of phasing
PPV	Proportion of true SNPs in the estimated SNPs	No account of phasing, missing SNPs, and wrong dosages
NGPS*	Number of gaps (interruptions) introduced in the estimated haplotype	No account of phasing, missing, wrong dosages and false positives
FR	Rate of failure for an algorithm	No account of the quality of the estimates

* *measure normalized by the size of original haplotypes, i.e. the number of original SNPs*

Some of the SNPs originally simulated were absent in the output set of each simulation, either due to mapping and variant calling errors or due to their not being phased by the algorithms, and therefore could not be included in the comparisons of true and estimated phasings. Instead, their proportion was calculated as *SNP missing rate* (*SMR*) and considered as the first measure of estimation quality. Similarly, spurious SNPs in the estimates were not included in the comparisons and the *Positive Predictive Value* (*PPV*) or the precision of the estimated SNPs was calculated as the proportion of genuine SNPs among the estimated SNPs as the next measure of estimation accuracy. The *incorrect dosage rate* (*IDR*) was also calculated as the proportion of SNPs that had an incorrect dosage in the estimated haplotypes in the set of SNPs that were common between the original haplotypes and the estimates.

Having excluded the missing and spurious SNPs, the *Pairwise Phasing Accuracy Rate* (*PAR*) [10] was computed as the proportion of all heterozygous SNP pairs for which the estimated phasing was correct. This measure captures the errors caused by chimeric elongation of the homologues during haplotype estimation, i.e. the elongation of a homologue by (part of) another homologue, as well as errors caused by incorrect dosage estimation.

One way to calculate the accuracy of phasing for more than just two SNPs is to consider the phasing accuracy rate for groups of three SNPs, four SNPs, etc. However, the phasing accuracies will no longer be independent for the groups of SNPs that have more than one SNP in common, leading to biased estimates of the accuracy rates. Instead, we calculated the *Vector Error Rate* (*VER*), also called the switch error rate, defined as the number of times a homologue is erroneously extended by part of another homologue [7]. Such erroneous extensions are also called switches between homologues, and the measure is equal to two times the number of wrong phasings for pairs of consecutive SNPs for diploids. For polyploids, the measure is calculated by finding the minimum number of crossing-overs needed to reconstruct the true haplotypes from the estimates [7]. To be able to compare of VER for different ploidy levels, genome lengths and SNP densities, we normalized it by the number of originally simulated SNPs as well as the ploidy level for each individual. The SNPs with a wrong estimated dosage of the alternative allele were omitted before applying this measure, as otherwise the true haplotypes could not be reconstructed by simple switching of the estimated homologues without considering allele flips from 0 to 1 or *vice versa*.

The last measure of estimation quality that we used was the *number of gaps per SNP* (*NGPS*) in the estimates, as the simulated continuous haplotypes can be broken into several disconnected blocks, causing gaps in the estimated haplotypes. This phenomenon happens if the connection between SNPs is lost due to low sequencing coverage or sequencing/variant calling errors at certain sites. Therefore, we calculated the number of break points, i.e. gaps, in the estimates (equal to the number of disjoint blocks minus one), and normalized it by the total number of simulated SNPs for each individual for the same reasons as for VER. In case gaps were present in the estimates, we calculated the other measures separately for each estimated block and reported the weighted average of the block-specific measures, weighted by the number of compared SNPs in each estimated block, or the number of possible pairwise phases in case of PAR.

Finally, the computational complexity of each of the haplotyping algorithms was

considered as a function of sequencing coverage, insert-size, ploidy, SNP density and homologue dosages. The applied haplotyping methods are memory intensive methods, increasingly consuming system resources with time, sometimes up to tens of gigabytes of virtual memory. To run the methods on a system with shared resources, and considering the fact that the algorithms require an increasing amount of virtual memory with time, a time limit of 900 seconds (2000 seconds for the analysis with various levels of ploidy) was imposed on each haplotyping algorithm, after which the algorithms were externally halted and the estimation considered a failure. This amount of time was deemed reasonable considering the $20kb$ length of the simulated genomic regions, and the number of time-out events was added to the number of times each algorithm failed to estimate any haplotypes due to the occurrence of internal errors. Total FRs are thus also reported for each haplotyping scenario.

2.2.4. Comparison of haplotyping algorithms

In order to compare the overall performance of the three haplotype estimation methods, we built three linear regression models with the mentioned quality measures as response and the haplotyping method as predictor, considering sequencing depth, sequencing technology and the (paired-end) library size as covariates in the model. As each of the simulated genomes was haplotyped simultaneously by the three estimation methods, the effect of the genome on the estimation quality was incorporated as a random effect in the model. Similarly, as 10 genomes were generated from each of the 25 randomly selected references, the effect of the common reference was added as the second random component to the model.

For each quality measure, a complete-case analysis was performed, including only the results of those simulations for which all the three estimation methods reported some value. The models were estimated by Restricted Maximum Likelihood (REML) [50] using the *lmer* function from the package *lme4* [6] in R 3.2.2 [47].

2.3. Results and Discussion

2.3.1. Haplogenerator produces realistic genomes

In order to investigate the compatibility of the simulated $20kb$ *S. tuberosum* genomic regions with the real regions sequenced by Uitdewilligen et al.(2013) [51] in terms of the density of bi-allelic SNPs, we obtained quantile plots (QQ-plots) of the distances between consecutive SNPs s_i, s_{i+1}, generated by the applied lognormal model versus the distances between consecutive bi-allelic SNPs (within the same RNA-capture region) in the combined data of 83 diverse cultivars from [51]. As shown in Figure 2.3, the two empirical distributions match well enough, although the distribution of real SNPs seems to have a heavier tail than lognormal (accounting for less than 2% of the total number of real bi-allelic SNPs). This heavier tail is plausibly explained by the presence of highly conserved regions in real genomes, subject to natural and artificial selection pressure, as well as the use of genome-wide RNA-baits to reduce the complexity of genome in [51] which can result in the exclusion of some SNPs in target regions that had poor capture success.

The proportions of simplex to quadruplex SNPs, i.e. the dosage proportions, were

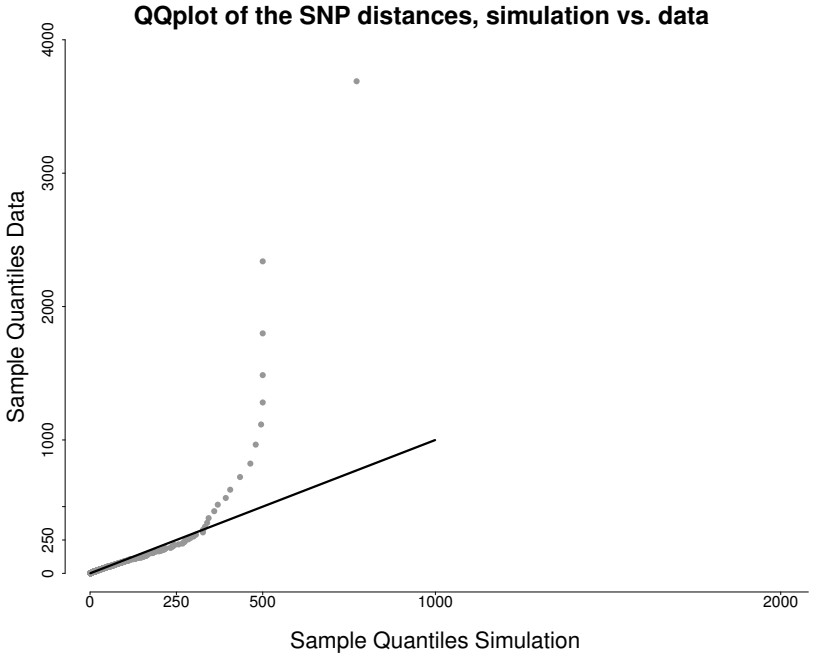


Figure 2.3: Quantile plots for the distances between successive SNPs obtained from simulation by *haplogenerator* using the lognormal distance model (horizontal axis) versus the one obtained from the data of 83 potato cultivars of Uitdewilligen et al. (2013) [51]. The two distributions match well, though a heavier tail is observed for the data of Uitdewilligen et al. (2013), accounting for less than 2% of the SNPs.

also almost identical as those obtained from Uitdewilligen et al. (2013) [51] (Section 2.2.2-c).

2.3.2. Sequencing depth is the major determinant of haplotyping quality

An important goal of the simulations was to observe to what extent sequencing strategies influence haplotyping results, because of the practical importance in setting up experiments and choosing a technology. As different technologies rely on different library preparation and nucleotide calling methods, their output is often different in terms of the average read length and sequencing error profile. Besides, the sequencing depth, average read length and paired-end insert-size can vary according to the user's requirements with the same technology. We found that the performance of all three haplotyping methods was considerably affected by the sequencing strategy, most notably by the sequencing depth.

Regardless of the used sequencing technology and the insert-size, a sequencing depth between 5-20× per homologue is required to obtain results satisfactory in terms of haplotype accuracy (PAR) and completeness (SMR) (Figure 2.4-a, b). Both improve continu-

ously with sequencing depth, but flatten out at 15×. A notable exception is HapTree, of which the increased failure rate (FR) at higher depths is reflected in a worse completeness (increasing SMR). Other quality measures (VER, PPV, IDR, NGPS) were not substantially influenced by sequencing depth at depths higher than 5× per homologue.

HapTree's failure rate (FR, Figure 2.4 c) was rather high for low and high sequencing depths. At lower depths, less than 2× per homologue, there is not enough information available for effective branching and pruning of the solution tree and time-out errors result in failures. In contrast, for high sequencing depths the relative likelihood values often become very small, due to the presence of many terms in the likelihoods of partial haplotypes, making a meaningful comparison impossible. This problem will be discussed further in Section 2.3.10.

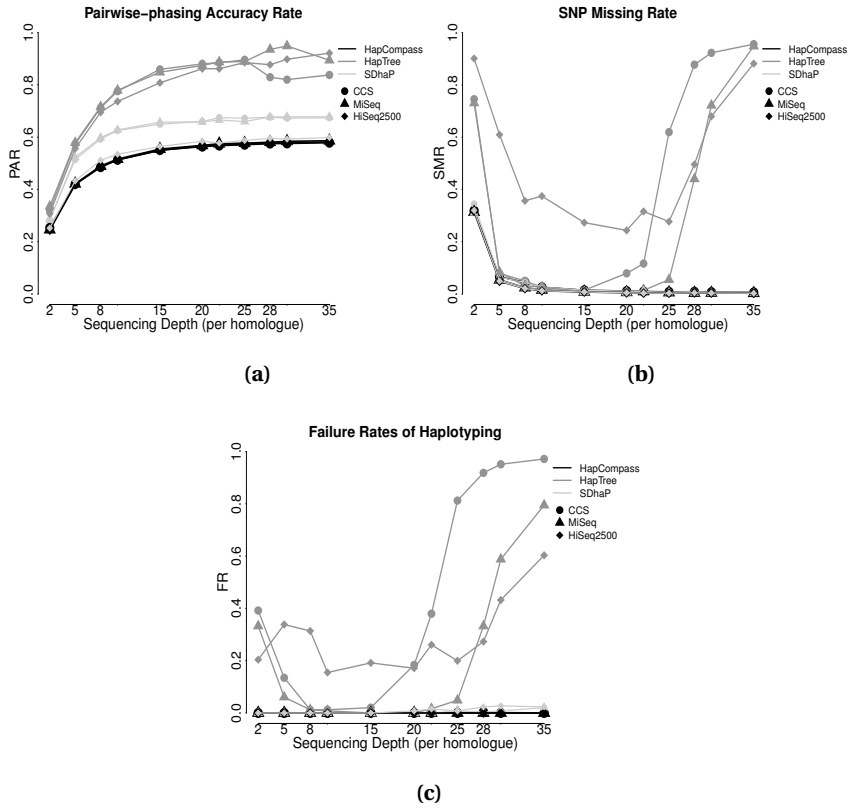


Figure 2.4: Plots of haplotype estimation quality measures: (a) SMR, (b) PAR and (c) FR as a function of sequencing depth per homologue using HapCompass (black), HapTree (gray) and SDhaP (light gray), for simulated 20*kb* tetraploid *S. tuberosum* genomes. Sequencing was performed in silico for paired-end MiSeq (triangle) and HiSeq 2500 (rhombus) with 800*bp* insert-size, as well as for PacBio-CCS of 1*kb* length (circle).

As sequencing depth is an important factor in determining the total cost of sequencing, these results show that extra cost can be avoided by choosing a moderate sequencing depth without sacrificing considerable haplotyping accuracy.

2.3.3. Large insert paired-end reads are competitive with long reads

In addition to sequencing depth, the insert-size of paired-end reads and the employed sequencing technology can have an impact on the estimation quality. These are also important factors to specify when designing a sequencing experiment, as they influence cost, throughput and quality. To quantify their effects, we simulated NGS reads for HiSeq 2500, MiSeq and CCS technologies at each sequencing depth, and simulated paired-end reads with various insert-sizes for HiSeq 2500 and MiSeq (Section 2.2.2-b).

Our results show that at the same sequencing depth, increasing the insert-size of paired-end reads was not markedly influential on the overall quality of haplotyping (Supp. Figure 1), except for the number of gaps that was expectedly reduced with larger inserts. Moreover, similar estimation qualities were obtained using the $1kb$ CCS reads and paired-end Illumina reads with a large insert-size ($800bp$) (Figure 2.4). At the same depth, the paired-end reads contain basically the same phasing information as the long reads.

Although libraries with large inserts are costly and difficult to obtain, they may be still easier to generate than long continuous reads and therefore can be a competitive option for designing haplotyping experiments.

2.3.4. HapTree is the most accurate method, but often fails

Different haplotyping algorithms yield different estimates for the same individual. With simulated individuals, it is possible to compare the quality of these estimates as the haplotypes are known a priori. To this end, we used linear regression models relating the performance measures to the algorithms used (Section 2.2.4). Because of the substantial difference between the estimation results using CLR compared to the other sequencing methods (Section 2.3.9), those results were excluded from the regression analysis.

Table 2.3 shows the 99% confidence intervals for the effects of estimation method and sequencing technology on the haplotyping accuracy for the tetraploid genomes, with HapCompass on CCS data taken as the baseline. HapTree is significantly more accurate (higher PAR and lower VER) than the other methods, but less complete (higher SMR) due to its frequent failure (Figure 2.4-c). SDhaP yields slightly, but significantly, worse dosage estimates (higher IDR). There was no significant relation between the haplotyping method and the continuity of haplotype estimates (NGPS) or precision of the SNPs (PPV).

Finally, the results were not significantly different using Illumina MiSeq reads and PacBio CCS, except for PPV which was slightly higher with MiSeq. On the other hand, HiSeq 2500 reads resulted in significantly higher VER, SMR and NGPS at α=0.01 compared to the other sequencing methods, with the most noticeable effect being on the number of gaps, which was expected considering the short single-read length of HiSeq.

Overall, these results confirm that HapTree is the most accurate method when it does not fail, and that Illumina and PacBio reads offer very similar performance.

Table 2.3: Point estimates and 99% confidence intervals for the effects of haplotyping and sequencing methods on the haplotyping quality measures

Paramter\Quality Measure	PAR	VER	SMR	IDR	PPV	NGPS
Intercept*	0.33(0.308;0.356)	0.25(0.206;0.285)	0.06(0.022;0.090)	0.20(0.178;0.222)	0.97(0.970;0.971)	0.01(0.007;0.013)
HapTree	0.23* (0.226;0.230)	-0.1* (-0.1;-0.09)	0.27* (0.263;0.270)	0.00(0.003;0.005)	0.00(0.000; 0.000)	0.00(-0.002;-0.001)
SDhaP	0.08* (0.079;0.082)	-0.01* (-0.014;-0.005)	0.00(-0.003;0.004)	0.01* (0.012;0.013)	0.00(0.000;0.000)	0.00(0.000;0.000)
HiSeq 2500	-0.005(-0.026;0.016)	0.06* (0.024;0.1)	0.04* (0.011;0.072)	0.01(-0.011;0.028)	0.03* (0.025;0.026)	0.01* (0.008;0.013)
MiSeq	-0.002(-0.023;0.019)	-0.02(-0.051;0.02)	0.02(-0.008;0.052)	0.01(-0.013;0.026)	0.02* (0.023;0.025)	0.00(0.001;0.006)

* *Statistically significant at* α=0.01

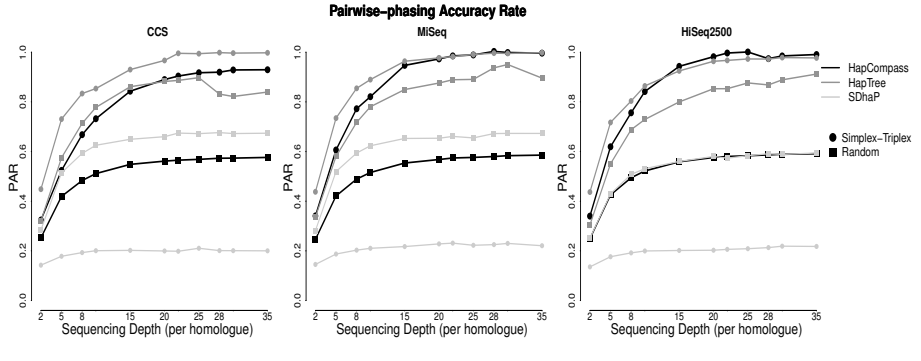
Point estimates and 99% Wald-type confidence intervals for the effects of haplotyping methods: HapTree, SDhaP and HapCompass (the reference) and the sequencing technologies: MiSeq, HiSeq2500 and CCS (the reference), on 5 measures of haplotyping quality: Phasing Accuracy Rate (PAR), SNP Missing Rate (SMR), Incorrect Dosage Rate (IDR), Positive Predictive Value of the called SNPs (PPV) and the Number of Gaps in estimates per SNP (NGPS).

2.3.5. Similarity between homologues eases haplotyping with Illumina

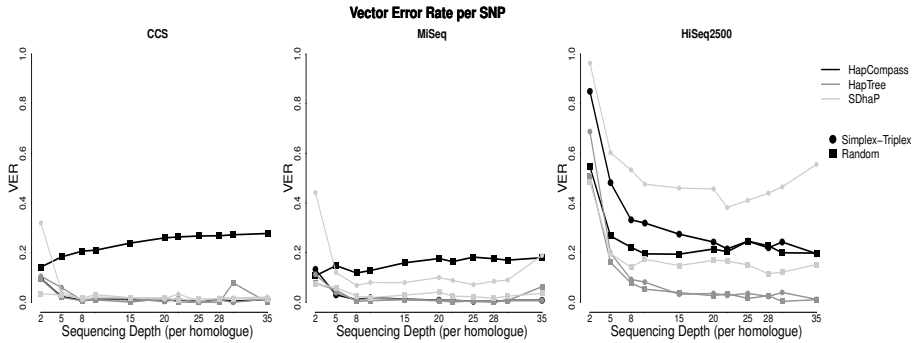
Similarity between homologues can have a large effect on haplotyping. This similarity often occurs when random mating is violated, e.g. in inbred or isolated populations. To investigate this, we simulated simplex-triplex individuals, i.e. tetraploid individuals consisting of two different genomes with dosages of 1 and 3. We generated paired-end MiSeq and HiSeq 2500 reads (800bp insert-size), as well as 1kb CCS read of PacBio, and evaluated the estimated haplotypes.

On this data, the performances of HapTree (with Illumina reads) and HapCompass improve over the original simulation, while the performance of SDhaP deteriorates significantly (Figure 2.5). In particular, the similarity between homologues resulted in a decreased accuracy for SDhaP (Figure 2.5-a, PAR around 0.2) and incorrect dosage estimates for more than half of the SNPs (Figure 2.5-c, IDR of 0.55), regardless of the sequencing method.

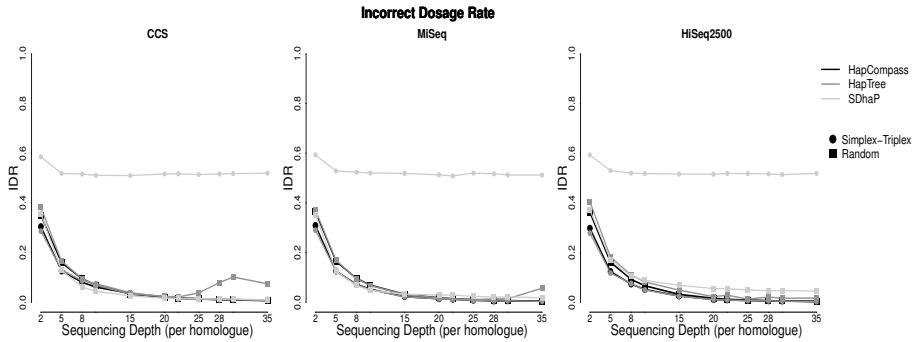
These results demonstrate the differences between the MEC (Minimum Error Correction) approach to haplotyping and other approaches. MEC is sensitive to (local) similarities between homologues, as they lead to approximately identical MEC scores for several different phasings, causing SDhaP to report a suboptimal solution. In contrast, the performances of HapCompass (MWER approach) and HapTree (relative likelihood approach) improve, at least when using Illumina sequencing (Figure 2.5-a, b). Having more similar fragments simplifies construction of the maximum spanning tree in the Compass graph and makes the branching and pruning of the solution tree of HapTree more accurate by enhancing the relative likelihoods of correct partial phasings. No improvement was observed, however, with HapTree using CCS reads, due to increasing failure rates (FR, Figure 2.5-e) caused by time-out errors.



(a)



(b)



(c)

Figure 2.5

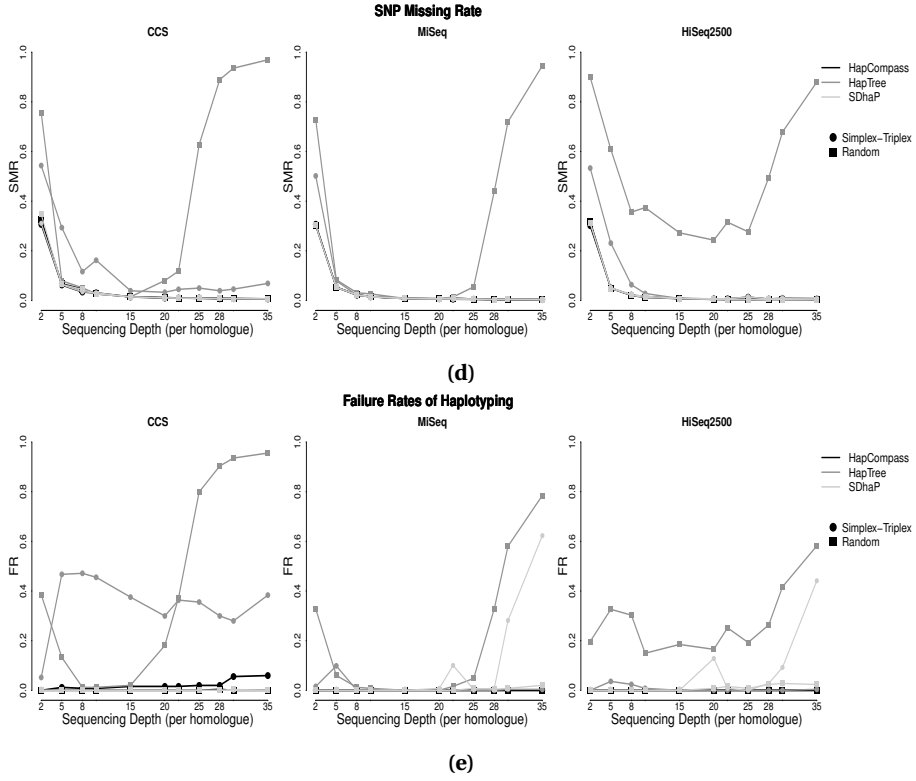


Figure (cont.) 2.5: Plots of haplotype estimation quality measures: (a) PAR, (b) VER, (c) IDR, (d) SMR and (e) FR as a function of ploidy level using HapCompass (black), HapTree (gray) and SDhaP (light gray), for simulated $20kb$ simplex-triplex tetraploid genomes (circle) compared to genomes with random haplotype dosages (square). Sequencing was performed in silico for paired-end HiSeq 2500 reads with $800bp$ insert-size.

These results show that the underlying algorithms lead to different sensitivities to homologue similarity, with MEC-based approaches yielding incorrect results and other methods demanding increasing computation time.

2.3.6. SNP density mainly influences continuity of haplotype estimates

In genomes with a lower SNP density than the highly heterozygous potato, *S. tuberosum*, fragments will overlap less often, which can influence the quality of haplotyping. To determine the effect of SNP density, we simulated tetraploid genomes with average SNP densities ranging from 1 SNP per 22 base pairs, the average density for potato, to 1 SNP per 110 base pair, and estimated the haplotypes using Illumina paired-end reads with an insert-size of $800bp$, as well as $1kb$ CCS reads of PacBio, at a sequencing depth of 15× per homologue. Increasing numbers of gaps (NGPS, Figure 2.6-a) and a decrease in completeness (SMR, Figure 2.6-b) were observed in the estimated haplotypes at lower

densities for all three haplotyping methods. The effect of the SNP density was, however, not manifest on the other haplotyping quality measures (Supp. Figure 5).

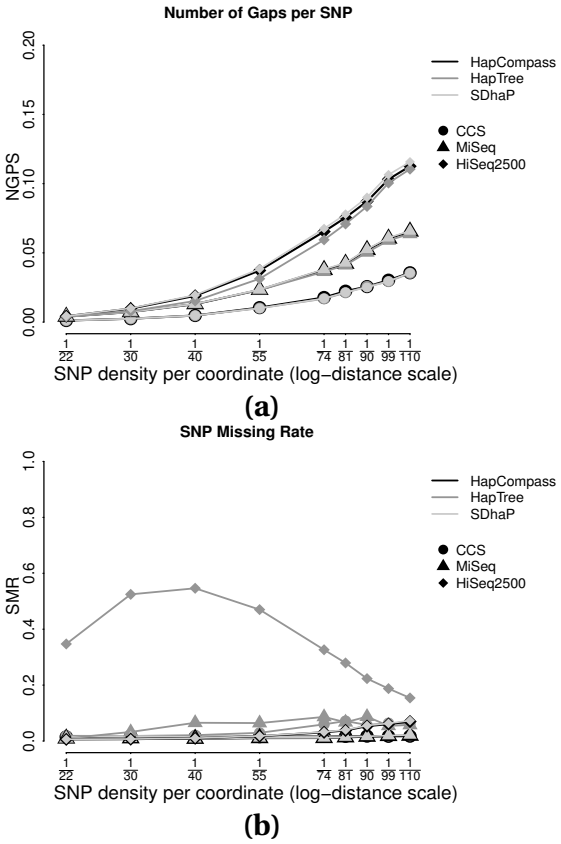


Figure 2.6: Plots of haplotype estimation quality measures: (a) NGPS, (b) SMR as a function of SNP density (at logarithmic distance scale) using HapCompass (black), HapTree (gray) and SDhaP (light gray), for simulated 20*kb* tetraploid *S. tuberosum* genomes. Sequencing was performed in silico for paired-end MiSeq (triangle) and HiSeq 2500 (rhombus) with 800*bp* insert-size, as well as for PacBio-CCS of 1*kb* length (circle), at a depth of 15×.

2.3.7. At higher ploidy levels, HapCompass is the best method to use

In order to investigate whether our findings for tetraploid genomes hold for other ploidy levels, we performed simulations with ploidy levels of 3-12 (Section 2.2.2-c). We simulated paired-end HiSeq 2500 reads with an insert-size of 800*bp*, as it gave high quality estimates in tetraploids and was more practical than the competitive sequencing options, at 5×, 15× and 20× sequencing depths per homologue.

The accuracy of HapTree and SDhaP decreases markedly with increasing ploidy level, up to 30% for 12*n* (PAR, Figure 2.7-a), while the performance of HapCompass remained stable. Likewise, the completeness of HapTree decreased (SMR, Figure 2.7-b) and failure

rates for both HapTree and SDhaP increased (Figure 2.7-c). Although the performance of the methods at each ploidy level was relatively better at higher sequencing depths, the deterioration of the haplotype estimation quality followed a similar pattern with the increase in ploidy, regardless of the depth.

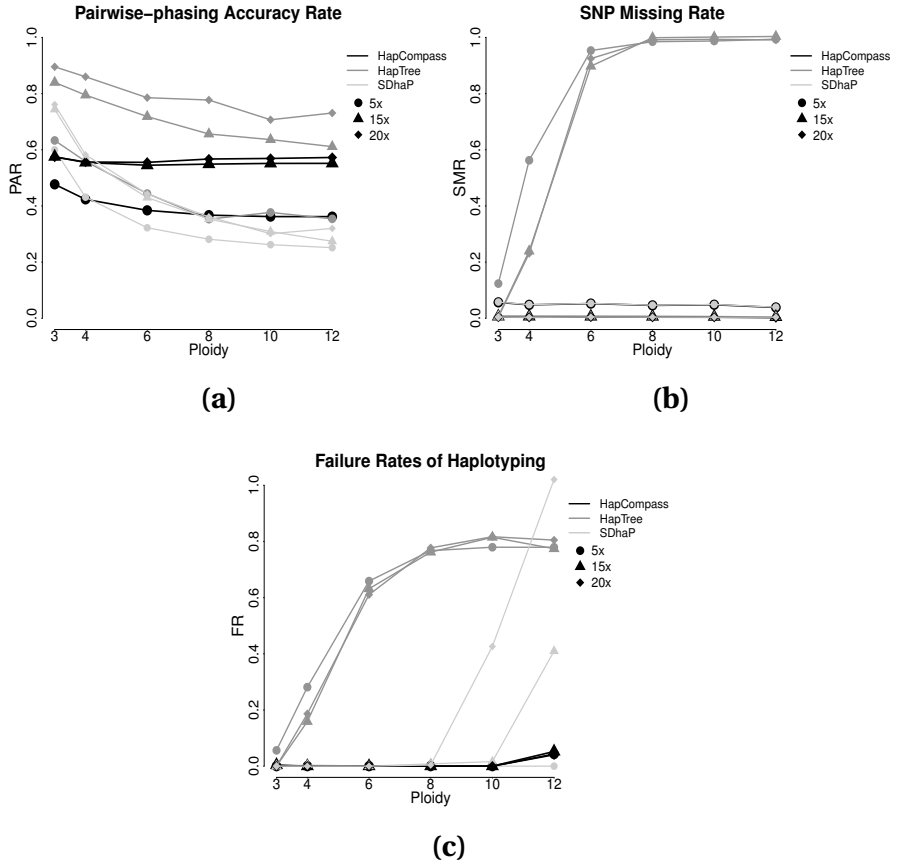


Figure 2.7: Plots of haplotype estimation quality measures: (a) PAR, (b) SMR and (c) FR as a function of ploidy level using HapCompass (black), HapTree (gray) and SDhaP (light gray), for simulated $20kb$ $3n$, $4n$, $6n$, $8n$, $10n$ and $12n$ genomes. Sequencing was performed in silico for paired-end HiSeq 2500 with $600bp$ insert-size. Three sequencing depth were used per homologue: 5x (circle), 15x (triangle) and 20x (rhombus).

Overall, none of the haplotyping methods is equipped to deal with high levels of ploidy: either they break down (HapTree, SDhaP) or are inaccurate (HapCompass).

2.3.8. SDhaP yields best results using long reads

Among the 3 tested haplotyping methods, SDhaP is the only method relying on MEC to select the best estimate. While MEC is an efficient criterion for diploid haplotyping, it

may not be able to distinguish the estimates in presence of more than two homologues as several estimates can have the same error correction score. This situation can especially occur when the homologues have local similarities and the SNP-fragment lengths are small. HapTree and HapCompass try to surmount this problem by considering complex criteria that result in more accurate estimates with short reads of Illumina and even $1kb$ reads of CCS. Nevertheless, these criteria lose their advantage with long sequence reads, $5kb$ CCS and $10kb$ CLR in our study. These longer reads also increase the failure frequency of HapTree. The MEC criterion of SDhaP performs in contrast very well when the reads are long enough to distinguish the homologues, requiring the least computation time (Figure 2.9-a) and providing accurate results using $5kb$ CCS (Figure 2.8-a).

2.3.9. Erroneous long reads lead to low accuracy, high SNP missing rate and many false SNPs in the estimates

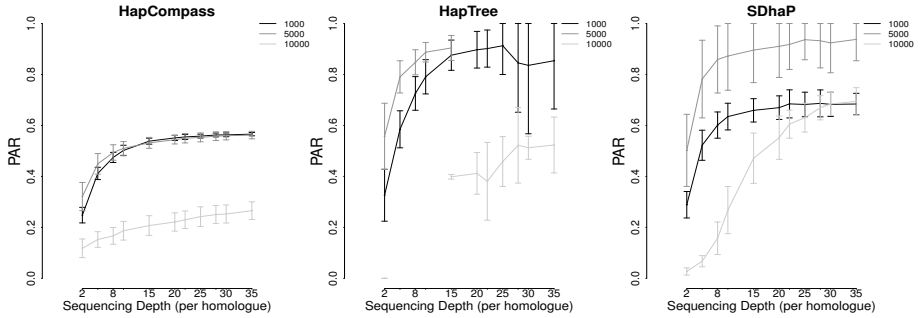
Recently, the generation of very long reads spanning tens of thousands of nucleotides has become a reality using technologies such as Oxford Nanopore and PacBio, at the expense of the precision of base calling [43]. Such lengthy reads are potentially ideal candidates for haplotyping as they can cover many variants and provide enough overlaps for accurate haplotype reconstruction [2]. However, our simulations using $10kb$ CLR reads of PacBio, with an average accuracy of 82%, show that these reads lead to inferior estimates for polyploids compared to paired-end Illumina reads and CCS reads. In particular, many spurious SNPs will be present (Figure 2.8-d) and many of the original SNPs will be missing in the estimates (Figure 2.8-e). In addition, wrong dosages abound in the estimated haplotypes (Figure 2.8-c). While increasing the coverage helps improve the estimation to some extent, especially with SDhaP (Figures 2.8-a, c, e), our results do not encourage the use of erroneous long reads for the estimation of polyploid haplotypes as achieving the extremely high coverages needed is usually not practical.

2.3.10. HapTree requires most resources

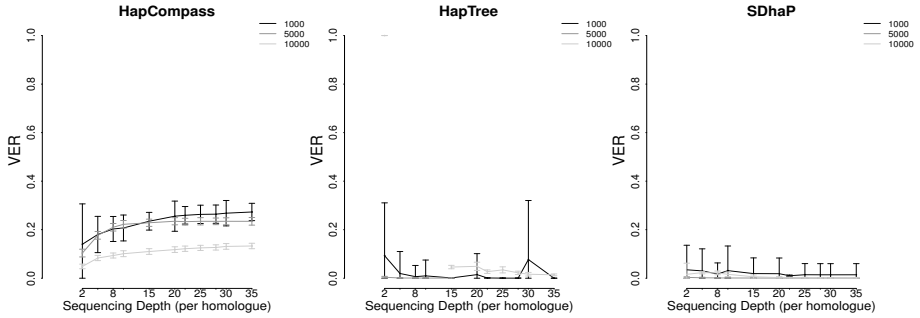
Computational efficiency is an important feature of every complex algorithm, such as the haplotyping algorithms discussed in this paper. Therefore, we measured the memory and time consumption of each algorithm for various ploidy levels, sequencing coverages and genome lengths. Using HiSeq 2500 and paired-end libraries with an insert-size of $800bp$ for the simulation of sequencing, we tested the effect of sequencing depth with tetraploid individuals and genomes of length $10kb$, and the effect of genome length with tetraploid individuals sequenced at an average depth of $10\times$ per homologue. Other settings were the same as for *S. tuberosum* (Section 2.2.2), for each condition generating 50 individuals from 50 randomly selected regions, i.e. one individual per region, with a time limit of 7200 seconds. Fixing the depth to $10\times$ per homologue and the genome length to $10kb$, the effect of ploidy was also investigated in a similar manner.

The analyses were run on multicore 2.6 GHz Intel-Xeon processors. For each run, the total CPU-time and physical memory consumption was measured using the Unix `getrusage` routine. HapTree clearly consumed most time and memory resources (Figure 2.10), increasing with genome length (Figure 2.10-a, b) and ploidy level (Figure 2.10-c, d). This increase was much less for HapCompass and SDhaP.

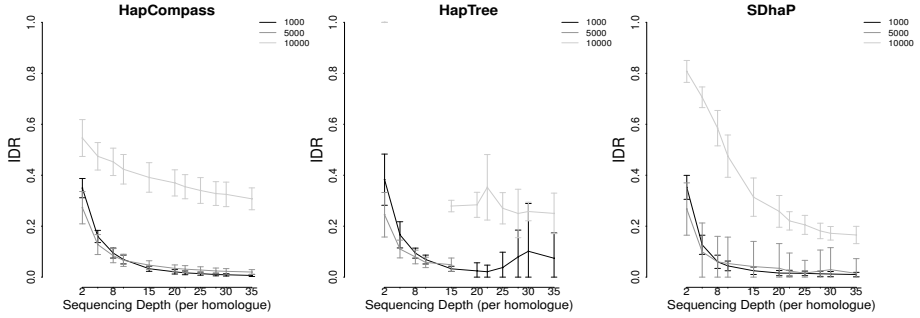
While sequencing depth was less influential, HapTree used much more time and



(a)



(b)



(c)

Figure 2.8

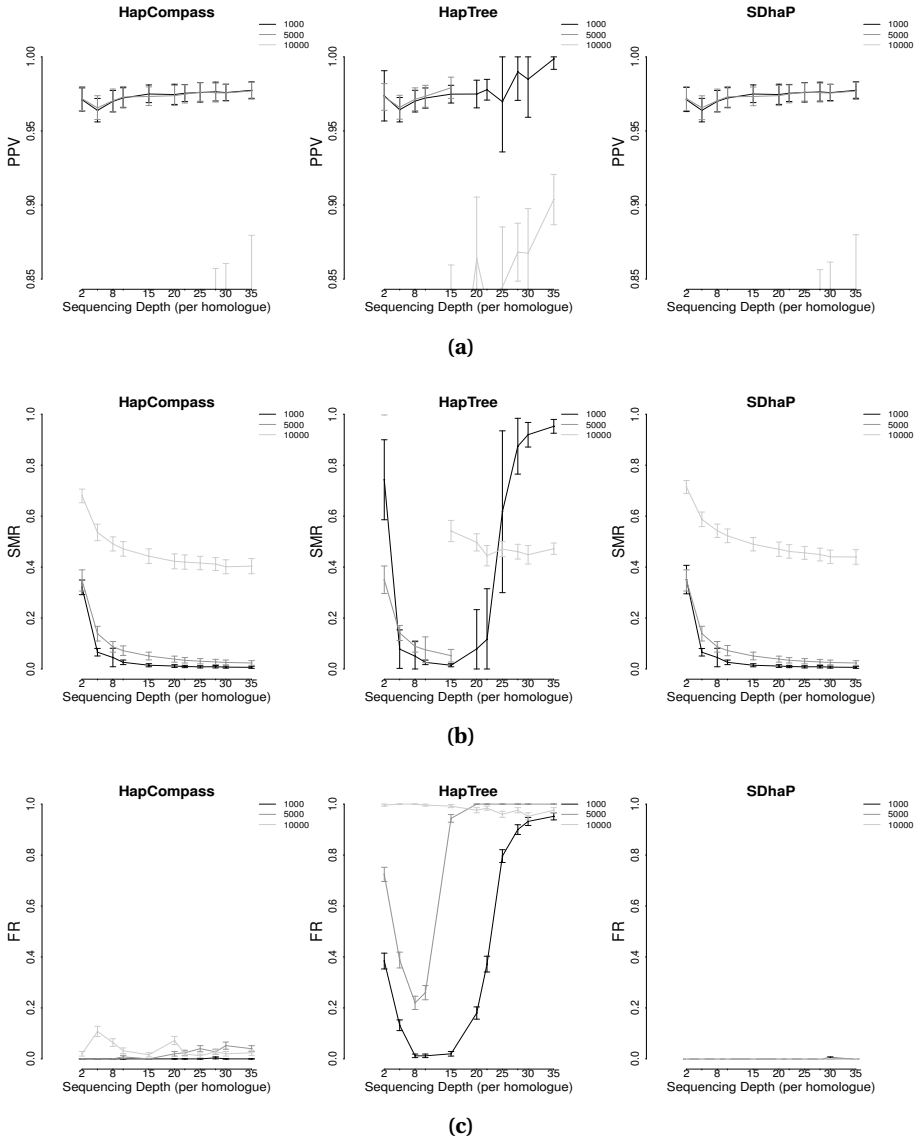
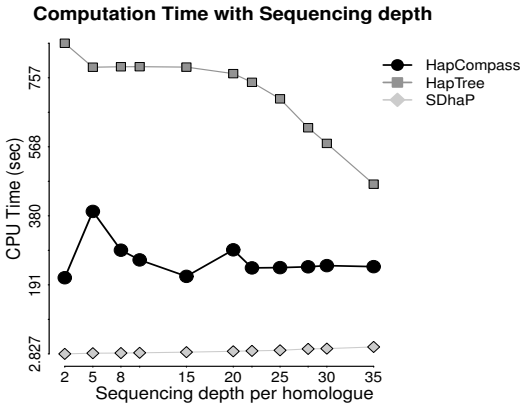


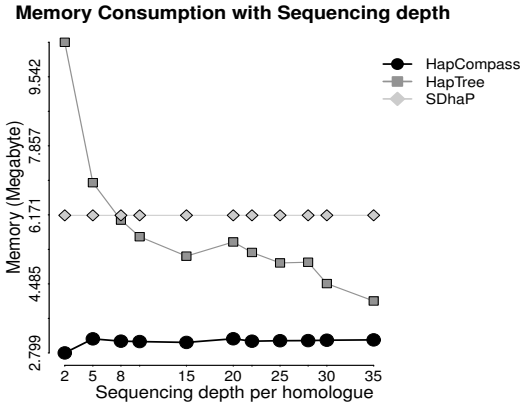
Figure (cont.) 2.8: Plots of haplotype estimation quality measures: (a) PAR, (b) VER, (c) IDR, (d) SMR, (e) PPV and (f) FR as a function of sequencing depth per homologue using HapCompass (left), HapTree (middle) and SDhaP (right), for simulated PacBio read: CCS 1*kb* (black), CCS 5*kb* (gray) and CLR 10*kb* (light gray).

memory at the lowest sequencing depth, 5× per homologue, falling rapidly with an increase in depth up to 10× per homologue and remaining almost constant afterwards up

to a depth of 20× per homologue (Figure 2.10-e, f). At depths of higher than 20× per homologue, the computation time fell rapidly again due to the premature failure of the algorithm as discussed in Section 2.3.2.



(a)



(b)

Figure 2.9: (a) Computation time (in seconds) (b) Physical memory consumption (in Megabyte) as a function of sequencing depth per homologue with three haplotype estimation softwares: HapCompass (black circle), HapTree (gray square) and SDhaP (light gray rhombus), using 10*kb* continuous longs reads of PacBio for tetraploid genomes of length 20*kb*.

2.4. Conclusion

We evaluated three algorithms for single individual haplotype estimation in polyploids: HapCompass, HapTree and SDhaP, and investigated the effects of SNP density, similarity between homologues, ploidy level, sequencing technology, sequencing depth and DNA

library size on the estimation quality using several measures of quality (Table 2.2) and through extensive simulation experiments. This yielded insight about the performance of haplotype estimation methods in practical situations. For this purpose, we have developed a realistic pipeline that can be used as basis for the benchmarking of single individual haplotyping softwares in future.

Our results show that HapTree can produce the best triploid and tetraploid haplotype estimates, followed by SDhaP and HapCompass. HapCompass is the best method to use with ploidy levels over $6n$, although its performance is not good in an absolute sense. We showed that sequencing depth was the most important factor determining the quality of haplotype estimation, and paired-end short reads of Illumina with a large insert can perform as well as long CCS reads of the same total size now possible with PacBio. For accurate haplotyping, we therefore suggest an average depth of between 5-20× per

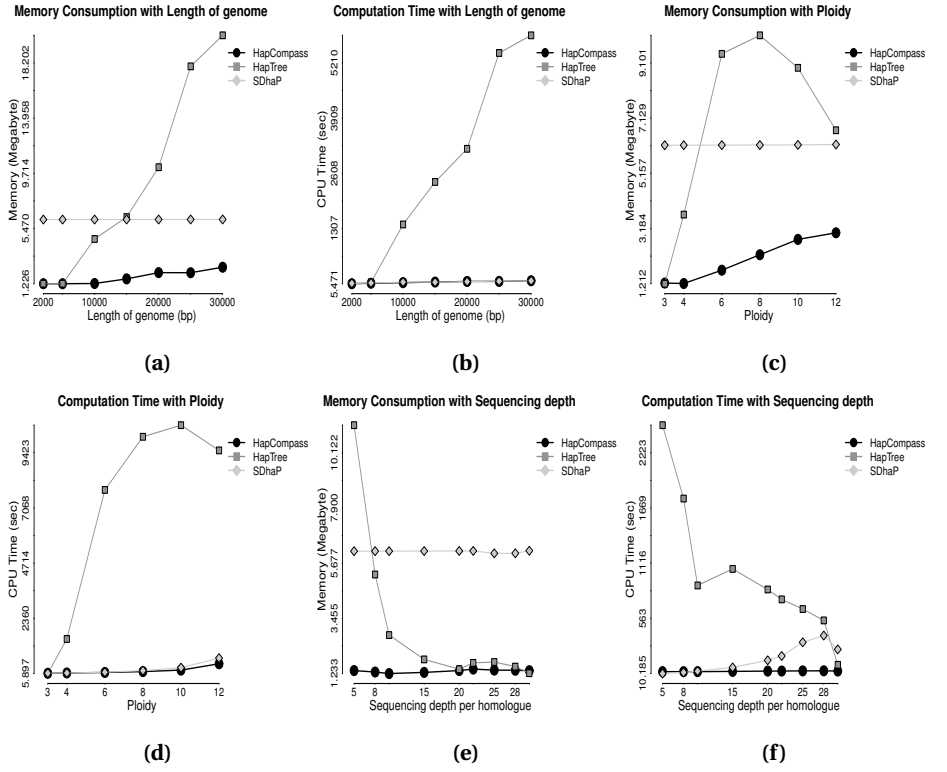


Figure 2.10: Plots of physical memory consumption (in Megabyte) with: length of genome (in base-pair) (a), ploidy (c) and sequencing depth (e), and plots of computation time (in seconds) with length of genome (in base-pair) (b), ploidy (d) and sequencing depth per homologue (f), for three haplotype estimation softwares: HapCompass (black circle), HapTree (gray square) and SDhaP (light gray rhombus). Sequencing was based on HiSeq 2500 paired-end technology with 800bp insert-size.

homologue with paired-end reads and an insert-size of 600-800*bp*.

In addition to the estimation quality, we investigated computation time and memory consumption of each algorithm under various settings to compare their efficiency. We showed that on average, HapTree requires the most computation time and memory, and its use of resources is highly dependent on the length of the genomic region, the ploidy level and the sequencing depth. Combined with the frequent failure to complete the estimation, this raises difficulties for applying HapTree on practical problems where the aim is to reconstruct long-range haplotypes.

Our findings show that while state-of-the-art single individual haplotype estimation algorithms produce promising results for triploid and tetraploid organisms over a limited genomic region, their performance rapidly decreases at higher ploidy levels and their resource use prohibits application to large genomic regions. The probability-based algorithm of HapTree produces the most accurate estimates but also requires the most computation time and memory. We believe it is worth investigating whether the HapTree approach can be made robust when faced with larger problems while maintaining its accuracy, e.g. using a divide-and-conquer approach or by adjusting the branching and pruning parameters according to the length of the genome, the ploidy level and the sequencing coverage. The variant calling error model could also be upgraded to be specific to the applied sequencing strategy and technology.

Finally, the performance of haplotyping methods on individual organisms could be greatly improved if it could also incorporate parental and sib information if available, e.g. in mapping populations relevant to plant and animal breeding studies. While the evaluated algorithms ignore these information, it can be extremely helpful to increase the precision of genotype calling when the average sequencing depth is low or to favor/disfavor some of the haplotypes a priori based on their expected frequency in the population. Such enhancements will prove essential to help understand the complex genetics found in many polyploid organisms and, in the long run, to better understand the rules governing genome organization.

Software

The simulation pipeline and its components can be downloaded at the software page of the Bioinformatics group, Wageningen University & Research: `http://www.bif.wur.nl`

Supplementary Figures and Data

The supplementary figures and data referenced in this chapter are available online at: `https://doi.org/10.1093/bib/bbw126`

References

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

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

[3] Aguiar, D., Wong, W. S., and Istrail, S. (2014). Tumor haplotype assembly algorithms for cancer genomics. In *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, page 3. World Scientific.

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

[5] Bansal, V., Halpern, A. L., Axelrod, N., and Bafna, V. (2008). An MCMC algorithm for haplotype assembly from whole-genome sequence data. *Genome Research*, **18**(8), 1336–1346.

[6] Bates, D., Sarkar, D., Bates, M. D., and Matrix, L. (2007). The lme4 package. *R package Version*, **2**(1).

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

[8] Birney, E. and Soranzo, N. (2015). Human genomics: The end of the start for population sequencing. *Nature*, **526**(7571), 52–53.

[9] Broad Institute ((Accessed: 2016/01/13; version 2.9.0)). Picard tools. http://broadinstitute.github.io/picard/.

[10] Browning, S. R. and Browning, B. L. (2011). Haplotype phasing: Existing methods and new developments. *Nature Reviews Genetics*, **12**(10), 703–714.

[11] Caputo, M., Rivolta, C. M., Gutnisky, V. J., Gruñeiro-Papendieck, L., Chiesa, A., Medeiros-Neto, G., González-Sarmiento, R., and Targovnik, H. M. (2007). Recurrence of the p. R277X/p. R1511X compound heterozygous mutation in the thyroglobulin gene in unrelated families with congenital goiter and hypothyroidism: haplotype analysis using intragenic thyroglobulin polymorphisms. *Journal of Endocrinology*, **195**(1), 167–177.

[12] Castiglione, C., Deinard, A., Speed, W., Sirugo, G., Rosenbaum, H., Zhang, Y., Grandy, D., Grigorenko, E., Bonne-Tamir, B., Pakstis, A., *et al.* (1995). Evolution of haplotypes at the DRD2 locus. *American Journal of Human Genetics*, **57**(6), 1445.

[13] Chin, C.-S., Peluso, P., Sedlazeck, F. J., Nattestad, M., Concepcion, G. T., Clum, A., Dunn, C., O'Malley, R., Figueroa-Balderas, R., Morales-Cruz, A., *et al.* (2016). Phased Diploid Genome Assembly with Single Molecule Real-Time Sequencing. *bioRxiv*, page 056887.

[14] Clark, A. G., Eisen, M. B., Smith, D. R., Bergman, C. M., Oliver, B., Markow, T. A., Kaufman, T. C., Kellis, M., Gelbart, W., Iyer, V. N., *et al.* (2007). Evolution of genes and genomes on the Drosophila phylogeny. *Nature*, **450**(7167), 203–218.

[15] Collins, F. S., Morgan, M., and Patrinos, A. (2003). The Human Genome Project: Lessons from large-scale biology. *Science*, **300**(5617), 286–290.

[16] Consortium, T. G. *et al.* (2012). The tomato genome sequence provides insights into fleshy fruit evolution. *Nature*, **485**(7400), 635–641.

[17] Das, S. and Vikalo, H. (2015). SDhaP: Haplotype assembly for diploids and polyploids via semi-definite programming. *BMC Genomics*, **16**(1), 260.

[18] Ganal, M. W., Altmann, T., and Röder, M. S. (2009). SNP identification in crop plants.

Current opinion in plant biology, **12**(2), 211–217.

[19] Garrison, E. and Marth, G. (2012). Haplotype-based variant detection from short-read sequencing. *arXiv preprint arXiv:1207.3907 [q-bio.GN]*.

[20] Genome Sequencing Consortium, P. *et al.* (2011). Genome sequence and analysis of the tuber crop potato. *Nature*, **475**(7355), 189–195.

[21] Gibbs, R. A., Belmont, J. W., Hardenbol, P., Willis, T. D., Yu, F., Yang, H., Ch'ang, L.-Y., Huang, W., Liu, B., Shen, Y., *et al.* (2003). The international HapMap project. *Nature*, **426**(6968), 789–796.

[22] Glusman, G., Cox, H. C., and Roach, J. C. (2014). Whole-genome haplotyping approaches and genomic medicine. *Genome Medicine*, **6**(9), 73.

[23] Goddard, M. E. and Hayes, B. J. (2009). Mapping genes for complex traits in domestic animals and their use in breeding programmes. *Nature Reviews Genetics*, **10**(6), 381–391.

[24] Groenen, M. A., Archibald, A. L., Uenishi, H., Tuggle, C. K., Takeuchi, Y., Rothschild, M. F., Rogel-Gaillard, C., Park, C., Milan, D., Megens, H.-J., *et al.* (2012). Analyses of pig genomes provide insight into porcine demography and evolution. *Nature*, **491**(7424), 393–398.

[25] Hamblin, M. T. and Jannink, J.-L. (2011). Factors affecting the power of haplotype markers in association studies. *The Plant Genome*, **4**(2), 145–153.

[26] Hickey, J. M., Gorjanc, G., Varshney, R. K., and Nettelblad, C. (2015). Imputation of Single Nucleotide Polymorphism Genotypes in Biparental, Backcross, and Topcross Populations with a Hidden Markov Model. *Crop Science*, **55**(5), 1934–1946.

[27] Hillier, L. W., Miller, W., Birney, E., Warren, W., Hardison, R. C., Ponting, C. P., Bork, P., Burt, D. W., Groenen, M. A., Delany, M. E., *et al.* (2004). Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature*, **432**(7018), 695–716.

[28] Howie, B., Fuchsberger, C., Stephens, M., Marchini, J., and Abecasis, G. R. (2012). Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nature genetics*, **44**(8), 955–959.

[29] Huang, W., Li, L., Myers, J. R., and Marth, G. T. (2012). Art: A next-generation sequencing read simulator. *Bioinformatics*, **28**(4), 593–594.

[30] Kaul, S., Koo, H. L., Jenkins, J., Rizzo, M., Rooney, T., Tallon, L. J., Feldblyum, T., Nierman, W., Benito, M. I., Lin, X., *et al.* (2000). Analysis of the Genome Sequence of the flowering plant Arabidopsis thaliana. *nature*, **408**(6814), 796–815.

[31] Kim, S., Park, M., Yeom, S.-I., Kim, Y.-M., Lee, J. M., Lee, H.-A., Seo, E., Choi, J., Cheong, K., Kim, K.-T., *et al.* (2014). Genome sequence of the hot pepper provides insights into the evolution of pungency in Capsicum species. *Nature Genetics*, **46**(3), 270–278.

[32] Krasileva, K. V., Buffalo, V., Bailey, P., Pearce, S., Ayling, S., Tabbita, F., Soria, M., Wang, S., Akhunov, E., Uauy, C., *et al.* (2013). Separating homeologs by phasing in the tetraploid wheat transcriptome. *Genome Biology*, **14**(6), 1.

[33] Kuleshov, V. (2014). Probabilistic single-individual haplotyping. *Bioinformatics*, **30**(17), i379–i385.

[34] LaFramboise, T. (2009). Single nucleotide polymorphism arrays: A decade of bi-

ological, computational and technological advances. *Nucleic Acids Research*, page gkp552.

[35] Lancia, G., Bafna, V., Istrail, S., Lippert, R., and Schwartz, R. (2001). SNPs problems, complexity, and algorithms. In *Algorithms–ESA 2001*, pages 182–193. Springer.

[36] Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv preprint arXiv:1303.3997*.

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

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

[39] Liu, L., Li, Y., Li, S., Hu, N., He, Y., Pong, R., Lin, D., Lu, L., and Law, M. (2012). Comparison of next-generation sequencing systems. *BioMed Research International*, **2012**.

[40] Lorenz, A. J., Hamblin, M. T., Jannink, J.-L., *et al.* (2010). Performance of single nucleotide polymorphisms versus haplotypes for genome-wide association analysis in barley. *PLoS One*, **5**(11), e14079.

[41] Mammadov, J., Aggarwal, R., Buyyarapu, R., and Kumpatla, S. (2012). SNP markers and their impact on plant breeding. *International Journal of Plant Genomics*, **2012**.

[42] McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., *et al.* (2010). The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, **20**(9), 1297–1303.

[43] Metzker, M. L. (2010). Sequencing technologies–the next generation. *Nature Reviews Genetics*, **11**(1), 31–46.

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

[45] Ozkan, H., Levy, A. A., and Feldman, M. (2001). Allopolyploidy-induced rapid genome evolution in the wheat (Aegilops–Triticum) group. *The Plant Cell*, **13**(8), 1735–1747.

[46] Quail, M. A., Smith, M., Coupland, P., Otto, T. D., Harris, S. R., Connor, T. R., Bertoni, A., Swerdlow, H. P., and Gu, Y. (2012). A tale of three next generation sequencing platforms: Comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics*, **13**(1), 341.

[47] R Core Team (2015). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

[48] Rizzi, R., Bafna, V., Istrail, S., and Lancia, G. (2002). Practical algorithms and fixed-parameter tractability for the single individual SNP haplotyping problem. In *Algorithms in Bioinformatics*, pages 29–43. Springer.

[49] Sachidanandam, R., Weissman, D., Schmidt, S. C., Kakol, J. M., Stein, L. D., Marth, G., Sherry, S., Mullikin, J. C., Mortimore, B. J., Willey, D. L., *et al.* (2001). A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature*, **409**(6822), 928–933.

[50] Thompson, R. (1980). Maximum likelihood estimation of variance components. *Statistics: A Journal of Theoretical and Applied Statistics*, **11**(4), 545–561.

[51] Uitdewilligen, J. G., Wolters, A.-M. A., D'hoop, B. B., Borm, T. J., Visser, R. G., and van Eck, H. J. (2013). A Next-Generation Sequencing Method for Genotyping-by-Sequencing of Highly Heterozygous Autotetraploid Potato. *PLoS One*, **8**(5), e62355.

[52] Visscher, P. M., Brown, M. A., McCarthy, M. I., and Yang, J. (2012). Five years of GWAS discovery. *The American Journal of Human Genetics*, **90**(1), 7–24.

[53] Vrijenhoek, R. C. (2006). Polyploid hybrids: Multiple origins of a treefrog species. *Current Biology*, **16**(7), R245–R247.

[54] Wang, R.-S., Wu, L.-Y., Li, Z.-P., and Zhang, X.-S. (2005). Haplotype reconstruction from SNP fragments by Minimum Error Correction. *Bioinformatics*, **21**(10), 2456–2462.

[55] Wang, R.-S., Wu, L.-Y., Zhang, X.-S., and Chen, L. (2006). A Markov chain model for haplotype assembly from SNP fragments. *Genome Informatics*, **17**(2), 162–171.

[56] Wang, Y., Feng, E., and Wang, R. (2007). A clustering algorithm based on two distance functions for MEC model. *Computational Biology and Chemistry*, **31**(2), 148–150.

[57] Wu, J., Wang, J., and Chen, J. (2009a). A practical algorithm based on particle swarm optimization for haplotype reconstruction. *Applied Mathematics and Computation*, **208**(2), 363–372.

[58] Wu, L.-Y., Li, Z., Wang, R.-S., Zhang, X.-S., and Chen, L. (2009b). Self-organizing map approaches for the haplotype assembly problem. *Mathematics and Computers in Simulation*, **79**(10), 3026–3037.

[59] Yabe, T., Ge, X., and Pelegri, F. (2007). The zebrafish maternal-effect gene cellular atoll encodes the centriolar component sas-6 and defects in its paternal function promote whole genome duplication. *Developmental Biology*, **312**(1), 44–60.

[60] Zhao, Y.-Y., Wu, L.-Y., Zhang, J.-H., Wang, R.-S., and Zhang, X.-S. (2005). Haplotype assembly from aligned weighted SNP fragments. *Computational Biology and Chemistry*, **29**(4), 281–287.

3

TriPoly: haplotype estimation for polyploids using sequencing data of related individuals

This chapter has been published with minor modifications in: Ehsan Motazedi, Dick de Ridder, Richard Finkers, Samantha Baldwin, Susan Thomson, Katrina Monaghan and Chris Maliepaard, **TriPoly: haplotype estimation for polyploids using sequencing data of related individuals**, Bioinformatics, Volume 34, Issue 22, 15 November 2018, Pages 3864-3872

Abstract

Knowledge of haplotypes, i.e. phased and ordered marker alleles on a chromosome, is essential to answer many questions in genetics and genomics. By generating short pieces of DNA sequence, high-throughput modern sequencing technologies make estimation of haplotypes possible for single individuals. In polyploids, however, haplotype estimation methods usually require deep coverage to achieve sufficient accuracy. This often renders sequencing-based approaches too costly to be applied to large populations needed in studies of Quantitative Trait Loci (QTL).

We propose a novel haplotype estimation method for polyploids, TriPoly, that combines sequencing data with Mendelian inheritance rules to infer haplotypes in parent-offspring trios. Using realistic simulations of both short and long-read sequencing data for banana (*Musa acuminata*) and potato (*Solanum tuberosum*) trios, we show that TriPoly yields more accurate progeny haplotypes at low coverages compared to existing methods that work on single individuals. We also apply TriPoly to phase SNPs on chromosome 5 for a family of tetraploid potato with 2 parents and 37 offspring sequenced with an RNA capture approach. We show that TriPoly haplotype estimates differ from those of the other methods mainly in regions with imperfect sequencing or mapping difficulties, as it does not rely solely on sequence reads and aims to avoid phasings that are not likely to have been passed from the parents to the offspring.

3.1. Introduction

Haplotypes are defined as sequences of consecutive nucleotides over a chromosome, which normally shares high similarity with $k-1$ other chromosomes in diploid ($k=2$) and polyploid ($k>2$) organisms. These k homologous chromosomes can nevertheless have important differences in the form of nucleotide substitutions or insertions and deletions, leading to genotypic (and phenotypic) diversity within an outcrossing population, e.g. of the diploid ($k=2$) human (*Homo sapiens*), tetraploid ($k=4$) African clawed frog (*Xenopus laevis*) or tetraploid potato (*Solanum tuberosum*), or between inbred lines of autogamous species, e.g. hexaploid ($k=6$) wheat (*Triticum aestivum*). The assignment of these variant forms, i.e. alleles, to the chromosomes is called *phasing* or *haplotyping*. In this context, phasing may also refer to the set of phased homologues, $H=\{h_1,h_2,...,h_k\}$ with k being the ploidy level and h_i ($i=1,...,k$) being the haplotype corresponding to the i^{th} homologue.

As phasing is uninformative at genomic positions with identical nucleotides over all the homologous chromosomes, i.e. at homozygous sites, haplotypes are usually defined as sequences of alleles at heterozygous sites over a chromosome. By this definition, 2^n haplotypes are theoretically possible for a region covering n bi-allelic Single Nucleotide Polymorphisms (SNPs), which is the most abundant form of genomic variation among individuals of the same species [25]. However, often far fewer haplotypes are actually found in a population.

While high-throughput genotyping assays, such as SNP arrays, exist for efficient determination of unphased SNPs, direct determination of haplotypes is much more complicated and usually requires laborious and expensive techniques such as bacterial cloning, chromosome microdissection or allele-specific PCR [11, 22, 29]. However, unphased SNPs provide less knowledge about an individual's phenotype compared to phased SNPs, as both gene expression and protein function can be affected by the heterozygous variants being in *cis* or *trans* with other variants [28]. Besides, haplotypes can be used as multi-allelic markers, offering more statistical power compared to single SNPs for genetic linkage and association studies [26].

Several computational methods have therefore been proposed to indirectly infer the phasing from available genotype data. These can be divided into three main categories. Methods in the first category, such as *Merlin* [1] and *TetraOrigin* [31], target pedigrees and aim to determine the most likely haplotypes using the segregation of marker alleles, taking into account the genetic distances between the marker loci. These methods can be applied to SNPs that are far enough apart to be informative about linkage, and are especially useful with large pedigrees. Methods in the second category, such as *Beagle* [6], *SHAPEIT* [9] and *Eagle* [20] target populations with unknown pedigrees and are based on coalescence theory, trying to obtain a set of highly frequent haplotypes in the population compatible with the genotype data. Methods in the third category, such as *HapCut* [3], *HapCompass* [2], *HapTree* [4] and *SDhaP* [8], use sequence read data and target single individuals, exploiting the fact that a sequence read that contains at least two SNPs reveals the phasing of the homologue from which it has originated at the contained SNP sites. The aim of these methods is therefore to assign the reads of a single individual to k groups, corresponding to the homologues of a k-ploid, and to obtain the consensus sequence of the reads within each group to reconstruct the haplotypes.

All of these approaches have limitations in terms of the ploidy level (k) and the required marker density. For the methods in the third category, sequencing depth and read length are also limiting factors. As an example, Merlin, Beagle, SHAPEIT, Eagle and HapCut can only phase diploids ($k = 2$) and the TetraOrigin algorithm is only applicable to bi-parental tetraploid populations ($k = 4$) for which a linkage map is available. Also, HapCompass, HapTree and SDhaP can fail to reconstruct haplotypes with high quality at low sequence depths or at ploidy levels higher than $k = 4$ [23].

In case parent-progeny relations exist in a sequenced population, it is possible to improve the quality of haplotype estimation by combining the information used in the first and the third categories under a unified scheme. With sequencing experiments becoming cheaper and more efficient, such an approach is of high practical importance as often whole populations are sequenced rather than only genotyped at specific marker loci. An implementation of this unifying framework, called *PedMEC*, has recently been reported by [15] for diploid *trios*, i.e. families with two parents and one offspring. Specifically, PedMEC extends the partial-phasing of sequence reads using their overlaps while penalising meiotic recombination events in each trio. However, the exact dynamic programming approach of [15] rapidly becomes intractable for polyploids, i.e. with $k > 2$, as its complexity increases exponentially with an increase in the ploidy level (Section 3.2).

Here we present a greedy algorithm, *TriPoly*, for phasing a set of SNPs connected by the sequencing reads in parent-offspring trios. Starting at the SNP site with the smallest genomic coordinate, TriPoly extends the phasing one SNP at a time, keeping only the most likely extended phasings to be worked out in the subsequent extension step. In determining the likelihood of each extension, TriPoly considers its compatibility with the sequence reads, as well as the number of recombination events observed by comparing the parental extensions with that of the offspring.

Using quantitative measures, we investigated the quality of haplotype estimates obtained by TriPoly in parent-offspring trios simulated under realistic assumptions with tetraploid × diploid and tetraploid × tetraploid parents. By comparing our results with those obtained using single individual haplotyping methods, we show that TriPoly yields substantially better estimates for the haplotypes of the progeny, especially at low sequencing depths.

Finally, we apply TriPoly to phase SNPs on chromosome 5 for a family of tetraploid potato with 2 parents and 37 offspring, sequenced with an RNA capture approach by paired-end Illumina HiSeq-2000 technology. We show that TriPoly phasings differ from those obtained by the other methods mainly in regions with imperfect sequencing or mapping difficulties, as TriPoly does not rely solely on sequence reads and aims to avoid phasings that are not likely to have been passed from the parents to the offspring.

3.2. Method

3.2.1. A Bayesian approach to obtaining phasing probabilities from sequence reads

In order to establish a probabilistic model for haplotypes, with the sequence reads as data and the base call error and recombination rates as parameters, we must first determine which reads are informative about the phasing. Informative reads need to cover at

least two variants, e.g. SNP sites which are heterozygous for at least one of the trio members (m, f, c), corresponding to mother, father and the offspring (child). As sites that are homozygous in all trio members retain no phasing information, we discard them from the sequence reads and keep only the base-calls corresponding to the variation positions. Therefore in the first step, the SNP sites, $s = 1, 2, ..., l$, are detected over a genomic region and the genotypes $G_s = (G_s^m, G_s^f, G_s^c)$ are estimated at these sites, using efficient algorithms such as FreeBayes [16]. The raw reads of each trio member are then replaced by the so-called *SNP fragments* of length l (Figure 3.1) that each correspond to a read and contain the numerically coded alleles, i.e. 0, 1, 2 or 3 representing the reference and alternative nucleotides, at the SNP sites covered by that read and '-' at positions not called or not covered. To reduce sequencing noise, the positions at which the base-calling quality is lower than a desired threshold can be set to '-' as well. Hereafter, by using the term sequence read, r, we refer to SNP fragments that contain at least two determined positions.

In the next step, one should assign the reads to k compatible sets in which all of the reads have the same allele at their overlaps, and obtain the consensus sequence of each set as the phasing. As shown in Figure 3.1, this process is straightforward for diploids in the absence of sequencing errors. With sequencing errors, however, such an assignment of reads to homologues will be possible only if mismatches are allowed. However, allowing mismatches at sites with no error can lead to incorrect haplotype estimates. Polyploidy results in further complexity, as there may be more than one way to assign the reads to $k > 2$ sets even when no error is present. This can happen for instance when several haplotypes are identical in a phasing solution, e.g. in a 3 SNP tetraploid phasing consisting of 4 homologues: $\left(\begin{smallmatrix}1&0&0\\1&0&0\\1&0&0\\0&1&1\end{smallmatrix}\right)$ in which three identical $(1\,0\,0)$ haplotypes are present. In this example, the reads will be compatible with any phasing as long as it contains both $(1\,0\,0)$ and $(0\,1\,1)$ haplotypes regardless of their dosages, for example $\left(\begin{smallmatrix}1&0&0\\1&0&0\\0&1&1\\0&1&1\end{smallmatrix}\right)$. Therefore, probabilistic models must take the uncertainty caused by the presence of several phasing possibilities and sequencing errors into account.

To build the probabilistic model, we assume an independent binomial error model at each SNP site [4] and assign an error vector, $\vec{\epsilon}_r$, of length l to each read containing the probability of erroneous base-calling at the SNP sites in that read. Using these error rates, the probabilities of possible maternal, paternal and offspring phasings in a trio, represented by H_m, H_f and H_c, respectively, can be derived from the set of sequence reads associated with the trio, $\mathbf{R}$ (consisting of maternal read $\mathbf{R}_m$, paternal reads $\mathbf{R}_f$ and offspring reads $\mathbf{R}_c$). In addition to the reads, we consider meiotic recombination probabilities, θ_s, between SNP $s-1$ and SNP s, represented by vector $\vec{\theta}$ for all $s > 1$ to adjust the probability assigned to each phasing using Mendelian inheritance rules as follows:

$$P(H_m, H_f, H_c|\mathbf{R}, \boldsymbol{\epsilon}, \vec{\theta}) = P(H_m|\mathbf{R}_m, \boldsymbol{\epsilon}_m) \cdot P(H_f|\mathbf{R}_f, \boldsymbol{\epsilon}_f) \cdot P(H_c|\mathbf{R}_c, H_m, H_f, \boldsymbol{\epsilon}_c, \vec{\theta}) \tag{3.1}$$

SNP 1	*SNP* 2	*SNP* 3	*SNP* 4	*SNP* 5	*SNP* 6	*SNP* 7	*SNP* 8	*SNP* 9	*SNP* 10	*SNP* 11	*SNP* 12	*SNP* 13	
0	0	0	0	0	0	0	0	0	0	0	0	0	REF
0	1	0	1	-	-	-	-	-	-	-	-	-	→ h_1
-	1	0	1	0	-	-	-	0	1	-	-	-	→ h_1
-	0	1	0	-	-	-	-	-	-	1	1	-	→ h_2
-	-	-	-	1	1	-	-	-	-	-	-	-	→ h_2
-	-	-	-	-	0	1	0	-	-	0	0	0	→ h_1
0	1	-	-	-	-	-	0	0	1	0	-	-	→ h_1
-	-	1	0	1	-	-	-	-	-	1	1	-	→ h_2
-	-	-	-	-	0	1	0	-	-	-	0	0	→ h_1
-	-	1	0	1	1	0	1	1	0	-	-	-	→ h_2
1	0	1	-	-	-	-	-	-	-	-	-	1	→ h_2
-	1	0	1	0	0	1	-	-	-	-	-	-	→ h_1
0	1	0	1	0	0	1	0	0	1	0	0	0	h_1
1	0	1	0	1	1	0	1	1	0	1	1	1	h_2

Figure 3.1: A set of SNP fragments aligned to a reference and the homologues, h_1 and h_2, from which the fragments originated. Fragments that have identical variants, specified by 0 (reference) and 1 (alternative), at their overlapping sites are assigned to the same homologue.

$$\mathbf{R} = \mathbf{R}_m \bigcup \mathbf{R}_f \bigcup \mathbf{R}_c$$
$$\boldsymbol{\epsilon} = \boldsymbol{\epsilon}_m \bigcup \boldsymbol{\epsilon}_f \bigcup \boldsymbol{\epsilon}_c$$

where $\boldsymbol{\epsilon}_m$, $\boldsymbol{\epsilon}_f$ and $\boldsymbol{\epsilon}_c$ are sets of error vectors associated with $\mathbf{R}_m$, $\mathbf{R}_f$ and $\mathbf{R}_c$, respectively. Assuming exchangeability of the offspring, it is straightforward to generalise Equation 3.1 to include n offspring as:

$$P(H_m, H_f, H_{c_1}, ..., H_{c_n} | \mathbf{R}, \boldsymbol{\epsilon}, \vec{\theta}) = \tag{3.2}$$
$$P(H_m | \mathbf{R}_m, \boldsymbol{\epsilon}_m) P(H_f | \mathbf{R}_f, \boldsymbol{\epsilon}_f) \prod_{i=1}^{n} P(H_{c_i} | \mathbf{R}_{c_i}, H_m, H_f, \boldsymbol{\epsilon}_{c_i}, \vec{\theta})$$
$$\mathbf{R} = \bigcup_{i=1}^{n} \mathbf{R}_{c_i} \bigcup \mathbf{R}_m \bigcup \mathbf{R}_f$$
$$\boldsymbol{\epsilon} = \bigcup_{i=1}^{n} \boldsymbol{\epsilon}_{c_i} \bigcup \boldsymbol{\epsilon}_m \bigcup \boldsymbol{\epsilon}_f$$

By calculating the righthand side of Equation 3.2, one can determine the likelihood of each possible phasing for a trio conditional on its sequence reads. However, as it is instead more convenient to calculate the probability of observing the reads conditional

on a phasing [4], we obtain each element of this equation using Bayes' formula according to:

$$P(H_p|R_p,\epsilon_p) = \frac{P(R_p|H_p,\epsilon_p)P(H_p)}{\sum_{H'_p} P(R_p|H'_p,\epsilon_p)P(H'_p)} \ , \ p \in \{m,f\} \tag{3.3}$$

$$P(H_{c_i}|R_{c_i},\epsilon_{c_i},H_m,H_f,\vec{\theta}) = \frac{P(R_{c_i}|H_{c_i},\epsilon_{c_i})P(H_{c_i}|H_m,H_f,\vec{\theta})}{\sum_{H'_{c_i}} P(R_{c_i}|H'_{c_i},\epsilon_{c_i})P(H'_{c_i}|H_m,H_f,\vec{\theta})}$$

where $P(H_p)$ is the prior probability of the parental phasing H_p and $P(H_{c_i}|H_m,H_f,\vec{\theta})$ is the prior probability of the phasing H_{c_i} for offspring c_i conditional on the parental phasings (H_m, H_f) and the recombination probability $\vec{\theta}$ (see Appendix A for the calculation of the read likelihoods and priors).

3.2.2. The TriPoly method

Following the Bayesian approach explained in Section 3.2.1, one has to calculate the likelihood of the reads conditional on every phasing possible. The computational cost of this brute-force approach, calculated in Appendix E, grows linearly with the sequencing depth but exponentially with the number of SNPs, l, rapidly rendering the solution intractable. To overcome this problem, we perform SNP-by-SNP reconstruction of haplotypes, starting from the leftmost SNP in the target region and keeping only a few most likely phasing extensions to the next SNP at each step (Figure 3.2, Appendix A: Equations 1-5). Following this approach, one will end up with a limited number of phasings that have passed the selection criteria during the extension procedure from $s = 1$ to $s = l$. Assuming the selection procedure effectively keeps the number of accepted solutions at each extension step bounded above by E_m and E_f for the mother and the father, respectively, the number of trio phasings at each extension will be bounded above by $\binom{k_m}{\frac{k_m}{2}}\binom{k_f}{\frac{k_f}{2}}E_mE_f$ and the total complexity will be $\mathcal{O}\left(lk_{max}\Omega_{max}\binom{k_m}{\frac{k_m}{2}}\binom{k_f}{\frac{k_f}{2}}E_mE_f\right)$, with k_{max} and Ω_{max} denoting the maximum ploidy level and the maximum sequencing coverage in the trio, respectively. This greedy method is therefore linear in terms of the number of SNPs, l. With parental ploidy levels, k_p ($p \in \{m,f\}$), in the range of 2 to 12 (covering most of the naturally occurring cases of polyploidy), $\binom{k_p}{\frac{k_p}{2}} < {k_p}^{2.75}$. Therefore, the computational complexity grows at a rate of $k_{max}^{6.5}$ with the ploidy level.

To implement this greedy method, which we call *TriPoly*, we employ *branching* and *pruning* steps similar to those in the HapTree algorithm [4] (Supplementary Figure S1). Starting at SNP site $s = 1$, the k alleles of each parent and the offspring are used as the base parental and offspring phasings, H_{bp} and H_{bc}. The base phasings are then extended step by step from SNP $s-1$ to SNP s for $s \geq 2$, until all of the SNPs have been phased as outlined in Appendix A. At each extension step, branching and pruning (Appendix: Procedure 3 and Appendix: Procedure 4) allow the algorithm to work with a limited number of possible phasings. This approach can be easily extended to include several offspring at the same time using Equation 3.2, a detailed description of which is given in Appendix A.

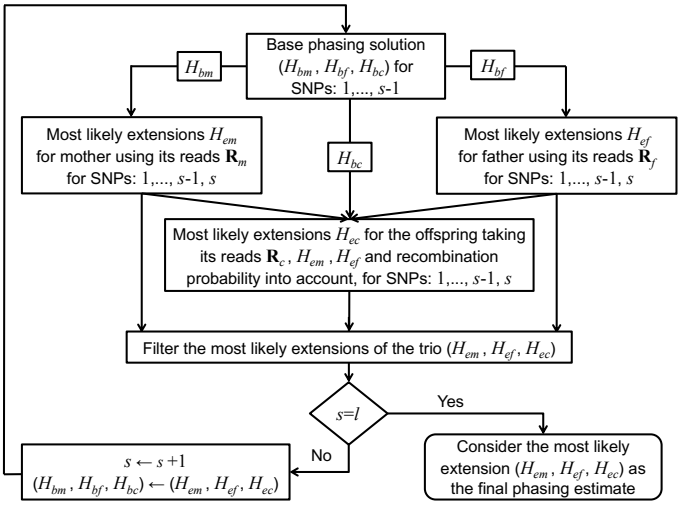


Figure 3.2: Overview of the SNP-by-SNP haplotyping method implemented in TriPoly for a trio consisting of two parents and one offspring, over a region containing l SNPs.

Note that this approach assumes working on the so-called phasing *blocks*, i.e. genomic regions in which each SNP, s, is connected to at least one other SNP, $s\prime$, through at least one of the reads in $\mathbf{R}$. In case the sequencing reads do not satisfy this condition for the whole set of SNPs in the region, it is straightforward to divide the SNP set into blocks prior to the phasing and phase each block separately, with the phasing being interrupted between the blocks.

3.3. Experimental setup

3.3.1. Simulation of polyploid trios

Before evaluating the performance of TriPoly through realistic simulations, we tested it on directly simulated SNP fragments to determine the upper bounds of its accuracy and to clearly show factors that influence its estimation quality. The advantage of this approach is bypassing of the intermediate base-calling, read alignment and variant calling steps that occur in reality and each add an undetermined amount of noise to the haplotyping process. Therefore, the direct simulation of SNP fragments lets us measure the accuracy in ideal situations and focus on the effects of sequencing depth, SNP fragment length and actual error rate in the SNP fragments. For this purpose, we simulated parental haplotypes corresponding to 1 kb regions according to the potato heterozygosity model [23], and randomly selected half of the haplotypes of each parent to simulate the offspring. The lengths of the SNP fragments, i.e. the number of SNPs contained in each fragment, was randomly chosen from uniform distributions within the ranges $[2,3]$ and $[2,l]$ (l being the total number of SNPs), resulting in average fragment lengths of 2.5 and $\frac{l}{2}+1$, respectively. Alleles at randomly selected SNP sites on a haplotype were included in each fragment and sufficient SNP fragments were generated to as-

sure the specified per homologue sequencing depths, 5-5-2, 5-5-5, 15-15-2 and 15-15-5 (maternal-paternal-offspring). Considering error rates of 0 (no error), 2% and 10%, SNP alleles were flipped by chance to introduce errors in the fragments. For each scenario, 100 trios were simulated and phased by TriPoly.

In the next step, we evaluated the performance of TriPoly, as well as three state-of-the-art single individual haplotyping algorithms: HapCompass, SDhaP and HapTree, by simulating realistic genomes and sequence data and following the read alignment and variant calling steps to obtain the SNP fragments for haplotype estimation. To this end, maternal and paternal genomes were independently simulated from a common reference using *Haplogenerator* [23], and offspring genomes were generated by passing recombinant parental chromatids at random considering a Poisson stochastic model for meiosis (see Appendix B for the details). In our simulations, we set the crossover rate (λ in Appendix B: Equation 9) to 3.07 cM/Mb, corresponding to the average recombination rate in potato [5, 13]. Using this approach, genomic regions of length 10 kb were simulated for 100 independent trios of tetraploid ($k_m = k_f = k_c = 4$) potato (*Solanum tuberosum*, $2n = 4x = 48$), based on 100 regions randomly selected from PGSC-DM genome, chromosome 5 (release version 4.03) [7] using a lognormal model to simulate genomic variation [23]. To fit the lognormal model, the SNP density of each parent was determined from empirical data [30] as described in [23], resulting in a mean distance of 21 bp between neighbouring SNPs with a standard deviation of 27 bp. The proportion of each parental marker type: simplex, duplex, triplex and quadruplex, in the total set of markers was also determined from [30] to be 0.5, 0.23, 0.14 and 0.13, respectively.

We also simulated crosses of tetraploid and diploid banana (*Musa acuminata*) yielding triploid offspring ($k_c = 3$), with the female parent being the tetraploid ($k_m = 4$) and the male parent being the diploid ($k_f = 2$), as the pollen of tetraploid banana is hardly viable [14]. In practice, commercial triploid bananas ($2n = 3x = 33$) are produced by such hybridisations, which have high consumer preference as their parthenocarpic fruits lack the large, hard seeds of fertilisation-induced fruits of diploid and tetraploid sorts. We used the sequence of chromosome 10 from the reference genome of DH-Pahang (a doubled haploid *M. acuminata*) [10], release version 2 [21], to simulate banana trios, applying the lognormal model to generate SNPs. To fit the model, we set the average SNP frequency to 1 per 200 bp with a standard deviation of 1194 bp. In the absence of population data like the one used for potato, we chose these compromise values so that we do not get many uninformative reads due to SNP sparsity (Section 3.2), while the average distance of 1394 bp reported for DH-Pahang SNPs [12] lies one standard deviation away from our used average distance and thus could still frequently occur in the simulations. As 1% recombination rate has been reported to correspond to 100 to 400 kb physical distance for banana (except at regions close to the centromere) [24, p. 130], we applied an average recombination rate of 0.04 cM/Mb for the simulation of meiosis. The proportions of parental marker types were set the same as that of potato.

For each simulated individual, sequence data were simulated according to Illumina HiSeq-2000 and PacBio CCS technologies, and the read alignment and variant calling steps were performed using conventional tools as explained in Appendix C.

3.3.2. Application to potato candidate gene sequencing data

We used TriPoly, HapCompass, SDhaP and HapTree to estimate the haplotype blocks of chromosome 5 in a mapping population of tetraploid potato consisting of 2 parents and 37 offspring (Appendix D). We used 1417 RNA capture probes for re-sequencing candidate genes by paired-end Illumina HiSeq-2000 technology with a median insert size of 316 bp per sample and a median absolute deviation (MAD) of 58 bp among the insert sizes of each sample's reads. The single reads within the paired fragments were 101 bp long. On average, the sequencing coverage for each sample was 58× (SD=15) on the captured regions. However, the coverage varied markedly with genomic position as expected with an RNA capture approach [27], with standard deviations from the mean over all of the positions ranging from 25.5 to 122 among the samples.

The sequence reads were mapped against the PGSC-DM genome (version 4.03) using bwa-mem [19], and 9762 SNPs were jointly called for all samples using FreeBayes [16]. A filtering step removed SNPs whose segregation ratios significantly violated those predicted by Mendelian rules according to Pearson's χ^2 test. In the end, 7994 SNPs were considered for phasing by each haplotyping approach.

3.3.3. Measures of phasing estimation quality

Knowing the true haplotypes in simulations, one can evaluate the performance of haplotyping methods by using measures that directly compare the estimates to the true haplotypes. We used the *reconstruction rate* (*RR*) [17] and the *pair-wise phasing accuracy rate* (*PAR*) [23] to evaluate the accuracy, and the SNP missing rate (SMR) [23] as well as the number of gaps per SNP (NGPS) to evaluate the completeness and continuity of haplotyping.

The first measure, RR, has been defined for diploids as the proportion of correctly phased markers in the phasing estimate of the target region [17]. However, to apply it for polyploids we have to generalise its mathematical formulation as haplotypes are not necessarily complementary in polyploids, making multiple correspondences possible between the original and estimated haplotypes.

Let $\hat{H} = \{\hat{h}_1, ..., \hat{h}_k\}$ be the estimated phasing and $H = \{h_1, ..., h_k\}$ be the correct phasing of a region containing l SNPs. We define RR as:

$$RR_{\hat{H},H} = 1 - \min_{p \in S_k} \frac{1}{kl} \sum_{i=1}^{k} D\left(h_i, \hat{h}_{\varphi(p,i)}\right) \tag{3.4}$$

where S_k represents the permutation group on $\{1, ..., k\}$ and φ denotes the group action on $\{1, ..., k\}$. In this definition, $D\left(h_i, \hat{h}_{\varphi(p,i)}\right)$ is the Hamming distance:

$$D\left(h_i, \hat{h}_{\varphi(p,i)}\right) = \sum_{s=1}^{l} d(h_i, \hat{h}_{\varphi(p,i)}, s) \tag{3.5}$$

$$d(h_i, \hat{h}_{\varphi(p,i)}, s) = \begin{cases} 1 & h_i^s \neq \hat{h}_{\varphi(p,i)}^s, \hat{h}_{\varphi(p,i)}^s \neq \text{"-"} \\ 0 & otherwise \end{cases}$$

where $\hat{h}_{\varphi(p,i)}^s = $ "-" means that SNP s has not been phased in $\hat{H}$.

As an alternative measure of estimation accuracy, PAR is defined as the proportion of all SNP pairs for which the inferred phasing is correct. It is important to note that PAR

takes into account phasings between any two SNPs and is not restricted to pairs of consecutive SNPs. While RR is an overall measure of accuracy based on the Hamming distance between the original haplotype and its estimate, PAR primarily shows the accuracy of long range phasing as it is highly affected by chimeric elongations of the haplotypes during estimation, i.e. the elongation of a homologue by part of another homologue. As the true haplotypes were not known for the empirical dataset (Section 3.3.2), we used RR and PAR to quantify the similarity between the haplotypes obtained by the various methods.

As haplotyping methods sometimes report phasings with high SNP exclusion, which nevertheless can have high RR and PAR, the average proportion of SNPs left out in the phasing estimates of each method (SMR) was calculated, measuring phasing completeness. Besides, in order to learn how fragmented the phasing estimates are for each method, which is not reflected in RR, PAR or SMR, the average number of interruptions, i.e. the number of blocks minus one, in the estimates of each method was calculated and normalised by the number of SNPs, l, as NGPS. Defined in this way, NGPS measures the continuity of phasing. All of the calculations to obtain these quality measures were performed using *hapcompare* [23].

In order to quantify the differences in haplotyping quality of the methods, we built regression models with the measures of haplotyping quality (RR, PAR, SMR and NGPS) as the dependent variables and the haplotyping method as the factor variable. To take the effect of sequencing depth into account, this was added as covariate to the regression models. We accounted for random variation among the simulated families by including a random intercept in the regression models.

3.3.4. Computational settings

All of the analyses were run using 2.90 GHz Intel Xeon processors. A time-limit of 1500 seconds was set for each haplotyping method during simulations, not to consume too much of the shared computational resources in case estimation became prohibitively difficult [23]. To achieve time-memory efficiency without losing much accuracy, we set the branching threshold of TriPoly, ρ, to 0.2 and its pruning threshold, κ, to 0.94, based on the results of pilot simulations. Besides, we forced TriPoly to keep no more than 11% of all possible phasing extensions at each step in case the pruning had not been able to discard as many with the value chosen for κ.

3.4. Results

To obtain the upper bounds of accuracy for TriPoly and to clearly observe the effects of SNP fragment length, sequencing depth and sequencing errors on its accuracy, we simulated SNP fragments with different lengths and error rates at various depths. We used the reconstruction rate (RR) and the pairwise phasing accuracy rate (PAR) (Section 3.3.3) to measure the accuracy of TriPoly in these simulations.

In order to assess the performance of TriPoly in practical situations and to compare it to HapCompass, HapTree and SDhaP, we next simulated realistic genomes and sequence reads for trios of tetraploid-diploid-triploid banana and tetraploid potato, and estimated the haplotypes by first aligning the reads and calling variants using conventional soft-

ware. In addition to the measures RR and PAR used with SNP fragment simulations, we used the number of gaps per SNP (NGPS) in each estimate as measure of phasing continuity and the fraction of unphased SNPs, the SNP missing rate (SMR), as measure of phasing completeness in the realistic simulations (both measures were zero in the SNP fragment simulations).

The mentioned haplotyping methods were also applied to a mapping population of tetraploid potato with 2 parents and 37 offspring, sequenced with paired-end Illumina HiSeq-2000 technology (Section 3.3.2). We compared the phasing estimates obtained by TriPoly to those of the other methods using PAR, RR, NGPS and SMR to detect agreements and conflicts. We also investigated which genomic regions are likely to be assigned to different phasings by different methods.

3.4.1. Performance on simulated data

TriPoly yields almost perfect haplotypes in ideal situations and improves the overall quality of phasing in practice

Figure 3.3 shows the performance of TriPoly with the simulated SNP fragments at various coverages and with different error rates and fragment lengths. As seen in Figure 3.3 (a), (b), the reconstruction rate (RR) and the pairwise phasing accuracy rate (PAR) are very close to 1 with an average of $\frac{l}{2}+1$ SNPs per fragment with low error rates (≤2%), hence a high phasing information content in the fragments, indicating the precision of TriPoly method in ideal situations. However, it is also evident from Figure 3.3 (c), (d) that with small fragment lengths (occurring in practice due to limited heterozygosity and restricted read lengths), the precision substantially drops, especially at high error rates, although higher sequencing coverages can compensate for this to some extent.

Through the realistic simulations of genomes and HiSeq-2000 reads, we showed 11% and 24% increases in reconstruction rate (RR) by using TriPoly compared to the other methods for the banana and potato offspring, respectively (Figure 3.4, Supplementary Tables S1-S2). The obtained average increases in accuracy were 15% and 28% for the simulations with PacBio CCS reads, with RR scores of over 95% with these long reads (Supplementary Figures S5-S6, Supplementary Tables S3-S4). These observed improvements in the overall phasing show that parental transmission is informative even for phasing between nearby SNPs, in which case the SNPs can be contained within a single read. However, TriPoly did not increase RR for the parents, especially compared to HapTree (Supplementary Figures S2-S6).

TriPoly markedly increases the accuracy of phasing between distant SNPs for the offspring

The realistic simulations with HiSeq-2000 reads showed 33% and 42% higher pair-wise phasing accuracy rates (PAR) obtained by TriPoly for banana and potato offspring, respectively, at the same SNP missing rates (SMR) compared to the other methods (Figure 3.5, Supplementary Figure S4, Supplementary Tables S1-S2). This increase was more manifest at low sequencing depths (Figure 3.5), as the parental transmission information used by TriPoly is especially advantageous when little information is provided by the reads. By penalising recombination events through the considered small recombination

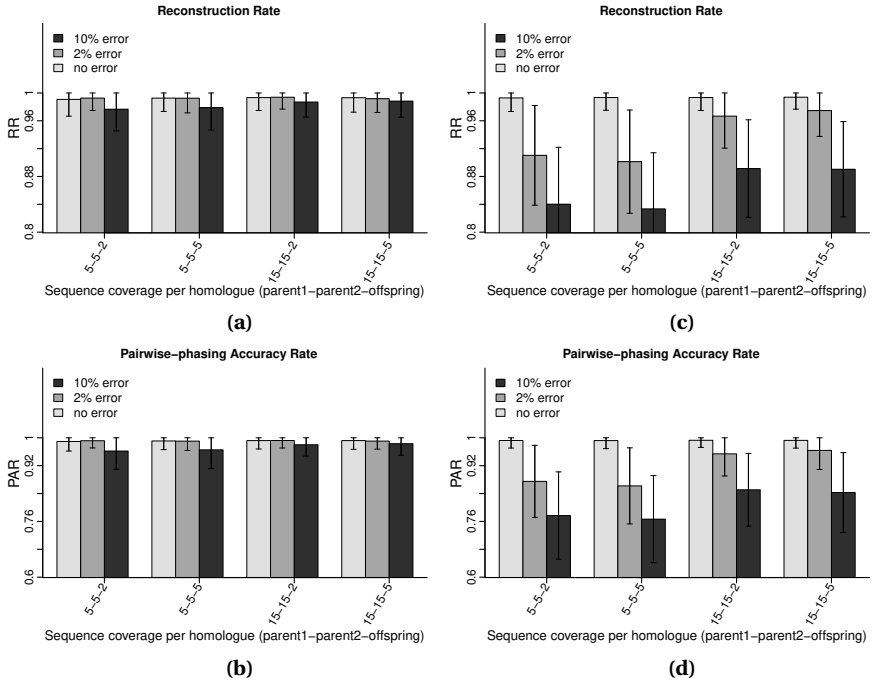


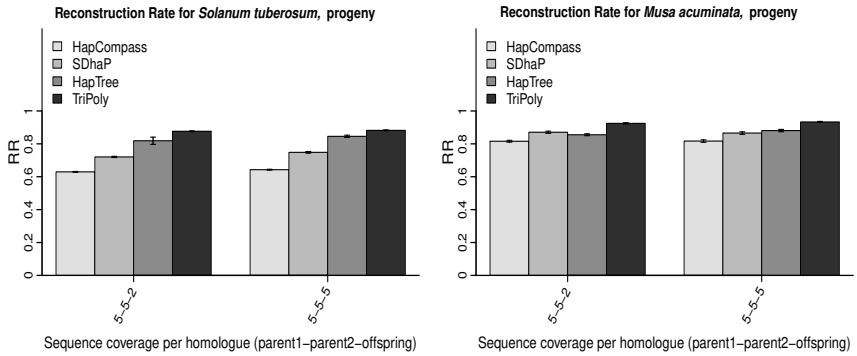
Figure 3.3: Average reconstruction rates (RR) and pair-wise phasing accuracy rate (PAR) obtained by TriPoly using the simulated SNP fragments for 100 trios with 0, 2% and 10% error rates at parent1-parent2-offspring coverages: 5-5-2, 5-5-5, 15-15-2, 15-15-5. (a) and (b) show the results with SNP fragment lengths in the range $[2, l]$ (l being the total number of SNPs), while (c) and (d) show the results for short fragment lengths in the range $[2, 3]$.

probability (Appendix A: Equation 6), TriPoly tends to reduce the chance of chimeric extensions and markedly improves the precision of phasing between distant SNPs in the offspring. However, TriPoly did not increase PAR for the parents (Supplementary Figures S2-S3).

With the simulated PacBio CCS reads, 31% and 45% increases in the average PAR were obtained by TriPoly for the banana and potato trios with the average PAR scores of TriPoly reaching 90% and 94%, respectively (Supplementary Tables S3-S4, Supplementary Figures S5-S6).

Fewer phasing interruptions are introduced in the haplotype estimates by TriPoly

As explained in Section 3.3.3, in read-based haplotyping the phasing is interrupted between two SNPs if there is no read that connects the two by covering both. With TriPoly, the blocks are determined once for the whole trio and therefore the SNPs can be connected through reads of any member in the trio. In this manner, two SNPs can be phased in an individual even if they are not connected by any reads of that individual as long as the other members in the trio provide the phasing information through their reads. This



(a) **(b)**

Figure 3.4: Average reconstruction rates (RR) for the progeny in the 100 trios simulated for a) potato and b) banana, obtained by HapCompass, SDhaP, HapTree and TriPoly at various sequencing depths using HiSeq-2000 reads.

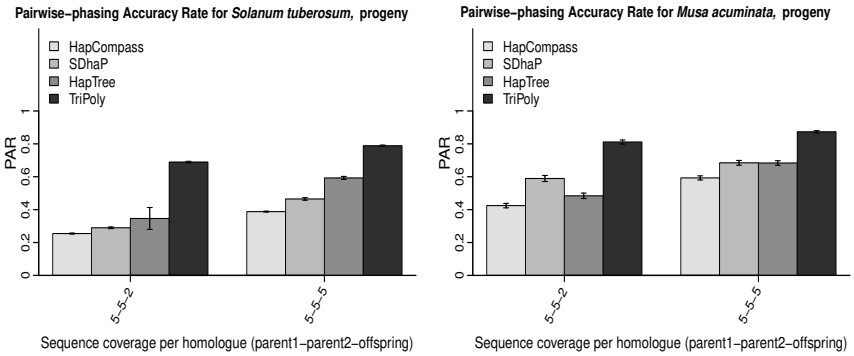
is especially beneficial for short Illumina reads, as PacBio reads contain more SNPs and hence often yield uninterrupted haplotypes even with single individual methods.

The regression analysis of NGPS for HiSeq-2000 simulations showed that the haplotypes obtained by TriPoly were significantly less interrupted compared to the other approaches, notably for banana (Supplementary Tables S1-S2, Figure 3.6). At lower SNP densities, as the average distance between subsequent SNPs will be larger a higher number of reads will be uninformative for phasing (Section 3.2) and therefore more interruptions can be introduced in the reconstructed haplotypes [23]. TriPoly proves to be especially beneficial in such situations, explaining the notable decrease in NGPS for banana compared to the slight decrease in NGPS for potato. With PacBio reads, the NGPS was (as expected) much lower on average (Supplementary Tables S3-S4) and even zero with all of the methods at 5-5-5 coverage for potato (Supplementary Figure S6). However, a substantial improvement was still observed with TriPoly especially for banana trios (Supplementary Table S3, Supplementary Figure S5).

Finally, the high standard deviation of NGPS for HapTree stands out in Figure 3.6 (a), which is a reflection of its high failure rate at low sequencing coverages for tetraploid potato [23]. As all of the SNPs belonging to a failed block are excluded from the final phasing, NGPS varies more across the simulated trios due to chance failures.

TriPoly has the smallest memory consumption with similar running times

As processing large genomic regions usually requires considerable amounts of CPU time and memory, it is important for a haplotyping algorithm to be efficient in terms of these two resources. Therefore, we measured the computation time and memory consumption of TriPoly for the simulated potato and banana trios at the applied sequencing depths and compared it to those of HapCompass, HapTree and SDhaP. As shown in the Supplementary Figures S7 and S8, TriPoly is the most memory-efficient algorithm compared to the others, while it requires more time compared to HapCompass and SDhaP for potato



(a) **(b)**

Figure 3.5: Average pairwise-phasing accuracy rates (PAR) for the progeny in the 100 trios simulated for a) potato and b) banana, obtained by HapCompass, SDhaP, HapTree and TriPoly at various sequencing depths using HiSeq-2000 reads.

with HiSeq-2000 read and more time compared to all of the algorithms for both banana and potato with PacBio CCS reads. However, the amount of time required by TriPoly was still close to that needed by the other algorithms.

3.4.2. Analysis of candidate gene sequencing data

As the true haplotypes were not known for the real dataset, we evaluated the performance of TriPoly by comparing its estimates to those obtained by HapCompass, SDhaP and HapTree. The previously introduced measures, PAR and RR, were used for this purpose by replacing the true phasing in the original definition of each with the estimates of single individual haplotyping methods.

This comparison revealed about 82% agreement between the pairwise phasings obtained by TriPoly and HapTree, and an overall similarity of 94% between the results of the two methods. Whilst almost the same agreement was observed between the estimates of HapCompass and TriPoly (PAR=80%, RR=93%), the phasings reported by SDhaP were largely different with only 46% similarity of pairwise phasing and an overall similarity score of 87% (Table 3.1).

Considering the SNP missing rate (SMR), HapTree suffered substantially higher rates (Table 3.1) as a result of its failure to generate estimates for many blocks due to instability, as reported before in a simulation study [23].

Among the applied methods, HapCompass, HapTree and SDhaP report the phasing the most compatible or the most likely with regard to the reads of each single individual as its phasing estimate. In contrast, TriPoly attempts to find the most likely haplotypes taking parental transmission probabilities into account in addition to the reads (Appendix A). The large agreement between the phasing estimates of TriPoly and HapTree (as well as HapCompass) suggests that TriPoly estimates are satisfactorily compatible with the reads, but are more accurate in presence of noise in regions with high base-calling error or low mapping/variant calling quality, since TriPoly penalises wrong

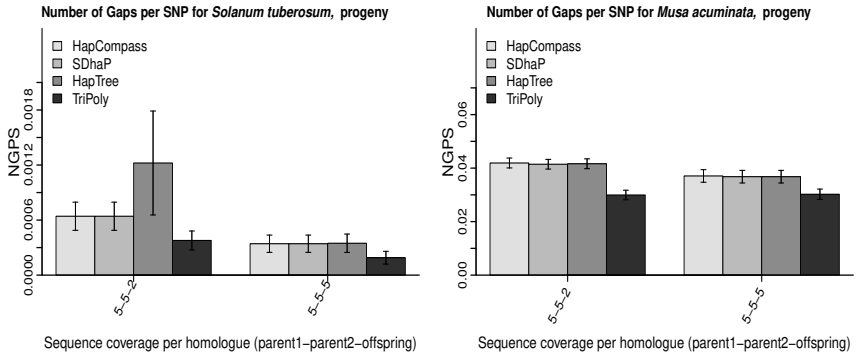


Figure 3.6: Number of Gaps per SNP (NGPS) in the phasing estimates of the progeny from the 100 trios simulated for a) potato and b) banana, using HapCompass, SDhaP, HapTree and TriPoly at various sequencing depths using HiSeq-2000 reads.

	PAR	RR	SMR	NGPS
TriPoly	—	—	0.017	0.18
HapCompass	0.80	0.93	0.015	0.18
SDhaP	0.45	0.87	0.025	0.18
HapTree	0.82	0.94	0.14	0.19

Table 3.1: Comparison of the phasings estimated by TriPoly to those estimated by HapCompass, SDhaP and HapTree for a family of tetraploid potato with 2 parents and 37 offspring.

extensions using the phasing information of the parents about the offspring. To test this hypothesis, we considered for each haplotype block the number of reads that were aligned back to its corresponding genomic region, as an indicator of sequencing depth and mapping success (and so indirectly the variant calling quality) of that region. As seen in Figure 3.7, the methods disagree mostly in regions with poor mapping, perhaps because of genomic divergence from the used reference (which occurs often in plant studies [18]), or in regions with poor capture success, where TriPoly is more reliable due to its taking parental transmission information into account.

We found a positive association between the phasing agreement scores of each block and its number of supporting reads, especially for distant SNPs as the minimum observed PAR increases exponentially with the number of aligned reads, at a rate corresponding at low coverages to 0.01 per 4 reads, i.e. one more read per chromosome for potato (Figure 3.7). Although this result is only descriptive, it points out the impact of sequencing depth and successful alignment on haplotype estimation.

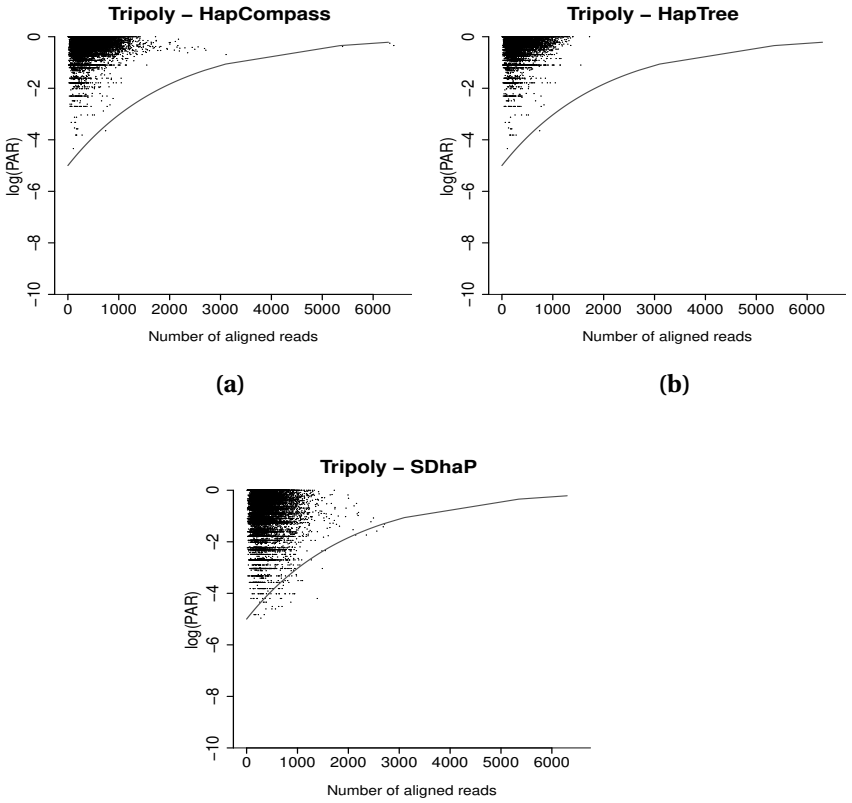


Figure 3.7: Agreement in pairwise SNP phasing, measured by the logarithm of PAR, between Tripoly and (a) HapCompass, (b) HapTree and (c) SDhaP at potato haplotype blocks against the number of aligned reads for each block. The grey curve shows the exponential fit to the minimum log PAR score at each alignment count, revealing an exponential increase in phasing agreement with an increase in the number of reads.

3.5. Conclusion and discussion

We propose a novel approach, called TriPoly, for estimating haplotypes in polyploid parent-offspring trios using sequencing data, while taking haplotype transmission from the parents to the progeny into account. TriPoly reconstructs the phasing of the SNPs over a genomic region simultaneously for the parents and for the offspring, starting from the SNP site with the smallest coordinate in the region, adding one SNP to the phasing at each step and greedily selecting the most likely extensions for the next extension step conditional on the sequence reads and recombination events. Through idealistic simulations, we show that TriPoly yields almost perfect haplotypes if the reads are long enough and accurate. Through realistic simulations of both short and long-read sequence data, we show that TriPoly significantly improves the haplotyping accuracy for the offspring compared to single individual approaches: HapCompass, SDhaP and

HapTree. Besides, we show that TriPoly estimates are more continuous compared to the other methods when the SNP density is low. TriPoly is also an efficient algorithm in terms of the memory consumption and CPU time.

We used TriPoly and the other methods to estimate haplotypes in a mapping population of tetraploid potato with 2 parents and 37 offspring sequenced using an RNA capture complexity reduction approach. We argue that in regions with imperfect sequencing or erroneous read mapping, TriPoly is more reliable compared to the other methods since it takes parental transmission probabilities into account to correct misleading read information.

In contrast to HapCompass, SDhaP and HapTree, TriPoly provides an option to keep homozygous or missing SNPs in the phasing estimates of individuals. In this way, haplotypes can be compared over the same set of SNPs in an F1-population and segregation patterns can be easily investigated. Moreover, haplotypes reported in this format can be coded as multi-allelic markers to be used in genetic analyses. Besides, TriPoly accepts input in the more convenient format of multi-sample BAM and VCF files, compared to the other methods that either require one-sample BAM/VCF (HapCompass) or the SNP fragment matrix (SDhaP and HapTree).

While TriPoly increases the accuracy of phasing for the offspring in a trio by incorporating parental recombination probabilities in the phasing likelihood (Equation 3.1), it assumes exchangeability of the offspring in families with more than one progeny (Equation 3.2) and therefore ignores the phasing information conveyed by one offspring about the others. By implementing more complex joint likelihood models, we can expect to achieve an enhancement in haplotyping accuracy for larger families. However, the computational burden is definitely a challenge in implementing such an approach. Another potential improvement in TriPoly is the phasing of the parents, the accuracy of which was shown to be inferior to that obtained by HapTree. An iterative approach of keeping a few surviving TriPoly solutions for the whole target region as the starting point for an Expectation Maximisation (EM) routine can be a way to tackle this problem, resulting in a refined set of most likely haplotypes in the population to which the reads of each individual can be mapped back to find its specific phasing. Like the joint likelihood approach, the computational challenge will be an important consideration here.

Software

TriPoly was written in Python 2.7.0 and can be freely downloaded (under license) from the software page of the Bioinformatics group, Wageningen University & Research: `http://www.bif.wur.nl`

Supplementary Figures and Data

The supplementary figures referenced in this chapter are available online at:
`https://doi.org/10.1093/bioinformatics/bty442`
The tetraploid potato DNA sequencing data described in this chapter is available at Sequence Read Archive (SRA) under the unique id 414303:
`https://www.ncbi.nlm.nih.gov/sra/?term=414303`

References

[1] Abecasis, G. R., Cherny, S. S., Cookson, W. O., and Cardon, L. R. (2002). Merlin—rapid analysis of dense genetic maps using sparse gene flow trees. *Nature Genetics*, **30**(1), 97.

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

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

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

[5] Bourke, P. M., Voorrips, R. E., Visser, R. G., and Maliepaard, C. (2015). The double-reduction landscape in tetraploid potato as revealed by a high-density linkage map. *Genetics*, **201**(3), 853–863.

[6] Browning, S. R. and Browning, B. L. (2007). Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *The American Journal of Human Genetics*, **81**(5), 1084–1097.

[7] Consortium, P. G. S. *et al.* (2011). Genome sequence and analysis of the tuber crop potato. *Nature*, **475**(7355), 189–195.

[8] Das, S. and Vikalo, H. (2015). SDhaP: Haplotype assembly for diploids and polyploids via semi-definite programming. *BMC Genomics*, **16**(1), 260.

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

[10] D'Hont, A., Denoeud, F., Aury, J.-M., Baurens, F.-C., Carreel, F., Garsmeur, O., Noel, B., Bocs, S., Droc, G., Rouard, M., *et al.* (2012). The banana (Musa acuminata) genome and the evolution of monocotyledonous plants. *Nature*, **488**(7410), 213–217.

[11] Doležel, J., Vrána, J., Cápal, P., Kubaláková, M., Burešová, V., and Šimková, H. (2014). Advances in plant chromosome genomics. *Biotechnology Advances*, **32**(1), 122–136.

[12] Droc, G., Lariviere, D., Guignon, V., Yahiaoui, N., This, D., Garsmeur, O., Dereeper, A., Hamelin, C., Argout, X., Dufayard, J.-F., *et al.* (2013). The banana genome hub. *Database*, **2013**, bat035.

[13] Felcher, K. J., Coombs, J. J., Massa, A. N., Hansey, C. N., Hamilton, J. P., Veilleux, R. E., Buell, C. R., and Douches, D. S. (2012). Integration of two diploid potato linkage maps with the potato genome sequence. *PLoS One*, **7**(4), e36347.

[14] Fortescue, J. and Turner, D. (2004). Pollen fertility in Musa: Viability in cultivars grown in Southern Australia. *Crop and Pasture Science*, **55**(10), 1085–1091.

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

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

[17] Geraci, F. (2010). A comparison of several algorithms for the single individual SNP haplotyping reconstruction problem. *Bioinformatics*, **26**(18), 2217–2225.

[18] Golicz, A. A., Bayer, P. E., Barker, G. C., Edger, P. P., Kim, H., Martinez, P. A., Chan, C.

K. K., Severn-Ellis, A., McCombie, W. R., Parkin, I. A., *et al.* (2016). The pangenome of an agronomically important crop plant *Brassica oleracea*. *Nature Communications*, **7**.

[19] Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv preprint arXiv:1303.3997*.

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

[21] Martin, G., Baurens, F.-C., Droc, G., Rouard, M., Cenci, A., Kilian, A., Hastie, A., Doležel, J., Aury, J.-M., Alberti, A., *et al.* (2016). Improvement of the banana "Musa acuminata" reference sequence using NGS data and semi-automated bioinformatics methods. *BMC Genomics*, **17**(1), 243.

[22] Michalatos-Beloin, S., Tishkoff, S. A., Bentley, K. L., Kidd, K. K., and Ruano, G. (1996). Molecular haplotyping of genetic markers 10 kb apart by allele-specific long-range PCR. *Nucleic Acids Research*, **24**(23), 4841–4843.

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

[24] Pillay, M., Ude, G., and Kole, C. (2012). *Genetics, Genomics, and Breeding of Bananas*. Science Publishers.

[25] Rafalski, J. A. (2002). Novel genetic mapping tools in plants: SNPs and LD-based approaches. *Plant Science*, **162**(3), 329–333.

[26] Simko, I., Haynes, K., Ewing, E., Costanzo, S., Christ, B., and Jones, R. (2004). Mapping genes for resistance to *Verticillium albo-atrum* in tetraploid and diploid potato populations using haplotype association tests and genetic linkage analysis. *Molecular Genetics and Genomics*, **271**(5), 522–531.

[27] Sims, D., Sudbery, I., Ilott, N. E., Heger, A., and Ponting, C. P. (2014). Sequencing depth and coverage: key considerations in genomic analyses. *Nature Reviews Genetics*, **15**(2), 121–132.

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

[29] Triplett, J. K., Wang, Y., Zhong, J., and Kellogg, E. A. (2012). Five nuclear loci resolve the polyploid history of switchgrass (panicum virgatum l.) and relatives. *PLoS One*, **7**(6), e38702.

[30] Uitdewilligen, J. G., Wolters, A.-M. A., D'hoop, B. B., Borm, T. J., Visser, R. G., and van Eck, H. J. (2013). A Next-Generation Sequencing Method for Genotyping-by-Sequencing of Highly Heterozygous Autotetraploid Potato. *PLoS One*, **8**(5), e62355.

[31] Zheng, C., Voorrips, R. E., Jansen, J., Hackett, C. A., Ho, J., and Bink, M. C. (2016). Probabilistic multilocus haplotype reconstruction in outcrossing tetraploids. *Genetics*, **203**(1), 119–131.

Appendix to Chapter 3

A) TriPoly algorithm

TriPoly is a greedy algorithm that aims to estimate haplotypes from DNA sequence data in mother-father-offspring trios, using a Bayesian maximum-likelihood approach. Starting with the first SNP site, $s = 1$, at the 5′ end of a target region that contains l SNPs, haplotype phasings are extended from $s-1$ to s for $2 \leq s \leq l$ by a greedy approach that only keeps the best extensions. The steps taken by the TriPoly algorithm are explained in detail in this section. The pseudocode of the algorithm is given in Algorithm 1.

Branching

To extend the phasing from $s-1$ to s ($2 \leq s \leq l$), the alleles at s are added to the base phasing of each parent, H_{bp} ($p \in \{m, f\}$), to obtain parental extensions, H_{ep}. As there exist multiple possibilities to allocate the alleles to the homologues, several extensions will be possible for each parent the probability of each is calculated conditional on the parental reads, $\mathbf{R}_p$, using Bayes' formula:

$$P(H_{ep}|H_{bp}, \mathbf{R}_p, \boldsymbol{\epsilon}_p) = \frac{P(\mathbf{R}_{sp}|H_{ep}, \boldsymbol{\epsilon}_p) P(H_{ep}|H_{bp})}{\Sigma_{H'_{ep} \in \{\text{ Extensions of } H_{bp} \text{ by } G^p_s\}} P(\mathbf{R}_{sp}|H'_{ep}, \boldsymbol{\epsilon}_p) P(H'_{ep}|H_{bp})} \quad , \; p \in \{m, f\} \tag{1}$$

where $\mathbf{R}_{sp}$ is a subset of $\mathbf{R}_p$ containing reads that cover s and at least one SNP position before s, and G^p_s denotes the genotype of parent p at position s. In other words, $\mathbf{R}_{sp}$ is the set that contains only the reads informative for extension from $s-1$ to s. Assuming $P(\mathbf{R}_{sp}|H_{ep}, \boldsymbol{\epsilon}_p)$ and $P(H_{ep}|H_{bp})$ are known (to which we shall return soon), $P(H_{ep}|H_{bp}, \mathbf{R}_p, \boldsymbol{\epsilon}_p)$ can be obtained for each parent according to Equation 1. Having these extension probabilities, only the phasings that have a probability greater than or equal to a *branching threshold*, ρ, are considered for extending the offspring haplotypes to s.

For extending the base phasing of the offspring, H_{bc}, all possible transmissions of alleles are considered from the surviving parental extensions, H_{em} and H_{ef}, limiting the transmissions to those compatible with the offspring genotype at position s, G^c_s. Without loss of generality, we may assume that in the extended phasing of the offspring, H_{ec}, the left-part of the homologues, i.e. h^1_{ec}, ..., $h^{\frac{k_m}{2}}_{ec}$ have been transmitted from the mother and the right-part, i.e. $h^{\frac{k_m}{2}}_{ec}$, ..., $h^{k_c}_{ec}$, are transmitted from the father. In case different parental transmissions result in the same extension for the offspring, the number of recombinations required by that extension is set to the minimum number inferred from those

transmissions. For each trio, we take inheritance into account by assigning a probability, $0 \leq \theta_s \leq 0.5$, to each recombination event between $s-1$ and s and calculate the probability of each H_{ec} in a Bayesian manner according to:

$$P(H_{ec}|H_{bc}, H_{em}, H_{ef}, \mathbf{R}_c, \boldsymbol{\epsilon}_c, \theta_s) = \frac{P(\mathbf{R}_{sc}|H_{ec}, \boldsymbol{\epsilon}_c) P(H_{ec}|H_{bc}, H_{em}, H_{ef}, \theta_s)}{\Sigma_{H'_{ec} \in \{H_{ec} \text{ obtained from } H_{em}, H_{ef}, G^c_s\}} P(\mathbf{R}_{sc}|H'_{ec}, \boldsymbol{\epsilon}_c) P(H'_{ec}|H_{bc}, H_{em}, H_{ef}, \theta_s)} \tag{2}$$

Having the probabilities for H_{em}, $H_{ef} and$ H_{ec}, a joint probability can be finally assigned to each trio extended to s using:

$$\begin{aligned} P(H_m, H_f, H_c|\mathbf{R}, \boldsymbol{\epsilon}, \vec{\theta}) &= P(H_m|\mathbf{R}_m, \boldsymbol{\epsilon}_m) \\ &\quad P(H_f|\mathbf{R}_f, \boldsymbol{\epsilon}_f) P(H_c|\mathbf{R}_c, H_m, H_f, \boldsymbol{\epsilon}_c, \vec{\theta}) \\ \mathbf{R} &= \mathbf{R}_m \bigcup \mathbf{R}_f \bigcup \mathbf{R}_c \\ \boldsymbol{\epsilon} &= \boldsymbol{\epsilon}_m \bigcup \boldsymbol{\epsilon}_f \bigcup \boldsymbol{\epsilon}_c \end{aligned} \tag{3}$$

Returning to the question of obtaining $P(\mathbf{R}_{st}|H_{et}, \boldsymbol{\epsilon}_t)$ for each trio member $t \in \{m, f, c\}$, i.e. the probability of the reads conditional on the extended haplotypes used in Equation 1 and Equation 2, it can be calculated based on the assumptions that: (*a*) the error probabilities are independent at the positions within each read and (*b*) the reads can originate from any of the homologues with equal a priori probabilities. Thus, one can obtain $P(\mathbf{R}_{st}|H_{et}, \boldsymbol{\epsilon}_t)$ according to:

$$P(\mathbf{R}_{st}|H_{et}, \boldsymbol{\epsilon}_t) = \prod_{r \in \{\mathbf{R}_{st}\}} \sum_{h_{et} \in H_{et}} \frac{1}{k_t} P(r|h_{et}) \qquad t \in \{m, f, c\}$$

$$P(r|h_{et}) = \prod_{\tau \in \{\text{SNP positions covered by } r\}} \frac{1}{3}\epsilon^\tau_r d(r, h_{et}, \tau) + \frac{1-\epsilon^\tau_r}{1-\frac{2}{3}\epsilon^\tau_r}\left(1 - d(r, h_{et}, \tau)\right) \tag{4}$$

$$d(r, h_{et}, \tau) = \begin{cases} 1 & r^\tau \neq h^\tau_{et} \\ 0 & r^\tau = h^\tau_{et} \end{cases}$$

where k_t is the ploidy level of member t, r^τ and h^τ_{et} denote the alleles appearing in read r and homologue h_{et} in H_{et} at position τ, respectively, and ϵ^τ_r is the error rate associated with r at position τ. In obtaining Equation 1, we assume an erroneous base call can be equally likely any of the 3 bases absent at each position. Hence is $\frac{1}{3}\epsilon^\tau_r$ the probability of calling the *actually observed* wrong allele. Accordingly, we adjust the correctness probability, $1-\epsilon^\tau_r$, with factor $\frac{1}{1-\frac{2}{3}\epsilon^\tau_r}$ as correctness is conditional on having the allele observed and the probability of observing the allele is qual to $1-\epsilon^\tau_r + \frac{1}{3}\epsilon^\tau_r$.

Having the probability of the reads conditional on phasing extensions, the last elements to be determined in Equation 1 and Equation 2 are the prior probabilities $P(H_{ep}|H_{bp})$ and $P(H_{ec}|H_{bc}, H_{em}, H_{ef}, \theta_s)$, respectively. As these priors are conditional on the base phasing at $s-1$, H_{bt} ($t \in \{m, f, c\}$), we build them considering only SNP positions $s-1$ and s.

Since parental haplotypes are extended prior to those of the offspring, we must first determine $P(H_{ep}|H_{bp})$ for $p \in \{m, f\}$. As changing the order of the homologues does not

change a phasing, several permutations of the alleles at $s-1$ and s can yield the same H_{ep}, determining the *a priori* weight of an extension at s. Therefore, we set $P(H_{ep}|H_{bp})$ proportional to the number of permutations that result in H_{ep}:

$$P(H_{ep}|H_{bp}) = \frac{\binom{k_p!}{m_1^{sp}!\ldots m_u^{sp}!}}{\Pi_{s-1}^p \Pi_s^p} \tag{5}$$

where Π_{s-1}^p and Π_s^p are the number of possible permutations of the alleles at $s-1$ and s, respectively, u is the number of distinct homologues, i.e. haplotypes, in H_{ep} regarding only positions $s-1$ and s, and m_i^{sp} for $i \in \{1,\ldots,u\}$ denotes the number of times an identical haplotype (regarding only positions $s-1$ and s) is present in H_{ep}. For example, with $k_p = 4$, $G_{s-1}^p = (1,1,0,0)$ and $G_s^p = (1,0,0,0)$, we have $\Pi_{s-1}^p = \binom{4!}{2!2!} = 6$ and $\Pi_s^p = \binom{4!}{3!1!} = 4$. The a priori probability of an extension with phasing $\binom{s-1:\ 1\ 1\ 0\ 0}{s:\ 0\ 0\ 0\ 1}$ will hence be $\frac{\binom{4!}{2!1!1!}}{24} = \frac{1}{2}$.

Similarly, we need to consider some value for $P(H_{ec}|H_{bc}, H_{em}, H_{ef}, \theta_s)$ to extend H_{bc}. While several models exist to describe the recombination process mathematically [12], we rely on a model assuming independent recombination events on each homologue [2, 6]:

$$P(H_{ec}|H_{bc}, H_{em}, H_{ef}, \theta_s) = P(H_{ec}|H_{bc})\theta_s^{N_\sigma}(1-\theta_s)^{N_o} \tag{6}$$

where N_σ is the number of recombination events at s and $N_o = k_c - N_\sigma$ is the number of linked transmissions of alleles at $s-1$ and s. In this formulation, $P(H_{ec}|H_{bc})$ is obtained in the same way as in Equation 5. To count the number of recombinations for each transmission, we keep track of the homologues passed from each parent to the offspring and check whether the homologues transmitted at s are the same as those transmitted at $s-1$. Each difference in the descent of the alleles at $s-1$ and s on an offspring homologue indicates either a true recombination event or a so-called chimeric extension caused by appending alleles from a different homologue to the current one. Using an assumed recombination rate, θ_s, this is translated into the prior probability for the offspring extension, as in Equation 6. This completes the steps to calculate the probabilities of extended trio phasings at each SNP site, s, conditional on the pedigree genotypes and sequencing reads.

The pseudocode of branching is given by Procedure 1 for each parent and Procedure 2 for the offspring. At the end, the obtained trio extensions are once more filtered by comparing their joint probability (Equation 3) against ρ (Procedure 3).

With more than one offspring, the steps to calculate $P(H_{ec}|.)$ (Procedure 2) are repeated separately for each offspring, assuming the exchangeability of the offspring, and the joint probability of the family phasing is calculated according to:

$$\begin{aligned} &P(H_m, H_f, H_{c_1}, \ldots, H_{c_n}|\mathbf{R}, \boldsymbol{\epsilon}, \vec{\theta}) = \\ &P(H_m|\mathbf{R}_m, \boldsymbol{\epsilon}_m)P(H_f|\mathbf{R}_f, \boldsymbol{\epsilon}_f)\prod_{i=1}^{n} P(H_{c_i}|\mathbf{R}_{c_i}, H_m, H_f, \boldsymbol{\epsilon}_{c_i}, \vec{\theta}) \end{aligned} \tag{7}$$

$$\mathbf{R} = \bigcup_{i=1}^{n} \mathbf{R}_{c_i} \bigcup \mathbf{R}_m \bigcup \mathbf{R}_f$$

$$\boldsymbol{\epsilon} = \bigcup_{i=1}^{n} \boldsymbol{\epsilon}_{c_i} \bigcup \boldsymbol{\epsilon}_m \bigcup \boldsymbol{\epsilon}_f$$

Dealing with missing information

As sequencing coverage, base calling precision or mapping quality of some regions might be insufficient to successfully call variants, missing genotypes can be reported for one parent or both of the parents in some regions. In case variant calling misses one or both of the parental genotypes at s, it is still possible to obtain the offspring extensions, ignoring transmission from the missing parent(s).

Besides missing parental genotypes, it can occur that no parental transmission is compatible with the called offspring genotype at s. In this case, we relax the restriction of extending H_{bc} in agreement with G_s^c, i.e. the offspring phasing is extended by only considering parental transmissions. It is noteworthy that in addition to sequencing and variant calling errors, such incompatibilities can also occur due to natural phenomena such as mutation or double reduction [13].

Pruning

After branching all of the base phasings, we proceed with pruning by setting the pruning threshold, $\kappa \leq 1$, and filtering out those family extensions whose joint probability (Equations 3, 8) is less than κ times the maximum probability among the whole set of branched extensions. In other words, an extension for a family with n offspring is accepted if:

$$P(H_{em},\ H_{ef},\ H_{ec_1},\ ...,\ H_{ec_n}) \geq \kappa \max_{(H'_{em},\ H'_{ef},\ H'_{ec_1},\ ...,\ H'_{ec_n}) \in \{\text{Accepted branches}\}} P(H'_{em},\ H'_{ef},\ H'_{ec_1},\ ...,\ H'_{ec_n}) \tag{8}$$

where $P(H'_{em},\ H'_{ef},\ H'_{ec_1},\ ...,\ H'_{ec_n})$ denotes the probability of the family extension $(H'_{em},\ H'_{ef},\ H'_{ec_1},\ ...,\ H'_{ec_n})$. The pseudocode for pruning is given by Procedure 4 for a trio.

B) Simulation of meiosis

In order to simulate meiosis, we calculated the Haldane frequency of recombination [4], μ, over the genomic region of interest from its average genetic distance, δ (cM/Mb) according to:

$$\mu = \frac{1}{2}(1 - e^{-0.02\delta}) \tag{9}$$

$$\delta = \lambda \frac{L}{10^6}$$

where δ is obtained assuming a uniform crossover rate, λ, over the region of length L bp. After determining μ, the number of recombination events for each parent, ν_p for $p \in \{m, f\}$, was randomly drawn from a Poisson distribution with mean μ. To take chiasma interference into account, we divided the target region of each parent into left-closed segments of length $\frac{L}{\nu_p}$ bp, so that one and only one chiasma is formed in each segment. In this way, ν_m and ν_f recombination spots were specified over the region, each placed in one of the segments according to a uniform spatial distribution over that segment. Finally, the genomic sequence from each recombination spot to the end of its containing segment was exchanged between two randomly selected non-sister chromatids to produce recombinants (corresponding to prophase I). To simulate the offspring, $\frac{k_m}{2}$ and $\frac{k_f}{2}$ chromatids were selected at random from the sets of maternal and paternal recombinants, respectively (corresponding to telophase II).

C) Simulation of sequence reads, read alignment and variant calling

For each simulated individual, sequence data were generated by simulating Illumina HiSeq 2500 technology using ART [5] and PacBio cicular consensus sequencing (CCS) [11]. Aiming for practical and efficient haplotyping, we simulated paired-end Illumina reads with an average end-to-end length of 600 bp (average single read length set to 125 bp) and CCS reads of length 2 kb [10]. Sequencing was simulated at average depths of 5× per homologue (moderate depth) for the parents, and at average depths of 2× per homologue (shallow sequencing) as well as 5× per homologue for the offspring. The simulated sequencing depths over each simulated genomic region followed a uniform distribution in the range between 0 and 2 times the given average depth [10].

The *in silico* reads were next aligned to their reference genome using bwa-mem [7] (with the default settings specified for Illumina and PacBio) and the alignments were pre-processed to remove duplicates by samtools [8] and Picardtools. Finally, SNPs were called by comparing the alignments to the reference using FreeBayes [3]. With equal ploidy levels in the trios, as was the case for potato, the SNP calling was performed by considering all trio members together. For banana trios, however, this was not possible as the ploidy levels differed between the trio members. The SNPs were therefore separately called for each parent and for the offspring and were merged afterwards into a multi-sample VCF file using a custom Python script. In this manually generated VCF file, the SNPs called in one member but not in another were considered homozygous reference calls for the missing member.

D) DNA extraction and sequencing

DNA was extracted from leaf samples of the parents and 37 F1 progeny derived from a “Karaka” × “1021/1” potato cross. Total genomic DNA was isolated using a nuclear lysis method with minor modifications [1]. Sequence capture services were provided by “RAPiD Genomics” (Gainesville, Florida, USA). In summary, the DNA was mechanically sheared to an average size of 300 bp. Next-Generation libraries were constructed by repairing the ends of the sheared fragments followed by the ligation of an adenine residue to the 3′-end of the blunt-end fragments. Next, barcoded adapters suited for

Illumina Sequencing platform were ligated to the libraries. Finally, ligated fragments were PCR-amplified for 9 cycles using standard cycling protocols (e.g. [9]). To prepare for the hybridisation, 16 barcoded libraries were pooled in equimolar amounts to a total of 500 ng. Target enrichment was performed using custom designed probes and protocols as suggested by Agilent (Palo Alto, California, USA). After enrichment, samples were re-amplified for additional 9 cycles. All enriched samples were sequenced using an Illumina HiSeq 2000 with paired-end 100 bp reads.

E) Computational complexity of the brute-force Bayesian maximum likelihood approach

To determine the computational complexity of finding the maximum likelihood phasing with a brute-force Bayesian approach taking all of the possible phasings into account, we begin by noting that the number of possible phasings for l SNPs in a k-ploid is bounded in the range:

$$\left(\max(1, \lfloor \frac{1}{k!} \prod_{s=1}^{l} \Pi_s \rfloor), \prod_{s=1}^{l} \Pi_s\right) \tag{10}$$

where Π_s denotes the number of possible permutations of the k homologues at position s. The $\frac{1}{k!}$ coefficient produces the lower bound, as the numbering of the homologues is arbitrary and therefore each phasing can be obtained by up to $k!$ combinations of the single SNP permutations (with $k!$ occurring when the phasing consists of k distinct haplotypes). As an example, for a tetraploid phasing that includes 3 SNPs ($1 \le s \le 3$) with genotypes: $G_1 = 1/1/0/1, G_2 = 0/0/1/0$ and $G_3 = 0/0/1/0$, Equation 10 gives lower and upper bounds equal to $\lfloor \frac{\binom{4}{3}\binom{4}{1}\binom{4}{1}}{4!} \rfloor = 2$ and $\binom{4}{3}\binom{4}{1}\binom{4}{1} = 64$, respectively, while 5 distinct phasings: $\begin{pmatrix}1&0&0\\1&0&0\\1&1&1\\0&0&0\end{pmatrix}$, $\begin{pmatrix}1&0&0\\1&1&0\\0&0&0\\1&0&1\end{pmatrix}$, $\begin{pmatrix}1&0&0\\1&0&0\\1&1&0\\0&0&1\end{pmatrix}$, $\begin{pmatrix}1&0&0\\1&0&0\\0&1&0\\1&0&1\end{pmatrix}$ and $\begin{pmatrix}1&0&0\\1&0&0\\1&0&0\\0&1&1\end{pmatrix}$ are actually possible, yielded by $12, 24, 12, 12$ and 4 combinations of the single SNP permutations, respectively.

With parental ploidy levels k_m, k_f and parental sequencing depths Ω_m, Ω_f, calculating the probability of each parental phasing conditional on its reads requires $\mathcal{O}(k_p l \Omega_p)$ computations for $p \in \{m, f\}$, as each determined allele in the reads must be compared to the corresponding allele on each of the k_p homologues and each SNP has been on average called in Ω_p reads. Assuming no recombination, at most $\binom{k_m}{\frac{k_m}{2}}\binom{k_f}{\frac{k_f}{2}}$ distinct haplotypes can be passed from the parents to the offspring through balanced meioses, yielding an offspring ploidy level $k_c = \frac{k_m + k_f}{2}$. Like the parental phasings, each offspring phasing requires $\mathcal{O}(k_c l \Omega_c)$ computations to calculate its probability conditional on the offspring reads at an average depth of Ω_c. Therefore, from Equation 10 it follows that a total computational cost of $\mathcal{O}\left(k_{max} l \Omega_{max} \binom{k_m}{\frac{k_m}{2}}\binom{k_f}{\frac{k_f}{2}} \prod_{s=1}^{l} \Pi_s^m \prod_{s=1}^{l} \Pi_s^p\right)$ is required to calculate Equation 10 assuming no recombination, with k_{max} the maximum parental ploidy level, i.e. $\max(k_m, k_f)$, and Ω_{max} the maximum sequencing depth for the trio members.

Allowing for recombination, different homologues may be passed to the offspring at each SNP position. To take all possible transmissions into account, we have to enumerate them separately at each SNP position. Thus, the order of computations increases to

$$\mathcal{O}\Big(k_{max} l \Omega_{max} \big(\tfrac{k_m}{\frac{k_m}{2}}\big)^l \big(\tfrac{k_f}{\frac{k_f}{2}}\big)^l \prod_{s=1}^{l} \Pi_s^m \prod_{s=1}^{l} \Pi_s^p\Big).$$

Procedure 1 BRANCHPARENT

procedure BRANCHPARENT(s, G_s^p, H_{bp}, $\mathbf{R}_{sp}$, $\boldsymbol{\epsilon}_p$, ρ)

$H_{bp} \leftarrow$ estimated phasing for parent $p \in \{m, f\}$ up to position $s-1$

$\mathbf{R}_{sp} \leftarrow$ semi-reads of parent p for position s

$k_p \leftarrow$ ploidy level of parent p

$\Pi_s^p \leftarrow$ all possible phasings of parent p *at* s

$\mathbf{H}_{ep} \leftarrow \{\}$

for $\pi \in \Pi_s^p$

$H_{ep} \leftarrow H_{bp} + \pi$

$u \leftarrow$ number of unique phasings between $s-1, s$ in H_{ep}

$m_i^{sp} \leftarrow$ number of times phasing $i \in \{1, ..., u\}$ between $s-1, s$ is repeated in H_{ep}

$$P(H_{ep}|H_{bp}) \leftarrow \frac{\binom{k_p!}{m_1^{sp}! \dots m_u^{sp}!}}{|\Pi_{s-1}^p||\Pi_s^p|}$$

$$P(H_{ep}|H_{bp}, G_s^p, \mathbf{R}_{sp}, \boldsymbol{\epsilon}_p) \leftarrow \frac{P(\mathbf{R}_{sp}|H_{ep}, \boldsymbol{\epsilon}_p)P(H_{ep}|H_{bp})}{\Sigma_{H'_{ep} \in \{H_{bp}+\pi', \pi' \in \Pi_s^p\}} P(\mathbf{R}_{sp}|H'_{ep}, \boldsymbol{\epsilon}_p)P(H'_{ep}|H_{bp})}$$

if $P(H_{ep}|H_{bp}, G_s^p, \mathbf{R}_{sp}, \boldsymbol{\epsilon}_p) \geq \rho$

$\mathbf{H}_{ep} \leftarrow \mathbf{H}_{ep} \cup \{H_{ep}\}$

end if

end for

return $\mathbf{H}_{ep}$

Procedure 2 BRANCHOFFSPRING

procedure BRANCHOFFSPRING(s, G_s^c, H_{bc}, H_{em}, H_{ef}, $\mathbf{R}_{sc}$, $\boldsymbol{\epsilon}_c$, ρ, $\vec{\theta}$)
  $h_m \leftarrow (1,...,k_m)$
  $h_f \leftarrow (1,...,k_f)$
  $\vec{H}_{ec} \leftarrow [\,]$
  $\mathbf{H}_{ec} \leftarrow \{\}$
  $\mathbf{N}_\sigma \leftarrow [\,]$
  $\mathbf{N}_o \leftarrow [\,]$
  $\theta \leftarrow \vec{\theta}[s]$
  $\Pi_s^c \leftarrow$ all possible phasings of the offspring at s
  for each $(h_{1t},...,h_{\frac{k_m}{2}t}) \in C^{h_m}_{\frac{k_m}{2}}$ **do**
    for each $(h_{(\frac{k_m}{2}+1)t},...,h_{k_c t}) \in C^{h_f}_{k_c-\frac{k_m}{2}}$ **do**
      if $(h_{1t},...,h_{k_c t}) \neq G_s^c$
        continue
      end if
      $H_{ec}^{'} \leftarrow H_{bc} + (h_{1t},...,h_{k_c t})$
      $N_o \leftarrow |\{h_i \in H_{ec} \mid h_i^{s-1} = h_{it}^{s-1}\}|$
      if $H_{ec}^{'} \notin \vec{H}_{ec}$
        $\vec{H}_{ec} \leftarrow \vec{H}_{ec} + [H_{ec}^{'}]$
        $\mathbf{N}_o \leftarrow \mathbf{N}_o + [N_o]$
        $\mathbf{N}_\sigma \leftarrow \mathbf{N}_\sigma + [k_c - N_o]$
      else
        $indx \leftarrow i : \vec{H}_{ec}[i] = H'_{ec}$
        $\mathbf{N}_o[indx] \leftarrow \max(\mathbf{N}_o[indx], N_o)$
        $\mathbf{N}_\sigma[indx] \leftarrow k_c - \mathbf{N}_o[indx]$
      end if
    end for
  end for
  for n **from** 1 **to** $|\vec{H}_{ec}|$
    $H_{ec} \leftarrow \vec{H}_{ec}[n]$
    $u \leftarrow$ number of unique phasings between $s-1, s$ in H_{ec}
    $m_i^{sc} \leftarrow$ number of times phasing $i \in \{1,...,u\}$ between $s-1, s$ is repeated in H_{ec}
    $P(H_{ec}|H_{bc}) \leftarrow \frac{\binom{k_c!}{m_1^{sc}!...\,m_u^{sc}!}}{|\Pi_{s-1}^c||\Pi_s^c|}$
    $P(H_{ec}|H_{bc}, G_s^c, H_{bc}, H_{em}, H_{ef}, \mathbf{R}_{sc}, \boldsymbol{\epsilon}_c, \theta) \leftarrow \frac{P(\mathbf{R}_{sc}|H_{ec},\boldsymbol{\epsilon}_c)P(H_{ec}|H_{bc})\theta^{\mathbf{N}_\sigma[n]}(1-\theta)^{\mathbf{N}_o[n]}}{\Sigma_{n'}P(\mathbf{R}_{sc}|\vec{H}_{ec}[n'],\boldsymbol{\epsilon}_c)P(\vec{H}_{ec}[n']|H_{bc})\theta^{\mathbf{N}_\sigma[n']}(1-\theta)^{\mathbf{N}_o[n']}}$
    if $P(H_{ec}|H_{bc}, G_s^c, H_{bc}, H_{em}, H_{ef}, \mathbf{R}_{sc}, \boldsymbol{\epsilon}_c, \theta) \geq \rho$
      $\mathbf{H}_{ec} \leftarrow \mathbf{H}_{ec} \cup \{H_{ec}\}$
    end if
  end for
  return $\mathbf{H}_{ec}$

Procedure 3 BRANCHTRIO

procedure BRANCHTRIO(s, G_s^m, G_s^f, G_s^c, H_{bm}, H_{bf}, H_{bc}, $\mathbf{R}_{sm}$, $\mathbf{R}_{sf}$, $\mathbf{R}_{sc}$, $\boldsymbol{\epsilon}_m$, $\boldsymbol{\epsilon}_f$, $\boldsymbol{\epsilon}_c$, ρ, θ)
  $\mathbf{H}_{eTrio} \leftarrow \{\}$
  for $H_{em} \in$ BRANCHPARENT($s, G_s^m, H_{bm}, \mathbf{R}_{sm}, \boldsymbol{\epsilon}_m, \rho$)
    for $H_{ef} \in$ BRANCHPARENT($s, G_s^f, H_{bf}, \mathbf{R}_{sf}, \boldsymbol{\epsilon}_f, \rho$)
      for $H_{ec} \in$ BRANCHOFFSPRING($s, G_s^c, H_{bc}, H_{em}, H_{ef}, \mathbf{R}_{sc}, \boldsymbol{\epsilon}_c, \rho, \vec{\theta}$)
        $P(H_{em}, H_{ef}, H_{ec} \mid .) \leftarrow P(H_{em} \mid .)P(H_{ef} \mid .)P(H_{ec} \mid H_{em}, H_{ef}, .)$
        if $P(H_{em}, H_{ef}, H_{ec} \mid .) \geq \rho$
          $\mathbf{H}_{eTrio} \leftarrow \mathbf{H}_{eTrio} \cup \{(H_{em}, H_{ef}, H_{ec})\}$
        end if
      end for
    end for
  end for
  return $\mathbf{H}_{eTrio}$

Procedure 4 PRUNE

procedure PRUNE($\mathbf{H}_{eTrio}$, κ)
  $\mathbf{H}_{pruned} \leftarrow \{\}$
  $p \leftarrow \max_{H'_{eTrio} \in \mathbf{H}_{eTrio}} P(H'_{eTrio} \mid .)$
  for $H'_{eTrio} \in \mathbf{H}_{eTrio}$
    if $P(H'_{eTrio} \mid .) \geq \kappa p$
      $\mathbf{H}_{pruned} \leftarrow \mathbf{H}_{pruned} \cup \{H'_{eTrio}\}$
    end if
  end for
  return $\mathbf{H}_{pruned}$

Algorithm 1 TRIPOLY

Input:

$G_s^m, G_s^f, G_s^c \quad s = 1, 2, \ldots, l$

$\mathbf{R}_m, \mathbf{R}_f, \mathbf{R}_c, \boldsymbol{\epsilon}_m, \boldsymbol{\epsilon}_f, \boldsymbol{\epsilon}_c$

$\vec{\theta} = (\theta_2, \ldots, \theta_l)$

κ, ρ

Output:

H_m, H_f, H_c

$\mathbf{H}_{Trio} \leftarrow \{\}$
for s **from** 1 **to** l
  $\mathbf{H}'_{Trio} \leftarrow \{\}$
  if $s = 1$
    $\mathbf{H}_{Trio} \leftarrow \mathbf{H}_{Trio} \cup \{(G_s^m, G_s^f, G_s^c)\}$
  else
    for $H'_{Trio} \in \mathbf{H}_{Trio}$
      $\mathbf{H}'_{Trio} \leftarrow \mathbf{H}'_{Trio} \cup \text{BRANCHTRIO}(H'_{Trio})$
    end for
    $\mathbf{H}_{Trio} \leftarrow \text{PRUNE}(\mathbf{H}'_{Trio})$
  end if
end for
$pmax \leftarrow \max_{H_{Trio} \in \mathbf{H}_{Trio}} P(H_{Trio})$
for $H_{Trio} \in \mathbf{H}_{Trio}$
  if $P(H_{Trio}) = pmax$
    $H_m, H_f, Hc \leftarrow H_{Trio}$
    break
  end if
end for
return H_m, H_f, H_c

	RR	PAR	SMR	NGPS
Intercept	0.813(0.803;0.823)	0.443(0.42;0.466)	0.624(0.613;0.636)	0.0412(0.0377;0.0448)
COV 5-5-5	0.008(0;0.015)	0.131(0.113;0.15)	-0.256(-0.263;-0.249)	-0.0035(-0.0057;-0.0013)
SDhaP	0.052(0.041;0.063)	0.128(0.102;0.155)	0.002(-0.008;0.011)	-0.0004(-0.0035;0.0027)
HapTree	0.052(0.041;0.062)	0.075(0.049;0.101)	0.001(-0.008;0.011)	-0.0003(-0.0034;0.0028)
TriPoly	0.113(0.102;0.123)	0.334(0.308;0.36)	0.012(0.003;0.022)	-0.0094(-0.0125;-0.0063)

Supplementary Table S1: 95% Confidence intervals for regression of quality measures on haplotype estimation variables for *M. acuminata* using HiSeq-2000 reads

	RR	PAR	SMR	NGPS
Intercept	0.628(0.622;0.635)	0.249(0.238;0.26)	0.399(0.388;0.411)	0.0006(0.0004;0.0008)
COV 5-5-5	0.016(0.01;0.023)	0.144(0.134;0.154)	-0.114(-0.125;-0.103)	-0.0002(-0.0004;-0.0001)
SDhaP	0.098(0.09;0.106)	0.056(0.043;0.069)	0(-0.014;0.014)	0(-0.0002;0.0002)
HapTree	0.2(0.19;0.209)	0.186(0.17;0.202)	0.08(0.062;0.097)	0(-0.0003;0.0002)
TriPoly	0.243(0.235;0.251)	0.418(0.405;0.431)	0.003(-0.011;0.017)	-0.0002(-0.0004;0)

Supplementary Table S2: 95% Confidence intervals for regression of quality measures on haplotype estimation variables for *S. tuberosum* using HiSeq-2000 reads

	RR	PAR	SMR	NGPS
Intercept	0.8(0.789;0.811)	0.496(0.468;0.525)	0.711(0.699;0.723)	0.00331(0.00248;0.00414)
COV 5-5-5	0.021(0.012;0.031)	0.144(0.12;0.168)	-0.159(-0.167;-0.151)	-0.00117(-0.00172;-0.00062)
SDhaP	0.091(0.077;0.104)	0.125(0.091;0.158)	0.002(-0.009;0.012)	-0.00025(-0.00103;0.00053)
HapTree	0.119(0.105;0.132)	0.149(0.115;0.182)	0(-0.011;0.011)	0(-0.00078;0.00078)
TriPoly	0.146(0.132;0.159)	0.308(0.274;0.342)	0.016(0.006;0.027)	-0.00098(-0.00176;-2e-04)

Supplementary Table S3: 95% Confidence intervals for regression of quality measures on haplotype estimation variables for *M. acuminata* using PacBio CCS reads

	RR	PAR	SMR	NGPS
Intercept	0.66(0.655;0.665)	0.267(0.253;0.281)	0.477(0.472;0.482)	2e-05(0;3e-05)
COV 5-5-5	0.033(0.029;0.037)	0.189(0.177;0.202)	-0.074(-0.077;-0.071)	-1e-05(-3e-05;0)
SDhaP	0.156(0.15;0.161)	-0.008(-0.025;0.01)	0(-0.004;0.004)	-1e-05(-3e-05;1e-05)
HapTree	0.246(0.24;0.252)	0.26(0.242;0.279)	0(-0.005;0.004)	0(-2e-05;2e-05)
TriPoly	0.282(0.276;0.287)	0.446(0.429;0.464)	0.004(0;0.008)	-1e-05(-3e-05;1e-05)

Supplementary Table S4: 95% Confidence intervals for regression of quality measures on haplotype estimation variables for *S. tuberosum* using PacBio CCS reads

References

[1] Bernatzky, R. and Tanksley, S. (1986). Genetics of actin-related sequences in tomato. *TAG Theoretical and Applied Genetics*, **72**(3), 314–321.

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

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

[4] Hartl, D. L., Clark, A. G., and Clark, A. G. (1997). *Principles of population genetics*, volume 116. Sinauer Associates Sunderland.

[5] Huang, W., Li, L., Myers, J. R., and Marth, G. T. (2012). ART: A next-generation sequencing read simulator. *Bioinformatics*, **28**(4), 593–594.

[6] Kojima, K., Nariai, N., Mimori, T., Takahashi, M., Yamaguchi-Kabata, Y., Sato, Y., and Nagasaki, M. (2013). A statistical variant calling approach from pedigree information and local haplotyping with phase informative reads. *Bioinformatics*, page btt503.

[7] Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv preprint arXiv:1303.3997*.

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

[9] Mamanova, L., Coffey, A. J., Scott, C. E., Kozarewa, I., Turner, E. H., Kumar, A., Howard, E., Shendure, J., and Turner, D. J. (2010). Target-enrichment strategies for next-generation sequencing. *Nature methods*, **7**(2), 111–118.

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

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

[12] Sousa, V. and Hey, J. (2013). Understanding the origin of species with genome-scale data: modelling gene flow. *Nature Reviews Genetics*, **14**(6), 404–414.

[13] Voorrips, R. E. and Maliepaard, C. A. (2012). The simulation of meiosis in diploid and tetraploid organisms using various genetic models. *BMC Bioinformatics*, **13**(1), 248.

4

Family-based haplotype estimation and allele dosage correction for polyploids using short sequence reads

This chapter has been published with minor modifications in: Ehsan Motazedi, Chris Maliepaard, Richard Finkers, Richard Visser and and Dick de Ridder, **Family-based haplotype estimation and allele dosage correction for polyploids using short sequence reads**, Frontiers in Genetics, Volume 10, Article 335, 16 April 2019, Pages 1-12

Abstract

DNA sequence reads contain information about the genomic variants located on a single chromosome. By extracting and extending this information using the overlaps between the reads, the haplotypes of an individual can be obtained. Using parent-offspring relationships in a population can considerably improve the quality of the haplotypes obtained from short reads, as pedigree information can be used to correct for spurious overlaps (due to sequencing errors) and insufficient overlaps (due to short read lengths, low genomic variation and shallow coverage).

We developed a novel method, PopPoly, to estimate polyploid haplotypes in an F1-population from short sequence data by taking into consideration the transmission of the haplotypes from the parents to the offspring. In addition, this information is employed to improve genotype dosage estimation and to call missing genotypes in the population. Through simulations, we compare PopPoly to other haplotyping methods and show its better performance. We evaluate PopPoly by applying it to a tetraploid potato cross at nine genomic regions involved in tuber formation.

4.1. Introduction

Genetic polymorphism is the key to understanding inheritance patterns of traits and to identifying genomic regions that affect a trait. While the traits of interest usually have medical importance in human genetics, in plant sciences these traits are often of importance for breeding and selection of the best varieties. Therefore, polymorphic genomic loci are used as genetic markers to investigate co-segregation of genetic variants (alleles) with qualitative traits, e.g. flower colour, in populations from crosses or in natural populations. These markers can also be used to investigate the genetic components of quantitative traits such as yield and the degree of tolerance to biotic or abiotic stresses.

The sequence of DNA marker alleles along a single chromosome is called a *haplotype*, of which a diploid organism possesses $k = 2$ versions while a polyploid has $k > 2$. To *phase* markers means to determine these k haplotypes, which might be identical (harbouring the same alleles) or different (having different alleles at some or all of the marker positions).

Among various types of genetic markers, Single Nucleotide Polymorphism (SNP) markers [7] are the most abundant and extensively used in genetic studies [2, 6]. While high-throughput assays such as SNP arrays exist for efficient determination of SNP alleles at single loci, direct determination of haplotypes usually requires laborious and expensive techniques such as bacterial cloning, allele-specific PCR or chromosome microdissection [9, 18, 27].

However, haplotypes can be used as multi-allelic markers in genetic studies offering more statistical power than single SNPs [23, 31], as both gene expression and protein function, i.e. the determinants of the phenotypes, can be affected by an allele being in *cis* or *trans* with other alleles [26]. Moreover, a marker allele which is on the same haplotype as a favourable causative allele is likely to be inherited together with that favourable allele, while the the co-transmission is unlikely if the alleles are on different haplotypes. This is important for genetic association analysis as well as for marker assisted selection.

Single individual haplotyping (SIH) methods use DNA-sequence reads to phase the SNPs of a single organism at positions covered by the reads, using the fact that the sequence of called alleles should be the same in the reads that originate from the same chromosome. To deal with sequencing errors, which can cause spurious differences between reads of the same chromosome and therefore can influence variant calling and haplotyping especially at low sequencing depths, these methods use probabilistic models or cost functions to prefer a certain phasing to others based on the observed reads [1, 3, 4, 8, 15, 30].

Recently, algorithms have been proposed that apply the rules of Mendelian inheritance to combine the information of reads and transmission in a cross in a cost function for diploids [11] or in a probabilistic model with arbitrary ploidy levels [20]. However, both of these approaches focus on trios consisting of two parents and one offspring, and therefore ignore the information provided by larger populations. In cross populations, the number of haplotypes is usually limited by the set of parental haplotypes, and therefore it is expected that we detect multiple occurrences of each haplotype across the population. This a priori information can be used to ease the estimation of haplotypes [24], but is not taken into account by the current methods. In addition, these methods accept recombinant haplotypes in the phasing estimate of the offspring (with

the recombination cost/probability being preset as desired), while recombination events have a very low probability between loci that are only a few thousands nucleotides apart, i.e. in the typical range of haplotypes obtained from short sequence reads. Sequencing and genotype calling errors can therefore be misinterpreted as recombination events by these methods and thus result in spurious haplotypes, especially in polyploids.

Here we propose a new haplotype estimation algorithm, PopPoly, that specifically targets larger F1-populations, which consist of two parents and several offspring, sequenced by short read sequencing technologies. Considering the short length of the reads, and hence the limitation of read-based phasing to a few hundreds to thousands of nucleotides, PopPoly is based on the assumption that all of the population haplotypes must be present in the parents. Therefore, all of the population reads are combined to estimate the parental haplotypes using a Bayesian probabilistic framework in the first step, and the offspring haplotypes are selected from the estimated parental haplotypes using the minimum error correction (MEC) criterion [17]. In addition, PopPoly uses the inheritance information to detect and correct wrongly estimated SNP dosages and to estimate missing genotypes in the population.

Through simulations of potato crosses with varying numbers of offspring and sequencing depths, we compare PopPoly to other haplotype estimation methods and show that it improves phasing and variant calling accuracy. Furthermore, two parents and 10 offspring of a potato cross were sequenced and subsequently analysed by PopPoly for 9 loci. For one of these loci (*StFKF1*), we selected haplotype tagging SNPs (*ht*SNPs) for the eight haplotypes proposed by PopPoly and developed a KASP assay [22] to assess the segregation in an offspring population of 181 individuals. Using genetic rules, we validated the haplotype solution proposed by PopPoly.

4.2. Material and Methods

Short-read sequencing technologies, such as Illumina, produce high-quality sequence reads of up to a few hundred bases in length, which are randomly positioned over the target genomic region and together cover each target position multiple times. By aligning these reads to some consensus reference, genomic variations can be detected and the variant alleles can be specified within each read. To resolve the succession of genomic variants on each chromosome, haplotype estimation or haplotyping methods aim to group the reads that have the same variants at the same positions as originating from the same chromosome. This approach requires overlap of the reads at the variation sites and the inclusion of at least *two* variation sites in a read, so that the flanking positions can be connected by the overlaps in between.

However, some of the reads do not meet the criterion of containing at least two variation sites, and the connection between the variation sites can be therefore broken at some positions. For this reason, current haplotyping algorithms start by detecting positions connected to each other through the sequence reads and aim to resolve the haplotypes over each obtained set of connected positions, i.e. the so-called "haplotype blocks" or solvable islands. With short sequence reads, haplotype blocks often include a few hundred up to a few thousand bases.

In our approach, we use the fact that recombination events are usually extremely unlikely over the short distances covered by the haplotype blocks obtained from short

reads. This usually confines the haplotypes observed in an F1 generation of small to moderate size to the haplotypes that exist in its parental cross. Assuming each parent transmits half of its chromosomes at random to its progeny, we combine all of the reads in an outcrossing F1-population that consists of two heterozygous parents and their F1 offspring, to estimate the haplotypes of the parents and determine the haplotypes of each offspring by selecting the phasing most compatible with its reads from the set of phasings offered by the transmission of the (already estimated) parental haplotypes.

To implement this method, we follow a greedy SNP-by-SNP extension approach (Figure 4.1), extending the base phasings H_{bm} and H_{bf} (for the mother and father, respectively) at each step by one SNP and choosing the most likely phasing extensions H_{em} and H_{ef} to continue with, as the base phasings of the next step, until all of the l SNPs within a haplotype block have been phased. Starting by the first two SNP positions in the block, the probabilities of the base and extended parental phasings, conditional on the reads and taking the observed offspring genotypes into account, are calculated using Bayes' formula. We use $s = 2$ to $s = l$ to denote the current extension SNP (as the starting base phasing is just the SNP genotype at $s = 1$), and denote the phasing extensions and called SNP genotypes by H^s_m, H^s_f, $H^s_{c_i}$ and G^s_m, G^s_f, $G^s_{c_i}$ for mother, father and offspring c_i ($i = 1, \cdots, n$) respectively. With these notations, the probability of each possible parental extension at s is related to its base phasing at $s-1$ according to:

$$P(H^s_m, H^s_f | H^{s-1}_m, H^{s-1}_f, G^s_m, G^s_f, G^s_{c_1}, \ldots, G^s_{c_n}, \mathbf{R}_{set}, \boldsymbol{\epsilon}_{set}) = \frac{P(\mathbf{R}_{set} | H^s_m, H^s_f, \boldsymbol{\epsilon}_{set}) P(H^s_m, H^s_f | G^s_m, G^s_f, G^s_{c_1}, \ldots, G^s_{c_n}, H^{s-1}_m, H^{s-1}_f)}{\sum_{(H^s_m, H^s_f)'} P\big(\mathbf{R}_{set} | (H^s_m, H^s_f)', \boldsymbol{\epsilon}_{set}\big) P\big((H^s_m, H^s_f)' \big| G^s_m, G^s_f, G^s_{c_1}, \ldots, G^s_{c_n}, H^{s-1}_m, H^{s-1}_f\big)} \tag{4.1}$$

where $\mathbf{R}_{set}$ denotes the set of all of the reads in the population and $\boldsymbol{\epsilon}_{set}$ stands for the set of base-calling error vectors, ϵ_j, associated with each $r_j \in \mathbf{R}_{set}$ ($1 \leqslant j \leqslant |\mathbf{R}_{set}|$). $P(\mathbf{R}_{set} | H^s_m, H^s_f, \boldsymbol{\epsilon}_{set})$ denotes the conditional probability of observing the reads given a pair of maternal and paternal extensions at s, (H^s_m, H^s_f), and the base-calling error probabilities given by $\boldsymbol{\epsilon}_{set}$.

The details of calculating Equation 4.1 are given in Appendix A. In order to get rid of improbable extensions and keep the number of stored phasings (almost) constant at each stage of the algorithm, at each s we discard those extensions that have a posterior probability less than $0 < \rho \leq 1$, i.e. we apply *branching* with hard thresholding. We then *prune* further the remaining extensions using a soft threshold $0 \leq \kappa \leq 1$ by discarding those with a posterior probability less than κP_{max}, where P_{max} denotes the maximum posterior probability among the branched extensions [4, 20]. The values of ρ and κ can be given by the user, and were set to 0.2 and 0.94, respectively, in our simulations.

This Bayesian framework for phasing extension can also be used to detect erroneous SNP genotypes, which result in zero probabilities for all extensions at a SNP position. We use a similar Bayesian approach to re-estimate these erroneous genotypes, as well as the uncalled SNP genotypes of the parents, by assigning probabilities to the possible population genotypes at a SNP position conditional on the reads and the segregation of parental alleles at the SNP position according to:

$$P(G_m^s, G_f^s, G_{c_1}^s, \cdots, G_{c_n}^s | \mathbf{R}_{set}, \boldsymbol{\epsilon}_{set}) = P(G_{c_1}^s, \cdots, G_{c_n}^s | G_m^s, G_f^s, \mathbf{R}_{set}, \boldsymbol{\epsilon}_{set}) P(G_m^s, G_f^s | \mathbf{R}_{set}, \boldsymbol{\epsilon}_{set}) \tag{4.2}$$

In order to calculate Equation 4.2, we first obtain the posterior probabilities of the parental genotypes, $P(G_m^s, G_f^s | \mathbf{R}_{set}, \boldsymbol{\epsilon}_{set})$, in a manner similar to that used in obtaining extension probabilities (Equation 4.1). We then assume conditional independence of the offspring genotypes given the parents, i.e. their exchangeability, to calculate:

$$P(G_{c_1}^s, \cdots, G_{c_n}^s | G_m^s, G_f^s, \mathbf{R}_{set}, \boldsymbol{\epsilon}_{set}) = P(G_{c_1} | G_m^s, G_f^s, \mathbf{R}_{c_1}, \boldsymbol{\epsilon}_{c_1}) \cdot \ldots \cdot P(G_{c_n} | G_m^s, G_f^s, \mathbf{R}_{c_n}, \boldsymbol{\epsilon}_{c_n}) \tag{4.3}$$

The details of calculating Equations 4.2 and 4.3 are given in Appendix B. The set of population genotypes with the highest likelihood is then assigned to each individual and used in Equation 4.1 for phasing extension.

After obtaining surviving phasing extensions at the last SNP position $s = l$, a phasing is chosen for each offspring from each set of parental phasing estimates by looking into the possible transmissions of the parental l SNP haplotypes. Assuming each parent transmits half of its haplotypes to each offspring, which of course requires balanced meiosis and even ploidy levels, $\binom{\frac{k_m}{2}}{2} \cdot \binom{\frac{k_f}{2}}{2}$ offspring phasings will be possible from each set of parental estimates, with k_m and k_f being the ploidy levels of the mother and the father, respectively. From this set of candidate phasings, we assign to each offspring the phasing that yields the smallest minimum error correction (MEC) score with respect to its individual sequence reads [17] (Appendix C).

Finally, each set of parental estimates and the offspring phasings deduced from them is ranked according to the relative likelihood of the parental phasings (compared to the other surviving phasings of the parents) and the sum of the MEC scores of the deduced offspring phasings. Thus, the output of the algorithm consists of sets of ranked phasing estimates for the whole population. In our simulations, we only kept the best set of population estimates for evaluation and comparison with other methods.

To examine the computational complexity of PopPoly and to see how it scales with respect to the maximum sequencing depth $d_{max} = \max(d_m, d_f, \max_{i=1}^{n} d_{c_i})$ (with d_m, d_f and d_{c_i} representing the sequencing depths of the mother, father and offspring c_i, respectively), population size $n+2$, and the number of SNPs l in the region of interest, we assume that the number of surviving extensions is effectively constant at each stage of the algorithm and denote it by η. Setting $k = \max(k_m, k_f)$, for each of the η base phasings at most $(k!)^2$ extensions must be examined at each extension step. For each of these extensions, Equation 4.1 requires $\mathcal{O}\big((n+2)d_{max}\big)$ calculations. To call the genotypes at a SNP position, Equation 4.2 requires calculations of the order $\mathcal{O}\big((k+1)^2 d_{max}^2 n\big)$, as the dosage of the alternative allele can vary from 0 to k in each parent (resulting in $\mathcal{O}\big((k+1)^2 d_{max}\big)$ complexity for the number of possible parental genotypes) and for each candidate pair of parental genotypes $\mathcal{O}\big(d_{max}\big)$ calculations are needed in each offspring to obtain the likelihood of its genotype conditional on the sequencing reads and the pair of parental genotypes (Equation 4.3). This adds up to:

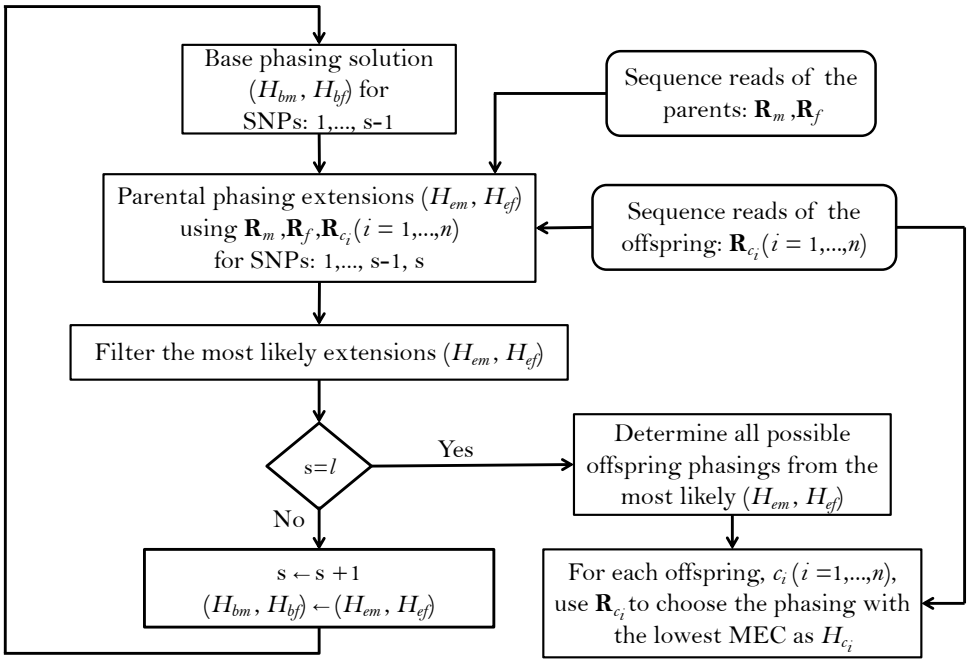


Figure 4.1: Summary of the PopPoly method to estimate haplotypes in an F1-population with two parents, (m, f), and n offspring, c_i ($i = 1,...,n$), using the sequence reads for a block including l SNPs.

$$\mathcal{O}\left(\eta(k!)^2(n+2)d_{max}^2\right) \tag{4.4}$$

complexity at each extension step. Multiplying the explained complexity by l, i.e. the number of extension steps, leads to the computational complexity of estimating parental phasings. The selection of offspring phasings using MEC scores at the end requires $\mathcal{O}\left(n\left(\frac{k}{2}\right)^2 l d_{max}\right)$ calculations for each surviving pair of parental estimates. Using $\left(\frac{k}{2}\right) < k!$ and $n+2 \leq 3n$ (as $n \geq 1$), the total complexity is:

$$\mathcal{O}\left(n\eta l(k!)^2 d_{max}^2\right) \tag{4.5}$$

which increases linearly with the number of SNPs l and the number of offspring n and quadratically with the sequencing depth d_{max}.

4.2.1. Performance evaluation by simulation

To evaluate the performance of PopPoly and compare it to other haplotyping methods, we simulated genomic regions for bi-parental F1-populations of tetraploid potato, as described in Motazedi *et al.* [20]. We simulated different scenarios, varying the number of offspring from 1 to 30. For each scenario, we randomly selected 100 regions of length

1 kb from the chromosome 5 sequence of the PGSC potato reference genome (release version 4.04) [21]. The genomes of the two parents were independently obtained for each region by introducing on average one bi-allelic SNP per 50 bp (SD=90 bp) according to the lognormal SNP density model and the dosage distributions described in Motazedi *et al.* [19], determined using data from a panel of tetraploid potato cultivars [28]. To simulate each offspring, two chromosomes were randomly selected from each parent. For the potato genome, typical ratios of genetic to physical distance vary in the range of 3 to 8 cM/Mb in different regions [5, 10]. Therefore, the assumption of improbable recombination holds for the simulated genomic regions and population sizes.

For each simulated population, paired-end Illumina HiSeq 2000 reads were generated *in silico*, with an average insert-size of 350 bp and single read length of 125 bp, using the sequencing simulator ART [13]. The simulated sequencing depth was 5× per homologue for each parent and 2× per homologue for the offspring. We also conducted simulations of families with 2, 6 and 10 offspring with higher sequencing depths, up to 30× per homologue for each individual, in order to evaluate the performance at higher coverages.

After mapping the simulated reads to their reference regions using BWA-MEM [16] and calling SNPs using FreeBayes [12], we estimated the phasing of the parents and the offspring in each F1-population using state-of-the-art SIH methods: SDhaP [8] and H-PoP [30], for comparison to PopPoly. We chose these two methods because of their computational efficiency and their allowing for SNP dosage correction, as well as the shown higher accuracy of H-PoP compared to the other state-of-the-art SIH methods [30]. We also estimated the haplotypes using the trio based method available for polyploids: TriPoly [20], and compared the obtained estimates to those obtained by PopPoly and the SIH methods.

We used several measures to compare the accuracy of haplotype estimation with the used methods. These include the *pair-wise phasing accuracy rate* (*PAR*), defined as the proportion of correctly estimated phasings for SNP-pairs [19], as well as the *reconstruction rate* (*RR*) defined to measure the overall similarity between the original haplotypes and their estimates using the Hamming distance [20].

As the quality of haplotype estimation depends not only on the accuracy of the estimated haplotypes, but also on the ability of the haplotyping method to phase as many SNPs as possible and to efficiently handle missing SNPs and wrong dosages, we calculated the *SNP missing rate* (*SMR*) and *incorrect dosage rate* (*IDR*) in the estimated haplotypes for each method.

Finally, to evaluate the continuity of phasing we measured the average number of phasing interruptions, i.e. the number of haplotype blocks minus one, in the estimates of each method and normalised it by the number of SNPs, l, as *number of gaps per SNP* (*NGPS*). The number of haplotype blocks for a set of SNPs, $\mathscr{S}$, is equal to the number of connected components in the *SNP-connectivity graph*, $\mathscr{G}_{\mathscr{S}} = (\mathscr{S}, E_{\mathscr{S}})$, in which each node represents a SNP ($|\mathscr{S}| = l$) and an edge is drawn between two SNP nodes, (s, s'), if s and s' are covered together by at least one sequence fragment.

Table 4.1: *S. tuberosum* loci selected for haplotyping

Gene	DNA sequence id	Chromosome: coordinates	Segregating bi-allelic SNPs
StCDF1	PGSC0003DMG400018408	chr05:4538880-4541736	38
StCDF2	PGSC0003DMG400025129	chr02:25588000-25591776	63
StCDF3	PGSC0003DMG400001330	chr02:46143998-46147444	75
StCDF4	PGSC0003DMG400033046	chr06:51598497-51601151	51
StCDF5	PGSC0003DMG400019528	chr03:55882564-55885296	100
StCO1	PGSC0003DMG401010056	chr02:45098374-45101578	57
StCO2	PGSC0003DMG402010056	chr02:45088023-45092647	66
StFKF1	PGSC0003DMG400019971	chr01:531784-536380	89
StGI1	PGSC0003DMG400001110	chr03:14265390-14266279	40

4.2.2. Haplotype estimation of tuberisation loci in potato

We used PopPoly to estimate haplotypes of the *S. tuberosum* loci involved in tuber formation reported by Kloosterman *et al.* [14], in an F1-population with 10 offspring obtained from the crossing of two *S. tuberosum* cultivars: Altus × Colomba ($A \times C$). The nine investigated loci (Table 4.1) belong mainly to the potato cycling DOF factor (*StCDF*) gene family, but also include other genes, such as CONSTANS (CO) genes CO1 and CO2, that are shown to be involved in *StCDF* regulation [14].

Sequence data for the parents and the offspring were obtained by whole genome sequencing (WGS) using Illumina HiSeq X Ten technology. Paired-end sequences were obtained with an average insert size of 380 bp (single read length of 151 bp) and aligned to PGSC-DM-v4.03 reference genome [21] using BWA-MEM [16]. Genomic variation within the boundaries of the selected genes was detected from the aligned reads using FreeBayes [12], with an average read depth of 85× (sd=30×) at the target loci. The sequence and variant calling data were used by PopPoly to estimate the phasing of the detected bi-allelic SNP sites (including SNPs obtained by collapsing FreeBayes complex variants).

To evaluate the accuracy of the estimated haplotypes, we selected 9 haplotype tagging SNPs (*ht*SNPs) for the parents at the *StFKF1* locus (Supplement B), and obtained their genotypes by the KASP genotyping platform [22]. The reason for choosing this specific locus was that it had 8 distinct haplotypes which could be uniquely tagged by a subset of the SNPs in the locus far enough from their neighbour variants, so that the KASP primers could be properly designed. To choose the *ht*SNPs, we considered those SNPs whose dosages in combination were compatible with one and only one of the 36 possible transmissions of the parental haplotypes in the offspring, with some redundancy to still be able to tag the haplotypes in case of low genotyping quality for some of the SNPs.

Table 4.2: Average values and 95% confidence intervals for the quality measures of each haplotyping method, obtained by simulation at the sequencing depth of 5-5-2× (mother-father-offspring) per homologue

	PopPoly	**TriPoly**	**H-PoP**	**SDhaP**
PAR	0.81(0.39;1)	0.71(0.35;1)	0.6(0.02;1)	0.44(0.04;0.93)
RR	0.95(0.8;1)	0.92(0.79;1)	0.89(0.7;1)	0.85(0.73;0.98)
SMR	0.1(0;0.33)	0.19(0;0.44)	0.33(0.04;0.64)	0.19(0;0.44)
IDR	0.09(0;0.31)	0.13(0;0.33)	0.2(0;0.69)	0.31(0;0.73)
NGPS	0.0009(0;0.001)	0.0009(0;0.001)	0.01(0;0.08)	0.01(0;0.08)

Using the KASP assay, allele specific probe signals were obtained from the parents and 181 offspring from the $A \times C$ cross (including the 10 re-sequenced offspring). To determine the genotypes, we used the *R* package fitPoly (a modified version of the package fitTetra [29]), which clusters the probe signals using a mixture of normal distributions corresponding to the marker dosages, taking the segregation of parental alleles into account. The Pearson correlation coefficient between the KASP and PopPoly dosages at these *ht*SNPs was calculated in the parents and in the 10 resequenced offspring, as a measure of the overall similarity between the true and the estimated haplotypes.

4.3. Results

4.3.1. Simulation study

To evaluate the performance of PopPoly, we simulated potato F1-populations with 1 to 30 offspring and estimated the population haplotypes using PopPoly as well as SDhaP [8], H-PoP [30] and TriPoly [20]. The estimated haplotypes were compared to the original haplotypes by *hapcompare* [19], using the measures introduced in Section 4.2.1. The overall values for the haplotyping quality measures of each method, i.e. the average of each measure over offspring sizes from 1 to 30, are given in Table 4.2 and the main conclusions are summarised below.

PopPoly yields more accurate offspring haplotypes

The average haplotype reconstruction rate (RR), which is a measure of overall phasing accuracy, obtained by PopPoly for the offspring was 0.96 (95% CI [0.87;1]) across different population sizes, which was higher than the other methods (Figure 4.2-a). The second measure of accuracy, the pairwise-phasing accuracy rate (PAR) which is especially sensitive to the accuracy of phasing between distant SNPs, had an average value of 0.84 (95% CI [0.5;1]) by PopPoly for the offspring, which was the best among the applied

methods (Figure 4.2-b). The improvement in PAR using PopPoly was, however, more manifest compared to RR.

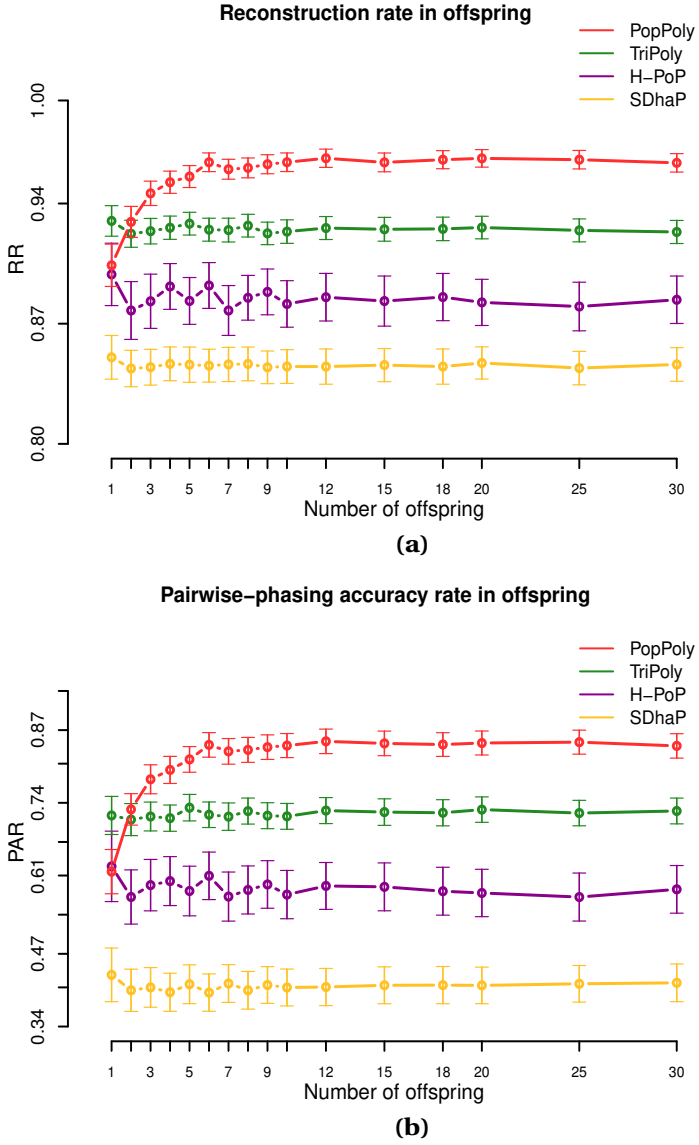


Figure 4.2: Haplotyping accuracy measures: (a) RR, (b) PAR in the offspring against the number of offspring in the population using PopPoly (red), TriPoly (green), H-PoP (purple) and SDhaP (yellow) for simulated tetraploid potato populations.

It was also noted that the accuracy of PopPoly depends on the population size, especially for distant phasing evaluated by PAR, although this dependence gradually dimin-

ishes as the number of offspring grows. As seen in Figure 4.2-b, PAR increases rapidly for PopPoly with an increase in the number of offspring from 1 to 3 and in fact, the highest offspring score for a trio, i.e. with only one offspring, is reported by TriPoly. Since an increase in the count of each parental haplotype in the population, through an increase in the number of the offspring, results in an increase in the number of reads coming from each haplotype (assuming no sequencing bias), the power of the PopPoly algorithm to detect the parental haplotype is boosted with more offspring. With a tetraploid trio, however, there is a chance that some of the parental haplotypes are not transmitted to the offspring, which causes the lower accuracy of PopPoly compared to TriPoly.

For the parents, the reported accuracy measures were very similar between the methods. However, H-PoP and PopPoly yielded the highest scores (Supplementary Figures S1-S2), with average PAR values of 0.64 (95% CI [0.2;1]) and 0.67 (95% CI [0;1]), and RR values of 0.89 (95% CI [0.73;1]) and 0.9 (95% CI [0.67;1]) for PopPoly and H-PoP, respectively.

While increasing the per homologue coverage from 5-5-2× (mother-father-offspring) to 30-30-30× yielded an average increase of 23-36% in PAR for TriPoly, H-PoP and SDhaP, the increase was only 14% for PopPoly (Supplementary Figures S3-S5), as combining the population reads already effectively augments the haplotyping coverage (the increase was actually less than 5% with 10 offspring, Supplementary Figure S5). Similarly, the difference in RR between the lowest and the highest coverage was 3% for PopPoly compared to 4-6% for the other methods (Supplementary Figures S6-S8).

Haplotype estimates of PopPoly include more SNPs than that of other methods

As seen in Table 4.2, the average SNP missing rate (SMR) of PopPoly was around 10%, which was 20% lower compared to H-PoP and around 10% lower compared to TriPoly and SDhaP (Figure 4.3). The reason for this is that combining individual NGS reads increases the chance to phase parental SNPs and choosing the offspring phasings from the estimated parental haplotypes leads to the inclusion of SNPs not sufficiently covered by the offspring reads, as well as to the imputation of SNPs uncalled in (some of) the offspring.

The 10% SMR of PopPoly can be explained by the algorithm's excluding a SNP position if the offspring genotypes at that position (either given as input or estimated anew) are incompatible with the surviving parental extensions. An example of this for a trio is the extension at $s = 2$, if the only surviving parental extensions are $H_m^2 = H_f^2 =$

$$\begin{pmatrix} & h_1 & h_2 & h_3 & h_4 \\ \text{s = 1:} & 0 & 0 & 1 & 1 \\ \text{s = 2:} & 1 & 1 & 0 & 0 \end{pmatrix}$$

while the offspring genotypes at $s = 1$ and $s = 2$ are $G_c^1 = H_c^1 = (0,0,0,1)$ and $G_c^2 = (1,1,1,1)$, respectively. In this case, G_c^2 is compatible with the parental genotypes at $s = 2$ (and therefore is accepted by the point-wise dosage estimation of PopPoly), but no H_c^2 can be obtained whose genotype at $s = 2$ is G_c^2, as haplotype

$$\begin{pmatrix} h_c \\ 1 \\ 1 \end{pmatrix}$$

cannot be transmitted to the offspring without meiotic recombination in either

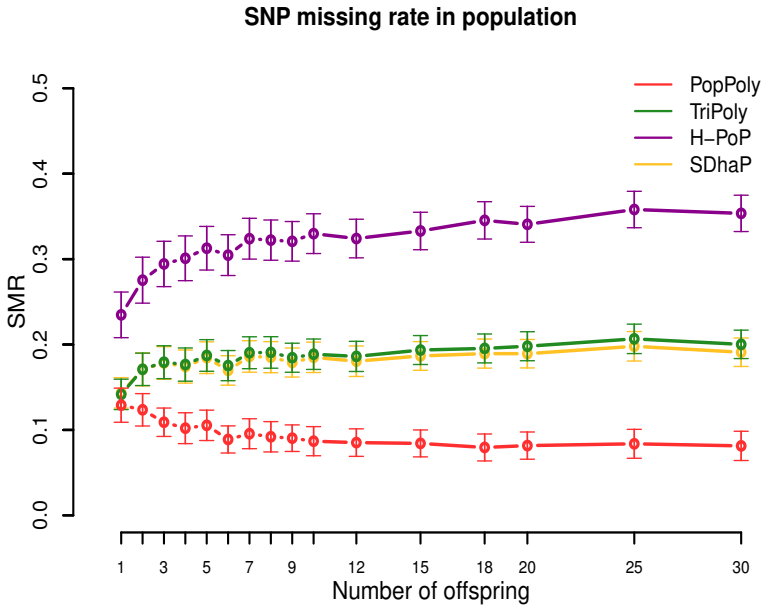


Figure 4.3: SNP missing rate (SMR) in the population against the number of offspring reported by PopPoly (red), TriPoly (green), H-PoP (purple) and SDhaP (yellow) for simulated tetraploid potato populations.

H_m^2 or H_f^2. Since PopPoly is based on the assumption of no recombination (Appendix A), it excludes the SNP site $s = 2$ from phasing.

Increasing the per homologue sequencing depth from 5-5-2× (mother-father-offspring) to 30-30-30× decreased the SMR by 16-17% for SDhaP, PopPoly and TriPoly, and by 26% for H-PoP (Supplementary Figures S9-S11).

PopPoly improves SNP dosage estimation

As shown in Table 4.2 and Figure 4.4, among the haplotyping methods PopPoly yielded the lowest incorrect dosage rate (IDR) in the phased SNPs, which was 9% on average.

The differences in the IDR between the methods is due to the differences in each algorithm's approach to handle genotype dosages. Specifically, H-PoP attempts to obtain an optimal partitioning of the reads into k groups corresponding to the homologues of a k-ploid, so that the difference between the reads assigned to the same homologue is minimised and the difference between the reads assigned to different homologues is maximised. The haplotypes are determined by taking a consensus of the reads within each group, and the dosages are determined by the estimated haplotypes. SDhaP on the other hand employs a gradient descent scheme with Lagrangian relaxation to find the best phasing (in the space of all possible phasings) according to the MEC criterion. Thus, its MEC solution determines the dosages of the SNP alleles.

In contrast to H-PoP and SDhaP, TriPoly and PopPoly use the input dosages as basis and make corrections to these based on parent-offspring relationships in the popula-

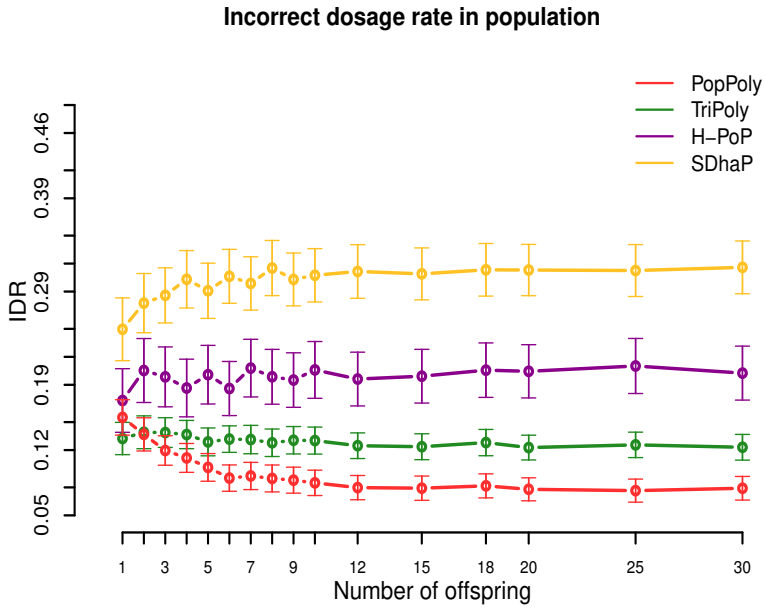


Figure 4.4: Incorrect dosage rate (IDR) in the population against the number of offspring reported by PopPoly (red), TriPoly (green), H-PoP (purple) and SDhaP (yellow) for simulated tetraploid potato populations.

tion. Specifically, if the genotype of an offspring in a trio is not compatible with the genotypes of the parents at position s, TriPoly obtains the offspring extension and hence the offspring genotype at s by considering all of the possible allele transmissions from the parents at s and by choosing the most likely trio extensions. The dosage correction method of PopPoly is explained in Appendix B.

The simulation results show that the dosage correction scheme of PopPoly is the most successful approach if there are at least two offspring in the population (Figure 4.4). For a trio, however, the most accurate dosages are reported by TriPoly. As discussed for the phasing accuracy, the ability of PopPoly to detect wrongly estimated dosages and to correctly (re)estimate dosages depends on the haplotype counts in the population. Due to the absence of some parental haplotypes in the offspring of a trio, the accuracy of PopPoly drops below that of TriPoly, which relies less on the parental haplotypes and more on the reads of the offspring to assign its dosages. With at least 6 offspring, the IDR of PopPoly drops below 10% (~7%).

Considering the sequencing coverage, SDhaP profited the most from the higher depths with a 24% lower IDR at 30-30-30× compared to 5-5-2× (per homologue), while this decrease in IDR was 12% for TriPoly and H-PoP and only 7% for PopPoly (Supplementary Figures S12-S14).

Continuity of haplotyping is improved by PopPoly compared to single individual methods

As shown in Table 4.2 and Figure 4.5, the expected number of phasing gaps (normalised by the number of SNPs) is much lower in the estimates of TriPoly and PopPoly compared to H-PoP and SDhaP, as a pair of SNPs has a higher chance of being connected when all of the population reads are used for the phasing of each individual compared to the case where for each individual only its own reads are considered. Sequencing coverage was not a determining factor for this (Supplementary Figures S15-S17).

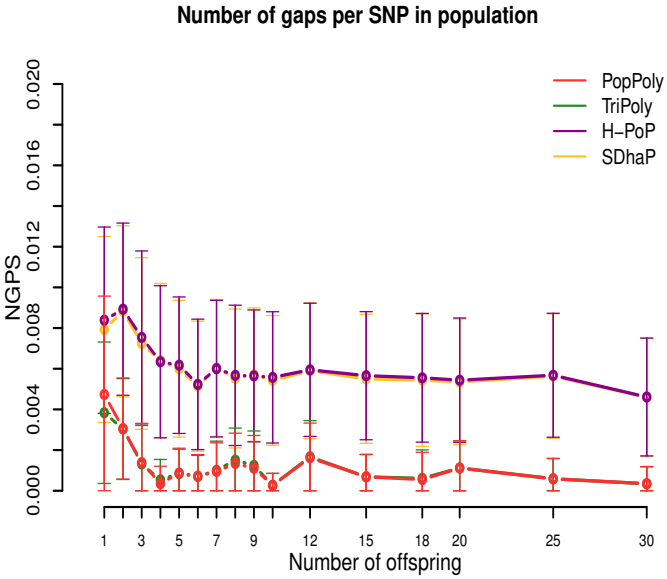


Figure 4.5: Number of phasing gaps normalised per SNP (NGPS) in the haplotype estimates of PopPoly (red), TriPoly (green), H-PoP (purple) and SDhaP (yellow) against the number of offspring in the population for simulated tetraploid potato populations.

4.3.2. Haplotypes of tuberisation loci in the tetraploid potato population

Using PopPoly, we phased all of the 579 segregating SNPs at 9 loci in the potato genome for a 10 offspring $A \times C$ cross (Supplement A). For each locus, we used the estimated haplotypes to calculate nucleotide diversity [25], i.e. the expected chance of a nucleotide difference per site between two randomly chosen haplotypes in the population. While the rather low nucleotide diversity values at the loci (mean=0.37, SD=0.06) showed high local similarity between the haplotypes, the numbers of distinct haplotypes were rather high, at 5 loci equal to the maximum of 8 (Table 4.3).

As evident from the median counts of the transmission of parental haplotypes to the offspring in Table 4.3, around half of the 58 distinct parental haplotypes (over all of the loci) were transmitted at least 5 times to the offspring. This is the expected transmission count of a haplotype in a tetraploid cross with 10 offspring if all of the parental haplotypes are distinct at the locus. However, larger sample sizes are needed to formally test

Table 4.3: Summary of SNP phasing at the potato loci introduced in Table 4.1

Gene	Number of distinct parental haplotypes	Transmission counts of parental haplotypes*	Nucleotide diversity
StCDF1	6	4-5-15	0.40
StCDF2	8	2-4.5-8	0.43
StCDF3	8	1-5-9	0.28
StCDF4	3	7-15-18	0.42
StCDF5	7	1-5-10	0.32
StCO1	3	8-11-21	0.40
StCO2	8	1-5-10	0.41
StFKF1	8	2-5-8	0.38
StGI1	8	1-4.5-9	0.29

* Minimum-Median-Maximum count of the distinct parental haplotypes observed in the offspring

whether the transmission patterns of the haplotypes are as expected under random segregation (Appendix A).

4.3.3. Validation of PopPoly estimated haplotypes

Based on the *ht*SNPs, a KASP assay was designed to investigate the eight distinct parental haplotypes of the *StFKF1* locus (Supplement B). We checked the segregation of these haplotypes using the *ht*SNPs in 181 offspring of the $A \times C$ cross, including the 10 sequenced offspring previously used in the estimation of the haplotypes with PopPoly. The obtained KASP signal ratios and the genotypes estimated by fitPoly are given in Supplement C. The KASP data was used to 1) calculate the correlation between the *ht*SNP dosages estimated from the whole genome sequencing data and the KASP dosages and 2) assess the transmission of the eight haplotypes in the 181 offspring individuals according to genetic rules, i.e. the expected transmission ratio of each maternal and paternal haplotype.

A correlation of 0.94 was observed in the comparison between the dosages of the *ht*SNPs observed in the sequencing data and the KASP data (varying within the range 0.85 to 1 per individual), in the 10 offspring assessed with both technologies. As the SNP dosages are estimated by fitting a probabilistic model for both the sequencing and the KASP assay, both approaches are prone to estimation error. The differences between the called dosages can also hinder choosing the transmitted parental haplotypes for the offspring in the larger KASP genotyped population. Therefore, some inconsistencies between the chosen haplotypes for each offspring and its KASP dosages are to be expected.

Within the larger KASP genotyped offspring population, we could assess the transmission counts of the eight haplotypes for *StFKF1* locus using genetic rules. 92% of the 181 offspring could be unambiguously phased, each consisting of two haplotypes from each parent. The 8% failure rate in uniquely choosing the haplotypes could be mainly attributed to the non-calling rate of around 2% observed for the *ht*SNPs in these individuals, as well as to inconsistencies between the dosage estimates of the *ht*SNPs obtained by the KASP assay and by PopPoly. As mentioned above, we therefore had to allow for some difference between the haplotypes and the KASP genotypes. Specifically, for each offspring we chose from the 36 possible parental transmissions the phasing that had the highest match in terms of the SNP dosages with its KASP genotypes (after eliminating SNP number 5, which had a very high inconsistency rate and was also redundant for tagging).

Subsequently, we assessed the consistency of the uniquely estimated phasings with the assumptions of random polysomic segregation. For this purpose, χ^2 goodness-of-fit tests were performed for the transmission of each haplotype from each parent, which showed no significant deviation at $\alpha = 0.05$. This suggests that the PopPoly prediction for each of the eight $StFKF1$ haplotypes is correct. However, the obtained results also show that accurate SNP dosage calling is challenging in polyploids.

4.4. Conclusion and Discussion

We present a novel algorithm, PopPoly, to exploit parent-offspring relationships for the estimation of haplotypes in an outcrossing F1-population that consists of two heterozygous parents and their F1 offspring, using short DNA sequence reads and SNP genotypes called in the population. In this approach, we first estimate the phasings of the parents by combining the sequence reads of the whole population. If necessary, SNP genotypes are also (re)estimated for the parents from the reads considering parent-offspring relationships. Having the parental phasings, we determine the phasing of each offspring by choosing from the possible transmissions of the parental haplotypes, such that the phasing chosen for each offspring has maximal compatibility with its individual reads. A natural advantage of obtaining offspring phasings from the parents is that the SNP genotypes uncalled in an offspring are imputed in its haplotypes, provided that these SNPs are included in the parental phasings.

The polyploid haplotyping problem is NP-hard and practical solutions thus by necessity depend on approximate optimisation methods. PopPoly takes a greedy approach based on Bayesian probability, extending haplotype estimates one position at a time starting from the leftmost position. While PopPoly is similar in this respect to TriPoly [20], its underlying model is quite different. As such, PopPoly is to our knowledge the first method that uses the information of siblings in estimating the haplotypes of each offspring.

Through simulations, we showed that PopPoly outperforms single individual haplotyping methods, which ignore family relationships. Besides, PopPoly yields better estimates compared to the trio based haplotyping method TriPoly when there are more than 2 offspring in the population. In addition, PopPoly uses Mendelian segregation to improve variant dosage estimation in the population at the detected SNP sites. We also show that the performance of PopPoly is influenced less by sequencing depth than

competing methods. While PopPoly assumes no limitation on the size of the population, computational resources become an important limitation when the number of offspring exceeds a couple hundred, which might require the division of a large population into smaller subpopulations for phasing. Also, the probability of observing recombinations in the F1 generation increases as the number of offspring grows. However, with genomic regions that are often at most 20 kb long, a typical maximum range for short read haplotyping, at least 500 offspring are needed to expect 1 recombination event in potato F1 populations, even at relatively high recombination rates of around 8 cM/Mb. This is not expected to have a substantial impact on the accuracy of the estimates of the parents and the other offspring.

To demonstrate the utility of PopPoly, we used it to phase 579 SNPs segregating at 9 tuberisation loci in an F1 population of tetraploid potato, the $A \times C$ cross, with 10 offspring. Using the KASP assay genotypes of a set of *ht*SNPs to represent the true haplotypes, we found a high correlation between the PopPoly estimates and the true haplotypes in the $A \times C$ population. We were able to uniquely determine the haplotypes at the tagged locus with a 92% success rate, using the parental haplotypes estimated by PopPoly and the KASP genotypes at the *ht*SNPs in 171 offspring of the $A \times C$ cross that had not been sequenced. We demonstrated that by sequencing the parents and a few offspring one can obtain the set of population (or family) haplotypes, from which the haplotypes of each individual can be determined using a set of genotyped *ht*SNPs. Such a strategy can be suitably adopted in QTL studies, with typical sizes of a few hundreds to a few thousands individuals, to increase the statistical power and to ease the interpretation of results.

Software

PopPoly was developed in Python 2.7.0 and is freely available (under license) on the software page of the Bioinformatics group, Wageningen University & Research: `http://www.bif.wur.nl`.

Supplementary Figures and Data

The supplementary figures and data referenced in this chapter are available online at: `https://doi.org/10.3389/fgene.2019.00335`
The DNA sequencing data of the $A \times C$ population described in this chapter is available at the software page on `http://www.bif.wur.nl`.

References

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

[2] Altshuler, D., Daly, M. J., and Lander, E. S. (2008). Genetic mapping in human disease. *Science*, **322**(5903), 881–888.

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

[4] Berger, E., Yorukoglu, D., Peng, J., and Berger, B. (2014). HapTree: A novel Bayesian

framework for single individual polyplotyping using NGS data. *PLoS Computational Biology*, **10**(3), e1003502.

[5] Bourke, P. M., Voorrips, R. E., Visser, R. G., and Maliepaard, C. (2015). The double-reduction landscape in tetraploid potato as revealed by a high-density linkage map. *Genetics*, **201**(3), 853–863.

[6] Braun, S. R., Endelman, J. B., Haynes, K. G., and Jansky, S. H. (2017). Quantitative trait loci for resistance to common scab and cold-induced sweetening in diploid potato. *The Plant Genome*, **10**(3).

[7] Brookes, A. J. (1999). The essence of SNPs. *Gene*, **234**(2), 177–186.

[8] Das, S. and Vikalo, H. (2015). SDhaP: haplotype assembly for diploids and polyploids via semi-definite programming. *BMC Genomics*, **16**(1), 260.

[9] Doležel, J., Vrána, J., Cápal, P., Kubaláková, M., Burešová, V., and Šimková, H. (2014). Advances in plant chromosome genomics. *Biotechnology Advances*, **32**(1), 122–136.

[10] Felcher, K. J., Coombs, J. J., Massa, A. N., Hansey, C. N., Hamilton, J. P., Veilleux, R. E., Buell, C. R., and Douches, D. S. (2012). Integration of two diploid potato linkage maps with the potato genome sequence. *PLoS One*, **7**(4), e36347.

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

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

[13] Huang, W., Li, L., Myers, J. R., and Marth, G. T. (2011). ART: a next-generation sequencing read simulator. *Bioinformatics*, **28**(4), 593–594.

[14] Kloosterman, B., Abelenda, J. A., Gomez, M. d. M. C., Oortwijn, M., de Boer, J. M., Kowitwanich, K., Horvath, B. M., van Eck, H. J., Smaczniak, C., Prat, S., *et al.* (2013). Naturally occurring allele diversity allows potato cultivation in northern latitudes. *Nature*, **495**(7440), 246–250.

[15] Lancia, G. (2016). Algorithmic approaches for the single individual haplotyping problem. *RAIRO-Operations Research*, **50**(2), 331–340.

[16] Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv preprint arXiv:1303.3997*.

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

[18] Michalatos-Beloin, S., Tishkoff, S. A., Bentley, K. L., Kidd, K. K., and Ruano, G. (1996). Molecular haplotyping of genetic markers 10 kb apart by allele-specific long-range PCR. *Nucleic Acids Research*, **24**(23), 4841–4843.

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

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

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

[22] Semagn, K., Babu, R., Hearne, S., and Olsen, M. (2014). Single nucleotide polymorphism genotyping using Kompetitive Allele Specific PCR (KASP): overview of the technology and its application in crop improvement. *Molecular Breeding*, **33**(1), 1–14.

[23] Simko, I., Haynes, K. G., Ewing, E. E., Costanzo, S., Christ, B. J., and Jones, R. W. (2004). Mapping genes for resistance to *Verticillium albo-atrum* in tetraploid and diploid potato populations using haplotype association tests and genetic linkage analysis. *Molecular Genetics and Genomics*, **271**(5), 522–531.

[24] Stephens, M., Smith, N. J., and Donnelly, P. (2001). A new statistical method for haplotype reconstruction from population data. *The American Journal of Human Genetics*, **68**(4), 978–989.

[25] Tajima, F. (1983). Evolutionary relationship of DNA sequences in finite populations. *Genetics*, **105**(2), 437–460.

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

[27] Triplett, J. K., Wang, Y., Zhong, J., and Kellogg, E. A. (2012). Five nuclear loci resolve the polyploid history of switchgrass (panicum virgatum l.) and relatives. *PLoS One*, **7**(6), e38702.

[28] Uitdewilligen, J. G., Wolters, A.-M. A., D'hoop, B. B., Borm, T. J., Visser, R. G., and van Eck, H. J. (2013). A Next-Generation Sequencing Method for Genotyping-by-Sequencing of Highly Heterozygous Autotetraploid Potato. *PLoS One*, **8**(5), e62355.

[29] Voorrips, R. E., Gort, G., and Vosman, B. (2011). Genotype calling in tetraploid species from bi-allelic marker data using mixture models. *BMC bioinformatics*, **12**(1), 172.

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

[31] Zhang, K., Calabrese, P., Nordborg, M., and Sun, F. (2002). Haplotype block structure and its applications to association studies: power and study designs. *The American Journal of Human Genetics*, **71**(6), 1386–1394.

Appendix to Chapter 4

A) Estimation of parental haplotypes

Inspired by the approach of Berger *et al.* [1], we start at the first SNP position in the target region ($s = 1$), and extend the maternal and paternal genotypes of this SNP, $G_m^1 = H_m^1$ and $G_f^1 = H_f^1$, respectively, to two-SNP phasings, H_m^2 and H_f^2. We consider every possible phasing between H_m^1 and H_f^1 and SNP position $s = 2$ in the region, and obtain the joint conditional probability of each extension pair, (H_m^s, H_f^s), at $s = 2$ given the sequence reads of the population and the parental genotypes, (G_m^s, G_f^s), as well as the offspring genotypes $G_{c_i}^s$ for $i = 1, \ldots, n$ (with n representing the number of offspring). Keeping only those parental extensions whose conditional probability exceeds or equals a pre-set *branching* threshold, $\rho \in (0, 1]$, we eliminate further the extensions whose probability is less than κP_{max}, where $\kappa \in [0, 1]$ is a pre-set *pruning* threshold and P_{max} is the maximum probability assigned to the candidate parental extensions. The surviving extensions at $s = 2$ are used in the next step as base phasings to obtain the extensions at $s = 3$ in a similar manner, and this procedure is iterated until the last SNP $s = l$ has been added to the parental extensions.

As it is not straightforward to directly calculate the conditional extension probabilities [7], we calculate instead the probability of the sequence reads conditional on each possible phasing and convert these probabilities to the desired extension probabilities using Bayes' formula:

$$P(H_m^s, H_f^s | H_m^{s-1}, H_f^{s-1}, G_m^s, G_f^s, G_{c_1}^s, \ldots, G_{c_n}^s, \mathbf{R}_{set}, \boldsymbol{\epsilon}_{set}) = \frac{P(\mathbf{R}_{set} | H_m^s, H_f^s, \boldsymbol{\epsilon}_{set}) P(H_m^s, H_f^s | G_m^s, G_f^s, G_{c_1}^s, \ldots, G_{c_n}^s, H_m^{s-1}, H_f^{s-1})}{\sum_{(H_m^s, H_f^s)'} P\big(\mathbf{R}_{set} | (H_m^s, H_f^s)', \boldsymbol{\epsilon}_{set}\big) P\big((H_m^s, H_f^s)' \big| G_m^s, G_f^s, G_{c_1}^s, \ldots, G_{c_n}^s, H_m^{s-1}, H_f^{s-1}\big)} \tag{6}$$

where $\mathbf{R}_{set}$ denotes the set of all of the reads in the population and $\boldsymbol{\epsilon}_{set}$ stands for the set of base-calling error vectors, ϵ_j, associated with each $r_j \in \mathbf{R}_{set}$ $(1 \leqslant j \leqslant |\mathbf{R}_{set}|)$. $P(\mathbf{R}_{set} | H_m^s, H_f^s, \boldsymbol{\epsilon}_{set})$ denotes the conditional probability of observing the reads given a pair of maternal and paternal extensions at s, (H_m^s, H_f^s), and the base-calling error probabilities given by $\boldsymbol{\epsilon}_{set}$.

To calculate $P(\mathbf{R}_{set} | H_m^s, H_f^s, \boldsymbol{\epsilon}_{set})$, we assume conditional independence of each read, $r_j \in \mathbf{R}_{set}$, from the other reads in $\mathbf{R}_{set}$ given $\boldsymbol{\epsilon}_{set}$, and use the fact that each read is either directly obtained from one of the parental samples or belongs to an offspring c_i $(i = 1, ..., n)$, in which latter case the read may have originated from either parent with equal probability. Under these assumptions, $P(\mathbf{R}_{set} | H_m^s, H_f^s, \boldsymbol{\epsilon}_{set})$ is determined according to:

$$P(\mathbf{R}_{set}|H_m^s, H_f^s, \boldsymbol{\epsilon}_{set}) = \prod_{j=1}^{|\mathbf{R}_{set}|} P(r_j|H_m^s, H_f^s, \boldsymbol{\epsilon}_{set}) =$$

$$\prod_{j=1}^{|\mathbf{R}_{set}|} \Big[P(r_j|H_m^s, \epsilon_j) U\big(\delta(r_j), m\big) + P(r_j|H_f^s, \epsilon_j) U\big(\delta(r_j), f\big) + \tag{7}$$

$$\frac{1}{2}\big(P(r_j|H_m^s, \epsilon_j) + P(r_j|H_f^s, \epsilon_j)\big) \sum_{i=1}^{n} U\big(\delta(r_j), c_i\big)\Big]$$

$$U(x,y) = \begin{cases} 1 & x = y \\ 0 & x \neq y \end{cases}$$

$$\delta : \mathbf{R}_{set} \longrightarrow \{m, f, c_1, ..., c_n\}$$

where the function $\delta(r_j)$ returns the origin of read r_j: mother (m), father (f), or one of the n offspring ($c_1, ..., c_n$).

Assuming independence of the sequencing errors at the SNP positions within each read, $P(r_j|H_m^s)$ and $P(r_j|H_f^s)$ in Equation 7 can be calculated according to Motazedi *et al.* [7]:

$$P(r_j|H_p^s, \epsilon_j) = \frac{1}{k_t} \sum_{h \in H_p^s} P(r_j|h, \epsilon_j) \qquad p \in \{m, f\}$$

$$P(r_j|h, \epsilon_j) = \prod_{\tau=1}^{s} \frac{1}{3}\epsilon_j^\tau d(r_j, h, \tau) + \frac{1-\epsilon_j^\tau}{1-\frac{2}{3}\epsilon_j^\tau}\big(1 - d(r_j, h, \tau)\big) \tag{8}$$

$$d(r_j, h, \tau) = \begin{cases} 1 & r_j^\tau \neq h^\tau, r_j^\tau \neq \text{"-"}, h^\tau \neq \text{"-"} \\ 0 & otherwise \end{cases}$$

where ϵ_j assigns a base-calling error probability to every SNP position in r_j, and h stands for each of the k_t homologues in the phasing extension H_p^s ($p \in \{m, f\}$). In Equation 8, we use the superscript τ in r_j^τ and ϵ_j^τ to represent the called base at SNP position τ and its associated error probability, respectively. Likewise, h^τ denotes the allele assigned to homologue h at SNP position τ. We use r_j^τ = "-" and h^τ = "-" to show that SNP position τ has not been called in r_j or is missing in h.

In obtaining $P(r_j|h, \epsilon_j)$ in Equation 8, we assume that an erroneously called base can with equal chance be any of the three wrong bases. Therefore, the probability of observing a specific wrong allele is $\frac{1}{3}\epsilon_j^\tau$. Also, the probability of no error is actually the probability that no error occurs ($1-\epsilon_j^\tau$), conditional on having observed either the reference or the alternative allele ($1 - \frac{2}{3}\epsilon_j^\tau$). Therefore, it is $\frac{1-\epsilon_j^\tau}{1-\frac{2}{3}\epsilon_j^\tau}$.

Equations 7 and 8 establish the procedure to calculate the likelihood in Bayes' formula in Equation 6. In order to solve Equation 6, one also needs to specify the prior, $P(H_m^s, H_f^s|G_m^s, G_f^s, G_{c_1}^s, \ldots, G_{c_n}^s, H_m^{s-1}, H_f^{s-1})$. While several ways can be thought of to specify this prior, we obtain it as follows. As the parental extensions (H_m^s, H_f^s) are confined

to those compatible with G^s_m and G^s_f, we set this prior to zero for every incompatible extension. For the compatible extensions, we look into the possible transmissions of the extended haplotypes (ignoring phenomena like aneuploidy [5], preferential chromosome pairing [3], recombination and double reduction [2]) to the offspring and for each offspring, c_i, we count the number of transmissions that agree with its genotype at s, $G^s_{c_i}$. Dividing this number by the total number of possible transmissions, $\binom{k_m}{\frac{k_m}{2}} \cdot \binom{k_f}{\frac{k_f}{2}}$, gives us $P(G^s_{c_i}|H^s_m, H^s_f)$. Calculating $P(G^s_{c_i}|H^s_m, H^s_f)$ for $i = 1, \ldots, n$, we obtain the average likelihood of an *observed* offspring genotype according to:

$$E_{H^s_m,H^s_f}[P(G^s_c|H^s_m, H^s_f)] = \sum_{i=1}^{n} \frac{P(G^s_{c_i}|H^s_m,H^s_f)}{P(G_{c_1}|H^s_m,H^s_f)+\cdots+P(G_{c_n}|H^s_m,H^s_f)} P(G^s_{c_i}|H^s_m, H^s_f)$$
$$= \frac{1}{\sum_{i=1}^{n} P(G_{c_i}|H^s_m,H^s_f)} \sum_{i=1}^{n} \left(P(G^s_{c_i}|H^s_m, H^s_f)\right)^2 \tag{9}$$

where $P(G^s_{c_i}|H^s_m, H^s_f)$ is the likelihood and $\frac{P(G^s_{c_i}|H^s_m,H^s_f)}{P(G_{c_1}|H^s_m,H^s_f)+\cdots+P(G_{c_n}|H^s_m,H^s_f)}$ is the probability of observing offspring c_i.

So far, we set the prior for each (H^s_m, H^s_f) to be proportional to $E_{H^s_m,H^s_f}[P(G^s_c|H^s_m, H^s_f)]$. However, as changing the order of the homologues does not change a phasing, several permutations of the alleles at $s-1$ and s can yield the same (H^s_m, H^s_f). Therefore, the prior should also be proportional to the number of permutations that result in (H^s_m, H^s_f). It can be thus set to:

$$P(H^s_m, H^s_f|G^s_m, G^s_f, G^s_{c_1}, \ldots, G^s_{c_n}, H^{s-1}_m, H^{s-1}_f) = E_{H^s_m,H^s_f}[P(G^s_c|H^s_m, H^s_f)] \frac{\binom{k_m!}{\omega_1^{sm}! \ldots \omega_{u_m}^{sm}!}}{\Pi^m_{s-1}\Pi^m_s} \frac{\binom{k_f!}{\omega_1^{sf}! \ldots \omega_{u_f}^{sf}!}}{\Pi^f_{s-1}\Pi^f_s} \tag{10}$$

where, for $p \in \{m, f\}$, Π^p_{s-1} and Π^p_s are the number of possible permutations of the alleles at $s-1$ and s, respectively, u_p is the number of distinct homologues, i.e. haplotypes, in H^s_p regarding only positions $s-1$ and s, and ω_i^{sp} for $i \in \{1, \ldots, u_p\}$ denotes the number of times an identical haplotype (regarding only positions $s-1$ and s) is present in H^s_p. Although it is possible to normalise the priors obtained this way over all of the possible extensions (to obtain a proper prior mass function), one does not need to do so as the discrete posteriors are normalised anyway at the end.

As an example, with tetraploid parents there will be $\binom{4}{2} \cdot \binom{4}{2} = 36$ possible haplotype transmissions to each offspring. With maternal and paternal extensions at $s = 3$ being equal to $H^3_m = \begin{pmatrix} & h_1 & h_2 & h_3 & h_4 \\ \text{SNP 1:} & 1 & 1 & 0 & 0 \\ \text{SNP 2:} & 1 & 0 & 0 & 1 \\ \text{SNP 3:} & 1 & 0 & 1 & 1 \end{pmatrix}$ and $H^3_f = \begin{pmatrix} & h_5 & h_6 & h_7 & h_8 \\ \text{SNP 1:} & 0 & 1 & 0 & 0 \\ \text{SNP 2:} & 0 & 0 & 1 & 1 \\ \text{SNP 3:} & 0 & 0 & 0 & 1 \end{pmatrix}$, respectively, and two offspring c_1 and c_2 with $G^3_{c_1} = (1\,0\,0\,0)$ and $G^3_{c_2} = (1\,0\,1\,0)$, only 9 out of 36 transmissions will be compatible with the genotype of c_1, while 18 transmissions will be compatible with

c_2. This results in $E_{H_m^s,H_f^s}[P(G_c^3|H_m^3,H_f^3)] = \frac{1}{\frac{3}{4}}\left((\frac{9}{36})^2+(\frac{18}{36})^2\right) = \frac{5}{12}$ for this extension. As $k_m = k_f = 4$, $G_m^2 = (1,0,0,1)$, $G_m^3 = (1,0,1,1)$, $G_f^2 = (0,0,1,1)$ and $G_f^3 = (0,0,0,1)$, we have $\Pi_2^m = \Pi_2^f = \binom{4!}{2!2!} = 6$ and $\Pi_3^m = \Pi_3^f = \binom{4!}{3!1!} = 4$. Considering only SNPs at $s-1=2$ and $s=3$, in each parent there is one haplotype present twice. The a priori probability of (H_m^3, H_f^3) is hence determined from Equation 10 to be $\frac{5}{12}\cdot\frac{\binom{4!}{2!1!1!}}{24}\cdot\frac{\binom{4!}{2!1!1!}}{24} = \frac{5}{48}$.

From Equations 7 and 10, the conditional probabilities of parental extensions at position s can be obtained using Equation 6 and the surviving extensions are used for the extension to $s+1$, as explained above.

B) Estimation of missing and erroneous genotypes

The SNP-by-SNP extension of the parental haplotypes using the sequencing reads of an F1-population is explained in Section A, assuming the SNPs have been accurately called for all of the population members. However, in practice every haplotyping algorithm has to handle missing and wrongly estimated SNP genotypes caused by sequencing and variant calling errors.

In presence of wrongly estimated genotypes (wrong dosages), it can occur that all of the offspring genotypes are incompatible with the parental extensions at some SNP position s. At these positions, the extension should either be skipped, as the prior weight of all candidate phasings will be zero, or the genotypes must be estimated anew. The extension at s will also be impossible if one or both of the parental genotypes are missing at s. To include these SNP positions in the extension, it is necessary to impute the missing genotypes.

In order to estimate the population genotypes at the missing or incompatible positions, we assume that the parents come from an infinite-size population at Hardy-Weinberg equilibrium. Limiting the attention to bi-allelic SNPs, the reference and alternative allele frequencies of the parents at position s can be estimated from the observed reads under the above assumption. Assuming a fixed sequencing error rate for all of the reads and nucleotide positions, $0 \leqslant \widehat{ER} < 0.5$, the frequency of the alternative allele can be obtained assuming a binomial model for the observed count of the alternative allele according to:

$$\begin{aligned} \xi &= |\{r_j \in \mathbf{R}_{set} | r_j^s = 1 \vee r_j^s = 0\}| \\ \psi &= \frac{|\{r_j \in \mathbf{R}_{set} | r_j^s = 1\}|}{\xi} \\ \hat{p} &= \frac{\psi - \widehat{ER}}{1 - 2\widehat{ER}} \end{aligned} \tag{11}$$

where ξ is the total sequencing coverage of the population at s and ψ is the proportion of the alternative allele among the observed alleles. As this observed frequency, ψ, depends on the latent true frequency, $\hat{p}$, through $\psi = (1-\widehat{ER})\hat{p} + \widehat{ER}(1-\hat{p})$, it is straightforward to show that $\hat{p}$ can be obtained as shown in Equation 11, with a standard error equal to $\frac{1}{(1-2\widehat{ER})}\cdot\sqrt{\frac{\psi(1-\psi)}{\xi}}$.

In case a specific base-calling error rate ϵ_j^s is assigned at each position s to each read r_j, e.g. by using the integer-rounded Phred (quality) scores reported by the sequencer [4], one can assume a Gaussian distribution for the probability of observing the alternative allele at s in each read, $f_s\left(P(r_j)|\hat{p},\hat{\sigma}^2\right) = \frac{1}{\sqrt{2\pi\hat{\sigma}^2}} e^{-\frac{(P(r_j)-\hat{p})^2}{2\hat{\sigma}^2}}$, and obtain $\hat{p}$ at each s according to:

$$\hat{p} = \frac{\sum_{\{r_j \in \mathbf{R}_{set} | r_j^s = 1 \vee r_j^s = 0\}} P(r_j)}{\xi}$$
$$\hat{\sigma}^2 = \frac{\sum (P(r_j) - \hat{p})^2}{\xi - 1} \tag{12}$$
$$P(r_j) = (1 - \epsilon_j^s) r_j^s + \epsilon_j^s (1 - r_j^s)$$

Having $\hat{p}$, a prior probability can be assigned to each of the 2^{k_m} and 2^{k_f} theoretically possible genotypes for the mother and the father, respectively, assuming a binomial model according to:

$$P(G_p^s) = \binom{k_t}{\nu} \hat{p}^{\nu} (1 - \hat{p})^{(k_t - \nu)} \tag{13}$$

where $p \in \{m, f\}$ and $0 \leqslant \nu \leqslant k_t$ is the dosage of the alternative allele in the candidate genotype, G_p^s. Assuming the parents have been independently chosen from a source population, a prior can be assigned to each (G_m^s, G_f^s) pair using $P(G_p^s)$ obtained from Equation 13, according to:

$$P(G_m^s, G_f^s) = P(G_m^s) \cdot P(G_f^s) \tag{14}$$

Given (G_m^s, G_f^s), a prior probability can be assigned to each specific offspring genotype, $G_{c_i}^s$, by counting the number of allele transmissions that result in that $G_{c_i}^s$. For example, with $(G_m^s, G_f^s) = \big((0,1,1,1),(1,0,0,0)\big)$, the prior $P(G_{c_1}|G_m^s, G_f^s)$ will be equal to 0, $\frac{9}{\binom{4}{2}\binom{4}{2}} = \frac{1}{4}$, $\frac{18}{\binom{4}{2}\binom{4}{2}} = \frac{1}{2}$, $\frac{9}{\binom{4}{2}\binom{4}{2}} = \frac{1}{4}$ and 0 for the offspring genotypes: $G_{c_1} = (0,0,0,0)$, $G_{c_1} = (1,0,0,0)$, $G_{c_1} = (1,1,0,0)$, $G_{c_1} = (1,1,1,0)$ and $G_{c_1} = (1,1,1,1)$, respectively.

To estimate the population genotypes, $(G_m^s, G_f^s, G_{c_1}^s, \cdots, G_{c_n}^s)$, we use the prior probabilities obtained as explained above, and assign a posterior probability to each population genotype by taking the sequencing reads into account. Noting that:

$$P(G_m^s, G_f^s, G_{c_1}^s, \cdots, G_{c_n}^s | \mathbf{R}_{set}, \boldsymbol{\epsilon}_{set}) = P(G_{c_1}^s, \cdots, G_{c_n}^s | G_m^s, G_f^s, \mathbf{R}_{set}, \boldsymbol{\epsilon}_{set}) P(G_m^s, G_f^s | \mathbf{R}_{set}, \boldsymbol{\epsilon}_{set}) \tag{15}$$

we separately obtain the posterior of the parental genotypes, $P(G_m^s, G_f^s|\mathbf{R}_{set}, \boldsymbol{\epsilon}_{set})$, and the conditional posterior of the offspring $P(G_{c_1}^s, \cdots, G_{c_n}^s|G_m^s, G_f^s, \mathbf{R}_{set}, \boldsymbol{\epsilon}_{set})$, from which the population posterior is derived using Equation 15. The posterior $P(G_m^s, G_f^s|\mathbf{R}_{set}, \boldsymbol{\epsilon}_{set})$ can be directly obtained from Equations 6 and 7 by substituting (H_m^s, H_f^s) with (G_m^s, G_f^s) in these equations and by using $P(G_m^s, G_f^s)$ (obtained by Equation 14) as the prior in Equation 6. Assuming conditional independence of the offspring genotypes given the parents, we obtain $P(G_{c_1}^s, \cdots, G_{c_n}^s|G_m^s, G_f^s, \mathbf{R}_{set}, \boldsymbol{\epsilon}_{set})$ by:

$$P(G_{c_1}^s, \cdots, G_{c_n}^s|G_m^s, G_f^s, \mathbf{R}_{set}, \boldsymbol{\epsilon}_{set}) = P(G_{c_1}|G_m^s, G_f^s, \mathbf{R}_{c_1}, \boldsymbol{\epsilon}_{c_1}) \cdot \ldots \cdot P(G_{c_n}|G_m^s, G_f^s, \mathbf{R}_{c_n}, \boldsymbol{\epsilon}_{c_n})$$
$$\mathbf{R}_{c_i} = \{r_j \in \mathbf{R}_{set} |\ \delta(r_j) = c_i\} \tag{16}$$
$$\boldsymbol{\epsilon}_{c_i} = \{\epsilon_j \in \boldsymbol{\epsilon}_{set} |\ \delta(r_j) = c_i\}$$

where $P(G_{c_i}|G_m^s, G_f^s, \mathbf{R}_{c_i}, \boldsymbol{\epsilon}_{c_i})$ is calculated according to:

$$P(G_{c_i}|G_m^s, G_f^s, \mathbf{R}_{c_i}, \boldsymbol{\epsilon}_{c_i}) = \frac{P(\mathbf{R}_{c_i}|G_{c_i}^s, \boldsymbol{\epsilon}_{c_i}) P(G_{c_i}^s|G_m^s, G_f^s)}{\sum\limits_{G_{c_i}'^s} P(\mathbf{R}_{c_i}|G_{c_i}'^s, \boldsymbol{\epsilon}_{c_i}) P(G_{c_i}'^s|G_m^s, G_f^s)} \tag{17}$$

and:

$$P(\mathbf{R}_{c_i}|G_{c_i}^s, \boldsymbol{\epsilon}_{c_i}) = \prod_{(r_j, \epsilon_j) \in \mathbf{R}_{c_i} \times \boldsymbol{\epsilon}_{c_i}} P(r_j|G_{c_i}^s, \epsilon_j) \tag{18}$$

where $\mathbf{R}_{c_i} \times \boldsymbol{\epsilon}_{c_i}$ represents the Cartesian product of $\mathbf{R}_{c_i}$ and $\boldsymbol{\epsilon}_{c_i}$, and (r_j, ϵ_j) denotes $r_j \in \mathbf{R}_{c_i}$ with its matched error rate vector, $\epsilon_j \in \boldsymbol{\epsilon}_{c_i}$. In Equation 18, $P(r_j|G_{c_i}^s, \epsilon_j)$ is obtained by replacing H_p^s with $G_{c_i}^s$ in Equation 8.

After calculating $P(G_m^s, G_f^s, G_{c_1}^s, \cdots, G_{c_n}^s |\ \mathbf{R}_{set}, \boldsymbol{\epsilon}_{set})$ from Equation 15, the most likely population genotypes at s can be assigned to the population members as genotype estimates.

C) Estimation of the offspring haplotypes

Having the set of all possible offspring phasings obtained by the possible transmissions of the parental haplotypes (Section A), we assign to each offspring c_i the phasing estimate $\hat{H}_{c_i}$ that yields the smallest number of required base-calling changes in the sequence reads, $\mathbf{R}_{c_i}$, in order to assign each $r_j \in \mathbf{R}_{c_i}$ to some homologue in $\hat{H}_{c_i}$. For each possible offspring phasing, $\hat{H}$, this required number of base-calling changes equals the so-called *minimum error correction* (*MEC*) score, defined as [6]:

$$MEC(\hat{H}, \mathbf{R}_{c_i}) = \sum_{r_j \in \mathbf{R}_{c_i}} \min_{\hat{h} \in \hat{H}} D(r_j, \hat{h}) \tag{19}$$

$D(r_j, \hat{h})$ is the Hamming distance between read $r_j \in \mathbf{R}_{c_i}$ and homologue $\hat{h} \in \hat{H}$ defined according to:

$$D(r_j, \hat{h}) = \sum_{\tau=1}^{l} d(r_j, \hat{h}, \tau) \tag{20}$$

where τ and l represent the SNP positions and the number of SNPs in the target region, respectively, and $d(r_j, \hat{h}, \tau)$ is defined in Equation 8. Thus, for each c_i we have $\hat{H}_{c_i} = \underset{\hat{H}}{\operatorname{argmin}}\, MEC(\hat{H}, \mathbf{R}_{c_i})$. If $\hat{H}_{c_i}$ is the same as the true phasing of c_i, its MEC score is expected to be close to the number of actual base-call errors in $\mathbf{R}_{c_i}$.

In case more than one set of parental haplotypes has the maximum probability (Section A), we infer the offspring haplotypes for each of them as explained above and finally choose the family whose total MEC score (summed over all offspring) is the smallest.

References

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

[2] Bourke, P. M., Voorrips, R. E., Visser, R. G., and Maliepaard, C. (2015). The double-reduction landscape in tetraploid potato as revealed by a high-density linkage map. *Genetics*, **201**(3), 853–863.

[3] Bourke, P. M., Arens, P., Voorrips, R. E., Esselink, G. D., Koning-Boucoiran, C. F., van't Westende, W. P., Santos Leonardo, T., Wissink, P., Zheng, C., Geest, G., *et al.* (2017). Partial preferential chromosome pairing is genotype dependent in tetraploid rose. *The Plant Journal*, **90**(2), 330–343.

[4] Edgar, R. C. and Flyvbjerg, H. (2015). Error filtering, pair assembly and error correction for next-generation sequencing reads. *Bioinformatics*, **31**(21), 3476–3482.

[5] Karp, A., Nelson, R., Thomas, E., and Bright, S. (1982). Chromosome variation in protoplast-derived potato plants. *TAG Theoretical and Applied Genetics*, **63**(3), 265–272.

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

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

5

AcroPoly: accurate estimation of multi-marker haplotypes using sequencing depth for finding trait loci in polyploids

This chapter is based on: Ehsan Motazedi, Arwa Shahin, Chris Maliepaard, Richard Finkers and Dick de Ridder, **AcroPoly: accurate estimation of multi-marker haplotypes using sequencing depth for finding trait loci in polyploids**, In preparation

Abstract

High throughput sequencing is becoming a convenient alternative to probe-based genotyping for finding loci associated with phenotypic traits in crop populations, also in polyploids. However, in polyploids determining the precise number of copies of the alleles (dosages) is challenging in the presence of sequencing noise.

We present an expectation-maximisation (EM) based approach, called AcroPoly, which uses read coverage to assign accurate probabilistic dosage scores to the alleles of multi-SNP haplotypes. Through simulations, we show that these scores reliably predict the actual dosages and can be directly applied to detect genotype-phenotype associations, providing more statistical power and precision compared to single SNP markers. For validation on real data, we used AcroPoly to estimate probabilistic dosage scores and to detect genotype-phenotype association to the genome-wide DNA sequence data of a moderate-sized F1-population derived from two heterozygous tetraploid Alstroemeria parents.

5.1. Introduction

Haplotypes are defined as sets of genomic variants at ordered and highly linked positions located on the same chromosome. For each set of positions there are $k = 2$ haplotypes in diploid species (such as humans) and $k > 2$ haplotypes in polyploid species, such as commercial potato (*Solanum tuberosum* L., $2n = 4x = 48$) and ornamentals like Peruvian lily (*Alstroemeria* spp. L., $2n = 4x = 32$). In segregating populations obtained by crossing, e.g. F1-populations with heterozygous parents, estimates of the phasing and recombination frequency are used to compute identity by descent (IBD) probabilities [51] for finding quantitative trait loci (QTL). Due to the limited number of meioses in such generations, even markers that are far from the actual trait locus (e.g. 10 cM) often remain strongly associated with it [32]. Besides, for outcrossing polyploids (such as potato) obtaining linkage maps is more complex than for diploids [7, 8].

Association mapping is an alternative approach to segregation mapping, which uses a dense set of markers and their identity by state (IBS) scores, and takes advantage of events that created correlations between causal variations affecting a trait and the marker alleles in the relatively distant past. For outcrossing crops, it can be assumed that many generations and therefore many meioses have elapsed since these events and thus associations with the causative loci are only found for markers at close genomic distance [32].

Compared to unphased genetic markers, multi-allelic haplotypes are more likely to uniquely associate with phenotypes and are therefore more powerful and robust for association mapping [13, 28, 34, 57], especially when the genetic component of the trait is determined by the interactions of a group of several variants [43, 45, 58]. Also, haplotypes can better reflect genetic diversity within and between populations [27, 52] and can help correct dosage calling errors using inheritance information [39, 47]. However, molecular determination of haplotypes is often costly and laborious as it requires differentiating between the chromosomes [29], while high-throughput methods exist to determine unphased markers, in particular single nucleotide polymorphisms (SNPs), using probe-based genotyping arrays or sequence data [23, 36, 49]. Therefore computational methods have been developed to estimate haplotype phasing from unphased genotypes in families [1, 54] and in random mating populations [11, 18, 41], or from unphased genotypes and aligned sequence reads for single individuals [4, 6, 16, 55] or for families [20, 47, 48].

Sequence-based methods have the advantage of being able to identify all of the variants that exist in a target genomic region, as well as providing phasing information for variants covered by the same read. However, the obtained reads must be first aligned to a common reference (either previously existing or obtained *de novo*), and true variations must be distinguished from base-calling errors [31]. While it is relatively easy to detect SNP positions by sequencing, accurate determination of SNP genotypes can be challenging in polyploids, as the number of possible allele dosages increases with the ploidy level and presence of complex variations might result in errors in the alignment [21, 49]. The problem of haplotype estimation, or SNP phasing, is also more difficult in polyploids. The available methods, which optimise either a likelihood [6, 47, 48] or a cost function [2, 16, 55], can converge to local optima and result in non-existent haplotypes or wrong haplotype dosages [46].

In association mapping, it is desirable to have a set of markers with accurate allele

dosages or frequencies that can reliably tag the causal genetic variants [3, 40, 58]. For a genomic region defined by l bi-allelic SNPs, 2^l haplotype alleles are possible, whose combinations and different possible dosages yield $\binom{2^l+k-1}{k}$ theoretically possible phasings for an individual with ploidy level k. These numbers grow quickly: regions of 4, 5 and 6 SNPs lead to 136, 528 resp. 2,080 possible phasings in diploids and to 3,876, 52,360 and 766,480 in tetraploids. This problem can be circumvented by breaking the region into scanning windows that each include only a few SNPs, enumerating all haplotype alleles in a window and assigning a dosage to each based on its read count. However, as the chromosome of origin, or the haplotype, is not directly observed for sequence reads, probabilistic models are needed to indirectly estimate the dosages from the number of reads compatible with each haplotype allele. This is an example of *latent class analysis* (*LCA*) or probabilistic clustering [5, 24], which yields fractional scores or probabilistic dosages (in contrast to the discrete dosages limited to $0, 1, \cdots, k$) for each haplotype. Using these probabilistic dosages, decision can be made on the actual (categorical) dosage of each haplotype in an individual or the frequency of haplotypes in a population. However, the probabilistic dosages can also be directly used in genetic analysis, as a way of taking into account that the estimates of the unobserved haplotype dosages are uncertain. Such an approach has successfully been used for genotype imputation [39] and for genotype calling from messy sequence data [22].

Here we propose an expectation maximisation (EM) approach, called AcroPoly, to determine probabilistic haplotype dosages of an individual using its aligned sequence reads. After calling SNPs in a target genomic region, we use scanning windows that each cover a fixed, small number of SNPs and assign probabilities to every possible haplotype in each window based on the number of supporting reads. An important advantage of this approach is that it can combine the sequence data of several individuals for dosage estimation, without being restricted to a certain population structure [3, 9, 47] or imposing *a priori* assumptions on the haplotypes present in the population [14, 33, 42].

We assess the accuracy of AcroPoly in predicting the actual dosages of haplotypes through simulations of tetraploid ($k = 4$), hexaploid ($k = 6$) and octoploid ($k = 8$) random mating populations, based on the reference genome of potato (*S. tuberosum*) cultivar DM [15], and compare it with other state-of-the-art single individual haplotyping (SIH) tools. Through simulations of continuous phenotypes and tetraploid loci, based on *S. tuberosum*, with various degrees of heritability, we show that the probabilistic multi-SNP haplotype dosages obtained by AcroPoly provide more statistical power than single SNP markers obtained by the traditional polyploid variant caller FreeBayes [21]. Finally, we apply AcroPoly to obtain multi-SNP haplotype markers in an F1-population with 82 offspring derived by the crossing of two heterozygous parents of tetraploid Alstroemeria, a popular outcrossing ornamental native to South America with a large and complex genome [25, 26]. We show that AcroPoly allows to detect scattered markers linked to the trait locus involved in disease resistance.

5.2. Material and Methods

We propose an Expectation-Maximisation (EM) approach to obtain allele dosages for haplotype alleles that consist of s bi-allelic SNPs using sequencing read depths. As the

number of possible alleles grows exponentially with s, we limit s to 3 or 4 and scan a genomic region with $l > s$ SNPs using windows of size s. The details of this EM approach, called AcroPoly, are given in Section 5.2.1. We compare the accuracy of AcroPoly in predicting the actual haplotype dosages, with optimisation based approaches HapTree [6] and SDhaP [16] (Section 5.2.2), and compare the power of testing for trait associations using its multi-SNP markers, with using single SNP markers estimated by the polyploid variant caller FreeBayes [21] (Section 5.2.3). Finally, we apply AcroPoly to detect genomic positions associated with resistance to leaf scotch in an outcrossing F1-population of tetraploid Alstroemeria with two heterozygous parents (Section 5.2.4).

5.2.1. Estimation of multi-SNP haplotype allele dosages from read depth

We assume a Poisson distribution for the unobserved read count of each possible haplotype in a window of size s, corresponding to $N = 2^s$ haplotype alleles with bi-allelic SNPs, and aim to estimate the rates of these Poisson distributions using the sequence reads. Within the window ω_t, the estimated Poisson rates, $\boldsymbol{\mu}_t = (\mu_{1t}, \mu_{2t}, ..., \mu_{Nt})^T$, are used to assign a dosage score to each possible haplotype marker, h_{it} $(i = 1, \ldots, N)$, according to:

$$P(h_{it}|\boldsymbol{\mu}_t) = \frac{\mu_{it}}{\sum_{j=1}^{N} \mu_{jt}} \tag{5.1}$$

Starting at the first SNP position in a region of interest, we can scan the whole region of l SNPs by shifting (sliding) the window. One can get overlapping windows by shifting the window by one SNP (or less than s SNPs) at a time, or non-overlapping windows by shifting by s SNPs at a time. With a 1 SNP shift, the total number of windows will be l-(s-1), while $\lceil l/s \rceil$ non-overlapping windows are required to cover the whole region.

For each window ω_t, we determine $\boldsymbol{\mu}_t$ using the Expectation-Maximisation (EM) algorithm [17, 37]. To this end, we first construct its compatibility matrix $\boldsymbol{M}_t$ indicating the compatibility of each read with the possible haplotypes in ω_t. We only consider those reads that contain at least one SNP position within ω_t, and call a read r_x compatible with a haplotype h_{it} if it contains the same alleles as h_{it} at the common SNP positions. $\boldsymbol{M}_t$ is accordingly defined as:

$$\boldsymbol{M}_t = \begin{bmatrix} m_{1,1} & m_{1,2} & \cdots & m_{1,N} \\ m_{2,1} & m_{2,2} & \cdots & m_{2,N} \\ \vdots & \vdots & \ddots & \vdots \\ m_{c_t,1} & m_{c_t,2} & \cdots & m_{c_t,N} \end{bmatrix}, \; m_{x,i} = \begin{cases} 1 & r_x \in h_{it} \\ 0 & r_x \notin h_{it} \end{cases} \tag{5.2}$$

where c_t denotes the total number of reads informative for ω_t and the notation $r_x \in h_{it}$ is used to denote the compatibility of r_x with h_{it}. From $\boldsymbol{M}_t$, we derive $\boldsymbol{p}_t = \boldsymbol{M}_t\boldsymbol{\mu}_t = (p_{1t}, p_{2t}, ..., p_{c_t t})^T$ and $\boldsymbol{\delta}_t = \boldsymbol{M}_t \text{diag}(\boldsymbol{\mu}_t)$, from which the count matrix $\boldsymbol{A}_t$ is constructed according to:

$$\boldsymbol{A}_t = \begin{bmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,N} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,N} \\ \vdots & \vdots & \ddots & \vdots \\ a_{c_t,1} & a_{c_t,2} & \cdots & a_{c_t,N} \end{bmatrix}, \; a_{x,i} = \frac{m_{x,i}\mu_{it}}{\sum_{i=1}^{N} m_{x,i}\mu_{it}} = \frac{\delta_{x,i}}{p_{xt}} \tag{5.3}$$

where $\delta_{x,i}$ denotes the (x,i) element of $\boldsymbol{\delta}_t$. The sum of the values in column i of $\boldsymbol{A}_t$, $\sum_{x=1}^{c_t} a_{x,i}$, gives the number of reads compatible with h_{it} taking into account that a read might be compatible with more than one haplotype. As an example, the read $(s_1 = 0, s_2 = 1)$ is compatible with two haplotypes: $(s_1 = 0, s_2 = 1, s_3 = 0)$ and $(s_1 = 0, s_2 = 1, s_3 = 1)$ (as the read does not include s_3), hence adding $\frac{1}{2}$ to the total read count of each haplotype. With this definition, it is evident that $\sum_{i=1}^{N}\sum_{x=1}^{c_t} a_{x,i} = c_t$.

Defining $\boldsymbol{k}_t = (k_{1t}, k_{2t}, ..., k_{Nt})^T$ as the unobserved read counts of the haplotypes, we use the notation presented above to formulate the log-likelihood of the reads using conditionally independent Poisson distributions according to:

$$\mathscr{L}(\boldsymbol{\mu}_t^T; M) = \sum_{i=1}^{N} \log\big(Pois(k_{it}; s\mu_{it})\big) \tag{5.4}$$
$$\boldsymbol{k}_t \sim Mult\Big(c_t, \frac{1}{c_t}\boldsymbol{J}_{1,c_t}\boldsymbol{A}_t\Big)$$

where $\boldsymbol{J}_{1,c_t}$ denotes the $1 \times c_t$ vector of ones. As the total number of reads that cover ω_t is fixed and equal to c_t, the unobserved numbers of reads k_{it} originating from each haplotype h_{it} follow a multinomial distribution with the probability corresponding to each haplotype being related to the dosage of that haplotype according to Equations 5.3 and 5.4. However, it is difficult to directly maximise the log-likelihood $\mathscr{L}(\boldsymbol{\mu}_t^T; M)$ in Equation 5.4, as $\boldsymbol{k}_t$ is unobserved. Therefore, we start by some initial value for $\boldsymbol{\mu}_t$, such as $\boldsymbol{\mu}_t^{(0)} = (\frac{1}{N}, \cdots, \frac{1}{N})^T$, and try to maximise $\mathscr{L}(\boldsymbol{\mu}_t^T; M)$ by replacing $\boldsymbol{k}_t$ with $\mathbb{E}[\boldsymbol{k}_t|\boldsymbol{A}_t]$ and maximising the modified log-likelihood:

$$\mathscr{Q}(\boldsymbol{\mu}_t^T; \mathbb{E}[\boldsymbol{k}_t|\boldsymbol{A}_t]) = \sum_{i=1}^{N} \log\big(Pois(\sum_{x=1}^{c_t} a_{x,i}; s\mu_{it})\big) \tag{5.5}$$

It is straightforward to show (see Appendix) that $\nabla\mathscr{Q}(\boldsymbol{\mu}_t^T; \mathbb{E}[\boldsymbol{k}_t|\boldsymbol{A}_t]) = \boldsymbol{0}$ for:

$$\boldsymbol{\mu}_t^T = \frac{1}{s}\boldsymbol{J}_{1,c_t}\boldsymbol{A}_t = \frac{1}{s}\mathbb{E}[\boldsymbol{k}_t|\boldsymbol{A}_t] \tag{5.6}$$

and the maximisation is performed by updating $\boldsymbol{\mu}_t^T$ according to Equation 5.6 using the current value of $\boldsymbol{A}_t$ at iteration z, $\boldsymbol{A}_t^{(z)}$. The E-step (Equation 5.6) and M-step (Equation 5.3) are iterated until the algorithm converges, i.e. if:

$$\Big|\mathscr{Q}\big(\boldsymbol{\mu}_t^{T^{(z+1)}}; \mathbb{E}[\boldsymbol{k}_t|\boldsymbol{A}_t^{(z+1)}]\big) - \mathscr{Q}\big(\boldsymbol{\mu}_t^{T^{(z)}}; \mathbb{E}[\boldsymbol{k}_t|\boldsymbol{A}_t^{(z)}]\big)\Big| < \boldsymbol{\varepsilon} \tag{5.7}$$
$$\boldsymbol{\mu}_t^{T^{(z+1)}} = \frac{1}{s}\boldsymbol{J}_{1,c_t}\boldsymbol{A}_t^{(z)}$$

with $\boldsymbol{\varepsilon}$ a preset convergence threshold. We used $\boldsymbol{\varepsilon} = 10^{-6}$ in our analyses.

In order to avoid uncertain marker scores at loci with low sequencing coverage, we set a threshold on the required coverage c_t for each window ω_t and estimate no haplotypes for ω_t if c_t falls below. In our analyses, we set this threshold equal to the ploidy level $k = 4$.

5.2.2. Evaluation of accuracy of haplotype dosages

We simulated 200 genomic loci of length 10 kb, considering ploidy levels $k = 4$, 6 and 8. The reference sequence of each locus was randomly selected from chromosome 5 of the PGSC potato genome [15], which is known to harbour important trait loci [35], and SNPs were introduced at random positions in the selected reference as described in Motazedi *et al.* [46]. For each simulated locus, we simulated paired-end Illumina HiSeq 2000 reads with an average insert-size of 350 bp (single read length of 100 bp) at an average depth of 15× per homologue (i.e. 60-120× per chromosome), according to a uniform distribution, using ART [30]. This average depth was chosen as it has been shown to yield high accuracy for haplotyping [46]. The simulated reads were aligned to the reference using bwa-mem [38]. After calling SNPs from the aligned reads using FreeBayes [21], we estimated the haplotypes using AcroPoly, setting the scanning window size to $s = 4$ SNPs and using non-overlapping phasing windows to scan the whole locus. For the sake of comparison, we also estimated the phasing of each simulated locus using the SIH methods HapTree [6] and SDhaP [16].

To evaluate the accuracy of AcroPoly dosage scores in predicting the actual phasing and to compare it with HapTree and SDhaP, we ranked its haplotype alleles upon their probabilistic dosages, then multiplied the dosages by the ploidy level k and rounded to the nearest integer. We determined the phasing by choosing haplotypes starting from the top rank allele, until the dosages of the chosen haplotypes added up to the ploidy level k. This procedure was followed for each scanning window ω_t. If the sum of integer dosages could not be set equal to k for a window, e.g. because of low probabilities for most of the haplotype alleles due to insufficient sequencing coverage, some of the haplotypes could not be clearly predicted in the phasing. In such cases, we considered the phasing missing for the window.

As the length of the haplotype blocks in AcroPoly is limited to its preset window size, i.e. 4 in our simulations, we used the reconstruction rate (RR) [48] to measure the overall local similarity of the estimated blocks to the true phasings. To assess the accuracy of the AcroPoly dosage scores in phasing prediction, we used the pair-wise phasing accuracy rate (PAR) [46], which shows the fraction of correctly predicted haplotype dosages of SNP pairs in a locus. We also compared the methods in terms of the ratio of SNPs that were phased by each to the total number of simulated SNPs.

5.2.3. Use of AcroPoly multi-SNP haplotypes as markers for detecting associations to traits

To show how the multi-SNP haplotype alleles obtained by AcroPoly can be used for detecting associations to traits in outcrossing polyploid populations, and to compare their power and mapping precision with sequence-based single SNP markers estimated by conventional polyploid variant callers, we simulated small populations of tetraploid potato with 20 offspring from a cross of 'Altus' and 'Colomba' parents [47]. We considered this small population size to demonstrate, with computational ease, the gain in power by using multi-SNP haplotype markers compared to using single SNPs.

In each simulation, we generated a phenotype influenced by one and only one of the 9 loci with genomic coordinates and numbers of SNPs shown in Table 5.1, with no other genotypic effects on the trait. For each of these loci, we had the phasing estimated in

Table 5.1: *S. tuberosum* loci selected for quantitative trait locus simulation

Gene	DNA sequence id	Chromosome: coordinates	Segregating bi-allelic SNPs	Distinct haplotypes
StCDF1	PGSC0003DMG400018408	chr05:4538880-4541736	38	6
StCDF2	PGSC0003DMG400025129	chr02:25588000-25591776	63	8
StCDF3	PGSC0003DMG400001330	chr02:46143998-46147444	75	7
StCDF4	PGSC0003DMG400033046	chr06:51598497-51601151	51	3
StCDF5	PGSC0003DMG400019528	chr03:55882564-55885296	100	7
StCO1	PGSC0003DMG401010056	chr02:45098374-45101578	57	3
StCO2	PGSC0003DMG402010056	chr02:45088023-45092647	66	6
StFKF1	PGSC0003DMG400019971	chr01:531784-536380	89	8
StGI1	PGSC0003DMG400001110	chr03:14265390-14266279	40	8

'Altus' and 'Colomba' from a previous study using PopPoly [47]. To simulate the genetic component of each phenotype, we randomly assigned two haplotypes at the causal locus from each parent to each offspring. Sequence reads were simulated for each offspring with an average depth of 10× per homologue and per locus, using the same approach as described in 5.2.2. The reads were aligned to the potato reference using bwa-mem [38] and SNP positions were identified using FreeBayes [21].

Quantitative phenotypes were generated by randomly selecting 5 causal SNPs with equal probabilities from the set of SNPs at each locus in each simulation. With the numbers of SNPs and genomic region lengths as shown in Table 5.1, in this way we simulated allelic heterogeneity with varying physical distances between the causal SNPs. The genetic component of each phenotype was simulated according to two models of trait inheritance: 1) the haplotype interaction model, corresponding to local epistasis between the SNPs [12, 53, 56] and 2) the independent SNP effect model, in which the effect of a SNP allele is independent from its host haplotype [50]. The haplotype interaction model represents situations where it is unlikely that a single marker SNP can perfectly tag the causative haplotype allele, and the independent SNP effect model represents the opposite, where it is likely that single marker SNPs can tag individual causative SNP alleles. As the diversity of parental haplotypes ranges from low (3) to the highest (8) at the chosen loci (Table 5.1), we believe that our simulations successfully reflect various degrees of markers' linkage disequilibrium with the causative alleles.

For simulating the haplotype interaction model, we randomly chose one of the occurring haplotype alleles of the 5 causal SNPs and associated it with a positive effect on the phenotype, while the other occurring haplotype alleles were assigned a negative effect of a magnitude equal to the positive effect. The genetic component of the phenotype was determined by multiplying the effect assigned to each haplotype allele by its dosage and then adding up the dosage effects of the occurring alleles in each individual. Gaussian noise with zero mean and a variance of σ_e^2 was simulated for each individual and

added to its genetic component to achieve the desired heritability h^2 in the trait using:

$$h^2 = \frac{\sigma_g^2}{\sigma_g^2 + \sigma_e^2} \tag{5.8}$$

to determine σ_e^2, where σ_g^2 is the variance of the simulated genetic component of the trait. To evaluate the dependency of the method's statistical power on trait heritability, we considered $h^2 = 0.1, 0.2, 0.3, 0.5, 0.6$ and 0.8. By setting σ_g^2 to zero, we also simulated phenotypes with $h^2 = 0$ to investigate the false positive rates of association tests. For each value of h^2, 15 populations were simulated for each locus (135 simulations in total).

For the independent SNP effect model, we assumed that at each of the 5 causal SNP sites the causative allele, i.e. one of the two occurring alleles randomly chosen with equal probability, has the same effect on the phenotype regardless of the haplotype on which it is located. The genetic component was thus determined by multiplying the effect of the causative allele at each SNP site by its dosage and then summing the dosage effects of the 5 SNPs in each individual. For simplicity, we considered an equal effect size for the causative alleles at all the SNP sites. For each value of h^2, 15 populations were simulated for each locus, as described for the haplotype interaction model.

Sequence reads were simulated for each individual as described in Section 5.2.2, mimicking whole genome sequencing (WGS) or amplicon sequencing of the targeted loci. To find back the genetic variants in each locus underlying the simulated quantitative phenotypes, we called SNPs by aligning the sequence reads and applying FreeBayes [21], as described in 5.2.2.

To test for association of single bi-allelic SNP markers to a trait, we built regression models per SNP site that related the simulated trait to the called dosage of the alternative allele. The significance of the estimated regression coefficients was tested using a likelihood ratio test (LRT) (compared to the null model) at $\alpha = 10^{-5}$, which is a significance level close to what is often used with a large set of genome wide markers. For the tests, we only considered those detected SNPs that had a minor allele frequency of at least 0.1, and a dosage calling failure rate of at most 0.4 in the simulated population. Individuals were excluded from the analysis at each simulation/locus if they had a SNP missing rate of more than 0.8 at the locus. These filtering steps are often taken in practice to ensure conclusions are made by only considering high quality polymorphic markers and individuals [44].

We used the probabilistic dosage scores obtained by AcroPoly for each haplotype allele in sliding windows of 3 SNPs to detect associations with the simulated traits at each locus, shifting the sliding windows by one SNP to cover the whole locus. In comparison to the simulations in Section 5.2.2, we set a smaller window size to obtain fewer, more accurate haplotypes, hence fewer, more reliable predictors in the trait models. Shifting windows one SNP at a time ensures that the density of haplotype markers becomes the same as the single SNPs on each locus. The parameters of the EM algorithm were set as in Section 5.2.2.

From the obtained haplotypes in each window, we filtered out from each trait model those alleles that had a score of zero in more than 80% of the individuals. Using the dosages of the remaining haplotype alleles in each individual, we built a multiple linear

regression model for the trait for each haplotype marker and used an LRT to test the significance of the marker-trait association ($\alpha = 10^{-5}$). By using the probabilistic dosages, we down-weighted the effect of uncertain alleles on these associations, while still including their information in the phenotype model.

While every association with a marker linked to the causative locus is a true positive in the genetic sense, we aimed to evaluate the precision of dense haplotype markers and single SNPs in reaching close to the causative variants within a locus. This precision is important in deciphering the mechanisms underlying a trait [10], and it is especially desirable with WGS to detect the actually causative variations and loci. To measure the precision on the simulated loci, which were relatively small (Table 5.1), we considered a significant association a true positive if its upstream or downstream distance to one of the causal SNPs was less than 500 bp, i.e. a distance on average equal to 20% of the lengths of the simulated loci. For the single SNP markers, the distance to a causal SNP was calculated from the marker positions, while for the haplotype markers this distance was defined from the start or stop position of the haplotype, whichever was the closest. The precision was obtained by dividing the number of true positive significant associations to the total number of significant associations. We also calculated the recall rate of each marker type, as the ratio of causal SNPs for which a significant marker was found within the 500 bp upstream or downstream distance, to the total number of causal SNPs. To estimate and compare the false positive rates (FPR) of AcroPoly and FreeBayes markers, we calculated the ratio of the significant associations found with $h^2 = 0$, which were all false positives, to the total number of markers in each simulation.

5.2.4. Detection of trait loci in Alstroemeria using estimated haplotype scores

Using AcroPoly, we estimated haplotype dosages from bait-capture exome sequence data in a cross of two tetraploid *Alstroemeria* cultivars, named 'ALS2' and 'ALS3', with 82 offspring. The RNA baits for capturing single copy genes were designed using a *de novo* assembly of independent RNA sequence data of the same cross, consisting of 5786 target contigs (with a median contig length of 1.474 kb). The DNA sequence reads were obtained using paired-end sequencing by Illumina HiSeq 2000 technology. After mapping the reads back to the target contigs, we called SNPs using FreeBayes and estimated haplotype markers using AcroPoly, requiring a sequencing coverage of at least 4×, i.e. on average 1× per homologue, over a window to call the SNP dosages and haplotype scores for an individual. We applied logistic regression using the probabilistic dosage scores of AcroPoly as predictors to find haplotype markers linked to disease resistance in a bi-parental population obtained from crossing a resistant parent with a susceptible parent.

To obtain AcroPoly haplotype markers, SNP positions were detected by FreeBayes, filtering out those with a calling failure rate of 0.3 or more in the population. This quality threshold was more stringent compared to the simulations (Section 5.2.3), as a much larger number of SNPs had been called, with many suspected false positives due to insufficient sequencing depths at some contigs. Probabilistic haplotype dosage scores were obtained over sets of 3 SNPs (yielding m^3 haplotype alleles for m-allelic SNPs), and the phasing window was shifted one SNP at a time to scan a whole contig. After obtaining

Table 5.2: Reconstruction rates (RR) obtained by phasing methods at various ploidy levels

Estimation method	$k = 4$	$k = 6$	$k = 8$
AcroPoly	0.91	0.96	0.96
HapTree	0.88	0.92	0.95
SDhaP	0.79	0.78	0.79

dosage scores of the haplotype alleles in all of the individuals, we kept only those alleles that had a non-zero score in at least 20% of the individuals and a missing score rate at most 0.4. We used multiple logistic regression per marker with haplotype allele dosages as the predictors and tested the association of each haplotype marker with the resistance trait using the LRT. We applied a Bonferroni correction for multiple testing and set the significance level for each marker according to the number of markers that passed the filtering steps, so that the genome-wide significance level $\alpha = 0.05$ was maintained.

5.3. Results

We evaluated AcroPoly through simulations of polyploid genomes with various levels of ploidy, as well as quantitative single locus phenotypes with various levels of heritability in tetraploid populations. Two objectives were aimed for in this evaluation: 1) the ability of AcroPoly to correctly predict multi-SNP haplotypes in polyploid individuals, 2) the usefulness of its probabilistic haplotype markers in finding trait associations in polyploid populations. For the first objective, we compared the performance of AcroPoly with state-of-the-art optimisation-based haplotyping approaches HapTree and SDhaP. For the second objective, we compared the precision and recall rates of linking traits to AcroPoly markers and to single SNP markers obtained by FreeBayes. Finally, we used the probabilistic scores of AcroPoly over genome-wide 3-SNP haplotypes to find loci associated with disease resistance in an outcrossing F1-population of Alstroemeria.

5.3.1. AcroPoly yields accurate dosage scores for multi-SNP haplotypes

Measuring the reconstruction rate (RR) of the true haplotypes showed that the phasings obtained by AcroPoly are substantially more accurate compared to SDhaP and HapTree, as shown in Table 5.2. While the accuracy scores of HapTree were still close to those of AcroPoly, AcroPoly had the important benefit of steady performance, especially at ploidy levels $k \geq 6$. Actually, HapTree was unable to produce any phasing estimate due to numerical instability at 15% and 40% of simulations at $k = 6$ and $k = 8$, respectively, and left 69% and 83% of the SNPs out of its estimates in the other cases. As a result, only 26% and 10% of the simulated SNPs were phased by HapTree at these ploidy levels, respectively, compared to around 73% by AcroPoly and 90% by SDhaP (Table 5.3).

It must be noted, however, that the probabilistic scores reported by AcroPoly for

Table 5.3: Proportions of simulated SNPs phased by each method at various ploidy levels

Estimation method	$k = 4$	$k = 6$	$k = 8$
AcroPoly	0.69	0.73	0.73
HapTree	0.73	0.26	0.10
SDhaP	0.87	0.89	0.90

each haplotype allele do not solely depend on the dosage of the haplotype alleles in the genome, but also on the certainty of assigning sequence reads to each haplotype. If, for example, a base is not called at a SNP position within a read, its ascription to a haplotype will not be unique as the read is compatible with several haplotype alleles. Such uncertainties are reflected in the read-count rates estimated by AcroPoly, but not if the estimated rates are rounded to predict the actual haplotype dosages in an individual. The increase in RR observed at higher ploidy levels (Table 5.2) shows in fact that this round-off error is less manifest at higher ploidy levels.

To assess the accuracy of the haplotype dosage scores predicted by AcroPoly, and to compare it with HapTree and SDhaP, we used the pair-wise phasing accuracy rate (PAR) which measures the fraction of correctly assigned haplotype dosages for SNP pairs located on a locus (Table 5.4). The highest accuracy was obtained by HapTree, but its failure rate was high at $k = 6$ and $k = 8$ and, as noted above, most of the SNPs were left out from its phasing at these ploidy levels. The dosage prediction accuracy of AcroPoly was quite close to that of HapTree (Table 5.4). Overall, AcroPoly was the most stable and accurate method for predicting haplotype dosages among the tested methods. Nevertheless, the haplotype estimates of AcroPoly are restricted in length, i.e. the number of SNPs included in a haplotype block, 4 in these simulations, where haplotype blocks estimated by HapTree and SDhaP often included 50 or more SNPs. While the higher accuracy of AcroPoly is partly due to this restriction in the haplotype length, the main advantage is that the restricted length allows AcroPoly to assign a score to every possible haplotype allele instead of choosing only k alleles. In this way, genetic diversity in a population is better reflected, especially for statistical analysis of phenotypes.

5.3.2. AcroPoly increases precision and recall rate of trait-locus association detection

By simulating single locus phenotypes with levels of heritability ranging from 0 to 0.8, we used the probabilistic scores obtained by AcroPoly for multi-SNP haplotypes to detect associations between phenotypes and their underlying causal variants in small populations of 20 individuals. We compared the results obtained by AcroPoly to the results obtained by using single SNP marker dosages obtained by FreeBayes. We compared the precisions of the two approaches in positioning causal variations within a locus, as well

Table 5.4: Ratio of the correctly predicted haplotype dosages for SNP pairs at various ploidy levels

Estimation method	$k=4$	$k=6$	$k=8$
AcroPoly	0.67	0.71	0.71
HapTree	0.77	0.76	0.76
SDhaP	0.62	0.45	0.35

as their recall and false positive rates.

Figure 5.1-(a) and Figure 5.1-(c) show the precision of the AcroPoly and FreeBayes markers in finding quantitative trait loci as a function of heritability. These results show that markers found by AcroPoly detect the causal loci more precisely than single SNP markers (estimated by FreeBayes), especially at low degrees of heritability. For both haplotype interaction and independent SNP effect models, the precision of AcroPoly already reaches around 0.8 for a heritability as low as 0.1, while FreeBayes reaches the same level only for $h^2 > 0.3$ with the independent SNP effect model and for $h^2 > 0.8$ with the haplotype interaction model.

These results indicate the superiority of multi-allelic short haplotype markers in detecting the causative variants, especially for the haplotype interaction model. When the effect of a causal SNP allele is independent of its phasing, as in the independent SNP effect model, bi-allelic SNP markers at close distances can still efficiently tag the causal SNPs. This is reflected in Figure 5.1-(c), as the precisions of AcroPoly and FreeBayes markers are both as high as around 80% when the heritability is more than 0.3. In contrast, in the haplotype interaction model the phasing of the causal SNPs' alleles determines the genetic component of the trait, hence the precision of single SNP markers in tagging the causal alleles substantially decreases and falls below that of the haplotype markers (Figure 5.1-(a)).

The recall rates of AcroPoly and FreeBayes at various levels of heritability are also shown in Figure 5.1. AcroPoly markers help detect more causal variants with both the haplotype interaction model (Figure 5.1-(b)) and the model with independent SNP effects (Figure 5.1-(d)). Both AcroPoly and FreeBayes, however, perform better with the independent SNP effect model, with AcroPoly reaching a recall of around 0.95 at $h^2 > 0.6$, compared to the haplotype interaction model of inheritance for which the maximum recall at the same levels of heritability ($h^2 > 0.6$) is around 0.6. This is to be expected, as the haplotype markers estimated by AcroPoly are still local and do not span beyond a few SNPs ($s = 3$), which means they may not be able to uniquely identify functional alleles that span tens of SNPs.

Using the simulations of phenotypic traits without a genetic component ($h^2 = 0$), we obtained the average values and the 99% confidence intervals for the false positive rates (FPR) of AcroPoly and FreeBayes to be 0.003(0;0.07) and 0.0005(0;0.039), respectively. The FPR was therefore very small for both methods and not significantly larger than zero

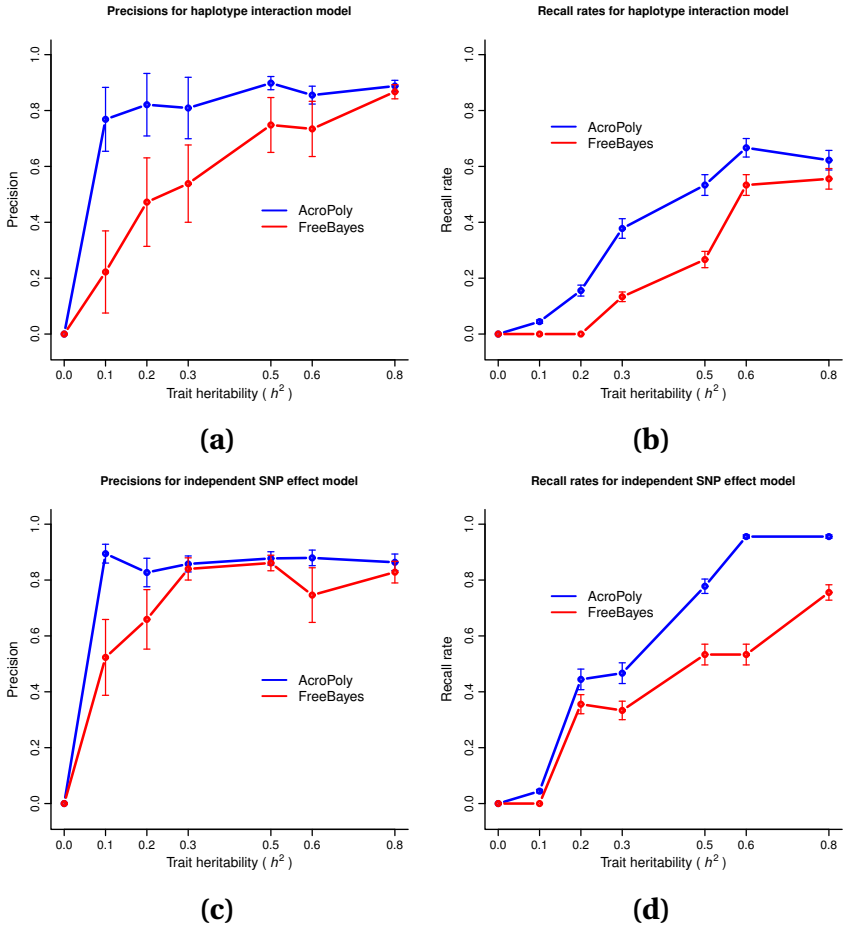


Figure 5.1: Precisions and recall rates, respectively, of AcroPoly and FreeBayes association analyses for (a),(b) haplotype interaction and (c), (d) independent SNP effect phenotype model, at different heritabilites in simulated populations of 20 individuals. The error bars indicate the standard errors of the precision averages over 135 simulations at each level of heritability.

at $\alpha = 0.01$, although it was on average slightly higher for AcroPoly.

These results show that the multi-allelic short haplotype markers scored by AcroPoly are more powerful compared to single SNP dosages and can also detect the causal variants with higher precision. Therefore, AcroPoly markers are better candidates for detecting trait associations.

5.3.3. Analysis of the Alstroemeria cross

We applied AcroPoly to paired-end Illumina sequence reads, obtained by a bait-capture approach, to investigate the genetic background of disease resistance in an outcrossing F1 population of Alstroemeria with heterozygous tetraploid parents and 82 offspring.

Table 5.5: Significant associations with disease resistance in the Alstroemeria cross

Contig name	Number of significant markers	Markers spanning coordinates (bp)	Average contig p-value*
Contig5481	2	980-989	2.16×10^{-8}
Contig10910	1	397-442	4.21×10^{-8}
conContig14495	3	1366-1389	2.29×10^{-8}
singconContig29807	1	156-167	4.46×10^{-8}
Contig8382	3	2851-2911	1.24×10^{-8}
conContig9987	1	982-989	4.69×10^{-8}
conContig21958	2	288-401	1.20×10^{-8}
singconContig40243	3	101-248	8.07×10^{-9}
conContig22185	1	325-333	2.91×10^{-8}
Contig4019	1	1041-1074	1.78×10^{-9}

* *Geometric mean of the p-values of significant markers*

After aligning the reads to the target contigs, the average sequencing coverage per individual was found to be 16× (SD=26×) with an average insert-size of 220 bp (SD=17 bp).

To reduce the effect of incorrectly mapped reads, especially those that might originate from non-target regions, we set the minimum mapping quality, as well as the minimum base calling quality, to 10 for variant calling and limited the number of variant alleles to 6 in FreeBayes for its scanning window [21]. In total, 1,357,668 SNPs were called using this approach. Haplotype dosages were estimated for 3-SNP haplotype windows, shifting the window by 1 SNP at a time. After filtering haplotype markers found in only a small subset of individuals (< 20%), 188,358 markers remained for regression analysis, as explained in Section 5.2.4. With this number of markers ending up in the trait analysis, we set the per marker significance level to $\alpha = 5 \times 10^{-8}$ (Section 5.2.4).

Using the LRT, 18 AcroPoly markers, scattered on 10 contigs, were found significantly associated with disease resistance (Table 5.5). For each contig with significant markers, an average p-value was calculated as the geometric mean of the p-values of all significant markers. The genetic distance between the 10 detected contigs was estimated from the combined sets of their SNPs using polymapR [8], which showed that all of the contigs were tightly linked and belonged to the same genomic locus, suggesting the resistance is a monogenic trait (data not shown).

5.4. Conclusion and Discussion

We present AcroPoly, an expectation maximisation (EM) approach to assign probabilistic dosages to multi-SNP haplotype alleles in polyploids. These dosages correspond to the estimated rates of Poisson distributions for the number of reads originating from each

haplotype allele. Through simulations, we showed that these probabilistic dosages can reliably predict the actual dosages even at high ploidy levels ($k \geq 6$) and can be used as powerful multi-allelic markers for finding associations to traits, provided that the read depth is sufficient to infer the latent Poisson rates and there is no sequencing bias affecting the coverages of the haplotypes. We applied AcroPoly to the genome-wide DNA sequence data obtained from an outcrossing F1-population of Alstroemeria with heterozygous tetraploid parents and 82 offspring, and found several haplotype markers significantly associated with disease resistance in this cross.

However, there are situations where single SNPs could be just as effective as the haplotypes, e.g. if the causal haplotypes are (almost) perfectly tagged by single SNPs or if the trait is shaped by multiple SNPs at unlinked loci. In such situations, our proposed approach can still offer the advantage of yielding accurate marker dosages and effectively taking the uncertainty of estimation into account, especially for polyploids, as it relies on a probabilistic latent class model that exploits most of the information contained in the read depth. Besides, the limit on the number of observed haplotypes at each locus in some populations, e.g. those derived from bi- or multi-parental crosses, can be easily incorporated as prior information in our flexible model (by adjusting the number of latent classes which correspond to possible haplotype alleles) and hence increase the dosage estimation accuracy.

An important point to consider about AcroPoly is the trade-off between the accuracy of the haplotype dosages and the length of the haplotypes, i.e. the number of SNPs included in each haplotype. While the limitation of computational resources usually imposes a restriction on the achievable length, a more important limiting factor is the length of the sequence reads and the heterozygosity rate of the organism, which determine the expected number of SNPs covered by each read. As the length of the haplotypes exceeds the average number of SNPs included in a read, the efficiency of the EM approach will also decrease as each read can only partially be matched to the possible haplotype alleles. Long read sequencing technologies, such as Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (ONT), can produce reads that are tens of kb long, which potentially allows for the estimation of dosage scores for much longer haplotypes in comparison to short read (Illumina) sequencing. To achieve long haplotypes using these technologies, however, the computational difficulties must still be overcome by a quick (*a priori*) exclusion of unlikely haplotype alleles, e.g. by applying a simple filtering step based on read counts or, if possible, by taking the population structure into account.

The accuracy of AcroPoly depends on the validity of its assumptions, notably the Poisson distribution for the number of reads matched to each haplotype. These assumptions might be violated for repetitive regions, as reads from other regions might be incorrectly mapped to the region of interest. Also, the EM method can result in biased estimates of the dosages if the library preparation and sequencing method artificially increase (decrease) the number of reads generated from some of the haplotypes. Alternative distributions, such as the negative binomial or Gamma, can replace the Poisson distribution in such cases, although this can render the implementation and the convergence of the EM algorithm more difficult.

While we applied AcroPoly to assign probabilistic dosages to individuals, its use is not limited to individuals as the EM algorithm can also estimate the frequencies of haplotype

alleles in pools of individuals. This application is especially important for crops that are bred, genetically studied and selected as family pools, such as ryegrass [19]. Genetic associations can be thus tested using the obtained haplotype frequencies as predictors of the desired phenotypes.

Appendix

Proof of Equation 5.6:

$$
\begin{aligned}
\nabla\mathcal{Q}(\boldsymbol{\mu}_t^T;\mathbb{E}[\boldsymbol{k}_t|\boldsymbol{A}_t]) &= (\frac{\partial\mathcal{Q}}{\partial\mu_{1t}},\cdots,\frac{\partial\mathcal{Q}}{\partial\mu_{Nt}}) \\
&\overset{\text{Equation 5.5}}{=} (\frac{\partial\log\big(Pois(\sum_{x=1}^{c_t}a_{x,1};s\mu_{1t})\big)}{\partial\mu_{1t}},\cdots,\frac{\partial\log\big(Pois(\sum_{x=1}^{c_t}a_{x,N};s\mu_{1t})\big)}{\partial\mu_{Nt}}) \\
&= \big(\sum_{x=1}^{c_t}a_{x,1}\frac{\partial\log(s\mu_{1t})}{\partial\mu_{1t}}-\frac{\partial(s\mu_{1t})}{\partial\mu_{1t}},\cdots,\sum_{x=1}^{c_t}a_{x,N}\frac{\partial\log(s\mu_{Nt})}{\partial\mu_{Nt}}-\frac{\partial(s\mu_{Nt})}{\partial\mu_{Nt}}\big) \\
&= \big(\frac{\sum_{x=1}^{c_t}a_{x,1}}{\mu_{1t}}-s,\cdots,\frac{\sum_{x=1}^{c_t}a_{x,N}}{\mu_{Nt}}-s\big) \\
\Longrightarrow \nabla\mathcal{Q}(\boldsymbol{\mu}_t^T;\mathbb{E}[\boldsymbol{k}_t|\boldsymbol{A}_t]) &= \mathbf{0} \text{ if } \frac{\sum_{x=1}^{c_t}a_{x,i}}{\mu_{it}}-s=0 \quad (i=1,\cdots,N) \\
\Longrightarrow \nabla\mathcal{Q}(\boldsymbol{\mu}_t^T;\mathbb{E}[\boldsymbol{k}_t|\boldsymbol{A}_t]) &= \mathbf{0} \text{ if } \boldsymbol{\mu}_t^T=\frac{1}{s}\boldsymbol{J}_{1,c_t}\boldsymbol{A}_t \overset{\text{Equation 5.4}}{=} \frac{1}{s}\mathbb{E}[\boldsymbol{k}_t|\boldsymbol{A}_t] \quad \mathbb{Q}.\mathbb{E}.\mathbb{D}.
\end{aligned}
$$

References

[1] Abecasis, G. R., Cherny, S. S., Cookson, W. O., and Cardon, L. R. (2002). Merlin-rapid analysis of dense genetic maps using sparse gene flow trees. *Nature Genetics*, **30**(1), 97.

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

[3] Ashraf, B. H., Jensen, J., Asp, T., and Janss, L. L. (2014). Association studies using family pools of outcrossing crops based on allele-frequency estimates from DNA sequencing. *Theoretical and Applied Genetics*, **127**(6), 1331–1341.

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

[5] Bartholomew, D. J., Knott, M., and Moustaki, I. (2011). *Latent variable models and factor analysis: A unified approach*, volume 904. John Wiley & Sons.

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

[7] Bourke, P. M., Voorrips, R. E., Kranenburg, T., Jansen, J., Visser, R. G., and Maliepaard, C. (2016). Integrating haplotype-specific linkage maps in tetraploid species using SNP markers. *Theoretical and Applied Genetics*, **129**(11), 2211–2226.

[8] Bourke, P. M., van Geest, G., Voorrips, R. E., Jansen, J., Kranenburg, T., Shahin, A., Visser, R. G. F., Arens, P., Smulders, M. J. M., and Maliepaard, C. (2018a). polymapR-linkage analysis and genetic map construction from F1 populations of outcrossing polyploids. *Bioinformatics*, **34**(20), 3496–3502.

[9] Bourke, P. M., van Geest, G., Voorrips, R. E., Jansen, J., Kranenburg, T., Shahin, A., Visser, R. G., Arens, P., Smulders, M. J., and Maliepaard, C. (2018b). polymapR-linkage analysis and genetic map construction from F1 populations of outcrossing polyploids. *Bioinformatics*, **34**(20), 3496–3502.

[10] Brodie, A., Azaria, J. R., and Ofran, Y. (2016). How far from the SNP may the causative genes be? *Nucleic Acids Research*, **44**(13), 6046–6054.

[11] Browning, S. R. and Browning, B. L. (2007). Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *The American Journal of Human Genetics*, **81**(5), 1084–1097.

[12] Bugawan, T. L., Mirel, D. B., Valdes, A. M., Panelo, A., Pozzilli, P., and Erlich, H. A. (2003). Association and interaction of the IL4R, IL4, and IL13 loci with type 1 diabetes among Filipinos. *The American Journal of Human Genetics*, **72**(6), 1505–1514.

[13] Calus, M. P., Meuwissen, T. H., Windig, J. J., Knol, E. F., Schrooten, C., Vereijken, A. L., and Veerkamp, R. F. (2009). Effects of the number of markers per haplotype and clustering of haplotypes on the accuracy of QTL mapping and prediction of genomic breeding values. *Genetics Selection Evolution*, **41**(1), 11.

[14] Cao, C.-C. and Sun, X. (2014). Accurate estimation of haplotype frequency from pooled sequencing data and cost-effective identification of rare haplotype carriers by overlapping pool sequencing. *Bioinformatics*, **31**(4), 515–522.

[15] Consortium, P. G. S. *et al.* (2011). Genome sequence and analysis of the tuber crop potato. *Nature*, **475**(7355), 189–195.

[16] Das, S. and Vikalo, H. (2015). SDhaP: haplotype assembly for diploids and polyploids via semi-definite programming. *BMC Genomics*, **16**(1), 260.

[17] Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodology)*, pages 1–38.

[18] Excoffier, L. and Slatkin, M. (1995). Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. *Molecular Biology and Evolution*, **12**(5), 921–927.

[19] Fè, D., Cericola, F., Byrne, S., Lenk, I., Ashraf, B. H., Pederson, M. G., Roulund, N., Asp, T., Janss, L., Jensen, C. S., and Jensen, J. (2015). Genomic dissection and prediction of heading date in perennial ryegrass. *BMC Genomics*, **16**(1), 921.

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

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

[22] Gerard, D., Ferrão, L. F. V., Garcia, A. A. F., and Stephens, M. (2018). Genotyping polyploids from messy sequencing data. *Genetics*, **210**(3), 789–807.

[23] Grandke, F., Ranganathan, S., Czech, A., de Haan, J. R., and Metzler, D. (2014). Bioinformatic tools for polyploid crops. *Journal of Agricultural Science and Technology*, **B 4**, 593–601.

[24] Hagenaars, J. A. and McCutcheon, A. L. (2002). *Applied latent class analysis*. Cambridge University Press.

[25] Han, T., Van Eck, H., De Jeu, M., and Jacobsen, E. (1999). Optimization of AFLP fingerprinting of organisms with a large-sized genome: a study on Alstroemeria spp. *Theoretical and Applied Genetics*, **98**(3-4), 465–471.

[26] Han, T.-H., De Jeu, M., Van Eck, H., and Jacobsen, E. (2000). Genetic diversity of Chilean and Brazilian Alstroemeria species assessed by AFLP analysis. *Heredity*, **84**(5), 564.

[27] Hauser, E., Cremer, N., Hein, R., and Deshmukh, H. (2009). Haplotype-based analysis: a summary of GAW16 group 4 analysis. *Genetic Epidemiology*, **33**(S1).

[28] Hayes, B., Chamberlain, A., McPartlan, H., Macleod, I., Sethuraman, L., and Goddard, M. (2007). Accuracy of marker-assisted selection with single markers and marker haplotypes in cattle. *Genetics Research*, **89**(4), 215–220.

[29] Huang, M., Tu, J., and Lu, Z. (2017). Recent advances in experimental whole genome haplotyping methods. *International Journal of Molecular Sciences*, **18**(9), 1944.

[30] Huang, W., Li, L., Myers, J. R., and Marth, G. T. (2011). ART: a next-generation sequencing read simulator. *Bioinformatics*, **28**(4), 593–594.

[31] Huang, X., Feng, Q., Qian, Q., Zhao, Q., Wang, L., Wang, A., Guan, J., Fan, D., Weng, Q., Huang, T., *et al.* (2009). High-throughput genotyping by whole-genome resequencing. *Genome Research*, **19**(6), 1068–1076.

[32] Jannink, J.-L. and Walsh, B. (2002). Association mapping in plant populations. *Quantitative Genetics, Genomics and Plant Breeding*, pages 59–68.

[33] Kessner, D., Turner, T. L., and Novembre, J. (2013). Maximum likelihood estimation of frequencies of known haplotypes from pooled sequence data. *Molecular Biology and Evolution*, **30**(5), 1145–1158.

[34] Kim, S., Park, K., Shin, C., Cho, N. H., Ko, J.-J., Koh, I., and Kwack, K. (2013). Diplotyper: diplotype-based association analysis. *BMC Medical Genomics*, **6**(2), S5.

[35] Kloosterman, B., Abelenda, J. A., Gomez, M. d. M. C., Oortwijn, M., de Boer, J. M., Kowitwanich, K., Horvath, B. M., van Eck, H. J., Smaczniak, C., Prat, S., *et al.* (2013). Naturally occurring allele diversity allows potato cultivation in northern latitudes. *Nature*, **495**(7440), 246–250.

[36] LaFramboise, T. (2009). Single nucleotide polymorphism arrays: a decade of biological, computational and technological advances. *Nucleic Acids Research*, **37**(13), 4181–4193.

[37] Li, B., Ruotti, V., Stewart, R. M., Thomson, J. A., and Dewey, C. N. (2009). RNA-seq gene expression estimation with read mapping uncertainty. *Bioinformatics*, **26**(4), 493–500.

[38] Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv preprint arXiv:1303.3997*.

[39] Li, Y., Willer, C. J., Ding, J., Scheet, P., and Abecasis, G. R. (2010). MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genetic Epidemiology*, **34**(8), 816–834.

[40] Lin, D. and Zeng, D. (2006). Likelihood-based inference on haplotype effects in genetic association studies. *Journal of the American Statistical Association*, **101**(473), 89–104.

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

[42] Long, Q., Jeffares, D. C., Zhang, Q., Ye, K., Nizhynska, V., Ning, Z., Tyler-Smith, C., and Nordborg, M. (2011). PoolHap: inferring haplotype frequencies from pooled samples by next generation sequencing. *PLoS One*, **6**(1), e15292.

[43] Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., McCarthy, M. I., Ramos, E. M., Cardon, L. R., Chakravarti, A., *et al.* (2009). Finding the missing heritability of complex diseases. *Nature*, **461**(7265), 747.

[44] Marees, A. T., De Kluiver, H., Stringer, S., Vorspan, F., Curis, E., Marie-Claire, C., and Derks, E. M. (2018). A tutorial on conducting genome-wide association studies: Quality control and statistical analysis. *International Journal of Methods in Psychiatric Research*, **27**(2), e1608.

[45] Morris, R. W. and Kaplan, N. L. (2002). On the advantage of haplotype analysis in the presence of multiple disease susceptibility alleles. *Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society*, **23**(3), 221–233.

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

[47] Motazedi, E., Maliepaard, C., Finkers, R., and de Ridder, D. (2018b). Family-based haplotype estimation and allele dosage correction for polyploids using short sequence reads. *bioRxiv preprint bioRxiv 318196*.

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

[49] Nielsen, R., Paul, J. S., Albrechtsen, A., and Song, Y. S. (2011). Genotype and SNP calling from next-generation sequencing data. *Nature Reviews Genetics*, **12**(6), 443.

[50] Orozco, G., Hinks, A., Eyre, S., Ke, X., Gibbons, L. J., Bowes, J., Flynn, E., Martin, P., Consortium, W. T. C. C., consortium, Y., *et al.* (2009). Combined effects of three independent SNPs greatly increase the risk estimate for RA at 6q23. *Human Molecular Genetics*, **18**(14), 2693–2699.

[51] Powell, J. E., Visscher, P. M., and Goddard, M. E. (2010). Reconciling the analysis of IBD and IBS in complex trait studies. *Nature Reviews Genetics*, **11**(11), 800.

[52] Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-

wide association studies. *Nature Genetics*, **38**(8), 904.

[53] Ushijima, K., Sassa, H., Dandekar, A. M., Gradziel, T. M., Tao, R., and Hirano, H. (2003). Structural and transcriptional analysis of the self-incompatibility locus of almond: identification of a pollen-expressed f-box gene with haplotype-specific polymorphism. *The Plant Cell*, **15**(3), 771–781.

[54] Williams, A. L., Housman, D. E., Rinard, M. C., and Gifford, D. K. (2010). Rapid haplotype inference for nuclear families. *Genome Biology*, **11**(10), R108.

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

[56] Zhang, J., Liang, F., Dassen, W. R., Doevendans, P. A., and de Gunst, M. (2003). Search for haplotype interactions that influence susceptibility to type 1 diabetes, through use of unphased genotype data. *The American Journal of Human Genetics*, **73**(6), 1385–1401.

[57] Zhang, K., Calabrese, P., Nordborg, M., and Sun, F. (2002). Haplotype block structure and its applications to association studies: power and study designs. *The American Journal of Human Genetics*, **71**(6), 1386–1394.

[58] Zhao, H., Pfeiffer, R., and Gail, M. H. (2003). Haplotype analysis in population genetics and association studies. *Pharmacogenomics*, **4**(2), 171–178.

6

General Discussion

A large part of the work presented in this thesis focused on the development of haplotype estimation algorithms and the evaluation of these algorithms. However, it is also important to investigate how these methods could obtain a wider applicability and which improvements are necessary for this purpose. In this closing chapter, I discuss the impact of emerging sequencing techniques on haplotyping for both diploids and polyploids, as well as the use of haplotype estimation methods in polyploid genetics and genomics, considering autopolyploids as well as allopolyploids. I conclude this chapter by mentioning prospective validation approaches and the significance of the methods developed in this thesis for the foreseeable future.

6.1. Issues and opportunities offered by emerging sequencing techniques

Since the emergence of next generation sequencing technologies in the late nineties, astonishing progress has been observed in the throughput, cost and read-length of sequencing [19]. Third generation sequencing technologies, such as Oxford Nanopore Technologies (ONT) [30] and PacBio [33], now offer sequence reads with lengths ranging from several to hundreds of kilobases. In addition to these long-read technologies, the multiplexed microfluidic DNA preparation method developed by 10X Genomics can be piggybacked on the traditional Illumina sequencing platform to generate short reads linked and separated by uncalled inserts within a single fragment that spans a long genomic distance [50]. Another DNA preparation approach, Hi-C, is capable of linking genomic loci that are distant in the linear genome, but are physically close through the 3D folding of chromatin [28]. The long reads or sequencing fragments obtained by these approaches are more likely to contain several genomic variants compared to short reads, and therefore reveal more information about the haplotypes [23]. However, the same challenges are present for phasing as with short reads, making a thorough evaluation of the estimation methods necessary using various measures of estimation quality, as discussed in Chapter 2.

Chin *et al.* [6] present a haplotype aware assembler for diploid genomes, FALCON Unzip, which is based on the classical three stages for *de novo* assembly: overlap, layout and consensus (OLC). Using the contigs obtained by the flexible primary assembler FALCON, FALCON Unzip looks into different possible phasings between overlapping contigs and chooses those supported the most by the reads. A novel method, FALCON Phase, tries to scaffold interrupted haplotype contigs of FALCON Unzip by mapping Hi-C short-reads to them, so that ultra-long-range haplotypes (>1Mb) can be obtained [26]. However, the success of these approaches depends on the sequencing depth, read length and base calling error, as well as on the rate and type of variation between the haplotypes. In particular, the rather high sequencing error rates can result in spurious haplotypes or interrupt the phasing. Due to higher sequencing error rates, this issue is specially a limiting factor with ONT sequencing [27]. Ghareghani *et al.* [18] present SaaRclust for reliable separation of ONT reads per chromosome before assembly, based on a latent variable model, which has only been evaluated on the human genome.

However, haplotype aware assembly is much more difficult for polyploid genomes. For allopolyploids, the problem can be reduced to diploid assembly by assuming an organism with $2n = kx$ chromosomes to be a diploid organism with $2n = 2(\frac{k}{2}x)$ chromosomes. Substantial differences between the subgenomes of an allopolyploid can make this approach possible, as has been applied to the allotetraploid blueberry, using a combination of 10X Genomics, Illumina and Hi-C data [10].

Recently, a pan-genomic approach has been developed by the NRGene company [31], which makes use of a combination of long and short reads to make a graphical database of all possible contigs and their mutual relationships for several cultivars within the same species. This approach has successfully been applied to the diploid corn [31] and is currently under development for polyploids, notably for potato. Using such a database, the read alignment, variant calling and haplotyping steps can all be integrated and improved for both long and short reads. Such an integrated approach is capable of revealing most of the SNPs, presence/absence variations (PAV) and chromosome rearrangements observed in a query genome with respect to the pan-genome reference [39, 42].

6.2. Use of haplotype estimation methods in polyploid genetics

When it comes to the application of sequence-based haplotyping to study the genetics of polyploid populations, the first question that must be faced in practice is get enough sequencing coverage. We showed in Chapter 2 that the efficiency of single individual haplotyping (SIH) heavily depends on the sequencing depth. In Chapters 3 and 4 we showed that using pedigree information results in higher accuracy at lower sequencing depths. However, a coverage of at least 2× per homologue is still needed for each individual to obtain satisfactory haplotype estimates using population-based methods such as TriPoly and PopPoly. This corresponds to a total coverage of 8× for tetraploids, while in large populations (consisting of a few hundreds or more individuals) the feasible total coverage, limited by time and budget considerations, using high-throughput (and relatively cheap) Illumina sequencing is often lower.

However, the interest is not always in the markers, and hence not in the haplotypes, of a single individual in genetic studies, for example when the aim is to compare frequencies of different alleles in a group of seedlings resistant to a phytopathogen with those in

a susceptible group [32, 44]. Such a strategy has also been applied to case and control groups in genetic epidemiology [21]. Therefore, an alternative strategy to deterministic phasing, as discussed in Chapter 5, can be applied in this situation to estimate the frequencies of haplotype alleles that are theoretically possible over a region of interest. This approach makes the pooling of sequence reads possible within each experimental group and thus compensates for the shallow coverage of each individual. However, it is still worth determining the phasing for a number of individuals to obtain a preliminary set of haplotype alleles that actually occur in the (unstructured) population, so that unlikely haplotype alleles are ruled out *a priori*. This is especially important if haplotypes contain many SNPs, and hence have many theoretically possible alleles, as the computational burden is considerably reduced and the precision of frequency estimates is improved. The PopPoly approach is in particular advantageous to obtain haplotypes that appear in an F1-population with low individual coverages, as it uses all of the population reads to estimate the parental haplotypes with high precision. The PopPoly method can also deliver several likely estimates of the phasing, instead of just the most likely, so that a broader set of haplotypes can be considered for the bulk analysis.

As sequencing costs are plummeting and throughput and precision increase, whole population sequencing is becoming an attractive alternative to the traditional marker genotyping and imputation methods. This can benefit polyploids the most, as linkage analysis and genetic imputation are much more complicated compared to diploids [3, 37, 49]. In sequenced segregating populations, haplotype estimation methods can reveal the inheritance patterns for a set of densely positioned markers with high precision and thus help fine mapping of the traits of interest. This landscape becomes more promising if cheap, high-throughput short-read sequence data can be complemented with long-range sequencing of selected regions or individuals, so that the gaps between haplotype blocks can be bridged and structural variations can be detected [38].

Haplotypes can also be very useful tools for genomic prediction, where the goal is to accurately predict traits of interest using a large set of dense genetic markers [5, 22]. However, errors in haplotype estimation might reduce the prediction accuracy with haplotypes compared to individual markers [35]. Approaches such as AcroPoly (Chapter 5) are capable of providing reliable haplotype scores to be used for prediction, incorporating the estimation uncertainty into probabilistic haplotype scores, for both diploid and polyploid populations.

6.3. Sequence-based phasing in populations

We presented two approaches that combine sequence data with inheritance information coming from the population structure to improve sequence-based haplotyping. The TriPoly method in Chapter 3 compares the phasing estimate of each offspring to the parental estimates according to the inheritance pattern expected in a trio with a prespecified recombination rate. The PopPoly method in Chapter 4 combines all of the reads in an F1-population to estimate the haplotypes of the parents and subsequently chooses the phasing of each offspring from the parental transmissions, assuming complete Mendelian inheritance with no recombination.

The complexity of haplotype estimation algorithms often requires heavy computations for large populations and with high sequencing depths. However, as sequence

reads are assumed independent of each other and the offspring conditionally independent given the parents, calculations are massively parallel and therefore can greatly benefit from GPU-computing, which makes use of powerful and highly-parallel programmable processors called graphics processing units (GPU). GPU-computing has successfully been used to accelerate sequence alignment [29] and has the potential of greatly speeding up haplotyping, rendering the methods scalable to large populations at not much extra cost.

The inheritance information provided by F1-populations can also be used to determine the phasing between haplotype blocks, i.e. to connect phasing interruptions caused by lack of informative overlaps between the reads. For the sake of illustration, consider a situation where in a tetraploid F1-population we have maternal haplotype a in block A, which is distinct from the three other haplotypes in the block, and maternal haplotype b in the closely located block B, which is also distinct. Looking into the offspring phasing estimates of block A and block B, we can conclude that haplotypes a and b are in coupling phase (located on the same chromosome), if they are always or most of the time inherited together. An observation of completely random co-inheritance, i.e. in $\frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$ of the offspring, leads us to conclude that a and b are in repulsion phase (located on different chromosomes). A generalisation of the population-based phasing methods, which currently handle only SNP markers [4, 49], allows systematic tackling of the long-range haplotyping problem in populations, even with short sequencing reads, and thus can be of particular interest for genomic prediction and genomic selection.

Finally, there is definitely a need to extend the methods developed for F1-populations to other types of (partially) outcrossing populations, such as a GWAS panel [14, 41]. In such populations, prior estimation of founder haplotypes and homozygous lines provides some information about the occurring haplotypes and their expected frequencies in the rest of the population, which can be combined with sequence data to construct improved likelihood models for phasing.

6.4. Haplotype estimation for studying genomic variation in allopolyploids

The methods of haplotype estimation discussed in this thesis were developed using the assumptions that for an individual, reads originating from any of the chromosomes can be accurately aligned to a common reference genome and that there is no preferential pairing between the chromosomes. These assumption are often violated in allopolyploid species, which consist of divergent subgenomes. However, the basic concepts underlying the methods can still be applied when regions dissimilar between the subgenomes are flanked by similar regions, e.g. orthologs, that contain SNP markers. For such regions, the flanking markers can be used to distinguish the occurring haplotypes. Knowing these haplotypes helps reveal phylogenetic origins of the subgenomes and is of great value for answering fundamental questions about molecular mechanisms [13, 45, 47].

If a common reference is available for the subgenomes, common SNPs can be called and their phasings can be estimated using the introduced haplotyping methods, even if the used reference is in parts divergent compared to the subgenomes [24]. The estimated SNP phasings can be used to group the reads according to their chromosomal origin. To

achieve this, the reference bases are first replaced with the SNP alleles of each haplotype to obtain a backbone sequence for each chromosome and the reads are aligned to these backbones using algorithms such as BLAST [48]. In this way, the full sequence of each chromosome can be determined over the region of interest by assembling the assigned reads. Structural variations and insertions/deletions can then be detected by comparing the assembled haplotype sequences to the common reference or to each other. Such an approach is especially applicable with long sequence reads, which are likely to contain several common SNPs. For allopolyploids, this approach can also be used to distinguish homologous, i.e. within subgenome, SNPs from homoeologous SNPs, i.e. SNPs that are homozygous within the subgenomes but bear different alleles on each. This method of detecting homologous SNPs can be applied to a broad range of allopolyploids, in contrast to methods such as SWEEP [8] and HAPLOSWEEP [9] that have been developed for self-fertilising allopolyploids.

6.5. Validation of haplotype estimation algorithms

In case the haplotypes of an organism can be set apart by experimental methods and thus separately sequenced [40], the performance of a haplotype estimation algorithm can be assessed with respect to the ground truth. This is, however, often not the case as the available procedures are laborious and costly and not always successful, especially for polyploids [12]. Simulation of polyploid genomes and sequence data is an alternative that also allows evaluating performance under varying heterozygosity rates, ploidy levels, sequencing approaches, read lengths and sequencing depths. Such a comparison of different scenarios is not easily possible with experimental datasets.

However, inferring real-world performance from simulation data is not straightforward, as the genomic complexity of polyploids goes far beyond the density of heterozygous SNPs and the distribution of their dosages. When compared to a common reference, distantly related genomes can host abundant insertions, deletions and chromosomal rearrangements. When primers and library preparation protocols are designed using a rather distant reference, these differences can heavily affect the upstream analysis of the sequence data including read alignment and variant calling. Some of the reads might be aligned to wrong genomic positions and therefore result in spurious variants or wrong dosage estimates. Some of the variant alleles might also be lost, or not amplified, during the DNA preparation step and therefore not detected by variant callers [2, 34].

In Chapter 2, we use an approach that integrates the simulation of polyploid genomes and sequence reads with conventional read alignment and variant calling, hence minimising the effect of the simplifying assumptions necessary for simulation. Using this approach, we show that the results obtained by more simplistic simulations do not always hold and may even be contradicted. For instance, while simplistic simulations show that SDhaP is more accurate than HapTree [11], we show in Chapter 2 that HapTree is the more accurate method with Illumina sequencing. However, our simulation approach is also inevitably based on simplifying assumptions about sequencing coverage and sequence similarity to the reference genome.

However, experimental approaches could be applied to a set of selected organisms and loci, providing a common database for the benchmarking of haplotype estimation methods that use high-throughput sequencing [36]. For example, the traditional Sanger

sequencing approach combined with allele specific PCR [43] or cloning [25] can yield highly accurate haplotypes, although at high cost and low throughput. The substantial cost and labour required for constructing such a database can be shared and provided by a consortium of research centres, which will eventually benefit the whole field [51]. High-throughput long-read and short-read sequence data can also be added to such a database at high depths, providing a standard for the evaluation of haplotyping, as well as genome assembly and alignment methods.

The insights obtained by such experimental haplotypes can also help improve simulation pipelines and genome modelling, so that simulations result in more reliable and more generalisable validations in various situations. The availability of standard data makes non-parametric simulation an option as well, as artificial polyploids can be made by combining available haplotypes and desired sequencing depths can be achieved by downsampling of the available high-throughput sequence data.

6.6. Concluding remarks

With recent advances in molecular techniques and statistical methods, plant genetics and breeding has observed a major shift in many aspects. Polyploidy has so become a centre of attention, with many questions still unanswered about its evolutionary consequences. Modern experimental methods and computational tools enable us to investigate polyploid genomes for finding trait loci and crop improvement [4, 15], among which high-throughput sequencing plays an increasingly important role [17, 20, 46]. Haplotyping is an important part of sequence data analysis in heterozygous polyploids and is tightly linked to read assembly/alignment and variant calling [7, 16, 17].

The first sequence-based haplotype estimation algorithm for polyploids was introduced by Aguiar and Istrail [1], about half a decade ago. While the state of the art has advanced since then, there is still ample room for improvement, especially considering the growth in the amount of available sequence data and the advances in computational resources. The work presented in the current thesis was one of the first attempts towards combining population structure and sequencing data for polyploid haplotyping, and using sequence-based haplotypes for studying phenotypic traits in polyploids. The short sequence reads targeted by the algorithms presented in this thesis will remain the main source of sequence data for large scale populations, at least in the foreseeable future, as these technologies still outperform the emerging long-read technologies in cost, throughput and precision.

With the continuous advances in sequencing, haplotypes are expected to become an essential tool for both genetic analysis and precision breeding in the near future.

References

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

[2] Aird, D., Ross, M. G., Chen, W.-S., Danielsson, M., Fennell, T., Russ, C., Jaffe, D. B., Nusbaum, C., and Gnirke, A. (2011). Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biology*, **12**(2), R18.

[3] Bourke, P. M., Arens, P., Voorrips, R. E., Esselink, G. D., Koning-Boucoiran, C. F., van't Westende, W. P., Santos Leonardo, T., Wissink, P., Zheng, C., Geest, G., *et al.* (2017). Partial preferential chromosome pairing is genotype dependent in tetraploid rose. *The Plant Journal*, **90**(2), 330–343.

[4] Bourke, P. M., Voorrips, R. E., Visser, R. G., and Maliepaard, C. (2018). Tools for genetic studies in experimental populations of polyploids. *Frontiers in Plant Science*, **9**.

[5] Calus, M., De Roos, A., Veerkamp, R., *et al.* (2008). Accuracy of genomic selection using different methods to define haplotypes. *Genetics*, **178**(1), 553–561.

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

[7] Clark, L. V., Lipka, A. E., and Sacks, E. J. (2019). polyRAD: Genotype calling with uncertainty from sequencing data in polyploids and diploids. *G3: Genes, Genomes, Genetics*, pages g3–200913.

[8] Clevenger, J. P. and Ozias-Akins, P. (2015). SWEEP: A tool for filtering high-quality SNPs in polyploid crops. *G3: Genes, Genomes, Genetics*, **5**(9), 1797–1803.

[9] Clevenger, J. P., Korani, W., Ozias-Akins, P., and Jackson, S. (2018). Haplotype-based genotyping in polyploids. *Frontiers in Plant Science*, **9**, 564.

[10] Colle, M., Leisner, C. P., Wai, C. M., Ou, S., Bird, K. A., Wang, J., Wisecaver, J. H., Yocca, A. E., Alger, E. I., Tang, H., *et al.* (2019). Haplotype-phased genome and evolution of phytonutrient pathways of tetraploid blueberry. *GigaScience*.

[11] Das, S. and Vikalo, H. (2015). SDhaP: haplotype assembly for diploids and polyploids via semi-definite programming. *BMC Genomics*, **16**(1), 260.

[12] Doležel, J., Vrána, J., Šafář, J., Bartoš, J., Kubaláková, M., and Šimková, H. (2012). Chromosomes in the flow to simplify genome analysis. *Functional & Integrative Genomics*, **12**(3), 397–416.

[13] Edger, P. P., Poorten, T. J., VanBuren, R., Hardigan, M. A., Colle, M., McKain, M. R., Smith, R. D., Teresi, S. J., Nelson, A. D., Wai, C. M., *et al.* (2019). Origin and evolution of the octoploid strawberry genome. *Nature Genetics*, page 1.

[14] Ferrao, L. F. V., Benevenuto, J., Oliveira, I. d. B., Cellon, C., Olmstead, J., Kirst, M., Resende Jr, M. F., and Munoz, P. R. (2018). Insights into the genetic basis of blueberry fruit-related traits using diploid and polyploid models in a GWAS context. *Frontiers in Ecology and Evolution*, **6**, 107.

[15] Gabur, I., Chawla, H. S., Liu, X., Kumar, V., Faure, S., von Tiedemann, A., Jestin, C., Dryzska, E., Volkmann, S., Breuer, F., *et al.* (2018). Finding invisible quantitative trait loci with missing data. *Plant Biotechnology Journal*, **16**(12), 2102–2112.

[16] Garrison, E. and Marth, G. (2012). Haplotype-based variant detection from short-read sequencing. *arXiv preprint arXiv:1207.3907 [q-bio.GN]*.

[17] Gerard, D., Ferrão, L. F. V., Garcia, A. A. F., and Stephens, M. (2018). Genotyping polyploids from messy sequencing data. *Genetics*, **210**(3), 789–807.

[18] Ghareghani, M., Porubskỳ, D., Sanders, A. D., Meiers, S., Eichler, E. E., Korbel, J. O., and Marschall, T. (2018). Strand-seq enables reliable separation of long reads by chromosome via expectation maximization. *Bioinformatics*, **34**(13), i115–i123.

[19] Goodwin, S., McPherson, J. D., and McCombie, W. R. (2016). Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics*, **17**(6), 333.

[20] Harper, A. L., Trick, M., Higgins, J., Fraser, F., Clissold, L., Wells, R., Hattori, C., Werner, P., and Bancroft, I. (2012). Associative transcriptomics of traits in the polyploid crop species Brassica napus. *Nature Biotechnology*, **30**(8), 798.

[21] Jeleñ, A. M., Salagacka-Kubiak, A., Ropicka, K., Zebrowska-Nawrocka, M. K., Galecki, P., Talarowska, M., Mirowski, M., and Balcerczak, E. I. (2019). Selected ABCB1 single nucleotide polymorphisms and its haplotype–connection with development of depression and treatment efficacy. *Int J Hum Genet*, **18**(4), 1–11.

[22] Jiang, Y., Schmidt, R. H., and Reif, J. C. (2018). Haplotype-based genome-wide prediction models exploit local epistatic interactions among markers. *G3: Genes, Genomes, Genetics*, pages g3–300548.

[23] Jiao, W.-B. and Schneeberger, K. (2017). The impact of third generation genomic technologies on plant genome assembly. *Current Opinion in Plant Biology*, **36**, 64–70.

[24] Kamneva, O. K., Syring, J., Liston, A., and Rosenberg, N. A. (2017). Evaluating allopolyploid origins in strawberries (Fragaria) using haplotypes generated from target capture sequencing. *BMC Evolutionary Biology*, **17**(1), 180.

[25] Kim, J.-H., Leem, S.-H., Sunwoo, Y., and Kouprina, N. (2003). Separation of long-range human TERT gene haplotypes by transformation-associated recombination cloning in yeast. *Oncogene*, **22**(16), 2452.

[26] Kronenberg, Z. N., Hall, R. J., Hiendleder, S., Smith, T. P., Sullivan, S. T., Williams, J. L., and Kingan, S. B. (2018). FALCON-Phase: Integrating PacBio and Hi-C data for phased diploid genomes. *bioRxiv*, page bioRxic preprint bioRxiv 327064.

[27] Laver, T. W., Caswell, R. C., Moore, K. A., Poschmann, J., Johnson, M. B., Owens, M. M., Ellard, S., Paszkiewicz, K. H., and Weedon, M. N. (2016). Pitfalls of haplotype phasing from amplicon-based long-read sequencing. *Scientific Reports*, **6**, 21746.

[28] Lieberman-Aiden, E., Van Berkum, N. L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B. R., Sabo, P. J., Dorschner, M. O., *et al.* (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, **326**(5950), 289–293.

[29] Liu, C.-M., Wong, T., Wu, E., Luo, R., Yiu, S.-M., Li, Y., Wang, B., Yu, C., Chu, X., Zhao, K., *et al.* (2012). SOAP3: ultra-fast GPU-based parallel alignment tool for short reads. *Bioinformatics*, **28**(6), 878–879.

[30] Loose, M., Malla, S., and Stout, M. (2016). Real-time selective sequencing using nanopore technology. *Nature Methods*, **13**(9), 751.

[31] Lu, F., Romay, M. C., Glaubitz, J. C., Bradbury, P. J., Elshire, R. J., Wang, T., Li, Y.,

Li, Y., Semagn, K., Zhang, X., *et al.* (2015). High-resolution genetic mapping of maize pan-genome sequence anchors. *Nature Communications*, **6**, 6914.

[32] Michelmore, R. W., Paran, I., and Kesseli, R. (1991). Identification of markers linked to disease-resistance genes by bulked segregant analysis: a rapid method to detect markers in specific genomic regions by using segregating populations. *Proceedings of the National Academy of Sciences*, **88**(21), 9828–9832.

[33] Quail, M. A., Smith, M., Coupland, P., Otto, T. D., Harris, S. R., Connor, T. R., Bertoni, A., Swerdlow, H. P., and Gu, Y. (2012). A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics*, **13**(1), 341.

[34] Ross, M. G., Russ, C., Costello, M., Hollinger, A., Lennon, N. J., Hegarty, R., Nusbaum, C., and Jaffe, D. B. (2013). Characterizing and measuring bias in sequence data. *Genome Biology*, **14**(5), R51.

[35] Solberg, T., Sonesson, A., Woolliams, J., and Meuwissen, T. (2008). Genomic selection using different marker types and densities. *Journal of Animal Science*, **86**(10), 2447–2454.

[36] Srivastava, K., Wollenberg, K. R., and Flegel, W. A. (2019). The phylogeny of 48 alleles, experimentally verified at 21 kb, and its application to clinical allele detection. *Journal of Translational Medicine*, **17**(1), 43.

[37] Su, S.-Y., White, J., Balding, D. J., and Coin, L. J. (2008). Inference of haplotypic phase and missing genotypes in polyploid organisms and variable copy number genomic regions. *BMC Bioinformatics*, **9**(1), 513.

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

[39] Sun, C., Hu, Z., Zheng, T., Lu, K., Zhao, Y., Wang, W., Shi, J., Wang, C., Lu, J., Zhang, D., *et al.* (2016). RPAN: rice pan-genome browser for 3000 rice genomes. *Nucleic Acids Research*, **45**(2), 597–605.

[40] Tu, J., Lu, N., Duan, M., Ju, A., Sun, X., and Lu, Z. (2016). Comparison of the experimental methods in haplotype sequencing via next generation sequencing. *Quantitative Biology*, **4**(2), 106–114.

[41] Uitdewilligen, J. G., Wolters, A.-M. A., D'hoop, B. B., Borm, T. J., Visser, R. G., and van Eck, H. J. (2013). A Next-Generation Sequencing Method for Genotyping-by-Sequencing of Highly Heterozygous Autotetraploid Potato. *PLoS One*, **8**(5), e62355.

[42] Valenzuela, D., Norri, T., Välimäki, N., Pitkänen, E., and Mäkinen, V. (2018). Towards pan-genome read alignment to improve variation calling. *BMC Genomics*, **19**(2), 87.

[43] Wang, L. and Xiao, P. (2017). Haplotype-contained PCR products analysis by sequencing with selective restriction of primer extension. *BioMed research international*, **2017**.

[44] Wilcox, P. L., Amerson, H. V., Kuhlman, E. G., Liu, B.-H., O'Malley, D. M., and Sederoff, R. R. (1996). Detection of a major gene for resistance to fusiform rust disease in loblolly pine by genomic mapping. *Proceedigs of the National Academy of Sciences*, **93**(9), 3859–3864.

[45] Yang, J., Moeinzadeh, M.-H., Kuhl, H., Helmuth, J., Xiao, P., Haas, S., Liu, G., Zheng,

J., Sun, Z., Fan, W., *et al.* (2017). Haplotype-resolved sweet potato genome traces back its hexaploidization history. *Nature Plants*, **3**(9), 696.

[46] Yang, X., Todd, J., Arundale, R., Binder, J. B., Luo, Z., Islam, M. S., Sood, S., and Wang, J. (2019). Identifying loci controlling fiber composition in polyploid sugarcane (Saccharum spp.) through genome-wide association study. *Industrial Crops and Products*, **130**, 598–605.

[47] Zhang, J., Esselink, G., Che, D., Fougère-Danezan, M., Arens, P., and Smulders, M. (2013). The diploid origins of allopolyploid rose species studied using single nucleotide polymorphism haplotypes flanking a microsatellite repeat. *The Journal of Horticultural Science and Biotechnology*, **88**(1), 85–92.

[48] Zhang, Z., Schwartz, S., Wagner, L., and Miller, W. (2000). A greedy algorithm for aligning DNA sequences. *Journal of Computational Biology*, **7**(1-2), 203–214.

[49] Zheng, C., Voorrips, R. E., Jansen, J., Hackett, C. A., Ho, J., and Bink, M. C. (2016a). Probabilistic multilocus haplotype reconstruction in outcrossing tetraploids. *Genetics*, **203**(1), 119–131.

[50] Zheng, G. X., Lau, B. T., Schnall-Levin, M., Jarosz, M., Bell, J. M., Hindson, C. M., Kyriazopoulou-Panagiotopoulou, S., Masquelier, D. A., Merrill, L., Terry, J. M., *et al.* (2016b). Haplotyping germline and cancer genomes with high-throughput linked-read sequencing. *Nature biotechnology*, **34**(3), 303.

[51] Zook, J. M., Catoe, D., McDaniel, J., Vang, L., Spies, N., Sidow, A., Weng, Z., Liu, Y., Mason, C. E., Alexander, N., *et al.* (2016). Extensive sequencing of seven human genomes to characterize benchmark reference materials. *Scientific Data*, **3**, 160025.

Summary

Haplotypes are sequences of ordered genomic variants over the same chromosome. Current sequencing technologies enable us to read the DNA and detect genomic variants, but cannot distinguish between the copies of the genome in diploids, one inherited from each parent. Therefore, it is not clear which alleles are found on the same chromosome. To detect inheritance patterns in populations, it is necessary to know the haplotypes, as alleles that are in linkage over the same chromosome tend to be inherited together. Besides, the allele effects sometimes depend on the alleles being located on the same copy of the genome. Mathematical optimisation algorithms have therefore been developed to indirectly estimate haplotypes by looking into overlaps between the sequence reads of an individual, as each sequencing read that contains more than two variation positions is representative of its haplotype of origin. However, such algorithms have to deal with sequencing errors and random variations in the counts of reads observed from each haplotype. Polyploid organisms possess more than two copies of their core genome and therefore contain k>2 haplotypes for each set of variation positions. Polyploidy occurs often within the plant kingdom, among others in important corps such as potato (k=4) and wheat (k=6). Haplotype estimation is much more difficult in polyploids compared to diploids and estimation algorithms are more prone to yield non-existing haplotypes.

Chapter 1 gives an overview of the use of haplotypes in plant genetics and breeding, and provides a thorough introduction to polyploidy and its origins. Basic mathematical concepts are also discussed, necessary for developing haplotype estimation algorithms.

In **Chapter 2**, we develop a simulation pipeline for polyploid genomes and introduce measures to assess the accuracy of polyploid haplotype estimation algorithms. The pipeline and the measures allow to investigate how haplotype estimation is influenced by the read length, library preparation and sequencing technology, heterozygosity rate and the ploidy level. Finally, the pipeline is used to evaluate and compare several state-of-the-art haplotyping algorithms. In **Chapter 3** and **Chapter 4**, two Bayesian algorithms are introduced to infer haplotypes in trios (TriPoly) and in families with several to many offspring (PopPoly). The Bayesian framework incorporates both inheritance information and partial haplotype information within the reads, and can therefore result in precise estimates. Missing alleles can also be imputed using this approach, as is done by PopPoly. We extend the simulation pipeline of **Chapter 1** to simulate parental crossing and show that both TriPoly and PopPoly significantly improve the quality of haplotyping in the offspring compared to single individual methods. In **Chapter 5** we propose an expectation-maximisation (EM) algorithm, AcroPoly, to infer the allele dosages of multi-SNP haplotype markers directly from the sequencing depth, and to associate these markers with traits.

The thesis is concluded in **Chapter 6** by discussing potential applications of the developed methods, as well as opportunities offered by emerging sequencing technologies for improved determination of haplotypes.

Samenvatting

Haplotypen zijn sequenties van geordende genomische varianten op hetzelfde chromosoom. Hoewel huidige sequencingtechnologieën ons in staat stellen om het DNA te lezen en daarin genomische varianten te ontdekken, kunnen ze geen onderscheid maken tussen de kopieën van het genoom in diploïden, waarvan één is geërfd per ouder. Daarom is het niet duidelijk welke allelen zich op hetzelfde chromosoom bevinden. Voor het detecteren van overervingspatronen in populaties is het noodzakelijk de haplotypen te kennen, omdat allelen die op hetzelfde chromosoom met elkaar verbonden zijn in het algemeen samen overerfd worden. Bovendien hangen de effecten van allelen er soms van af of ze zich op dezelfde kopie van het genoom bevinden. Daarom zijn er wiskundige optimalisatiealgoritmen ontwikkeld die haplotypes indirect proberen te schatten, door te kijken naar overlap tussen de DNA reads van een individu. Deze aanpakken zijn gebaseerd op het feite dat elke gelezen sequentie die twee of meer variatieposities bevat, het haplotype dat bij die sequentie hoort vertegenwoordigt. Zulke algoritmen hebben echter te maken met sequencingfouten en toevallige afwijkingen in het aantal waargenomen reads per haplotype. Polyploïde organismen bezitten meer dan twee kopieën van hun kerngenoom en bevatten daarom k>2 haplotypen per set variatieposities. Polyploïdie komt vaak voor binnen het plantenrijk, onder andere in belangrijke gewassen zoals aardappel ($k = 4$) en tarwe ($k = 6$). In vergelijking met diploïden is haplotypeschatting veel moeilijker in polyploïden en geven schattingsalgoritmen vaker niet-bestaande haplotypen.

Hoofdstuk 1 geeft een overzicht van het gebruik van haplotypen in plantengenetica en veredeling. Dit hoofdstuk biedt ook een grondige inleiding tot polyploïdie en de oorsprong ervan. Daarnaast worden de basale wiskundige begrippen besproken die noodzakelijk zijn voor het ontwikkelen van haplotypeschattingsalgoritmen.

In **Hoofdstuk 2** ontwikkelen we een simulatiepijplijn voor polyploïde genomen en stellen we maten voor om de nauwkeurigheid van polyploïde haplotypeschattingsalgoritmen te beoordelen. De pijplijn en de voorgestelde maten maken het mogelijk om te onderzoeken hoe de haplotypenschatting wordt beïnvloed door de lengte van de gelezen sequenties, het maken van de sequentie-library en sequencingtechnologie, de frequentie van heterozygositeit en het ploidieniveau. Ten slotte wordt de pijplijn gebruikt om verschillende huidige haplotypeschattingsalgoritmen te evalueren en te vergelijken. In **Hoofdstuk 3** en **Hoofdstuk 4** worden twee Bayesiaanse algoritmen geïntroduceerd om haplotypes af te leiden in trio's (TriPoly) en in families met meerdere tot vele nakomelingen (PopPoly). Het Bayesiaanse kader maakt het mogelijk de overervingspatronen en de gedeeltelijke informatie over de haplotypen in de reads te combineren, resulterend in preciezere schattingen van de haplotypen. Bovendien kunnen op deze manier ontbrekende varianten worden afgeleid, zoals door PopPoly. We breiden de simulatiepijplijn van **Hoofdstuk 1** uit om de kruising tussen de ouders te simuleren en laten zien dat zowel TriPoly als PopPoly leiden tot aanzienlijk betere schattingen in

de nakomelingen in vergelijking met de methoden die enkele individuen beschouwen. In **Hoofdstuk 5** stellen we een expectation-maximization (EM) algoritme voor, AcroPoly, om de alleldoseringen van multi-SNP haplotypemerkers direct af te leiden uit de sequencingdiepte. We gebruiken deze merkers in een populatie om associaties met fenotypische eigenschappen te detecteren.

Het proefschrift wordt afgesloten in **Hoofdstuk 6**, waarin potentiële toepassingen van de ontwikkelde methoden besproken worden, evenals kansen die opkomende sequencingtechnologieën bieden voor een betere bepaling van de haplotypen.

Epilogue

The study of inheritance, in its foundation, was to find out relations between observable traits and the kinship. This culminated in the experiments of Gregor Mendel (1822-1844) and his well known laws of inheritance. Today, we know much more about gene expression, genomic recombination and the 3D structures of the life molecules. The technology has so evolved that the whole DNA sequence, albeit blurry and fragmented, can be measured within a single cell. The organization of the eukaryotic DNA in (partially) homologous haplotypes is an important feature which must be known in order to fully understand the regulatory processes and interactions between genes.

However, making sense out of the immense, fragmentary and noisy data, ranging from the DNA to RNA and amino-acid sequences and their 3D structures, requires mathematical simplification, optimization and approximation. As Bertrand Russel (1872-1970) once put into words, attention must be paid not to confound the mathematical properties that we can discover with the underlying ontology.

The current thesis has been the result of a joint effort, and it should be acknowledged as such. First and foremost, I need to thank my promotor, *Dick de Ridder*, and my co-promotors *Chris Maliepaard* and *Richard Finkers*. The amount of support and inspiration that I have received from them has been beyond my expectation. The combination of your knowledge and expertise in multiple disciplines gave this thesis its unique character and way of handling the subject by combining inheritance information with sequencing and focusing on applicability. What you taught me I will carry with myself in the rest of my career and life. Dick, you are more than a supervisor to your students, someone who manages and cares. Chris, your modesty, scientific rigour and enthusiasm will always be a model to me. Richard, you are as kind as you are serious and bright. Thank you all!

The research presented in this thesis was conducted under the auspices of The Graduate School Experimental Plant Sciences (EPS), Wageningen. I would like to thank EPS for the generous financial and educational support. I especially would like to thank the past and current PhD program coordinators of EPS: *Douwe Zuidema* and *Susan Urbanus* for their help and advice all through my PhD.

I would like to thank *Richard Visser*, the head of WUR Plant Breeding, for his support of my work, and of polyploid research in general, as well as for being involved in Chapter 4. My gratitude goes also to my previous teachers at the Erasmus Medical Centre (EMC), Rotterdam, *Henning Tiemeier* and *Emmanuel Lesaffre,* for encouraging and supporting me to start a PhD in the first place. What I learned from them appears all over this thesis.

I express my sincere gratitude to the TKI polyploid project initiative and all involved therein, for the fruitful gatherings, workshops and discussions that also resulted in the collaborations appearing in this thesis. In particular, I would like to thank *Arwa Shahin*, without whom Chapter 5 would not have been finished, *Paul Arens, Mohammad Ghanbari, Jan de Boer, Annemarie Castricum, Julie Ho, Nick de Vetten* and *Heleen Bastiaanssen*. I would like also to thank *Samantha Baldwin, Susan Thomson* and *Tim Millar* from the

New Zealand's Plant and Food Research for collaboration in Chapter 2 and for their inspiring visit to WUR.

Finally, my thanks go to all of my colleagues at WUR Plant Breeding and Bioinformatics. I would like to especially mention the secretaries of the two departments: *Maria Augustijn*, *Marie-Jose van Iersel* and *Nicole Trefflich-Luit*, who make life much easier for everyone. As the names are too many and the space is limited, I have to mention only a handful, which of course does not undermine those who do not appear but have contributed to this thesis. Sincere thanks to: *Mas Muniroh*, *Miguel Correa Marrero*, *Peter Bourke*, *Roeland Voorrips*, *Rens Holmer*, *Siavash Sheikhizadeh Anari*, *Mehmet Akdel*, *Herman van Eck*, *Johan Willemsen*, *Charlotte Prodhomme*, *Vittorio Tracanna*, *Raul Wijfjes*, *Naser Askari*, *Aalt-jan van Dijk*, *Sandra Smit*, *Harm Nijveen*, *Agata Gulisano*, *Geert van Geest*, *Marnix Medema*, *Giorgio Tumino*, *Michela Appiano*, *Samin Hosseini Farhangi*, *Behzad Rahsidi*, *Kim Magnée*, *Danny Esselink*, *Eric van de Weg*, *Yanlin Liao*, *Michiel Klaassen*, *Eric Schranz* and *Rene Smulders*. It seems fit to me to also thank my former colleagues at the Department of Biostatistics, EMC, especially *Kazem Nasserinejad*, *Nicole Erler*, *Nahid Mostafavi*, *Eline van Gent-Geertsema*, *Elrozy Andrinopoulou* and *Dimitris Rizopoulos*. Thank you all for being there!

Education Statement of the Graduate School

Experimental Plant Sciences

Issued to: Ehsan Motazedi
Date: 07 November 2019
Group: Bioinformatics & Plant Breeding
University: Wageningen University & Research

1) Start-Up Phase		date	cp
▶	**First presentation of your project**		
	Haplotype estimation in polyploids using next generation sequencing and genetics	18 April, 2016	1,5
▶	**Writing or rewriting a project proposal**		
▶	**Writing a review or book chapter**		
▶	**MSc courses**		
	Principles of Plant Breeding	25 Mar, 2015	3,0
	Plant Breeding: Basic Concepts and their Applications	27 Mar, 2015	3,0
	Subtotal Start-Up Phase		7,5

2) Scientific Exposure		date	cp
▶	**EPS PhD student days**		
	EPS PhD student days	28-29 Jan, 2016	0,6
	EPS PhD student days	9-10 Feb, 2017	0,6
▶	**EPS theme symposia**		
	Theme 4 'Genome Biology', UvA	15 Dec, 2015	0,3
	Theme 4 'Genome Biology', WUR	16 Dec, 2016	0,3
	Theme 4 'Genome Biology', UvA	25 Sep, 2018	0,3
▶	**Lunteren Days and other national platforms**		
	Experimental Plant Sciences Meeting, Lunteren	13-14 Apr, 2015	0,6
	Bioinformatics & Systems Biology conference, Lunteren	20-21 May, 2015	0,6
	Experimental Plant Sciences Meeting, Lunteren	11-12 Apr, 2016	0,6
	Bioinformatics & Systems Biology conference, Lunteren	19-20 Apr, 2016	0,6
	Bioinformatics & Systems Biology conference, Lunteren	4-5 Apr, 2017	0,6
	Experimental Plant Sciences Meeting, Lunteren	10-11 Apr, 2017	0,6
	Experimental Plant Sciences Meeting, Lunteren	9-10 Apr, 2018	0,6
	Bioinformatics & Systems Biology conference, Lunteren	15-16 May, 2018	0,6
▶	**Seminars (series), workshops and symposia**		
	Polyploid Software Workshop III, WUR	12-13 Dec, 2016	0,5
	WURomics Symposium	15 Dec, 2016	0,3
▶	**Seminar plus**		
▶	**International symposia and congresses**		
	European Mathematical Genetics Meeting 2015, Brest, France	16-17 Apr, 2015	0,6
	EUCARPIA Biometrics Meeting 2015, Wageningen, The Netherlands	9-11 Sep, 2015	0,9
	European Mathematical Genetics Meeting 2016, Newcastle upon Tyne, UK	11-12 May, 2016	0,6
	European Conference on Computational Biology 2016, The Hague, The Netherlands	5-7 Sep, 2016	0,9
	EUCARPIA Biometrics Meeting 2018, Ghent, Belgium	3-5 Sep, 2018	0,9
▶	**Presentations**		
	Poster: 'Haplotype Assembly in Polyploids using NGS data, Challenges and Opportunities', Experimental Plant Sciences Meeting, Lunteren	13-14 April, 2015	1,0
	Oral: 'Exploiting next generation sequencing to solve the haplotyping puzzle in polyploids: a simulation study', Experimental Plant Sciences Meeting, Lunteren	12 April, 2016	1,0
	Poster: 'An Evaluation of Haplotyping Algorithms for Polyploids', EMGM 2016, Newcastle Upon Tyne, UK	12 May, 2016	1,0
	Poster: 'TriPoly: a SNP-phasing method for polyploid F1-populations using NGS data', Bioinformatics & Systems Biology conference, Lunteren	4 Apr, 2017	1,0
	Poster and flash presentation: 'Family based haplotype estimation using DNA sequence reads', EUCARPIA 2018, Ghent, Belgium	3 Sep, 2018	1,0
	Oral: 'Family based haplotype estimation using DNA sequence reads', Bioinformatics & Systems Biology conference, Lunteren	15 May, 2018	1,0
	Oral: 'Family based haplotype estimation using DNA sequence reads in polyploid crops', EPS Theme 4 Symposium, UvA	25 Sep, 2018	1,0
▶	**IAB interview**		
▶	**Excursions**		
	Subtotal Scientific Exposure		18,6

3) In-Depth Studies		date	cp
▶	**Advanced scientific courses & workshops**		
	Genome Assembly, WUR	28-29 Apr, 2015	0,6
	Python course for plant breeding, WUR	17-19 May, 2016	0,9
	RNA-seq data analysis, LUMC	26-28 Sep, 2017	0,9
	Optimisation Techniques in Bioinformatics and Systems Biology	12-16 Feb, 2017	3,0
	Statistical analysis for new phenotyping techniques	25-27 Jun, 2018	0,9
▶	**Journal club**		
	Bioinformatics Journal Club, WUR	2015-2018	1,5
	Quantitative methods in plant breeding, WUR	2016-2018	1,5
▶	**Individual research training**		
	Subtotal In-Depth Studies		9,3

4) Personal Development		date	cp
▶	**General skill training courses**		
	EPS Introduction Course	22 Sep, 2015	0,2
	PhD Competence Assessment	6 April, 2016	0,3
	WGS PhD Workshop Carousel	8 April, 2016	0,3

	Interpersonal Communication for PhD candidates	24-25 Nov, 2016	0,6
	Brain Training	8 Feb, 2017	0,3
	Project & Time Management	18 Jan - 1 Mar, 2017	1,5
▶	**Organisation of meetings, PhD courses or outreach activities**		
	Bioinformatics annual group retreat	12-13 Oct, 2015	1,5
	Polyploid Software Workshop IV, WUR	26 Jun, 2018	1,5
▶	**Membership of EPS PhD Council**		
		Subtotal Personal Development	6,2

TOTAL NUMBER OF CREDIT POINTS*	**41,6**

Herewith the Graduate School declares that the PhD candidate has complied with the educational requirements set by the Educational Committee of EPS with a minimum total of 30 ECTS credits.

** A credit represents a normative study load of 28 hours of study.*

List of publications

1. **Motazedi, E.**, Maliepaard, C., Finkers, R., Visser, R. G. F., and de Ridder, D. (2019). Family-based haplotype estimation and allele dosage correction for polyploids using short sequence reads. *Frontiers in Genetics*, **10**(335), 1–12

2. Kooijmans, H., Post, M., **Motazedi, E.**, Spijkerman, D., Bongers-Janssen, H., Stam, H., and Bussman, H. (2019). Exercise self-efficacy is weakly related to engagement in physical activity in persons with long-standing spinal cord injury. *Disability and Rehabilitation*, pages 1–7

3. **Motazedi, E.**, de Ridder, D., Finkers, R., Baldwin, S., Thomson, S., Monaghan, K., and Maliepaard, C. (2018). TriPoly: haplotype estimation for polyploids using sequencing data of related individuals. *Bioinformatics*, **34**(22), 38643872

4. **Motazedi, E.**, Finkers, R., Maliepaard, C., and de Ridder, D. (2017). Exploiting next-generation sequencing to solve the haplotyping puzzle in polyploids: a simulation study. *Briefings in bioinformatics*, **19**(3), 387–403

5. Pappa, I., St Pourcain, B., Benke, K., Cavadino, A., Hakulinen, C., Nivard, M. G., Nolte, I. M., Tiesler, C. M., Bakermans-Kranenburg, M. J., Davies, G. E., Evans, D. M., Geoffroy, M.-C., Grallert, H., Groen-Blokhuis, M. M., Hudziak, J. J., Kemp, J. P., Keltikangas-Jrvinen, L., McMahon, G., Mileva-Seitz, V. R., **Motazedi, E.**, Power, C., Raitakari, O. T., Ring, S. M., Rivadeneira, F., Rodriguez, A., Scheet, P. A., Seppl, I., Snieder, H., Standl, M., Thiering, E., Timpson, N. J., Veenstra, R., Velders, F. P., Whitehouse, A. J., Smith, G. D., Heinrich, J., Hypponen, E., Lehtimki, T., Middeldorp, C. M., Oldehinkel, A. J., Pennell, C. E., Boomsma, D. I., and Tiemeier, H. (2016). A genome-wide approach to children's aggressive behavior: the EAGLE consortium. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*, **171**(5), 562–572

6. Bais, B., Kubat, B., **Motazedi, E.**, and Verdijk, R. M. (2015). β-Amyloid precursor protein and ubiquitin immunohistochemistry aid in the evaluation of infant autopsy eyes with abusive head trauma. *American Journal of Ophthalmology*, **160**(6), 1285–1295

7. Van Mil, N. H., Steegers-Theunissen, R. P., **Motazedi, E.**, Jansen, P. W., Jaddoe, V. W., Steegers, E. A., Verhulst, F. C., and Tiemeier, H. (2015). Low and high birth weight and the risk of child attention problems. *The Journal of Pediatrics*, **166**(4), 862–869

The research described in this thesis was financially supported by the Graduate School Experimental Plant Sciences (EPS), Wageningen University & Research.

Cover design by Natalya Nikolayevna, Ehsan Motazedi

Printed by Ipskamp drukkers, Enschede