Digital Comprehensive Summaries of Uppsala Dissertations
from the Faculty of Science and Technology 2354

A computational and statistical framework for cost-effective genotyping combining pooling and imputation

CAMILLE CLOUARD

ACTA UNIVERSITATIS
UPSALIENSIS
2024

ISSN 1651-6214
ISBN 978-91-513-2006-9
urn:nbn:se:uu:diva-519887

UPPSALA
UNIVERSITET

Dissertation presented at Uppsala University to be publicly examined in 101195 (Heinz-Otto Kreiss), Ångströmlaboratoriet, Lägerhyddsvägen 1, hus 10, Uppsala, Friday, 8 March 2024 at 10:15 for the degree of Doctor of Philosophy. The examination will be conducted in English. Faculty examiner: Prof. Christine Baes (Department of Animal Biosciences, University of Guelph).

Abstract

Clouard, C. 2024. A computational and statistical framework for cost-effective genotyping combining pooling and imputation. *Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology* 2354. 81 pp. Uppsala: Acta Universitatis Upsaliensis. ISBN 978-91-513-2006-9.

The information conveyed by genetic markers, such as single nucleotide polymorphisms (SNPs), has been widely used in biomedical research to study human diseases and is increasingly valued in agriculture for genomic selection purposes. Specific markers can be identified as a genetic signature that correlates with certain characteristics in a living organism, e.g. a susceptibility to disease or high-yield traits. Capturing these signatures with sufficient statistical power often requires large volumes of data, with thousands of samples to be analysed and potentially millions of genetic markers to be screened. Relevant effects are particularly delicate to detect when the genetic variations involved occur at low frequencies.

The cost of producing such marker genotype data is therefore a critical part of the analysis. Despite recent technological advances, production costs can still be prohibitive on a large scale and genotype imputation strategies have been developed to address this issue. Genotype imputation methods have been extensively studied in human data and, to a lesser extent, in crop and animal species. A recognised weakness of imputation methods is their lower accuracy in predicting the genotypes for rare variants, whereas those can be highly informative in association studies and improve the accuracy of genomic selection. In this respect, pooling strategies can be well suited to complement imputation, as pooling is efficient at capturing the low-frequency items in a population. Pooling also reduces the number of genotyping tests required, making its use in combination with imputation a cost-effective compromise between accurate but expensive high-density genotyping of each sample individually and stand-alone imputation. However, due to the nature of genotype data and the limitations of genotype testing techniques, decoding pooled genotypes into unique data resolutions is challenging.

In this work, we study the characteristics of decoded genotype data from pooled observations with a specific pooling scheme using the examples of a human cohort and a population of inbred wheat lines. We propose different inference strategies to reconstruct the genotypes before devising them as input to imputation, and we reflect on how the reconstructed distributions affect the results of imputation methods such as tree-based haplotype clustering or coalescent models.

Camille Clouard, Department of Information Technology, Division of Scientific Computing, Box 337, Uppsala University, SE-751 05 Uppsala, Sweden.

ISSN 1651-6214
ISBN 978-91-513-2006-9
URN urn:nbn:se:uu:diva-519887 (http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-519887)

To anyone who likes to question
the whys and hows of things.

List of abbreviations

1KGP:	1000 Genomes Project
AAF:	Alternate Allele Frequency
bp:	base pair
DNA:	DeoxyriboNucleic Acid
EM:	Expectation-Maximisation
GBS:	Genotyping By Sequencing
GS:	Genomic Selection
GWAS:	Genome-Wide Association Studies
HMM:	Hidden Markov Model
LD :	Linkage Disequilibrium
MAF:	Minor Allele Frequency
MAGIC:	Multiparent Advanced Generation InterCross
MCMC:	Markov Chain Monte Carlo
ML:	Maximum Likelihood
MML:	Maximum Marginal Likelihood
M(C)AR:	Missing (Completely) At Random
MNAR:	Missing Not At Random
NGS:	Next-Generation Sequencing
NGT:	Nonadaptive Group Testing
NORB:	Nonadaptive Overlapping Repeated Block
ONT:	Oxford Nanopore Technologies
PCR:	Polymerase Chain Reaction
QTL:	Quantitative Trait Loci
SNP:	Single Nucleotide Polymorphism
STD:	Shifted Transversal Design
WGS:	Whole Genome Sequencing

List of Papers

This thesis is based on the following papers, which are referred to in the text by their Roman numerals:

I C. Clouard, K. Ausmees, and C. Nettelblad. A joint use of pooling and imputation for genotyping SNPs. BMC Bioinformatics. October 2022. DOI: 10.1186/s12859-022-04974-7.

II C. Clouard and C. Nettelblad. Consistency study of a reconstructed genotype probability distribution via clustered bootstrapping in NORB pooling blocks. Technical report No. 2022-005 in IT series, June 2022.

III C. Clouard and C. Nettelblad. Genotyping of SNPs in bread wheat at reduced cost from pooled experiments and imputation. Theoretical and Applied Genetics. January 2024. DOI: 10.1007/s00122-023-04533-5.

IV C. Clouard and C. Nettelblad. Using feedback in pooled experiments augmented with imputation for high genotyping accuracy at reduced cost. *Submitted to* PLoS Computational Biology. *Preprint:* `https://www.biorxiv.org/content/10.1101/2023.12.12.571203v1`.

Reprints were made with permission from the publishers.

Moreover, all chapters in this manuscript are largely based on the following Licentiate thesis:

> C. Clouard. Computational Statistical Methods for Genotyping Biallelic DNA Markers from Pooled Experiments. Licentiate thesis from the Department of Information Technology, Uppsala University. November 2022.
> Available at: `http://www.it.uu.se/research/publications/lic/2022-003/`.

Contents

1. Large-scale DNA sequencing and genotyping for applications in biomedical research and agriculture

The study of the genome of living organisms has been boosted by the rise of new technologies for collecting the genetic data, as well as better-performing computational tools have been developed for processing these data. In the era of 'big data', many research fields have relied on the increasing inflow of genetic data in order to deepen the understanding of various biological phenomena. Examples where the genetic underpinnings are being elucidated include susceptibility to various diseases in medicine, and yield and grain quality in crop breeding.

In this chapter, we first define and explain a few genetic terms, and how these concepts are usually represented as numerical data for computational purposes. We also give a few examples of statistical studies based on human or crop genetic data which are conducted in modern genomic medicine and plant breeding.

The work presented in this thesis has been conducted using two very different populations as demonstration examples. The first one being a human cohort and the second one a collection of bread wheat lines. We briefly present some features and characteristics of these populations, and to what extent this can affect their genetic structure. This structure may then influence the models and computational methods used in our experiments.

As the quality of the data are crucial to the reliability of the results in later analyses, we briefly review some current technologies for collecting the genotype data. Overall, beyond depending on the quality of data, the quality of the results of the statistical analyses usually benefits from larger volumes of data that ensure statistical significance. A current challenge in all species is the cost of collecting and processing larger data sets, both in terms of the number of samples and the number of genetic positions surveyed.

1.1 Genetic data: DNA, nucleotides, and chromosomes

The visible characteristics, or phenotype, of an organism as well as its metabolism are the result of the expression of the genetic code. The genetic code of living organisms is encoded within genetic material which consists of *DNA*

molecules. DNA molecules have remarkable stability properties when replicated, divided and shared through the sexual reproduction. This behaviour is a fundamental assumption in parentage and population genetic studies. A DNA molecule consists of two complementary strands, each being a sequence of *nucleotides* . The nucleotide (or base) adenine (A) on one strand is always paired to thymine (T) on the other strand, while cytosine (C) is paired to guanine (G). Therefore, the DNA molecule is usually denoted a sequence of base pairs (*bp*). The strands are oriented, by convention the reference strand has a *forward* orientation.

The *chromosomes* are sequences of DNA that are stored in a compacted form in the cell nucleus. In mammalian species, the chromosomes have a length with an order of magnitude of 10^8 bp. In the case of diploid species such as human, each chromosome has a corresponding homolog, which shares the same structural features and the same genes at the same loci (genetic positions). The exact DNA sequence of each homolog depends however on its parental origin. Thus, homologous chromosomes can carry different *alleles* at corresponding loci. The combination of alleles is described as a *genotype*. Genetic diversity results from the processes of mutation and recombination that occur in DNA over generations.

Millions of loci in the human DNA have been identified as known positions of genetic variation, which are referred to as *variant* positions or markers. One type of genetic markers are Single Nucleotide Polymorphisms (SNPs), which have positions in the DNA that match a single pair of nucleotides. Most SNPs are biallelic, which means that each individual can carry at most two different alleles at any SNP locus on the homologous chromosomes. f The possible nucleotides on the reference strands at this specific genetic position are called morphisms, and they are the same for all the individuals of the species. The morphisms can be any combination from the set A, T, G, C and vary with each SNP.

The complementary nature of the two DNA strands allows to simplify the genome representation as a single strand of nucleotides as shown on Figure 1.1. In general, SNPs are not adjacent to each other in the genetic sequence.

Usually, the two alleles of a SNP are arbitrarily typed as *reference* or *alternate* (Figures 1.1 and 1.2). This choice can be unrelated to the allele frequencies in a population, but it is more common that the reference allele refers to the most frequent or the ancestral one. The degree of statistical correlation between alleles at two loci is called Linkage Disequilibrium (LD). A higher level of LD indicates that specific alleles in two SNPs are more strongly linked and thus found in the same individuals. The resulting series of alleles that derive from the same parent constitute a *haplotype*.

From a computational perspective, SNPs have the advantage that their allele within a haplotype can be represented as a single binary value, where 0 would denote the reference allele and 1 the alternate allele. Commonly, the genotype of a SNP is represented as a ternary entity representing the total allele count

in the locus, with the possible values of $\{0, 1, 2\}$. If an individual carries the same allele in both haplotypes, i.e. its genotype is 0 or 2, it is said to have a *homozygous* genotype, else it has a *heterozygous* genotype (genotype 1).

1.1.1 Example of a human sample from a natural population

Figure 1.1 shows a simplified representation of a pair of human chromosomes and their sequence of nucleotides. The genetic data contains both heterozygous and homozygous loci.

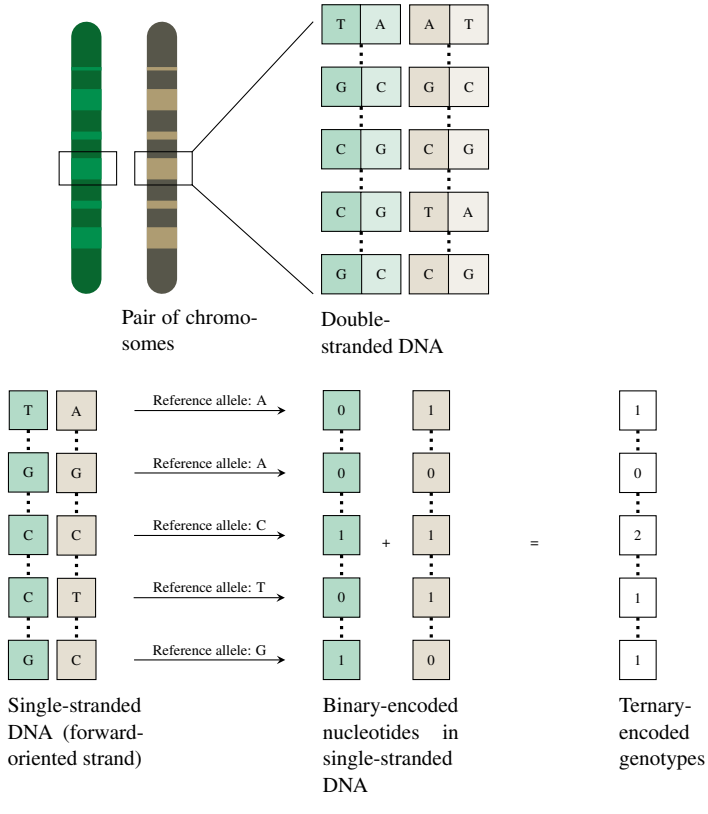

Figure 1.1. **Simplified representation of chromosomes, DNA, and nucleotides for an outbred sample in a diploid species.**

1.1.2 Example of a sample in an inbred line of wheat

Bread wheat is a non-model organism which has 21 pairs of chromosomes. The chromosomes are derived from 3 subgenomes which were hybridised throughout the evolutionary history of the species. Although wheat is for-

mally allohexaploid, in practice it can be treated as diploid for each of the subgenomes A, B, and D [25]. The 3 subgenomes are highly similar to each other and contain about 85% repetitive elements. Together they form a large *pangenome* [53] of about 16 Gb, which is complex and diverse, and remains largely unknown [81].

Figure 1.2 shows 3 pairs of wheat chromosomes from the 3 subgenomes A, B, and D. While wheat is formally allohexaploid due to these highly similar hybridised subgenomes, the genome of wheat can be in practice treated as a diploid one by distinguishing each subgenome. From a computational point of view, this allows for modeling the SNPs data of wheat in the same way as for human data, and parse the genotypes at any marker as a pair of alleles, a reference and an alternate one. The example proposed represents the chromosomes in an inbred sample, in which the breeding process has eliminated any heterozygosity. Therefore, at any locus, the ternary-encoded genotypes are assumed to have values in $\{0, 2\}$ only.

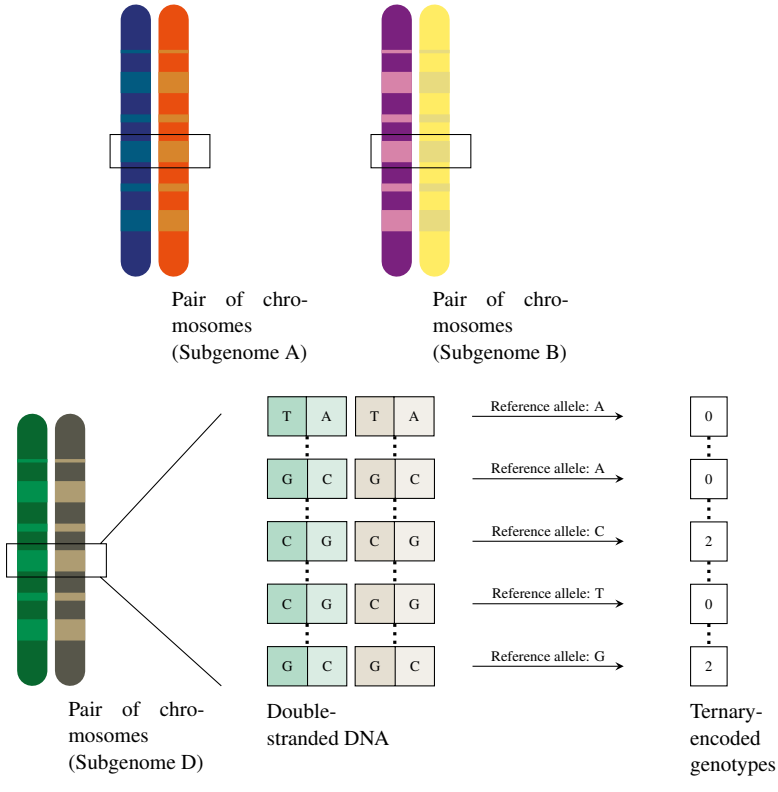

Figure 1.2. **Simplified representation of chromosomes, DNA, and nucleotides for an inbred sample in a hexaploid species.**

1.2 Applications and uses of the genetic data in life sciences

1.2.1 Biomedical research and pharmacogenetics towards personalised medicine

In biomedical research, genetic data can be used in Genome-Wide Association Studies (GWAS) to investigate the susceptibility and the heritability of both common and rare human diseases. The aim of these studies is to find statistical correlations between genotypes and phenotypes [8] in a selected population. Often, the phenotype of interest is the status for a disease such as diabetes. In the case statistically significant correlations are found, the results of GWAS can be interpreted as the risk of developing the disease being associated with a particular genetic profile, commonly expressed quantitatively as a polygenic risk score [24]. In GWAS, the size and the diversity of the cohort studied as well as the density in genetic markers that are screened, are critical factors that condition the accuracy and the quality of the conclusions. Moreover, the rare variants — defined as those having a minor allele frequency (MAF) less than 5%, may provide crucial profiling information to draw conclusions in GWAS [8, 52], and require special attention in genotyping. Large-scale genomic studies also open up prospects for future applications such as the personal genome project [71] and personalised medicine based on genetic profiles.

1.2.2 Genomics-empowered plant breeding

The aim of plant breeding is to actively and efficiently select individuals with the genetic basis for superior phenotypes to become candidates for new competitive varieties. Usually, the selection process spans over several generations, also known as breeding cycles [54]. Selection in conventional plant breeding is based on observable characteristics (phenotypes) of the plants, so that individuals with favourable and competitive profiles gradually emerge over several generations. In bread wheat (*Triticum aestivum*), for example, plant breeders have for example tried to breed varieties with higher yields, higher protein content, and improved drought resistance. Selection practices have resulted in a large number of cultivated varieties of bread wheat, or *cultivars*, out of which only a limited number are considered as *elite cultivars* because of their excellent and stable agronomic and nutritional characteristics.Traits with high heritability are suitable for efficient phenotype-based selection of candidate founder individuals to be crossed into the next generation. After one or more crossing cycles, several generations of selfing are carried out. Selfing, i.e. the mating of an individual with itself, results in purely homozygous inbred lines, which are usually preferred because they are genetically more stable. It can easily take ten generations, i.e. ten years, to develop a new bread wheat variety.

Modern plant breeding has been going beyond direct phenotype-based selection, increasingly adopting genetic data to accelerate the creation and the release of suitable wheat cultivars since the 1990s [53, 62]. In the last few decades, both plant breeders and plant biologists have made increasing efforts to understand the genome structure of bread wheat and the traits that can be improved through selection, such as yield, disease resistance, protein content, and baking quality. Genetic data, such as the genotypes of high-density single nucleotide polymorphisms (SNPs), allow the crop science community to establish correlations between parts of the genome and the expression of specific phenotypes. In other words, certain sequences of markers can be viewed as genetic signatures associated with specific traits. Detection and analysis of marker–trait associations have been developed both in quantitative trait loci (QTL) mapping experiments and more recently in GWAS. In GWAS, the traits of interest are assumed to correlate to markers which are either part of the genes that control the genes of interest themselves, or in LD with them. QTL analysis and mapping can enrich and complement, or even replace, phenotypic data for selection purposes through marker-assisted selection (MAS), as well as genomic selection (GS) based on GWAS [7, 16, 32, 40, 54, 63, 72, 83].

Nevertheless, the QTL mapping technique often fails to capture the architecture of more complex traits that may be controlled by small-effect genes, and mapping accuracy is generally lower with less well defined traits [20, 72]. Gene-trait mapping becomes even more challenging when the complex traits of interest have low-heritability because they are are subject to adaptation to multiple environmental factors, such as when attempting to characterise populations growing in different ecosystems around the world [20]. For example, some traits that are major components of wheat yield, such as the thousand kernel weight and the kernel size, are complex because they involve multiple loci and genes that are also influenced by the environment and cannot be accurately mapped to the traits of interest via QTL [67].

In comparison to discrete QTL mapping, GS is well suited and effective for understanding complex traits [83] and captures the genetic profiles at a larger scale. In addition, GS has demonstrated high accuracy for predicting traits with low heritability [48, 16]. GS is a prediction method borrowed from animal breeding [50] that relies on a so-called genetic estimated breeding value (GEBV) to decide which individuals to use as parents for the next generation. GS offers promising outlook for accelerating plant breeding schemes [63, 72], as well as for decreasing their cost. A wide range of suitable statistical models have been investigated, mainly random regression but also GBLUP, Bayesian regression, and machine learning approaches [63, 72]. Generally, the accuracy of the GEBV is higher when using large genetic data sets with high marker density [48, 54, 63], and large populations [49].

While the use of low-density sets of markers in GS and in QTL mapping studies would limit the cost of genotyping, it has been shown that the GEBV prediction accuracy and the QTL mapping precision are improved with high-

density marker sets. Marroni et al. also argued that investigating the role of rare variants in GS could enlighten the missing heritability observed for complex traits in GWAS [48]. Pooling strategies, such as the ones explored in this thesis, are not only cost-effective for large-scale genotyping, but they also have the advantage of being accurate in identifying rare but valuable variants. Therefore, pooling may represent a relevant genotyping solution in plant breeding, allowing high-density genotyping at limited cost, including for the low-frequency variants.

1.3 Populations of interest and genetic data resources

Genetic data used in biomedical research and in agriculture can be collected from various sorts of populations. The conditions and mechanisms that have shaped a population, as well as their relationships to each other, may be relevant to know in order to understand the performance and the results of genotype imputation presented in Chapter 3. In this thesis, our work has been guided by studies using genetic data collected in a human *natural population* on the one hand, and in an *experimental population* of bread wheat on the other.

Natural populations are not deliberately created by humans and random mating between the individuals is often assumed. In the 1000 Genomes Project data set and in many other human natural populations, the assumption of Hardy-Weinberg equilibrium (HWE) holds. According to the HWE principle, the allele and the genotype frequencies observed remain constant over the generations. In the case of biallelic SNPs, HWE can be expressed as an equation that describes the distribution of the genotypes at any locus with respect to the alternate allele frequency at this locus. The data sets available for human populations fall in the category of natural populations and some populations of crop species can be found too. Natural populations are frequently characterised by a high level of genetic diversity that is the result of many recombination events over numerous generations and the wide panel of available individuals at each generation. Experimental populations, on the contrary, are generated artificially through deliberate breeding choices. They are much smaller in size than most natural populations and the offspring in each generation often derive from an active selection of the parents. These generating processes of populations are one of the reasons for explaining the very different genetic profiles and genetic structure of the humans vs. crop species.

Data resources for human populations

The human genome has been one of the most studied ones among the living organisms, in various research fields such as in population or evolutionary genetics, but also in biomedical and epidemiological research, and in particular

in GWAS. Active research on human genetic data has led to the creation of various data resources, for instance a reference genome, several releases of genome assemblies, and multiple linkage or recombination maps. Overall, the genetic profile of the human populations is well-known.

Several large and well-documented public data sets of genotypes exist, such as *1000 Genomes Project* (1KGP) [75] that we have used in Paper I, the HapMap consortium [80], the UK BioBank [74], and the SweGen project [5]. They consist of individuals from natural populations that were sampled around the world. Beyond the privacy concerns that have emerged with such data, for instance with human genetic testing [41] or with accessing personal data e.g. GDPR, it has become crucial to get access to large data sets. Up to thousands of samples and million of SNPs and diverse populations are desirable to achieve relevant statistical studies.

The cost per sample for genotyping human genomes has decreased significantly thanks to the advanced technologies and the numerous manufacturers which offer various solutions for collecting the data. In scenarios where most of the cost of including additional samples is related to genotyping, the large-scale impact of such costs can still be prohibitive. In breeding programs, growing another plant can be cheap, and animals can be reared for other purposes, so the marginal cost of including them in a program with genetic testing would be determined by the genotyping cost. We believe that our methods of combining pooling and imputation would be a useful addition in such circumstances.

Plant material for modern crop genomics and breeding

Plant breeders and crop scientists have developed breeding programs and experimental populations that are suitable for more precise trait mapping and thus make the selection more effective. In a breeding program, the parents in the first generation are called founders. Some of the cycles, usually in the first generations, are crossing cycles, where selected parents are mated to produce the next generation of plants. The later generations are often selfing cycles, which consist in self-pollinating the plants in order to obtain lines that are homozygous at all loci and therefore genetically stable. Selfing corresponds to the most extreme form of inbreeding. In such case, offspring in the last generation are often called inbred lines. The limited number of breeding cycles in experimental populations, and their small size compared to natural cohorts, lead to a relatively low genetic diversity and a lack of fine-grained genetic structure. These specific characteristics, together with the homozygosity of individuals, can affect the statistical models used to analyse the population.

Bi-parental populations, recombinant inbred lines (RILs), and advanced intercross lines (AILs) are examples of experimental populations designed to improve the precision of genetic mapping [19, 54, 68]. RILs can follow a multi-parental scheme and therefore have a higher genetic diversity than bi-parental populations, while the AIL design was created to address the problem

of the low recombination rates. Yet, all these experimental populations have a poor mapping resolution for the complex traits.

The multi-parent advanced generation intercross (MAGIC) breeding scheme was developed to combine the high level of genetic diversity found in natural populations with the advantages of synthetic populations [46, 68]. Initially studied in animal species for mapping purposes, the MAGIC scheme was mentioned in crop species in 2007 with the intention of improving QTL mapping [19, 46]. Since then, MAGIC wheat populations have demonstrated their suitability for QTL analysis [29], for example allowing for successful QTL mapping of the hectolitre weight complex trait [37]. More genes of interest, especially complex ones, and more functional variants are yet to be discovered and understood in wheat, which is likely to increase the demand for larger and higher quality data sets in the coming years [20]. In this regard, MAGIC populations represent a useful tool for genomics-assisted plant breeding, which requires appropriate methods for large-scale and cost-effective collection of genetic data.

In Papers III and IV, we apply our genotyping strategies to a subset of the data released in 2020 by the National Institute for Applied Botany (NIAB), which published the curated genotypes of approximately one million SNPs for a MAGIC population of 504 inbred lines derived from sixteen parents [69].

1.4 Technologies and computational methods for sequencing the DNA and genotyping markers

In this section, we briefly present some current techniques for collecting and processing the genetic data.

Historical evolution of DNA sequencing and successive improvements

DNA sequencing consists in determining the nucleotide sequence of selected portions of the DNA of an organism. DNA sequencing technologies are usually divided in successive generations that have been developed since the late 1970s. First-generation sequencing, originally synonymous with Sanger sequencing, allows scientists to identify readouts after separating DNA fragments on polyacrylamide gel [23, 30, 33].

In second-generation DNA sequencing, significant cost reductions were realised with the introduction of so-called Next-Generation Sequencing (NGS) technologies, which have been used with both human and crop DNA [7, 48, 53]. The NGS machines perform massive parallelisation of the sequencing for numerous individuals [32] by using multiplexed schemes for the DNA probes. NGS has drastically reduced the cost of sequencing and facilitated Whole Genome Sequencing (WGS) projects, although the cost of preparing the multiplexed samples, called libraries, can remain prohibitive.

Current developments in third-generation sequencing technologies [84] give the research community hope of achieving ultra-low-cost sequencing. Pacific Biosciences (PacBio) sequencing and the Oxford Nanopore Technologies (ONT) are emerging alternatives that produce long reads with high accuracy [45]. This type of output may be particularly relevant to studies aimed at sequencing plant genomes, which often contain a large number of repeated elements that can be difficult to order correctly from short reads.

Technologies for genotyping variants of interest in DNA

The genotyping technologies are designed to determine what alleles are found in a particular variant, i.e. which nucleotide is present at single genetic positions in the case of SNPs. The term *genotyping density* refers to how many SNPs per kbp are genotyped and can be understood as the average physical distance between consecutive markers.

Specific processing techniques and selective amplification can be used to extract the genotype of single positions from sequence data generated by sequencing technologies. This approach is called genotyping-by-sequencing (GBS) [66]. In recent years, GBS has gained popularity in the plant breeding community to achieve high-throughput and low-cost marker genotyping [86], particularly due to its flexibility in variant discovery experiments. One of the drawbacks of GBS is the extensive bioinformatic analysis required to obtain usable data, which can be technically challenging [2].

Alternatively, genotyping can be performed using DNA array technologies, such as fluorescence-based SNP microarrays, which are specifically designed for a fixed set of previously identified variants. All the research described in this thesis has been carried out in simulated data representing pooled genotyping experiments on such microarrays. The fluorescent detection of nucleotide on DNA arrays was developed in the late 1990s and 2000s. Figure 1.3 [1] shows an example of array technology that detects which alleles are present at each variant. In the context of pooled genotyping, it is important to note that the colour intensity of the signal measured when reading the array is not linearly proportional to the allele concentration in the samples, and some recalibration might be needed to properly derive all mixing proportions.

Microarrays can carry a high cost for designing an array with a new variant set [66, 86], but they are well known tools with established bioinformatics pipelines that remain cheaper in cost per data point on a longer term for routine high-throughput genotyping [38]. The processing steps in the laboratory to generate data from arrays and to analyse these data are cheaper and easier to run than for NGS data [44]. In addition, microarrays provide measurements with higher confidence levels than NGS as typically used for genotyping [62], making them a suitable and reliable option for producing data intended for imputation [40]. Although less flexible compared to the GBS techniques, breeders still find the hybridisation-based SNP arrays well suited for routine genotyping [40].

Today, the three dominant manufacturers of SNP microarrays are Affymetrix, Illumina, and Thermofisher, offering a wide range of specific equipment and services for human data as well as for crop species such as bread wheat. For example, we performed the simulations in Paper I in filtered low- and high-density sets of markers that are offered by the bead-chip products *Infinium OmniExpress-24 Kit* and *Infinium Omni2.5—8 Kit* from Illumina. In Papers III and IV, we model a hypothetical array from a data set obtained by careful post-processing and stringent filtering of sequencing reads. Designing a microarray for a polyploid species such as wheat can be challenging in practice, but some products exist, for example the Wheat Breeder's array from Affimetrix [4].

Despite recent advances, the current sequencing technologies still have limitations for statistical studies based on the sequenced data. For example, NGS technologies produce only short reads [33], which makes the technology unsuitable for genome assembly purposes and for the analysis of genomic variations that might be observed in longer segments. Irrespective of the technology, the single positions read from SNP genotyping and their sparsity cannot capture all the structural variants in the genome either, and sometimes fail to capture the LD [30]. In addition to their aforementioned advantages, PacBio and ONT technologies also provide better data on LD information and could therefore address this specific issue.

1.5 Overview of the remaining chapters

This thesis explores computational approaches that can tackle the difficulty of achieving cost-effective genotyping in large populations, for both rare and common genetic variations, and thus contribute to future advances in biomedical genomic research and plant breeding. Indeed, these two fields, among others, are increasingly using larger volumes of genetic data in their analyses. Technological advancements in sequencing and genotyping have considerably eased the collection and accessibility of such data. However, we believe that scientific computing can alleviate some of the remaining weaknesses of certain genotyping techniques, such as the difficulty in identifying loci with rare genetic variations, whilst also reducing the cost of genotyping.

In our work, we have mainly considered two examples of real, and very different populations. Specifically, these are a population reflecting actual worldwide human genetic diversity, and a set of inbred lines of bread wheat. The simulations we performed in the papers supporting this thesis illustrate potential applications in genomic research and plant breeding of our computational hybrid approach to genotyping, which combines pooling and imputation. Chapter 1 gives an overview of these practical applications, along with the data and technologies they require. Our models for pooling augmented by imputation consider the nature of the data i.e. biallelic SNPs, the technology

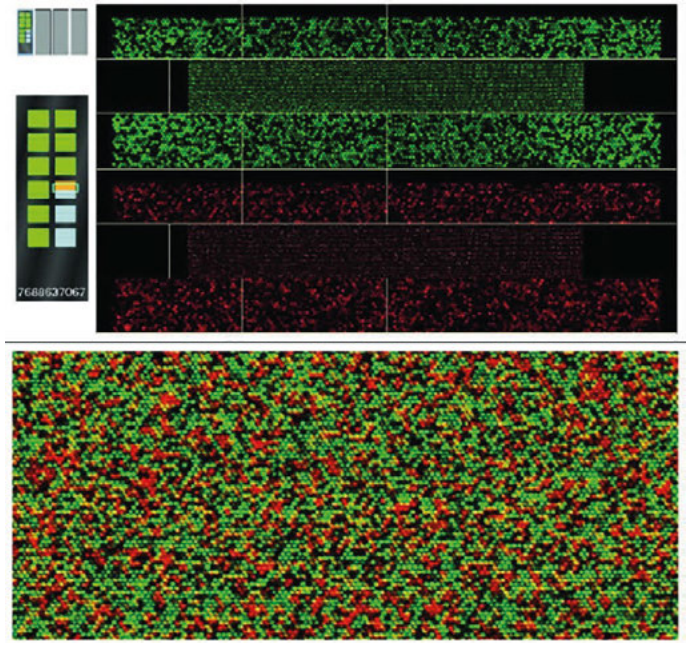

Figure 1.3.

Example of an Illumina BeadChip microarray for SNP genotyping. A microarray is a glass plate with one well for each SNP to be genotyped, one individual can therefore be tested for thousands or millions of positions on a single plate. On each chip, there is a collection of short fragments of synthetic DNA probes (less than 100 bp in size) that are complementary to the sequences where the SNP of interest is located. The probes are attached on a chip [30]. After denaturation, amplification, and fragmentation of the DNA, the allele discrimination is done by hybridising nucleotides marked with a fluorescent dye (dNTP*). The SNP alleles are determined based on the color of the fluorescent signal on the chip. Usually, a sample which is homozygous at the locus of interest returns either a red or a green light signal depending on the allele which is detected. For the heterozygotes, both the reference and the alternate alleles are detected on the chip. The returned light signal combines both red and green fluorescence, resulting in a yellow colour.

Upper panel: The BeadChip in this example is made of 12 cells on a glass plate and marked with a identifying number at its bottom. Each cell has thousands of micro wells where the DNA probes are fixed. After hybridisation, the cells are scanned with red and green fluorometry.

Lower panel: Zoom-in on a cell. The software combines and interprets the results as colors.

that is used to collect the data i.e. SNP microarrays, its limitations, and the particularities and characteristics of the populations studied, in particular the influence of their genesis and design on their genetic structure.

Chapter 2 outlines key aspects of group testing theory and introduces our pooling design methodology. In essence, the pooling strategy efficiently cap-

tures the rare variants, which is particularly important in certain approaches, such as GWAS. Previous research has highlighted the important role of rare variants that may affect complex diseases in humans, or, in crop species, complex traits of agricultural relevance. However, pooled genotyping frequently leads to missing genotype data for common variants. We define the typology of these data to determine which methods can be used to infer the missing genotypes from the combinatorial constraints imposed by the chosen pooling design. We propose a few examples of relevant inference methods to estimate the genotype probabilities in pooled data, which can be tailored to accommodate the genetic composition of the population.

In Chapter 3, we examine how genotype imputation is interacting with pooling in two imputation methods, and how this scenario differs from the more commonly used case of genotype imputation from low-density data. We elaborate on this by providing minimal examples to illustrate how pooled genotypes influence the imputation model in contrast to standard low-density data sets. We also outline that a likelihood framework for expressing the genotype distributions offers a suitable versatility for taking advantage of the complementary strengths of pooling and imputation in different applications. By using a probabilistic representation of the genotype data and proper algorithms, computational methods which augment genotype pooling with imputation can suit various types of population, and enable a cost-conscious and accurate genotype reconstruction of both rare and common SNP variants.

All papers pertain to several chapters in this thesis. In each chapter, we refer to the parts of the papers that are relevant to the topic treated in that chapter, along with highlighting the particularities of each study. Section 5 provides more details about the contents of each paper and the contributions of the different authors to them.

2. Pooled experiments with genotype data and probabilistic decoding methods

The general pooling problem is to accurately and efficiently identify a few deviant items in a population in an accurate and efficient way [56] by testing groups of samples rather than performing individual tests.

A first major research question in group testing is which strategy to choose for constructing the groups to be tested. This can be seen as an optimisation problem, which can be tackled by finding a design that minimises the number of pools, or that limits the size of the pools.

Although we have explored some design variations for allocating the samples to pools, this problem has not been the focus of this thesis. In the following sections, we analyse one specific pooling strategy that we have proposed for genotyping markers on microarrays, which consists in a row-column design that we refer to as the Nonadaptive Overlapping Repeated Block (NORB) design. A second research question in group testing is the algorithms to confidently retrieve each individual result from the pooled observations pools, in other words how to resolve the pools in the most accurate way.

We present tailored inference methods for estimating the missing genotype data in cases where direct inference of the genotypes without ambiguity is not possible in a NORB pooling design. In this sense, the pooling algorithm can be defined as the missing data mechanism that underpins which genotypes can or cannot be determined from the pools on the level of individual samples. The specific characteristics of the structure of the missing data in pooled experiments determine which statistical inference methods are appropriate, as well as their potential caveats due to the peculiar dependence structure in the data set. The characteristics of the genetic data being pooled, such as the level of diversity and the distribution of the genotypes, affect the decoding power of the design, as our simulations with outbred human samples (Paper I), and inbred lines of bread wheat (Papers III and IV) show. The decoding results can in turn affect the accuracy of imputation in various ways. This is presented in Chapter 3.

2.1 Group testing for sequencing and genotyping

Group testing, or pooling, is relevant to applications that require large volumes of data, and where the total cost of sequencing and genotyping can be the

limiting factor to up-scaling. This situation might arise in genetic studies of non-model organisms, where the cost per test tends to be higher. Cheaper alternative strategies that aim to collect genetic data in sparser marker panels have the disadvantage of often missing the rare but highly informative variants. Beyond the technological efforts to parallelise and automate sequencing and genotyping, pooling represents an additional strategy to reduce the cost of large-scale genomic testing [30].

DNA pooling has been used successfully since the 1990s for large-scale association studies of human diseases [70] and later for breeding and selection purposes as in rice [22] or cattle [3]. A well-known application of pooled sequencing is the detection of the carriers of rare variants that can have a significant impact in GWAS [30].

2.1.1 Categorisation of group testing schemes

Principle of pooling

In group testing, the samples to be evaluated are mixed together, or pooled, and tested jointly. This reduces the total number of tests performed compared to individual testing. Typically, tests have a binary outcome, e.g. the infection status for a disease. A defective item, for example an infected individual, returns a positive test result, whereas the other items return a negative one. The definition of 'defective' items depends on the test being performed, in our case these are samples carrying the least frequent allele in a pooling block at a given locus. In a natural human population as in Paper I, the defective genotypes can be either heterozygous or homozygous for the least frequent allele, i.e. the genotyping test has a ternary outcome. For inbred wheat lines as in Papers III and IV, we have assumed complete homozygosity for each sample, thus simplifying the defective status to homozygote for the least frequent allele.

A key assumption underlying pooling is that the result of a test for a pool is positive if at least one of the items in the pool is positive. For instance, numerous studies have recently been published on practical applications of pooling for population screening and the identification of groups of individuals that were infected with the severe acute respiratory syndrome coronavirus 2 (SARS-Cov-2) [27, 64]. If the test result for a group was positive, all individuals in the group group were considered to be possibly infected, or at least exposed. This type of testing was most commonly used in the US in some schools across the country.

In pooled genotyping on a microarray, a mixture of DNA fragments from all individuals belonging to the same pool is deposited on the chip. As described in Figure 1.3, the light signal returned by the chip depends on the allelic composition of the DNA mixture but is not proportional to the allelic dosage. Within our papers, we have assumed error-free genotyping, that is the allele detection on the chip is fully reliable.

Families and categories of group testing schemes

Two main categories of group testing are commonly mentioned in the literature. The first one is combinatorial group testing, which assumes that the maximum number of defective items in the population to be tested is known and fixed to some integer. The second category is probabilistic group testing, where a fixed probability is set for any item to be defective.

If pooling and testing are repeated s times. and each new iteration depends on the results of the previous one, the pooling design is said to be s-staged, or adaptive. If the procedure for forming the pools and testing them is specified independently of any other results and for only one stage, the pooling design is nonadaptive [56].

Each SNP chip is manufactured for a predetermined set of SNPs, allowing the genotypes of thousands of SNPs to be tested simultaneously. This setup does not allow for adaptive testing of individual SNPs or a subset of the SNPs targeted. Therefore, only nonadaptive group testing (NGT) algorithms can be used for SNP genotyping purposes [17, 87].

Strategies for constructing deterministic designs

Various methods have been studied for constructing deterministic pooling designs such as pooling-deconvolution, shifted transversal design (STD) and its hypergraph extensions, multiplexed schemes, or compressed sensing [17, 21, 28, 30, 56]. Within this thesis, we have used a nonadaptive approach [79], which arranges the samples into equally sized blocks and constructs overlapping pools in each block. This block construction is repeated across the population so that each sample to be genotyped is assigned to one block. Section 2.2.1 in this chapter provides more details on pooling terminology. The designs used for DNA library screening or rare variant frequency estimation are not necessarily overlapping [82], severely limiting the possibility to decode the identity of individual carriers.

2.1.2 Properties and parameters of deterministic nonadaptive pooling designs

Definitions and notation

A pooling design defines an algorithm that determines the encoding and decoding rules. The assignment of samples to pools corresponds to an encoding step [28]. The process of determining the test results for each individual from the pooled results is referred to as the decoding step of the pooling design. For example, in some studies that use nonadaptive overlapping designs [79, 85], the decoding step follows a pattern consistency rule.

A nonadaptive group testing design with repeated blocks can be compactly described for each block by a design matrix, which we denote M. M is a matrix with binary entries of dimensions (T,B), where each of the T rows

represents a pool and each of the B columns represents a sample. The entry 1 at coordinates (i, j) indicates that the sample j belongs to the pool i, otherwise the entry is 0. An example of a design matrix, largely based on the STD and the DNA Sudoku, is given in the next section of this chapter.

Let $y = [y_1, \ldots, y_T]$ be the vector indicating the test results for the pools. Likewise, we denote $x = [x_1, \ldots, x_B]$ the vector representing the test results for the individuals. The relationship between the outcomes of the pooled tests in y and the decoded results for each sample represented by x can be modelled as the ceiled result of the multiplication of the design matrix M by the outcome vector y

$$y = \lceil M \cdot x \rceil \tag{2.1}$$

Solving the pooling problem consists in finding the vectors x that satisfy 2.1 given the observed outcomes in the vector y. Typically, it is desirable for the design matrix M to be d-disjunct. The design then guarantees exact reconstruction of the vector x if there are at most d defective items in it. d is also called *decoding robustness*.

Performance-critical parameters of the pooling design
Let us define the reduction factor $\rho = \frac{B}{T}$ based on the suggestion in [87] for an under-sampling ratio. Optimising the pooling design consists in finding a trade-off between the reduction factor and the decoding robustness. Ideally, both ρ and d should be as large as possible. If there are more than d defective samples in the population to be pooled, some items are missing after pooling because they cannot be decoded.

2.2 Example of a Nonadaptive Overlapping Repeated Block design for reconstructing SNPs genotypes

This section introduces the simple case of STD that is used in all our papers to simulate pooled genotyping experiments. Given the characteristics of the design, we choose to refer to it as a Nonadaptive Overlapping Repeated Block (NORB) pooling design.

2.2.1 NORB parameters and design matrix

A NORB pooling design can be described with the following properties:

- The population to be tested is divided into **blocks** of equal size B [87]. In our experiments, we chose $B = 16$.
- This means that if the study population consists of 160 individuals, a block unit is **repeated** 10 times.

- Within each block, we assign the individuals to pools, such that each pool consists of 4 samples and each sample is part of $W = 2$ pools. In other words, there are 2 pools **overlapping** for each sample. Moreover, each of the $T = 8$ pools in the block intersects any other pool at most $\lambda = 1$ time.
- The blocks and the pools are allocated only once for one testing stage, that is the algorithm is **nonadaptive**.

This NORB scheme can be represented by the following design matrix M, where the horizontal line separate the row- and the column-pools:

$$
\mathrm{M} = \begin{array}{c} \\ R_1 \\ R_2 \\ R_3 \\ R_4 \\ \hline R_5 \\ R_6 \\ R_7 \\ R_8 \end{array}
\begin{array}{c}
\begin{array}{cccccccccccccccc} C_1 & C_2 & C_3 & C_4 & C_5 & C_6 & C_7 & C_8 & C_9 & C_{10} & C_{11} & C_{12} & C_{13} & C_{14} & C_{15} & C_{16} \end{array} \\
\left(\begin{array}{cccccccccccccccc}
1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\
1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\
0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1
\end{array}\right)
\end{array}
$$

With this NORB design, the matrix M is said to be d-disjunct if Equation 2.2 with the maximal value for λ and the minimum value for W is satisfied:

$$d = \left\lfloor \frac{W_{min} - 1}{\lambda_{max}} \right\rfloor \tag{2.2}$$

The decoding robustness of the design d is defined as the maximum number of carriers of the alternate allele in the block that can be identified with certainty. Our design has a decoding robustness $d = \frac{2-1}{1} = 1$.

In the case of genotype testing, identifying the carriers does not mean that the exact genotype of these items can be resolved, as they could be either heterozygous or homozygous for the alternate allele. The reduction factor of the pooling design is $\rho = 2$, which means that half the number of tests are needed to genotype the pools compared to doing one test per individual.

2.2.2 Representation of a pooling block

As an alternative to the design matrix, we have used within our papers a more intuitive, graphical representation of pooling blocks. Figure 2.1 shows this block representation as a 4×4 square grid. The $B = 16$ individuals fill the cells of the grid. Each row of the grid consists of 4 samples that belong to the same pool, and likewise for the columns of the grid. That is, there are $T = 4 + 4 = 8$ pools in a block. Each sample intersects one row and one column of the grid, which corresponds to the weight of the design, $W = 2$.

	P_5	P_6	P_7	P_8
P_1	I_1	I_2	I_3	I_4
P_2	I_5	I_6	I_7	I_8
P_3	I_9	I_{10}	I_{11}	I_{12}
P_4	I_{13}	I_{14}	I_{15}	I_{16}

Representation of a NORB pooling block as a square grid.

Figure 2.1. The 16 samples $(I_1, I_2, ..., I_{16})$ are assigned to the 8 pools $(P_1, P_2, ..., P_8)$ and each sample belongs to 2 distinct overlapping pools.

2.2.3 Algorithms for encoding and pattern-consistent decoding

Previous implementations of overlapping pooling schemes [18, 82, 85] were interested in identifying carriers of a rare variant. We initially explored pooling in ternary genotypes (heterozygote and two types of homozygote) in the case of outbred human data, and later in binary genotypes (two opposite homozygotes) for inbred wheat lines.

Algorithm 1 provides pseudocode for simulating the genotype of a pool based on the genotype of the samples in that pool (encoding step). This algorithm uses only integer-valued genotypes. In Paper IV, we propose a different procedure for simulating pooling, which determines the expected genotype of a pool as a distribution based on the alleles that are detected in it, given the prior genotype probabilities of the samples in the pool.

The rules for decoding ternary genotypes $G_{ij} \in \{0,1,2\}$, as in Paper I and II, are described in Algorithm 2. Any individual is resolved as homozygote for an allele iff it belongs to at least one pool which is homozygote for that allele. We want to point out that with the rules we use for encoding, it is impossible that two overlapping pools are opposite homozygotes. Given the symmetry property for the reference and the alternate alleles, the decoding procedure is similar for the genotypes $\{0,2\}$. If both overlapping pools have heterozygous genotypes, then the genotype of the individual belonging to these pools cannot be resolved with certainty. Intuitively, the larger the MAF, the more likely such an unresolved situation is to occur. Unresolved genotypes are reported as completely missing ($\emptyset$) if neither allele is assayed, or partially missing ($\{0,1\},\{1,2\}$) if the presence of one allele is definite but the other allele is indeterminate. In Paper I, all entries in a pooling block are represented with integer-valued genotypes and fully or partially indeterminate entries are assigned to the value -1. For binary genotype outcomes, such as in Papers III and IV where the inbred lines are assumed to be fully homozygous at each lo-

Algorithm 1 Genotype encoding with a NORB pooling design

P_{jk} is the genotype of the kth pool at the jth marker
G_{ij} is the genotype of the ith individual at the jth marker
k is the kth pool
for all j **do**
  for all k **do**
    if $\{G_{ij} = 0\}, i \in k$ **then**
      $P_{jk} \leftarrow 0$
    else if $\{G_{ij} = 2\}, i \in k$ **then**
      $P_{jk} \leftarrow 2$
    else
      $P_{jk} \leftarrow 1$
    end if
  end for
end for

cus, there is less ambiguity in decoding heterozygous pools, resulting in fewer missing data. For example, $\{P_{ijk} = 1\}, i \in k \cap \{P_{ijk} = 0\}, i \notin k$ implies $G_{ij} \leftarrow 2$, and likewise, $\{P_{ijk} = 1\}, i \in k \cap \{P_{ijk} = 2\}, i \notin k$ implies $G_{ij} \leftarrow 0$. The decoding robustness of the pooling design is the same with binary as with ternary genotypes, but the homozygosity condition on the genotype values allows a better resolution of the pools. A more realistic scenario with noisy pooled results would require the implementation of an appropriate decoding algorithm so that the procedure accommodates testing errors [56].

The nested tests in Algorithms 1 and 2 are computationally expensive. Depending on the programming language, more efficient alternatives can be implemented in practice, as suggested in [85]. For example, we developed a code used in Papers I, II, III and IV which performs the first decoding and encoding steps with vector-matrix computations. Encoding in a simulation context as well as decoding for simulated or actual data can be performed independently for different genetic markers, making this task suitable for parallel execution.

2.3 Structure and characteristics of the missingness in NORB pooled data

2.3.1 Minimal example of a NORB pooling design

For readability, we use hereafter a smaller example than the 4×4 pooling block. The 2×2 minimal example is intended for illustrative purposes only — for one thing, its dimensions imply that pooling will not reduce the number of tests performed in this scenario. This pooling design would require a total of $2 + 2$ tests for rows and columns, which is equivalent to the number of

Algorithm 2 Genotype decoding with a NORB pooling design for samples that can be homozygotes or heterozygotes

P_{ijk} is the genotype of the kth pool at the jth marker in which the individual i participates
G_{ij} is the genotype of the ith individual at the jth marker
k is the kth pool
for all j **do**
  for all i **do**
    if $P_{ijk} = 0, i \in k$ **then**
      $G_{ij} \leftarrow 0$
    else if $P_{ijk} = 2, i \in k$ **then**
      $G_{ij} \leftarrow 2$
    else if $\{P_{ijk} = 1\}, i \in k\} \cap \{\{P_{ijk} = 0\}, i \notin k\}$ **then**
      $G_{ij} \leftarrow \{1,2\}$
    else if $\{P_{ijk} = 1\}, i \in k\} \cap \{\{P_{ijk} = 2\}, i \notin k\}$ **then**
      $G_{ij} \leftarrow \{0,1\}$
    else
      $G_{ij} \leftarrow \emptyset$
    end if
  end for
end for

tests needed for traditional per-individual testing. Furthermore, the examples below only apply to genotype pooling with ternary outcomes. There would be no ambiguity if the samples were fully homozygous.

The square block representation for a 2×2 pooling design with the definitions used in Chapter 1 is shown in Figure 2.2.

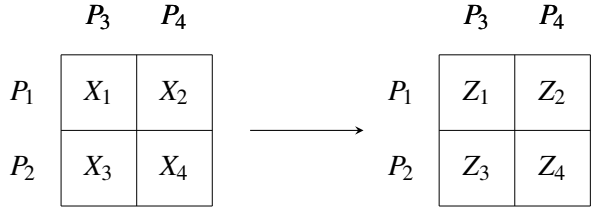

Figure 2.2. **Square block representation of the true and the decoded genotypes for a 2×2 pooling design.**
X is the true complete genotype data for each individual I. Z is the possibly incomplete data after pooling and decoding.

2.3.2 Graph representation of the pooling algorithm as a missingness mechanism

As described by Mézard et al. [55] for investigating missingness before inferring missing data, Directed Acyclic Graphs (DAGs) can be used to represent a missingness mechanism. NORB pooling can be interpreted as a missingness mechanism. Figure 2.3 shows the DAG representation of the 2×2 NORB pooling design.

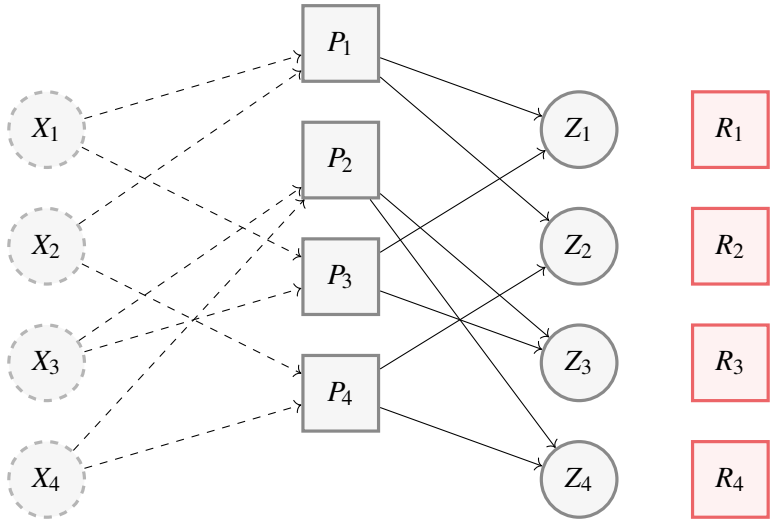

Figure 2.3.

DAG representation of a 2×2 NORB pooling design.

We use notations similar to those proposed by Mézard et al. [55]. The nodes P_i are represented as square nodes since their value corresponds to the direct result of a genotyping test. The variables X_i and Z_i, represented with circle nodes, are individual genotypes in the block. X represents the true data. The dotted lines express that this data is only accessible if we simulate a pooling experiment from previously collected data, otherwise only the data P is known. Z stands for the pooled and decoded data, which is possibly missing. R is a variable indicating the missingness status of Z. The edges on the left-hand side of the DAG indicate what samples X_i belong to which pools P_i, this corresponds to the encoding stage of the pooling algorithm. On the right-hand side, the edges connect the pools from which the genotyping results are combined to be decoded into an individual genotype Z_i.

2.3.3 Classification of the missingness mechanism

There are three main categories of missingness described in the literature, which are defined based on the dependency structure in a data set with missing items. As introduced on Figure 2.3, we use the variable Z to represent the

pooled and decoded data that may be missing. That is, a realisation of Z can generate both determined and indeterminate data. The nature of the dependence between the missingness status R_i of any item i in a pooling block and both the other determined and indeterminate items in the block allows us to distinguish the following categories:

- Missing Completely At Random (MCAR): the missingness status is independent of both the determined or indeterminate data.
- Missing At Random (MAR): the missingness status depends only on the determined data.
- Missing Not At Random (MNAR): the data are neither MCAR nor MAR.

In this section, we use the 2×2 study case shown in Figure 2.4 to illustrate the dependencies in NORB pooled data. Figure 2.4 shows a heterogeneous example with a mix of all genotypes, where 1 pool (P_4) is observed to be homozygous for the reference allele. After decoding, both Z_2 and Z_4 can therefore be identified as homozygotes for the reference allele, whereas Z_1 and Z_3 are missing.

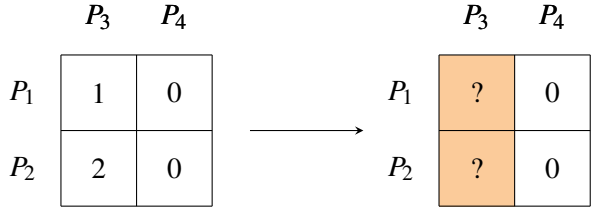

Figure 2.4. **Example of missing data in a 2×2 block after pooling and decoding.** The left block shows the values for the true data X and the right block shows the pooled and decoded values Z. All three possible genotypes are present among the items X, but by chance the homozygotes for the reference allele are placed in the same column-pool. In this example, the pooling pattern is $\psi = ((0,2,0),(1,1,0))$ as both P_1 and P_2 are tested with genotype 1, as is P_3, and P_4 is homozygous for the reference allele. The individual genotypes Z_1 and Z_3 are missing after pooling.

For any NORB pooling block, we define its pooling pattern $\psi = (n_{G_{rows}}, n_{G_{columns}})$, where $n_{G_{rows}}$ (resp. $n_{G_{columns}}$) is a triplet of integers denoting, in this order, the number of row-pools (resp. column-pools) having genotype 0, 1, and 2. For instance, the pooling pattern in Figure 2.4 is $\psi = ((0,2,0),(1,1,0))$.

Dependency between missingness status and observed data

The result of the pooled genotyping test for P_1 affects the decoded value for both the genotypes Z_1 and Z_2. Therefore, given a particular outcome for Z_2 and the jointly observed result for the pool P_1, some values of Z_1 are inconsistent. For example, if $Z_2 = 0$ is observed, as in Figure 2.4, this constrains $Z_1 \neq 0$. Indeed, if $X_1 = 0 \wedge X_2 = 0$, the pooling algorithm generates the result shown in Figure , that is the observed pooling pattern is $\psi = ((1,1,0),(1,1,0))$.

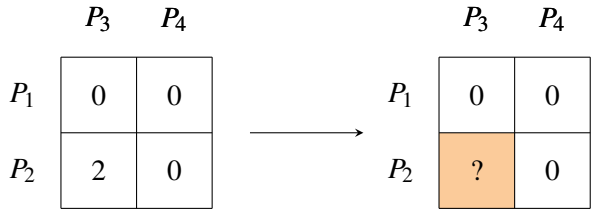

Figure 2.5.

Example of missing data in a 2×2 block after pooling and decoding.
The left block shows the values for the true data X and the right block shows the pooled and decoded values Z. Both homozygous genotypes are present among the items X, but no heterozygous ones. By chance, the homozygotes for the reference allele are placed in the same column-pool P_4 and the same row-pool P_1. In this example, the pooling pattern is $\psi = ((1,1,0),(1,1,0)) \neq ((0,2,0),(1,1,0))$ and only the genotype Z_3 is indeterminate.

In other words, the missingness status of Z_1 is conditioned on the determined value of Z_2.

Dependency between the missingness status and the unobserved data
The test result for the pool P_3 affects the decoded values for both Z_1 and Z_3. Hence, the pooling algorithm imposes $Z_3 = 2 \implies Z_1 \neq 2$. Indeed, if $X_1 = 2 \wedge X_3 = 2$, the pooling algorithm generates the result shown in Figure 2.6, that is the observed pooling pattern is $\psi = ((0,2,0),(1,0,1))$ and the pooling block is fully decoded.

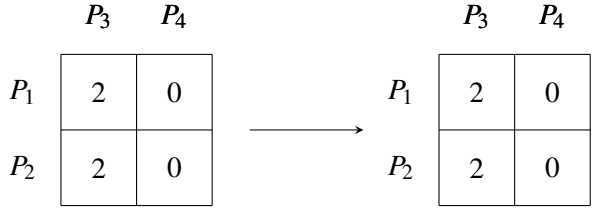

Figure 2.6.

Example of complete data in a 2×2 block after pooling and decoding.
The left block shows the values for the true data X and the right block shows the pooled and decoded values Z. Only homozygous genotypes 0 and 2 are present among the items X, and by chance the homozygotes for the same allele are placed in the same column-pools P_3 and P_4. In this example, the pooling pattern is $\psi = ((0,2,0),(1,0,1)) \neq ((0,2,0),(1,1,0))$ and all individual genotypes can be decoded.

Thus, the missingness status of Z_1 is also conditioned on the unobserved value of Z_3.

Figure 2.7 shows the DAG equivalent to Figure 2.4 with the dependencies between the variables highlighted in blue. Using the example of the relation-

ships between Z_1, Z_2, and Z_3, we show that the missingness status of an item in a NORB pooling block is dependent on both the other observed and unobserved items. Therefore, in accordance with the definitions given at the beginning of this section, we can characterise the undecoded items in NORB pooling as MNAR.

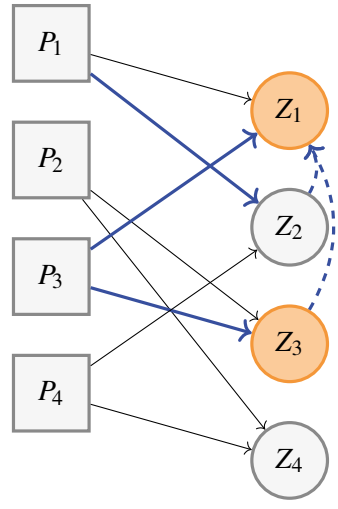

Figure 2.7. **DAG representation of the dependencies in decoded data in an example of a 2×2 block.**
Only the decoding step from the pools to individual genotypes is shown, since the behaviour during decoding is the focus of this chapter. The variable R is not represented, instead the node Z_i is coloured in orange if its value is missing. These choices for the representation are made in order to align the matrix representation of the pooling block with the corresponding DAG. The plain edges highlighted in blue indicate which items are involved in the decoding algorithm for the missing item Z_1. The dashed blue arrows indicate that the value of Z_1 is conditioned on both Z_2 and Z_3. The value for Z_1 is obtained from the genotyping result of P_1 and P_3. Since X_2 (not represented here) also affects the test result for P_1, the observed value Z_2 indirectly affects the missingness status of Z_1. Similarly, the value of the missing variable Z_3 indirectly affects the underlying value of Z_1 because P_3 involves both X_1 and X_3. This example of pooling block illustrates the MNAR mechanism imposed by NORB pooling, as the missingness status of the decoded genotypes depends on observed and on missing variables participating in the same pools.

The examples in Figures 2.4, 2.5, and 2.6 show valid and invalid configurations for X given a pooling block pattern.

2.4 Tailored inference methods for pooled genotype data

In the research related to this thesis, we have investigated the implementation of custom probabilistic decoding methods for NORB genotype pooling for the following reasons:

- In applications in genetics, a deterministic decoding method is proposed for the scenario of detecting the individuals carrying the alternate allele in any variant, regardless of whether one or two copies of the allele are carried [85]. For our genotyping purposes, we need to extend the decoding algorithm so that it returns ternary outcomes — namely two opposite heterozygous genotypes and one heterozygous. Since some genotypes cannot be fully resolved from the pools, the decoded outcomes should preferably be in a form that can represent uncertainty.
- The pooling strategies we propose for genotype data and the related decoding methods should be devised with the understanding that they are intended to be augmented by imputation. For one thing, this means that the output of decoding must be usable as input to the imputation methods. Second, expressing the decoded data as genotype probabilities can reflect small variations that may arise between the outcomes of different strategies, which would not be possible with integer-valued genotypes. This is important because the extent to which the pooling mechanism affects the performance of imputation is unknown.
- Other studies have demonstrated the effectiveness of a probabilistic framework in hybrid genotyping approaches combining pooled sequencing and imputed array data [31, 34].

2.4.1 Likelihood framework for estimating the missing items in pooled data with a NORB design

As the statistical framework that we have used is largely described in Papers I and II, this section only briefly presents some elements of that framework.

Vector notation of the data

The pooling mechanism consists in simulating pooled observations from individual data and resolving the pooled observations. Let us model this mechanism as the data mapping t,

$$\begin{aligned} t\colon \mathscr{X} &\longrightarrow \mathscr{Z} \\ \mathbf{x} &\longmapsto \mathbf{z} \end{aligned}$$

where $\mathbf{x} = (x_1, x_2, ..., x_n)$ and $\mathbf{z} = (z_1, z_2, ..., z_n)$ are vectors of genotypes. The true genotype data at any marker for a sample i is represented by a probability simplex $x_i = [p_{0i}, p_{1i}, p_{2i}]^\top$. The three simplex values represent, in order, the

probability that the genotype is a homozygote for the reference allele, a heterozygote, and a homozygote for the alternate allele. Similarly, the decoded genotype data for the same sample i is denoted by $z_i = [\tilde{p}_{0i}, \tilde{p}_{1i}, \tilde{p}_{2i}]^\top$. Several different sequences $\mathbf{x}$ can lead to the observation of the same pooling pattern ψ, which is decoded into only one sequence $\mathbf{z}$. In this sense, t is a many-to-one mapping and decoding $\mathbf{z}$ into estimated genotype probabilities is an inference problem.

Formulation of the inference problem in NORB pooled data

The inference problem can be decomposed into a series of maximum likelihood estimation problems conditioned on each missing data pattern, i.e. each pooling block pattern ψ. For arbitrarily complex models, it is common to solve likelihood maximisation problems iteratively using Expectation-Maximisation (EM) approaches [26]. For each pattern ψ, we are interested in studying possible inversions of the mapping t to estimate $\mathbf{z}$ as being the most likely sequence of genotype probabilities for $\mathbf{x}$.

2.4.2 Expectation-Maximisation based methods

A detailed description of the EM-based estimation methods we have used can be found in Papers I and II, as well as a few numerical examples.

Therefore, here we only illustrate the general idea for a single iteration (m) in Figures 2.8 and 2.9 and highlight some specific features of our EM-based estimation method for pooled genotype data. Paper II involves several variations of the main steps presented here. Overall, the various EM versions perform rescaling and marginalisation of the expected frequencies of genotypes and/or alleles in different ways.

Expectation step

The expectation step, or E-step, enumerates all data completions of $\mathbf{z}$ for a pooling block with the pattern ψ, as shown in Figure 2.8 (I) and 2.9 (II). Some of the enumerated completions might be invalid in the sense that they map to a decoded vector of genotypes which is inconsistent with ψ. Figure 2.8 (II) gives a few examples of invalid completions for the pooling pattern $\psi = ((2,2,0),(2,2,0))$ in a 4×4 pooling block.

From an algorithmic point of view, the enumeration can be implemented as a dynamic recursion in a ternary tree with n_B levels where each node has a genotype value in $\{0,1,2\}$. The invalid completions correspond to branches in the tree that are pruned, which makes the complexity of the algorithm unpredictable to a certain extent. This dynamic and recursive task poses some computational challenges, not only because of the size of the search space (the tree might have up to 4.3×10^7 terminating leaves), but also because of the irregular length of the branches depending on the pooling pattern considered.

Other more efficient strategies could be chosen for the computation, such as a Forward-Backward algorithm.

The genotype probabilities for each item in the block to be decoded can be initialised to any probability simplex. For each enumerated data completion $\mathbf{x}$, its expected proportion $\mathbb{E}[\mathbf{x}|\mathbf{z};\psi]^{(m)}$ is computed as in Equation 2.3.

$$\mathbb{E}[\mathbf{x}|\mathbf{z};\psi]^{(m)} = \frac{Pr(\mathbf{x}|\mathbf{z};\psi)^{(m)}}{\sum_{\mathbf{x}} Pr(\mathbf{x}|\mathbf{z},\psi)^{(m)}} = \frac{Pr(\mathbf{z}|\mathbf{x};\psi)\,Pr(\mathbf{x})^{(m-1)}}{\sum_{\mathbf{x}} Pr(\mathbf{z}|\mathbf{x};\psi)\,Pr(\mathbf{x})^{(m-1)}} \tag{2.3}$$

In the case of invalid data completion, $Pr(\mathbf{z}|\mathbf{x};\psi) = 0$, otherwise $Pr(\mathbf{z}|\mathbf{x};\psi) = 1$.

$Pr(\mathbf{x})^{(m-1)} = \prod_{b=1}^{B} Pr(x_b)^{(m-1)}$ is the probability of $\mathbf{x}$ computed from the individual posterior probabilities in the iteration $(m-1)$.

$\mathbb{E}[\mathbf{x}|\mathbf{z};\psi]$ is the expected proportion of every valid data completion, given that we observe the pattern ψ.

Maximisation step

The maximisation step, or M-step, calculates for every item in $\mathbf{x}$ the probability of each genotype from the expected frequencies computed in the E-step:

$$(\tilde{p_0},\tilde{p_1},\tilde{p_2})^{(m)} = \frac{\mathbf{x}\,\mathbb{E}[\mathbf{x}|\mathbf{z};\psi]^{(m)}}{\sum_{\mathbf{x}} \mathbf{x}\,\mathbb{E}[\mathbf{x}|\mathbf{z};\psi]^{(m)}} \tag{2.4}$$

where $(\tilde{p_0},\tilde{p_1},\tilde{p_2})^{(m)}$ are vectors of estimated genotype frequencies for all samples in the block in iteration m.

Rescaling step

Consecutively to the usual E- and M-step, we implement rescaling operations as follows:

- First, each item in $(\tilde{p_0},\tilde{p_1},\tilde{p_2})^{(m)}$ is divided by its individual prior and the result is normalised to ensure consistency of the method.
- Second, the probabilities for heterozygotes are explicitly upscaled by a factor of 2. This relates to our representation of a single heterozygous state, while there are actually two distinct heterozygous genotypes. Thus, even with a uniform prior, the heterozygotes should be twice as common. We refer to this effect as *heterozygote degeneracy*.
- Once the convergence criterion is met, a final downscaling of the estimated genotype probabilities is performed. This step is implemented in view of using the genotype probabilities as input in genotype imputation algorithms that internally double the probabilities for the heterozygotes. Downscaling avoids an over-representation of the heterozygous genotypes in imputation.

2.4.3 Iterative point-wise correction of the decoded probabilities with feedback from imputation

Using a likelihood framework for decoding the pooled observations makes possible to inject information about the genotype distribution into the decoding process [31, 34]. In Paper IV, we explore such a scenario, in which these information is featured in the genetic structure of an auxiliary cohort that is assumed to be similar to the population that was pooled. We find that sequential corrections in the decoded data and repeated augmentation with imputation improve the quality of genotype reconstruction.

By implementing an iterative scheme with cycles of 'repooling' and imputation, genetic features such as the LD between markers and similar genetic sequences shared by different samples are indirectly embedded in the decoding procedure. In each cycle and in each pooling block, the 'repooling' algorithm simulates the composite genotype of every pool from the last imputed genotype probabilities. These composite imputed genotypes are used as feedback to adjust the decoded probabilities. A correction is applied to the decoded likelihood whenever the 'repooled' imputed data are inconsistent with the pooled observations, and the corrected data is used as input to run a new iteration of imputation. The notion of consistency we use refers to what alleles are detected or not in the pooled observations and in the 'repooled' imputed data. The likelihoods for positive and negative detection of each allele are computed for the repooled genotypes, assuming that the imputed posterior genotype probabilities are independent from each other. This approach based on a feedback structure and likelihoods can be viewed as a coupling mechanism between the pooled observations and the population-wide haplotype structure available through imputation.

2.4.4 Quality of the decoded genotype probabilities

We showed that the NORB pooled data are MNAR, so that the EM-based methods might produce biased estimates of the unresolved genotypes [65]. Paper II focuses on the nature and the extent of this bias by examining the consistency of the genotype distributions reconstructed based on pooled observations using EM methodology. We use a divergence metric to assess the distributional consistency of the decoded data with respect to the true data. This divergence is a first alternative to evaluate the quality of the genotypes reconstructed with our different decoding strategies.

A second alternative pertains to our findings in Paper IV and can also be linked to a discussion point we raise in Paper II. An appropriate definition of quality in our context might therefore not be how close they are to the ground truth, but rather how much the reconstructed distribution influences accuracy of imputation based on the decoded data. Paper IV suggests that sequential adjustments to the decoded data, if based on adequate external information,

can improve the imputation accuracy without explicitly targeting the reconstruction of the true data distribution in decoding.

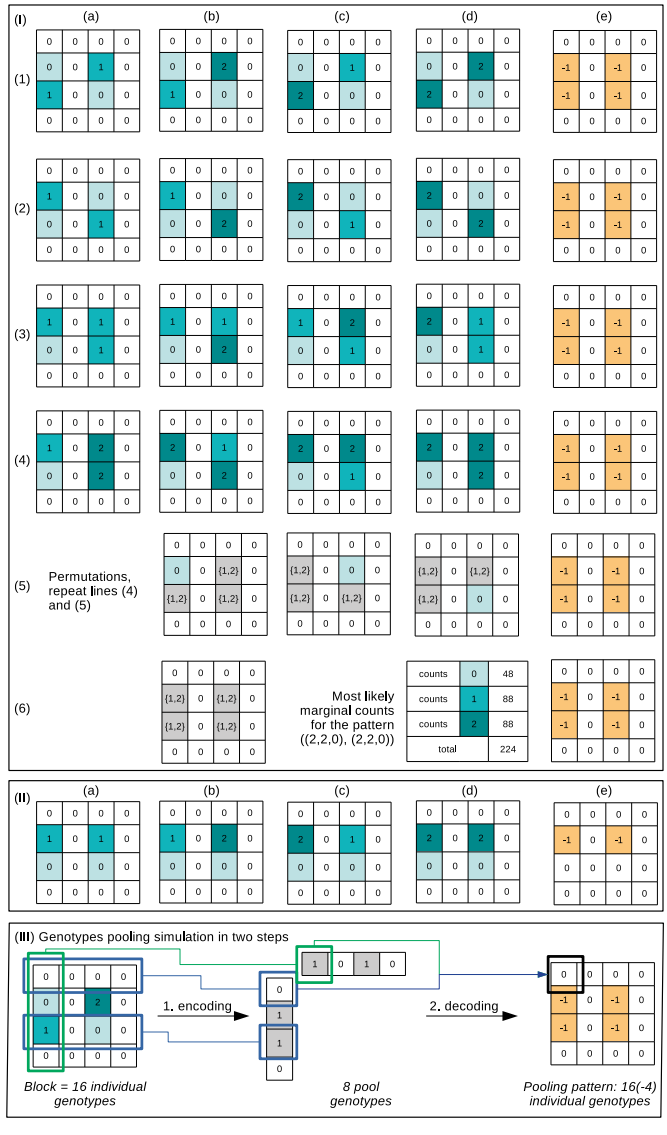

Figure 2.8.

A Maximum-Likelihood method for decoding pooling blocks.

Enumeration example for a 4×4 block with a pattern $\psi = ((2,2,0),(2,2,0))$.

(I) Enumerating the valid layouts compatible with this pattern results in 56 configurations. Over these combinations, the homozygotes having genotype 0 (resp. the heterozygotes 1 and the opposite homozygotes 2) appears 48 times (resp. 88 and 88), such that the estimated genotypes distribution fitted to the layout is $(0.214, 0.393, 0.393)$. This corresponds to a Maximum Marginal Likelihood estimation.

(II) For a given set of genotypes, some permutations result in a genotype vector that is not compatible with the observed pooling pattern ψ.

(III) Simulating pooling consists in a first encoding step which resolves the genotype of the row- and column-pools: 2 rows have genotype 0, 2 have genotype 1, none has genotype 2, and similarly for the column-pools. The second step decodes the pooled data into individual genotypes. After decoding, the 4 missing items are placed on two distinct pairs of row and column, while the other items are decoded homozygotes (genotype 0).

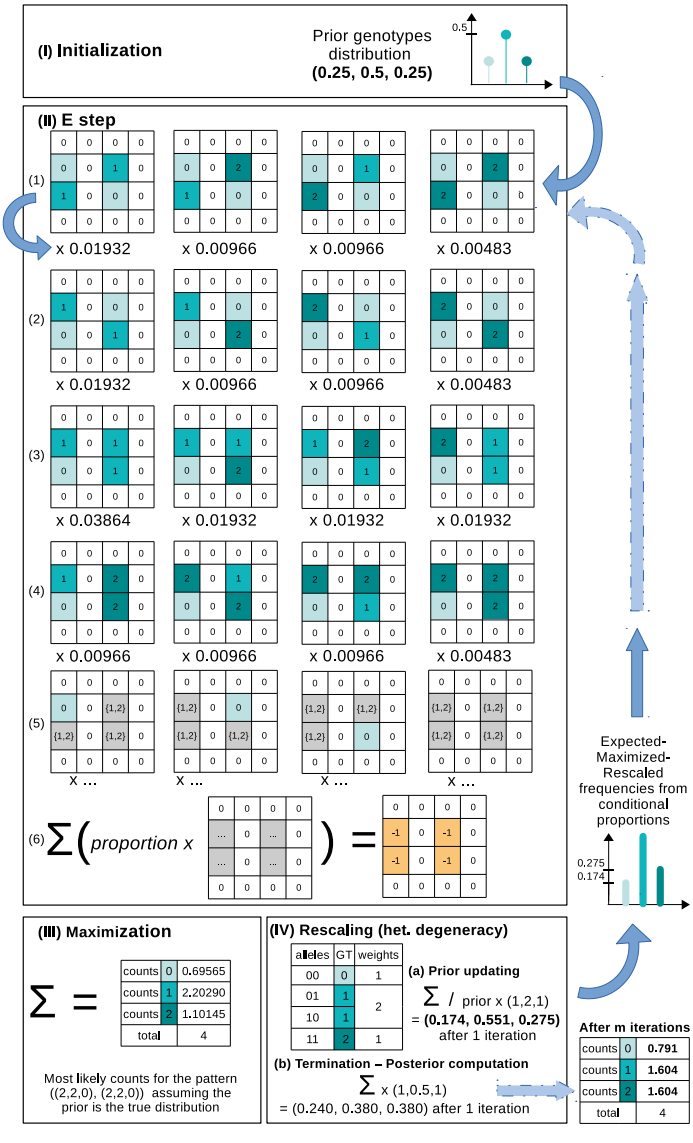

Figure 2.9.

A self-consistent method for resolving ambiguous pooled genotypes with heterozygotes degeneracy.

The figure shows the same pooling block as in Figure 2.8.

(I) The genotype probabilities for any sample in the pooling block are initialized to a prior value of $(0.25, 0.5, 0.25)$.

(II) The enumeration of the valid data completions is executed in the same way as in Figure 2.8. For each valid completion, the prior genotype probabilities are used to compute the likelihood of the given completion. The likelihood of each completion is later used as a weighing factor.

(III) The most likely genotypes counts are computed based on the likelihood of every valid completion.

The second step decodes the pooled data into individual genotypes. After decoding, the 4 missing items are placed on two distinct pairs of row and column, while the other items are decoded homozygotes (genotype 0).

(IV) Rescaling is applied for accounting for heterozygotes degeneracy and layouts collapsing, as well as a final down-scaling step if the computed estimates are to be used in genotype imputation.

3. Statistical genotype imputation for missing markers in large populations

In this chapter, we present statistical computational methods for performing genotype imputation. Imputing the genotype of markers consists in inferring the most likely genotype at these markers when they are missing, based on the known genotype data available for other markers. Data can be missing for different reasons: the DNA material might for example be damaged as with ancient samples, or the genotype data is not reliable after the genotyping technique returned noisy results or of poor quality (e.g. low coverage or low calling rate).

Imputation methods are most commonly used for decreasing the cost of large-scale studies based on the genotype data of markers, e.g. in GWAS. Given a chosen set of markers of interest in a study population, only part of these markers will be assayed with genotyping techniques. For the remaining part which is unassayed, computational methods are used for imputing the data. The best-performing imputation methods have shown high accuracy, however they usually give less accurate results for rare variants. One proposed strategy to improve the results for rare variants is to increase the size of the reference panel. For one thing, this involves higher costs for collecting such data. Very large reference panels may also need additional processing into a specific format that has lower memory requirements, and reading these data may become the computational bottleneck in imputation [11].

In our application, the rare variants are accurately captured regardless of the size of the reference panel, but genotype data for the more common variants are mostly missing due to the pooling technique used. The characteristics of group testing, which were discussed in previous chapters, pose new challenges for current imputation algorithms.

3.1 Introduction

On a general level, the imputation problem can be formulated as resolving ambiguous or unknown genotypes in a study population by using genetic information that is extrinsic to the observed genotypes for the individual to be imputed. The resolution of the genotypes involves various probabilistic predictions [36]. The predictions are derived from different information types available, commonly a set of densely genotyped and subsequently phased individuals serving as as a reference for estimating the unassayed genotypes of

study individuals, and relatedness between the individuals if such data are provided. If a reference genome is available, the set of SNPs that are targeted can be positioned on a so-called genetic map with respect to this genome. A reference genome is obtained from a genome assembly experiment which relies on the sequenced data in a cohort of individuals. For instance, in the *1000 Genomes Project* (1KGP), the genomic data were positioned on maps derived from the GRCh37 and GRCh38 genome assemblies [75, 76] released by the Genome Reference Consortium. GRCh38 corresponds to an improved version of GRCh37 with updated annotations that better covers different regions of the genome. For bread wheat, organisations like the International Wheat Genome Sequencing Consortium (IWGSC) have released a series of improved versions of reference genomes [39] in the past decade. Thanks to the latest sequencing technologies, genome assembly studies in various non-model organisms, e.g. plant species, have been ongoing research in the recent years [60], but high-quality assemblies may still be lacking for some species. In particular, for crop species with polyploid genomes and a high intra-specific diversity, the assembly of a reference genome requires the resolution of polyploid genome sequences, which is more accurate using a pangenome graph rather than single reference genome sequences, but also more difficult to perform. It is also usually more challenging to design reliable microarrays for polyploid organisms and also to carry out imputation in these populations, especially when the imputation method necessitate a genetic map derived from a reliable reference genome.

We focus on population-based methods, designed for dealing with unrelated individuals. There are also family-based methods including pedigree information in the imputation process, but we have not considered them here. They would be worth investigating in further research work associated to breeding, where ancestry and descent information of lines are available over several generations. This can be even more true in many livestock programs than in plant breeding.

While other approaches can be found in the literature, we describe only two groups within the population-based methods that have been dominating in the field of genotype imputation [15, 36]. The first group encompasses the coalescence-based models, such as MaCH and Impute2, as well as a locally developed implementation called *prophaser* [6], closely modelled on MaCH, but tailored for being used with genotype probabilities as input data. The second group is illustrated by the tree-clustering models, such as Beagle. Both approaches are iterative and they have been reported [15, 36] among the best performing ones. MaCH and Impute2, as well as Beagle, have been essentially designed for solving the imputation problem in populations where the genotype data is fully missing for all individuals at some markers, which is equivalent to the genotypes being missing (completely) at random (M(C)AR). To remind the reader, this means that the propensity of the genotypes to be

missing is not dependent on the actual genotype values or any of the observed data.

3.1.1 Definitions and notations

Let us denote Θ being a set of n_h template haplotypes at n_j loci. Depending on the imputation strategy used, Θ is built upon the reference panel and/or the n_i individuals from the study population. For each of the n_i individuals of the study population, H_{ij} $j \in [1, n_j]$ is a pair of haplotypes from Θ at marker j for the i-th individual and we denote $\mathbf{H}_i$ the sequence of n_j haplotypes over all markers. Similarly, $G_{i,j}$ is the genotype (pair of alleles) at marker j for the i-th individual and $\mathbf{G}_i$ the sequence of n_j genotypes. In other words, any study sample i is modelled as a sequence of either haplotype states or genotypes.

3.1.2 Mathematical formulation of the imputation problem

Both population-based models are typically used as iterative statistical methods that yield probabilistic predictions of the genotypes for the missing marker data. For each marker imputed at a locus j for an individual i, the probabilistic predictions calculated for the genotype can be formulated as in Equation 3.1. Commonly, this prediction is discretised as the *best-guess genotype* value [36] and formatted as GT. The GT value is the genotype having the highest probability in the predicted probability tuple. The GT format can be a *phased* genotype if each of the alleles is attributed to a specific haplotype. If not, the genotype is said to be *unphased*.

$$p_{ija} = Pr(G_{ij} = a|\Theta, \mathbf{G}_i),\ a \in \{0,1,2\},\ \sum_a p_{ija} = 1 \tag{3.1}$$

That is, the genotype of any marker at locus j for the individual i is conditioned both on the other haplotypes Θ in the population that are used as templates, and on the genotypes observed at the other loci in the sequence $\mathbf{G}_i$.

3.1.3 Hidden Markov Models for modelling haplotypes and sequences of genotypes

Both coalescent and tree-clustering methods implement Hidden Markov Models (HMMs). A graphical representation of a generic HMM for genotype imputation is given in Figure 3.1.

Using a notation consistent with the classical HMM treatment by Rabiner [61], the HMMs used in imputation models can be characterised as follows:

1. The number of states n_h in the model equals the number of template haplotypes, or the number of pairs of template haplotypes. The hidden state i is denoted s_i in Figure 3.1.

2. There are 2 distinct observation symbols per haplotype i at the locus j, one for each allele e.g. $G_{ij} = 0$ or $G_{ij} = 1$ in Figure 3.1. If the hidden state is a pair of haplotypes, there are 2^2 observation symbols for each hidden state, each of them corresponding to a phased genotype.
3. The transition probability distribution from state s_{i_1} to s_{i_2} is $\mathscr{F} = \{f_{s_{i_1},s_{i_2}}\}$. $\mathscr{F}$ is either explicitly parametrised with a recombination rate ρ as in coalescent models, or implicitly captured through the counts of haplotype clusters when building the tree in the Beagle model. The transition from one haplotype state to another between two consecutive markers mimics a historical recombination event. It is correlated to the LD between markers.
4. The probability of emitting the symbol a_j from the state s_i is $\mathscr{G} = \{g_{s_i}(a_j)\}$. The observed genotypes model possibly erroneous copies of the haplotypes and hence express mutation events. These events are explicitly parametrised in coalescent models with the mutation rate μ.
5. The initial states are determined or randomly assigned based on the observed haplotypes in the population to impute (see examples in the Sections 3.2.3 and 3.3.2). We denote $\mathscr{S}$ the initial distribution of the states.

An HMM model designed for genotype imputation is typically used for solving three problems for any study sample i [43], which are:

Problem 1 Given a sequence of observation, in our case a sequence of genotypes $\mathbf{G}_i$ and the model parameters $(\mathscr{F},\mathscr{G},\mathscr{S})$, the HMM lets one compute the probability of the sequence $Pr(\mathbf{G}_i|\mathscr{F},\mathscr{G},\mathscr{S})$. The computation is executed with the Forward-Backward algorithm which computes the probability of observing $\mathbf{G}_i$ summed over all possible sequences of hidden states e.g. haplotypes.

Problem 2 Given a sequence of genotypes $\mathbf{G}_i$ and the model parameters $(\mathscr{F},\mathscr{G},\mathscr{S})$, the Viterbi algorithm determines the most likely sequence of haplotypes $\mathbf{H}_i$ from which $\mathbf{G}_i$ derives.

Problem 3 The Baum-Welch algorithm adjusts the model parameters $(\mathscr{F},\mathscr{G},\mathscr{S})$ with the intent to maximise $Pr(\mathbf{G}_i|\mathscr{F},\mathscr{G},\mathscr{S})$.

In coalescent models, the most likely sequence of genotypes for each study individual is computed iteratively by using a combination of these three algorithms, and sometimes a random sampling step, at each iteration. The Impute2 model however uses fixed model parameters and hence simplifies the **Problem 3**, whereas MaCH reevaluates the model parameters at each iteration. *prophaser* as used in Paper I executes only one iteration of the MaCH model, directly determining genotype probabilities using the forward-backward algorithm.

Tree-based clustering, in the form implemented in Beagle, is an empirical model determined by the counts of similar segments found across the template haplotypes.

For both the coalescent and the tree-based models, the hidden states underlying the Markov chain of the HMM are defined by single or aggregated

template haplotypes. The strategy for choosing the template haplotypes used as hidden states notably differs between the coalescent and the tree-clustering approaches. An illustrated example is given later in this chapter (Sections 3.2.3 and 3.3.2).

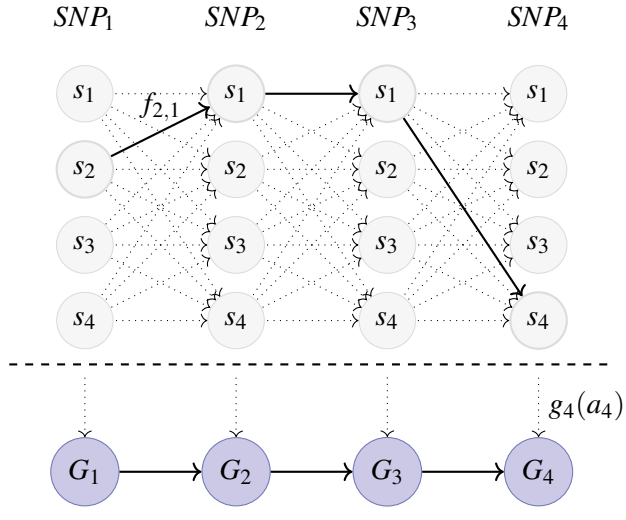

Figure 3.1.

Trellis of the observation sequence (G_1, G_2, G_3, G_4) for an HMM with 4 states.

The thick arrows indicate the most probable transitions. Each state s_i represent a template haplotype. $f_{s_{i_1}, s_{i_2}}$ is the probability to transition from the hidden state s_{i_1} to the hidden state s_{i_2} which depends on the linkage disequilibrium (LD) between the two successive loci. $g_{s_i}(a_j)$ is the probability to emit the symbol a_j from the state s_i. In the case of coalescent models, the recombination and the mutation probabilities are modelled with the explicit parameters ρ and μ. The most likely sequence of states (s_2, s_1, s_1, s_4) can be seen as a mosaic of template haplotypes.

3.1.4 Factors affecting the accuracy of genotype imputation

Studies in various human and plant populations have shown that the choice of an appropriate reference panel with respect to the study population to be imputed has a major impact on the accuracy of the imputed genotypes [51, 88]. Changing the reference panel directly modifies the template haplotypes Θ and their number n_h, which in turn affect the initial distribution of the states $\mathscr{S}$ and the distribution $\mathscr{F}$. In addition, proper parameterisation of the imputation methods is necessary to obtain optimal predictions, particularly if the model was primarily intended for a type of population that differs from the one under study [58]. When used, a recombination map affects the transition probabili-

ties between the hidden states, and the use of an unsuitable map can notably worsen the imputation results. The effective population size may also be modified to account for different ranges of LD in the population under study. For instance, the effective population size used to impute the inbred wheat lines in Papers III and IV is much smaller than the value set in the case of the natural human population in Paper I. Indeed, the population of inbred wheat lines involves fewer generations and recombination events, so that the average LD is larger. In our method *prophaser*, changing the effective population size has an impact on the transition probabilities calculated with the forward-backward algorithm.

HMMs are powerful tools for reducing noise in signals for example, as they have an architecture that effectively handles uncertainty. The nature of imputation HMMs and their associated algorithms makes it reasonable to express uncertainty about the observed genotypes in terms of genotype probabilities. These are implicitly assumed to be independent of the genotype in the other markers, and in other individuals. We are interested in studying the response of imputation HMMs to input data genotyped on a microarray and "perturbed" by pooling. It is also unclear what order of magnitude in the perturbations in the sequence of observations can improve, or degrade, the quality of the imputation. These questions are addressed in the simulation studies we have conducted, and we have aimed to find a suitable strategy for informing imputation with decoded data. That is, the decoded data are intended solely as input to the imputation, and therefore the strategy for calculating genotype probabilities should not be evaluated for its inherent accuracy, but for the accuracy of the resulting imputed genotypes.

For a constant reference panel and a fixed recombination map specific to the example data set studied, we have explored various strategies for decoding pooled genotypes into individual genotype probabilities. The variations in the decoded genotype probabilities computed with different strategies modify the initial sequence of observations $Pr(\mathbf{G}_i|\mathscr{F},\mathscr{G},\mathscr{S})$ for a given individual to be imputed. The initial sequence of observations is used as input to **Problem 1** presented in Section 3.1.3, and the effects of the modifications are propagated to **Problem 2** and **3**.

The first strategy computes unbiased estimates for the decoded genotypes that should approximate the true genotype distribution. The *simpool* algorithm, in its various versions, implements this strategy, which is illustrated in Papers I and II. Paper III also complements the first strategy with simulations in fully homozygous wheat inbred lines and genotypes decoded using maximum likelihood estimation.

The second strategy we have investigated aims to decode the pooled genotypes in a way that ensures consistency between these and the imputed genotypes. Due to its complex structure, it is difficult to assess the exact sensitivity of the HMM to changes in specific genotype probabilities. Therefore, this approach is based on using existing genotype probabilities and gradually

changing them in a direction that promotes results consistent with the observed pool genotypes. This strategy is presented in Paper IV and implemented in the *repool* algorithm.

The minimal examples presented hereafter illustrate the impact of pooling on the initial sampling in the different models . They use an integer representation of genotypes rather than genotype probabilities, in order to simplify the treatment. In many implementations, this is also the form used in part of the processing in order to realize higher efficiency for typical usage modes. Our implementation *prophaser* is a notable exception to this, since it was explicitly intended for uncertain data.

3.2 Coalescent models

3.2.1 The coalescence principle

The models in this family rely on the so called principle of *coalescence* [35, 43] which asserts that the haplotypes in a homogeneous population tend to be similar. The variations found between the haplotypes are explained through a combination of the genetic events of recombination and mutation over time. These events are assumed to be rare over small genetic distances and limited time-spans, there are therefore great similarities between haplotypes in different individuals within a population [15, 73]. MaCH and Impute2 exploit the linkage disequilibrium (LD) between markers for capturing the genetic patterns across haplotypes.

3.2.2 Specific aspects of the coalesent models

For each sample i of the study population, the coalescent models computes the probability of the observation sequence $\mathbf{G}_i$ based on Equation 3.2. Impute2 and MaCH proceed by sampling sequences of states through the trellis of haplotypes as in Figure 3.1.

The hidden states underlying the Markov chain of the HMM are the haplotypes (single haplotypes or haplotype pairs depending on the model) which are selected from a set of template haplotypes. The way this set of template haplotypes is constituted varies with the imputation method used.

$$Pr(\mathbf{G}_i|\Theta,\mu,\rho) = \sum_{\mathbf{H}} Pr(\mathbf{G}_i|\mathbf{H},\mu) \cdot Pr(\mathbf{H}|\Theta,\rho) \tag{3.2}$$

The factor $Pr(\mathbf{G}_i|\mathbf{H},\mu)$ models mutation (μ is an explicit parameter or not depending on the model type) along the Markov chain of hidden states, and the factor $Pr(\mathbf{H}|\Theta,\rho)$ models recombination (whether ρ is explicit or not). Mutation represents a hidden state that emits at a marker a symbol which is

different from the allele in the haplotype. Recombination corresponds to a transition of haplotype between two consecutive markers.

Impute2 uses fixed probabilities of recombination events that are provided in a fine-scale recombination map as LD values between the markers. These values depend on the physical distance between the markers [42, 47, 73]. The distance between the markers is provided in the form of a genetic map that is calculated from a genome assembly.

MaCH reevaluates the recombination and mutation probabilities at each iteration once all the study samples have been processed, based on the outcome of the haplotype sampling process in that iteration.

Selection of the template haplotypes

Impute2 selects the template haplotypes from the reference panel and the study population based on similarity to the individual being phased ('informed selection' of conditioning states) [35, 36]. MaCH randomly selects a subset from the reference and the study population [43]. The subsetting strategy maximises the use of available information while limiting the size of the state space in the Markov chain. Our method *prophaser* is implemented for using all reference and study haplotypes as templates. However, since *prophaser* is executed separately for each study sample in our papers, only the reference haplotypes are used as templates in practice.

Haplotype phasing

In every iteration, the Impute2 algorithm consists of two main steps, haplotype phasing and actual genotype imputation [47]. In the HMM used for phasing, the transition probabilities are the probabilities that the hidden state switches between two consecutive assayed markers (observed genotypes). At the first iteration, the haplotypes in the study population are randomly phased and the initial transition probabilities are equal for all hidden states. Phasing the haplotypes of every study sample is executed in the Impute2 model sampling the most likely state path in a Markov Chain Monte Carlo (MCMC) scheme. The resulting path can be seen as a compound of template haplotypes, therefore the expression "mosaic of haplotypes" is frequently employed [36, 43, 58].

The MaCH model performs stochastic backward path sampling, which differs from the regular Viterbi algorithm that also proceeds backwards, but deterministically by choosing the most likely state. At each locus, MaCH uses the forward probabilities of the possible paths through the templates, weighted by the likelihood of the current estimate, to randomly sample an updated sequence of haplotypes. This technique adjusts the likelihood of the sampled sequence locally at each marker without recomputing the likelihood of the entire path. The final sequence of n_j haplotypes sampled for each study individual is used in its turn as one of the templates in the processing of additional individuals.

Genotype imputation

In the Impute2 model, the genotype imputation step reuses the results of the computations in the phasing step for computing the marginal probability of each genotype $0, 1, 2$ for any missing item. The model assumes that the phased haplotypes were sampled from a population that conforms to Hardy-Weinberg Equilibrium (HWE). The genotype probabilities are derived from the allelic probabilities [35] in the entire population.

MaCH does not directly compute the genotype probabilities at each marker, but the path sampling for each individual is performed in a way such that the sequences of haplotypes are edited consistently with the observed genotypes (**Problem 3**). The genotype probabilities at missing markers are deduced after the last iteration from the counts of sampled genotypes over all iterations.

Complexity and computational performance

Impute2 and MaCH form the HMM hidden states by selecting n_h template haplotypes in both the reference and the study population, such there is a constant number n_h^2 hidden states at each of the n_j diploid markers. Thanks to a memory-saving technique implemented in the forward-backward algorithm, both methods have a memory complexity $\mathcal{O}(\sqrt{n_j})$ for each individual. The time complexity grows linearly as the size of the study population and quadratically with the number of template haplotypes [35]. Several papers point out computational time issues with MaCH [15, 36, 57] when compared to the other methods mentioned. One reason is the reevaluation of the crossover and the mutation rate parameters after each iteration.

By contrast, Beagle operates a dimension reduction of the hidden states space thanks to its clustering approach, which has been shown to be particularly efficient when imputing large data sets. The successive releases for Beagle have improved the software performance in this direction [10, 11, 12, 14, 15].

3.2.3 Minimal examples of phasing and imputation in randomly missing and pooled genotype data

The illustrations presented in this section are based on the example used by Howie and Marchini [36, 47].

The reference panel consists of phased haplotypes from individuals. Each haplotype is a sequence of alleles at the markers of interest, inherited from the mother or the father.

The study sample consists of genotypes with sparse data at the same markers, where the haplotypes are unphased. Let us define two marker sets as follows:

- The set of markers $\mathcal{T}$ which consists of markers for which the genotypes are known in both the reference panel and the study population,

- The set of markers $\mathscr{U}$ which consists of markers for which the genotypes are assayed in the reference panel only and missing in the study population.

Figures 3.2 and 3.4 illustrate the definitions for the haplotypes and the marker sets. We consider the examples of two different study populations at the same loci:

1. A study population where the genotype data is missing fully at random. Whenever a marker is assayed, the genotypes are known for all samples, and conversely when a marker is unassayed, the genotype data is missing for all study samples (M(C)AR data).
2. A study population where the genotype data is missing due to a NORB pooling process. The markers are likely to be missing for only some samples, or entirely missing at common variants (MNAR data).

The HMM employed for phasing uses haplotypes from the reference as panel as well as those currently in $\mathscr{T}$ as templates. For simplicity, our figures will show template haplotypes chosen exclusively from the reference population. Figure 3.3 shows the phased haplotypes of three study samples after one iteration of the phasing-imputation algorithm with M(C)AR data, as well as the resulting imputed genotype for one sample.

If prior genotype probabilities are provided for any missing genotype, they are specified with the factor $Pr(\mathbf{G_i}|H, G, \mu)$ in Equation 3.2. The prior genotype probabilities affect the phasing step and the resulting mosaic of haplotypes. In Paper I, we have investigated how pattern-adaptive estimates of the genotype probabilities in pooled data can improve the accuracy of the phasing step and consequently benefit genotype imputation.

3.3 Tree-based haplotype clusters models

Beagle is a prominent implementation of a tree-based haplotype clustering approach for imputation and is commonly used in the literature. It has been developed and improved by Browning and Browning since 2006 [9, 10, 11, 12, 13, 14, 15]. The different versions of Beagle have shown competitive accuracy and computational performance in various settings, including very large data sets. The software has been tested on human [36] as well as on animal and crop species genomic data [59]. Thanks to the clustering approach that reduces the state space, Beagle has been shown being particularly efficient on large data sets and the successive releases have improved the method performance in this direction. Browning and Browning have adopted an alternative approach to coalescence for exploiting sequence variations that feature a given genetic structure in a population.

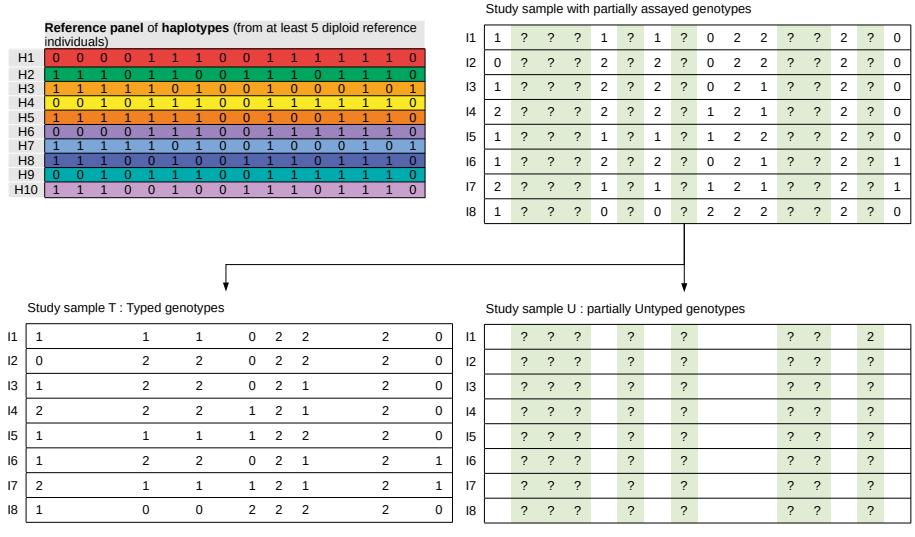

Figure 3.2.

Data sets involved in phasing and imputation for a coalescent model with M(C)AR data.

The data sets consist of a 10 haplotype reference panel and an 8 individual study population. The marker genotypes at 16 loci are represented as integers being the sum of their two alleles for readability. In this particular examples, the template haplotypes come from the reference panel only. The study population is split into a set $\mathscr{T}$ with assayed genotypes and the complementary set $\mathscr{U}$ with unassayed genotypes. With M(C)AR genotype data, the markers are either fully missing for all study individuals, or fully assayed.

3.3.1 Specific aspects of the Beagle model

Construction of the template haplotypes

At each iteration of Beagle, the algorithm includes a preliminary model-building step which uses all haplotypes available in the reference panel and the study population. More recent versions of Beagle implement an iterative weighing of the reference vs. the study haplotypes, such that the reference panel affects the model building more in the initial iterations [10]. The model-building step consists in fitting an HMM with n_j levels to the observed haplotype data. The levels correspond to an ordered sequence of n_j markers. The resulting model that is built can be described as a Variable-Length Markov Chain (VLMC) where the number of template haplotypes that condition phasing and imputation varies at each marker. This feature is a notable difference to the coalescent models where the number of template haplotypes used is constant along the sequence of markers.

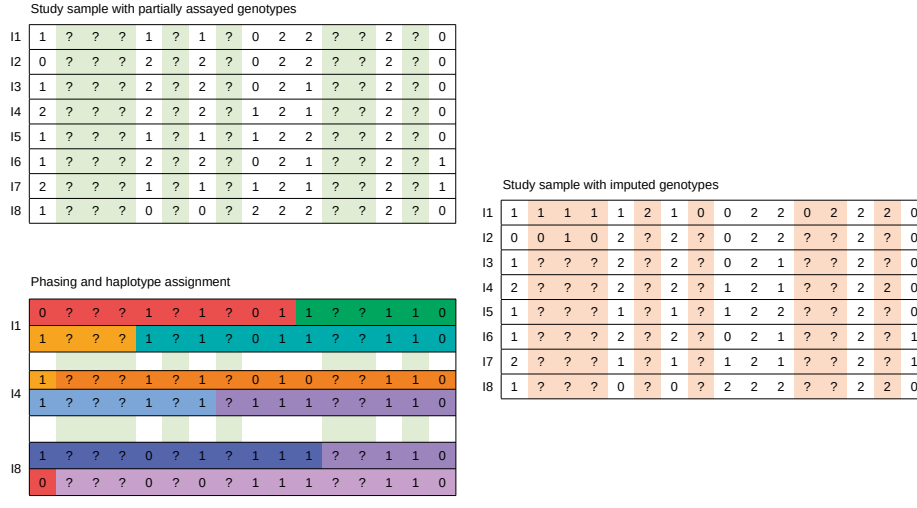

Figure 3.3.

Mosaic of haplotypes from phasing and imputation for a coalescent model with M(C)AR data.

The phasing step computes the most likely pair of mosaic haplotypes for any study sample, based on the template haplotypes (only the reference panel here) and the assayed genotypes. At each locus, the likelihood of every possible pair of haplotypes is computed, which results in $n_h^2 \times n_j = 10^2 \times 16 = 1600$ operations for every study sample. The missing genotypes are imputed as the most likely emitted symbol from the phased haplotypes. The imputed genotypes are represented as the sum of the alleles that are carried by the two haplotypes at the locus e.g. I_1 has genotype $1 = 0 + 1$ for the three first imputed markers since the red segment of haplotype carries the allele 0 at these loci, and the orange haplotype carries the allele 1.

A minimal example is provided in the Section 3.3.2 of this chapter.

At each level j of the tree, the child nodes at level $j+1$ are derived by splitting the observed haplotypes according to their alleles at the current marker. For biallelic markers, any node will have up to two children, depending on what alleles are actually present in the considered haplotypes at marker $j+1$. The tree is extended at each locus such that two loci are connected by an edge.

After processing the last marker at locus n_j, the edges of the tree are weighted by the number of observed haplotypes passing through them. Every template haplotype initially has a unit weight [10]. At initialisation, all haplotypes available in the reference panel and in the study population are used. Nodes are merged at each level of the tree according to a threshold calculated from downstream haplotype frequencies [9, 13, 14].

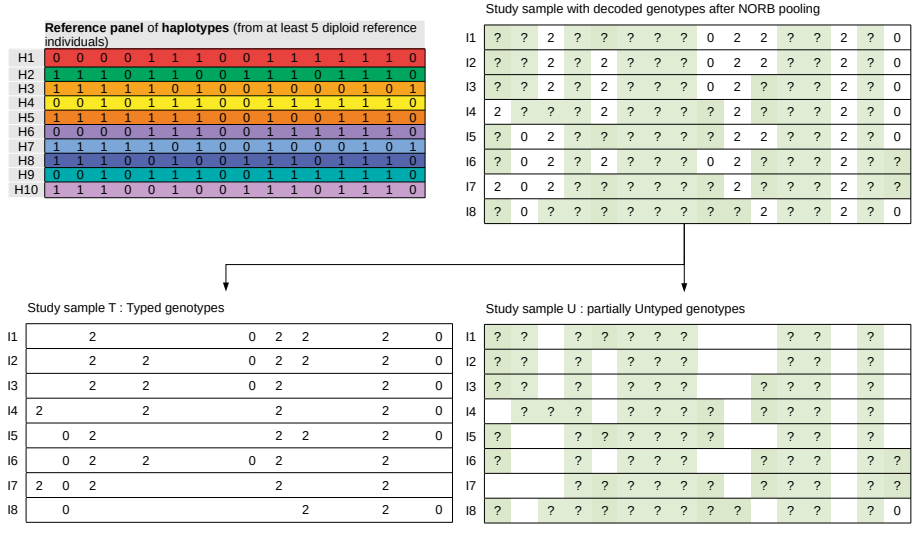

Figure 3.4.

Data sets involved in phasing and imputation for a coalescent model with MNAR data.

The data sets are the reference panel of 10 haplotypes and study population of 8 individuals. The marker genotypes at 16 loci are represented as integers being the sum of their two alleles for readability. In this particular example, the template haplotypes come from the reference panel only. The study population is decoded from a pooled genotype testing, with the same split as used in Figure 3.2. With MNAR genotype data, some markers can be missing for only a subset of individuals, partially dependent on their actual genotype, something that affects the results of phasing imputation.

The merging process results in haplotypes that are clustered based on a frequency criterion of the allele sequences [10]. Node mergers will occur depending on the linkage disequilibrium between successive loci, so that the number of nodes collapsed together locally increases with the linkage disequilibrium [12, 15].

Haplotype phasing and genotype imputation

Beagle can perform both phasing and imputation simultaneously, but phasing can also be done beforehand, either with Beagle, or with other software, e.g. Impute2. For each target individual, phasing is done by sampling the most likely haplotypes with the Viterbi algorithm from the clustered tree, conditioned on the observed genotypes. Missing alleles are randomly imputed according to the observed allele frequencies, which are themselves derived from the haplotype estimates. The newly sampled haplotypes are used as es-

timates in the next iteration of the algorithm for updating the haplotype tree. The genotype probabilities at each locus are eventually computed from the last estimated tree.

By applying merging, weighting and pruning in the successive trees, Beagle captures the population-specific diversity through the haplotype patterns, without explicitly modeling recombination or mutation events as sources of genetic variation [15].

Complexity and computational performance
The number of template haplotypes obtained with clustering is less than the initial number of haplotypes in the reference panel and the study population. Therefore, the size of the state space of the HMM used for imputation is decreased, which is a key factor of the computational efficiency of Beagle in terms of memory as well as time consumption.

3.3.2 Minimal examples of a leveled HMM from M(C)AR and MNAR data

The reference panel and the study population are identical to the examples previously shown for the coalescent models in order to facilitate comparisons between these two families of imputation models. Figures 3.5 and 3.6 show the initiation of the model building step in the case of imputation of M(C)AR data, in accordance with the model by Browning and Browning [10, 13]. Figures 3.7 and 3.8 contain the corresponding illustrations for the MNAR case of decoding pooled NORB data.

Genotype data missing fulling at random: M(C)AR data
This example corresponds to the classical imputation scenario from Paper I, where it is compared against imputation of decoded pooled data.

After sampling alleles at unknown markers and randomly phasing the genotypes, the reference panel and the study population would correspond to the state shown in Figure 3.5. The tree shown in Figure 3.6 is derived from the counts presented in Table 3.1.

Beagle is designed to be used with a large number of haplotypes (several hundred), so that the clustering model has sufficient statistical power [15], especially for imputation of rare variants. In the Figures accompanying this section, the number of template haplotypes is kept small for the sake of the example.

Genotype data missing not at random: pooled data
This example corresponds to the joint pooling and imputation scenario studied in Paper I.

After sampling alleles at unknown markers and randomly phasing the genotypes, the reference panel and the study population would correspond to the

Reference panel of **haplotypes** (from at least 5 diploid reference individuals)

H1	0	0	0	0	1	1	1	0	0	1	1	1	1	1	1	0
H2	1	1	1	0	1	1	0	0	1	1	1	0	1	1	1	0
H3	1	1	1	1	1	0	1	0	0	1	0	0	0	1	0	1
H4	0	0	1	0	1	1	1	0	0	1	1	1	1	1	1	0
H5	1	1	1	1	1	1	1	0	0	1	0	0	1	1	1	0
H6	0	0	0	0	1	1	1	0	0	1	1	1	1	1	1	0
H7	1	1	1	1	1	0	1	0	0	1	0	0	0	1	0	1
H8	1	1	1	0	0	1	0	0	1	1	1	0	1	1	1	0
H9	0	0	1	0	1	1	1	0	0	1	1	1	1	1	1	0
H10	1	1	1	0	0	1	0	0	1	1	1	0	1	1	1	0

Study sample with randomly phased haplotypes and alleles sampled

I1	1	0	1	0	1	?	0	?	0	1	1	?	?	1	?	0
	0	1	1	0	0	?	1	?	0	1	1	?	?	1	?	0
I2	0	1	1	0	1	?	1	?	0	1	1	?	?	1	?	0
	0	1	1	0	1	?	1	?	0	1	1	?	?	1	?	0
I3	1	0	0	1	1	?	1	?	0	1	0	?	?	1	?	0
	0	0	1	0	1	?	1	?	0	1	1	?	?	1	?	0
I4	1	0	1	0	1	?	1	?	1	1	0	?	?	1	1	0
	1	0	1	1	1	?	1	?	0	1	1	?	?	1	1	0
I5	1	0	1	0	0	?	1	?	1	1	1	?	?	1	?	0
	0	1	1	1	1	?	0	?	0	1	1	?	?	1	?	0
I6	0	0	1	0	1	?	1	?	0	1	1	?	?	1	?	0
	1	0	1	0	1	?	1	?	0	1	0	?	?	1	?	1
I7	1	1	0	0	1	?	1	?	1	1	0	?	?	1	?	0
	1	0	1	1	0	?	0	?	0	1	1	?	?	1	?	1
I8	1	1	1	1	0	?	0	?	1	1	1	?	?	1	?	0
	0	0	0	0	0	?	0	?	1	1	1	?	?	1	?	0

Figure 3.5.

Example of initiation of the VLMC with sparse M(C)AR data.
The unassayed genotypes in the study population to be imputed were randomly phased and the alleles chosen proportionally to the observed allele frequency at each marker. For instance, at the second marker (2nd column of the reference panel and the study population), the genotypes are fully unassayed. The observed frequency of allele 0 is the one observed in the reference panel only, which is equal to $\frac{4}{10} = 0.4$ (0.6). In the study population, the 16 unknown alleles are randomly assigned in these proportions, that is to say $16 \times 0.4 \sim 6$ haplotypes carry allele 0.

state shown in Figure 3.7. The tree representation is shown in Figure 3.8, based on the counts in Table 3.2.

Pooling notably modifies the genotype frequencies in the decoded data relative to the true frequencies. In the case of *prophaser*, these modifications affects the computations performed in the HMM to solve the **Problems 1, 2, and 3**. In the case of Beagle, modifying and discretising genotype frequencies affects the model building step and subsequent haplotype sampling. The number of haplotypes and the length of marker sequences considered in the examples are too small to fully demonstrate the effect of pooling on the node merging step. Nevertheless, the trees shown in Figures 3.6 and 3.8 indicate clear variations in the counts of haplotypes, resulting in VLMC with very different structures.

Haplotype	Count
0000	3
0001	0
0010	4
0011	0
0100	0
0101	0
0110	3
0111	1
1000	0
1001	1
1010	4
1011	2
1100	1
1101	0
1110	3
1111	4

Haplotype counts in M(C)AR data.

Table 3.1. *The counts are obtained after completing the missing data based on the observed allele frequencies at each marker.*

Haplotype	Count
0000	2
0001	0
0010	3
0011	1
0100	0
0101	0
0110	0
0111	1
1000	0
1001	0
1010	9
1011	3
1100	0
1101	0
1110	4
1111	3

Haplotype counts in MNAR data.

Table 3.2. *The counts are obtained after completing the missing data based on the observed allele frequencies at each marker.*

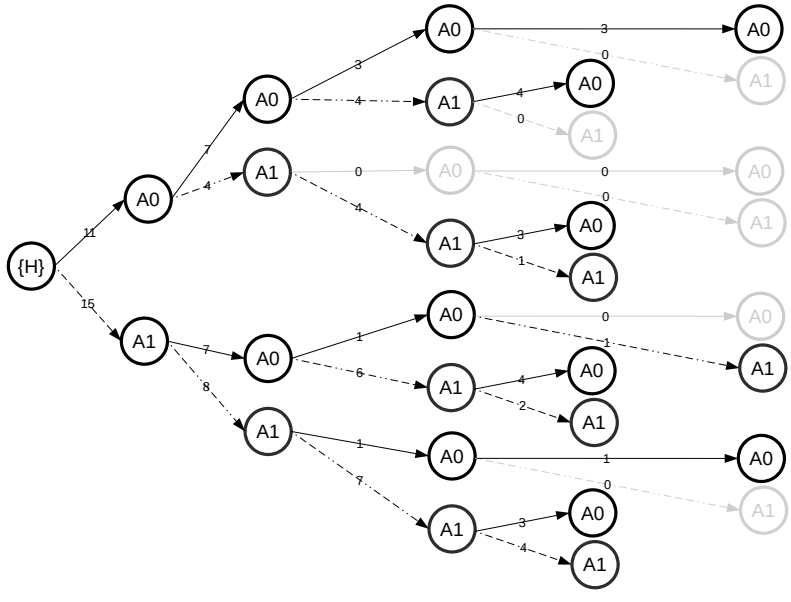

Figure 3.6. **Example of VLMC with haplotypes from sparse M(C)AR data.** The tree is formed from haplotype counts for the 4 first markers in Figure 3.5. The root of the tree, $\{H\}$, is not a marker. $A0$ represent the allele 0 and $A1$ the allele 1. Grey nodes and branches indicate that the allele sequences that were not observed in the available set of haplotypes.

3.4 Conclusion

In this chapter, the illustrated examples with a coalescent model in section 3.2.3, and the Beagle model in section 3.3.2, show the effect of MNAR genotype data on haplotype phasing and genotype imputation, compared to M(C)AR genotype data in usual imputation settings. Paper I presents larger simulations of the MNAR and M(C)AR scenarios in a human population and investigate more thoroughly the performance of both Beagle and the coalescent method *prophaser* with pooled decoded genotype probabilities. Papers III and IV explore the MNAR scenario in inbred lines of wheat.

We want to emphasise the importance of using a likelihood framework to support genotype imputation in pooled experiments. In particular, Paper IV demonstrates that a probabilistic formulation of the decoded genotypes from the pools and an appropriate computational strategy can overcome the issue of biased genotype frequencies when the data are MNAR due to pooling. How much the pooled genotype frequencies are biased relative to the true ones depends on the allele frequency at the markers. This relationship is not linear, but rather related to the hypergeometric distribution [17]. By analogy with signal processing and error-correcting codes in message passing, the likelihood

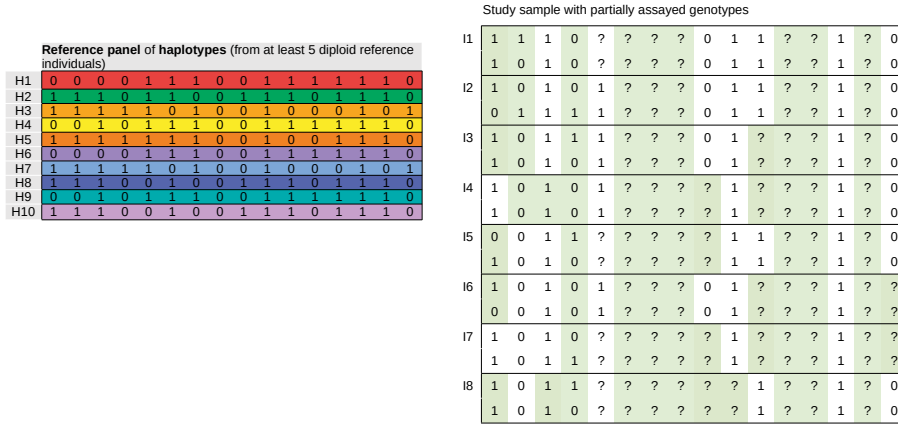

Figure 3.7.

Example of initiation of the VLMC with sparse MNAR data.
The unassayed genotypes in the study population to be imputed were randomly phased and the alleles chosen proportionally to the observed allele frequency at each marker. For instance, at the second marker (2nd column of the reference panel and the study population), the genotypes are partially unassayed. The observed frequency of the allele 0 is the one observed in the reference panel and for 8 haplotypes from the study population, which is equal to $\frac{12}{18} \sim 0.7$. In the study population, the 16 unknown alleles are randomly assigned in these proportions, that is to say $8 \times 0.7 \sim 6$ haplotypes carry the allele 0. As a result, the allelic proportions are notably different relative the ones in Figure 3.6.

framework treats the genotypes as "soft" inputs and outputs. This approach is computationally more costly than using integer genotypes, but it is also more flexible and allows to explore the effect of small changes in the prior genotype probabilities that are passed to the imputation HMMs. We find that the coalescent model *prophaser* is particularly responsive to slight variations in the prior genotype probabilities. This indicates a good sensitivity to the likelihood framework, which is a suitable property in our case. Preliminary to phasing and imputation, Beagle first converts the genotype likelihood to integer-valued genotypes, which cancels out small adjustments in the genotype frequencies resulting from the different decoding strategies. We believe that this discretisation of the prior genotypes may explain the low responsiveness of the Beagle model to the adaptive decoding strategies, and thus makes it less suitable for imputing pooled genotypes.

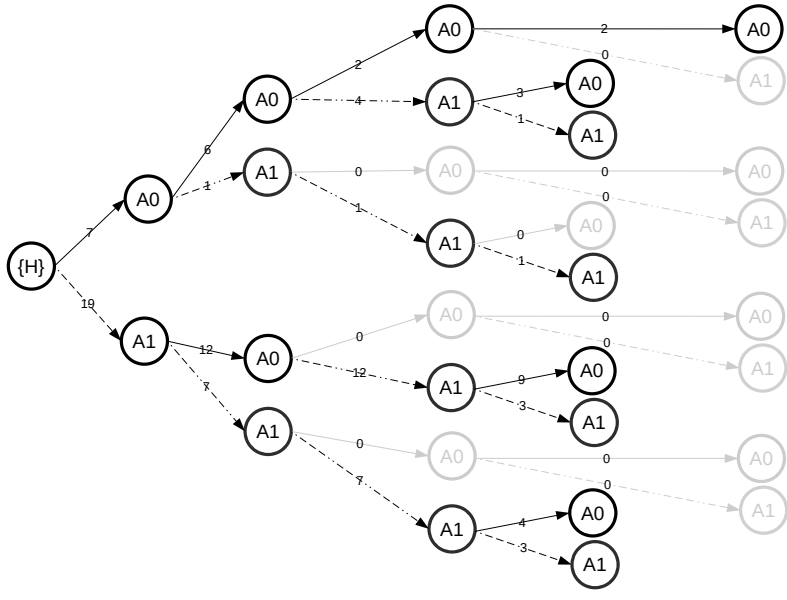

Figure 3.8.

Example of VLMC with haplotypes from sparse MNAR data.
The tree is formed from haplotypes counts at the 4 first markers in Figure 3.7. The root of the tree $\{H\}$ is not a marker. $A0$ represent the allele 0 and $A1$ the allele 1. Grey nodes and branches indicate that the allele sequences that were not observed in the available set of haplotypes. The underlying haplotypes supporting the tree include the specific realisations sampled from the pooled study population. Because of the different allelic proportions at each marker, the tree of haplotypes is looking different than the tree form M(C)AR data. Some haplotypes e.g. 0110 are missing compared to the previous example in Figure 3.5, while other ones are over represented, e.g. the haplotype 1010. This might have a significant impact on the later node merging step and notably modify the template haplotypes used for imputation, which in turn will affect the accuracy of the imputation results.

In addition to its versatility, the likelihood framework is also the key to optimally combining pooling and imputation, so that we can fully exploit the complementarity between these techniques. Indeed, expressing genotypes in terms of likelihood underpins the coupling mechanism that we study in Paper IV. The feedback structure that we implement re-informs the decoded data with genetic information gained at the population level through imputation. Standalone decoding, as we implement it, is not able to take advantage of the intra-individual genetic information contained in the genetic map, nor is it designed to incorporate the inter-individual genetic variation captured by imputation with the reference panel. Typically, accurate inference in MNAR data can be obtained if the distribution of the missing data can be incorporated

in the model [77]. By indirectly embedding the genotype frequencies in the population into decoding, we believe that a coupled model can specify the missingness mechanism correctly and thus achieves more accurate inference. Figure 3.9 shows the improvement in genotyping accuracy through a series of cycles in a setup implemented in Paper IV.

Usually, genotype imputation with larger reference panels produce more accurate results. However, some studies have nuanced this and have demonstrated that the positive effect of the reference panel on the quality of imputation is the strongest when the panel contains a subset of individuals which are structurally similar to the study population to be imputed [15, 57]. Papers III and IV suggest that accurate imputation can be performed even with a small number of reference individuals, if these individuals suffice for explaining all the genetic variation encountered in the study samples. In our case, the set of only 16 founders covers all haplotype variation in the lines in the study population.

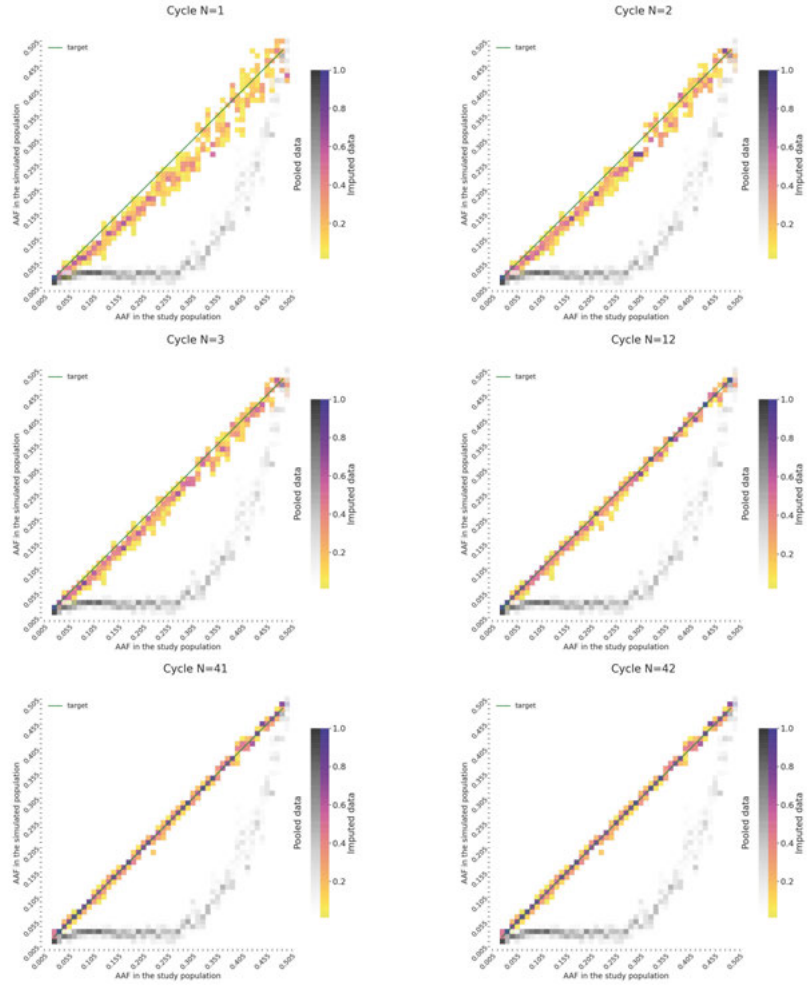

Figure 3.9. **Improvement in the predicted genotypes through the cycles 1, 2, 3, 12, 41, and 42 with a coupled model of corrected decoding and imputation.**

The improvement is represented as the proportions of markers per AAF bin in the pooled and imputed population in relation to the AAF bin in the true study population. The simulated population corresponds on the one hand to the pooled and decoded study population (in grey scale), on the other hand to the imputed study population with prophaser (in colour). The target line represents represents the ideal situation, where there is an exact correspondence between the AAF of the markers in the true study population and in the simulated population. The darkest colour on the heatmap, near the origin of the plot, is close to the target line, which indicates that a high proportion of markers with low AAF (AAF below 5%) are decoded at the correct frequency. The yellow shades, on the other hand, indicate a lower concentration of markers, and the increased spread of the coloured cells indicates that the that the imputation accuracy has a higher variance. The corrections in the decoded data in each cycle are too small to be rendered as visible differences in the subfigures, such that the pooled data (in grey) appears to be unchanged. The corrections are nonetheless large enough to positively affect the imputation method. The imputed genotypes translate to allelic frequencies that are very close to the true frequencies (closer to the target line) for almost all markers since the heatmap becomes narrower and darker. The largest improvements are observed through the first cycles, whereas almost no improvements are visible in the last cycles, which can be interpreted as a convergence phenomenon.

4. Conclusion and outlooks

We introduce genotyping as a method for collecting data about genetic markers and the relevance of these data to research fields such as medical genomics and plant breeding. In both domains, the cumulative cost for testing large cohorts of samples can be a barrier to the feasibility of some studies. The usual techniques for reducing the cost of genotyping have the drawback of often sacrificing valuable genetic information featured in uncommon variations. Using two examples of very different populations, we have explored how a pooling technique augmented by imputation can produce accurate genotype data at limited cost, irrespective of the frequency of the genetic variations.

Pooling is a testing technique that has two major advantages in genotyping applications. For one thing, testing pools rather than individual samples decreases the expenses for genotyping. Second, a pooled setup can capture the genotypes in rare variants with high precision. Nonetheless, for more common variants, the pooled observations are most frequently not possible to resolve into determined genotypes. The resolution power depends, among other things, on the pooling design that is used and how to construct it optimally. We have not explicitly covered those questions within this thesis. For concrete applications, one will need to find a suitable strategy for pool assignment, where practical concerns can lead to individuals sampled from the same location at the same time being tested together. Simultaneous testing of the pools is referred to as nonadaptive group testing. Using a different strategy for pool assignment might influence performance, for instance larger pool sizes would increase the savings but also decrease the decoding robustness of the design.

In our work, we choose a specific pooling design and investigate the decoding problem, i.e. how to reconstruct the individual genotypes from the pooled observations. We have been particularly interested in how to infer the indeterminate genotypes that follow nonrandom missingness patterns and thus challenge the usual inference techniques for missing data. Within a likelihood framework, we have proposed different strategies to decode the pools into genotype probabilities and simulate these strategies in the example populations. The reconstruction methods we suggest are tailored for genotype data that may be ternary-valued. They also take into account the unusual missingness patterns, and produce outputs that are compatible with a usage in imputation. We argued that, beyond considerations about their computational performance, the decoding strategies can be evaluated in terms of the quality of the reconstructed genotypes through two main angles. The first angle favours the

consistency of the reconstructed genotypes with the true data and the second angle focuses on how well the reconstructed data can inform imputation and increase its accuracy, regardless of the consistency with the true genotypes.

We briefly illustrate how decoded the decoded data is handled by two currently available imputation methods, which have been primarily developed for augmenting low-density data assayed on SNP arrays to high density, that is a scenario where genotypes are randomly missing. We demonstrate that the specific dependencies in the reconstructed genotype data poses some difficulties for imputation, both with coalescent and clustered tree methods. Nonetheless, within a likelihood framework and with an appropriate model, such as a coalescent method, imputation can perform very well. The complementary nature of pooling to imputation especially improves the genotyping accuracy of the rare variants.

Our simulations assume error-free genotyping, which is not encountered in practice with SNP arrays and noise in the measurements should be expected [84]. It would be ideal to apply our approach to the data of actual pools, and while such efforts were planned, they failed to materialise within a timeframe making it possible to include such studies within the thesis. However, we conducted preliminary experiments of imputation from pooled data, in which we added artificial noise with varying intensity to model possible genotyping inaccuracies (unpublished research). Noise was simulated for each pool and at each marker in the pooling step using a procedure that involved random sampling from a continuous uniform distribution. If the sampled value exceeded the calling rate of the marker, the simulated composite genotype of the pool was altered by randomly flipping one of the alleles, for instance the genotype 0 was changed to 1. The noise intensity would be adjusted and raised by narrowing the support of the probability density function of the continuous uniform distribution. The lower bound of the interval was set to be equal to the noise intensity and the upper bound equal to the call rate. In the subsequent decoding step, pattern inconsistency between overlapping pools may arise because of the noise added. In this case, the decoded genotype of every sample in the affected pools was assigned the genotype probabilities $(0.33, 0.33, 0.33)$. We observed that reasonable levels of this type of noise do not significantly degrade the accuracy of imputation, which indicates the robustness of imputation methods to such a perturbation.

It would be possible to actively model perturbations in the decoding method. For instance, the algorithm *repool* presented in Paper IV assumes that all samples in a pool equally contribute to its allelic composition and that the allele detection is sensitive enough to be reliable even at low allelic dosages. This becomes a more crucial issue in pooled genotyping by sequencing, where it would not be possible to guarantee that each sample in a pool has generated the same number of reads. Consequently, the pooled reads might over- or undersample one of the alleles, and the probabilistic model used for decoding should take this uncertainty into account.

In Chapter 1, we mention sequencing technologies as an alternative method to SNP arrays for collecting genotype data. During the thesis work, genotyping by sequencing techniques have matured significantly, and are now a cost-effective option to arrays in many cases. Testing and developing our pooling strategy for the genotype data of SNPs obtained with the GBS technology would be relevant with respect to the demand of the plant breeding industry, as suggested for example by Technow and Gerke [78]. The decoding method may need to accommodate possible variations in sequencing coverage and other sources of inaccuracies that are specific to GBS techniques. We have noted that some studies propose approaches that explicitly model pooling noise, uneven coverage, and sequencing errors [31]. These results would provide a relevant basis for comparing extensions of our approach to noisy data.

We have established a partnership with a plant breeding company and we are optimistic about the possibility of applying our approach to their data, which will enable us to explore data from actual pooling experiments in a GBS setting. As the model of pooling augmented by imputation that we propose for cost-effective determination of genotypes is flexible, we believe it may be relevant to other fields beyond biomedical research and plant breeding. We have not used any pedigree data, however, if need be, the likelihood framework would make it possible to account for such information. For example, investigating the feasibility and the performance of our model in animal populations for selection and breeding purposes could provide valuable insights. The feedback mechanism proposed in Paper IV could especially fruitful in that setting, since pedigree-based imputation is commonplace in animal breeding. With pedigree information available for each individual in a pool, the feedback signal to the decoding will be more specific.

5. Summary of papers

Paper I

In Paper I, we conduct a study comparing two scenarios of genotype imputation in a study population sampled from the *1000 Genomes Project*.

The first scenario simulates a situation in which the data set to be imputed consists of markers that are either fully assayed in the study population to be genotyped, or fully missing for all samples. In this usual setup for genotype imputation, the genotype data are missing at random (MAR data). Existing methods for genotype imputation, such as the coalescent models of MaCH or Impute2, as well as the Beagle model, have been developed for a usage in this scenario, and they have shown very good accuracy and computational performance. The second scenario simulates genotype pooling with a 4×4 NORB design in the study population in a first step, followed by genotype imputation in a second step. The genotype data in this setup are not randomly missing, which can affect the performance of imputation.

We propose two new tools to address the particularities of the missing data in the case of pooling. First, a self-consistent iterative algorithm (*simpool*) to reconstruct the genotypes from the pooled observations. The reconstruction consists in inferring the most likely genotype, expressed as genotype probabilities, of any missing item in a pooling block. Second, an extended coalescent method (*prophaser*), which is able to make use of the estimates computed by *simpool* for improving the accuracy of imputation with pooled data.

In both scenarios, we evaluate the accuracy of imputation performed with the Beagle software and with *prophaser*. While imputation in the usual settings performs best over all markers, we find that the pooled strategy outperforms classical imputation for genotyping the rare variants.

Contributions

The second author developed the initial version of *prophaser* and provided advice on its use. In addition to making changes in the code of *prophaser* to improve its performance, the last author conceived the study and provided guidance on the design of the experiments. The author of this thesis developed the pipeline for simulating pooling, collaborated to setting up and running the experiments, and conducted the overall analysis. All authors contributed to the conclusions, as well as they edited and proofread the manuscript.

Paper II

As the inference strategies for MNAR data are not guaranteed to produce unbiased estimates, Paper II proposes to investigate the consistency of the genotype distribution reconstructed with different versions of the *simpool* algorithm. We evaluate the distributional consistency based on a divergence criterion. The new insights provided by this study allow us to improve the original algorithm, so that the later versions of *simpool* produce genotype probabilities that are more consistent with the true genotype distribution. However, a systematic bias remains in the reconstructed genotype distribution. Therefore, the quality of the reconstructed data should be rather interpreted with respect to the accuracy of imputation from the reconstructed genotypes, which is performed in a subsequent step.

Contributions

The author of this thesis designed the study, implemented the methods and executed the experiments, and drafted the manuscript. The results and their interpretation and the ideas underlying the conclusions were discussed jointly by both authors.

Note:

The acronym NORB used for nonoverlapping repeated block that describes the pooling design we use in Paper II is not an established terminology, but rather corresponds to the characteristics of the design which are defined in this thesis. Formally, this pooling scheme falls into the category of the shifted transversal designs.

Paper III

As we noticed the interest of the plant breeding scientists in cost-effective methods for large-scale genotyping to support genomic selection, Paper III applies our pooling strategy augmented by imputation to a MAGIC population of inbred lines of bread wheat. We use sequenced data that was strongly filtered and curated, such that it can reasonably model genotypes tested on a hypothetical microarray.

Consistent with the findings of Paper I, pooling prior to imputation improves the accuracy of the genotype predictions for rare variants and imputation with a coalescent model performs better than with a clustered haplotypes model. We also find that using a small reference panel without any pedigree information can suffice to obtain highly accurate imputed genotypes, if this panel has a composition that can explain all the genetic structure and variations found in the study population. A small panel has the advantage of maintaining a low computational cost for the imputation task, which is part of a cost-conscious genotyping strategy.

The code supporting the simulations carried out in this paper is implemented in as a workflow using Snakemake and relies on a publicly available data set. This makes the results presented reproducible.

Contributions
CN devised the study and made changes in *prophaser* to fit the study case. CC wrote the code necessary to coordinate the experimental steps into a reproducible workflow. CC performed the analyses and all authors engaged in interpreting the results and formulating the findings presented. The draft manuscript was proposed by CC, but all authors read, amended, and approved the final version of the manuscript.

Paper IV

Paper IV presents an iterative coupled model that progressively adjusts the genotype probabilities resolved from the pools by correcting the decoded estimates using the imputed outcomes. In each iteration, the coupling is implemented as a feedback mechanism in which the likelihood of detecting the alleles is used as a consistency criterion. If the same alleles are likely to be detected in both the pooled and in the imputed data, the outcomes of imputation are considered consistent with the pooled observations. In the case of inconsistency, the deviation in the allele detection likelihoods between the pooled and the imputed data is used to update the decoded genotype probabilities. The updated decoded data is then processed again with imputation. In essence, the updates in the data aim to favour the expected alleles so that consistent genotypes are imputed in the next cycle.

The simulations are performed in the same MAGIC population of bread wheat as in Paper III, in order to illustrate a relevant application of our method in the context of plant breeding. We substantially extend the workflow developed in Paper III with a method called *repool*, which performs re-pooling of the imputed data and, if necessary, adjustments to the decoded genotype probabilities.

We find that repeated cycles of correction and imputation in the decoded data can fully exploit the advantages of both pooling and imputation. The genotyping accuracy is greatly increased across the entire MAF spectrum, and the genotype predictions for common variants are particularly improved. From a cost perspective, our iterative approach is effective as each cycle has a very low computational time and negligible memory requirements only.

Contributions
CN suggested the idea of the study and the setup for experiments. CN also developed the code for *repool*. CC implemented the scripts to extend the workflow of the model into coupled iterations, performed the experiments,

and drafted the manuscript. The analysis of the results and the formulation of the conclusions were carried out in cooperation between both authors, who also both edited and approved the final manuscript.

6. Sammanfattning på svenska

Genetiska data används numera ofta i olika forskningsfält och tillämpningsområden, till exempel genomisk medicin, förädling av kulturväxter och boskapsavel. Den genetiska informationen för såväl människor som djur och växter finns inkodad i cellernas DNA-molekyler. Det är ofta tillräckligt att endast analysera specifika korta avsnitt i DNA för att karakterisera stora delar av en individs genetiska särart. Dessa små avsnitt kan kallas genetiska markörer och finns på samma position i genomet i de allra flesta individer av samma art. Det finns många olika typer av genetiska markörer, men så kallade SNPar ("snippar") är bland de vanligaste. SNP står för det engelska *Single Nucleotide Polymorphism*, där nukleotid är den minsta elementära komponenten i DNA.

Med lämpliga statistiska modeller och korrekt analys, kan genotypdata ibland bidra till att avslöja nya kopplingar mellan vissa sjukdomar och specifika genetiska profiler, eller att med högre precision identifiera förädlingslinjer som kan ge bättre kvalitetsegenskaper och på så vis göra så kallad genomisk selektion mer effektiv. Bland kvalitetsegenskaper kan till exempel avkastning och tålighet mot torr odlingsmljö eller skadedjur nämnas. En viss markörsekvens kan i så fall ses som en genetisk signatur. Helgenomassocationsstudier, förkortade som GWAS på engelska, siktar på att upptäcka sådana genetiska signaturer.

Insamling av SNP-genotypdata kallas ofta helt enkelt genotypning. Det finns flera labbtekniker för att göra genotypning. De har samtliga utvecklats och förfinats radikalt över de senaste decennierna så att det nu är möjligt att utföra genotypning snabbt och kostnadseffektivt. Nuförtiden kan SNP-arrayer och sekvenseringsmetoder användas för att undersöka genotyperna i tiotusentals markörer i hundratals eller tusentals individer. Verkligt storskalig genotypning av ett stort antal individer kan ändå innebära en betydande merkostnad och begränsa storleken på vissa studier.

Så kallade imputationsmetoder för genotypdata har studerats och utvecklats med syfte att sänka kostnaden för genotypning. De är avancerade beräkningsmetoder för statistisk inferens som med hög noggrannhet kan uppskatta genotyperna på mycket fler markörer i en studiepopulation, baserat på glesa markörgenotyper från SNP-arrayer för den populationen, tillsammans med en separat panel med mer fullständiga genotyper för referensindivider. Inferensprocessen i imputation använder den genetiska strukturen i referenspanelen och i studiepopulationen, samt likheter i genotypmönster mellan individer, för att uppskatta de okända genotyperna. En låg markörtäthet innebär att markörvarianter som är ovanliga i populationen ofta inte testas direkt. Samtidigt kan

just dessa vara svåra att matcha rätt för imputationsmetoderna. Precisionen i genotypuppskattningarna kan vara väsentlig för kvaliteten på GWAS-resultat. Dessutom kan sällsynta varianter ibland påverka den egenskap man vill studera markant.

Att på ett effektivt sätt identifiera sällsynta egenskaper inom en population är ett problem som dyker upp i många kontexter som inte har med genotypning att göra. Detta kan lösas med grupptestningstekniker, som också kallas för poolningstekniker. Kärnidén i poolning är att testa grupper av individer i stället för enskilda individer var för sig, vilket minskar antalet test som ska utföras.

I detta avhandlingsarbete presenteras olika poolningsstrategier som är anpassade för genotypdata, kombinerade med genotypimputation. Hur individerna optimalt ska delas in i pooler kan vara en forskningsfråga i sig, som inte är i fokus i denna avhandling. Fokus har i stället varit avkodningsmetoder för studera poolningsstrategier, d.v.s. hur genotypen i varje prov kan uppskattas utifrån observationer som görs i pooler. Resonemangen baseras på en specifik poolningsdesign som kan representeras som oberoende block av överlappande pooler. Studiepopulationen som ska testas delas upp i sådana poolningsblock. Inom varje block ingår varje enskild individ alltid i två olika pooler. Vi har simulerat poolade data som sedan har avkodats till individuella genotypsannolikheter. Dessa kan i sin tur sedan utgöra indata till en imputationsmetod. Poolning minskar kostnaden för tät genotypning och fångar de sällsynta varianterna direkt, med hög noggrannhet. Oftast kan de vanligare varianterna inte avkodas med lika hög noggrannhet och kan på så sätt betraktas som saknade data. Imputationsmetoder kompletterar genoptypningen genom att ge hög precision för inferens av vanliga varianter utan att några ytterligare labbtester behövs.

Man kan faktiskt jämföra vår avkodningsprocess med att lösa Sudoku. Ibland kan man otvetygt härleda vissa utfall utifrån siffrorna som står i andra block samt på andra rader och kolumner. I andra fall kan flera olika siffror passa in i en ruta. Dessa är tvetydiga fall som kan hanteras genom att uttrycka de möjliga utfallen som sannolikheter. Om exempelvis endast en siffra kan passa in i rutan betyder det att den siffran har 100 % chans att stå där, med andra ord en sannolikhet lika med 1. Om man däremot drar slutsatsen att två olika siffror är lika möjliga utfall i en ruta blir den motsvarande probabilistiska formuleringen att var och en kan förekomma med sannolikhet 0.5, d.v.s. med 50 % chans. Våra strategier för att avkoda poolerna följer samma princip och beräknar för varje individ sannolikheterna för olika genotyputfall. Varje poolningsblock är ett rutnät som består av 4 rader och 4 kolumner där de möjliga genotyperna i en enskild markör kan uttryckas med heltalen 0, 1 och 2. Varje rad och varje kolumn utgör en pool, så varje individ tillhör två pooler. Totalt undersöks genotyperna hos 16 individer med bara 8 tester.

I avhandlingens sammanfattning beskrivs kontexten för poolad genotypning och en del teori relaterad till detta forskningsfält. Med hjälp av några ko-

rta illustrerade exempel undersöker vi de speciella egenskaper som uppträder i saknade data i samband med poolavkodning, samt i vilken mån dessa egenskaper påverkar två utvalda imputationsmetoder. I de bifogade artiklarna och den tekniska rapporten presenteras olika empiriska undersökningar i form av simulerad poolning följd av imputation i två studiepopulationer som har mycket olika egenskaper. Den första är en människopopulation som kunde vara en tillämpning av våra metoder i genomisk medicin. Den andra populationen omfattar inavlade linjer av vete, i ett scenario som mer påminner om praktiska tillämpningar inom växtförädling.

I avhandlingen som helhet framgår att trots flera utmaningar med poolade genotypdata kan dessa användas i en imputationsmetod om man väljer en lämplig avkodningsmetod för poolerna. De slutliga genotypuppskattningarna kan då bli mycket norgranna och erhålls till betydligt lägre kostnad i jämförelse med att utföra tät genotypning separat på varje individ.

7. Résumé en français

La recherche médicale ainsi que la sélection génomique végétale et animale ont aujourd'hui fréquemment recours aux données génétiques, encodées dans l'ADN de chaque individu, pour améliorer leurs résultats et accélérer leur obtention. Le long du génome, ce sont d'infimes sections de l'ADN qui sont decryptées et analysées, mais sans subir de modifications. Ces minuscules segments génomiques correspondent à des marqueurs génétiques situés à des positions bien précises et partagées par presque tous les individus d'une même espèce. Les marqueurs génétiques d'intérêt varient selon les usages visés et l'organisme étudié. L'identité de chaque marqueur est appelée génotype. Dans certains cas, une séquence de génotypes peut être corrélée à un profil à risque pour une maladie comme le diabète, ou bien au contraire indiquer une prédisposition génétique telle que la résistance à la sécheresse chez une variété de céréales. La séquence de génotypes correspond alors en quelque sorte à une signature génétique.

Les SNPs, de l'anglais *Single Nucleotide Polymorphisms*, sont une catégorie de marqueurs génétiques couramment utilisée. La recherche moderne a mis au point des technologies performantes telles que les puces à ADN ou des méthodes de séquençage et de génotypage à haut débit qui permettent de récolter avec précision les génotypes de dizaines de milliers voire de millions de SNPs chez des centaines d'individus. Toutefois, le coût relatif des tests en laboratoire pour de telles études reste encore parfois un obstacle majeur au déploiement du génotypage à très grande échelle, notamment pour les espèces animales et végétales.

Certaines méthodes de calcul scientifique et d'inférence statistique spécifiques aux génotypes, regroupées sous le terme de méthodes d'imputation génétique, peuvent être utilisées pour réduire le coût du génotypage. L'imputation génétique est utilisée en particulier dans les études d'associations à l'échelle du génome, abrégées par l'acronyme GWAS en anglais. De façon générale, l'inférence statistique vise à produire des estimations pour des valeurs inconnues dans un jeu de données. À partir des données génétiques partielles et obtenues par génotypage pour un coût raisonnable, les méthodes d'imputation génétiques calculent les génotypes d'autres SNPs en se fondant sur une population de référence, permettant ainsi d'augmenter le volume de données disponibles sans avoir à tester en laboratoire tous les marqueurs génétiques pour l'ensemble de la population étudiée. Les variations génétiques les plus rares sont souvent absentes de ces jeux de données partiels. La précision des génotypes calculés est déterminante pour la qualité des analyses ultérieures telles

que la détection d'associations génétiques. Bien que les méthodes d'imputation aient démontré une grande précision pour la plupart des marqueurs génétiques, leur faiblesse demeure le calcul des génotypes pour les variantes peu fréquentes, alors même que ces variantes peuvent avoir une signification majeure dans les signatures génétiques étudiées.

Identifier avec efficacité au sein d'une population des individus présentant un profil rare est un problème qui dépasse largement le cadre du génotypage. Ce problème peut être traité avec des techniques de test groupé, appelé en anglais *pooling*. Les tests groupés ont par exemple été employés pour le dépistage massif de maladies à faible prévalence et identifiables par tests immunologiques, sérologiques ou antigéniques. Au lieu de tester des échantillons individuels un par un, le principe de ces techniques est de tester des groupes d'échantillons mélangés, ou pools, puis de décoder les données aggrégées, c'est-à-dire de reconstituer le résultat du test pour chaque individu à partir des résultats observés pour les groupes. Idéalement, le nombre de groupes est très inférieur au nombre d'individus, si bien que moins de tests sont nécessaires pour couvrir l'ensemble de la population, ce qui peut permettre de réaliser d'importantes économies. Toutefois, il n'est pas toujours possible de déduire avec certitude tous les résultats individuels et ceci se traduit par des données manquantes.

Le travail exposé dans cette thèse porte sur les stratégies de génotypage groupé complété par l'imputation génétique. Plus exactement, le cœur des travaux de recherche présentés concerne des méthodes de calcul inférentiel pour non seuleument décoder les résultats groupés, mais aussi en vue d'utiliser ces génotypes comme données d'entrée pour l'imputation. Comment construire des groupes de façon optimale est une question de recherche à part entière qui n'est pas traitée dans cette thèse. Les études et réflexions proposées se fondent sur des exemples utilisant un schéma chevauchant de tests groupés en blocs indépendants.

Le principe utilisé pour décoder les groupes génotypés comporte des similarités avec la résolution de grilles de Sudoku. Dans ce jeu, en raisonnant à partir des chiffres déjà placés dans les carrés, lignes et colonnes, il est parfois possible de déduire directement et sans équivoque le seul chiffre pouvant se trouver dans une case donnée. Dans d'autres cas, plusieurs chiffres peuvent être compatibles avec ceux déjà inscrits. Ces déductions, totales ou partielles, peuvent être exprimées sous forme de probabilités : un chiffre déterminé sans ambiguité a 100 % de chances, soit une probabilité égale à 1, d'occuper une case ; si deux chiffres sont envisageables, alors chacun d'eux a une probabilité égale à 0,5. Dans le schéma de test groupé que nous avons étudié, chaque bloc indépendant peut être représenté comme une grille constituée de 4 lignes et de 4 colonnes. Chaque ligne et chaque colonne correspond à un groupe, si bien que tout échantillon appartient à deux groupes chevauchants et 16 individus sont testés avec seulement 8 groupes. Les génotypes, groupés ou individuels, peuvent prendre la valeur 0, 1, ou 2. Les méthodes inférentielles étudiées cal-

culent la probabilité de chacune de ces valeurs pour chaque individu, que son génotype soit déductible directement ou ambivalent.

La première partie de cette thèse est une introduction détaillée du contexte du génotypage groupé ainsi que d'une partie de la théorie liée à ce domaine de recherche. Cette introduction propose également, à travers de courts exemples illustrés, une étude formalisant les particularités des génotypes décodés et dans quelle mesure ces particularités affectent deux méthodes choisies d'imputation génétique. La deuxième partie de la thèse est constituée par les articles de recherche et rapport technique qui ont contribués aux résultats et conclusions présentés. Ces travaux empiriques sont des simulations de scénarios de génotypage groupé combinés à l'imputation, conduits dans deux populations exemples aux caractéristiques très différentes. La première est une population humaine pouvant illustrer un cas d'application de nos méthodes en recherche génomique médicale. La deuxième est une population de lignées de blé obtenues par croisements et autofécondation, telle qu'on en trouve en sélection et amélioration variétale. À travers ces différents exemples théoriques et empiriques, nous montrons qu'avec des modèles et procédures adaptés, le génotypage groupé et combiné à l'imputation génétique peut permettre d'obtenir des données précises tout en limitant les dépenses liées aux tests.

8. Acknowledgments

I would like to express my sincere thanks to my supervisor, Carl Nettelblad, for his patient and dedicated guidance, for his responsiveness and reliability in unfailingly answering every question I had, and for the always thorough and insightful discussions.
To all my TDB colleagues, please be assured of my gratitude for the moral support, the scientific and intercultural insights, and the valuable advice on writing and teaching. Thanks to the few ones who kept coming to the office during the pandemic, in a deserted and ghost ITC, and thus made the days more lively and somewhat more tangible in those surreal times.
To all my climbing friends, thank you for the lead (climbing) on whatever routes, be they shared academic experiences, relaxed - if not lazy - indoor climbing sessions, or proper outdoor adventures. You have preserved me from despair and you have also helped to improve my (climbing) problem solving skills a lot! Many of you are "gypsy" scientists, expatriates in Uppsala like me, I believe you know how much it means to find a community that you feel you belong to, when everything else is foreign, and how important it is for someone's well-being.
To the family and friends who received a knitted or crocheted gift in the autumn of 2023, thank you for indirectly contributing to the writing of this thesis. If you ever read it carefully, you may notice the interwoven yarn between the lines in some paragraphs.

The work presented in thesis has been conducted thanks to the funding of the Swedish Research Council Formas (grant No. 2017-00453) and the computing resources provided by the National Academic Infrastructure for Supercomputing in Sweden (NAISS, formerly SNIC prior to 2023) through Uppsala Multidisciplinary Center for Advanced Computational Science (UPPMAX).

References

[1] A. Adler, G. Wiley, and P. Gaffney. Infinium assay for large-scale SNP genotyping applications. *Journal of Visualized Experiments*, 81(e50683), 2013.

[2] S. F. Alamary. The use of SNP genotyping for QTL/candidate gene discovery in plants discovery in plants. *Journal of Engineering and Applied Sciences*, 7(2), 2020.

[3] P. A. Alexandre, L. R. Porto-Neto, E. Karaman, S. A. Lehnert, and A. Reverter. Pooled genotyping strategies for the rapid construction of genomic reference populations. *Journal of Animal Science*, 97(12):4761–4769, 2019.

[4] A. M. Allen et al. Characterization of a Wheat Breeders' Array suitable for high-throughput SNP genotyping of global accessions of hexaploid bread wheat (triticum aestivum). *Plant Biotechnology Journal*, 15(3):390–401, 2017.

[5] A. Ameur et al. Swegen: a whole-genome data resource of genetic variability in a cross-section of the swedish population. *European Journal of Human Genetics*, 25:1253–1260, 2017.

[6] K. Ausmees and C. Nettelblad. Achieving improved accuracy for imputation of ancient DNA. *Bioinformatics*, 39(1):btac738, 2023.

[7] J. Bhat, S. Ali, R. Salgotra, Z. Mir, S. Dutta, V. Jadon, A. Tyagi, M. Mushtaq, N. Jain, P. Singh, G. Singh, and K. Prabhu. Genomic selection in the era of next generation sequencing for complex traits in plant breeding. *Frontiers in Genetics*, 7(221), 2016.

[8] W. Bodmer and C. Bonilla. Common and rare variants in multifactorial susceptibility to common diseases. *Nature Genetics*, 40:695–701, 2008.

[9] B. L. Browning and S. R. Browning. Efficient multilocus association testing for whole genome association studies using localized haplotype clustering. *Genetic Epidemiology*, 31:365–375, 2007.

[10] B. L. Browning and S. R. Browning. 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*, 84:210–223, 2009.

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

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

[13] S. R. Browning. Multilocus association mapping using variable-length markov chains. *The American Journal of Human Genetics*, 78:903–913, 2006.

[14] S. R. Browning and B. L. Browning. 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*, 81:1084–1097, 2007.

[15] S. R. Browning and B. L. Browning. Haplotype phasing: existing methods and new developments. *Nature Reviews Genetics*, 12, 2011.
[16] M. P. L. Calus, T. H. E. Meuwissen, A. P. W. de Roos, and R. F. Veerkamp. Accuracy of Genomic Selection Using Different Methods to Define Haplotypes. *Genetics*, 178(1):553–561, 01 2008.
[17] C. Cao, C. Li, Z. Huang, X. Ma, and X. Sun. Identifying rare variants with optimal depth of coverage and cost-effective overlapping pool sequencing. *Genetic Epidemiology*, 37(8):820–830, 2013.
[18] C. Cao, C. Li, and X. Sun. Quantitative group testing-based overlapping pool sequencing to identify rare variant carriers. *BMC Bioinformatics*, 15(195), 2014.
[19] C. Cavanagh, M. Morell, I. Mackay, and W. Powell. From mutations to magic: resources for gene discovery, validation and delivery in crop plants. *Current Opinion in Plant Biology*, 11(2):215–221, 2008. Genome studies and Molecular Genetics, edited by Juliette de Meaux and Maarten Koornneef / Plant Biotechnology, edited by Andy Greenland and Jan Leach.
[20] C. R. Cavanagh et al. Genome-wide comparative diversity uncovers multiple targets of selection for improvement in hexaploid wheat landraces and cultivars. *Proceedings of the National Academy of Sciences*, 110(20):8057–8062, 2013.
[21] H.-B. Chen and F. Wang. A survey on nonadaptive group testing algorithms through the angle of decoding. *Journal of Combinatorial Optimization*, 15:49–59, 2008.
[22] X. Chi, X. Lou, M. Wang, et al. An optimal DNA pooling strategy for progressive fine mapping. *Genetica*, 135(267), 2009.
[23] F. S. Collins, M. Morgan, and A. Patrinos. The human genome project: Lessons from large-scale biology. *Science*, 300(5617):286–290, 2003.
[24] J. A. Collister, X. Liu, and L. Clifton. Calculating polygenic risk scores (PRS) in UK Biobank: A practical guide for epidemiologists. *Frontiers in Genetics*, 13, 2022.
[25] A. Cseh, P. Poczai, T. Kiss, et al. Exploring the legacy of central european historical winter wheat landraces. *Scientific Reports*, 11(23915), 2021.
[26] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, 39(1):1–22, 1977.
[27] E. Esteves, A. Mendes, M. Barros, C. Figueiredo, J. Andrade, J. Capelo, et al. Population wide testing pooling strategy for SARS-CoV-2 detection using saliva. *PLoS ONE*, 17(1), 2022.
[28] H. Gao, F. K. Hwang, M. T. Thai, W. Wu, and T. Znati. Construction of d(H)-disjunct matrix for group testing in hypergraphs. *Journal of Combinatorial Optimization*, 2006.
[29] K. Gardner, L. Wittern, and I. Mackay. A highly recombined, high-density, eight-founder wheat magic map reveals extensive segregation distortion and genomic locations of introgression segments. *Plant Biotechnology Journal*, 14(6):1406–1417, 2016.
[30] A. Gomes and B. Korf. Chapter 5 - genetic testing techniques. In N. H. Robin and M. B. Farmer, editors, *Pediatric Cancer Genetics*, pages 47–64. Elsevier, 2018.
[31] D. He et al. Genotyping common and rare variation using overlapping pool

sequencing. *BMC Bioinformatics*, 12(6), 2011.
[32] J. He, X. Zhao, A. Laroche, Z.-X. Lu, H. Liu, and Z. Li. Genotyping-by-sequencing (GBS), an ultimate marker-assisted selection (mas) tool to accelerate plant breeding. *Frontiers in Plant Science*, 5:484, 2014.
[33] J. M. Heather and B. Chain. The sequence of sequencers: The history of sequencing dna. *Genomics*, 107:1–8, 2016.
[34] F. Hormozdiari et al. Efficient genotyping of individuals using overlapping pool sequencing and imputation. *2012 Conference Record of the Forty Sixth Asilomar Conference on Signals, Systems and Computers (ASILOMAR)*, pages 1023–1027, 2012.
[35] B. Howie, P. Donelly, and J. Marchini. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genetics*, 5(6), 2009.
[36] B. Howie and J. Marchini. Genotype imputation for genome-wide association studies. *Nature Reviews Genetics*, 11, 2010.
[37] B. E. Huang, A. W. George, K. L. Forrest, A. Kilian, M. J. Hayden, M. K. Morell, and C. R. Cavanagh. A multiparent advanced generation inter-cross population for genetic analysis in wheat. *Plant Biotechnology Journal*, 10(7):826–839, 2012.
[38] P. J. Hurd and C. J. Nelson. Advantages of next-generation sequencing versus the microarray in epigenetic research. *Briefings in Functional Genomics*, 8(3):174–183, 06 2009.
[39] International Wheat Genome Sequencing Consortium (IWGSC). Shifting the limits in wheat research and breeding using a fully annotated reference genome. *Science*, 61(6403), 2018.
[40] G. Keeble-Gagnère, R. Pasam, K. Forrest, D. Wong, H. Robinson, J. Godoy, A. Rattey, D. Moody, D. Mullan, T. Walmsley, H. Daetwyler, J. Tibbits, and M. Hayden. Novel design of imputation-enabled SNP arrays for breeding and research applications supporting multi-species hybridization. *Frontiers in Plant Science*, 12, 2021.
[41] A. Kho, L. Rasmussen, J. Connolly, et al. Practical challenges in integrating genomic data into the electronic health record. *Genetics in Medicine*, 15:772–778, 2013.
[42] N. Li and M. Stephens. Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics*, 165(4):2213–2233, 2003.
[43] Y. Li, C. J. Wille, J. Ding, P. Scheet, and G. R. Abecasis. Mach: Using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genetic Epidemiology*, 34(8):816–834, 2010.
[44] H. Liu, M. Bayer, A. Druka, et al. An evaluation of genotyping by sequencing (GBS) to map the Breviaristatum-e (ari-e) locus in cultivated barley. *BMC Genomics*, 15(104), 2014.
[45] G. Logsdon, M. Vollger, and E. Eichler. Long-read human genome sequencing and its applications. *Nature Reviews Genetics*, 21:597–614, 2020.
[46] I. Mackay and W. Powell. Methods for linkage disequilibrium mapping in crops. *Trends in Plant Science*, 12:57–63, 2007.
[47] J. Marchini, B. Howie, S. Myers, G. McVean, and P. Donnelly. A new

multipoint method for genome-wide association studies by imputation of genotypes. *Nature Genetics*, 39:906–913, 2007.

[48] F. Marroni, S. Pinosio, and M. Morgante. The quest for rare variants: Pooled multiplexed next generation sequencing in plants. *Frontiers in Plant Science*, 3, 2012.

[49] T. H. Meuwissen. Accuracy of breeding values of 'unrelated' individuals predicted by dense SNP genotyping. *Genetics Selection Evolution*, 41(1):1–9, 2009.

[50] T. H. Meuwissen, B. J. Hayes, and M. Goddard. Prediction of total genetic value using genome-wide dense marker maps. *genetics*, 157(4):1819–1829, 2001.

[51] M. Mitt, M. Kals, K. Pärn, S. B. Gabriel, E. S. Lander, A. Palotie, S. Ripatti, A. P. Morris, A. Metspalu, T. Esko, R. Mägi, and P. Palta. Improved imputation accuracy of rare and low-frequency variants using population-specific high-coverage WGS-based imputation reference panel. *European Journal of Human Genetics*, 25:869–876, 2017.

[52] Y. Momozawa and K. Mizukami. Unique roles of rare variants in the genetics of complex diseases in humans. *J Hum Genet*, 66:11–23, 2021.

[53] J. D. Montenegro, A. A. Golicz, P. E. Bayer, B. Hurgobin, H. Lee, C.-K. K. Chan, P. Visendi, K. Lai, J. Doležel, J. Batley, and D. Edwards. The pangenome of hexaploid bread wheat. *The Plant Journal*, 90(5):1007–1013, 2017.

[54] S. P. Moose and R. H. Mumm. Molecular Plant Breeding as the Foundation for 21st Century Crop Improvement. *Plant Physiology*, 147(3):969–977, 07 2008.

[55] M. Mézard, M. Tarzia, and C. Toninelli. Group testing with random pools: Phase transitions and optimal strategy. *Journal of Statistical Physics*, 131:783–801, 2008.

[56] H. Q. Ngo and D.-Z. Du. A survey on combinatorial group testing algorithms with applications to DNA library screening. *DIMACS Series in Discrete Mathematics and Theoretical Computer Science*, pages 49–59, 2000.

[57] M. Nothnagel, D. Ellinghaus, S. Schreiber, M. Krawczak, and A. Frankeg. A comprehensive evaluation of SNP genotype imputation. *Human Genetics*, 125:163–171, 2009.

[58] T. Pook, M. Mayer, J. Geibel, S. Weigend, D. Cavero, C. Schoen, and H. Simianer. Improving imputation quality in beagle for crop and livestock data. *Genes Genomes Genetics*, 98:116–126, 2019.

[59] E. Porcu, S. Sanna, C. Fuchsberger, and L. G. Fritsche. Genotype imputation in genome-wide association studies. *Current Protocols in Human Genetics*, 1.25.1, 2015.

[60] B. Pucker, I. Irisarri, J. de Vries, and B. Xu. Plant genome sequence assembly in the era of long reads: Progress, challenges and future directions. *Quantitative Plant Biology*, 3:e5, 2022.

[61] L. R. Rabiner. A tutorial on Hidden Markov Models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

[62] A. Rasheed and X. Xia. From markers to genome-based breeding in wheat. *Theoretical Applied Genetics*, 132:767–784, 2019.

[63] C. Robertsen, R. Hjortshøj, and L. Janss. Genomic selection in cereal breeding. *Agronomy*, 9:95, 02 2019.

[64] N. Salcedo, A. Harmon, and B. B. Herrera. Pooling of samples for

SARS-CoV-2 detection using a rapid antigen test. *Frontiers in Tropical Diseases*, 2, 2021.
[65] J. L. Scafer and J. W. Graham. Missing data: Our view of the state of the art. *Psychological Methods*, 7(2):147–177, 2002.
[66] A. Scheben, J. Batley, and D. Edwards. Genotyping-by-sequencing approaches to characterize crop genomes: choosing the right tool for the right application. *Plant Biotechnology Journal*, 15:149–161, 2017.
[67] M. Schierenbeck, A. Alqudah, U. Lohwasser, et al. Genetic dissection of grain architecture-related traits in a winter wheat population. *BMC Plant Biology*, 21(417), 2021.
[68] M. Scott, O. Ladejobi, S. Amer, et al. Multi-parent populations in crops: a toolbox integrating genomics and genetic mapping with breeding. *Heredity*, 125:396–416, 2020.
[69] M. F. Scott, N. Fradgley, A. R. Bentley, T. Brabbs, F. Corke, K. A. Gardner, R. Horsnell, P. Howell, O. Ladejobi, I. J. Mackay, R. Mott, and J. Cockram. Limited haplotype diversity underlies polygenic trait architecture across 70 years of wheat breeding. *Genome Biology*, 22(137), 2021.
[70] P. Sham, J. Bader, I. Craig, et al. DNA pooling: a tool for large-scale association studies. *Nature Reviews Genetics*, 3:862–871, 2002.
[71] J. Shendure and al. Advanced sequencing technologies: Methods and goals. *Nature Reviews Genetics*, 5:335–344, 2004.
[72] L. Skøt and N. Grinberg. Genomic selection in crop plants. In B. Thomas, B. G. Murray, and D. J. Murphy, editors, *Encyclopedia of Applied Plant Sciences (Second Edition)*, pages 88–92. Academic Press, Oxford, second edition edition, 2017.
[73] M. Stephens, N. J. Smith, and P. Donnelly. A new statistical method for haplotype reconstruction from population data. *The American Journal of Human Genetics*, 68:978–989, 2001.
[74] C. Sudlow, J. Gallacher, N. Allen, V. Beral, P. Burton, J. Danesh, et al. Uk biobank: An open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Medicine*, 12(3), 2015.
[75] P. Sudmant, T. Rausch, E. Gardner, et al. An integrated map of structural variation in 2,504 human genomes. *Nature*, 526:75–81, 2015.
[76] Y. J. Sung, L. Wang, T. Rankinen, C. Bouchard, and D. Rao. Performance of genotype imputations using data from the 1000 Genomes Project. *Human Heredity*, 73:18–25, 2012.
[77] N. Tang and Y. Ju. Statistical inference for nonignorable missing-data problems: a selective review. *Statistical Theory and Related Fields*, 2(2):105–133, 2018.
[78] F. Technow and J. Gerke. Parent-progeny imputation from pooled samples for cost-efficient genotyping in plant breeding. *PLoS ONE*, 12(12), 2017.
[79] N. Thierry-Mieg. A new pooling strategy for high-throughput screening: the shifted transversal design. *BMC Bioinformatics*, 7(28), 2006.
[80] G. Thorisson, A. Smith, L. Krishnan, and S. LD. The international hapmap project web site. *Genome Research*, 15:1592–15933, 2005.
[81] S. Walkowiak, L. Gao, C. Monat, et al. Multiple wheat genomes reveal global variation in modern breeding. *Nature*, 588:277–283, 2020.
[82] J. Wang et al. Investigation of rare and low-frequency variants using

high-throughput sequencing with pooled DNA samples. *Nature Scientific Reports*, 6(33256), September 2016.

[83] S.-X. Wang, Y.-L. Zhu, D.-X. Zhang, H. Shao, P. Liu, J.-B. Hu, H. Zhang, H.-P. Zhang, C. Chang, J. Lu, X.-C. Xia, G.-L. Sun, and C.-X. Ma. Genome-wide association study for grain yield and related traits in elite wheat varieties and advanced lines using SNP markers. *PLoS ONE*, 12(11):1–14, 11 2017.

[84] K.-C. Wong. Letter to the editor: Big data challenges in genome informatics. *Biophysical Reviews*, 11:51–54, 2019.

[85] A. G. Y. Erlich, K. Chang et al. DNA Sudoku—harnessing high-throughput sequencing for multiplexed specimen analysis. *Genome Research*, 19:1243–1253, 2009.

[86] J. Zhang, J. Yang, L. Zhang, et al. A new SNP genotyping technology target SNP-seq and its application in genetic analysis of cucumber varieties. *Scientific Reports*, 10(5623), 2020.

[87] P. Zhang, F. Krzakala, M. Mezard, and L. Zdeborova. Non-adaptive pooling strategies for detection of rare faulty items. *Lecture Notes in Computer Science and Workshop on Algorithms and Data Structures 2005: Algorithms and Data Structures*, 2013.

[88] C. Zheng, M. P. Boer, and F. A. van Eeuwijk. Accurate Genotype Imputation in Multiparental Populations from Low-Coverage Sequence. *Genetics*, 210(1):71–82, 07 2018.

Acta Universitatis Upsaliensis

Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology 2354

Editor: The Dean of the Faculty of Science and Technology

A doctoral dissertation from the Faculty of Science and Technology, Uppsala University, is usually a summary of a number of papers. A few copies of the complete dissertation are kept at major Swedish research libraries, while the summary alone is distributed internationally through the series Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology. (Prior to January, 2005, the series was published under the title "Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology".)

Distribution: publications.uu.se
urn:nbn:se:uu:diva-519887

ACTA UNIVERSITATIS
UPSALIENSIS
2024