LD-based SNP and genotype calling from next-generation sequencing reads

Androniki Menelaou

St. Anne's College

Trinity term 2012

Department of Statistics

University of Oxford

A thesis submitted for the fulfilment of the requirements

for the degree of Doctor of Philosophy

Supervisor: Jonathan Marchini

LD-based SNP and genotype calling from next-generation sequencing reads

Androniki Menelaou

St. Anne's College

Department of Statistics

University of Oxford

D.Phil thesis, Trinity term 2012

Abstract

Next-generation sequencing is revolutionising in genetics, where base-by-base information for the whole genome is available for a large sample of individuals. This type of data is becoming commonly used and will continue to be in the near future. One of the first questions arising is the identification of novel variants and subsequently genotype calling of the individuals in the sample. However, given the cost of sequencing, so far most projects are sequencing individuals in low to medium coverage.

In this thesis, we present two distinct methods for SNP and genotype calling from low-coverage sequencing data, TreeCall and MVNcall, that combine sequencing and Linkage Disequilibrium (LD) information. We begin by describing the pipeline for next-generation sequencing analysis and existing methods for SNP and genotype calling using low-coverage sequencing information. Subsequently, we present the two novel LD-based methods for SNP and genotype calling. The two methods developed assume a study design where the individu-

als are both genotyped and sequenced at low-coverage. The genotypes are used to construct a haplotype scaffold, where the LD information is extracted, either by the construction of genealogical trees (TreeCall), or the approximation of a windows of contiguous SNPs of the scaffold by a multivariate normal distribution (MVNcall).

Both methods have been applied on real datasets from the 1000 Genomes project and compared to other LD-based methods applied on the same datasets, mainly in terms of genotype calling and phasing. Whereas TreeCall gives lower genotype concordance rates than the other methods, MVNcall provides the highest genotype concordance rates for a dataset with a small sample size (Low-coverage pilot of the 1000 Genomes project).

Applying the MVNcall on a larger dataset (Phase 1 of the 1000 Genomes project), it achieves an overall genotype discordance rate of 0.58%, whereas SNPTools achieves an overall genotype discordance rate of 0.57%, Thunder 0.56%, and BEAGLE 0.61% (comparison based on Axiom chip). The main advantage of MVNcall is in terms of phasing accuracy, where by using a haplotype scaffold, and especially in the case where the haplotype scaffold is phased using pedigree information, it provides accurate haplotypes. MVNcall is also extended to incorporate trio information for genotype calling. Experiment on a deeply sequenced trio leads to an accurate set of haplotypes of the trio with switch error rates as low as $\sim$0.28 for the parents and $\sim$0.12 for the offspring.

Acknowledgements

I would first like to thank my supervisor Jonathan Marchini for his guidance and support throughout my DPhil. I am grateful to all the nice people I have been surrounded during my DPhil in Oxford. My officemates Claire Churchhouse, Valentina Iotchkova, Elena Frangou, and Marie Forest created a great and fun environment in the office all these years. Special thanks also to Olivier Delaneau, Jason Liu, Christopher Hallsworth, Bryan Howie and Adam Auton for helpful advice in different stages of my DPhil. Good friends made Oxford a home, Emmanuela Repapi and Melissa Ouellet is my surrogate family in Oxford and would like to thank them for all their support and great time we had together, as well as friends from home, Marianna Charalambous, Antonia Christofi and Maria Kofterou.

My studies in Oxford were made possible thanks to the funding body EPSRC, for funding all my years of DPhil. I also thank my examiners Gil McVean and Vincent Plagnol for considering this thesis.

My journey through my studies would never be feasible without the great support from my family. I am deeply grateful to my parents, for their unconditional love and support all these years, my brothers Theodosis and Marios for making me laugh, my brother Michalis that his memory is always with me, and my grandparents that gave me widsom advice. This thesis is dedicated to my family, the least I could do to thank them for everything they did all these years.

Contents

Chapter 1

Introduction and Background

1.1 Introduction

Heritability is the concept that encapsulates the proportion of phenotypic variation due to genetic factors. Mendel was the first to investigate heritability using pea plants, in which he found patterns of inheritance[1]. In humans, tall parents tend to have tall children, tall family members and so on. It turns out that what we have inherited from our parents, our genetic code, influences our predisposition to certain phenotypes[2]. Our genetic code consists of the complete set of DNA which is packed in the nucleus of the cells. The DNA consists of 23 pairs of chromosomes: one chromosome comes from the father, and the other one from the mother. In humans, this is equal to a three billion base pair sequence with four chemical components (nucleotides): Adenine (A), Guanine (G), Thymine (T), and Cytosine (C). Out of the three billion base pairs, 99.9% of the DNA

[1]Mendel, G., 1866, Versuche ber Pflanzen-Hybriden (Experiments in Plant Hybridization). Verh. Naturforsch. Ver. Brnn 4: 347 (in English in 1901, J. R. Hortic. Soc. 26: 132)

[2]Phenotypes are affected by both genetic and environmental factors. In this thesis, we only consider genetic factors.

is identical between humans, and only 0.1% differs (Human Genome Project, 2010). One of the main goals when studying genetics is to investigate the relationship between the genetic variation and different phenotypes, which can include physical and disease traits.

The relationship between disease traits and genetic variation is of great interest for medicine, since by understanding the relationship between different phenotypes and genotypes, personalized medicine can be developed that will correspond to each individual's genome, rather than the one-medicine-fits-all approach that has been applied until now.

For this to happen, detailed exploration of the genomes up to a base-to-base level is necessary so all genetic variation is captured. This requires technologies that sequence fast and cheaply a large number of genomes, and a set of different algorithms that extract the information and detect different types of variants. This thesis is dedicated to the study of "next generation sequencing " data and to how this information can be used to identify and genotype single base mutations (Single Nucleotide Polymorphisms or SNPs).

1.2 From the Human Genome Project to the 1000 Genomes Project

For the past 20 years, an abundance of projects have been undertaken to unveil the properties of our genetic code. These efforts began in 1990 when a consortium was formed aiming to sequence a complete human genome, the *Human*

Genome Project (HGP) (The International Human Genome Sequencing Consortium, 2001). The HGP not only recorded a complete human genome, but also identified important properties of it, including the indentification of genes and the combinatorial architecture of proteins (The International Human Genome Sequencing Consortium International, 2004). In 2004, a revised version of the complete human genome (NCBI35), consisted of 2.85 billion nucleotides interrupted by only 341 gaps (The International Human Genome Sequencing Consortium International, 2004).

The HGP provided information across one whole genome but still the question of genetic differences across individuals with different traits or from different population backgrounds needed the investigation of the DNA differences in a sample of individuals. In 2005, the HapMap project launched with the aim of studying human polymorphism in individuals from different population backgrounds (Asian, African, and European). The main objective of the project was to "create a genome wide database of *common* variation, providing information needed as a guide to genetic studies of clinical phenotypes" (International HapMap Consortium, 2005). In the three phases of the project, 3.5 million SNPs were identified. The project induced the development of low-cost genotyping technologies, and increased the interest in the design of chips for genotyping SNPs. After the completion of the project, a reference panel was created that could be used for association studies.

In the years that followed the HapMap project, there have been numerous studies on the association of genetic variation and a specific phenotype, these

are referred to as Genome-Wide Association Studies (GWAS). These studies collect two groups of individuals: one with the phenotype of interest (case group) and another without (control group). Using a statistical framework, they test the association between genetic variants and the trait of interest. Examples of GWAS from the past few years include the Wellcome Trust Case Control Consortium (WTCCC), which investigated the genetic basis of seven common complex diseases (Crohn's disease, coronary artery disease, hypertension, bipolar disorder, rheumatoid arthritis, type 1 diabetes, and type 2 diabetes) (The Wellcome Trust Case Control Consortium, 2007) as well as neuropsychiatric diseases such as schizophrenia and autism (Walsh et al., 2008). GWAS have been performed by various groups from all over the world, and SNP-trait associations are catalogued in the National Human Genome Research Institute as "A catalogue of published Genome Wide Association Studies" (Hindorff et al., 2010) .

The first generation of GWAS is well powered only for SNPs with Minor Allele Frequency (MAF) >5% and based on the assumption that common diseases could be explained by common variants, the so-called common disease-common variant hypothesis (CD/CV) (Ku et al., 2010). However, the effect size of the loci identified to be associated with a phenotype is relatively small in the majority of the cases, and they usually only explain a small proportion of the heritability (Manolio et al., 2009). After the first wave of GWAS, the question arose regarding where the "missing" heritability is hidden. It has been argued that one of the components which could explain part of the missing heritability are rare variants (Ku et al., 2010; Manolio et al., 2009; Schork et al., 2009).

In order to discover the rare variants, it is required to explore the genome in further detail. Nevertheless, the genotyping technologies used in the HapMap project, and the GWAS subsequently, were able to capture only a subset of the genetic variants, usually common variants, and the first generation sequencing was prohibitevely expensive and time consuming to sequence a large sample of individuals.

A recent advance in sequencing technologies, the so-called *next-generation sequencing* (NGS) technologies, have allowed the detailed exploration of genetic variation by moving from genotyping techniques that only capture specific positions to sequencing technologies that can read the genome base-by-base. These technologies can process millions of reads in parallel, providing an enormous amount of data at an increasing speed and decreasing costs. In 2008, an international consortium, the 1000 Genomes Project (1KGP), set out to explore the genetic variation amongst 2,500 individuals from different geographical backgrounds in order to create a detailed catalogue of human polymorphism (The 1000 Genomes Project Consortium, 2010).

The NGS technologies have brought a new era in genetics and now provide the tool for a detailed exploration of genetic variation in a population level. The 1KGP is aiming to create a detailed catalogue of human polymorphism that can be used as a reference panel for the projects to come. Given the decreasing cost of sequencing and the obvious advantages gained from sequencing individuals towards personalized medicine, a series of projects have commenced that sequence individuals at a national level including:

1. UK10K[3] aims at understanding the connection between rare variants and genetic diseases by sequencing $4,000$ controls and $6,000$ cases with different diseases in U.K.,

2. the SardiNIA project[4] studies individuals from Sardinia, an isolated island in the Mediterranean without much immigration, to find genes that affect the traits on this "founder" population,

3. *Genome of the Netherlands*[5] aims to sequence 750 related individuals (250 trios) from different parts of Netherlands to understand the genetic differences between the different regions of the country as well as detecting rare variants specific to Netherlands.

1.3 The 1000 Genomes project

The 1000 Genomes project involves a huge spectrum of scientists, including population and statistical geneticists, who work at several sequencing centres, technology companies and universities trying to create a detailed catalogue of human polymorphism. The primary goals are to i) identify >95% of the variants with MAF >1% in the sequenceable parts of the genome with 95% certainty, ii) identify >95% of rare variants with MAF >0.1% in exons with 95% certainty,[6] and iii) discover structural variants like Copy Number Variants (CNVs) and insertions/deletions (indels) (The 1000 Genomes Project Consortium, 2010).

[3] http://www.uk10k.org/
[4] http://sardinia.nia.nih.gov/
[5] http://www.nlgenome.nl/wiki/GonlStart
[6] i.e. False Discovery Rate (FDR) <5%

	Low-coverage pilot	Trio pilot	Exon pilot
Sample size (number of individuals)	179 (unrelated)	6 (2 trios)	697
Populations	YRI, CEU, JPT	YRI, CEU	YRI,LWK,CEU, TSI,CHB,JPT, CHD
Targeted region	whole genome (85% accessible)	whole genome (93% accessible)	8,140 exons from 906 random genes
Average mapped coverage	3.6x	42x	56x
Number of SNP calls	14.4 million	5.9 million	12,758
Number of indel calls	1.3 million	650,000	96
Number of CNV calls	20,000	14,000	–

Table 1.1: Summary information of the three pilot projects conducted by the 1KGP.

Since the cost, although decreasing, is still high to sequence a large sample of individuals on high coverage, the 1KGP studied three pilot projects, where for each pilot the number of samples and average coverage per genome were varied, to determine the design of the full project. Summary information for the three pilots examined are provided in Table 1.1 (The 1000 Genomes Project Consortium, 2010). In the low-coverage pilot, 179 unrelated individuals from Asia, Africa, and Europe are sequenced (whole genome) at low-coverage (3.6x) and detected 14.4 million SNPs, 1.3 million indels and, 20,000 CNVs; in the trio pilot the whole genome of two family trios from Africa and Europe ancestry was sequenced at 42x and discovered 5.9 million SNPs, 650,000 indels and 14,000 CNVs; and in the exon pilot the exomes of 697 individuals are sequenced at 56x and discovered 12,758 SNPs and 96 indels.

The pilot project of the 1KGP gave some first results regarding the information

extracted from the NGS data and a more detailed picture of the human genome. In particular, the variation detected by the 1KGP is not uniformly distributed across the genome. In terms of the known SNPs (e.g. dbSNPs), 56% were found in all populations in the low-coverage pilot. Conversely, 84% of the novel SNPs are present only in one population, where the majority of the novel variants are private to African populations. Comparing the power of discovering SNPs across the different pilots, it was found that the discovery of SNPs is affected by the depth of coverage. In particular, the trio mothers from trio pilot were also present in the low-coverage pilot. In trio pilot, 96.9% and 95.0% of the SNPs detected in the CEU and YRI mother respectively, are also detected in the low-coverage pilot; in terms of genotype accuracy, 93.8% and 88.4% of the genotypes are correctly called by the low-coverage pilot. The low-coverage pilot had high power to detect SNPs present in five or more of the 120 samples ($\sim$90%), reaching complete detection of variants with more than ten copies of the alternative allele, whereas the power to detect sites with less than ten copies of the alternative allele was lower with a power to detect singletons falling as low as $\sim$ 25%. It should be noted, that even though we speak about "whole genome" analysis, the project has only made calls in the "sequenceable" part of the genome (83% in the low-coverage pilot and 93% in the trio pilot), i.e. regions that are highly repetitive like the telomeres and the centromeres are excluded from the analysis. The current Phase I of the project involves the sequencing of 1,092 individuals from different population backgrounds including admixed populations and $\sim$40 million SNPs have been detected.

In the next three sections, we present in more detail the process during which the sequencing reads are generated using NGS technologies, the alignment procedure where the sequencing reads are aligned to a reference genome, and how we can approach the problem of detecting and genotyping novel variants. Here we focus only on SNPs from a design with a large sample size sequenced at low-coverage.

1.4 Next-generation sequencing pipeline

The pipeline for the analysis of NGS data employed by the 1KGP involves sequencing the genome in small fragments and then given a reference genome the reads are mapped to the corresponding part of the genome. When the sequencing reads have all been mapped to the reference genome, base-by-base nucleotide information for the genome of the individual sequenced is available, and we can combine the information from the reads at each site to identify the genetic variants including SNPs, structural variants and indels.

1.4.1 Generation of next-generation sequencing reads

Sequencing was first utilized in the HGP (The International Human Genome Sequencing Consortium, 2001), and it was developed by Frederick Sanger (Sanger et al., 1977). After more than three decades of exploration of the Sanger sequencing (or first generation sequencing), the per base "raw" accuracy reached 99.999% (Shendure and Ji, 2008). After the completion of the HGP, more major

organisms have been sequenced, and a reference genome library was created for various species. The benefits for sequencing were apparent, but the sequencing costs, although decreased by a 100-fold (Collins et al., 2003), were still high to be applied to a large sample. In the meantime, advances in different fields of chemistry, computational algorithms, and data storage opportunities made feasible a "next-generation" of sequencing technologies which produce an abundance of data within a short time period, running thousands of runs in parallel, and at decreasing costs. Different companies have developed different chemistry techniques for sequencing, the most popular at the moment being: Roche (454) GS FLX sequencer (Margulies et al., 2005), Illumina Genome Analyser (Bentley et al., 2008), Applied Biosystems SoLiD sequencer, Polonator, (Shendure et al., 2005) and Heliscope Single Molecule Sequencer Technology (Harris et al., 2008). The 1KGP used the first three aforementioned technologies to sequence the individuals in the project. We do not go into detail regarding how the technologies function in this thesis. Below, we only give a brief summary of the main features of these three methods in terms of method specific errors.

Roche (454) GS FLX sequencer

The length of the sequencing reads are ∼1,000 bp[7] and each run takes around 23 hours creating ∼1,000,000 reads. A challenge faced by this technology is the situation of long homopolymers. For example, if in a sequencing read there are

[7]http://454.com/products/gsflxsystem/index.asp, accessed May 2012

6 Adenines consecutively (e.g. CGAAAAAATG), the technology finds challenging to distinguish the exact number of A's. Hence, the method is prone to base insertions and deletions errors. On the other hand, the substitution error is very low (Magi et al., 2010) and the length of the read is larger than the other two technologies at the moment.

Illumina Genome Analyser

The length of sequencing reads from Illumina can reach ∼150 bp[8], where for the pilot phase of the 1KGP the read length was shorter (on average 35bp). This method does not have the problem with homopolymers as Roche, but it has a higher rate of substitution error and biases on the GC content. The overall misscall error rate is around 1%[9] (Nielsen et al., 2011).

Applied Biosystems SoLiD sequencer

The read length of SoLiD technology is up to 75 base pairs[10] and sequence a whole genome at 30x in one run. This tehcnology uses a color scheme, where each di-nucleotide corresponds to a different color. Moving one base at a time and testing a di-nucleotide at each step, means that each nucleotide is tested twice, which is an advantage of this technology. A drawback of this technology arises in biases due to the color scheme that appear in later machine cycles (Nielsen et al., 2011).

[8] http://www.illumina.com/systems/genome_analyzer_iix/performance_specifications.ilmn, accessed May 2012

[9] Newer versions of the sequencing machines have lower error rates

[10] http://www.appliedbiosystems.com/absite/us/en/home/applicationstechnologies/solidnext-generationsequencing/nextgenerationsystems.html, accessed May 2012

The NGS technologies have the great advantage that they can process millions of sequence reads in parallel opposed to 96 at a time with Sanger sequencing (Mardis, 2008). However, the accuracy of the NGS technologies is still not well understood and different technologies have different types of biases. Further work is required to fully understand the sources of biases. Also, due to those biases, the NGS technologies can not produce long reads at the moment in contrast to Sanger sequencing, although this is changing. Nevertheless, the NGS technologies give us for the first time the opportunity to sequence a large sample of genomes that was not viable before.

1.4.2 Alignment and Assembly of sequencing reads

The output from the sequencing technologies is the read sequence and the accompanying Phred[11] quality scores (Ewing et al., 1998; Ewing and Green, 1998) for each base on the read, based on the noise of the image analysis each sequencing technology performs [12]. The Phred quality score for base b is defined by

$$Q = -10log_{10}\mathbb{P}(\hat{b} \neq b) \tag{1.1}$$

where $\mathbb{P}(\hat{b} \neq b)$ is the probability that the base read is an error. For example, when Q=10, there is a 10% probability the base read is an error and when Q=30,

[11]Phred quality scores where firstly used by the program *Phred* in the Human Genome project to access the quality of the DNA sequencing. Phred created quality score look-up tables for sequences where the truth was known, and then for each base sequenced different parameters were calculated that corresponded to the look up table.

[12]In the 1KGP, this information is provided in BAM format (Li et al., 2009a)

the probability of the base read being an error is 0.1%. This information is vital for the following steps including alignment, SNP, and genotype calling.

Alignment

When a reference assembly exists (e.g. human, mouse), different alignment methods have been developed to map those reads to the reference genome. In a toy example in Fig.1.1, we show eight reads which are aligned to a reference genome. Reads 3 and 4 show some evidence of polymorphism for position 11 with two alleles, C and T. Another form of variation we can detect when the reads are aligned on the reference genome are insertions and deletions. For example, read 5 shows a 2bp deletion with respect to the reference genome. However, failing to detect indels can lead to wrongly aligned reads and false positive SNPs nearby indels (read 8).

Various alignment algorithms have been developed including: MAQ (Li et al., 2008a), ELAND (Cox, 2007), Novoalign (Novocraft, 2009), Mosaik (Quinlan et al., 2008), Stampy (Lunter and Goodson, 2011), SOAP (Li et al., 2008b), SOAP2 (Li et al., 2009b), BOWTIE (Langmead et al., 2009), and BWA (Li and Durbin, 2009). As we discussed above, each sequencing technology has a different source of errors and thus different mapping algorithms correspond to different sequencing technologies, e.g. MAQ, STAMPY, ELAND, Bowtie, SOAP, SOAP2, Novoalign[13] aligns Illumina sequencing reads, Mosaik aligns both Roche 454, SoLiD, Illu-

[13]also aligns paired-end reads from Roche 454

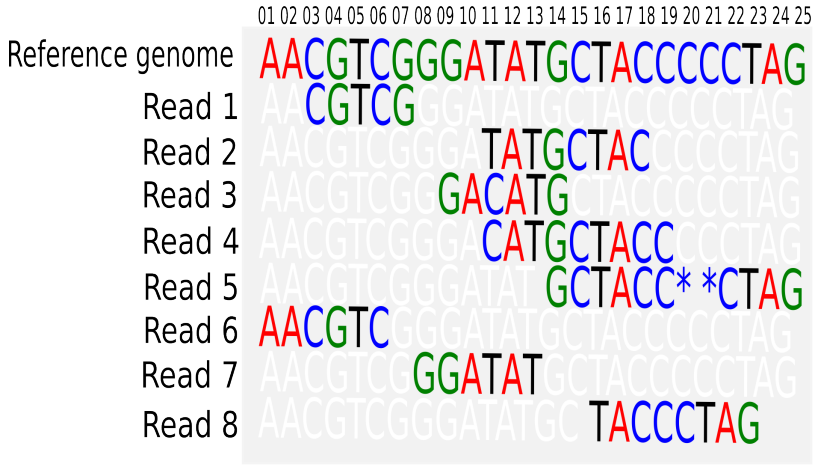

Figure 1.1: Toy example of alignmnent on a reference genome. The first row represents the reference genome and below are the reads that are aligned on the reference genome. Read 5 represents a deletion compared to the reference panel. At position 11, there is an indication that the position might be a SNP: reads 2 and 7 have a T at the position where reads 3 and 4 have a C at that position.

mina sequencing reads and provides experimental support for Helicos.

The alignment algorithms usually follow a two-step procedure. On the first step, they scan the reference genome to find potential mapping regions. On the second step, a more complex model is used to map the read in one of the candidate regions (Flicek and Birney, 2009). In addition, the algorithms also perform the so-called "gap-alignment", where they try to identify indels and prevent alignmnent errors like read 8 at Fig.1.1. This step greatly improves the variant calling accuracy in subsequent steps (Li and Homer, 2010).

In addition to gap alignment, alignment algorithms make use of the *paired-end* information provided by the NGS technologies. Two reads are paired when it is known that they are seperated by an approximate distance in the genome (Magi et al., 2010). The information from the paired-end reads helps with the mapping procedure, since when one of the two paired-reads maps uniquely to

the genome, then the other read is known to be at some approximate distance from the other mapped read.

After a read is mapped to the reference genome, the algorithms also output a *mapping quality* which is a Phred-like probability that a read alignment is incorrect. When a read falls in repetitive regions, the algorithm finds multiple positions to map the read. In these cases, the alignment algorithms give a low mapping score and map the read randomly at one of the matching regions.

Recalibrating the phred quality scores

The accuracy of the base calls and the quality scores (base and mapping), play a crucial role on the subsequent steps for variant calling. Therefore, the base quality scores are recalibrated so that they more accurately correspond to Eq. 1.1. The quality base accuracy depends on the flanking sites, the read base, the cycle (the position of the base on the read, the later it is on the read, the less accurate), and on the mapping quality. Different models were developed for quality base recalibration and the one employed by the 1KGP is implemented by GATK (McKenna et al., 2010; DePristo et al., 2011).

Assembly

When there is no reference assembly for the organism sequenced, assembly algorithms have been developed to reconstruct the genome up to a chromosome length (Miller et al., 2010). Using sequencing reads, those methods try to reconstruct the genome by joining the reads. Different methods have been employed

on this problem based on graph structures including greedy graphs, OLC (Overlap/Layout/Consensus) methods, and *De Bruijn* graphs. Numerous assemblers have been developed, we only name few here including SSAKE (Warren et al., 2007), Newbler (Margulies et al., 2005), ABySS (Simpson et al., 2009), and Cortex (Iqbal et al., 2012).

Assembly algorithms face a lot of challenges given the small length of the sequencing reads from NGS which makes assembly even more challenging. A possible improvement given the short reads is deep sequencing, but still repetitive regions will be indistinguishable. This issue could be resolved with longer reads that will span outside the repetitive regions. Also, the coverage of the genome is not uniform and some regions may have no coverage at all which will be omitted from the assembled genome (Alkan et al., 2011). However, the assembly algorithms give the opportunity to analyse organisms with no reference assembly at the moment. The advance of the NGS technologies will enhance the improvement of those methods as well.

1.4.3 Calling variants from NGS

After sequencing, alignment, and recalibration, information regarding the bases mapped and the corresponding qualities in terms of base and mapping quality, is given for each site on the genome. Different methods have been developed for the detection of different types of genetic variants (e.g. SNPs, indels and CNVs). For the rest of this thesis, we only focus on methods for detecting and genotyp-

ing SNPs from low coverage data. For a review of the methods developed for detecting indels and structural variants, refer to the supplementary material of The 1000 Genomes Project Consortium (2010).

1.5 Description of the SNP and genotype calling problem from low-coverage NGS data

After the sequencing reads are mapped to a reference genome, the information for a physical position in the genome for one individual is summarized with the alleles that have been mapped to that position and the accompanying base and mapping qualities. Note, that coverage is not uniform across the genome, and across different individuals, and there might be positions covered by many reads and positions with no reads at all. Using information from all individuals at a position, we can infer positions that are covered by more than one allele or that are covered by an allele different than the reference allele. In those cases, there is a possibility that the position is polymorphic. SNP detection and genotyping algorithms can be first divided into two categories: algorithms that consider one individual at a time, and algorithms that consider jointly a sample of individuals.

The first category ideally requires the deep sequencing of a single individual. The sequencing reads are mapped into the reference genome and sites are called polymorphic if bases mapped on the reference genome differ from the reference allele. Quality measures are taken into account to filter out false pos-

itives. The first deep sequencing of a single individual is the genome of Craig Venter (Levy et al., 2007). Sanger sequencing was used to sequence the genome and 4.1 million polymorhisms were detected, 3.2 million of them being SNPs. Subsequent whole genome sequencing of a single individual using NGS technologies (Wheeler et al., 2008; Wang et al., 2008; Bentley et al., 2008; Kim et al., 2009) revealed the potential advantages of NGS in terms of information and cost.

Low coverage/large sample versus high coverage/small sample

Nevertheless, studying one individual at a time does not answer questions about population structure and differences in allele frequencies in different groups of individuals with different phenotypes, which is important for GWAS studies. Also, even though the cost of sequencing has decreased considerably with the advent of the NGS, it is still costly to deep sequence a large sample of individuals. A possible design is to deep sequence target regions, but given that the majority of GWAS signals have been found in "unexpected" regions, makes the design of deep sequecing of target regions only less attractive. In addition, the demand for larger sample sizes suggests that a low to medium coverage designs will be more desirable (Nielsen et al., 2011).

So, there is a question regarding the design of different projects: with a fixed sequencing effort, what is the best design? Sequence few individuals in high coverage or a large sample of individuals in lower coverage. This is one of the main questions 1KGP investigated with the pilot phase of the project. If we se-

quence individuals in higher coverage, we will confirm with higher accuracy the variants present in the sample but we fail to discover variants that are not present in the sample (Li, 2011a). On the other hand, with low coverage, there is a higher probability that one of the two chromosomes in an individual is not sequenced, and thus we are missing some true polymorphism present in the sample. The benefits of low-coverage sequencing for a sample of individuals is examined in the low-coverage pilot of the 1KGP. Given a sample of 179 individuals, the 1KGP identified 14.4 million SNPs versus $\sim$6 million SNPs identified by the second pilot where 2 trios were sequenced by high coverage. The genotype accuracy at low-coverage is lower for rare SNPs but reaches the accuracy of the high-coverage design for common SNPs. Therefore, with a low coverage/large sample design, we maximize the number of variants discovered but we have a lower power to detect rare variants and achieve lower genotype accuracy rates, whereas with a high coverage/small sample we discover accurately all the variants present at the sample with high genotype accuracy but we miss the variation not present in the sample. The benefit of the different designs has been recently studied for their impact in association studies (Li et al., 2011) and it was demonstrated that low-coverage designs are more informative than high coverage designs when we want to detect modest effect sizes in a large sample.

However, both in the low-coverage pilot of the 1KGP and the study by Li et al. (2011), the benefit of the low-coverage designs, is maximized by the development of algorithms that combine information from low-coverage sequencing reads and LD information especially for SNP genotyping. These methods are

the main focus of this thesis.

Low coverage & LD information

We illustrate the advantage of sequencing a larger sample of individuals at lower coverage, given a fixed amount of resources, with a schematic representation in Fig.1.2. We present the haplotypes of 5 diploid individuals and we colour the haplotypes according to their ancestry. Each haplotype is sequenced at 2x coverage. However, given that the haplotype blocks with the same colour share common ancestry, we can combine the information across those haplotypes and obtain more information. This is shown in Fig.1.2 for the haplotypes indicated by the arrows. Given that four haplotypes share common ancestry, if we combine the information, then we have 8x coverage for this haplotype. Therefore, by studying jointly a sample of individuals at once, we gain information from the LD patterns.

The question then reduces to how to model the LD information across samples. The coalescent model and approximations of it have been extensively used to model the LD patterns in a sample of individuals for phasing and imputation problems (e.g. PHASE (Stephens et al., 2001; Stephens and Scheet, 2005), MACH (Li et al., 2006), IMPUTE (Marchini et al., 2007)),. We thus present in the next section a brief summary of the main features of the coalescent model and one of its most popular approximations, the Li and Stephens model (Li and Stephens, 2003).

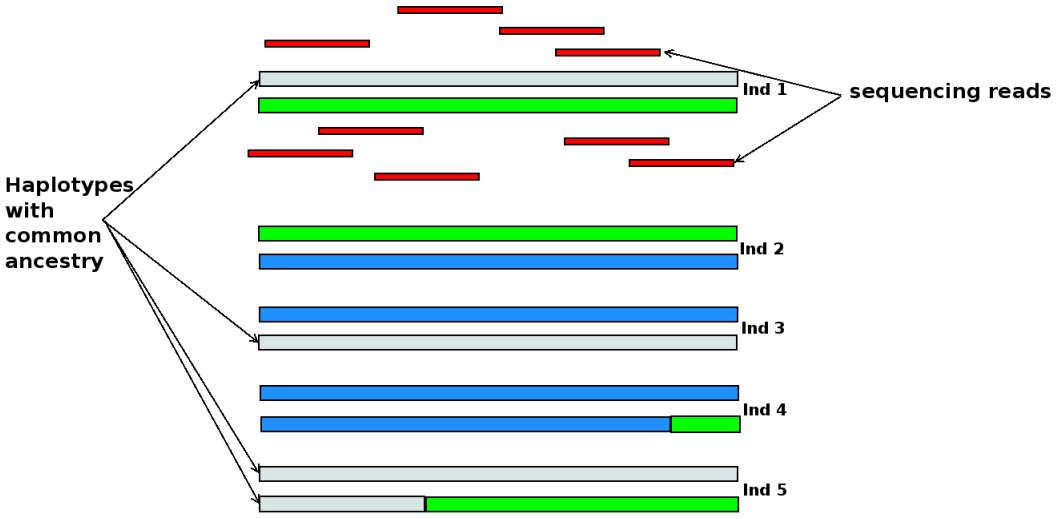

Figure 1.2: Schematic drawing of population-scale genome sequencing. The haplotypes of five diploid individuals are presented which have 2x coverage. The haplotypes are coloured according to the ancestral haplotype they have descended from. If we merge the information coming from the same ancestral haplotypes we gain more information. On the drawing, the grey haplotype is sequenced on average 8 times.

1.6 The coalescent model and its approximations

In the previous section we discussed the incorporation of LD in SNP and genotype callers to enhance mainly the genotyping step. How can the haplotypes of the sample aid the detection of SNPs? Haplotypes contain information about the genealogical history of the samples and thus on the relationships between haplotypes. The genealogical history of the samples includes mutation and recombination events, population structure, forces of natural selection, and random transmission of genetic material (Wakeley, 2008). By understanding the forces that shaped the present day haplotypes, we can answer questions regarding those forces (in our case, mutation events). Unfortunately, the genome of any organism undergoes highly complex random processes and it is extremely challenging to reconstruct the true genealogy. We thus rely on statistical meth-

ods to approximate the genealogy of the sample.

1.6.1 The Wright-Fisher model

The first and most basic model for the transmission of genetic material is the *Wright-Fisher* model (Wright, 1931; Fisher, 1930) which makes the following assumptions:

- the population size N is constant and finite
- the generations are non-overlapping
- migration and selection is not permitted
- the offsprings in population $k+1$ are randomly sampled from generation k with replacement

The model assumes haploid organisms and the existence of only one mating type. The assumptions oversimplify the genetic process but it allows it to be expressed mathematically and yields results which can be used as approximations of the true genealogy. The model considers the history of the offsprings from a prospective point of view, from the past to the present. However, this poses some difficulties since some parents may have not reproduced and thus have no offsprings after some generations. In Fig.1.3 on the left plot, we can see that some lineages have been lost. The coalescent model, or the *n coalescent* as coined by Kingman (1982), abandons the idea of following the future evolution of the

sample and instead studies the history of the process given the present day offsprings in a retrospective manner. The coalescent follows the same principles as the *Wright-Fisher* model when the population size $N \to \infty$.

It should be noted that in the course of time, the rather unrealistic assumptions made in the *Wright-Fisher* model have been abandoned, and more complex scenarios have been incorporated, including variable population size, recombination, migration, population structure, and selection. However, the aim of this section is to present the idea of the coalescent model and the simplified version is preferred.

Let us consider a population of size 2N. Given the assumptions above, the probability that two offsprings choose the same parent is $\frac{1}{2N}$ and thus the probability they choose a different parent is $1 - \frac{1}{2N}$. If T is the time in generations until two offsprings choose the same parent, then

$$\mathbb{P}(T > t) = \left(1 - \frac{1}{2N}\right)^{t-1} \tag{1.2}$$

where for $t - 1$ generations the two offsprings choose different parents. T follows a geometric distribution with $p = \frac{1}{2N}$. Thus, the expected time until the two offsprings choose the same parent is 2N.

Now let us assume that we have a sample of n offsprings where $n \ll 2N$. The probability that the n offsprings choose different parents from the previous generation is given by

$$\frac{2N-1}{2N} \times \frac{2N-2}{2N} \times \cdots \times \frac{2N-n+1}{2N} = \prod_{i=1}^{n-1} \left(1 - \frac{i}{2N}\right) \tag{1.3}$$

since the first offspring has no restriction of choosing its parent, the second offspring has to choose one of the $2N-1$ remaining parents, and so on. Expanding the right hand side of Eq.1.3, we get

$$1 - \sum_{i=1}^{n-1} \frac{i}{2N} + O\left(\frac{1}{4N^2}\right) = 1 - \frac{\binom{n}{2}}{2N} + O\left(\frac{1}{4N^2}\right) \tag{1.4}$$

where the third term vanishes as $N \to \infty$. If we rescale time in generations with $\tau = \frac{t}{2N}$, then the time T_n until 2 of the n offsprings choose the same parent is

$$\mathbb{P}(T_n > \tau) = \left[\lim_{N\to\infty} \left(1 - \frac{\binom{n}{2}}{2N}\right)^{2N}\right]^{\tau} = e^{-\binom{n}{2}\tau} \tag{1.5}$$

which follows an exponential distribution $T_n \sim exp\left(\binom{n}{2}\right)$, where T_i is an independent random variable. Using this result we can show that the expected time of coalescence in a sample size n (two offsprings choose the same parent, i.e. merge) is $\frac{2}{n(n-1)}$. When the last two offsprings coalesce, the sample has reached the Most Recent Common Ancestor (MRCA).

Using the distribution of T_n, we can calculate the expected height H_n, and the total branch length L_n of the coalescent tree, for a sample of size n. The height of the coalescent tree, or the Time to the Most Recent Common Ancestor (TMRCA), is given by

$$H_n = \sum_{i=1}^{n} T_i \tag{1.6}$$

and its expectation is

$$E(H_n) = \sum_{i=1}^{n} E(T_i) = \sum_{i=1}^{n} \frac{2}{i(i-1)} = 2\left(1 - \frac{1}{n}\right) \tag{1.7}$$

An interesting observation is that the coalescent time of the last two offsprings follows $T_2 \sim exp(1)$ and thus most of the height of the tree involves only two offprings.

The total branch length of the tree is

$$L_n = \sum_{i=1}^{n} iT_i \tag{1.8}$$

and its expectation following the same arguments as in Eq.1.7 is

$$E(L_n) = \sum_{i=1}^{n} iE(T_i) = 2\sum_{i=1}^{n} \frac{1}{i-1} = 2\sum_{i=1}^{n-1} \frac{1}{i} \tag{1.9}$$

Until now, we have only considered the coalescent events that form a genealogy through the random sampling of parents from the previous generation, also known as genetic drift. An important feature of both the *Wright-Fisher* model and the coalescent is that, when mutations are selectively neutral, they can be considered seperately from the coalsecent events. An example of a Wright-Fisher population genealogy and the corresponding genealogical tree is given in Fig.1.3, where the population size remains constant across generations. The

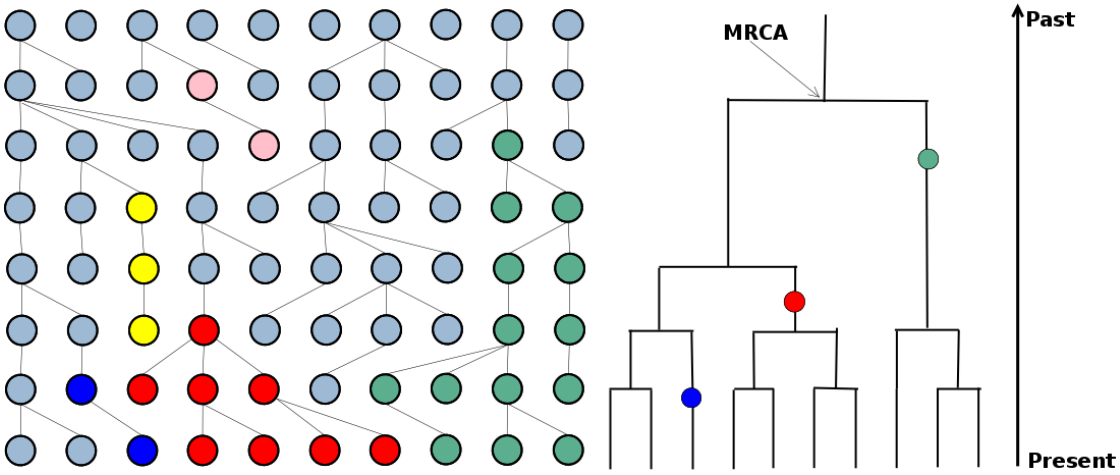

Figure 1.3: Pictorial example of a population process on the left and the corresponding genealogical tree on the right. Time goes from present time to the past. The left hand side graph represents the evolution of a sample of size n =10 for 8 generations. Each generation chooses its parents randomly with replacement. Mutation events are represented by the change of colour. In the present day sample only three out of the five mutations are maintained. On the right hand side we display the corresponding genealogical tree. As we move backwards in time, the lineages coalesce until the MRCA is reached. The mutation events are represented by circles.

genealogical tree of the present time sample can be constructed ignoring the mutation events (change of colour on the left hand side graph). The mutation events can then be added to the genealogical tree.

The above model assumes asexual organisms which is not the case for humans. If we want to construct genealogies for sexually reproducing organisms, we need to consider the effects of *recombination*. Recombination is the exchange of genetic materials between two chromosomes. By incorporating recombination in the coalescent model, we now assume that each offspring can have more than one parent. Hence, if there is a coalescent event, the number of lineages in the previous generation decreases by one, but in the case of recombination, it increases by one. When recombination is added in the population process, the

genealogy of a sample of segments can not be represented by a single tree but rather in a graph, the so-called Ancestral Recombination Graph (ARG) (Griffiths and Marjoram, 1996). Then, from the ARG, we can sample marginal genealogical trees for each site that might not be the same. The coalescent model is central in the population genetics field, however, its complexity and computational cost induced the development of approximations of the model.

In the remainder of this section we are going to present different methods that model the LD structure. We commence with the tree-based models, that they construct genealogical trees either in a heuristic way or by applying the coalescent model, and make inference based on the sampled trees. Approximation of the coalsescent model follow in an order of simplicity. By making further approximations we gain computational speed on the cost of simplifying assumptions that are discussed in each of the models presented. A first widely used approximation of the coalescent, that benefits a lot of its features, is the Li and Stephens PAC model which assumes that a newly seen haplotype is a mosaic construction of previously seen haplotypes. A further approximation of the LD is that of the representation of the haplotypes on a graph. Lastly, a further approximation of the coalescent met in the literature is that of the multivariate normal approximation, which assumes that a newly seen haplotype follows a multivariate normal distribution, with parameters estimated from the previously seen haplotypes. For each of the models, we give examples of softwares that use those models and how those are adapted for the genotype calling from next-generation sequencing reads.

1.6.2 Tree based methods

Tree based methods use genealogical trees for inference. The genealogical trees are constructed using the corresponding models that make assumptions about the mutation and recombination events Hallsworth (2010) or a more heuristic way Minichiello and Durbin (2006). Those methods assume that the haplotypes on a set of sites are known a priori and can be used for the build of the genealogical trees. Then, using the genealogical trees, different mutation events can be superimposed on the tree and the likelihood of each of the mutation events estimated. There are two tree based methods developed for genotype calling: QCALL and TreeCall[14].

QCALL (Le and Durbin, 2011) uses the infinite sites assumption where there is at most one mutation at a given position. QCALL assumes that a scaffold of haplotypes is available for the individuals in the study at a set of known sites. Using MARGARITA (Minichiello and Durbin, 2006), ARGs are built using the haplotype scaffold, and then genealogical trees are sampled from the ARGs for each candidate site. Using those trees, the likelihood of a genotype configuration can be calculated given an ancestral and a mutational allele.

Since the true genealogy of the samples is unknown, multiple trees are sampled and the results are averaged over the sampled trees.

More formally, for a given site, let us denote Δ to be the case there is a mutation on the tree, $\bar{\Delta}$, the case there is no mutation, T the genealogical tree, and R, the information from the sequencing data. Then, using Bayes' rule, the posterior

[14]This method will be presented in detail in the following chapter.

probability of a mutation given the tree T and the sequencing data R is

$$\mathbb{P}(\Delta \mid R, T) = \frac{\mathbb{P}(R \mid \Delta, T)\mathbb{P}(\Delta \mid T)}{\mathbb{P}(R \mid \Delta, T)\mathbb{P}(\Delta \mid T) + \mathbb{P}(R \mid \bar{\Delta}, T)\mathbb{P}(\bar{\Delta} \mid T)} \tag{1.10}$$

where the prior $\mathbb{P}(\Delta \mid T) = \theta \sum_{i=1}^{2m} 1/i$. Let us consider first the likelihood of no mutation. Assuming the individuals are unrelated, the likelihood is a simple product of their genotype likelihoods. Note that at a given position there is also information coming from the reference genome and the site is monomorphic if the genotypes in the study sample are the same as the reference allele r. Thus, the likelihood of no mutation is given by

$$\mathbb{P}(R \mid \bar{\Delta}, T) = \sum_{r} \mathbb{P}(r) \prod_{i=1}^{n} \mathbb{P}(R_i \mid G_i = rr) \tag{1.11}$$

where $\mathbb{P}(r) = 1 - \epsilon$ if r is the reference allele and $\mathbb{P}(r) = \epsilon/3$ otherwise[15]. To calculate the likelhood of a mutation, the likelihood of each possible mutation is considered and it is averaged over all sampled trees t_k, thus

$$\mathbb{P}(R \mid \Delta, T) = \sum_{r} \mathbb{P}(r) \sum_{t_k} \mathbb{P}(R \mid \Delta, t_k, r)\mathbb{P}(t_k) \tag{1.12}$$

where $\mathbb{P}(R \mid \Delta, t_k, r)$ corresponds to the sum of likelihoods of all possible mutations at tree t_k with ancestral allele r. Using a threshold in Eq. 1.10, a decision is made whether a site is polymorphic or not. For the sites called as polymorphic, the posterior probabilities of the genotypes of the individuals under study are calculated by

[15]Using empirical experiments in 1KGP $\epsilon = 2 \times 10^{-5}$.

$$\mathbb{P}(G_i = ab \mid R, \Delta, T, r) = \frac{\mathbb{P}(R, G_i = ab \mid \Delta, T, r)}{\sum_{a',b'} \mathbb{P}(R, G_i = a'b' \mid \Delta, T, r)} \tag{1.13}$$

where $\mathbb{P}(R, G_i = ab \mid \Delta, T, r)$ is the sum over likelihoods of an ancestral and a mutation pair that results in $G_i = ab$ for individual i.

The tree based methods benefit from the fact that they model the genealogical history and can incorporate genetic events like mutation, recombination, migration and selection. However, at the current stage the models for the construction of genealogical trees are compuationally intensive and can not be applied to very large samples. In addition, all sampled trees will always inherit some error that will be propagated in the genotype calling step. QCALL and TreeCall have been applied on the low-coverage pilot of the 1KGP.

1.6.3 Li and Stephens Model

A well-known model that can be used to capture salient features of the coalescent with recombination is that of Li and Stephens (2003). The idea that governs this approximation is that a new haplotype can be reconstructed as an imperfect mosaic of the previous seen haplotypes (see Fig.1.4). This idea is expressed in the *Product of Approximate Conditionals* (PAC) likelihood. Given a recombination rate ρ, the probability of a sample of n haplotypes $h_1, h_2, ..., h_n$ is given by

$$\mathbb{P}(h_1, h_2, ..., h_n \mid \rho) = \mathbb{P}(h_1 \mid \rho)\mathbb{P}(h_2 \mid \rho; h_1) \ldots \mathbb{P}(h_n \mid \rho; h_1, h_2, ...h_{n-1}) \tag{1.14}$$

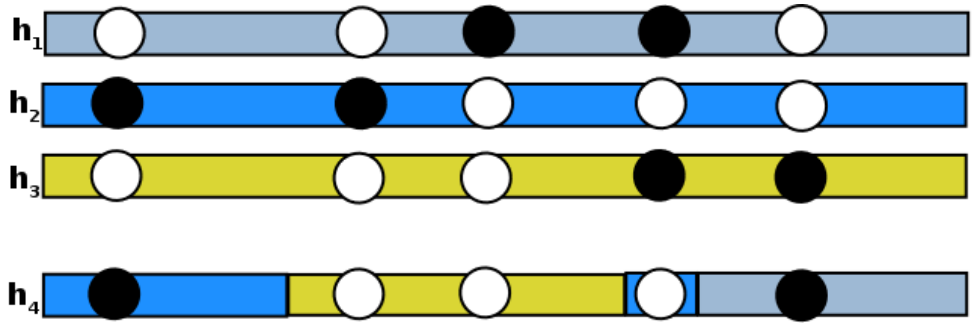

Figure 1.4: Li and Stephens pictorial example. In this diagram, we have previously observed haplotypes h_1, h_2 and h_3 where the dots represent mutations. We construct h_4 as an imperfect mosaic of the previous haplotypes. h_4 copies from h_2, h_3, h_2, and h_1 with a mutation on the last SNP.

where the conditional distributions are unknown and are approximated by distribution π. The distribution π needs to satisfy the following properties:

1. The next haplotype is more probable to match a previously seen haplotype that has been met many times before than one that has been met less frequently,

2. As n increases, the probability of seing a new haplotype decreases,

3. The probability of seing a new haplotype increases as the mutation rate θ increases,

4. If a new haplotype does not match perfectly a previously seen haplotype, then it differs by a number of mutations from a previously seen haplotype rather than being unique compared to previously seen haplotypes,

5. When the recombination rate is low, the probability of matching previous haplotypes is higher over large genomic regions and vice versa.

Fearnhead and Donnelly (2001) provided an approximation that satisfies all the aforementioned properties. The Li and Stephens model uses a simplification of the conditionals proposed by Fearnhead and Donnelly (2001) that still satisfy all the five properties.

Thus, the PAC likelihood is expressed

$$L_{PAC} = \pi(h_1 \mid \rho)\pi(h_2 \mid \rho; h_1)...\pi(h_n \mid \rho; h_1, h_2, ...h_{n-1}) \tag{1.15}$$

where $h_i = \{h_{i1}, h_{i2}, ...h_{iL}\}$, assuming the window of interest consists of L biallelic SNPs. The conditionals of the Li and Stephens model are calculated using a HMM. We begin by sampling the first haplotype, which is independent of ρ, with probability $\pi(h_1 \mid \rho) = 1/2^L$. Subsequently, for $n > 1$, denote the sequence of copied states for haplotype h_{n+1} as $Z_{n+1} = (Z_{n+1,1}, ...Z_{n+1,L})$.

Recombination is modelled by transition probabilities where the current haplotype switches from copying from one haplotype to copying from another. The first state is sampled from a uniform distribution. For the subsequent transitions the probability is given by

$$\mathbb{P}(Z_{n+1,l} \rightarrow Z_{n+1,l+1}) = \begin{cases} e^{\frac{\rho_l}{n}} + \frac{1-e^{\frac{\rho_l}{n}}}{n} & Z_{n+1,l} = Z_{n+1,l+1} \\ \frac{1-e^{\frac{\rho_l}{n}}}{n} & Z_{n+1,l} \neq Z_{n+1,l+1} \end{cases}$$

where $\rho_l = 4N_e r_l$, where N_e is the effective population size and r_l is the genetic distance between sites l and $l+1$.

The mutation process is modelled by allowing the new haplotype to copy imperfectly an observed haplotype, presented as the emission probabilities which

are given by

$$\mathbb{P}(h_{n+1,l} \mid Z_{n+1,l}, h_1, ..., h_n) = \begin{cases} \frac{n}{n+\theta} + 0.5\frac{\theta}{n+\theta} & h_{n+1,l} = h_{Z_{n+1,l},l} \\ 0.5\frac{\theta}{n+\theta} & h_{n+1,l} \neq h_{Z_{n+1,l},l} \end{cases}$$

This model has been extensively applied to imputation models (Stephens and Scheet, 2005; Li et al., 2006; Marchini et al., 2007). MACH has been implemented for phasing and imputation, and it is now adopted to handle sequence information (Thunder). To begin with, a vector of unphased genotypes $\vec{G}$ is to be phased given the information of L flanking phased sites H. In addition, let us define $\vec{Z} = \{Z_1, Z_2, ..., Z_L\}$ variables indicating the mosaic state of $\vec{G}$ (Li et al., 2010). The joint probability of $\vec{G}$ and $\vec{Z}$ is thus given by

$$\mathbb{P}(\vec{G}, \vec{Z}) = \mathbb{P}(Z_1) \prod_{j=2}^{L} \mathbb{P}(Z_j \mid Z_{j-1}) \prod_{j=1}^{L} \mathbb{P}(G_j \mid Z_j) \tag{1.16}$$

where the first term is the prior probability of the first mosaic state and it is uniform across all possible states. The first product multiplies over a possible path from Z_1 to Z_2, from Z_2 to Z_3 and so on, the so called transition probabilities that take into account the recombination rate, and the second product considers the emission probabilities $\mathbb{P}(G_j \mid Z_j)$ that accounts for mutation, genotype error and gene conversion (Li et al., 2010). The algorithm is Markovian since the current state Z_j only depends on the previous state Z_{j-1}. The method begins by assigning an initial set of haplotypes for each idividual that agrees with the unphased genotype and then at each iteration step an individual is updated given the population allele frequencies.

The method is extended to handle sequence data. In the case of sequence data, we do not know the *true* genotypes at a position, but we have the genotype likelihoods. Therefore, the genotype likelihoods are fed into the emission probabilities. The emission probability is now given by

$$\mathbb{P}(R_j \mid Z_j) = \mathbb{P}(G_j \mid Z_j)\mathbb{P}(R_j \mid G_j) \tag{1.17}$$

where R is the sequence read information.

A further method that utilizes the idea of the Li and Stephens model is SNPTools. *SNPTools* is a new method for SNP and genotype calling[16]. It is a self-sufficient package since it inputs the raw sequence information and calculates its own genotype likelihoods that are used in the SNP and genotype model. The imputation model is coined as *Constrained "Fearnhead, Donnelly, Li and Stephens" (FDLS) distribution* model where each haplotype is a mutated recombinant of only two haplotypes from the population. The method employs a Metropolis-Hastings sampler to sample the "parental" haplotypes, and then the haplotypes of the individual are sampled according to the haplotypes of the parents. At the time of submission of this thesis the method remains unpublished.

1.6.4 Localized clustering models

Another class of models for representing LD structure are the localized clustering models (Greenspan and Geiger, 2004; Kimmel and Shamir, 2005; Scheet and Stephens, 2006). This class of models has been used for phasing problems with a

[16]http://www.hgsc.bcm.tmc.edu/cascadetechsoftware_snp_toolsti.hgsc

good example being fastPhase. The core idea of the localized haplotype clustering model is to combine similar haplotypes together and summarize the information of each cluster. Here, we present a widely used model of this category that has been used for association mapping as well as phasing and imputation, and now further extended for handling genotype likelihoods, BEAGLE (Browning, 2006; Browning and Browning, 2007b,a). The initial representation of the haplotypes is made on a bifurcating tree. For example, in Fig.1.5, the initial node represents the whole set of haplotypes prior to to any processing, and the final node represents all haplotypes after all markers are processed. Commencing from left to right, each of the edges represents an allele which is weighted by the number of haplotypes carrying that allele at that position. The same procedure is repeated for all SNPs on the window analyzed and a weighted bifurcating tree is constructed that represents all the haplotypes present in the set. (see for example Fig. 1.5). BEAGLE models recombination by merging nodes if given the allele at position t, the probability of seing a set of alleles at markers $t+1$,$t+2$,$t+3$,.. is similar. Therefore, a directed acyclic graph (DAG) is created from the bifurcating tree. After the DAG is constructed, each edge on the graph represents a local haplotype cluster.

In the case of phasing, given that the haplotypes are unknown, an initial set of haplotypes that agree with the given genotypes is constructed, H^0. Then, at each iteration step t, using the current set of haplotypes H^t, a DAG is constructed M^t. Then, for each individual the probability of the haplotype given the genotypes and the DAG is calculated, i.e. $\mathbb{P}(H_i \mid G_i, M^t)$.

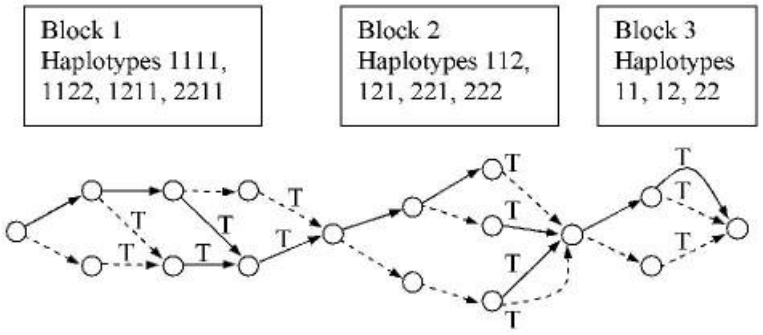

Figure 1.5: Figure extracted from Browning (2006) where the haplotypes are represented on tha graph with the corresponding haplotypes in the boxes above. In the graph, allele 1 is represented by a solid line whereas allele 2 by a dashed line.

The benefits of BEAGLE lies primarily on the speed and on the ability to handle a large number of samples. Also, the model does not assume a specific population genetic model (Browning and Browning, 2007a).

1.6.5 Multivariate Normal approximation model

A further approximation suggested in Wen and Stephens (2010) is instead of assuming that a new haplotype is a mosaic of the previously seen haplotypes, the conditional distribution of the new haplotype h^*, given a set of known haplotypes M, follows a multivariate normal distribution where the model parameters μ and Σ are estimated from M. A connection with the HMM-based methods remains however, in how the estimates of μ and Σ are calculated. Shrinkage is applied on the covariance matrix Σ that accounts for the recombination rate between any two SNPs and thus reduces the noise in the data. With this approximation, only the pairwise correlation between sites is considered, compared to all the aformentioned methods. In addition, the approximation is not directional and the sites are exhangable, a feature that is not desirable. However, by apply-

ing shrinkage on the covariance matrix directionality is imposed on the model. With this model, no mutation model is required and it is a simple off-the-shelf statistical model that can be applied on genetic data. We will discuss in further detail this model in Chapter 3 as it is the basis of the second genotype model presented in this thesis.

1.7 Pipeline for SNP and genotype calling from low-coverage data

Following sequencing, mapping, and recalibration (see Section 1.4), the sequencing information for each individual at each position is summarized by the (i) bases mapped at this position, (ii) the accompanying base qualities, and (iii) the mapping qualities of the reads mapped at this site. This information is used to calculate the probability of the reads given a potential underlying genotype, for all ten possible genotypes, i.e. $g \in \{AA, AC, AG, AT, CC, CG, CT, GG, GT, TT\}$. This probability is termed the *genotype likelihood* (Nielsen et al., 2011). The genotype likelihood for a position for genotype g_i is defined as

$$GL(g_i) = \mathbb{P}(R, Q \mid G = g_i) \tag{1.18}$$

$\forall i = 1, 2...10$ where R is the information from the sequencing reads, i.e. the bases overlapping to the position, and Q represents the corresponding map-

ping and base qualities[17]. The genotype likelihoods are usually reported in log-scale units. For example, assume that for a particular position $-10log_{10}\mathbb{P}(R, Q \mid AA) = 0$, $-10log_{10}\mathbb{P}(R, Q \mid AG) = 10$, and $-10log_{10}\mathbb{P}(R, Q \mid GG) = 20$. This can be interpreted as genotype "AA" is ten times more likely than genotype "AG" and a hundred times more likely than "GG". When a position is not covered by any reads, the genotype likelihoods are equal for all possible genotypes. Two methods that estimate the genotype likelihoods are Samtools (Li et al., 2008a), GATK[18] (DePristo et al., 2011) and SNPTools[19].

Using the sequencing information we then move on the SNP discovery and SNP genotyping steps. The former, involves non-LD based methods, that are usually faster than LD-based methods (see Appendix A for a description of various methods), that use only sequencing information to detect putative SNPs. In the SNP discovery procedure, it is also important to detect error modes in the data generation and alignment, in order to exclude false positives arising from sequencing[20] or alignment[21] artifacts. Hence, the non-LD methods play a crucial role in the SNP discovery step of the NGS pipeline, and provide a set of SNPs that will then be fed into LD-methods for genotype calling. Using the set of SNPs, the role of the LD-based methods (for examples models that have been introduced in the previous section) is to determine the genotypes of the individuals at those sites.

[17] In subsequent sections we drop Q and write the genotype likelihood as $\mathbb{P}(R \mid g)$ for notation simplicity.

[18] GATK uses a similar model to that of Samtools for SNP calling

[19] http://www.hgsc.bcm.tmc.edu/cascadetechsoftware_snp_toolsti.hgsc

[20] For example strand bias, tail distance bias

[21] For example very low/high coverage at a site, misalignment due to nearby indels

1.8 Summary

In this chapter, we introduced the advent of the new generation sequencing data and the urge to develop statistical models to identify and genotype genetic variants. We also presented the results from the 1KGP consortium, the first large project that uses NGS data to provide a detailed catalogue of human polymorphism, similar to that of the HapMap project, but with variants reaching as low as 1% MAF as well as indels and other structural variants. We then focused on the SNP and genotype calling problem from multi-sample low coverage sequencing data, which is to be the main focus for the next chapters. A two step procedure was adopted for variant calling. Firstly, a fast non-LD based method scans the genome and excludes clearly monomorphic sites given only the sequencing information. Then, LD-based methods are applied on the candidate set created, to refine the SNP calls[22] and/or to provide the genotypes at the called sites.

In the following two chapters, we will introduce our own SNP and genotype calling methods and then compare them with other competing methods, all of which were discussed in this chapter. In Chapter 2, we will present a SNP and genotype calling method that employs marginal geneological trees. In Chapter 3, we present a second SNP and genotype calling method based on an approximation of the distribution of haplotypes. In Chapter 4, we compare our new approaches with other competing methods using data from the low-coverage pilot of the 1 KGP, in terms of SNP and genotype calling. In Chapter 5, we make

[22]This was the pipeline applied for the first wave of data from 1KGP but not in the subsequent phases.

comparisons in terms of genotype calling on the Phase 1 of the 1KGP.

Chapter 2

SNP and genotype calling using marginal genealogical trees

Genetic polymorphism in a sample of chromosomes is the result of mutations occuring on the genealogical tree that links a sample of chromosomes (Nordborg and Tavaré, 2002). Chromosomes that carry the mutated allele are more closely correlated than chromosomes without the mutated allele, because they have coalesced more recently in the past. Polymorphic loci surrounding the mutated site are passed together, subject to recombination events, to the next generation and thus in LD with the mutated site. The coalescent theory presented in Section 1.6, models the genetic processes that occured in a sample of individuals backwards in time represented in the form of a genealogical tree. Thus, using genealogical trees, we can infer LD information in the sample studied.

A key assumption of the method presented, is the availability of known haplotypes at a subset of sites for the individuals in the study. Hence, we assume that the study design involves both the sequencing and genotyping of the indi-

viduals in the study. Using the genotypes from the genotyping step, we can use a phasing algorithm to phase the genotypes (from the genotyping technologies) and obtain a set of known haplotypes. Then, those known haplotypes can be used to contruct genealogical trees.

The algorithm described in this chapter combines information from NGS reads in the form of genotype likelihoods and LD information, through genealogical trees, to detect and genotype segregating sites. Using these two sources of information and assuming the infinite sites model, we can calculate the likelihood there is a mutation by summing up the likelihoods of all possible mutations on the tree, and comparing it to the likelihood there is no mutation on the tree, and thus a monomorphic site. The model presented, TreeCall, is built under a Bayesian framework where the Bayes Factor (BF) of the two models, with and without mutation, is calculated for each candidate site. By applying a threshold on the BF, depending on the prior information, we decide whether a site is segregating or not.

We begin this chapter by introducing the idea of the Bayesian model formulated giving toy examples. Subsequently, we provide a detailed derivation of the Bayes Factor as well as two procedures we developed for genotype calling using our model. In the last section, we discuss briefly the main features of the algorithm used for the construction of genealogical trees used in our method.

2.1 Description of the proposed model

A formal description of the new method for SNP and genotype calling follows. First, it is necessary to introduce some notation that will be used throughout the chapter:

- N : number of individuals,
- V_i : number of technologies that have sequenced individual i,
- T : tree at site of interest,
- R : all sequencing reads covering the site of interest in all study individuals,
- R_{ij} : reads at site of interest from sequencing technology j of individual i,
- g = genotype, $g \in \{AA, AC, AG, AT, CC, CG, CT, GG, GT, TT\}$,
- $G_{ij}(g) = -10log_{10}\mathbb{P}(R_{ij} \mid g)$: phred-scaled genotype likelihood for genotype g from the j^{th} sequencing technology of individual i,
- A_0 : reference allele,
- A_k : the remaining three alleles that are not equal to the reference allele, $k \in \{1, 2, 3\}$,
- M_0 : model of no mutation at site,
- M_1 : model of a mutation at site,
- $r \in \{A, C, G, T\}$: base at root of tree,

- b : branch of tree,
- l_b : relative length of the tree branch b (scaled so that the sum over all branches equals 1, $\sum_b l_b = 1$),
- $m \in \{A, C, G, T\}$: mutation allele,
- $c_h^l = \{c_h^{l_1}, c_h^{l_2}\}$: inferred genotype for individual l at configuration h, $l = 1, ..., N$,
- $c_h = \{c_h^1, ..., c_h^N\}$: h^{th} genotype configuration for $h = 1, .., M$, M is the number of all possible configurations, $M = 2(2N-1) \times 12 + 4$; where $2(2N-1)$ is the total number of branches on the tree; and there are 4 cases of a root allele and 3 cases of mutation allele (12) and 4 configurations that correspond to the case where all genotypes are homozygous (4 possibilities of root alleles). The allele configuration corresponds to the set of the inferred genotypes for all individuals under study given a mutation base, a root base, and the branch where the mutation occurs,
- E : event where the reference allele (A_0) is observed with an error, $\mathbb{P}(E) = p$,
- u : unobserved branch associated with the reference allele,
- F : event where the mutation lies on u, $\mathbb{P}(F) = q$.

We test whether a given position is segregating or not by comparing two models: model M_0, where there is no mutation on the tree at the site and thus

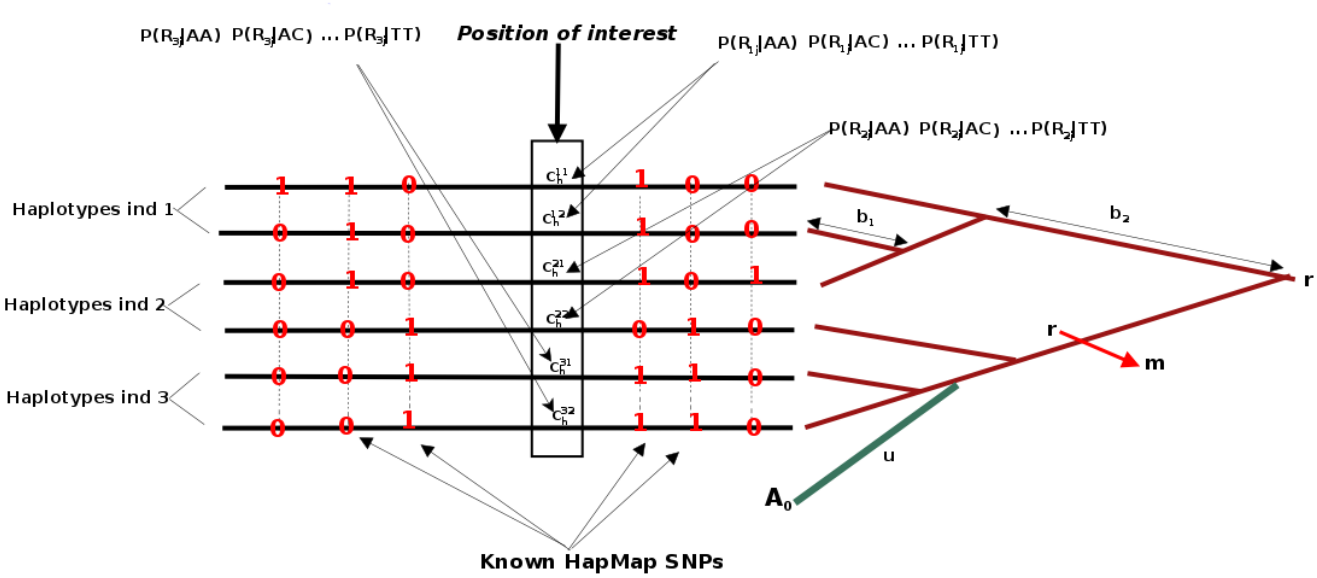

Figure 2.1: Schematic representation of SNP calling algorithm. The haplotypes of 3 individuals are displayed (with 3 SNPs at either side of the position of interest). Those SNPs are employed for the construction of the marginal genealogical tree at the position of interest. The genotype likelihoods at each individual corresponds to the information extracted by the NGS output technologies about the genotype of the individual at that position and it could be used to calculate the likelihood of genotype configurations inferred by a mutation on the tree. Also, the unobserved branch on the tree u, correspond to the reference allele.

the position is monomorphic, against model M_1, where there is a mutation on the tree and thus the position is polymorphic[1].

We first present pictorially the main idea of the method with $N = 3$ and $V_i = 1, \forall i$. In Fig. 2.1, the haplotypes of three individuals are shown. We incorporate known haplotypes at either side of the position of interest to construct a marginal genealogical tree at that position. Under M_0, there is no mutation on the tree, and given that the root base is r, then all the inferred genotypes for

[1]We assume the infinite-sites model holds where at most one mutation occurs at a given position.

all individuals will be "rr". We can calculate the likelihood of all genotypes being "rr" by summing the corresponding genotype likelihoods $\mathbb{P}(R_{ij} \mid G = rr)$ $\forall i$. The configuration of the alleles will be $\{c_h^{11}, c_h^{12}, c_h^{21}, c_h^{22}, c_h^{31}, c_h^{32}\} = \{r, r, r, r, r, r\}$. Under M_1 there is at most one mutation on the tree (under the assumption of infinite sites model). The mutation can occur in any of the $2(2N-1)$ branches on the tree. In Fig. 2.1, we give an example of a mutation occurring on a branch of the tree where r is mutated to m, where $r \neq m$. Assuming this mutation on the tree, the inferred alleles will be $\{c_h^{11}, c_h^{12}, c_h^{21}, c_h^{22}, c_h^{31}, c_h^{32}\} = \{r, r, r, m, m, m\}$ and the corresponding genotypes are $c_h^1 = rr$ genotype for individual 1, $c_h^2 = rm$ genotype for individual 2, and $c_h^3 = mm$ genotype for individual 3. We can then calculate the likelihood of the mutation lying on that branch by summing over the corresponding genotype likelihoods for each individual. This procedure is repeated for all possible root bases r and all possible mutation bases m. Note that the probability of a mutation lying on different branches of the tree is not equal, since the larger the branch the longer the time between the coalescent events, and subsequently the higher the probability of a mutation occuring on that branch. Going back to Fig. 2.1, the probability of a mutation is greater in branch b_2 than in branch b_1.

Up to now, we have only incorporated the information from the genotype likelihoods for the position of interest, and we have used the flanking haplotypes of the individuals under study to construct the marginal genealogical tree at the position of interest. We now introduce the information coming from the reference allele into the model. Let us consider the reference genome as an extra

individual not in the study sample. The reference allele is incorporated on the tree as an "unobserved" branch u, where we don't know where it lies on the tree, but which we can estimate its expected length using standard coalescent results. We should note that the only information we have for the reference genome is the reference allele, and that there is no NGS information for it. Also, in the construction of the genealogical tree, the haplotypes of the reference genome are not used. Considering only bi-allelic SNPs, if the reference allele is observed with no error then either the root base r or the mutation base m must be equal to A_0. The mutation can occur either in one of the tree branches or on the unobserved branch. The examples that follow assume $A_0 = A$. The possible mutations, given the reference allele is observed with no error, are given in Fig. 2.2. On the left, we consider the case where the mutation lies on the tree branches. We illustrate the two possible scenarios with two different colours. On the right, we depict the case where the mutation occurs on u. Assuming C as the alternative allele, we see that the genotypes of all individuals are monomorphic but different to the reference allele.

If the reference allele is observed with an error, when the mutation lies on a tree branch, the root base or the mutation base must be equal to the "true" reference allele as we see in Fig. 2.3 (a). Back to the example, if the reference allele is observed with an error it will be either C, G, or T ($A_0 = A$). We examine these cases separately. For example, if the alternative allele is C then the possible cases are the ones given in red in Fig. 2.3 (a), where either the root base or the mutation base includes the base C. We give an example of inferred alleles for the

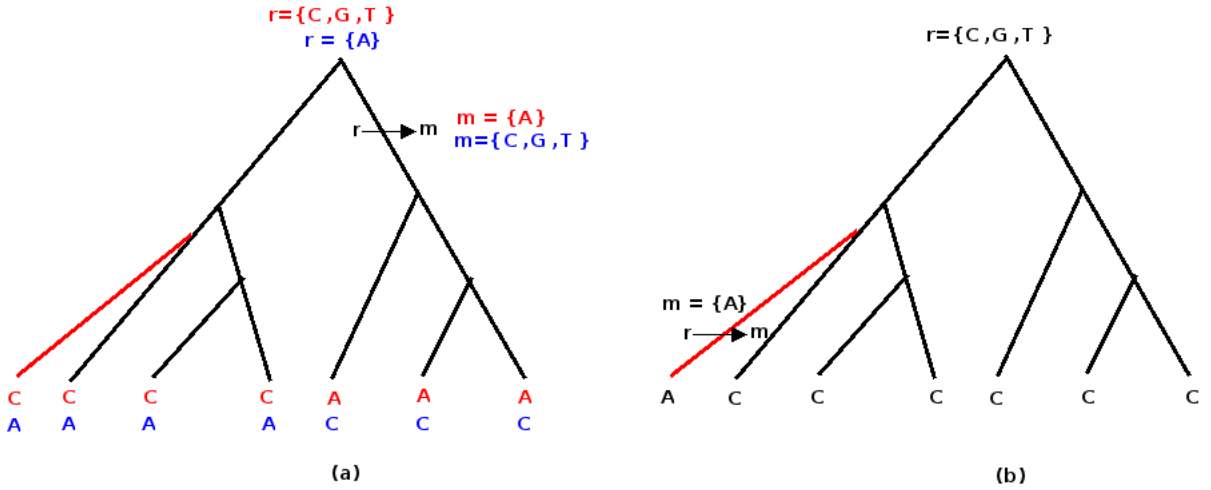

Figure 2.2: Model M_1 when reference allele is observed with no error. (a) Mutation lies on the tree. Assume $A_0 = A$, then to have a bi-allelic SNP either the root base (ancestral allele) or the mutation base must be equal to A_0. i) red : mutation base is equal to A_0, ii) blue: root base is equal to A_0. (b) Mutation lies on the unobserved branch of the tree. Since the reference allele is read with no error then the mutation base is fixed. The root base can be any of the other three bases. On the bottom of the tree, we show an example of the inferred alleles considering the first element of each set as the alternative allele.

three cases, and we can see that the polymorphic sites are bi-allelic. As shown in Fig. 2.3(b), when the reference allele is observed with an error and the mutation occurs on the unobserved branch of the tree, the mutation base is going to be one of the three alternative alleles. The root base can then be any of the bases but the mutation base.

In our new method, we propose to combine the LD information, in the from of genealogical trees, at a position of interest with the genotype likelihoods from the NGS to firstly decide whether a position is segregating by comparing the two scenarios, whether there is a mutation on the genealogical tree or not, and then by calling the genotypes at the study sample. For the scenario where there is one mutation on the tree, we consider all the different combinations of ancestral and mutation alleles and weigh them according to the length of the branch, where

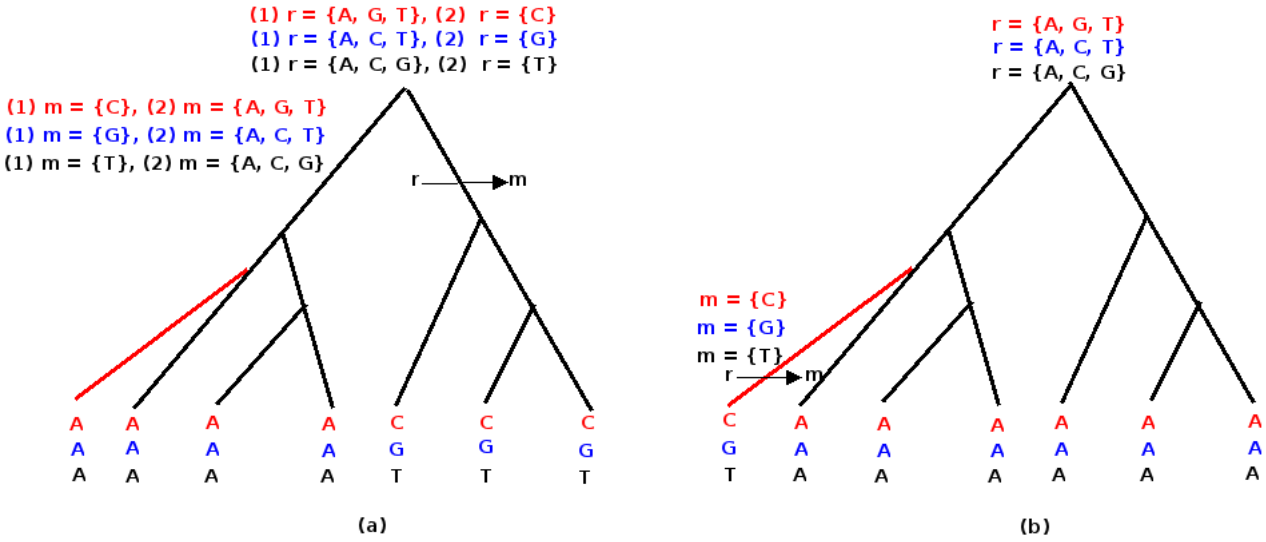

Figure 2.3: Model M_1 when reference allele is observed with error. If we assume that $A_0 = A$ is an error, then the reference allele will be either C, G, or T. To have a bi-allelic SNP, either the root base or the mutation base has to be equal to the reference allele (cases (1) and (2) on the tree). The three colours correspond to the three possible reference alleles: i) red: the true reference allele is C, ii) blue: the true reference allele is G, iii) black: the true reference allele is T. (a) corresponds to the case where the mutation lies on the tree, and (b) where the mutation lies on the unobserved branch. On the bottom of the tree, we show an example of the inferred alleles considering the first element of each set as the alternative allele.

the mutation allele occur as well as to whether the mutation or ancestral allele is the reference allele or not. In the following section, we present a detailed statistical description of our model.

2.2 Statistical framework of proposed SNP and genotype calling model

In this section, we formalise the idea of the SNP and genotype calling method introduced in the previous section. The method is developed to answer two questions: whether a position is segregating, and if the position is segregating

then, what are the genotypes of the individuals are under study. We consider models M_0 and M_1 corresponding to the cases where there is zero and one mutation at the site of interest respectively. We would like to compute the posterior odds

$$\frac{\mathbb{P}(M_1 \mid R, T)}{\mathbb{P}(M_0 \mid R, T)} = \frac{\mathbb{P}(R \mid T, M_1)}{\mathbb{P}(R \mid T, M_0)} \times \frac{\mathbb{P}(M_1)}{\mathbb{P}(M_0)} \tag{2.1}$$

The prior odds are based on our prior beliefs of the frequency of SNPs in the human genome. Therefore, we want to calculate the Bayes Factor,

$$BF(M_1, M_0) = \frac{\mathbb{P}(R \mid T, M_1)}{\mathbb{P}(R \mid T, M_0)} \tag{2.2}$$

We start solving Eq. 2.2 by calculating the likelihood of the model with no mutation at the position of interest. We approximate the prior probability that the reference allele is read with an error to be $p = 1/1000$. We again consider the two cases, reading the reference allele with an error or without. If the reference allele is read with **no** error, then the root base has to be equal to the reference allele, i.e. $r = A_0$, and all the genotypes are going to be A_0A_0. If the reference allele is read with an error, then there is equal probability that the position is going to be monomorphic and the root base is going to be $r = \{A_1, A_2, A_3\}$. Let $Z(T, r, i)$ be the inferred genotype on tree T with root base r at individual i. Then the probability of obtaining read R_{ij} given the implied genotype $Z(T, r, i)$ for individual i is given by

$$\mathbb{P}(R_{ij} \mid Z(T,r,i)) = 10^{\frac{-G_{ij}(Z(T,r,i))}{10}} \tag{2.3}$$

Thus, the probability of no mutation is

$$\begin{aligned}
\mathbb{P}(R \mid T, M_0) &= \mathbb{P}(R \mid \bar{E}, T) + \mathbb{P}(R \mid E, T) \\
&= \mathbb{P}(R \mid r = A_0, \bar{E}, T)\mathbb{P}(r = A_0, \bar{E} \mid T) + \sum_{k=1}^{3} \mathbb{P}(R \mid r = A_k, E, T)\mathbb{P}(r = A_k, E \mid T) \\
&= (1-p)\mathbb{P}(R \mid r = A_0, \bar{E}, T) + \frac{1}{3}p \sum_{k=1}^{3} \mathbb{P}(R \mid r = A_k, E, T) \\
&= (1-p)\mathbb{P}(R \mid Z(T, A_0)) + \frac{1}{3}p \sum_{k=1}^{3} \mathbb{P}(R \mid Z(T, A_k)) \\
&= (1-p)10^{\alpha^*_{T,A_0}} + \frac{1}{3}p \sum_{k=1}^{3} 10^{\alpha^*_{T,A_k}}
\end{aligned}$$

where $\alpha^*_{T,A_k} = \sum_{i=1}^{N} \sum_{j=1}^{V_i} \frac{-G_{ij}(Z(T,A_k,i))}{10}$

Since we are summing over small numbers[2], we define $Q^* = max(\alpha^*_{T,A_k})$, $\forall k = \{0,1,2,3\}$ and $\beta^*_{T,A_k} = \alpha^*_{T,A_k} - Q^*$.

Then,

$$\mathbb{P}(R \mid T, M_0) = 10^{Q^*}\left((1-p)10^{\beta^*_{T,A_0}} + \frac{p}{3}\sum_{k=1}^{3} 10^{\beta^*_{T,A_k}}\right)$$

and taking the logarithm of on both sides, we get

[2]This step is included for computational reasons and avoid rounding errors

$$log_{10}(\mathbb{P}(R \mid T, M_0)) = Q^* + log_{10}\Big((1-p)10^{\beta^*_{T,A_0}} + \frac{p}{3}\sum_{k=1}^{3} 10^{\beta^*_{T,A_k}}\Big) \tag{2.4}$$

In order to estimate the probability $\mathbb{P}(R \mid T, M_1)$, we scan through all possible mutations under each of the four possible scenarios described in the previous section. In detail,

$$\begin{aligned}\mathbb{P}(R \mid T, M_1) \quad &= \mathbb{P}(R \mid T, \bar{E}, \bar{F})\mathbb{P}(\bar{E}, \bar{F} \mid T) \\ &+\mathbb{P}(R \mid T, E, \bar{F})\mathbb{P}(E, \bar{F} \mid T) \\ &+\mathbb{P}(R \mid T, E, F)\mathbb{P}(E, F \mid T) \\ &+\mathbb{P}(R \mid T, \bar{E}, F)\mathbb{P}(\bar{E}, F \mid T)\end{aligned} \tag{2.5}$$

As we have mentioned above, F corresponds to the case where the mutation lies on the unobserved branch of the tree u with probability $\mathbb{P}(F) = q$. The probability of a mutation lying on a branch is equal to the scaled branch length so the sum of the scaled lengths is equal to 1. From the standard coalescent, the total branch length L_n is given in Eq.1.8, and its expectation is equal to the right hand side of Eq.1.9. Using this result, the expected length of the unobserved tree branch *u* is $E(L_{n+1}) - E(L_n)$. Hence, this yields

$$\mathbb{P}(F) = q = 1 - \frac{2\sum_{j=1}^{n-1}\frac{1}{j}}{2\sum_{j=1}^{n}\frac{1}{j}} \tag{2.6}$$

Assuming that $\mathbb{P}(E)$ and $\mathbb{P}(F)$ are independent, we can calculate straightfor-

wardly the following:

$$\mathbb{P}(\bar{E}, \bar{F} \mid T) = (1-p)(1-q)$$

$$\mathbb{P}(E, \bar{F} \mid T) = p(1-q)$$

$$\mathbb{P}(E, F \mid T) = pq$$

$$\mathbb{P}(\bar{E}, F \mid T) = (1-p)q$$

Subsequently, we expand each of the likelihoods. The first term in Eq. 2.5 corresponds to the case where the reference allele is read with no error and the mutation lies on a tree branch. Hence, either the mutation base or the root base will be equal to the reference allele A_0. Giving an equal probability to the two cases we have:

$$\begin{aligned}
\mathbb{P}(R \mid T, \bar{E}, \bar{F}) = \frac{1}{2}\Big(& \sum_{i=1}^{3}\sum_{b} \mathbb{P}(R \mid r = A_0, m = A_i, b, \bar{E}, \bar{F}, T)\mathbb{P}(r = A_0, m = A_i, b \mid \bar{E}, \bar{F}, T) \\
& + \sum_{i=1}^{3}\sum_{b} \mathbb{P}(R \mid r = A_i, m = A_0, b, \bar{E}, \bar{F}, T)\mathbb{P}(r = A_i, m = A_0, b \mid \bar{E}, \bar{F}, T)\Big) \\
= \frac{1}{6}\Big(& \sum_{i=1}^{3}\sum_{b} \mathbb{P}(R \mid r = A_0, m = A_i, b, \bar{E}, \bar{F}, T) l_b \\
& + \sum_{i=1}^{3}\sum_{b} \mathbb{P}(R \mid r = A_i, m = A_0, b, \bar{E}, \bar{F}, T) l_b\Big)
\end{aligned}$$

The second term in Eq. 2.5 refers to the case where the reference allele is read with an error and the mutation lies on a tree branch. If the reference allele is read

with an error, the reference allele could be any of A_1, A_2 or A_3 with probability $p/3$ for each of the three alleles. Thus, in order to have only bi-allelic sites, the root base or the mutation base have to be equal to the "true" reference allele. Therefore,

$$\begin{aligned}
\mathbb{P}(R \mid T, E, \bar{F}) = \frac{1}{3}\Big(& \sum_{\substack{j \wedge k = 1 \\ j \neq k}} \sum_{b} \mathbb{P}(R \mid r = A_j, m = A_k, b, E, \bar{F}, T)\mathbb{P}(r = A_j, m = A_k, b \mid E, \bar{F}, T) \\
& + \sum_{\substack{j \wedge k = 2 \\ j \neq k}} \sum_{b} \mathbb{P}(R \mid r = A_j, m = A_k, b, E, \bar{F}, T)\mathbb{P}(r = A_j, m = A_k, b \mid E, \bar{F}, T) \\
& + \sum_{\substack{j \wedge k = 3 \\ j \neq k}} \sum_{b} \mathbb{P}(R \mid r = A_j, m = A_k, b, E, \bar{F}, T)\mathbb{P}(r = A_j, m = A_k, b \mid E, \bar{F}, T)\Big) \\
= \frac{1}{18}\Big(& \sum_{\substack{j \wedge k = 1 \\ j \neq k}} \sum_{b} \mathbb{P}(R \mid r = A_j, m = A_k, b, E, \bar{F}, T) l_b \\
& + \sum_{\substack{j \wedge k = 2 \\ j \neq k}} \sum_{b} \mathbb{P}(R \mid r = A_j, m = A_k, b, E, \bar{F}, T) l_b \\
& + \sum_{\substack{j \wedge k = 3 \\ j \neq k}} \sum_{b} \mathbb{P}(R \mid r = A_j, m = A_k, b, E, \bar{F}, T) l_b\Big)
\end{aligned}$$

where we assume an equal probability for each of the alternative alleles. Also, the third term in Eq. 2.5 is expanded to

$$\mathbb{P}(R \mid T, E, F) = \frac{1}{3}\Big(\sum_{i \in \{0,2,3\}} \mathbb{P}(R \mid r = A_i, m = A_1, u, E, F, T)\mathbb{P}(r = A_i, m = A_1, u \mid E, F, T)$$

$$
\begin{aligned}
&+ \sum_{i\in\{0,1,3\}} \mathbb{P}(R \mid r = A_i, m = A_2, u, E, F, T)\mathbb{P}(r = A_i, m = A_2, u \mid E, F, T) \\
&+ \sum_{i\in\{0,1,2\}} \mathbb{P}(R \mid r = A_i, m = A_3, u, E, F, T)\mathbb{P}(r = A_i, m = A_3, u \mid E, F, T)\Big) \\
&= \frac{1}{9}\Big(\sum_{i\in\{0,2,3\}} \mathbb{P}(R \mid r = A_i, m = A_1, u, E, F, T) \\
&+ \sum_{i\in\{0,1,3\}} \mathbb{P}(R \mid r = A_i, m = A_2, u, E, F, T) \\
&+ \sum_{i\in\{0,1,2\}} \mathbb{P}(R \mid r = A_i, m = A_3, u, E, F, T)\Big)
\end{aligned}
$$

Finally, the last term in Eq. 2.5, corresponds to the case where the reference allele is read with no error and the mutation is on the unobserved branch of the tree, therefore

$$
\begin{aligned}
\mathbb{P}(R \mid T, \bar{E}, F) &= \sum_{i=1}^{3} \mathbb{P}(R \mid r = A_i, m = A_0, E, F, T)\mathbb{P}(r = A_i, m = A_0 \mid E, F, T) \\
&= \frac{1}{3}\sum_{i=1}^{3} \mathbb{P}(R \mid r = A_i, m = A_0, E, F, T)
\end{aligned}
$$

Putting all four scenarios together we have

$$
\begin{aligned}
\mathbb{P}(R \mid T, M_1) = &\frac{1}{6}(1-p)(1-q)\sum_r\sum_m\sum_b \mathbb{P}(R \mid T, r, m, b, \bar{E}, \bar{F})l_b \\
&+ \frac{1}{18}p(1-q)\sum_r\sum_m\sum_b \mathbb{P}(R \mid T, r, m, b, E, \bar{F})l_b \\
&+ \frac{1}{9}pq\sum_r\sum_m \mathbb{P}(R \mid T, r, m, E, F)
\end{aligned}
$$

$$
\begin{aligned}
&+ \frac{1}{3}(1-p)q \sum_r \sum_m \mathbb{P}(R \mid T, r, m, \bar{E}, F) \\
&= \frac{1}{6}(1-p)(1-q) \sum_r \sum_m \sum_b 10^{\alpha_{T,r,m,b,\bar{E},\bar{F}}} \\
&+ \frac{1}{18}p(1-q) \sum_r \sum_m \sum_b 10^{\alpha_{T,r,m,b,E,\bar{F}}} \\
&+ \frac{1}{9}pq \sum_r \sum_m 10^{\alpha_{T,r,m,E,F}} \\
&+ \frac{1}{3}(1-p)q \sum_r \sum_m 10^{\alpha_{T,r,m,\bar{E},F}}
\end{aligned}
$$

where

$$\alpha_{T,r,m,b,\bar{E},\bar{F}} = log_{10}(\mathbb{P}(R \mid T, r, m, b, \bar{E}, \bar{F})l_b) \tag{2.7}$$

$$\alpha_{T,r,m,b,E,\bar{F}} = log_{10}(\mathbb{P}(R \mid T, r, m, b, E, \bar{F})l_b) \tag{2.8}$$

$$\alpha_{T,r,m,E,F} = log_{10}(\mathbb{P}(R \mid T, r, m, E, F)) \tag{2.9}$$

$$\alpha_{T,r,m,\bar{E},F} = log_{10}(\mathbb{P}(R \mid T, r, m, \bar{E}, F)) \tag{2.10}$$

Following the same reasoning as above, for each combination of root base (r) and mutation base (m) on tree (T), there is a different set of inferred alleles. Let $Z(T, r, b, m, i, \bar{E}, \bar{F})$ denote the inferred genotype for individual i considering a possible mutation from r to m when the reference allele is read with no error and the mutation lies on the tree. Then,

$$\mathbb{P}(R \mid T, r, b, m, \bar{E}, \bar{F}) = \prod_{i=1}^{N} \prod_{j=1}^{V_i} \mathbb{P}(R_{ij} \mid Z(T, r, b, m, i, \bar{E}, \bar{F})) \tag{2.11}$$

$$= \prod_{i=1}^{N} \prod_{j=1}^{V_i} 10^{\frac{-G_{ij}(Z(T,s,b,m,i,\bar{E},\bar{F}))}{10}} \tag{2.12}$$

which is the product of the phred-scaled genotype likelihoods for all individuals. Hence, Eq. 2.7 can be calculated by

$$\alpha_{T,r,m,b,\bar{E},\bar{F}} = log_{10} l_b + \sum_{i=1}^{N} \sum_{j=1}^{V_i} \frac{-G_{ij}(Z(T, s, b, m, i, \bar{E}, \bar{F}))}{10}$$

and the same reasoning follows the calculations of Eqs. 2.8-2.10. We define Q to be

$$Q = max(\alpha_{T,r,m,b,\bar{E},\bar{F}}, \alpha_{T,r,m,b,E,\bar{F}}, \alpha_{T,r,m,E,F}, \alpha_{T,r,m,\bar{E},F})$$

and we then define

$$\beta_{T,r,m,b,\bar{E},\bar{F}} = \alpha_{T,r,m,b,\bar{E},\bar{F}} - Q$$

$$\beta_{T,r,m,b,E,\bar{F}} = \alpha_{T,r,m,b,E,\bar{F}} - Q$$

$$\beta_{T,r,m,E,F} = \alpha_{T,r,m,E,F} - Q$$

$$\beta_{T,r,m,\bar{E},F} = \alpha_{T,r,m,\bar{E},F} - Q$$

Substituting α with β, the probability of the reads given there is a mutation

on tree T is

$$\begin{aligned}\mathbb{P}(R \mid T, M_1) = &\frac{10^Q}{6}(1-p)(1-q)\sum_r\sum_m\sum_b 10^{\beta_{T,r,m,b,\bar{E},\bar{F}}} \\ &+ \frac{10^Q}{18}p(1-q)\sum_r\sum_m\sum_b 10^{\beta_{T,r,m,b,E,\bar{F}}} \\ &+ \frac{10^Q}{9}pq\sum_r\sum_m 10^{\beta_{T,r,m,E,F}} \\ &+ \frac{10^Q}{3}(1-p)q\sum_r\sum_m 10^{\beta_{T,r,m,\bar{E},F}}\end{aligned}$$

So, the log-scaled likelihood of a mutation event is

$$\begin{aligned}log_{10}\mathbb{P}(R \mid T, M_1) = Q + log_{10}\Big(&\frac{1}{6}(1-p)(1-q)\sum_r\sum_m\sum_b 10^{\beta_{T,r,m,b,\bar{E},\bar{F}}} \\ &+ \frac{1}{18}p(1-q)\sum_r\sum_m\sum_b 10^{\beta_{T,r,m,b,E,\bar{F}}} \\ &+ \frac{1}{9}pq\sum_r\sum_m 10^{\beta_{T,r,m,E,F}} \\ &+ \frac{1}{3}(1-p)q\sum_r\sum_m 10^{\beta_{T,r,m,\bar{E},F}}\Big)\end{aligned}$$

All of the above calculations have been performed for a fixed tree T. The marginal genealogical trees are sampled from the distribution of genealogies (see Section **??**). We thus decide to construct K trees for each position of interest, giving an equal probability to each marginal tree, and calculate the mean BF for each position, which is given by

$$
\begin{aligned}
BF_{mean} &= \frac{\mathbb{P}(R \mid T, M_1)}{\mathbb{P}(R \mid T, M_0)} \\
&= \frac{\sum_{k=1}^{K} \mathbb{P}(R \mid T_k, M_1)\mathbb{P}(T_k)}{\sum_{k=1}^{K} \mathbb{P}(R \mid T_k, M_0)\mathbb{P}(T_k)} \\
&= \frac{\sum_{k=1}^{K} \mathbb{P}(R \mid T_k, M_1)}{\sum_{k=1}^{K} \mathbb{P}(R \mid T_k, M_0)} \\
&= \frac{\sum_{k=1}^{K} \sum_{c \in s_{T_k}} \mathbb{P}(R \mid T_k, c, M_1)\mathbb{P}(c)}{\sum_{k=1}^{K} \sum_{c \in s_0} \mathbb{P}(R \mid c, M_0)\mathbb{P}(c)} \\
&= \frac{\sum_{k=1}^{K} \sum_{c \in s_{T_k}} \mathbb{P}(R \mid T_k, c, M_1)\mathbb{P}(c)}{K\mathbb{P}(R \mid M_0)} \\
&= \frac{1}{K} \sum_{k=1}^{K} \frac{\mathbb{P}(R \mid T_k, M_1)}{\mathbb{P}(R \mid M_0)} \\
&= \frac{1}{K} \sum_{k=1}^{K} BF_k
\end{aligned}
$$

where *c* is the set of all possible inferred genotype configurations, s_{T_k} is the set of inferred genotype configurations in tree T_k, and s_0 refers to the four configurations that correspond to homozygous genotypes for all individuals.

2.2.1 Genotype calling

Each time a root base and a mutation base are placed on the tree, we obtain the inferred alleles for all individuals. Henceforth, there are $2N(n_b - 1) \times 12 + 4$ possible configurations at each tree, where N is the number of individuals, n_b is the number of branches on the tree, and 12 are the possible combinations of the 4 bases plus the 4 configurations that are homozygous in all individuals. We present two possible procedures to make genotype calls using our model:

- *best* configuration : choose the configuration of genotypes with the maximum likelihood across all trees
- marginal posterior probability of genotype : determine the genotypes of the individuals by calculating the posterior probability of the genotype for each individual and choose marginally the genotype with the maximum posterior probability for each individual.

Further discussion of the second method is given below. We first consider the calculation of the posterior probability of a genotype at one tree. Let c_i , $i = 1, ..., M$ be the set of possible genotype configurations and c_i^j the inferred genotype for individual j at configuration i, $j = 1, ..N$.

We calculate the likelihood y_i for configuration c_i

$$y_i = log(\mathbb{P}(R \mid c_i)\mathbb{P}(c_i))$$

then scale the likelihood $s_i = e^{y_i - max(y)}$, and as a result obtain the weight for the configuration to be $w_i = \frac{s_i}{\sum_i s_i}$.

Hence, if α_j is the genotype for individual j and β is the set of all possible genotypes, then $\beta \in \{AA, AC, AG, AT, CC, CG, CT, GG, GT, TT\}$, the posterior probability of the genotype α_j, is given by:

$$\mathbb{P}(\alpha_j = \beta \mid R) = \sum_{i=1}^{M} I(c_i^j = \beta) w_i \tag{2.13}$$

where I is the indication function. The posterior probability of the genotype is averaged over all trees. We then infer the genotypes of the individuals

by choosing marginally the genotype with the maximum posterior probability. Both genotype calling procedures are applied in real data and their performance is compared in Chapter 4.

2.3 Summary

In this chapter, we presented a novel SNP and genotype calling model, TreeCall, developed for calling polymorphic sites using low coverage sequencing data from NGS technologies. The model incorporates the flanking haplotypes for the individuals under study to construct marginal genealogical trees and takes into account the reference allele at the position of interest to give an initial indication of the possible alleles at that position. Assuming the infinite sites model, we calculate the likelihood of all possible mutations on the tree using the NGS genotype likelihoods. We presented in detail the formulas carried out for the calculation of the BF considering different scenarios as well as two methods for determining the genotypes in the samples studied. Lastly, we briefly described the algorithm used for the construction of the marginal genealogical trees. In Chapter 4, we will apply the method in real data and compare the performance of the method with other competing methods.

Chapter 3

A genotype calling algorithm using a Multivariate-Normal approximation

The increasing number of projects with a large sample of individuals, pose a new challenge for the analysis of sequencing data where the sample size will not be in the order of few hundreds of individuals, but of few thousands in the near future. This restricts the use of highly computational models for SNP and genotype calling using NGS data. For example, MARGARITA (Minichiello and Durbin, 2006) uses a heuristic approach to construct ARGs for a sample of sequences and it has been reported that it gets locked in incorrect structures as the sample size increases (Le and Durbin, 2011), which makes QCALL unable to be used in large projects. In addition, the method presented in Chapter 2 for constructing trees, requires 2,830 sec CPU time to construct one tree (with 20 flanking SNPs either site) for 500 individuals, which makes TreeCall restricted in large sample sizes as well. Therefore, current methods that use genealogical trees for SNP and genotype calling are unable to handle more than a few hun-

dreds of individuals. On the other hand, methods that use HMMs to detect and genotype SNPs, e.g. MACH and IMPUTE, can only use a subset of haplotypes from the reference panel in the HMM procedure, in order to be computationally feasible to carry out the calculations (default value 120 in IMPUTE) which restricts the amount of information used.

Our proposed method, MVNcall, assumes that a scaffold of haplotypes at known SNPs is available for the individuals in the study, therefore, apart from the sequencing data available for the sample considered we require that a subset of SNPs is genotyped and those SNPs are phased using a phasing algorithm. This assumption is valid for the individuals sequenced in 1KGP. At each potential variant site, we use a Gibbs sampling phasing scheme to update each individual's haplotypes at that site, given the haplotypes from the rest of the individuals. Rather than using a Hidden Markov Model (HMM) in the phasing updates, we use a simple, fast, and novel method based on approximating a new haplotype as a draw from an appropriately chosen multivariate normal distribution. Since the haplotypes at the known sites are static, many calculations involved can be efficiently implemented so that the method is fast and scales well. The evidence for variation at the site and estimates of individual genotypes can be obtained by summarizing the MCMC output. However, we still assume that a non-LD method has been applied beforehand and a set of candidate sites is chosen for genotyping. Although this method outputs a probability of the site being polymorphic, the main focus is still on SNP genotyping.

3.1 Description of proposed SNP and genotype caller

Let the sample size of the individuals be n with corresponding haplotypes $h_{1,1}, h_{2,1}, \ldots h_{1,n}, h_{2,n}$. The output from the non-LD methods is a set of candidate sites with the corresponding genotype likelihoods for each individual (dotted lines in Fig.3.1). Also, we assume that the individuals are genotyped in a subset of sites and are phased, i.e. the haplotype scaffold (sites in yellow in Fig.3.1). Our method analyses one position at at a time, taking into account the genotype likelihoods at the position of interest and a pre-specified set of flanking SNPs from the haplotype scaffold.

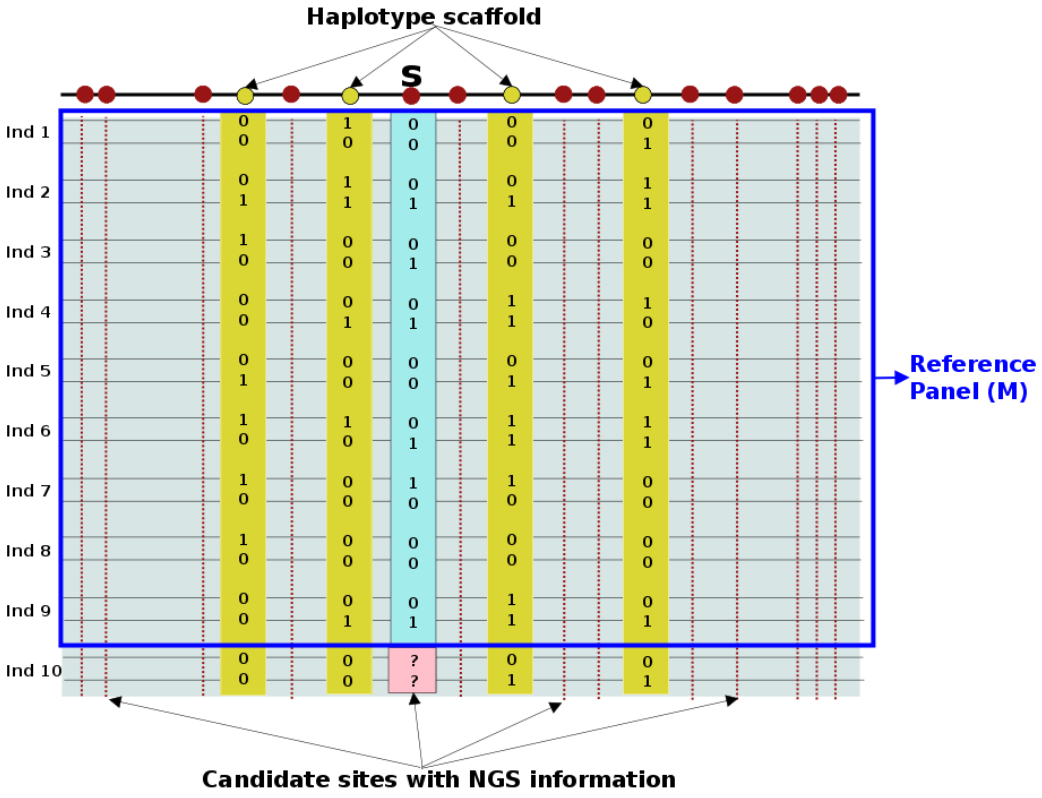

Figure 3.1: Pictorial example of an iteration step of MVNcall, number of individuals in the sample is 10. At the example above, we want to calculate the posterior probability of the genotype for individual 10 given the flanking sites (in yellow) and the fixed phased genotypes at the site of interest for the rest of the individuals (in blue).

We begin by initializing the haplotypes at the candidate site (site S in Fig.3.1), using solely the genotype likelihood information jointly for all individuals. After the initialization step, we employ an MCMC procedure where the phased genotype of an individual i at site S, (in schematic Fig. 3.1 i =10) at each iteration, is updated given the reference panel (M) which consists of the haplotypes of $n-1$ individuals, $H_{-i} = \{h_{1,1}, h_{2,1}, \ldots h_{1,i-1}, h_{2,i-1}, \ldots h_{1,i+1}, h_{2,i+1}, \ldots h_{1,n}, h_{2,n}\}$ and the genotype likelhoods of the individual at the site of interest. The haplotypes at site S in the reference panel are assumed fixed with the current haplotype guesses.

In particular, we treat the haplotypes $h_{1,i}$ and $h_{2,i}$ as independent and identically distributed (iid) draws from the reference panel, $\mathbb{P}(h_{1,i}, h_{2,i} \mid M)$, which is approximated by a multivariate normal distribution. The posterior probabilities of each phased genotypes are calculated by considering both the LD-information and the genotype likelihoods. The posterior probabilities of the four possible phased genotypes are normalized and fitted to a multinomial distribution, we sample a genotype, and we replace the current individual's genotype with the sampled genotype. This MCMC procedure is repeated for a number of iterations, where a small number of iterations at the beginning are used as burn in, followed by a larger number of iterations for inference. Convergence of the MCMC can be tested by running the method multiple times and comparing the results from the different runs or run the method for a large number of iterations and compare intermediate results. The output of the MCMC algorithm consists of the average marginal posterior probabilities for all individuals that are used

for inference.

By representing the LD structure by a multivariate normal distribution, we are losing some information compared to other methods discussed in Section 1.6. To begin with, we account only pairwise correlation between the sites in the window we study instead of the correlation of all SNPs simultaneously as Thunder, SNPTools and BEAGLE. In addition, this approximation has not directionality, since by permuting the SNPs in the window, the multivariate normal distribution will remain unchanged, which is not a desirable feature as the order does matter. On the other hand, in the genotype calling problem from sequencing data, by analyzing one site a time, and keeping a haplotype scaffold fixed, we avoid convergence issues that are met in the other methods that analyze a window of sites simultaneously. This issue can be realized in the case of Thunder, where it requires a good starting point in order to converge. In addition, in cases where there are false positives, due to misalignments or sequencing steps, the methods that analyze a set of windows of sites will get confused and propagate the uncertainty to the nearby sites. In contrast, the approximation where one site a time is considered and the errors on the scaffold are independent of the sequencing errors, the suggested method will not suffer from these situations.

In the following section, we present in detail the model and how we can use it to make inference.

3.2 Statistical framework

In this section, we present the statistical results governing the proposed method. We begin by introducing some notation that will be used for the description of the method. Some notation is repeated from Chapter 2, but we re-define them here for a complete and independent description of this method. Therefore, let

- N : number of individuals under study
- p : candidate site
- V_i : number of technologies that have sequenced individual i, i.e. number of GLF files for individual i
- R_i: sequencing reads at position p at individual i
- $R = \{R_i, ..., R_n\}$: sequencing reads at the position of interest for all individuals
- $G = \{G_1, ..., G_N\}$: vector of phased genotypes at position of interest
- μ : mean vector of allele frequencies of SNPs in the haplotypes studied
- Σ : covariance matrix of allele frequencies of SNPs in the haplotypes studied
- $h_i = \{h_{1i}, h_{2i}\}$: pair of haplotypes of individual i
- $H(-h_{ki})$: all the haplotypes in the study except haplotype h_{ki}
- $\tilde{g}$, phased genotype in binary format, $\tilde{g} = \{00, 01, 10, 11\}$

Let us assume we have $2N$ haplotypes $H = \{h_{1i}, h_{2i}, ...h_{1N}, h_{2N}\}$ from N individuals in the study sample, and the reference panel M consists of $m = N - 1$ of those individuals, with the corresponding haplotypes $H(-h_{1i}, -h_{2i})$ for a flanking window of d SNPs. We assume that the haplotype $h_{1i} = h^*$ (as well as h_{2i}) are iid draws from a conditional distribution $\mathbb{P}(h^* \mid M)$. We approximate the distribution $\mathbb{P}(h^* \mid M)$ with a multivariate normal distribution

$$h^* \mid M \sim MVN(\mu, \Sigma) \tag{3.1}$$

where the mean and the variance-covariance matrix can be estimated from the reference panel. The mean is essentially a vector of the allele frequencies of the flanking sites and the current estimate of the allele frequency at the candidate site, and the variance-covariance matrix gives information about the local pairwise LD structure in the window of flanking SNPs. The mean and variance-covariance matrix are then given by

$$\mu = \mathbb{E}(h^* \mid M) \tag{3.2}$$

and

$$\Sigma = \frac{1}{2m} Var(h^* \mid M) \tag{3.3}$$

where μ is of dimension $d \times 1$ and Σ is of dimension $d \times d$. The empirical mean vector and variance covariance matrix can be estimated by

$$\hat{\mu}_e = \frac{1}{2m} \sum_{i=1}^{m} \sum_{k=\{1,2\}} h_{ki} \tag{3.4}$$

where $\hat{\mu_e}^i$ turns out to be the estimated allele frequency of SNP i. Also, the variance-covariance matrix can be estimated by

$$\hat{\Sigma}_e = \frac{1}{2m-1} \sum_{i=1}^{m} \sum_{k=\{1,2\}} (h_{ki} - \hat{\mu}_e)(h_{ki} - \hat{\mu}_e)^T \tag{3.5}$$

However, the empirical mean vector $\hat{\mu_e}$ and variance-covariance matrix $\hat{\Sigma_e}$ can not be used straight forwardly in the model. Consider the case where a flanking SNP is monomorphic in the sample studied. The empirical mean will be equal to zero as well as its variance. This poses two problems: the $\hat{\Sigma_e}$ is not invertible and secondly, with zero variance the possibility of the site being polymorphic is restricted. Thus, in order to allow polymorphism, we should truncate the mean vector so none of the means is either 0 or 1, and we should also add a penalty on the variance-covariance matrix to ensure invertibility. Also, from a statistical point of view, the empirical variance-covariance matrix is not considered as a good approximation of the true covariance matrix when $d \gg N$ (or even when $d \approx N$) and shrinkage approaches are suggested (Schäfer and Strimmer, 2005). An alternative of shrinkage, is to assign prior distributions on the parameters of interest. In the following two sections, we consider these two alternatives: i) estimating the empirical means and variance-covariance matrix and apply some shrinkage and ii) assign some prior distributions on μ and Σ, and derive the conditional posterior distribution of the haplotype given the ref-

erence panel.

3.2.1 Point estimates for μ and Σ

The multivariate normal approximation has been recently used in another genetic data framework to impute allele frequencies from summary or pooled genotype data. Wen and Stephens (2010) developed a method to impute allele frequencies given only summary data for the reference panel, and have also extended the model for individual level genotype imputation, given full or summary data on the reference panel. Using the conditional distribution $\mathbb{P}(h \mid M)$ from Li and Stephens (2003), they provide analytical results for the estimates of $\hat{\mu}$ and $\hat{\Sigma}$. Specifically, $\hat{\mu}$ is given by

$$\hat{\mu} = (1-\theta)\hat{\mu}_e + \frac{\theta}{2}1 \tag{3.6}$$

where I ($d \times d$) is the identity matrix. The variance-covariance matrix is

$$\hat{\Sigma} = (1-\theta)^2 S + \frac{\theta}{2}(1-\frac{\theta}{2})I \tag{3.7}$$

In Eq. 3.7, the matrix S applies some shrinkage on the covariances according to the recombination rate between SNP$_i$ and SNP$_j$. So, S is given by

$$S^{ij} = \begin{cases} \Sigma_e^{ij} & i = j \\ e^{-\frac{\rho_{ij}}{2m}}\Sigma_e^{ij} & i \neq j \end{cases} \tag{3.8}$$

where ρ_{ij} is an estimate of the population-scaled recombination rate between

SNPs i and j. When ρ_{ij} is large, the shrinkage parameter $e^{-\frac{\rho_{ij}}{2m}}$ will shrink the covariance towards zero. This results to a sparse matrix, which eliminates the noise, and greatly simplifies the calculation of the inverse.

Also, θ is the parameter that relates to the mutation rate, and they use the estimate suggested by Li and Stephens (2003) which is

$$\theta = \frac{(\sum_{i=1}^{2m-1} \frac{1}{i})^{-1}}{2m + (\sum_{i=1}^{2m-1} \frac{1}{i})^{-1}} \tag{3.9}$$

As a consequence, by adding a shrinkage on the covariance matrices we implicitly add directionality in the model, which is one of the disadvantages of the Multivariate normal approximation, where without applying the shrinkage, the SNPs on the scaffold are exchangeable.

However, in a small flanking window and/or when the SNP density in the haplotype scaffold is dense, we can assume that the probability of recombination between the SNPs is very small, and thus the shrinkage parameter on the covariances will be close to 1. An alternative procedure that still assures that the matrix is invertible, is to add a penalty parameter λ on the diagonal of the variance-covariance matrix. Hence, a simpler alternative estimate is:

$$\hat{\Sigma} = \hat{\Sigma}_e + \lambda I \tag{3.10}$$

3.2.2 Assigning prior distributions on μ and Σ

In the previous section, we calculated point estimates of μ and Σ. By doing so, we do not explore the space of possible values of those parameters. In addition, as we discussed above, assigning prior distributions on the parameters provides implicitly some shrinkage. However, the choice of priors will affect both the performance of the method and the computational time. In order to maintain the great advantage of the proposed method which is computational simplicity, we choose conjugate priors for μ and Σ, and we carry out the calculations to find a closed form for the conditional distribution of $\mathbb{P}(h \mid M)$. We thus define a hierarchical model, where the prior distribution of Σ is an inverse-Wishart distribution

$$\Sigma \sim INW(\nu_0, \Lambda_0) \tag{3.11}$$

where Λ_0 and Σ are $d \times d$, and the conditional distribution of μ given Σ is

$$\mu \mid \Sigma \sim MVN(\mu_0, \Sigma/\kappa_0) \tag{3.12}$$

where μ is $d \times 1$.

Under those prior distributions, the conditional distribution $\mathbb{P}(h \mid M)$ is

$$h \mid H_M \sim t_{\nu'-d+1}\left(\mu', \frac{\kappa'+1}{\kappa'(\nu'-d+1)}\Lambda'\right) \tag{3.13}$$

where the parameters are

$$\begin{aligned}
\mu' &= \frac{\kappa_0}{\kappa_0 + 2m}\mu_0 + \frac{2m}{\kappa_0 + 2m}\hat{\mu}_e \\
\kappa' &= \kappa_0 + 2m \\
\nu' &= \nu_0 + 2m \\
\Lambda' &= \Lambda_0 + (2m-1)\hat{\Sigma}_e + \frac{2\kappa_0 m}{\kappa_0 + 2m}(\hat{\mu}_e - \mu_0)(\hat{\mu}_e - \mu_0)^T
\end{aligned} \tag{3.14}$$

where κ_0, μ_0, Λ_0, and ν_0 are hyperparameters from the prior distributions.

Therefore, instead of substituting the mean and the variance-covariance matrix with point estimates, we assign prior conjugate distributions to μ and Σ, and calculate the probability of the haplotype given the data which now follows a multivariate-t distribution. In this way, we do not restrict the values of the parameters μ and Σ and we explore the whole parameter space. A t-distribution is a symmetric distribution like the normal distribution, but with heavier tails that incorporates the case of more extreme events.

3.3 A new algorithm for SNP and genotype calling

In this section, we display the algorithm we developed for SNP and genotype calling that employs the multivariate-normal (or multivariate-t) approximation presented in the previous section. The algorithm commences by initializing the haplotypes at the candidate site, and then it performs a Gibbs sampling scheme where each individual's genotype is called given (i) flanking sites from the haplotype scaffold and (ii) the fixed haplotypes at the position of interest of the rest of the individuals.

3.3.1 Initialization of haplotypes at candidate site

The first step of the algorithm is to initialize the haplotype vector at the position of interest. A natural choice would be to choose for each individual the genotype with the highest likelihood, but in some cases the individual has no reads covering that position, and thus in these cases the genotype likelihood distribution is flat. However, by combining the information from all individuals, we could pick the genotype that is more likely given the genotype likelihoods of the individual at that position and the population allele frequency. To do so, we developed an EM algorithm that calculates the posterior genotype at the E-step and maximizes the population allele frequencies in the M-step. In particular, we initialize the allele frequencies of the two alleles $\delta = \{\delta_A, \delta_B\} = \{0.5, 0.5\}$. Define $g \in \{AA, AB, BB\}$ to be the set of possible genotypes given the two alleles. Then, at the **E-step** of the EM algorithm we calculate for each individual i

$$\mathbb{P}(G_i = g_k \mid \delta, R_i) \propto \mathbb{P}(R_i \mid G_i = g_k)\mathbb{P}(G_i = g_k \mid \delta) \quad \forall k = 1, 2, 3 \tag{3.15}$$

The first term is the genotype likelihood and the second term can be calculated assuming the Hardy-Weinberg equilibrium, i.e.

$$\mathbb{P}(AA \mid \delta) = \delta_A^2$$
$$\mathbb{P}(AB \mid \delta) = 2\delta_A\delta_B$$
$$\mathbb{P}(BB \mid \delta) = \delta_B^2$$

Then at the **M-step** we estimate the allele frequencies. The allele frequency

at iteration $t+1$ is given by

$$\hat{\delta}_s^{t+1} = \sum_{n=1}^{N} \frac{I(s \in g)\mathbb{P}(G_i = g \mid \delta_s^t, R_i)}{2N}, \quad \forall s \in \{A, B\} \tag{3.16}$$

where I is the indication function. When the EM algorithm converges we use the estimated allele frequencies to calculate the posterior probability of the genotypes of individual i given the estimated allele frequencies and we choose the genotype with the maximum posterior probability, i.e.

$$G_i = \arg\max_g \mathbb{P}(G_i = g \mid R_i, \hat{\delta}) \tag{3.17}$$

The EM algorithm described here is the single population version of the Multi-population algorithm presented in Section A.

After the initialization step, we need to phase the initial genotypes at the candidate site. The phasing of homozygous genotypes is trivial but that does not apply for the case of heterozygous sites. An intuitive solution is to phase randomly the heterozygous genotypes. An alternative procedure in our framework is the following: we set as the reference panel the haplotypes of all individuals with homozygous genotypes at the candidate site. Then, we phase in a random order the heterozygous sites. At each step, when a heterozygous genotype is phased, the corresponding haplotypes are added on the reference panel, and the process is repeated until all heterozygous genotypes are phased. Both procedures are examined in the Chapter 5.

3.3.2 Gibbs sampling iteration scheme

The objective in our setting is to calculate the joint density of the phased genotypes $\mathbb{P}(G_1, G_2, ..., G_N \mid R, H)$. However the calculation of the joint density is cumbersome and computationally intensive since the search space is very large. However, the conditional distribution of an individual's haplotype is of closed form. We are interested in the conditional distribution of the genotype of individual i given the sequencing reads R_i, the haplotype panel $H(-\{h_{1i}, h_{2i}\})$, which also includes the fixed phased genotypes at the candidate site p, and the flanking haplotypes of individual i: h_{1i}^{-p}, h_{2i}^{-p}. Hence the conditional probability of the genotype G_i at position p is given by

$$\begin{aligned} &\mathbb{P}(G_i = \tilde{g} \mid R_i, H(-\{h_{1i}, h_{2i}\}), h_{1i}^{-p}, h_{2i}^{-p}) \\ &\propto \mathbb{P}(R_i \mid G_i = \tilde{g}) \mathbb{P}(G_i = \tilde{g} \mid H(-\{h_{1i}, h_{2i}\}), h_{1i}^{-p}, h_{2i}^{-p}) \end{aligned} \tag{3.18}$$

As a side note, when there is more than one genotype likelihood file[1] for the same individual then

$$\mathbb{P}(R_i \mid G_i = \tilde{g}) = \prod_{k=1}^{V_i} \mathbb{P}(R_{ik} \mid G_i = \tilde{g}) \tag{3.19}$$

Assuming the haplotypes are *iid*, we can rewrite the right hand side of Eq.3.18 as

$$\mathbb{P}(R_i \mid G_i = \tilde{g}) \mathbb{P}(h_{1i}^{p} \mid h_{1i}^{-p}, H(-\{h_{1i}, h_{2i}\})) \mathbb{P}(h_{2i}^{p} \mid h_{2i}^{-p}, H(-\{h_{1i}, h_{2i}\})) \tag{3.20}$$

[1] this corresponds to the case where the individual has been sequenced by more than one method

If we partition μ and Σ from Eq.3.1

$$\mu = \begin{bmatrix} \mu_p \\ \mu_{-p} \end{bmatrix} \text{ with sizes } \begin{bmatrix} 1 \times 1 \\ (d-1) \times 1 \end{bmatrix}$$

and

$$\Sigma = \begin{bmatrix} \Sigma_{p,p} & \Sigma_{p,-p} \\ \Sigma_{-p,p} & \Sigma_{-p,-p} \end{bmatrix} \text{ with sizes } \begin{bmatrix} 1 \times 1 & 1 \times (d-1) \\ (d-1) \times 1 & (d-1) \times (d-1) \end{bmatrix}$$

then the conditional distributions in Eq.3.20 follow a normal distribution

$$\mathbb{P}(h_{1i}^{p} \mid h_{1i}^{-p} = \alpha, H(-\{h_{1i}, h_{2i}\})) \sim N(\mu^c, \Sigma^c) \tag{3.21}$$

where the mean and variance can be calculated using the standard results

$$\begin{aligned} \mu^c &= \mu_p + \Sigma_{p,-p}\Sigma_{-p,-p}^{-1}(\alpha - \mu_{-p}) \\ \Sigma^c &= \Sigma_{p,p} - \Sigma_{p,-p}\Sigma_{-p,-p}^{-1}\Sigma_{-p,p} \end{aligned} \tag{3.22}$$

Hence the posterior probabilities of the four different phased genotypes $\tilde{g} = \{00, 01, 10, 11\}$ are proportional to

$$\begin{aligned}
&\mathbb{P}(R_i \mid G_i = 00)\mathbb{P}(h_{1i} = 0 \mid h_{1i}^{-p}, H(-\{h_{1i}, h_{2i}\}))\mathbb{P}(h_{2i} = 0 \mid h_{2i}^{-p}, H(-\{h_{1i}, h_{2i}\})) \\
&\mathbb{P}(R_i \mid G_i = 01)\mathbb{P}(h_{1i} = 0 \mid h_{1i}^{-p}, H(-\{h_{1i}, h_{2i}\}))\mathbb{P}(h_{2i} = 1 \mid h_{2i}^{-p}, H(-\{h_{1i}, h_{2i}\})) \\
&\mathbb{P}(R_i \mid G_i = 10)\mathbb{P}(h_{1i} = 1 \mid h_{1i}^{-p}, H(-\{h_{1i}, h_{2i}\}))\mathbb{P}(h_{2i} = 0 \mid h_{2i}^{-p}, H(-\{h_{1i}, h_{2i}\})) \\
&\mathbb{P}(R_i \mid G_i = 11)\mathbb{P}(h_{1i} = 1 \mid h_{1i}^{-p}, H(-\{h_{1i}, h_{2i}\}))\mathbb{P}(h_{2i} = 1 \mid h_{2i}^{-p}, H(-\{h_{1i}, h_{2i}\}))
\end{aligned} \tag{3.23}$$

We thus employ a Gibbs sampler, where at each iteration each individual's posterior genotype probabilities are estimated given the haplotypes at the flanking sites and the current haplotype at the candidate site of the individuals in the reference panel. The algorithm commences by assigning initial phased genotypes for each indiviual at the candidate site. Given the initial guesses, we perform a number of MCMC iterations. At each iteration step t we update individual i according to the following procedure

1. The reference panel consists of the haplotypes of all study individuals apart from individual i. We estimate the parameters $\hat{\mu}$ (Eq. 3.4) and $\hat{\Sigma}$ (Eq. 3.5) of the multivariate-normal (or multivariate-t) model using the reference panel and apply the shrinkage method (Eq. 3.6,3.7 or 3.10),

2. We calculate the posterior probabilities for the four phased genotypes as in Eq. 3.23. We scale the four posterior probabilities so they sum up to 1,

3. The scaled posterior probabilities follow a multinomial distribution and we draw a genotype from the multinomial distribution g^* and we denote the genotype $G_i^{(t+1)} = g^*$.

3.4 Extending the model to handle large samples of individuals

The simplicity of the model allows the analysis of a large sample of individuals and in this section we aim to present how the model can be extended to this case. When we update an individual's genotype given the reference panel, we could use all individuals in the reference panel without a large computational cost since the size of the mean vector and the variance-covariance matrix depend solely on the number of flanking sites. However, by pooling information from all individuals, we implicitly assume that we obtain the same information from each haplotype, and that the individuals come from the same population, which might not be the case. A possible solution for that is to choose the closest haplotypes for the haplotype of interest and use only those for inference, since haplotypes that are not close will not provide any further information. Below we discuss different measures to select close haplotypes and how this information can be incorporated in our model.

3.4.1 Distance measures to select the closest haplotypes from a reference panel

If we knew the genealogical tree of a sample of haplotypes, we would like to choose the haplotypes that have coalesced more recently in the past with the haplotypes of the individual of interest. Those haplotypes will share a more common ancestry with the haplotypes of interest since their MRCA is more

recent and thus there has been less time for recombination and/or mutation events.

Distance measures have been previously used to choose from a reference panel the "closest" haplotypes for a haplotype of interest. The requirement of choosing a subset of haplotypes for an inference problem has been also met in imputation and phasing methods. For example, IMPUTE v2 (Howie et al., 2009) chooses a subset of haplotypes by calculating the *hamming distance* between the haplotypes in the reference panel and the haplotypes of the individual. In particular, for each haplotype, the hamming distance is calculated for the two haplotypes and the minimum of the two distances is saved. The k first haplotypes with the smallest Hamming distance are the conditioning states and the other haplotypes are ignored. As it has been noted in (Howie et al., 2009), there might be the situation where the set of haplotypes chosen are closer to one haplotype than the other, i.e. the number of chosen haplotypes is not symmetric. This works well for phasing problems where the genotype is known, therefore if we can phase one haplotype then the phasing of the other haplotype is trivial. Nevertheless, in the genotype calling scenario, the genotype is unknown and thus information for both haplotypes is required, i.e. symmetric information. Other methods have employed weights on the haplotypes on the reference panel through cross-validation (Huang et al., 2009), or ancestry estimation (Pasaniuc et al., 2010). Below we present two simple haplotype distance measures presented in the literature.

Hamming distance is a measure that has been successfully used for choos-

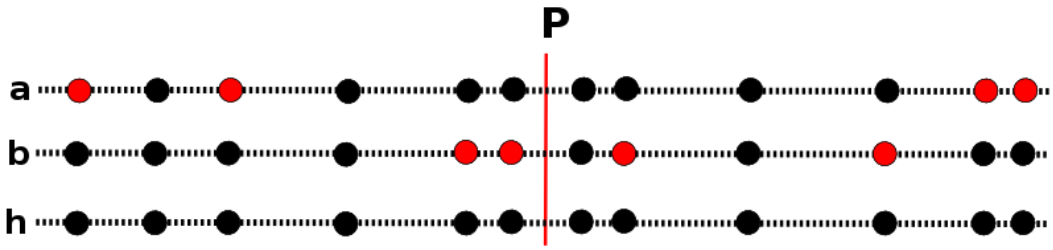

Figure 3.2: In this drawing, P is the site of interest and a and b are two haplotypes that are compared with the haplotype of interest h. The red dots represent the sites the haplotypes a and b are different from haplotype h. For both haplotypes a and b the Hamming distance measure is equal to 4.

ing the closest haplotypes for methods that phase a window of SNPs jointly. At a given window, the greater the Hamming distance, the bigger the number of mutation events between the two haplotypes, and the longer the time until the MRCA. Hamming distance accounts only for the mutation events, but not for any recombination events. Consider the following scenario given in Fig.3.2: haplotypes a and b are two haplotypes that are compared with the haplotype of interest h with the same Hamming distance value (4). The red dots correspond to mismatches. The Hamming distance on its own does not account for differences that might be due to historical recombination events (Li and Jiang, 2005). We thus propose another metric to capture the closest haplotypes that correspond to the length of consecutive sites that are identical to the haplotype of interest, commencing from the site of interest and counting on the right and on the left of site P. This metric is not new in the literature, it has been used for haplotype-sharing methods for association studies, under the assumption that at a given risk locus, cases should have a similar haplotype background surrounding the risk locus (Beckmann, 2010; Tzeng et al., 2003; Li and Jiang, 2005;

Yu et al., 2004). Formally, we call this new measure "*Perfect match*" (PM). Let us assume on a given set of haplotypes, we are interested to call the genotypes for the k^{th} site, then the PM distance between haplotype a and b is given by

$$PM_k(a,b) = \sum_{i \in S_k^R} I(a_i = b_i) + \sum_{i \in S_k^L} I(a_i = b_i) \tag{3.24}$$

where I is the indication function and $S_k^R = \{k+1, k+2, ...\}$ is the largest set of continuous markers that a and b share the same allele on the right of the site of interest and $S_k^L = \{k-1, k-2, ...\}$ is the largest set of continuous markers that a and b share the same allele on the left of the site of interest. For example, in Fig. 3.2, the PM measure between h and a is 7 and between h and b is 1.

Different versions and extensions of this metric have been proposed, including weighting the PM measure according to the physical or the genetic distance between the first and the last position that are identical (Morton et al., 2001). Another extension is the definition of a combined measure that includes both the PM and the Hamming distance measure, by applying appropriate weighting to each measure (Li and Jiang, 2005).

Weighting the conditioning haplotypes according to distance measure

In the case of IMPUTE V2, where the reference panel consists of the k closest haplotypes, the accuracy reaches a plateau after all the "informative" haplotypes have been added to the reference panel (Howie et al., 2009). A possible

explanation is that even though extra haplotypes are added on the reference panel, the haplotypes of interest do not "copy" from them[2]. In a similar notion, we want to ensure that we obtain more information from closest haplotypes. Up to now, our model gives equal weight to all haplotypes. Hence if we order the haplotypes by their haplotype distance from the haplotype of interest, the first and the last haplotype will weight equally in the model. Nevertheless, we could assign different weights to the haplotypes according to their distance to the haplotype of interest, giving higher weight for the closest haplotypes and less weight for the more distant haplotypes. We use the distance measures to calculate the weight for each conditioning haplotypes. In particular, if we want to give different weights to the k conditioning haplotypes $h_i, ..., h_k$ in the reference panel then the estimated mean will be given by

$$\hat{\mu_e^w} = \sum_{i=1}^{k} w_i h_i \tag{3.25}$$

and the estimated variance-covariance matrix is

$$\hat{\Sigma_e^w} = \frac{V_1}{V_1 - V_2} \sum_{i=1}^{k} w_i (h_i - \hat{\mu_e^w})(h_i - \hat{\mu_e^w})^T \tag{3.26}$$

where $V_1 = \sum_{i=1}^{k} w_i$ and $V_2 = \sum_{i=1}^{k} w_i^2$.

[2]IMPUTE uses the Li and Stephens (2003) model, see Section 1.6 for a brief summary of the model

3.4.2 Gibbs sampling iteration scheme for large samples

Let us assume we are at iteration step t, and we would like to update the genotypes of individual i. The reference panel consists of the haplotype set H_{-i}. The two first steps described in the Gibbs sampling algorithm (see Section 3.3.2) above change to

1. Calculate the Hamming distance of the haplotypes in the reference panel and choose the k closest haplotypes for haplotype h_{i1} (H_{k_1}) and k closest haplotypes for haplotype h_{i2} (H_{k_2}). Calculate the estimated μ_{k_1} and Σ_{k_1} using H_{k_1}, and μ_{k_2} and Σ_{k_2} using H_{k_2}

2. Calculate the conditional probability $Pr(G_i = \tilde{g} \mid R_i, H_{k_1}, H_{k_2}, H(\{h_{1i}^{-p}, h_{2i}^{-p}\}))$ where $h_{1i} \mid H_{k_1} \sim MVN(\mu_{k_1}, \Sigma_{k_1})$ and $h_{2i} \mid H_{k_2} \sim MVN(\mu_{k_2}, \Sigma_{k_2})$.

3.4.3 Model averaging

A standard statistical procedure includes the choice of a model from a class of models where inference is based. However, by selecting one model from a class of models, the uncertainty concerning model selection is ignored. A possible procedure to decrease the uncertainty on model selection is to consider a class of different models and then combine the information across them (Claeskens and Hjort, 2008; Hoeting et al., 1999; Draper, 1995). However, there is a trade-off between the number of models to average over and the computational time. In the case of MVNcall, the computational burden lies on the computation of the inverse of the covariance matrices, but it is still feasible to consider more than

one set of parameters.

In our model, we consider instead of only one model with one specific number of conditioning haplotypes, to run the model for different number of conditioning haplotypes and average over the models. The benefit of performing model averaging will become apprent in the Chapter 5.

3.4.4 Timing considerations

A major advantage of MVNcall is the fact that the haplotype scaffold is static and a number of calculations have to be done once. In particular, for the implementation of the method, we have used the following features of the model to considerably decrease the computational time

1. The haplotype distance measure is only calculated once from the haplotype scaffold, and the same set of conditioning sites is used for all iterations for each individual. This is because the haplotype scaffold is static and at each iteration the only site that changes is the site of interest.

2. In Eq.3.22, the estimates of the mean and variance involve the multiplication of three matrices. Firstly, the inverse matrix in that expression involves only sites in the haplotype scaffold, and thus static across the iteration procedure. Thus, we compute the inverse matrices once and we can store them, without needing to recalculate them, which would have been computationaly intensive.

3. The haplotype scaffold is the same for all candidate sites that lie within

two haplotype scaffold sites. This allows us to use the same haplotype distance measures and the inverse matrices from (1) and (2) for all sites in that interval.

4. The model involves the calculation of mean and covariances that involve the multiplication with many zero values. In our code we have represented the data in the form of a Yale Matrix (Golub and Van Loan, 1996), which allows us to loop over only non-zero values. This greatly speeds-up the implementation of the code [3].

5. Since the model considers one position at a time and the result of one position does not influence the result of the other positions inferred, it makes it highly parallelizable. Also, for the same reasons, the model does not have high memory requirements.

3.5 Incorporating trio information for genotype calling

The model presented above assumes that the individuals in the study are unrelated. In the case where the individuals are related, this information can be incorporated in the model to aid genotype calling and phasing. Here, we consider the situation where a trio is sequenced and genotyped on a microarray chip and that the microarray genotypes have been phased using an algorithm that utilises the trio information. At each iteration step we update each member of the trio jointly. To do this we assume that natural assumption that the offspring hap-

[3] For the Phase 1 setting this decreased the computational time by ∼5-fold.

lotypes are conditionally independent of the reference panel given the parental haplotypes. We also assume that the transmitted and untrasmitted haplotypes from the parents to the offspring are known and we assume that in the window of SNPs being analysed there is no recombination between parents and child. Let $g_C = (g_{1C}, g_{2C})$ be the two unobserved alleles of the child, $g_P = (g_{1P}, g_{2P})$ the unobserved parental alleles, and $g_M = (g_{1M}, g_{2M})$ the unobserved maternal alleles. The haplotypes of the child at the scaffold sites are h^s_{1C}, h^s_{2C}, and the parental and maternal haplotypes at the scaffold sites are h^s_{1P}, h^s_{2P} and h^s_{1M}, h^s_{2M} respectively.

Therefore, the posterior probability of the genotype of the trio is

$$P((g_{1C}, g_{2C}), (g_{1P}, g_{2P}), (g_{1M}, g_{2M}) \mid R_C, R_P, R_M, H_{-T}, h^s_{1C}, h^s_{2C}, h^s_{1P}, h^s_{2P}, h^s_{1M}, h^s_{2M}) \tag{3.27}$$

where (g_{1C}, g_{2C}) are the alleles of the child, (g_{1P}, g_{2P}) are the parental alleles, (g_{1M}, g_{2M}) are the maternal alleles, R_C,R_P and R_M the sequencing reads of the trio, and $T = \{h_{1C}, h_{2C}, h_{1P}, h_{2P}, h_{1M}, h_{2M}\}$.

Assuming that (g_{1C}, g_{2C}) is conditionally independent of H_{-T}, given (g_{1P}, g_{2P}) and (g_{1M}, g_{2M}), then

$$\begin{aligned} &P((g_{1C}, g_{2C}), (g_{1P}, g_{2P}), (g_{1M}, g_{2M}) \mid R_C, R_P, R_M, H_{-T}, h^s_{1C}, h^s_{2C}, h^s_{1P}, h^s_{2P}, h^s_{1M}, h^s_{2M}) \\ &\propto \left(\prod_{i \in \{C,P,M\}} P(R_i \mid g_i) \right) P(g_C \mid g_P, g_M) P((g_{1P}, g_{2P}) \mid H_{-T}, h^S_{1P}, h^S_{2P}) P((g_{1M}, g_{2M}) \mid H_{-T}, h^S_{1M}, h^S_{2M} \end{aligned} \tag{3.28}$$

where $P(R_i \mid g_i)$ is the genotype likelihood, $P(g_C \mid g_P, g_M)$ is the probability of the child's genotype given the parental genotypes that are calculated according to Mendel's laws. We make the assumption that we know the transmitted haplotypes from the parents to the offspring. Let the first haplotype from the parents to be the transmitted haplotypes, i.e. h_{1P} is the transmitted haplotype to haplotype h_{1C} and h_{1M} is the maternal transmitted haplotype to the offspring haplotype h_{2C}. Then,

$$P(g_C \mid g_P, g_M) = \begin{cases} (1-10^{-8})^2 & g_{1C} = g_{1P} \wedge g_{2C} = g_{1M} \\ (1-10^{-8})10^{-8} & g_{1C} \neq g_{1P} \oplus g_{2C} \neq g_{1M} \\ 10^{-16} & g_{1C} \neq g_{1P} \wedge g_{2C} \neq g_{1M} \end{cases}$$

where the probability of a mutation in one generation is set to 10^{-8}. The last two terms in Eq.3.28 correspond to the probability of the parental haplotypes given the reference panel, which follows the same notion as the description above, since the parents are assumed to be unrelated.

3.6 Summary

In this chapter we presented a novel method for SNP and genotype calling, MVNcall, which assumes that the conditional distribution of a new haplotype given a reference panel is multivariate normal. The main characteristics of this method are (1) the use of a haplotype scaffold and (2) it analyses one site a time. The method commences by initializing the haplotypes at the site of interest us-

ing only the sequencing information via the genotype likelihoods, and then it uses a Gibbs sampling scheme, where the posterior probability of a genotype given the sequencing information and the scaffold is estimated for each individual. The model has been also extended to large population samples, where we introduced the idea of conditioning haplotypes and weighting. Also, given the computational efficiency of the model, we also discussed the idea of model averaging. In the following two chapters, we investigate the performance of this method on real data.

Chapter 4

SNP and genotype calling results for low-coverage pilot data

Over the course of Chapters 2 and 3 we have developed two distinct methods for SNP and genotype calling from low-coverage next-generation sequencing reads, assuming only bi-allelic SNPs. In this chapter we are going to apply the two methods, TreeCall and MVNcall, on the CEU individuals sequenced in the low-coverage pilot data of the 1KGP. In particular, using this dataset we are going to

- Investigate different settings under the TreeCall model,
- Investigate different tuning parameters for the MVNcall,
- After learning about the parameters from our two models, we are going to compare the TreeCall and MVNcall with other competing LD-methods applied to this dataset, in terms of both SNP (only TreeCall) and genotype calling (TreeCall and MVNcall).

4.1 Data processing

We use the CEU population sample that consists of 60 unrelated individuals and we analyse a 10Mbp region on chromosome 20 (chr20:30,000,000-40,000,000). A subset of 55 individuals are unrelated and present in HapMap3. Since our methods require a haplotype scaffold, which in this dataset is the set of haplotypes from HapMap3, we only consider those 55 individuals in the subsequent analysis.

The BAM files used are from the August 2009 release[1], where 35 out of the 55 individuals were sequenced by two sequencing technologies, and the rest by one technology. Illumina Solexa sequencing platform was mostly used (45 genomes), followed by SoLid (27 genomes) and then by 454 (18 genomes). The BAM files were mapped using the corresponding three alignment methods: MAQ, Corona Lite and SSAHA to the reference sequence NCBI36. Afterwards, base qualities were recalibrated using GATK and remapped using the recalibrated qualities. Following mapping, Samtools was used to calculate the genotype likelihoods for each individual at each position. Using the genotype likelihoods, we run the vanilla version of the Multi-population EM algorithm (Liu and Marchini, personal communication) to exclude clearly monomorphic sites. The output consisted of a set of candidate sites with the two most probable alleles, the ones with the highest allele frequencies, and the corresponding three genotype likelihoods for each individual. The candidate set in the region analysed, included 64,433 candidate sites. Comparing the candidate set with the

[1] ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/pilot_data/data/

calls made by any of the three methods presented above, few positions were not present in our candidate set. In order to have a head-to-head comparison between the SNP calls, we incorporated those sites to our candidate set.

4.2 Assessing TreeCall in low-coverage pilot data

We use the HapMap3 haplotypes as the haplotype scaffold. To build the marginal genealogical trees, we use 20 flanking sites (from the scaffold) at either side from the site of interest. We then build 50 marginal genealogical trees at the midpoint position between two scaffold SNPs, and use these trees for all sites that lie within that interval. After running TreeCall, we calculate the $log_{10}BF_{mean}$[2]. The prior probability of the site being polymorphic under neutral population genetics theory is given in Eq.A.9, where $\theta = 0.001$ for humans. Thus, we set the prior odds $\frac{\mathbb{P}(M_0)}{\mathbb{P}(M_1)} = 0.001$[3], and the threshold applied for SNP calling is $log_{10}BF$ >3. Further filters are applied for reducing the false positive calls due to alignment and coverage artifacts, discussed below.

In this section we investigate the following:

- **Genotype likelihood information**: how the method performs when we input all 10 genotype likelihoods versus 3 genotype likelihoods, that corresponds to the situation where we apply a non-LD method that provides us with the two most probable alleles at each candidate site,

[2]For simplicity we refer to this measure as $log_{10}BF$.
[3]See Eq.2.1

- **Genotype calling step**: we compare the genotype accuracy performance of the two methods proposed in Section 2.2.1,
- **Filtering**: we apply different filters for creating a refined candidate set.

4.2.1 Comparison of the method using all likelihoods versus three likelihoods

In the initial description of our TreeCall method (refer to it as setting A), we consider all 12 possible mutations for the four alleles [4] on the genealogical tree. If, however, we know *a priori* the two alleles, i.e. by running a non-LD method that produces a list of candidate sites with the associated alleles at each site, we can reduce the calculations, by only considering two possible mutations (setting B). In this section, we compare the results of our method when all 10 genotype likelihoods are employed versus if we only employ 3 genotype likelihoods that correspond to the two most probable alleles at each position.

In Fig.4.1, we compare the log_{10}BF between the two different settings. The log_{10}BFs in the majority of the positions are equal, with few positions deviating from the straight line for small values of log_{10}BF. In all cases the log_{10}BF from the run using the 10 genotype likelhoods is greater than the log_{10}BF from the run using the 3 genotype likelhoods, which corresponds to the loss of some information when we use 3 instead of 10 genotype likelihoods. We examined the positions where the discrepancy between the log_{10}BF from the two runs was

[4] This can be the situation where a non-LD method has not been ran first, or that we only have a set of candidate sites from a non-LD method but we have no information of the most probable alleles on the site.

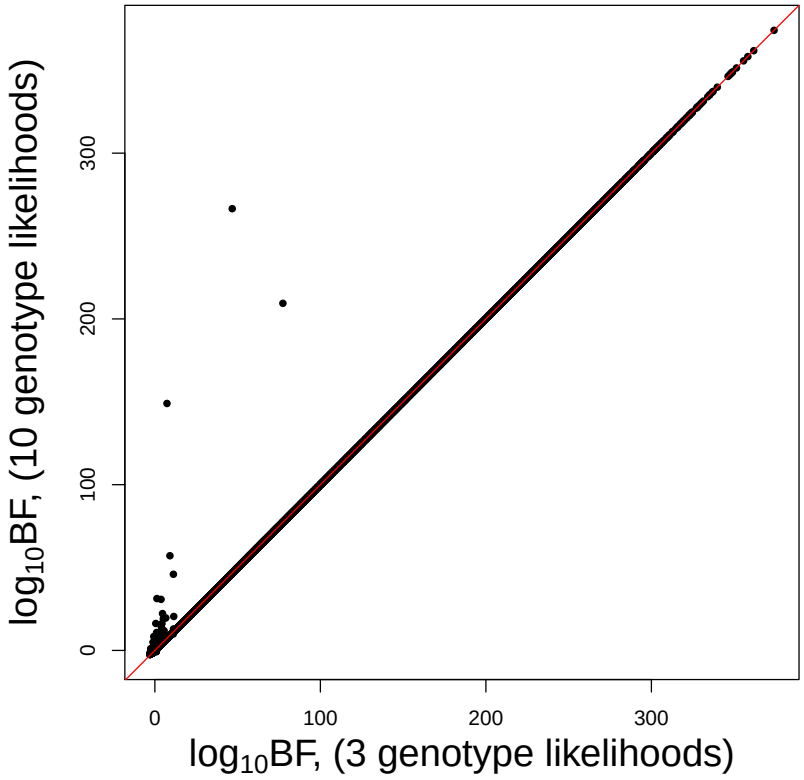

Figure 4.1: Comparison of log_{10}BF between two settings: Setting A considers all 4 nucleotides and thus 12 possible mutations (use of all 10 genotype likelihoods) in the y-axis versus Setting B which considers only two alleles selected by a non-LD method ($R \rightarrow A$ and $A \rightarrow R$) where R is the reference allele and A the alternative allele in the x-axis.

large. For those positions, the two different settings report a different alternative allele. When we checked the genotype likehoods at those sites, we noticed that there is some evidence that the SNP might actually be tri-allelic, which explains the different choice of the alternative allele. Also, for small values of log_{10}BF we see some sites that have slightly higher log_{10}BF in setting A than in setting B. We notice that in these positions, the heterozygous genotype likelihoods that involve the reference allele, are similar with no clear distinction towards a certain alternative allele. In setting A, we take into account all possible genotypes, and

thus the method captures that there is a possible polymorphism, by summing up the likelihoods involving different alternative alleles. However, in setting B, this information is lost, since we only consider three likelihoods (and only one alternative allele), and thus we do not have enough power to detect the polymorphism. Given that the method provides consistent results for the two different settings investigated above, we can infer that the method is able to correctly pick the same alleles as the non-LD methods in the majority of the cases. We also compared the genotype discordance between the two settings, in HapMap2 sites not in HapMap3, and we get the same accuracies. For the rest of this chapter we present results from setting B.

4.2.2 Comparison of genotype calling methods with TreeCall

In Section 2.2.1, we presented two procedures that were adapted to the TreeCall method for genotype calling. The first method (A) proposed, chooses among all trees the mutation with the highest likelihood, and uses the inferred alleles to call the genotypes of each individual. The second method (B), averages over all trees to calculate the posterior probability of the genotype, and for each individual chooses marginally the genotype with the highest posterior probability. In Table 4.1, we present percent discordance rate comparisons between the two methods, for HapMap2 sites not present in HapMap3, assuming HapMap genotypes as the *truth*, of 39 overlapped samples between the 1KGP low-coverage pilot CEUs and HapMap3 CEUs.

To view the improvement of the genotype calls as we include more informa-

Method	HomR %	Het %	HomA %	R^2
GLF	3.88	26.73	8.77	0.6181
GLF & AF	2.77	20.26	13.10	0.9073
A (TreeCall)	2.70	4.85	5.31	0.9620
B (TreeCall)	3.02	7.95	5.71	0.9536

Table 4.1: Percent genotype discordance comparison between (i) using only genotype likelihood information, (ii) using genotype likelihood and allele frequency information, (iii) method A from TreeCall and (iv) method B from Treecall for genotype calling. *HomR %* corresponds to the percent discordance rate for homozygous reference genotypes, *Het %* for heterozygous genotypes, *HomA %* for homozygous alternative genotypes and R^2 is the Pearson correlation coefficient between the called and the true genotypes. The comparison is based on 3,531 SNPs present in HapMap3 but not in HapMap2.

tion, we begin by taking into account the genotype likelihoods at each individual independently and call the genotype with the highest genotype likelihood. When the genotype likelihoods are flat, we randomly choose one genotype. The genotype calling discordance rates are shown on Table 4.1 as "GLF". We notice that the genotype discordance for the heterozygous genotypes is very high, with 26.73% discordance rate. Following, we add the allele frequency information for calculating the posterior probability of genotype for each individual at each site. To do so, we use the EM algorithm presented in Chapter 3 and it is denoted on Table 4.1 as "GLF & AF". By including the allele frequency information we see a great improvement in the R^2, increasing from 0.6181 to 0.9073. The most apparent decrease is in terms of the heterozygous genotypes, which seem to have the highest discordance rates. Then, we are adding the LD information using our TreeCall model. We notice that the heterozygous genotype discordance benefits the most, with a discordance rate decreasing from 20.26% to 4.85%. The homozygous genotype discordance does not vary greatly between the different

methods whereas the homozygous alternative seem to also benefit from the LD information.

Now, in terms of the two genotype calling methods proposed using the TreeCall model, Method A has a lower genotype discordance rate than Method B. Method B seems to miss mostly the heterozygous genotypes compared to method A ($\sim$3.1% difference). The homozygous reference genotype accuracies differ by $\sim$0.32% and for homozygous alternative genotype by $\sim$0.4%. A possible explanation of the performance of method B is the following: when we average over all trees, bearing in mind that the genealogical inference will always have some errors, the errors will accumulate and result in lower accuracy than using a single "best" tree. Also, in method B, we choose *marginally* the genotype for each individual, without taking into account the genotypes of the other individuals, which may also count towards the lower accuracy rates. For all subsequent comparisons we choose method A to call genotypes from TreeCall.

4.2.3 Filtering

For the low-coverage pilot, different LD-based methods have used different non-LD based methods to create a candidate set. Further filters were applied to reduce the FDR from the candidate set. We follow the same strategy and we apply (i) a method-specific filter, by calling sites polymorphic when the calculated log_{10}BF>3[5] and (ii) some filters regarding coverage and misalignment artifacts.

[5]With the exception of sites that are monomorphic with the alternative allele.

Some of the filters suggested by the other methods and considered here are

- **s50**: More than 50% of the individuals have no read information at the candidate site,
- **high/low coverage**: exclude sites with average total depth per individual less than 0.5x or greater than 20x.
- **fw10**: There are three or more candidate SNPs in 10bp,
- **distance from known indels**: exclude candidate SNPs that are within 5bp from known indels.

25,884 out of the 64,433 candidate sites have log_{10}BF>3. The Ts/Tv ratio of this set is 1.97 lower than the expected genome-wide Ts/Tv of ∼2.1, which indicates some false positive sites. Out of the 25,884 sites in the refined call set, 169 fail the s50 filter, 2,208 fail the fw10 filter, 114 sites have extremely low/high average coverage and 2,101 are within 5bp of a known indel (w5bp)[6].

The two filters regarding coverage lead to a Ts/Tv of 1.98 whereas the fw10 filter leads to a Ts/Tv of 2.06 compared to 2.03 of the w5bp filter. We apply both coverage filters and we choose the fw10 filter for indels since it retains more of the known SNPs than the w5bp filter.

[6]Indel set used : ftp://ftptrace.ncbi.nih.gov/1000genomes/ftp/pilot_data/release/2010_07/low_coverage/indels/

4.3 Assessing MVNcall in low-coverage pilot data

We now shift our attention to the MVNcall method and we examine different features of the method. We begin by examining the performance of the method under various parameter settings including: the number of iterations, the initialization procedures discussed in Section 3.3.1, the parameter λ, and the number of flanking sites.

4.3.1 Performance of MVNcall under various parameter settings

We begin by examining how the different tuning parameters affect the performance of our model by considering genotype calling performance under different parameter settings.

Number of iterations and burn-in

To check the required number of iterations and burn-in in MVNcall, we carry a convergence analysis where we vary the number of iterations and the number of burn-in iterations. In particular, we randomly choose 1,000 SNPs that are present in HapMap2 but not in HapMap3, and we run MVNcall for iterations $=\{50, 100, 500, 1000\}$ and burn-in $= \{10, 50, 200, 500\}$, where iteration $>$ burn-in. For all runs we use λ=0.05 and we set the number of flanking sites at either side to 40. For each of the 10 runs we compare the genotype accuracy by considering the HapMap2 genotypes as the "truth".

Instead of applying a single threshold on the maximum posterior probability

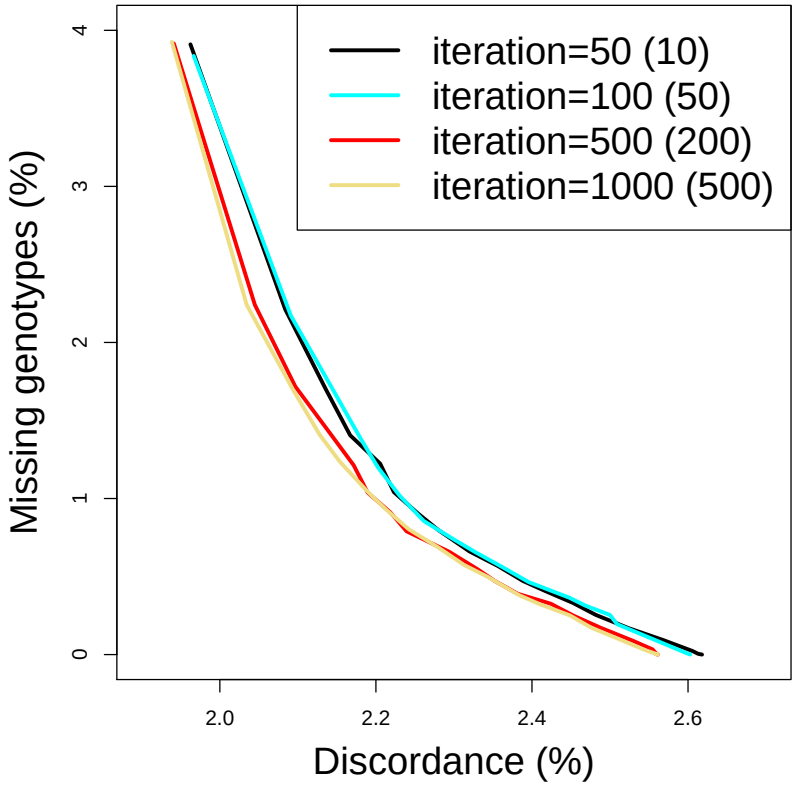

Figure 4.2: Percent discordance versus percent missing genotypes for MVNcall method with $\lambda = 0.05$ and number of flanking SNPs at either site to 40. We present results for 50, 100, 500, 1000 iterations with 10, 50, 200, 500 burn-in iterations respectively. The results correspond to threshold ranging from 0.33 to 0.99. The comparison is made on 1,000 SNPs present in HapMap2 but not in HapMap3.

of the genotypes to create a final genotype set and perform genotype comparisons, we instead consider a range of threshold values. For each threshold (ranging from 0.33 to 0.99 with increments of 0.03 in Fig.4.2), we calculate the percent dicordance for all genotypes that pass the threshold on the x-axis and on the y-axis we plot the percentage of genotypes with maximum posterior genotype probability less than the threshold (set as missing genotypes). In Fig.4.2, we present the curves for different number of iterations and burn-in. We notice that for a fixed number of iterations, the curves for the different number of burn-in

are indistinguishable, with a slight preference to the iteration run with the highest number of burn-in examined. Thus, for clarity, we only present the curves for each iteration tried with the largest burn-in. The curves that are closer to the origin have a lower percentage of missing genotypes and lower percent discordance rates than curves that are further away from the origin. The curves for 50 and 100 iterations are overalapping, and we see an improvement as we increase the number of iterations to 500. Increasing further the number of iterations to 1000, shows a marginal improvement. For the rest of the low-coverage analysis we set the number of iterations to 500 and the number of burn-in iterations to 200.

Initializing the haplotypes at the candidate site

In Section 3.3.1 we presented an alternative procedure for initializing the phasing of hererozygous genotypes. The procedure for phasing the initial heterozygous genotypes is based on the idea of incorporating information from the homozygous genotypes for phasing, as it is discussed in Section 3.3.1. This procedure will have some limitations: firstly, when the number of homozygous genotypes is small, the initial reference panel will only consist of a small number of haplotypes, and thus little information is available for phasing the heterozygous genotypes. Secondly, when all the homozygous genotypes involve the same allele, and thus the second allele is not present, there will be a bias towards the present allele. This initialization procedure is introduced as an alternative to random phasing of the heterozygous genotypes. Of course, if the

MCMC converges, it should "forget" its initial values and converge to the stationary distribution.

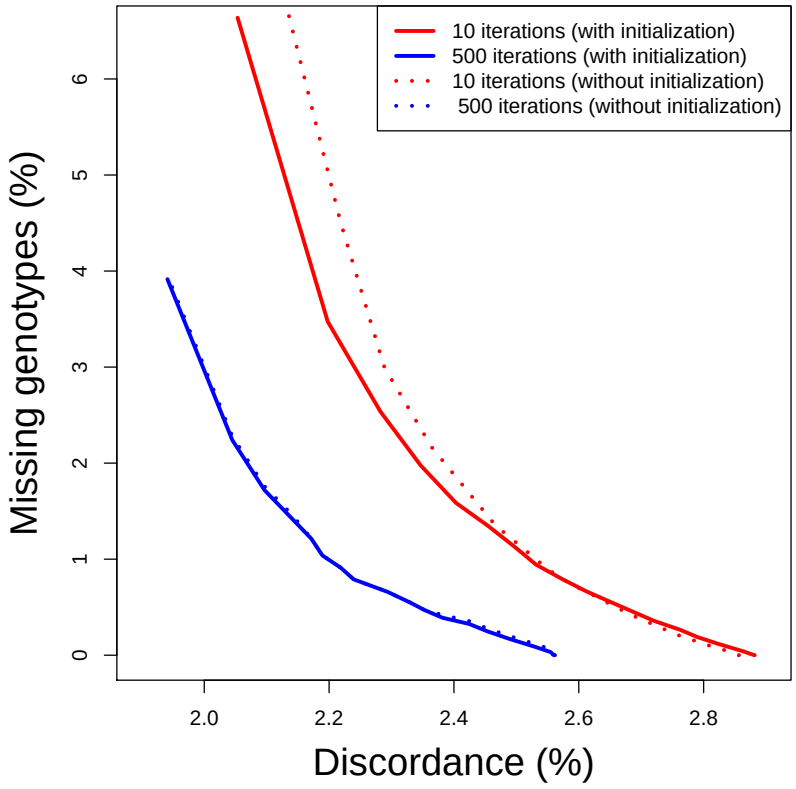

Figure 4.3: Comparison of two proposed procedures for initial phasing of the heterozygous genotypes. We propose an initialization procedure that incorporates the MVN model to phase in turn the heterozygous genotypes given the already phased homozygous genotypes prior to the MCMC step of the method. "without initialization" on the plot refers to the random phasing of the heterozygous genotypes. We plot the curve discordance versus percent of missing genotypes for 10 and 500 iterations.

We fix all other parameters, and we run the algorithm with and without the initialization step. We run the algorithm for 500 iterations (200 burn-in) and output intermediate results at the 10^{th} iteration. In Fig.4.3, we plot the percent discordance versus percent missing genotype curves for the two runs. After 10 iterations, the performance of the initialization step procedure outperforms that

of random phasing, though after 500 iterations the curves are overlapping.

The suggested initialization procedure for phasing the heterozygous genotypes is preferred contra the random phasing since it gives a better initial starting point for the algorithm. However, this will not have an effect in the long run, since when we run the chain long enough it does converge to the stationary distribution.

Parameter λ

The parameter λ was introduced in the model to ensure the covariance matrix is invertible, and thus allows the calculation of the density. From experiments, a value of λ in the range between 0.04-0.1 is required in our algorithm. When λ is above this range it ends up considering only the information from the genotype likelihoods and not the LD information, thus giving results similar to the non-LD based methods. On the other extreme, if the value of λ is below the range, only the LD information is taken into account and thus gives poorer results, since none of the sequencing information is used.

We first try to understand the role of λ in the model. The parameter λ is added on the variances of the variance-covariance matrix Σ, i.e. it increases the variances compared to the covariances, thus making the SNPs less dependent on their nearby SNPs. By increasing the value of λ, the SNPs are more independent, which directly means that less LD information is taken into account, and vice versa. Therefore, λ is a parameter that tunes the amount of information to be incorporated from the LD information, and as a consequence from

the sequencing data, i.e. the genotype likelihoods. A very small number of λ corresponds to a model that takes most of its information from the LD structure and less from the genotype likelihoods. At the other extreme, if $\lambda \to \infty$, then the variance increases, the multivariate-normal distribution is flattened and it can be approximated by a uniform distribution. The value of λ can also be considered as the error rate we expect from the sequencing data. Whereas, in the Wen and Stephens (2010), the parameter added on the diagonal of the covariance matrix corresponds to the expected mutation rate, in our model the value of λ corresponds to the error rate from the genotype likelihoods, i.e. from the sequencing data. This can explain why the value considered in our model is larger than the value considered in the Wen and Stephens (2010). It is reassuring to see that the optimal value of λ does not vary greatly for different reference panel sizes and different population groups, as we will see in the next chapter.

In Fig. 4.4, we plot λ versus overall genotype discordance for $\lambda \in$ [0.03,0.07] for 3,531 SNPs present in HapMap2 but not in HapMap3.

The overall percent genotype discordance[7] reaches a minimum of 2.45% for λ=0.05, where for λ=0.04 and λ=0.06 the overall discordance is slightly higher 2.47% and 2.49% respectively, which justifies the choice of λ=0.05 for the present analysis.

[7]We also checked the non-reference overall discordance, to see how the method performs on heterozygous and homozygous non-reference genotypes, and the minimum genotypre discordance reaches a minimum at λ=0.05.

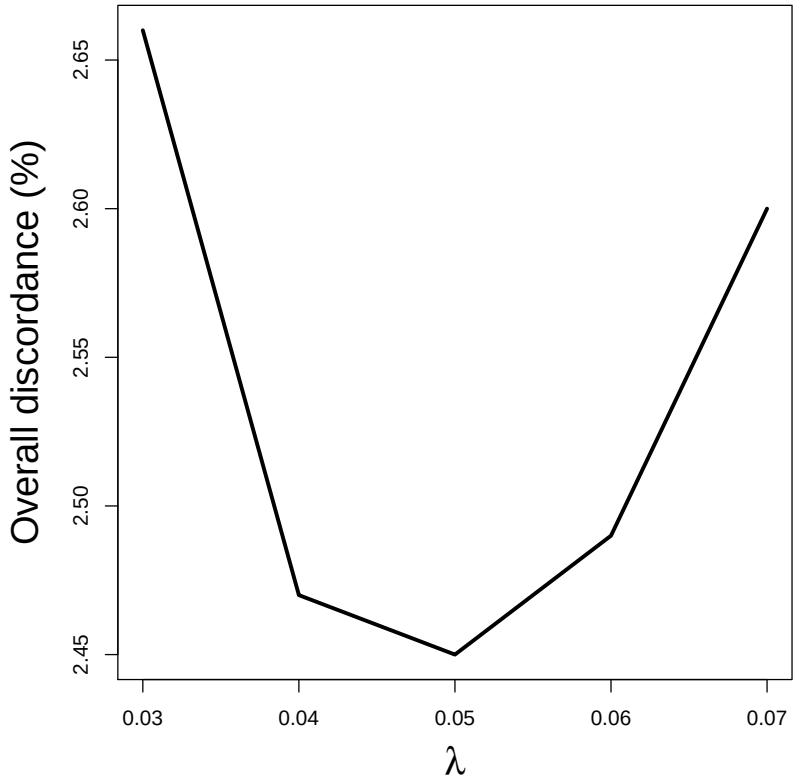

Figure 4.4: λ versus overall genotype discordance for $\lambda \in [0.03,0.07]$ for 3,531 SNPs present in HapMap2 but not in HapMap3 for 39 individuals that overalp between HapMap3 and the low-coverage pilot of 1KGP.

Number of flanking sites

MVNcall takes a fixed number of flanking sites (from the scaffold) from the position of interest. The number of flanking sites greatly affects the computational time of the algorithm since the number of flanking sites determines the size of the covariance matrix in the model. In this section we vary the number of flanking sites and we measure the percent overall genotype accuracy.

In Table 4.2 we present the percent overall discordance rate for different number of flanking sites. The method gives more accurate genotypes as we increase

Number of flanking sites	20	30	40	50	60
Percent overall genotype discordance	2.49	2.48	2.45	2.46	2.45

Table 4.2: Percent overall discordance for different number of flanking sites at either side of the position of interest. The number of flanking sites varies from 20 to 60 with increments of 10.

the number of flanking of sites, which is expected, since as we increase the number of flanking sites we incorporate more information. However, we see that the genotype discordance rate reaches a plateau after we incorporate more than 40 flanking sites at either side. As we go further away from the site of interest, the LD between the SNP of interest and the additional SNPs is decreased, and thus the addition of extra SNPs will not add more LD information. Of course, this depends on whether the site is in a low LD region or not. In the case we are in a low LD region, by adding more flanking sites, we are adding more noise, which explains why we see a slight increase when the number of flanking sites at either side is 50.

Comparison of point estimates of the covariance matrix Σ

As we are adding sites from the scaffold that are further away from the position of interest, LD decreases and the probability of a recombination event between the two sites increases. Given a genetic map, we can take into account the probability of recombination between two sites, and shrink the covariance between the two sites if the probability of recombination between them is high and vice versa. This idea is incorporated in the estimation of the variance-covariance matrix in E.q.3.7. We thus compare the method when we apply shrinkage on the

covariance entries of the variance-covariance matrix estimates versus the vanilla version where we only add a penalty value on the diagonal of the covariance matrix (λ) as described in E.q. 3.10. The percent overall genotype discordance falls from 2.45% to 2.41% when we perform shrinkage on the off-diagonal element of the variance-covariance matrix. We conclude that there is a benefit in applying shrinkage as this reduces the noise in the data. We also considered the case where instead of calculating point estimates to take a Bayesian approach to parameter estimation (see Section 3.2.2) but the genotype accuracy achieved was lower than the ones presented in this analysis (data not shown).

4.4 SNP and genotype calling comparisons in low-coverage pilot, chromosome 20 data

The following methods were applied on the low-coverage pilot of the 1KGP (The 1000 Genomes Project Consortium, 2010):

- **QCALL**[8]: A dynamic programming algorithm is ran to create a candidate set. Then QCALL was subsequently applied on the candidate set. The genome is divided in windows of 1Mbp, and for each window 20 ancestral recombination graphs (ARGs) are built using MARGARITA, using genotype or haplotype information (HapMap3) when available. For each candidate site, 40 marginal ancestral trees are sampled from the ARGs and used in QCALL which outputs the posterior probability of being polymorphic,

[8] ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/pilot_data/release/2010_03/pilot1/supporting/CEU.SI.pilot1.vcf.gz

genotype calls, and phased haplotypes for the individuals in the study. Sites with a posterior probability less than 0.9 are filtered out. Coverage filters[9] are also applied. After QCALL is ran, the fw10 filter is also applied.

- **Thunder**[10]: GlfMultiples is the non-LD method applied to create a candidate set for Thunder. The posterior probability of a site being polymorphic is calculated with GlfMultiples, and sites with a posterior probability greater than 0.9 are considered as candidate SNPs. Coverage[11] and indel[12] filters are applied. Then Thunder (MACH) is run for 100 iterations for the set of candidate sites. After Thunder is applied on the candidate set for genotype calling, the r^2 statistic that estimates the correlation between the estimated and the true underlying genotypes is calculated and sites with r^2 <0.5 are excluded.

- **GATK/BEAGLE**[13]: GATK is the non-LD method used to create a candidate set, where sites with a posterior probability of site being polymorphic greater than 0.9 are included in the candidate set. Then, BEAGLE was applied on the candidate set for genotype calling.

An integrated call set from the low-coverage pilot was created, that included SNPs that were called in at least two of the three aforementioned methods (we

[9] s50 filter

[10] ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/pilot_data/release/2010_03/pilot1/supporting/CEU.UMich.pilot1.vcf.gz

[11] s50 and high/low coverage filter

[12] distance from known indels

[13] ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/pilot_data/release/2010_03/pilot1/supporting/CEU.BI.pilot1.vcf.gz

refer to this final callset as the "2-3 way set"[14]). Also, a consensus of genotype calls was created using the three methods, by selecting the genotype preferred by the majority of the methods. For genotype comparison, we compare the genotype calls with SNPs that are present in HapMap2 but not in HapMap3 to avoid biases. Out of the 55 individuals analysed, 39 of them are present also in HapMap2, and we base our genotype comparisons on this set.

4.4.1 Comparison of methods on SNP calling

We now move on comparing our two methods with the other three existing methods for SNP and genotype calling used in the low-coverage pilot of the 1KGP. In this section we focus only on SNP calling. The comparison is based on the 55 individuals present in our methods and the other methods. The final call sets from all methods are considered.

Interpretation of results

Summaries of the final call sets from the different methods applied on the low-coverage pilot data[15] are presented in Table 4.3. The different methods report a final number of SNPs that ranges from 24,176 (GATK) to 24,367 (TreeCall). To assess the performance of the different call sets, we compare the number of SNPs called that are present in publicly available databases including SNPs from dbSNP (in this analysis we use dbSNP129) as well as HapMap sites. Since TreeCall,

[14]The final call set is termed *2-3 way* because a voting system was used where a SNP was included in the final call set if at least two of the three methods applied on this dataset, called the site polymorphic.

[15]CEU individuals, chromosome 20 (30,000,000-40,000,000)

Method	TreeCall	QCALL	Thunder	GATK
# SNP calls	24,367	24,216	24,314	24,176
# dbSNP129	15,714	15,890	15,485	15,328
# Hap2 (not in Hap3)	3,537	3,723	3,643	3,626
# Novel SNPs	8,563	8,326	8,828	8,848
Ts/Tv	2.06	2.24	2.02	2.07
Ts/Tv (known)	2.25	2.29	2.21	2.27
Ts/Tv (novel)	1.77	2.15	1.73	1.76

Table 4.3: Comparison of TreeCall, QCALL, Thunder and GATK in terms of SNP discovery. We present the total number of SNP calls, the number of SNPs present in dbSNP129, HapMap2 (not in HapMap 3), the number of novel SNPs, and the Ts/Tv ratio.

and QCALL have used the HapMap3 haplotypes as a scaffold, we compare the call sets against SNPs that are present in HapMap2 but not in HapMap3.

The percentage of the SNPs in the final call set that are present in dbSNP in ascending order are 63.4% for GATK, 63.7% for Thunder, 63.8%, for TreeCall, and 65.6% for QCALL. QCALL also has the highest intersection with SNPs that are present in HapMap2 but not in HapMap3, followed by Thunder, GATK, and TreeCall. The number of SNPs that are novel to each method is an indication of possible false positives. Thunder has the highest number of novel calls with 8,828, followed by GATK (8,848), TreeCall (8,563), and QCALL (8,326). Considering however the percentage of the SNPs called by each method that are novel GATK has the highest percentage of novel SNPs with 36.6% and QCALL the lowest with 34.4%. Another way to assess the quality of the SNP sets is to calculate the Ts/Tv ratio and compare it to the expected $\sim$2.0-2.1. All methods have a Ts/Tv ratio in the expected interval apart from QCALL that has an overall Ts/Tv of 2.24. Breaking down the Ts/Tv ratio for known and novel SNPs, we figure that the Ts/Tv ratio for known SNPs is similar to all five methods ranging

from 2.21 to 2.29 whereas the Ts/Tv ratio for novel SNPs is much lower for all methods apart from QCALL. The lower Ts/Tv ratio of the novel SNPs can be explained by false positives SNPs, as well as the different Ts/Tv ratio for low frequency variants (DePristo et al., 2011), since the allele frequency spectrum of novel variants includes rarer SNPs, accounting for the fact that most of the known variants are common.

TreeCall provides comparable results with the other competing methods on SNP discovery. QCALL achieves the highest intersection with the known SNPs and the lowest number of novel SNPs, an indication of lower false positive rate than the other methods. It should be noted that the other methods have used all 60 individuals in their analysis, compared to our methods that only 55 individuals have been used. We believe that the larger sample size will give a benefit on SNP discovery.

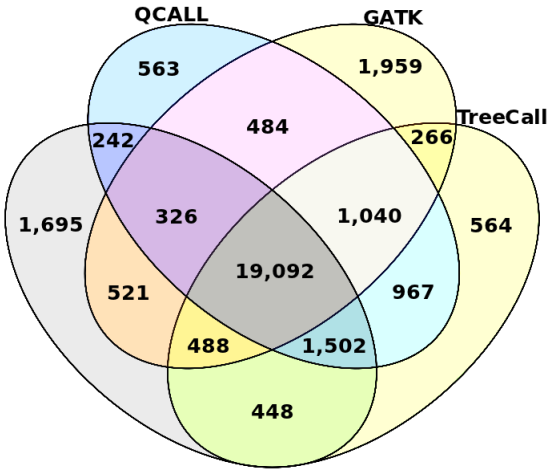

Figure 4.5: Four-way Venn diagrams comparing the intersection of the three methods with TreeCall. We compare only the filtered sets.

A further comparison on SNP discovery apart from the comparison with known SNPs, is the comparison of intersections between the different call sets.

In the low-coverage pilot, a consensus of SNP calls achieves a Ts/Tv ratio closer to the expected and the rediscovery rates are higher than the corresponding ones from each one of the methods individually. We thus compare the intersection of TreeCall with the other three methods. To present the results in a clearer form, we provide two four-way Venn diagrams, comparing each time the three existing methods with one of our methods. In Fig.4.5, on the left panel, we display the Venn diagram of the competing methods with TreeCall method. The intersection of all four methods is 19,092 sites (78% of the TreeCall calls), where 326 calls made by the other three methods are not made by TreeCall. Comparing the three-way intersection for the other methods, GATK does not call 1,502 SNPs called by the other three methods, followed by Thunder (1,040), QCALL (488), and TreeCall (326). From the two-way interactions, we see that the largest two-way intersection is between QCALL and TreeCall, with 967 SNPs, which can be justified by the similarity of the two methods. The number of calls private to each method in ascending order are : 563 by QCALL, 564 by TreeCall, 1,695 by Thunder and 1,959 by GATK.

We find it encouraging that our method has a high intersection with the other three methods and also have low number of calls private to its method, the second lowest in both cases. Both methods grasp a high proportion of the calls made by the other three methods ($\sim$98%) and have a low percentage of calls made only by its method (2.3% for TreeCall).

We should note here that the different methods have applied different filters, and some discrepancies on the call sets are the result of the different filters

applied. Considering the sites private to TreeCall, 506 out of the 564 sites are present in the unfiltered set of QCALL, and 417 in Thunder, which shows the different effect of filters on the different data sets.

4.4.2 Comparison of methods on genotype calling

The main use of the LD-based methods is for genotype calling. We compare the genotype calls from the different LD-based methods in Table 4.4. The comparison is based on HapMap2 SNPs, not in HapMap3, a set of 3,531 SNPs, for 39 individuals that are present in both projects.

Method	HomR %	Het %	HomA %	Overall %	R^2
TreeCall	2.70	4.85	5.31	3.86	0.9620
MVNcall	**1.99**	**2.39**	**3.37**	2.41	0.9705
QCALL	2.39	2.96	3.71	2.84	0.9658
Thunder	2.02	3.54	5.19	2.82	0.9641
GATK	2.02	4.02	5.58	3.34	0.9624

Table 4.4: Percent genotype discordance comparison between TreeCall, MVNcall, QCALL, Thunder and GATK. HomR corresponds to homozygous reference genotypes, Het to heterozygous genotypes and HomA to homozygous alternative genotypes. Comparisons based on HapMap2 sites not in HapMap3 on 39 individuals.

MVNcall achieves the lowest percent overall genotype discordance rate of 2.41%, followed by Thunder, QCALL, GATK, and TreeCall. The main advantage of MVNcall compared to the other methods is on the calling of heterozygous genotypes, with a percent genotype discordance 0.57% less than the second best method which is QCALL. TreeCall has the highest discordance rate for homozygous reference genotypes where GATK (BEAGLE) has the highest genotype discordance for homozygous alternative genotypes. We also compare

the genotype accuracy versus the minor allele frequency and we do not see any striking differences between the different methods (see Appendix B.2).

Given the small sample size and the low-coverage information from the sequencing reads, the incorporation of information from known haplotypes, greatly aids the genotype calling of the sites, as in the case of MVNcall and QCALL. Furthermore, the haplotypes used in this analysis are from the HapMap3 project, derived using trio information, and thus are highly accurate. Therefore, MVNcall and QCALL try to place the genotypes of the novel sites in a set of very accurate haplotypes. We believe that a reason that MVNcall performs better than the tree-based methods (QCALL and TreeCall) is due to the fact that in all underlying genealogies some branches will be wrong and this will lead to lower accuracies. By using a simple representation of the LD structure, as the MVNcall, we were able to represent the LD relationships among the SNPs, in a more diffuse way and achieve better accuracies.

TreeCall has the highest overall genotype discordance rate compared to the other methods. Different reasons can explain this observation. Firstly, we construct marginal genealogical trees on the midpoint of two scaffold sites and use these trees for all sites falling in that interval. In this way, we implicitly made the assumption that there is no recombination event in that interval, which might not be the case. Moving further away from the site we built the trees, the probability of recombination increases. In the event of recombination, the genealogical history changes, and the trees constructed do not represent the genealogy of the site. This situation is not met in the case of QCALL, since ancestral trees

are sampled from ARGs for each candidate site. Also, we did not study how the performance of the method varies for different parameters for constructing the trees, including the number of flanking sites to build the trees and the option to choose the flanking sites according to genetic distance. We believe that if this investigation was studied further, the genotype accuracy of TreeCall will improve. However, given the computational cost of TreeCall, both for tree construction and running TreeCall, and the limitation of the method to be applied in larger datasets, we did not carry a further analysis.

4.5 Summary

In this Chapter we applied TreeCall and MVNcall on the CEU individuals from the low-coverage pilot of 1KGP. We first presented the investigation of different parameter settings for the two methods and the filtering step followed to create a final call set. We then moved on and compared the call sets between our methods and the competing methods applied on the same dataset. Both methods provide high intersection of calls with the other methods and achieve the expected Ts/Tv ratio. We then compared the genotype calling accuracy of the different methods. MVNcall had the lowest genotype discordance rate from all methods whereas TreeCall had the highest genotype discordance.

Chapter 5

Genotype calling results for phase 1 data

After the completion of the pilot phase of the 1KGP, the project is in the process of sequencing 2,500 individuals from different population backgrounds at 4x coverage. During the course of this thesis, phase 1 of 1KGP is completed where 1,092 individuals from 14 population backgrounds are sequenced. Phase 1 provides a larger sample from diverse populations which made some methods applied in the low-coverage pilot unattractive. QCALL[1] and similarly our method TreeCall, were computationally expensive to be applied to this dataset. Thus, TreeCall was not applied on this dataset and in this section we only apply MVNcall on the Phase 1 dataset.

In Phase 1, a common candidate set was used by all LD-based methods, and the LD-based methods were solely applied for genotype calling and phasing.

[1]MARGARITA (Minichiello and Durbin, 2006) uses a heuristic approach to construct ARGs for a sample of sequences and it has been reported that it gets locked in incorrect structures as the sample size increases (Le and Durbin, 2011), which makes QCALL unable to be used in large projects

Thus, in this chapter we compare the different methods in terms of genotype and phasing accuracy only.

In this chapter we present the following:

- Exploration of the MVNcall tuning parameters for the current dataset,
- Genotype calling comparison of MVNcall with other competing methods,
- Phasing accuracy comparison of the different methods, using trio phased HapMap3 haplotypes, where available, as the truth,
- Imputation analysis using the resulting haplotypes from the different methods the reference panels,
- Application of the extension of the model to handle trio information for genotype calling on a deeply sequenced CEU trio,
- Comparison of genotype accuracy of MVNcall for different SNP density scaffolds.

5.1 Data processing

In this analysis, we investigate the performance of the different methods on genotype calling on the individuals from Phase 1. Although the full project involves the sequencing of trios (father, mother, offspring) from different populations, in Phase 1 only founder individuals have been sequenced. The populations can be divided into four population groups according to their ancestry: Asian ancestry (ASN), European ancestry (EUR), African ancestry (AFR)

and Admixed ancestry (ADM). The 14 populations divided by the population groups (and the number of individuals in brackets) are the following

- Asian ancestry
  - **CHB** (97): Han Chinese in Beijing, China
  - **CHS** (100): Han Chinese South, China
  - **JPT** (89): Japanese in Tokyo, Japan
- European ancestry
  - **IBS** (14): Iberian populations, Spain
  - **FIN** (93): HapMap Finnish individuals, Finland
  - **GBR** (89): British individuals from England and Scotland, United Kingdom
  - **CEU** (85): Utah residents (CEPH) with Northern and Western European ancestry
  - **TSI** (98): Tuscan individuals, Italy
- African ancestry
  - **YRI** (88): Yoruba in Ibadan, Nigeria
  - **LWK** (97): Luhya in Webuye, Kenya
- Admixed ancestry
  - **ASW** (61): HapMap African ancestry from South West, United States of America

- **MXL** (66): HapMap Mexican individuals, Los Angeles California, United States of America
- **PUR** (55): Puerto Rican, Puerto Rico
- **CLM** (60): Colombian in Medellin, Colombia

In total 1,325 BAM files are generated, 858 individuals are sequenced in one platform, 232 individuals in two plarforms and 1 individual in three plarforms. Illumina was used for the sequencing of 975 genomes, ABI SoLiD for 338 genomes, and 454 for 12 genomes. The sequencing reads have been aligned by different alignment algorithms, according to the sequencing technology, i.e. Illumina reads have been aligned using BWA (Li and Durbin, 2009), 454 reads with SSAHA (Ning et al., 2001) and SoLiD reads with BFAST (Homer et al., 2009) using the NCBI37 reference genome. Six different non-LD based methods have been applied to create a candidate list of sites, which were then filtered using GATK (VQSR version) to create a list of SNPs. The genotype likelihoods were produced from SNPTools using the BAM-Specific Binomial Mixture Model (BBMM)[2].

As part of the 1KGP, all individuals of the full project that are sequenced, they are also genotyped on the OMNI2.5 chip[3]. The family of omni microarrays is developed in collaboration with the 1KGP, where the OMNI2.5 chip consists of

[2]Both Samtools and SNPTools have been applied to generate genotype likelihoods for this set. Comparing the genotype accuracies for this two different genotype likelihood sets, all LD-based methods perform better with the SNPTools GLs, the same observation was made when we used the different genotype likelihoods in MVNcall, and thus those genotype likelihoods are used by all the methods in the final call set.

[3]http://www.illumina.com/documents/products/datasheets/datasheet_gwas_roadmap.pdf

$\sim$2.5 million SNPs with MAF $>$2.5%. The OMNI2.5 chip includes novel variants discovered by the 1KGP pilot analysis, and it is designed to give the maximum information for diverse populations.

Subsequently, the 1KGP provides both genotype and sequencing information for all individuals, which are the prerequisites to use MVNcall. The SNPs from the OMNI2.5 chip serve as a scaffold for MVNcall. We used the phasing software SHAPEIT (Delaneau et al., 2011), which appears to be the best available method for phasing at the moment, able to phase a whole chromosome at once, to create the haplotype scaffold of our method from the OMNI genotypes. It should be noted, that genotypes from 2,123 individuals are used by SHAPEIT, which contained 327 trios, 42 duos and 1,058 unrelated at 2,177,885 SNPs for phasing. It is the case, that the haplotypes of some founder individuals in the Phase 1 are phased using trio information.

The analysis presented in this chapter is on chromosome 20, with a total number of 824,954 SNPs. The number of SNPs in OMNI2.5 for chromosome 20 is 54,241.

5.2 Genotype calling

Using the data from 1KGP, we first explore the different tuning parameters for MVNcall in this setting on a 10Mbp region. After learning about the parameters of MVNcall, we apply the method on the whole chromosome 20 and we compare the genotype accuracy of MVNcall with three other methods that have

been applied on the same dataset.

5.2.1 Exploration of MVNcall parameters

To begin with, we proceed as above for the low-coverage pilot analysis and we investigate the tuning parameters of MVNcall. However, in this analysis, the extended model of MVNcall is applied for large samples (see Section 3.4), where apart from the tuning parameters, λ and number of flanking sites, we also need to choose the k closest haplotypes for each individual's haplotype for inference. We also investigate different metrics for choosing the closest haplotypes suggested in Chapter 3, as well as the idea of model averaging. We also test whether analysing all populations together versus analysing the populations seperately, a strategy adopted by Thunder on Phase 1 analysis, is preferred under our model.

We run MVNcall under different parameter settings to investigate the performance of the method, on a 10Mbp region on chromosome 20 (chr20:40,000,000-50,000,000[4]) and we base the genotype accuracy comparisons on SNPs that are present in HapMap3 but not in OMNI (2,935 SNPs), since OMNI SNPs are used to construct the haplotype scaffold. Given that we compare the results with HapMap3, we only consider individuals present both in HapMap3 and 1KGP (591 individuals). The genotype likelihoods used are from the March release of the SNPTools genotype likelihoods[5].

[4]We choose to analyse a different region than the region analysed on the low-coverage pilot analysis, so the parameters are not tuned to one specific region and to ensure that the parameters can be applied to the whole genome.

[5]ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/working/20110329_wgs_genotypes/bcm/

In this section we compare (i) the performance of MVNcall for different haplotype distance measures for choosing the conditioning haplotypes, (ii) how the method performs when we analyse all the populations together versus by population group, and (iii) the improvement achieved when we perform model averaging. All the comparisons presented use the following tuning parameters: λ (0.06), number of flanking sites (50), number of conditioning haplotypes (100), number of iterations (100) and number of burn-in iterations (10).

Perfect match versus Hamming distance

In Section 3.4.1, we discussed two different haplotype distance measures for choosing the closest haplotypes namely the *Hamming distance* and the *Perfect match distance*. The former distance takes into account the number of mutations between two haplotypes and the latter takes into account recombination since it counts the maximum length of identical share between two haplotypes. It should be noted, that for the Perfect match distance, we would like to have symmetric information surrounding the site of interest. We use a simple procedure for those cases and haplotypes with a mismatch within two sites of the site of interest are down-weighted. In Table 5.1 we present the percent genotype discordance between the two metrics seperated by population group and genotype type to investigate the effect of the different metrics on different populations.

Perfect match outperforms Hamming distance measure for all populations with an apparent advantage on the heterozygous genotypes where the performance improves by 0.36% for populations with African ancestry, 0.23% for pop-

		Percent Genotype discordance			
		HomR	Het	HomA	Overall
African	Hamming distance	0.21	1.23	0.84	0.62
	Perfect match distance	0.25	0.87	0.82	0.54
Asian	Hamming distance	0.18	0.89	0.73	0.46
	Perfect match distance	0.18	0.66	0.65	0.39
European	Hamming distance	0.16	0.83	0.68	0.42
	Perfect match distance	0.15	0.62	0.63	0.35

Table 5.1: Comparison of the haplotype distance measures for choosing the closest haplotypes: Hamming distance and Perfect match distance by population group. Comparison based on HapMap3 SNPs not in OMNI.

ulations with Asian ancestry and 0.21% for populations with European ancestry. The only situation where Hamming distance performs better than the Perfect match distance is for the homozygous reference genotypes in the African population group by 0.04% and the same genotype discordance is achieved for the homozygous reference genotypes in the Asian population group (0.18%).

The first advantage of the Perfect match distance over the Hamming distance measure is the fact that it accounts for the position of the site of interest in the window studied, and calculates the distance starting from that site on the right and on the left of the site. Since our model analyses one position at a time and not a window of SNPs simultaneously, the Perfect match distance accommodates this information. On the other hand, Hamming distance measure does not account for this, and it gives equal weight to mismatches that are close to the site of interest and to mismatches that are further away from the site of interest which appears not to be ideal in our model.

A distance measure that could account for both the number of mutations and length of the sharing haplotypes would be ideal. However, the issue of

combining those two metrics, and the amount of weight given to each of the two metrics at a given region, it is not trivial. Experiments with different weights did not achieve better genotype accuracies than the Perfect match distance. We also tested our method by applying a Perfect match distance that allows more than one mismatch, or weights the mismatches according to physical and genetic distance from the site of interest, (results not shown), but we did not achieve a higher genotype accuracy than the vanilla version of the Perfect match distance. Therefore, the rest of the analysis is performed using the Perfect match distance to choose the closet haplotypes.

Analysing all populations together versus analysing them by population group

In Phase 1, the populations were divided into four population groups according to their ancestry (see Section 5.1). In this section we compare the genotype accuracy performance of MVNcall if we run all populations together or if we run Asian & Admixed, European & Admixed, and African & Admixed[6] seperately. This strategy was followed by other methods for analysing the Phase 1 dataset. The percent genotype discordance for the two runs are presented in Table 5.2.

Running each population group seperately is preferred in MVNcall since the percent genotype discordance for all population groups are decreased compared to the scenario where we run all the populations together. The most apparent difference is met on the populations with African ancestry, where the overall ac-

[6]A slightly different population grouping is now adopted by the 1KGP for this dataset, including the ASW population in the African population group and naming the Admixed population group as Americas.

		Percent Genotype discordance			
		HomR	Het	HomA	Overall
African	All populations	0.26	0.90	0.90	0.58
	African & Admixed	0.25	0.87	0.82	0.54
Asian	All populations	0.20	0.69	0.68	0.41
	Asian & Admixed	0.18	0.66	0.65	0.39
European	All populations	0.15	0.65	0.62	0.36
	European & Admixed	0.15	0.62	0.63	0.35

Table 5.2: Percent genotype discordance comparison between running all populations together (All populations) versus each population group seperately with the admixed population group (Population group & Admixed). Perfect match distance was used to choose the closest haplotypes.

curacy differs by 0.04%, compared to 0.02% for populations with Asian ancestry and 0.01% for populations with European ancestry. For the individuals with admixed ancestry that were used in all runs, we find that they achieve the highest genotype accuracy in the run with individuals with African ancestry, and we thus use this run to call the genotypes for the admixed individuals.

The difference of the two scenarios lies on the choice of the closest haplotypes for each individual's haplotypes. The only difference on the two scenarios on the model, is mainly the choice of the closest haplotypes. By running the populations separately, we implicitly ban the method on choosing haplotypes from distant populations for inference. The haplotype distance measure adapted does not account for population diversity, as it only measures local similarity between haplotypes, which can justify why the method performs better when we only use haplotypes from the same population group. For this dataset, we decide to run the populations separately.

Model averaging runs for different number of conditioning haplotypes k

In the aforementioned analysis we used 100 conditioning haplotypes. Here, we further explore the behaviour of the MVNcall for different number of conditioning haplotypes. We thus run the method for different values of k, $k = \{10, 50, 100, 200, 300, 400, 500, 600, 700, 800, 900, 1000\}$ and we compare the genotype discordance from different runs (vanilla version). Here we present results for the Asian population, the other populations follow the same pattern. We further investigate the performance of the method for different values of k by stratifying the overall percent discordance by genotype type. In Fig.5.1 (red line) we plot on the x-axis the number of conditioning haplotypes and on the y-axis the percent discordance. The four plots correspond to the overall genotype discordance (top left), homozygous reference genotypes discordance (top right), heterozygous genotypes discordance (bottom right) and homozygous alternative genotype discordance (bottom right). We observe that the homozygous reference discordance decreases as we increase the number of conditioning haplotypes, and reaches a plateau for k >500. A similar behaviour is observed for the homozygous alternative genotypes. However, for the heterozygous genotypes, we reach the minumum genotype discordance for k=100 and then the discordance increases for k >100.

From a sequencing point of view, at low-coverage, the genotype likelihoods of the heterozygous genotypes are expected to be less confident than the corresponding genotype likelihoods for the homozygous genotypes, because for heterozygous genotypes we need to observe both alleles, instead of one in the

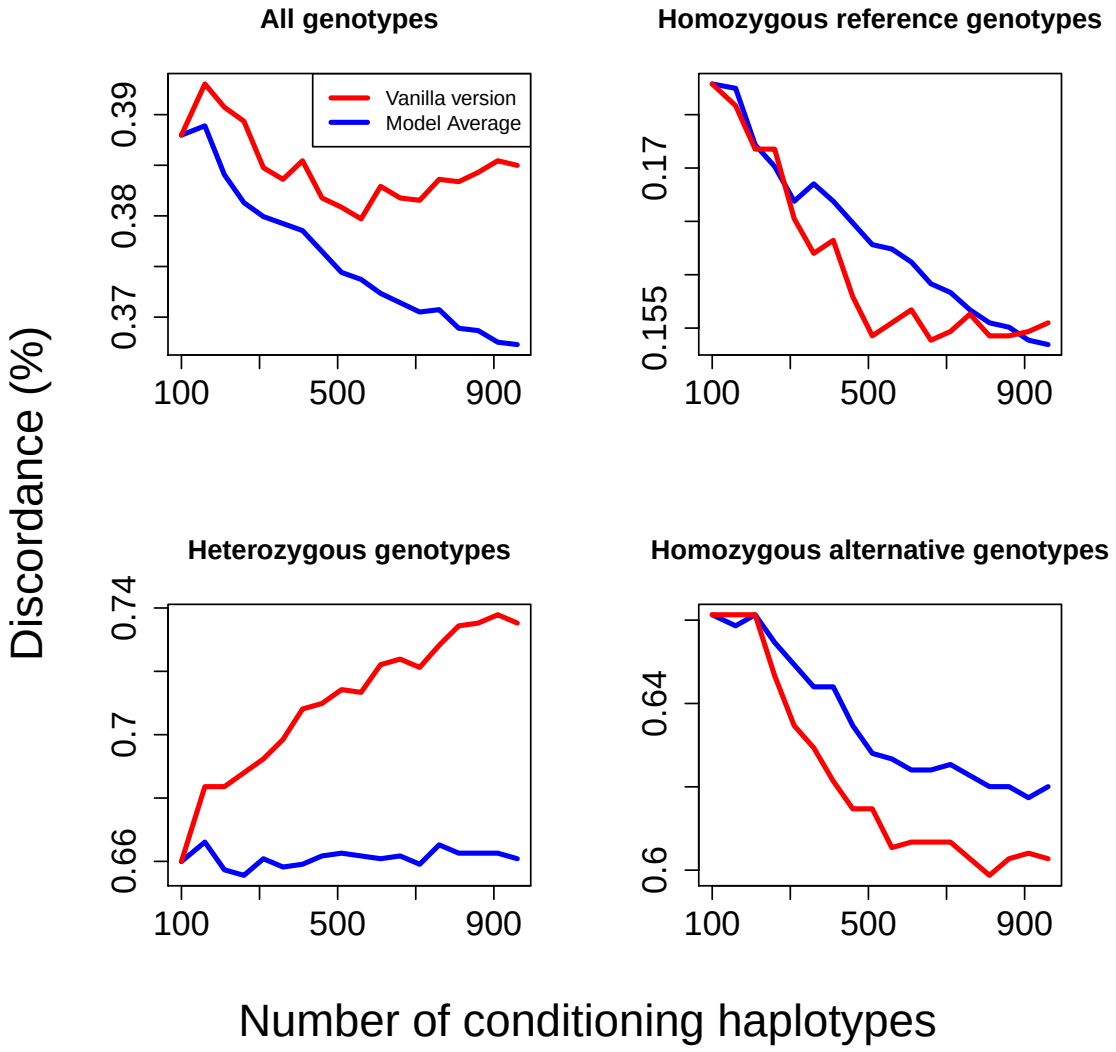

Figure 5.1: Percent discordance comparison of two different versions of the MVNcall model, the first version is the vanilla version and the second version is model averaging between two runs with different value of k, $\lambda = 0.06$, number of flanking sites at either side (50), iterations (100), burn-in (10), population group (Asian)

case of the homozygous genotypes. Now, in the case of MVNcall, by adding more haplotypes, we add more information when the haplotypes are closely related with the haplotype of interest and at the same time noise when the haplotypes added are less related with the haplotype of interest. Noise will distort

the genotype calls when the genotype likelihood for the true genotype is not confident, which is mostly the case for heterozygous genotypes. For example, we display some examples in Fig.5.2 where for a small value of k ($k = 100$) we call the heterozygous genotype correctly, whereas for a large value of k ($k = 1200$) we call the heterozygous incorrectly. For each of the three examples, we give the physical position, as well as the genotype likelihoods of the heterozygous genotype of interest (hom ref, het, hom alt). In terms of the genotype likelihoods, the maximum genotype likelihood is given to the homozygous reference genotype. Hence, only strong evidence in favour of the heterozygous genotype will aid the calling of the heterozygous genotype correctly.

For each of the three examples, we plot the posterior probability of each of the two haplotypes, where 0 corresponds to the referene allele and 1 to the alternative allele. For position 41169938, the alternative allele is carried by haplotype 2. By adding more haplotypes, we increase the amount of noise. Given that the sequencing information supports the homozygous reference allele, the genotype is called incorrectly. Similar behaviour is met to the other two positions.

A possible solution is to run the algorithm for different values of k, and average over the runs. There is of course a trade-off for model averaging, which is the number of models to average over versus the computational time. To keep the best of both worlds, we average over two models, the one with k=100, since it is the value with the minimum genotype discordance for the heterozygous genotypes, and we combine it with runs with varying value of k. As a note, by "averaging over" the two models, we just take the average posterior from the

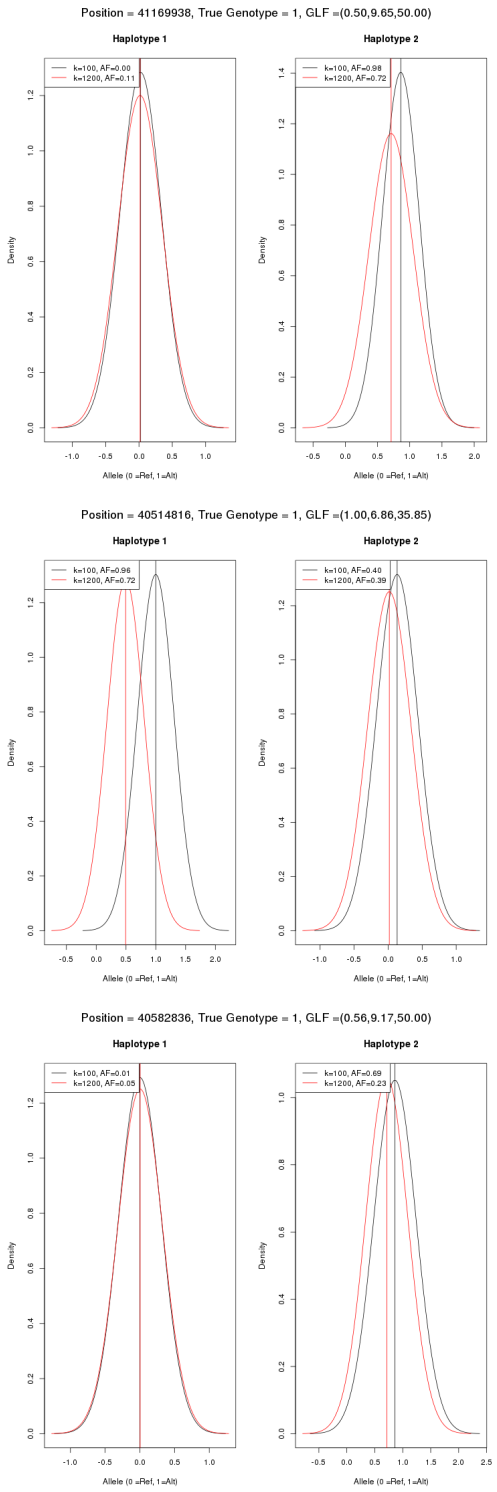

Figure 5.2: Density plots of heterozygous genotypes that are called correctly when $k = 100$ and incorrectly when $k = 1200$. Physical position as well as the genotype likelihoods are provided for each example.

two runs. In Fig. 5.1 the blue line denotes the method where we average over two runs (on the x-axis for this strategy, one run is with k=100 and the other run with k=y, on the x-axis we plot y). As we can see, by averaging over two runs, the genotype discordance decreases as the k increases for the homozygous genotypes, although for the homozygous alternative genotypes the discordance is higher than the vanilla version. The benefit of model averaging is apparent on genotype discordance for the heterozygous genotypes, which remains constant as we increase the value of k, and thus making the model more robust on the value of k. As a consequence, the overall discordance keeps decreasing as the value of k increases.

This assures that the method can benefit from all the available information from the data; for k=100, the model captures the local structure of the underlying genealogical tree, where when we use all the haplotypes in the reference panel, it captures the overall structure of the underyling genealogical tree. These two pieces of information seem to help the model. We use the model averaging idea for the full analysis presented below.

Exploration of the other tuning parameters and the number of MCMC iterations are summarized in Appendix C.1. MVNcall uses the following parameters and settings for the full analysis: λ=0.06, number of flanking sites at either side (50), Perfect match distance measure for choosing the conditioning haplotypes, model averaging option with two runs one with k=100 and the other with all the haplotypes in the reference panel, population groups are ran seperately, number of iterations (100), burn-in iterations (10).

5.2.2 Genotype calling comparison of methods

After learning about the parameters of MVNcall, we run the method on chromosome 20. In this section, we compare the genotype accuracies achieved by MVNcall with other competing methods applied on the same dataset, namely

- **Thunder**: Thunder has used the latest releases the genotype likelihoods from SNPTools and have applied the method by separating the populations by population groups as it was explained above. Thunder requires to run BEAGLE first to initialize the haplotypes. The release used for genotype comparison can be found in ftp://ftptrace.ncbi.nih.gov/1000genomes/ftp/technical/working/20120117_new_phase1_intgrated_genotypes/. The number of individuals in this release is 1,092.

- **SNPTools**: SNPTools has used the latest release of genotype likelihoods and the genotype comparison presented below is based on the release which can be found here ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/technical/working/20111028_bcm_phase1_integration/. SNPTools runs all populations together. The number of individuals in this release is 1,041.

- **BEAGLE**: We applied BEAGLE using the genotype likelihoods from the Thunder vcf file above. We split the chromosome in 3Mbp chunks with 200Kb overlap. We fed BEAGLE with all populations at once as it was applied in previous analysis. The number of individuals analysed is 1,092.

MVNcall has also used the genotype likelihoods provided in the Thunder vcf file and analysed 1,092 individuals. Even though Thunder and SNPTools have

also made indel calls, here we only consider the SNP calls they have made, for all subsequent analysis.

Below, we are going to compare the genotype accuracies on two different sets where there is genotype information from chips on a subset of the individuals present in the 1KGP. Since MVNcall uses a scaffold and in this analysis SNPs from the OMNI chip, our method will not call any SNPs in the scaffold, assumming that the scaffold is correct. To have an unbiased comparison between the different methods, we only compare SNPs that are present in other chips but not in OMNI. In particular, the two sets of genotypes we will compare the methods to are[7]:

- **HapMap3**: 591 analysed by all aforementioned methods are also present in HapMap3, specifically, 160 with European ancestry, 167 with Asian ancestry, 162 with African ancestry and 102 with Admixed ancestry. The number of SNPs in HapMap3 but not in OMNI2.5M, found in all four calls is 18,155.

- **Axiom**: A subset of 601 individuals was also genotyped on the Affymetrix Axiom chip: 167 with European ancestry, 162 with Asian ancestry, 165 with African ancestry and 107 with Admixed ancestry. The number of SNPs in Axiom genotype set not in OMNI, found in all four call sets is 85,858. The Axiom genotypes used can be found here ftp://ftptrace.ncbi.nih.gov/1000genomes/ftp/technical/working/20120208_axiom_genotypes/

[7]We also compared the different methods on a set of validation SNPs, the comparison is presented in Appendix C.2

Interpretation of results

We begin by comparing the percent genotype discordance between the different methods on HapMap3 sites not in OMNI. In Table 5.3, we present the percent genotype discordance divided by population group. To get a better understanding on the genotype errors, we break down the percent genotype discordance by genotype type, i.e. Homozygous reference genotypes (HomR), heterozygous genotypes (Het) and homozygous alternative genotypes (HomA). We also report the overall percent genotype discordance as well as the percent genotype discordance excluding the homozygous reference genotypes (Non-ref).

		Percent Genotype discordance					
		HomR	Het	HomA	Overall	Non-ref	R^2
African	BEAGLE	0.26	0.97	0.70	0.55	0.86	0.9951
	MVNcall	**0.20**	0.79	**0.62**	**0.45**	0.72	0.9959
	Thunder	0.26	**0.69**	0.67	0.46	**0.68**	0.9958
	SNPTools	0.29	0.68	0.74	0.49	0.70	0.9956
Admixed	BEAGLE	0.18	0.93	0.51	0.45	0.77	0.9958
	MVNcall	**0.16**	0.76	**0.47**	0.38	0.65	0.9963
	Thunder	0.18	0.71	0.49	0.38	0.62	0.9963
	SNPTools	0.18	**0.63**	0.51	**0.36**	**0.58**	0.9965
Asian	BEAGLE	0.12	0.70	0.38	0.31	0.56	0.9971
	MVNcall	0.12	0.63	0.37	0.29	0.51	0.9972
	Thunder	**0.10**	0.56	**0.34**	**0.26**	0.46	0.9975
	SNPTools	0.12	**0.50**	0.36	**0.26**	**0.44**	0.9975
European	BEAGLE	0.11	0.66	0.40	0.30	0.56	0.9970
	MVNcall	0.11	0.53	0.40	0.27	0.48	0.9972
	Thunder	**0.09**	0.49	**0.35**	0.24	0.43	0.9975
	SNPTools	**0.09**	**0.44**	0.37	**0.23**	**0.41**	0.9975

Table 5.3: Percent genotype discordance comparison of **HapMap3** SNPs not in OMNI on chromosome 20. Results are divided by population group and in the five coloumns we report the percent genotype discordance of: **HomR**: homozygous reference genotypes, **Het**: heterozygous genotypes, **HomA**: homozygous alternative genotypes, **Overall**: the overall genotype discordance, **Non-ref**: the percent discordance on heterozygous and homozygous alternative genotypes, and R^2 is the pearson correlation coefficient between the estimated and true genotypes.

In the African population group, MVNcall has the lowest overall discordance rate (0.45%) which is mainly driven by the difference on the genotype discordance on the homozygous genotypes, where MVNcall has reached a genotype discordance of 0.20% for the homozygous reference genotypes and 0.62% for the homozygous alternative genotypes, whereas the other methods have a genotype discordance above 0.26% and 0.67% respectively. MVNcall's non-ref discordance rate is higher than Thunder's and SNPTools by 0.04% from the former and 0.02% from the latter due to the higher genotype discordance on the heterozygous genotypes. MVNcall still outperforms BEAGLE by 0.18% on the heterozygous genotype discordance and achieves 0.1% lower overall discordance and 0.14% lower non-ref discordance rate. A similar pattern is observed on the Admixed population group, where MVNcall achieves again the lowest genotype discordance rates for the homozygous genotypes (0.16% for homozygous reference and 0.47% for homozygous alternative) and has the second lowest overall discordance rate of 0.38% with Thunder, after SNPTools that achieves an overall discordance of 0.36%. The lowest discordance rate on heterozygous genotypes is again achieved by SNPTools with 0.63%, followed by Thunder with 0.71%, MVNcall with 0.76% and BEAGLE with 0.93%.

For the Asian and European population groups, we notice that the genotype discordance rates decrease in all methods with overall genotype discordance rates ranging from 0.23% to 0.30% for the European group and from 0.26% to 0.31% for the Asian group. Thunder achieves the lowest genotype discordance rates for the homozygous genotypes in these population groups,

where SNPTools still outperforms all other methods on calling heterozygous genotypes. MVNcall has 0.05% higher discordance rate both for the overall and non-reference discordance rates than the second best. The main difference again lies in the heterozygous genotype discordance. Still, MVNcall outperforms BEAGLE with a significant improved genotype calls on the heterozygous genotypes; for the European group MVNcall has 0.13% lower discordance rate than BEAGLE and for the Asian group 0.07%.

		Percent Genotype discordance					
		HomR	Het	HomA	Overall	Non-ref	R^2
African	BEAGLE	0.27	1.62	1.63	0.71	1.62	0.9898
	MVNcall	**0.23**	1.48	**1.59**	0.65	1.52	0.9904
	Thunder	0.29	**1.24**	1.62	**0.65**	**1.38**	0.9905
	SNPTools	0.30	1.35	1.69	0.68	1.48	0.9901
Admixed	BEAGLE	0.29	1.25	**1.77**	0.73	1.45	0.9887
	MVNcall	**0.28**	1.07	**1.77**	0.68	1.34	0.9892
	Thunder	0.31	1.01	**1.77**	0.68	1.30	0.9892
	SNPTools	0.30	**0.92**	1.78	**0.66**	**1.25**	0.9893
Asian	BEAGLE	0.14	2.08	1.32	0.54	1.75	0.9910
	MVNcall	0.14	2.06	1.29	0.54	1.73	0.9912
	Thunder	**0.13**	**1.90**	**1.29**	**0.51**	**1.64**	0.9915
	SNPTools	**0.13**	**1.90**	1.31	**0.51**	1.65	0.9915
European	BEAGLE	0.15	1.46	1.47	0.52	1.46	0.9911
	MVNcall	0.15	1.34	1.46	0.50	1.38	0.9914
	Thunder	**0.14**	1.27	**1.44**	**0.48**	1.33	0.9916
	SNPTools	**0.14**	**1.24**	**1.44**	**0.48**	**1.31**	0.9917
All populations	BEAGLE	0.19	1.66	1.51	0.61	1.60	0.9904
	MVNcall	0.18	1.55	1.49	0.58	1.52	0.9908
	Thunder	0.19	1.40	1.49	0.56	1.44	0.9910
	SNPTools	0.20	1.42	1.51	0.57	1.46	0.9909

Table 5.4: Percent genotype discordance comparison on **Axiom** SNPs not in OMNI, divided by population group.Results are divided by population group and in the five coloumns we report the percent genotype discordance of: **HomR**: homozygous reference genotypes, **Het**: heterozygous genotypes, **HomA**: homozygous alternative genotypes, **Overall**: the overall genotype discordance, **Non-ref**: the percent discordance on heterozygous and homozygous alternative genotypes, and R^2 is the pearson correlation coefficient between the estimated and true genotypes.

It should be stressed that the HapMap3 SNPs are not representative of the full genome, since the SNPs are common and usually fall in easily sequencable parts of the genome. In order to have a more representative set of sites for comparison, we also compare the genotype discordance rates of the methods on the Axiom genotypes. The Axiom genotype dataset contains 5.4 million SNPs, including SNPs reported in the HapMap project, the 1000 Genomes project, and present in the public database of dbSNPs, and thus contains a set of SNPs with different MAFs[8]. The genotype discordance rates are presented in Table 5.4, which follows the same type of information as the previous table, but this time the comparison is made against the dataset from the Axiom chip. A first observation to make, comparing Tables 5.3 and 5.4, is that the genotype discordance rates have increased, which can be justified by the inclusion of SNPs that are more difficult to call than the SNPs present in HapMap3. MVNcall keeps the lead on calling the homozygous genotypes on the African and Admixed population groups, and has the second lowest overall genotype discordance rate (0.01% from the lowest overall discordance rate from all methods). MVNcall seems to have elavated discordance rates for the heterozygous genotypes with 1.48% compared to Thunder (1.24%) and SNPTools (1.35%) in the African population. BEAGLE has the highest genotype discordance rates for heterozygous genotypes. In terms overall discordance rate, (i) in the African population group, Thunder and MVNcall have the lowest followed by SNPTools and BEAGLE, (ii) in the Admixed population, the ranking is SNPTools, MVNcall and

[8]http://www.affymetrix.com/support/technical/sample_data/axiom_db/axiomdb_data.affx

Thunder come second, and lastly BEAGLE and for (iii) the Asian and European populations the ranking is SNPTools/Thunder, MVNcall and BEAGLE. In terms of the non-reference discordance rates, Thunder and SNPTools have the lowest discordance rates, followed by MVNcall and BEAGLE.

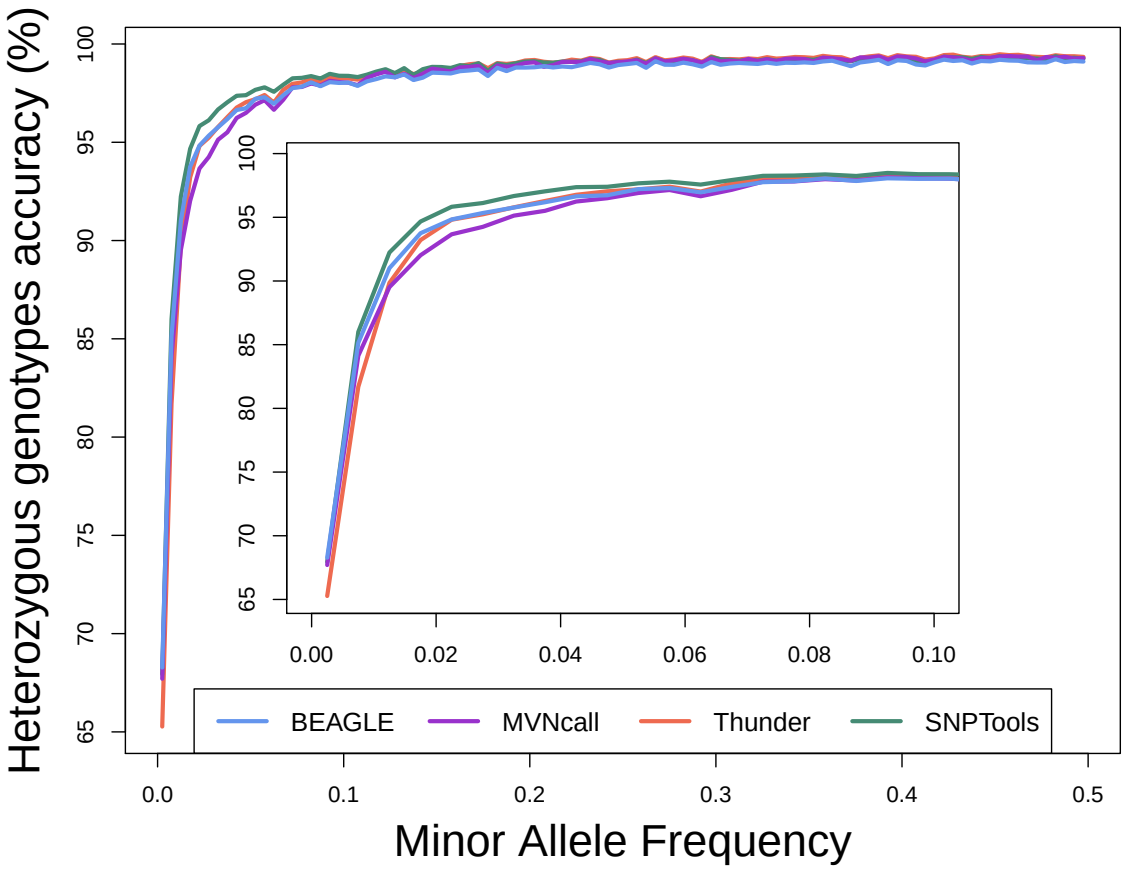

Figure 5.3: Plot of the percent heterozygous genotype accuracy versus the MAF for all populations together, for the four different methods applied in the dataset. The subplot zooms in the low minor allele frequency SNPs, and ranges from 0 to 0.1. The comparison is based on sites present in Axiom but not in OMNI.

MVNcall seems to be competitive on the African and Admixed population groups in terms of genotype accuracy. A possible explanation is that in those population the LD is lower than the case of the Asian and European popula-

tions, with more recombination events. In those situations, representing the LD structure in a more diffuse way seems to help, since it captures the local LD structure which might not follow the genetic assumptions made by the other models.

The main challenge for MVNcall is on the heterozygous genotypes, which has lower genotype accuracies than Thunder and SNPTools. For low coverage data used here, the genotype likelihoods for the heterozygous genotypes are less confident than the corresponding genotypes for homozygous genotypes. In those situation, more information is required from the LD structure.

Lastly, we examine how the genotype accuracy varies with the Minor Allele frequency. For rarer SNPs, using low coverage data, the genotype accuracy is low compare to the more common SNPs, an observation also observed in the low-coverage pilot data. We use the population MAF from the integrated release from Phase 1 of 1KGP, and by using the corresponding MAF for each population, we count the number of correct genotype called for different MAF bins. We divide the MAF spectrum in 50 intervals, and we plot the midpoint of the bin versus the percent genotype accuracy. For the homozygous genotypes we only observe minor differences between the different methods, so we only present the plot for the heterozygous genotypes in Fig.5.3. The SNPs employed in this figure are from Axiom SNPs not in OMNI. Thunder has the lowest accuracy for SNPs with MAF between 0 and 0.005 (0.65%), whereas SNPTools, MVNcall and BEAGLE reach an accuracy of 68%. SNPTools outperforms all the other methods for SNPs with low MAF, as we can see clearer from the subplot in Fig.5.3.

MVNcall outperforms Thunder for SNPs with MAF<0.01 but Thunder obtains better accuracies for 0.01<MAF<0.08. MVNcall has the lowest genotype accuracy for SNPs with MAF between 0.02 and 0.06. The difficulty for calling the genotypes on rare sites by MVNcall lies on the fact that for those sites the covariance between the candidate site and the flanking sites will be very low, and thus no information can be obtained from the LD structure.

5.3 Phasing accuracy comparisons

In addition to the genotype accuracy comparison, we want to get a sense of the phasing accuracy of the methods on this dataset. We do this by examining (i) the switch error rate on individuals that are trio phased in the HapMap project and are also present in the Phase 1 dataset, and (ii) by using the reference panels from the different methods to run an imputation analysis and compare the imputation accuracy using the different reference panels.

5.3.1 Phasing accuracy comparison based on HapMap3 haplotypes

A subset of 170 individuals that are present in the Phase 1 analysis are also trio phased in the HapMap3 project; 105 out of the 170 individuals have trio information on the scaffold used by MVNcall, since genotype information was available from the genotype data, 7 have duo information on the scaffold and the remaining 58 are unrelated on the scaffold. As it was mentioned before, the sequencing on phase 1 was performed only on the founder individuals, so all

individuals with sequencing information are unrelated. However, our method, benefits from the trio information coming from the phasing of the scaffold. The individuals in this analysis come from four different populations: MXL, YRI, CEU, and ASW.

For those individuals, we calculate the *switch error rate*, the percentage of possible switches in haplotype orientation, used to recover the correct phase in an individual. For each method, we only consider heterozygous sites that are correctly called by each method. On average, there is one heterozygous site per 10kb.

The results of this analysis are presented in Fig.5.4. For each of the four methods we plot the median and the error bars are extended to the 25% and 75% interquartile. The median switch error rates in ascending order are: MVNcall 1.0357, SNPTools 2.6817, Thunder 3.3375 and BEAGLE 3.9902. MVNcall has the highest phasing accuracy compared to the other methods. However, as it was discussed previously, our method uses a scaffold that is trio phased, where this information was not used by any of the other competing methods. To examine if there is any effect, we compare the switch error rates for two groups: the first group involves the set of individuals that the scaffold used trio or duo information to phase it, and the second group involves individuals that are phased on the scaffold without pedigree information.

The median switch error rate for individuals in the first group is 0.8938 and for the second group is 2.6522. As we can see, when the scaffold was phased using pedigree information, the resulting haplotypes were more accurate, and this

was reflected in our method when we subsequently tried to phase the genotypes on the candidate SNPs.

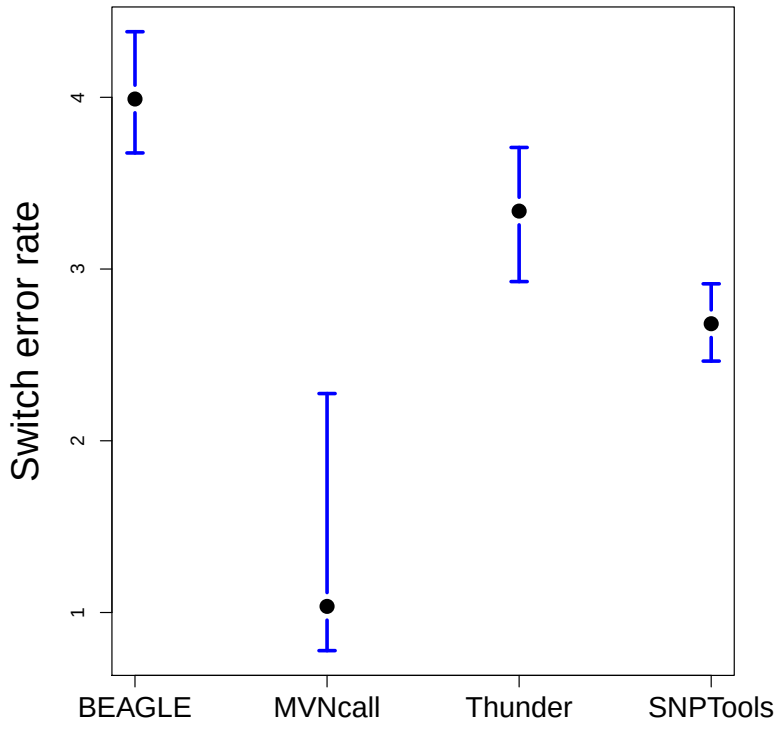

Figure 5.4: Phasing error rates for 170 individuals that are trio phased in the HapMap3 project and present in Phase 1 of 1KGP. The results are summarized with the media, shown as a dot on the plot, and the error bars are extended in the interquartile range.

Hence, when an individual is accurately phased using trio information, this will provide a more accurate scaffold for MVNcall, which in turn provides more accurate phase of the candidate sites. In addition, even for the subset of individuals that are phased without pedigree information on the scaffold, MVNcall (2.6443) still provides lower median switch error rates than Beagle (3.6020) and Thunder (3.0072) and very close to SNPTools (2.5710). Going one step back, we compared the genotype accuracy of the individuals that were pedigree phased

on the scaffold and compared them to the genotype accuracy of the individuals that were phased in the scaffold without pedigree information, and we do not observe a singificant difference in the corresponding genotype accuracies.

The comparison is based on a small sample of individuals and on a small number of SNPs. Also, the comparison of switch error rates does not account for the genotype accuracy between the different methods, since it only accounts the heterozygous sites that are correctly called by each method. We thus proceed to an additional comparison of the reference panels produced by the different methods in the following section. Nevertheless, it should be streesed here, that an accurately phased scaffold greatly helps MVNcall for phasing the genotypes on the candidate sites.

5.3.2 Imputation accuracy comparison

The haplotype sets produced by the HapMap project were widely used as reference panels in imputation analysis. With the completion of the Phase 1 of the 1KGP, a larger reference panel, both in the number of individuals and the number of sites, is produced by the different methods. In the previous section, we compared the phasing accuracy on a set of individuals on a small number of sites, but this does not give us information about the phasing accuracy on all sites. An alternative procedure to examine the quality of the haplotypes produced by the different methods, is to use the resulting haplotype sets as reference panels in imputation, and compare the imputation accuracy using the different reference panels. This is the main focus of this section.

Data processing

Complete Genomics Inc. (Complete Genomics, 2012) has deeply sequenced (on average 80x) 69 individuals. After excluding the individuals that are present in Phase 1, 16 individuals come from three of the population groups of the phase 1 of 1KGP. In particular 9 individuals have African[9] ancestry , 4 have European[10] ancestry and 3 have Admixed[11] ancestry. The genotypes from this source will serve as the "true" genotypes in the analysis. We then run an imputation analysis where genotypes are masked and then predicted. Here, we use different imputation scaffolds derived from different chips (with the number of SNPs in brackets): Affy500k (12,238), Affy6.0 (23,022), Illumina1M (25,939) and OMNI2.5M (47,317). For each imputation scaffold, we measure the imputation accuracy on the imputed genotypes on SNPs that are not present in the chip. We use IMPUTEv2[12] (Howie et al., 2009, 2011), to perform imputation on the CGI dataset using the reference panels from SNPTools, Thunder, MVNcall and BEAGLE in turn. The reference panels consisted of the same number of individuals, in this analysis this totals to 907 individuals. Even though Thunder and SNPTools have included in their reference panels indel calls, we excluded those calls, to have a head-to-head comparison with the other reference panels.

A standard measure of imputation accuracy is the Pearson R^2 correlation coefficient which measures the correlation between the true genotypes ,which can

[9]LWK (3), YRI (2), MKK, Maasai individuals from Kenya, (4)
[10]CEU (3), TSI (1)
[11]MXL (3)
[12]We split the chromosome 20 in 5Mbp chunks, and run IMPUTEv2 with the default parameters.

take {0,1,2} values, and the imputed dosages, i.e. defined as $\sum_{x=0}^{2} \mathbb{P}(G = x)x$ and can take values [0,2], at each SNP. It is common practice to measure mean R^2 versus MAF, since we want to compare the imputation accuracy especially on rare SNPs, which is the main challenge. However, the R^2 metric behaves poorly in a setting where the sample size is small. This is because, when the true genotypes of the individuals are the same for a given SNP, the R^2 is undefined. To avoid the problem of calculating per-SNP metrics for small sample sizes, instead, we divide the imputed SNPs according to MAF, and then for each bin we calculate an aggregate metric. We thus consider an alternative measure to calculate the imputation accuracy. In this analysis, instead of calculating the per-SNP R^2, we calculate the aggregate non-reference concordance, which can be computed with any sample size. This measure is in accordance with the imputation analysis carried out in the 1KGP (Bryan Howie, personal communication).

Interpretation of results

We present the results per population group, using the different reference panels, and the Affy500k chip as the imputation scaffold. For plots regarding the other imputation scaffolds see Appendix C.3. In Fig.5.5 we plot on the x-axis the allele frequency of the imputed SNPs in log scale and on the y-axis the non-reference genotype concordance. The x-axis is divided into bins, and the aggregate non-reference genotype concordance is calculated for all SNPs falling in that bin (y-axis), where on the x-axis we plot the midpoint of the interval. We consider smaller intervals for the rarer SNPs, to have a finer resolution of

the results for the rare SNPs. For the European population group, MVNcall has the highest non-reference concordance rate for SNPs with alternative allele frequency (AAF) <0.005% of 0.2558, followed by SNPTools with 0.2460, Thunder 0.2448 and BEAGLE with 0.2368. For SNPs with AAF between 0.005 and 0.05 SNPTools, Thunder, and MVNcall have similar non-reference concordance rates, and in different bins the ranking differs, without a clear distinction towards one method. BEAGLE comes last for all different AAF bins. The ranking between the different methods is the same when we use the Affy6.0 imputation scaffold, whereas when we use the Illumina1M and Omni2.5M imputation scaffold, MVNcall clusters with BEAGLE than with the other two methods, which remain indistinguishable (see Appendix C.3). For the individuals with admixed ancestry, bottom left plot in Fig. 5.5, we observe a similar pattern. SNPTools performs marginally better than MVNcall for rare SNPs, followed by Thunder and then by BEAGLE. For the individuals with African ancestry, MVNcall has a clear lead compared to the other methods. The non-reference concordance rate for SNPs with AAF <0.005 is 0.2418 by MVNcall, 0.1939 by SNPTools, 0.1881 by Thunder, and 0.1661 by BEAGLE. MVNcall outperforms the other three methods in this population group, with the most apparent difference on SNPs with AAF<0.05.

As it was discussed in the previous section, the phasing accuracy of the resulting haplotypes from MVNcall depends on whether the individuals have been phased as unrelated or by using pedigree information in the haplotype scaffold. In Table 5.5 we report the phasing information used by SHAPEIT for

phasing the individuals present in this analysis, focusing only on the individuals from the three population groups we are considering in this analysis. As we can see, for the European population group there is only one individual phased using pedigree information, and the rest of the individuals are phased as unrelated. Conversely, for the African population, 115 out of the 199 individuals with African ancestry are duo/trio phased in the haplotype scaffold. This could explain the patterns observed: MVNcall outperforms the other methods in the African population group where for the European group it achieves similar accuracies than the other methods. For the Admixed population, 148 out of the 149 individuals have been duo/trio phased.

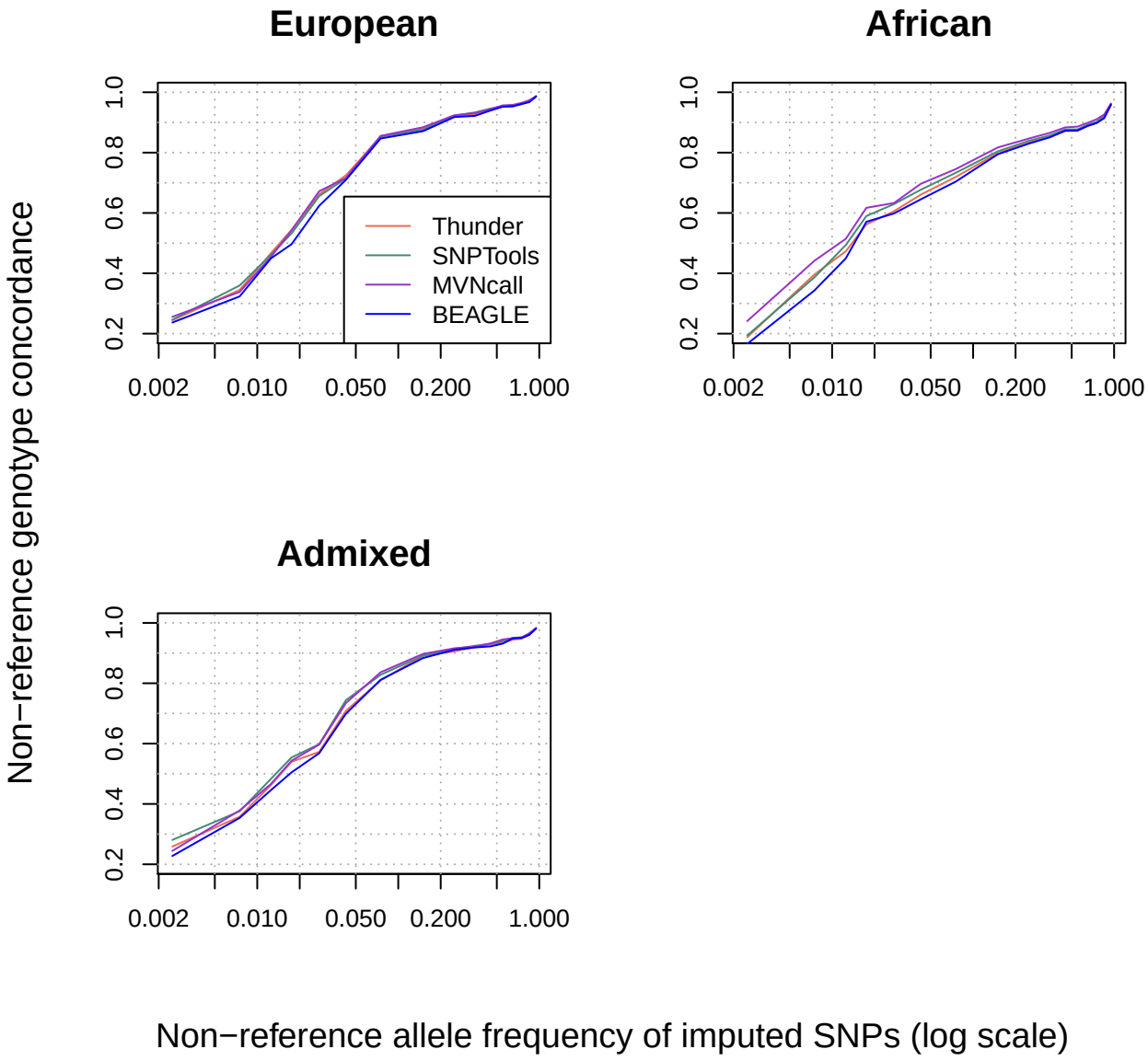

Figure 5.5: Allele frequency at imputed SNPs (log scale) versus the non-reference genotype concordance. Different lines correspond to different reference panels. The results are stratified by population group: European (top left), African (top right), and Admixed (bottom left). Imputation scaffold used is the Affy500k.

The difference between MVNcall and the other methods on the Admixed population is marginal, but this can be explained by the fact that the condition-

ing haplotypes chosen by IMPUTE for the admixed individuals might include individuals with European ancestry, that are phased as unrelated.

Population	Unrelated	Duo/Trio	Total
CEU	75	0	75
FIN	75	0	75
GBR	66	1	67
TSI	96	0	96
EUR	312	1	313
YRI	3	73	76
LWK	72	2	74
ASW	9	40	49
AFR	84	115	199
CLM	1	49	50
PUR	0	45	45
MXL	0	54	54
ADM	1	148	149

Table 5.5: Pedigree information for phasing the haplotype scaffold using SHAPEIT for the individuals in the reference panel used by MVNcall for imputation. The information is stratified by population group.

MVNcall provides competitive imputation accuracies compared to the other methods that applied the Phase 1 dataset. We should note here that given Thunder and SNPTools have also incorporated indel and structural variants for genotype calling, this might help the calling and phasing steps of those methods, even though the indel calls were excluded from the reference panels for the imputation analysis. Nevertheless, MVNcall, using the haplotype scaffold, provided close and sometimes better accuracies than those methods. Of course, the metric used to measure accuracy is an aggregate metric, and we should better try to understand the correlation of this metric with the Pearson R^2 metric used for measuring imputation accuracy. The use of the different reference pan-

els on larger datasets can provide a further comparison of the reference panels provided by the different methods.

5.4 Extension of the model to handle trio information

In Section 3.5, we extended our model to incorporate trio information for genotype calling. To do this, we used a deeply sequenced trio to examine the performance of MVNcall.

Data processing

In this section, we examine the performance of our method in terms of genotype and phasing accuracy on a deeply sequenced CEU trio (NA12891 (father), NA12892 (mother), NA12878 (offpring)) sequenced at ∼40x per individual. To tailor the coverage to the coverage of the rest of the individuals in the study, we subsample the reads so we have ∼4x coverage per individual. Using the resulting reads, we ran SNPTools on the candidate set from the 1KGP and we obtain genotype likelihoods for the trio. In addition, the same trio was genotyped by OMNI and phased by SHAPEIT, thus we also have the scaffold for the trio. We thus incorporate this trio into the European & Admixed run.

We compare our results to the most likely genotypes derived from the deeply sequenced information, taking into account only the highly confident genotypes. This ends up in a set of 562,633 sites. Also, using again the read information from the deep coverage data, the trios are phased and we consider this set as the true set of haplotypes.

Interpretation of results

Apart from examining the performance of the method when the trio information is used for genotype calling, we want to examine the benefit of adding pedigree information, either in the phasing of the scaffold or our model for genotype calling, and see the effect of this information in terms of the genotype and phasing accuracy on the trio under consideration. To do so, we study the four following scenarios (see Table 5.6): for the first two scenarios we only add the parents to the European & Admixed panel, and we use two different scaffolds: in scenario A we consider only the founder individuals for phasing, and thus no pedigree information is used for phasing the scaffold, whereas for scenario B we add the offpsrings in the set and thus trio information for phasing the scaffold. In Scenarios C and D we add the offspring in the analysis, and we vary the model applied for genotype calling: in Scenario C we treat the trio as unrelated in the model for genotype calling and in Scenario D we add the pedigree information in the model, and consider the trio information in the genotype calling.

Scenario	Individuals	Phasing Scaffold	Genotype calling model
A	parents	founders	unrelated
B	parents	founders & offsprings	unrelated
C	trio	founders & offsprings	unrelated
D	trio	founders & offsprings	related

Table 5.6: Scenarios examined for genotype calling of the trio; the "individuals" column corresponds to the individuals in the trio added in the analysis, in the "Phasing scaffold" column, the information incorporated in the phasing step of the scaffold, and in the column "Genotype calling model" the information considered in MVNcall for calling the individuals in the second column.

In Table 5.7, we compare the number of genotypes correctly called by each of

the four scenarios by individual. In scenarios A and B, only the parents of the trio are added in the analysis. The number of incorrectly called genotypes for the father is 635 in scenario A and 624 in scenario B, where for the mother is 568 in scenario A and 557 in scenario B. Therefore, phasing the scaffold using the offprings increases slightly the number of correctly called genotypes. In scenario C, by adding the child in the analysis, but still treating the trio as unrelated in the model, we see a decrease of 56 and 34 of the number of incorrect calls for the father and the mother respectively. Adding the child in the analysis, even though it is treated as unrelated, aids the calling of genotypes of the parents. This can be explained by the fact that the child has close haplotypes to the parental haplotypes and thus it was chosen in the set of closest haplotypes when inferring the genotypes of the parents. Lastly, in scenario D, by adding the trio information in the model we observe a considerable decrease in the number of incorrect genotypes compared to scenario C: 102 for the father, 97 for the mother, and 206 for the offpring. The offpring has the largest benefit of the trio information, which is expected, since the parents are unrelated, but the child is conditionally independent from the scaffold given the parents, and the genotype calls are made according to the calls of the parents. Looking further to the type of genotypes correctly called by scenario D and incorrectly by scenario C, a big proportion of the gain is observed on the heterozygous genotypes. In particular, 77% of the parental genotypes and 67% of the offspring genotypes that are wrongly called in scenario C and correctly called by scenario D are heterozygous genotypes. We thus see a great benefit from incorporating the trio information for calling

the genotypes of the trio.

Scenario	NA12891 (F)	NA12892 (M)	NA12878 (C)
A	635	568	N/A
B	624	557	N/A
C	568	523	529
D	**466**	**426**	**323**

Table 5.7: Number of incorrectly called genotypes for each of the individual in the trio for each of the four scenarios.

We then move on to compare the phasing accuracy of the trio for the four different scenarios. To do so, we calculate the switch error rates on the heterozygous genotypes correctly called in each scenario for each individual, taking as a true set of haplotypes the haplotypes derived from the high coverage reads. In Table 5.8 we report the switch error rates by individual by scenario. Comparing the switch error rates between scenario A and scenario B, we see a small improvement in the switch error rate by using a scaffold that is phased using pedigree information, the switch error of the father decreases from 1.4502 to 1.3685 and for the mother from 1.5827 to 1.5176. We should note here, that out of the 967 European & Admixed individuals present in the genotype data used for phasing the scaffold, 439 are unrelated, and, 98 out of the 104 CEU individuals in the sample are unrelated, which also explains why we do not observe a great difference of using the two different scaffolds. However, by just adding the child in the analysis, we see a greater improvement, for the same scaffold. The switch error rate of the father drops from 1.3685 to 1.2605 and for the mother from 1.5176 to 1.2354. Again, we see a big benefit by adding the trio relationship in the model. The switch error rate of the parents decreases $\sim$4 times and the

switch error of the offpring decreases by ∼7-fold. The switch error rates for the father in Scenario D is 0.2734, for the mother 0.2959 and for the offpring 0.1248.

Scenario	NA12891 (F)	NA12892 (M)	NA12878 (C)
A	379/26134 (1.4502%)	394/24894 (1.5827%)	N/A
B	359/26233 (1.3685%)	378/24907 (1.5176%)	N/A
C	331/26260 (1.2605%)	308/24932 (1.2354%)	250/27113 (0.9221%)
D	**72/26339 (0.2734%)**	**74/25011 (0.2959%)**	**34/27251 (0.1248%)**

Table 5.8: Switch error comparison between the four different scenarios per individual in the on the heterozygous genotypes correctly called.

To better understand where the switches occur, we also calculate the flip errors. Flip errors are defined as the number of two consecutive switches. The higher the number of flip errors, the longer the stretches of haplotypes correctly phased. As we notice from Table 5.9, the number of flip errors consists of the majority of the switch errors[13]. For example, for scenario D 70 out of the 72 switch errors for the father are flip errors, whereas for the mother and the child all switch errors are flip errors.

Scenario	NA12891 (F)	NA12892 (M)	NA12878 (C)
A	175	187	N/A
B	165	169	N/A
C	158	137	117
D	35	37	17

Table 5.9: Flip error comparison between the four different scenarios per individual in the trio on the heterozygous genotypes correctly called.

Lastly, we count the number of sites where the trio genotypes are mendelian inconsistent. For scenario C there are 302 sites where the genotypes are mendelian inconsistent, for scenario D the number falls to 4. Adding the trio information

[13]# of switches = 2 x # flip-errors + # non-flip errors.

in the genotype calling step is an important extension of the model, since the incorporation of this information greatly improves both the genotype and phasing accuracy of the individuals. The resulting estimated haplotypes of trios from a a model that jointly takes into account the LD, genotype likelihoods and trio information provides highly accurate haplotypes.

5.5 Varying the SNP density in the scaffold

The Phase 1 of the 1KGP genotyped all individuals in the study using the latest genotype chips with 2.5 million SNPs across the genome. This set of SNPs was used to create the scaffold of the MVNcall for the analysis of the 1KGP. However, different projects have genotyped the samples in the past using a different chip, where previous chips have a less dense set of SNPs. Here, we want to examine what is the effect of the scaffold density on the genotype accuracy results of MVNcall.

Data processing

We use the OMNI set of sites to subsample different sets of SNPs, where the allele frequency distributions mimic those chips; i.e. we favour more the sampling of common SNPs than rare SNPs. Given that chromosome 20 represents $\sim$2% of the genome, we sample 10,000, 20,000, 30,000 and 40,000 SNPs for chip densities 500k, 1M, 1.5M and 2M (We also present 2.5M which is the OMNI chip used in the 1KGP which has $\sim$ 54,000 SNPs). We run the region chr20:40,000,000-50,000,000, and compare the genotype accuracy on HapMap3 SNPs not in OMNI.

We use the same parameter settings for MVNcall (iteration(100), burn-in(10), λ=0.06, number of flanking sites (50)).

Interpretation of results

We present the overall genotype discordance rate on SNPs that are present in HapMap3 but not in OMNI in Fig. 5.6, stratifying the results by the three main population groups.

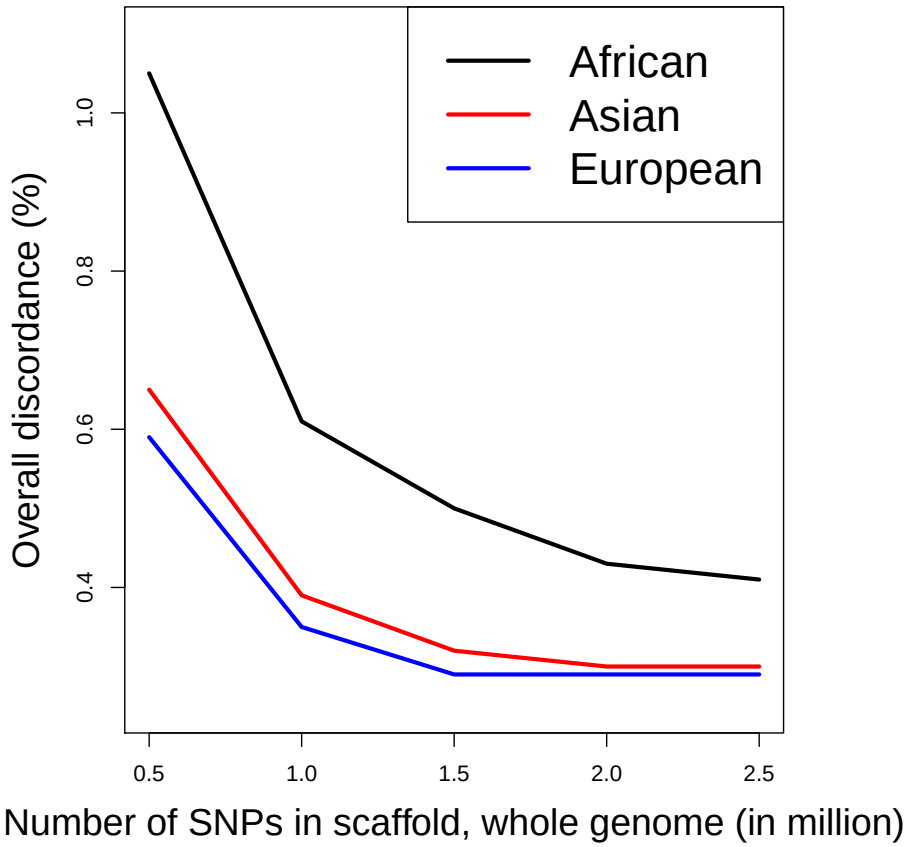

Figure 5.6: Overall percent genotype discordance for different scaffold SNP density. Results are presented by population groups: African, European and Asian.

As we can see, by having a more dense scaffold, the genotype accuracy im-

proves, as it was expected. The overall discordance of the African population is higher than the overall discordance of the Asian and European individuals, and the rate of change of the overall discordance is more apparent on the African discordance rates than the Asian and European. The overall genotype discordance is decreasing from 1.05% for a 500k scaffold density to 0.41% for a 2.5M scaffold density for the African population. Increasing the scaffold density from 2M to 2.5M, there is an improvement of 0.02% on the overall genotype discordance. On the other hand, for the Asian and the European populations, we see an improvement as we increase the scaffold density from 500k to 1.5M, and then it plateaus for 1.5M, 2M and 2.5M scaffold densities. This behaviour can be explained by the difference in the LD in the different population groups: the African population has low LD and high recombination rates, thus by having a more dense scaffold, MVNcall gets more information about the local structure. On the other hand, the Asian and European populations have higher LD and thus even with a less dense set of SNPs, the LD structure can be captured. Although we have not carried out the experiment for this dataset, if the scaffold has a lower density than the one used in this analysis, the option of adding the shrinkage on the covariance of the matrices, as it was applied in the analysis of the low-coverage pilot dataset, can further improve the genotype accuracy.

5.6 Simulation study varying the reference panel size

Looking forward to the future projects, as the sequencing costs decrease, the number of the individuals studied will increase. It is thus of interest to study how the method will behave as the reference panel size increases. To do so, we simulate haplotypes for 10,000 individuals and the accompanied sequencing reads, and we investigate the genotype calling performance as we increase the reference panel size. In particular, we study five different reference panel sizes (2000, 4000, 6000, 8000, 10000) and we measure the overall genotype discordance rate for the different reference panels on a set of 2000 randomly chosen SNPs.

Data processing

We simulated 20000 haplotypes using MaCs (Chen et al., 2009) with the same population parameters provided in Chen et al. (2009). We then randomly paired haplotypes to create 10000 individuals. We mapped the 5Mbp chunk simulated to chromosome 20. We then use ART (Huang et al., 2012) to simulate reads using the parameters used for the simulation of reads in the 1KGP[14]. Given that we compare MVNcall for different reference panel sizes, we make the assumption that the sequenced reads map perfectly on the reference genome and thus omit the mapping step. Hence, after having obtained the simulated reads, we assume perfect mapping and we run SNPTOOLS to obtain the genotype likelihoods on the segregating sites. We then mimic the procedure followed in Le and Durbin (2011) and created the haplotype scaffold. We randomly choose 10 haplotypes

[14]We used the parameters for simulating Illumina 51-bp paired-end reads

and identify SNPs present in those haplotypes, and we select the same number of SNPs as those present in the OMNI2.5 chip. We do so by taking the nearest site seen twice in the 10 simulated haplotypes to each true OMNI2.5 site. Finally, we randomly choose individual haplotypes to form the different sets of reference panels with different sizes. We randomly choose 2000 SNPs from the 5Mbp region that are not present in the scaffold, and we run MVNcall using the parameter settings used in the Phase 1 of the 1KGP.

Interpretation of results

For each of the reference panel sizes we calculate the overall genotype discordance rate, presented in Fig.5.7. The first observation to make, is that as we increase the number of reference panel size we can see a decrease of the overall discordance rate. For the reference panel size of 2000, the overall discordance rate is 0.27479%. When we increase the reference panel size to 4000, the overall genotype discordance rate drops to 0.25740%. As we continue increasing the reference panel size we observe a further decrease of the overall discordance rate, reaching 0.24961% when the reference panel size is 10000.

We thus notice, that even though we summarize the data in a crude way, we still benefit from the increased reference panel size. A feature of the method that mainly takes advantage of the increased reference panel size is the selection of the closest haplotypes. By choosing the closest haplotypes, we zoom in the tree on the branches that are closer to the haplotype of interest. It should however be noted that as the reference panel size increases the number of inverse ma-

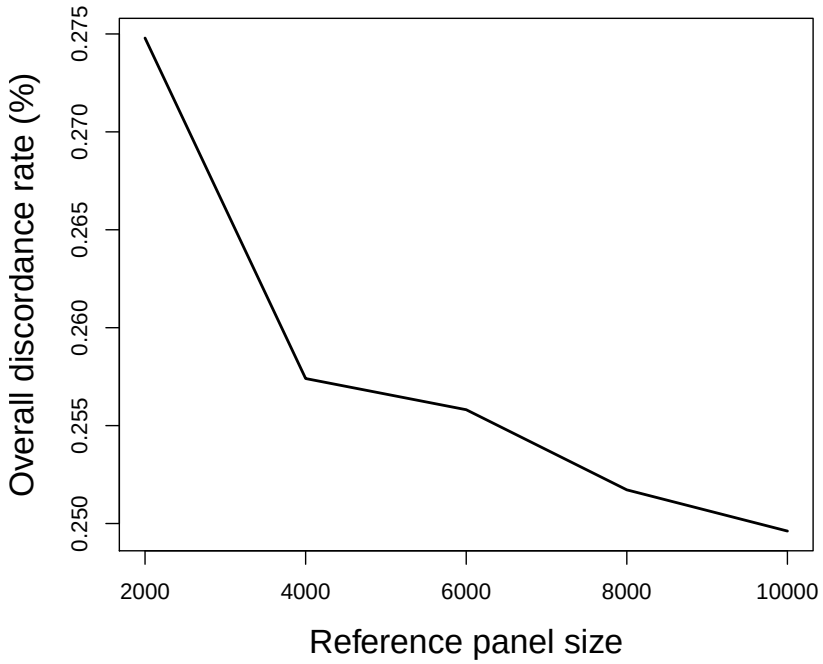

Figure 5.7: Overall percent genotype discordance for different reference panel sizes.

trices calculations increase, which is the heaviest computational burden of the method, and thus the computational time is high. However, it is of great interest to realise that MVNcall can be potentially used for large studies.

5.7 Summary

In this chapter we applied MVNcall on the Phase 1 dataset of the 1KGP on chromosome 20. We first investigated the different tuning parameters and extensions of the model for large sample sizes. We then procceeded and compared the

genotype calling accuracy of MVNcall with Thunder, SNPTools, and BEAGLE. MVNcall provides competing genotype accuracies, with genotype accuracies better than BEAGLE for all population groups. MVNcall has lower genotype accuracies than Thunder and SNPTools, mainly because of higher heterozygous genotype discordance. We then compared the methods in terms of switch error rates. MVNcall has the lowest median switch error in this dataset. Individuals that have been trio phased on the scaffold have lower switch error rates than the individuals that are phased as unrelated. We then carried out an imputation analysis using the different reference panels crated by the different methods. MVNcall has a clear advantage on the African population, which is the result of the higher number of trio phased African individuals on the scaffold compared to the other populations. The trio information was also incorporated for genotype calling with MVNcall, and the results showed an improvement both in terms of genotype and phasing accuracy.

Chapter 6

Summary and Discussion

6.1 Summary

In the previous two chapters we have used two datasets from the 1KGP to test our two methods developed for SNP and genotype calling from low-coverage next-generation sequencing reads. We investigated specific features and tuning parameters on each of the two proposed methods and then we compared the genotype and phasing accuracy of those methods with other competing methods that were applied on the same set of data.

In chapter 4, we began by assessing TreeCall on the low-coverage pilot data. We concluded that feeding the method with either the genotype likelihoods of all possible genotypes or with the three genotype likelihoods that correspond to the most probable alleles, outputted from a non-LD method, gives very similar BFs and the same resulting genotypes which assures that the method is able to correctly pick the correct alleles. In the situation where the alleles are known, the method can take as input only the three likelihoods and the time is decreased

by 6-fold. We then tested different filters on the candidate set from TreeCall and chose a set of filters that account for coverage and mapping artifacts due to indels. The choice of filters was mostly based on the resulting Ts/Tv ratio of the filtered set and the number of retained known sites. Lastly, we compared the two methods proposed for genotype calling from the TreeCall method. The method that chooses the configuration of genotypes resulting from the highest likelihood, for a given reference allele, mutation allele and branch where the mutation occurs, is chosen.

We then moved on assessing MVNcall on the low-coverage pilot data for different sets of tuning parameters. We tested convergence of the method by outputting intermediate results from a single run and comparing the results. We then investigated the optimal values for the diagonal penalty λ as well as the number of flanking sites. Applying a shrinkage both on the diagonals and the off-diagonals entries of the variance-covariance matrix was chosen in this setting, which was reducing noise and producing more accurate genotype calls. Filtering was also applied to create a refined candidate set.

The comparison of our methods with other competing methods, (QCALL, Thunder, and BEAGLE) was based on a 10Mbp region on chromosome 20 on 55 unrelated CEU individuals. We compared both the SNP calling sets as well as the genotype calls made by the different methods, taking as "true" genotypes from HapMap2. The SNP calling sets produced by the different methods had a high intersection and the SNP calls of both TreeCall and MVNcall had a high intersection with the low-coverage pilot the calls made by the other methods. The

percentage of known SNPs discovered was competitive with the other methods and the resulting Ts/Tv ratio of the filtered sets was close to the expected. In terms of genotype accuracy comparisons, MVNcall provides the most accurate genotype calls, followed by QCALL, Thunder, BEAGLE, and TreeCall. We observe that using a haplotype scaffold, as it is the case of MVNcall, QCALL is beneficial in the situation where the sample size is small and we only have low-coverage sequencing information. TreeCall does not perform as well as the other LD-based methods. Possible reasons for this is the parameters used for building the genealogical trees, and the number of the trees built. When we update the haplotype scaffold with 2-3 way haplotypes (see Appendix B.1), we observe an increase in genotype accuracy especially in the TreeCall method, which explains the reason why we believe that a more dense set of trees might increase the genotype accuracy in the initial analysis.

In chapter 5, we analysed a larger sample from individuals from diverse population backgrounds. The phase 1 of the 1KGP involved the sequencing of 1,092 individuals from 14 population groups, stratified in four major population groups, sequenced at $\sim$4x. TreeCall and QCALL were not applied on this dataset since the larger sample made the methods used for constructing genealogical trees unattractive in terms of computational time. We thus applied only MVNcall in this dataset and compared it to other LD-based methods applied to this set of individuals, Thunder, SNPTools, and BEAGLE. Following the advances in the pipeline of the 1KGP, non-LD based methods have been developed after the completion of the pilot project, that construct a candidate set that

is then used by all LD-based methods. That means, that the main focus of the analysis that followed was on the genotype calling and phasing. The comparison presented was based on the chromosome 20 on a basis of ∼800,000 SNPs.

We first compared the performance of the MVNcall under different extensions suggested in the case of large samples, based on a 10Mbp region on chromosome 20. We found that the perfect match distance was a better haplotype distance measure than the Hamming distance measure in our model, with an apparent advantage on the African populations. Perfect match distance takes into account the position of interest and captures the local LD structure around it, which is preferred to hamming distance that counts the mismatches in a given window. Also, we observe that seperating the populations in population groups and running each one separately, together with the admixed individuals, we were getting higher genotype accuracies than analysing all the individuals at once. Thus, by separating the groups in individuals, we restrict the method from choosing closest haplotypes from a distant population. This might have an effect on rare SNPs present in one population, but common in another population, but the overall genotype accuracy was better in the case of seperating the populations, which was then adopted in the whole chromosome analysis. Also, in the analysis for choosing the number of conditioning haplotypes k, we notice that for homozygous genotypes the genotype discordance was decreasing as we were increasing the value of k, but this was not the situation for the heterozygous genotypes, where the discordance was reaching a minimum for k =100 and then it was increasing for k >100. By running the method twice, one by choos-

ing the 100 closest haplotypes from the reference panel and another run by using all the haplotypes in the reference panel, we manage to keep the homozygous genotype discordance low and the heterozygous genotype discordance constant at the minimum genotype discordance achieved by MVNcall.

The comparison in terms of genotype calling was based on two sets of genotype data, one from the HapMap project and the other from the Axiom Affymetrix chip, were a subset of the individuals in the analysis have been genotyped. MVNcall outperforms BEAGLE in all population settings for both dataset comparisons, with an apparent advantage on African and Admixed populations. The overall genotype discordance of MVNcall is 0.05% lower than the corresponding ones from BEAGLE for the African and Admixed population and 0.02% for the European population. For the Asian population, both BEAGLE and MVNcall reach the same overall genotype accuracy. MVNcall provides lower non-reference discordance rates than BEAGLE. Given that both models do not apply a genetic-based model, our method seems to capture better the LD structure. Comparing MVNcall with Thunder and SNPTools, MVNcall provides the lowest discordance rates for homozygous genotypes for the same population groups. MVNcall has a disadvantage on calling heterozygous genotypes compared to those methods, where they apply versions of the Li and Stephens model. Given that the genotype likelihoods for heterozygous genotypes is less confident than the corresponding for homozygous genotypes, more information is taken from the LD structure; a better representation of the LD than the diffusive way we use in our model, seems to be beneficial in this case.

To investigate the quality of the haplotypes produced by the different methods, we compared the estimated haplotypes with haplotypes that are trio phased in HapMap3, when available. For the set of individuals studied, MVNcall provides the lowest median switch error rate, followed by SNPTools, Thunder, and BEAGLE. Investigating further the clear advantage of our method in the phasing step, we notice that individuals that have been phased in the scaffold using pedigree (duo/trio) information, has a considerably lower switch error rates than individuals that were phased without pedigree information in the scaffold. Therefore, the highly accurately phased sites in the scaffold aid the phasing of the rest of the candidate sites subsequently. Furthermore, to test the accuracy of the phasing on the whole set of the estimated haplotypes, we use them as reference panels in an imputation analysis, and we compare the non-reference concordance across the different reference panels, i.e. the different methods. MVNcall has a consistent advantage in the case of individuals from African populations in the range of rare SNPs, followed by SNPTools/Thunder and lastly BEAGLE. For individuals from European and Admixed ancestry the imputation accuracy differences were marginal between SNPTools, Thunder, and MVNcall.

The extension of the model to incorporate trio information for genotype calling was examined using a deeply sequenced CEU trio. We used the information to call the most likely genotypes and construct the haplotypes of the trio that was used as a reference. By subsampling, we tailored the trio information to 4x coverage and added them in the analysis. The advantage of adding the trio information in the model is highly beneficial both in terms of genotype calling

and phasing. The number of incorrect genotype calls decreases considerebly for the parents, but mostly for the child, and the majority of the genotypes called correct with the trio information are heterozygous sites. This is important in our model, since we observe that by adding extra information on the model, we increase considerably the concordance rate for heterozygous sites, which are the main challenge of our algorithm. The phasing accuracy of the estimated haplotypes is very high, reaching a switch error of 0.12% for the child and 0.27-0.30% for the parents.

Lastly, looking ahead to other project designs that may have not genotyped their samples in chips other than the OMNI2.5M, we studied the genotype accuracy performance of the method for scaffolds with different SNP density. As it was expected, for the African population group, the overall genotype discordance keeps falling as the SNP density increases in the scaffold, whereas for the Asian and European population groups, we see a plateau after a scaffold with 30,000 SNPs on chromosome 20, i.e. a chip with 1.5M SNP density. This assures that the method can be confidently applied on other studies with chip data in less dense chips.

6.2 Discussion

In the first three chapters we have introduced the SNP and genotype calling problem in the face of low-coverage sequencing read information in a sample of individuals and we presented methods that have been developed for SNP and

genotype calling. All methods have a common feature: they employ LD information. This triggered the development of two methods that are LD-based for this problem, TreeCall and MVNcall. Both of our methods use the idea of a scaffold. This prerequisites that all the individuals in the study that have been sequenced, have also been genotyped on a subset of sites. Then, using those fixed genotypes, we perform a phasing step in order to phase the genotypes from the chip data, and we use this haplotype scaffold in the models. Another feature of the methods developed, is that they analyse one position at a time, in contrast to the other LD-based methods presented that analyse a window of SNPs simultaneously[1]. This has an apparent advantage in terms of convergence compared to methods that analyse a window of SNPs simultaneously. When analyzing a set of SNPs jointly, if some genotypes are uncertain, this will affect the speed of convergence in nearby SNPs. Also, the MVNcall converges irrespective to the starting haplotypes at the position of interest, something that does not apply to Thunder, since it requires to run BEAGLE first to initialize the haplotypes.

Considering the spectrum of models and how they capture the LD structure, the tree-based models directly construct genealogical trees using the coalescent model and can incorporate events including mutation, recombination, migration, and selection. Those methods tend to be computationally intensive and in the application for genotype calling they suffer from the fact that some branches will always be wrong, since the state space of possible trees is large, and these mistakes will be directly related to the genotype calls. TreeCall and QCALL are

[1] Apart from QCALL.

examples of tree-based methods for genotype calling.

A widely used approximation of the coalescent, is that of the model of Li and Stephens (Thunder and SNPTools), which carries a lot of the features of the coalescent model, but it is more computationally feasible. The model considers a new haplotype as a mosaic from the previous seen haplotypes. Mutation events are added in the form of miscopying and the recombination events are modeled as switching from copying from one haplotype to another. The Li and Stephens model has been successfully applied to a variety of problems including phasing and imputation. In the case of genotype calling from sequencing, the model enjoys a variety of the benefits met in the phasing and imputation problems and the models that used this approximation reached the highest genotype concordance rates. However, a good starting point is required for the model to converge, given that some positions will not have confident genotype likelihoods. Analysing a window of SNPs at once, bears the danger that in the case that in the scenario where some genotype likelihoods are flat, will affect the convergence of all positions in the window of interest. In addition, in the situation where false positives arise due to mapping misalignments, this will also affet the calling of nearby sites in the same window.

A different representation of the LD structure, is in the form of graphs, and in the case of BEAGLE, Directed Acyclic Graphs (DAG). This model outperforms all the aforementioned models in terms of speeds, since it is linear in the number of individuals. The haplotypes are represented in the form of graphs, and haplotypes that are similar are grouped and represent a cluster of haplotypes.

A benefit of this approximation, is that directionality is preserved, and building the tree from right to left or vice versa does not lead to the same graph. However, by approximating the recombination events by clustering, the genotpe discordance when applied to genotype calling using sequencing data, is lower than the models that use the Li and Stephens approximation.

Moving to a further approximation of the representation of the coalescent, we end up in the MVNcall model, which assumes that a newly seen haplotype comes from a multivariate normal distribution, with the parameters estimated from the reference panel. The model does not account for the relationship of a window of SNPs simultaneously, in contrast, only the pairwise covariance between the SNPs in the window studied are taken into account, and the allele frequency of the sites are summarized by the mean. The multivariate normal approximation is not directional, since the permutation of the SNPs on the reference panel will lead to the same estimated density. This disadvantage is corrected when shrinkage is applied on the covariance that accounts the probability of recombination between two SNPs using the genetic map. The model does not require an assumption about the mutation events. Analysing one position at a time, although ignoring information of nearby SNPs with genotype likelihoods not present in the scaffold, it avoids convergence issues and it is not affected by false positives due to misalignments errors.

Our first SNP and genotype calling method, TreeCall, provided a comparative call set but the genotype accuracy was lower than all the other LD-based methods on the low-coverage pilot dataset. We believe that a tuning of the pa-

rameters used might have improved the genotype accuracy of the method. In particular, we have only constructed genealogical trees with 20 flanking sites at either side and we did not explore how the performance varies with different number of flanking sites. Also, a more dense set of genealogical trees could be useful. This is because, the genealogical trees were constructed for a particular position and the same set of trees were used for all flanking sites. This would work well for nearby sites, but as we get further away from the site, the probability of a recombination event increases, and when a recombination event occurs, the tree does not represent the genealogical history of that site anymore. Given that the genotype calling models between QCALL and TreeCall follow the same idea, we believe that the main difference between the two methods lies on the marginal genealogical trees used. QCALL uses MARGARITA to build ARGs and then it samples marginal genealogical trees for each candidate site. On the other hand, in TreeCall we build marginal genealogical trees for a site that is in the middle of two known sites (from the haplotype scaffold), and use the trees built for that site for all sites falling in the interval between the two known sites. We therefore make an implicit assumption that that there is no recombination event in that interval, which might not be the case. Furthermore, using the inferred genotypes from the most probable configuration, although it has a higher sensitivity on calling rare genotypes, it also carries the mistakes of the estimated genealogical tree, since the estimated genealogical tree is not the true genealogical tree. The method was only used on the low-coverage pilot data since the computational time, especially of the method for constructing the genealogical

trees, was restrictive.

On the other hand, MVNcall was successfully applied on the low-coverage pilot data and gave the highest genotype accuracy across all methods. MVNcall confirmed the benefit of using a known haplotype set, the same idea as QCALL which had the highest genotype accuracy in the low-coverage pilot, and how representing the LD structure in a more diffuse way, can be beneficial in this type of problem. The model was able to be applied on a larger dataset, and we applied our method on the phase 1 dataset of 1KGP. The method provided more accurare genotype calls than BEAGLE, and provided competitive genotype accuracies for the African and Admixed population groups, compared to Thunder and SNPTools. The main challenge in the case of larger sample sizes, is the choice of the closest haplotypes. We applied a haplotype distance metric that only accounts for recombination events in a heuristic way, but a haplotype distance that accounts for both recombination and mutation events would be preferred. Nevertheless, there is a trade-off between the amount of resources spent on the calculation of the haplotype distance and computational time, so for this analysis we chose a fast haplotype distance measure. Also, in this study we chose to seperate the populations to population groups and run them separately. This implicitly means we restricted the choice of some haplotypes from distant populations in the step of choosing the closest haplotype step.

MVNcall has a lower genotype accuracy for heterozygous genotypes than SNPTools and Thunder. We believe that this observation is due to the genotype likelihoods for the heterozygous genotypes, that are usually less confident than

the corresponding ones for homozygous genotypes, since we need to observe both alleles to have a higher genotype likelihood for heterozygotes. Given the lower confidence on the genotype likelihoods, the heterozygous genotypes are more sensitive to the conditioning haplotype set. A way around this was to average over two values of conditioning haplotypes, a small value of k that takes into account only the close haplotypes and a set with all the haplotypes that accounts for the overall information in the haplotype set. It is however important to note that the tuning parameters used for all population groups are the same which means that the model is independent of the population structure of the individuals in the study.

We should note here, that MVNcall requires a set of haplotypes for a scaffold, which requires some external information apart from the sequencing information in order to be applied, which restricts the application of the model when only sequencing information is available. Firstly, most of the studies have already genotyped the samples in the study in the past, and are now moving on sequencing the same set of individuals. We have shown that even with a less dense set of sites, depending to the population studied, we can use the genotype data for creating a haplotype scaffold, and use MVNcall with a small decrease in the genotype accuracy. Otherwise, depending on the study design different ways can be adopted to create a haplotype scaffold: we could go from a first pass with a non-LD method to detect sites where all individual genotypes are confident, and use this set to create a scaffold. We should of course make sure that the allele frequency spectrum includes common SNPs.

MVNcall is slower than BEAGLE, which is the only method we run in this study. Since we divide the population groups and run them seperately whereas BEAGLE runs all population groups jointly, we do not have a head-to-head comparison of timings. The rest of the results are obtained from the 1KGP website and we do not have an accurate estimate of the running times. We did not include running time comparisons at this stage because the algorithm is in ongoing development. However, the method is highly parallelizable, since we analyse one position at a time, which means that the candidate set can be divided to small chunks and applied in parallel. This is not the case for the other methods that analyse a window of SNPs at a time, which puts a restriction on the size of the chunk to have good results. Also, for the methods analysing a window of SNPs simultaneously, we need to ensure that the chunks have some overlapping regions to avoid edge effects, a situation that is not met in our method.

Several questions remain open in this work. We need to better understand and investigate how to choose the set of conditioning haplotypes. Given that we use a fixed scaffold, and we thus only calculate the haplotype distance once, we should investigate other ways for choosing the conditioning haplotypes. We have seen a great improvement on the genotype accuracies when choosing different haplotype distance measures, and we believe that this can give a boost to the method. Also, by ranking the conditioning haplotypes with a better haplotype distance measure, we may be able to find one optimal value for the set of conditioning haplotypes, and thus decrease the computational time by a factor of two. Also, it is of interest to see how the method performs in datasets with

different depth of coverage, since methods with both lower and higher coverage than the 1KGP are emerging.

We also view some further extensions of the MVNcall. The model can be potentially used for indel calling, although we have not yet tested the performance. Also, the method can be extended to multi-allelic SNPs and indels. Looking forward after the completion of the 1KGP and the creation of a new reference panel, MVNcall can adapt this reference panel in the method, and include this set of haplotypes as a reference panel.

6.3 Conclusions and future scope

In this thesis, we presented two LD-based methods for SNP and genotype calling, focusing more on the second method, MVNcall. We introduced the idea of using a haplotype scaffold in the genotype calling problem that arose with the advent of sequencing data, and presented the advantages of using a scaffold for this type of problems. We tested the method in terms of genotype calling and phasing accuracy and realised the advantage of using a scaffold especially in terms of phasing accuracy.

The exploration of different approaches to this problem is important and it is an ongoing research, since we have only recently started using next-generation sequencing data. Advances in all the pipeline for analysing sequencing data, from the sequencing technologies, to alignment and mapping, and recalibration, are going to be reflected on the accuracy of the genotype calling methods as

well. It is of great interest to continue investigating the different features of sequencing data and develop methods to capture most of the information. In our case, we tried to incorporate previous algorithms and knowledge we have gained with genotype data, with the new sequencing data. Lots of challenges lie ahead, but we look forward to meet the challenges and try to understand the hidden information in the data.

Appendices

Appendix A

Non-LD methods for SNP discovery

In this appendix chapter, we present different methods that have been developed for SNP discovery. Those models incorporate information from the sequencing reads, usually in the form of genotype likelihoods. The non-LD methods employ different ways to calculate the probability the site is a SNP, assuming a neutrally evolving population under the coalescent. The main challenge in this calculation is the huge space of possible genotype configurations. Different algorithms, approach this problem differently as we will see below.

A.1 Samtools

Samtools is a library and software package that reads files in SAM/BAM format (Li et al., 2009a) including alignment viewing, calculation of genotype likelihoods, as well as SNP and short indel calling. Here we only focus on the SNP calling model of Samtools. Samtools estimates the allele frequency spectrum (AFS) aiming to reduce false positives. Let us assume in a given set, there is

information for N sites and the study sample consists of n individuals. Let $R = \{\vec{R_1}, ..., \vec{R_N}\}$ be the sequencing information for the sites. The same notion applies for the true genotypes $G = \{\vec{G_1}, ..., \vec{G_N}\}$ and the number of reference alleles at each site $X = \{X_1, ...X_N\}$. Define $\Phi = \{\phi_0, ..., \phi_{2n}\}$ as the allele frequency spectrum (AFS), where ϕ_i is the proportion of sites with i alternative alleles, and $\sum_k \phi_k = 1$. To estimate the AFS for the N sites, it needs to find Φ that maximizes $\mathbb{P}(R \mid \Phi)$, using an EM algorithm. Hence, at step t, the E-step is given by

$$Q(\Phi \mid \Phi_t) = \sum_x \mathbb{P}(X = x \mid R, \Phi_t) log \mathbb{P}(R, X = x \mid \Phi) \tag{A.1}$$

and the M-step is

$$\phi_k^{(t+1)} = \frac{1}{N} \sum_\alpha \mathbb{P}(X_\alpha = k \mid \vec{R_\alpha}, \Phi_t) \tag{A.2}$$

where $\mathbb{P}(X_\alpha = k \mid \vec{R_\alpha}, \Phi_t)$

After calculating Φ, the probability of a site α having $X = x$ alternative alleles given the AFS and the sequence data is

$$\mathbb{P}(X = x \mid \vec{R_\alpha}, \Phi) = \frac{\phi_k \mathbb{P}(\vec{R_\alpha} \mid X = k)}{\sum_l \phi_l \mathbb{P}(\vec{R_\alpha} \mid X = l)} \tag{A.3}$$

A site is polymorphic if $\mathbb{P}(X = 0 \mid \vec{R}, \Phi)$ exceeds a threshold. The method has been extended to take into account base quality alignment to reduce the FDR of SNPs caused by misalignments of reads to regions that include indels (Li, 2011b).

A.2 QCALL non-LD dynamic programming

This non-LD method is a pre-processing step of QCALL (presented below) (Le and Durbin, 2011). Given a sample of size n, the genotypes of the individuals in the study at this site is $\vec{G} = (G_1, ..., G_n)$, where G_i is the genotype of individual i. A position is monomorphic if all n genotypes $\vec{G} = (G_1, ..., G_n)$ are homozygous reference, and it is polymorphic otherwise. Hence, given R, the data from the sequencing reads, the probability of site s being polymorphic given the sequencing reads is

$$\mathbb{P}(s = SNP \mid R) = 1 - \frac{\mathbb{P}(R \mid \vec{G})\mathbb{P}(\vec{G})}{\sum_{\vec{G'}} \mathbb{P}(R \mid \vec{G'})\mathbb{P}(\vec{G'})} \tag{A.4}$$

If we assume that the sequencing reads are independent between the individuals, then the probability of the data given the genotypes is a product of the individual genotype likelihoods, i.e.

$$\mathbb{P}(R \mid \vec{G} = (G_1, ..., G_n)) = \prod_{i=1}^{n} \mathbb{P}(R_i \mid G_i) \tag{A.5}$$

The prior distribution of a genotype configuration $\mathbb{P}(\vec{G})$, assuming a neutrally evolving distribution under the coalescent, is given by (Fu, 1995)

$$\mathbb{P}(\vec{G}) = \begin{cases} \frac{\theta}{2}\left(\frac{1}{n_\alpha} + \frac{1}{2n-n_\alpha}\right)\frac{1}{C^{2n}_{n_a(\vec{G})}} & 2n > n_\alpha > 0 \\ \frac{1}{2}\left(1 - \theta\sum_{i=1}^{2n-1} 1/i\right) & \text{otherwise} \end{cases} \tag{A.6}$$

where n_α is the number of chromosomes with allele a, $2n - n_a$ the number of chromosomes with allele b, and θ is the population scaled mutation rate. The challenge in the calculation of Eq. A.4, is the calculation of the normalization

factor that involves the calculation of all possible configurations. Rewriting the normalization factor as

$$\sum_{\vec{G}} \mathbb{P}(R \mid \vec{G}) = \sum_{k} \mathbb{P}(k) \sum_{\vec{G}:n_a(\vec{G})=k} \mathbb{P}(R \mid \vec{G}) = \sum_{k} \mathbb{P}(k) Q_{n,k} \tag{A.7}$$

where

$$\mathbb{P}(k) = \theta\big(\frac{1}{k} + \frac{1}{2n-k}\big)\frac{1}{C_k^{2n}}$$

and

$$Q_{n,k} = \sum_{\vec{G}:n_a(\vec{G})=k} \mathbb{P}(R \mid \vec{G})$$

that is the probability of all possible configurations that include k alleles a. This is calculated efficiently using dynamic programming in $O(n^2)$ steps. Afterwards, the posterior probability of the site s being monomorphic can be calculated, and by choosing a threshold they exclude sites that have a posterior probability less than that threshold.

A.3 glfMultiples

glfMultiples inputs the genotype likelihoods and calculates the likelihood the position is polymorphic and the likelihood the position is monomorphic (Li et al., 2011). The posterior probability that is calculated is

$$\mathbb{P}(M = 1_{\{a,b\}} \mid R) \propto \mathbb{P}(M = 1_{\{a,b\}})\mathbb{P}(R \mid M = 1_{\{a,b\}}) \tag{A.8}$$

where $M = 0$ when the site is monomorphic and $M = 1_{\{a,b\}}$ when the site is polymorphic with alleles a and b. The prior for the site being polymorphic is based on population genetics results and it is given by (Hudson, 1990)

$$\mathbb{P}(M = 1) = \frac{\theta}{2} \sum_{h=2}^{n} h\mathbb{E}(T_h) = \theta \sum_{h=1}^{n-1} \frac{1}{h} \tag{A.9}$$

where n is the sample size and θ is defined as the per pair base heterozygosity. This prior assumes a constant rate neutral mutation process and the infinite-sites assumption.

The prior favours transitions over transversions, i.e.

$$\mathbb{P}(M = 1_{\{a,b\}}) \propto \mathbb{P}(M = 1) \times \begin{cases} 2/3 & \text{a=ref and b is a } \textit{transition} \text{ mutation} \\ 1/6 & \text{a=ref and b is a } \textit{transversion} \text{ mutation} \\ 1/1000 & \text{otherwise} \end{cases} \tag{A.10}$$

The last term in Eq.A.8 is the likelihood which is calculated by

$$\mathbb{P}(R \mid M = 1_{\{a,b\}}) = \prod_{i=1}^{n} \mathbb{P}(R_i \mid M = 1_{\{a,b\}}) \tag{A.11}$$

$$\propto \prod_{i=1}^{n} \sum_{g} \mathbb{P}(G_i = g \mid M = 1_{\{a,b\}})\mathbb{P}(R_i \mid G_i = g) \tag{A.12}$$

where g is the set of all possible genotypes and G_i is the genotype of individual i. The first term in E.q.A.12 is calculated assuming the Hardy-Weinberg equations and the second term is the genotype likelihood. glfMultiples has been applied in the analysis of the low-coverage pilot data of 1KGP and a threshold of 0.9 was applied on the posterior probability to call polymorphic sites.

A.4 GATK (Genome Analysis Toolkit)

GATK applies a Bayesian algorithm for discovering variant sites (DePristo et al., 2011). Define $q_i = \{0, 1, 2\}$ as the count of alternative alleles at individual i. The sum of q_i, $q = \sum_{i=1}^{n} q_i$, over all n individuals in the study sample, represents the number of alternative alleles at the site of interest. Therefore, a site is monomorphic if $\mathbb{P}(q = 0 \mid R)$ is above a threshold, where R is the information coming from the sequencing reads mapped at the site of interest.

The probability $\mathbb{P}(q = X \mid R)$ can be estimated by

$$\mathbb{P}(q = X \mid R) = \frac{\mathbb{P}(q = X)\mathbb{P}(R \mid q = X)}{\sum_Y \mathbb{P}(q = Y)\mathbb{P}(R \mid q = Y)} \tag{A.13}$$

where the prior probability $\mathbb{P}(q = X)$ is the expected probability under the infinite-sites neutral expectation[1] and is given by

$$\mathbb{P}(q = X) = \begin{cases} \theta/X & 2n > X > 0 \\ 1 - \theta \sum_{i=1}^{2n-1} 1/i & \text{otherwise} \end{cases}$$

[1]The same prior as the one given in Eq. A.6

The likelihood is given by

$$\mathbb{P}(R \mid q = X) = \sum_{\vec{G} \in \Gamma} \prod_{i=1}^{2n} \mathbb{P}(R_i \mid G_i)$$

where G_i is the genotype of individual i, R_i is the information from the sequencing reads for individual i, and Γ is the set of all different genotype vectors for the individuals under study that contain X alternative alleles. The number of possible combinations increases exponentially (3^n) and a heuristic method is employed to approximate it. The main idea to reduce the search space of the possible genotype configurations in this method is for $X \in \{1, ..., 2n\}$, for a given configuration with $q = X - 1$ alternative alleles, to find the most likely individual that have an extra alternative allele, and thus $q = X$. This greedy algorithm is linear with the number of individuals. Then, the logarithm of the posterior probability (i.e. QUAL score) is

$$QUAL = -10log_{10}[\mathbb{P}(q = 0 \mid R)] \tag{A.14}$$

and GATK calls a site polymorphic if QUAL>50 for deep coverage sequencing reads if and QUAL>10 for low coverage reads (DePristo et al., 2011).

The aforementioned three methods have been used for the analysis of 1KGP low-coverage pilot data. GATK and the QCALL non-LD dynamic programming differ only on the calculation of the normalizing constant that is challenging given the large search space, and otherwise make the same assumptions,

i.e. the priors used. glfMultiples on the other hand, also applies priors favouring transitions over transversions. Samtools, differs by the other methods by calculating the allele frequency spectrum and incorporating this information to calculate the probability of site being polymorphic.

A.5 GATK (revised) - VQSR (Variant Quality Score Recalibration)

GATK has proposed another model for choosing putative variants that lies on the idea that if a variant has similar characteristics with a known variant, e.g. from the HapMap project, then it is more likely to be a *true* variant compared to variants that do not have similar characteristics and might have been the result of a sequencing or mapping error (DePristo et al., 2011). A variational Bayes Gaussian mixture model (GMM) is fitted to a training set of known variants and the parameters are estimated using the training set. Then the likelihood of each candidate variant can be calculated under the fitted model.

In particular, assume a variant ν_i has a corresponding vector $\vec{\nu_i}$ of covariates for different metrics including base, mapping quality, strand bias, and depth for that site. Then, the probability of the variant site given the model is

$$\mathbb{P}(\nu_i \mid GMM) = \sum_{\kappa=1}^{K} \pi\kappa N(\bar{\nu_i} \mid \mu_\kappa, \Sigma_\kappa)$$

where K is the total number of clusters fitted to the data. The prior of $\vec{\pi}$ is a Dirichlet distribution,

$$\mathbb{P}(\bar{\pi}) = Dir(\bar{\pi} \mid \alpha_0) = C(\alpha_0) \prod_{\kappa=1}^{K} \pi_{\kappa}^{\alpha_0 - 1}$$

the conjugate prior distribution of the mean vector $\bar{\mu}$ is multivariate Normal, and the conjugate prior distribution of the variance-covariance matrix Σ_k, follows an inverse Wishart distribution given by

$$\mathbb{P}(\bar{\mu}, \Sigma) = MVN\Big(\bar{\mu} \mid \bar{m_0}, \Sigma/\beta_0\Big) INW(\Sigma \mid W_0, \tau_0)$$

where m_0, β_0, W_0, and τ_0 are hyperparameters. After the model is fitted using the training data, then the likelihood of a putative variant can be estimated by

$$\begin{aligned} L(\nu_i \mid GMM) &= \mathbb{P}(\nu_i)\mathbb{P}(\vec{\nu_i} \mid GMM) \\ &= (1 - FP_{\text{singleton}})^{AC} \mathbb{P}(\text{novelty of } \nu_i) \sum_{\kappa=1}^{K} \pi\kappa N\big(\bar{\nu_i} \mid \mu_\kappa, \Sigma_k\big) \end{aligned}$$

where $FP_{\text{singleton}}$ is the false positive rate for singletons and it is fixed to 50%, AC is the estimated count of the alternative allele, and the probability of the variant being novel depends on whether it is present in a known variant or not with probabilities fixed to[2]

$$\mathbb{P}(\text{novelty of } \nu_i) = \begin{cases} 97\% & \nu_i \text{ is in HapMap3} \\ 37\% & \text{otherwise} \end{cases}$$

[2]The training set used for the analysis up to now are variants present in HapMap3.

A.6 Multi-population calling

All the non-LD methods presented above, assume that all samples come from one population, since they do not account for the different population allele frequency spectrum when considering more than one population at once. However, by considering only one population at a time, the intercorrelations between the allele frequencies from the different populations are not incorporated in the model, and variants that are rare in one population might be missed, even though the site might be a common SNP in another population. Liu and Marchini (personal communication) have developed an EM algorithm for SNP calling considering all individuals from diverse population backgrounds jointly. The main idea of this multipopulation model is to calculate the allele frequencies of all the possible alleles at each population, and assume that those allele frequencies are interconnected with an ancestral allele frequency that is unknown. The distance between a present population and the ancestral population (i.e. F_{st}) is calculated using known genotype data (e.g. from HapMap, OMNI) and are fixed in the model. Consider at a single site s,

- X_{ki}: number of copies of the chosen allele for individual i in population k
- G_{ki}: genotype for individual i in population k;
- θ: allele frequency of ancestral population;
- θ_k: allele frequency in population k;
- c_k: F_{st}-like parameter, to measure the differentiation between population k

and the ancestral population.

and also define

$$\alpha_k = \frac{1 - c_k}{c_k} \tag{A.15}$$

Thus, at site s for each individual i in population k, consider a beta-binomial model (Marchini et al., 2004)

$$X_{ki} \sim \text{Bin}(2, \theta_k) \tag{A.16}$$

where

$$\theta_k \sim \text{beta}(\theta\alpha_k, (1-\theta)\alpha_k) \tag{A.17}$$

Hence the full likelihood at site s is given by:

$$L(\theta|\mathbf{G}, \mathbf{X}, \theta_k, \alpha_k) = \prod_{k=1}^{K} \left(\prod_{i=1}^{N_k} \mathbb{P}(G_{ki}|X_{ki}) \mathbb{P}(X_{ki}|\theta_k) \right) \mathbb{P}(\theta_k|\alpha_k, \theta) \tag{A.18}$$

where N_k denotes the number of individuals in population k and the total number of individuals is $N = \sum_{k=1}^{K} N_k$.

The allele frequencies are estimated using an EM algorithm where in the E-step the expectation is

$$Q(\theta) = \mathbb{E}_Q \left[\log L(\theta|\mathbf{G}, \mathbf{X}) \right] = \mathbb{E}_Q \left[l(\theta|\mathbf{G}, \mathbf{X}) \right] \tag{A.19}$$

and in the M-step the maximum likelihood of the population specific allele frequencies are estimated, where

$$\theta_k^* = \frac{\sum_{i=1}^{N_k} \sum_{j=0}^{2} jQ_{kij} + \theta\alpha_k - 1}{2N_k + \alpha_k - 2} \tag{A.20}$$

The stopping condition requires the difference between the full likelihood at two successive iterations to be less than a pre-specified threshold, and calls are made when the minor allele frequency is greater than a threshold.

Appendix B

Supplementary material from pilot 1 analysis

B.1 Updating the scaffold with 2-3 way phased genotypes

The final call set of the low-coverage pilot, included sites that have been detected by at least two of the three methods, and after a filter on mapping quality, the final call set is created (2-3 way call set). For genotype calling, a voting strategy has been adopted, and when the methods were in disagreement, QCALL genotypes were preferred, followed by Thunder and GATK, which can be justified by the genotype accuracies in Table 4.4, and phasing has been performed where it was possible. This created a phased call set of known and novel SNPs. The genotype accuracy of the 2-3 way was higher than any of the genotype accuracies of any single method. In this section, we update the haplotype scaffold to include phased sites that are not present in the HapMap3, and we then rerun TreeCall and MVNcall with the updated haplotype scaffold, to examine the

effect on the SNP and genotype calling accuracies.

With the updated haplotype scaffold, MVNcall detects 219 extra sites, detecting 119 extra sites present in the 2-3 way call set, not detected in MVNcall with the HapMap3 scaffold. In the case of the TreeCall method, the call set includes 265 sites not present in the initial run, the majority of those in 2-3 way call set. This shows that a more dense haplotype scaffold will aid the detection of more true positives, and that the phased haplotypes from 1KGP aim towards this direction.

Method (scaffold)	HomR %	Het %	HomA %
TreeCall (Scaffold = Hap3)	3.61	10.43	13.57
TreeCall (Scaffold = Hap3 + 1KGP)	3.50	9.36	13.00
MVNcall (Scaffold = Hap3)	2.01	5.59	10.27
MVNcall (Scaffold = Hap3 + 1KGP)	1.99	5.29	10.25

Table B.1: Percent genotype discordance comparison when we use a haplotype scaffold with HapMap3 SNPs (Hap3 in table) and when we use a haplotype scaffold with HapMap3 SNPs and the phased pilot1 2-3 way SNPs (Hap3 + 1KGP in table). We compare both TreeCall and MVNcall percent genotype discordance rates when using the two haplotype scaffolds on SNPs that are present in HapMap2 but not in HapMap3 nor 2-3 way (1,104 SNPs)

In Table B.1, we show the genotype accuracy metrics for TreeCall and MVNcall when we use the two scaffolds, HapMap3 haplotypes and HapMap3 and 1KGP haplotypes. We compare them to sites that are present in HapMap2 but not present in either HapMap3 nor the phased call set of 1KGP. We observe that the percent genotype discordances is higher than the corresponding percent genotype discordances in Table 4.4 which can be explained by the fact that those positions are in harder to call regions. Updating the scaffold with the 1KGP, we see an improvement of 1.07% and 0.3% on TreeCall and MVNcall respectively,

on heterozygous genotype accuracy, and for homozygous alternative genotypes 0.57% and 0.02% respectively and for homozygous reference genotypes we observe a small improvement in the TreeCall method 0.11% whereas for MVNcall 0.02%. We notice that TreeCall benefits more than MVNcall on genotype accuracies when we update the scaffold with the 1KGP SNPs. This might be because, with the increased number of SNPs in the scaffold, we end up constructing more genealogical trees in a more dense way, and thus benefit from that extra information in TreeCall, where in MVNcall we use the same number of flanking sites and the benefit will be more apparent when the updated SNPs are in high LD with the sites used here for calculating genotype accuracies.

B.2 Plots of MAF versus genotype accuracy

In Fig.B.1, we display all five methods genotype accuracy curves for heterozygous genotypes. A common feature across all methods is the lower genotype accuracy for low minor allele frequency sites, as we can see at the ends of the curves, where the genotype frequency falls for rare variants, which is expected given the power to detect rare variants with low-coverage information. Zooming in the low allele frequency variants, we see that Thunder and GATK have the lowest accuracy for singletons followed by QCALL, MVNcall and TreeCall.

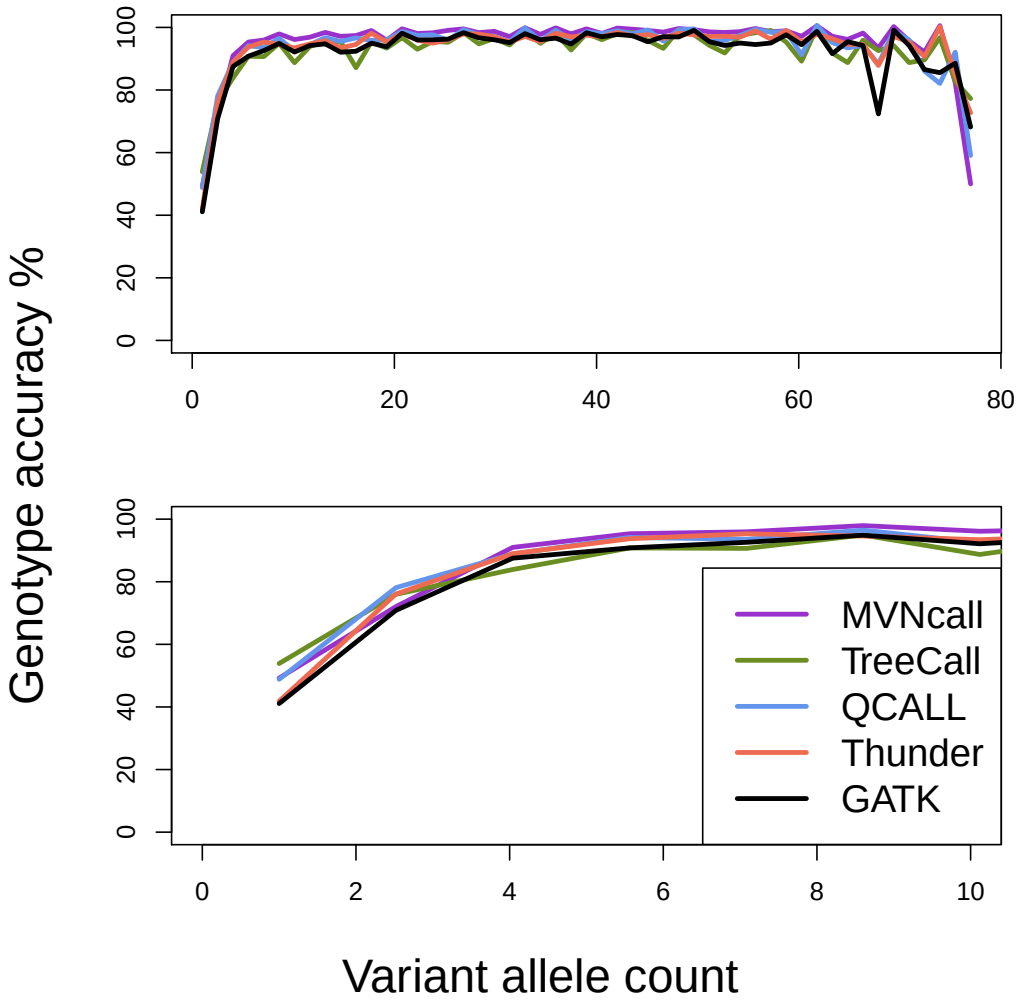

Figure B.1: Genotype accuracy for heterozygous genotypes by allele frequency. On the top panel we present the curves for the whole allele frequency spectrum and on the bottom panel we zoom in on the variant allele counts between 1 and 10. Different colours correspond to different methods: black for MVNcall, red for TreeCall, blue for QCALL, green for THUNDER and orange for GATK.

It is interesting to observe that for singletons our method for calling genotypes on TreeCall method outperforms all the other four methods. This might be because we do not average over the information from all trees, and thus lose some signal, but we rather choose the tree and mutation with the highest likelihood. For common SNPs, MVNcall gives better accuracies than all other meth-

ods. We repeat the same analysis for the genotype accuracies curves for homozygous genotypes (plots not shown), and we find similar patterns in terms of the order of performance of the different methods.

Appendix C

Supplementary material from phase 1 analysis

C.1 Performance of MVNcall for different parameter settings for λ, k and number of flanking sites

To understand how the different parameter settings affect the performance of MVNcall, we choose a 10Mbp region on chromosome 20 (chr20:40,000,000- 50,000,000) and we apply different combinations of three tuning parameters (λ, number of flanking sites, and number of conditioning haplotypes k) and record the genotype accuracies on SNPs present in HapMap3 but not in OMNI (2,935 SNPs). As it is noted in Chapter 5, after investigation of whether we should analyse all populations together versus divide the populations in population groups and run them seperately, we observe that in the case of our method it is preferred to run the populations by groups. Also, comparing the two haplotype distance measures for choosing the closest haplotypes, the perfect match distance is pre-

ferred to the hamming distance measure. Thus, the results presented below follow the aformentioned scenario, population grouping and the perfect match distance to choose the closest haplotypes.

We notice that the optimal tuning parameters for the three population groupings, European & Admixed, African & Admixed and Asian & Admixed where the same and thus below we only present the results from analysing the third population group, although the same pattern was observed for the other two population groups. We run MVNcall for the different combinations where $\lambda = \{0.03, 0.04, 0.05, 0.06, 0.07, 0.08\}$, number of flanking sites at either side $\{20, 30, 40, 50\}$ and $k = \{10, 50, 100, 150, 200, 250, 300\}$ for 300 iterations with 100 burn-in. In Tables C.1,C.2 and C.3 we report some of the genotype accuracies results for a subset of the runs. In particular, in Table C.1, we vary λ and number of flanking sites while fixing $k = 10$, in Table C.2 we vary againλ and number of flanking sites while fixing $k = 100$ and in Table C.3 we vary λ and number of flanking sites while fixing $k = 200$. For $k = 10$, for different number of flanking sites, the genotype discordance decreases for larger values of λ, which is expected since with only 10 conditioning haplotypes, there is not much LD information, and by allowing more information from the genotype likelihoods, we end up in better accuracies[1]. In Table C.2, we notice that by increasing the number of flanking sites we see a decrease in the percent genotype discordance. We should notice that for $\lambda > 0.06$ and 50 flanking sites at either side, the percent overall discordance is 0.39. Lastly, in Table C.3, we observe the lowest overall discordance for 50 flank-

[1]This follows the discussion of the λ parameter in Section 4.3.

ing sites at either side for $\lambda > 0.06$. However, comparing the percent genotype accuracies by type, we notice that by increasing the value of k, the percent discordance for the homozygous genotypes is decreasing, but for the heterozygous genotypes is increasing. We deal with this observation by running the model for different values of k and averaging over the two runs discussed in the main text (see Section 5.2.1).

To also run the method for > 50 flanking sites at either side, and we observe that the percent overall discordance is reaching a plateau after 50, see Fig.C.1. Lastly, we apply the shrinkage on the off-diagonal of the covariance matrix, as it was applied on the low-coverage pilot data, but we observe no benefit on the genotype accuracies. A possible explanation is that the scaffold used in phase 1, is more densed than the scaffold used in the low-coverage pilot analysis, and thus, by using the a similar window size of SNPs, we actually consider a smaller region and thus the probability of recombination between the flanking sites is small.

After the study of the tuning parameters, we run the algorithm for 300 iterations with different number of burn-in iterations. The algorithm converged after 100 iterations. We thus decide to run the algorithm for 100 iterations with 10 burn-in in this dataset.

# SNPs / λ	0.04	0.05	0.06	0.07	0.08
20	0.59	0.55	0.53	0.52	0.51
	(0.32,0.97,0.91)	(0.31,0.91,0.88)	(0.29,0.87,0.83)	(0.27,0.86,0.82)	(0.27,0.86,0.79)
30	0.66	0.56	0.52	**0.50**	**0.50**
	(0.35,1.13,0.96)	(0.31,0.95,0.84)	(0.28,0.84,0.78)	(0.28,0.81,0.78)	(0.27,0.82,0.74)
40	0.84	0.62	0.55	0.52	**0.50**
	(0.46,1.41,1.19)	(0.36,1.08,0.92)	(0.31,0.88,0.84)	(0.28,0.84,0.78)	(0.27,0.82,0.76)
50	1.04	0.71	0.59	0.53	0.52
	(0.57,1.76,1.48)	(0.36,1.11,0.99)	(0.32,0.96,0.89)	(0.29,0.88,0.82)	(0.28,0.84,0.79)

Table C.1: Genotype discordance comparison under different combinations of tuning parameters (λ and number of flanking sites), for k=**10**. We report the percent overall discordance and in brackets we break down the overall discordance by genotype type (homozygous reference, heterozygous and homozygous non-reference). Results are based on 2,935 SNPs present in HapMap3 but not in OMNI, for 167 Asian individuals present in HapMap3 and 1KGP.

# SNPs/ λ	0.04	0.05	0.06	0.07	0.08
20	0.45	0.43	0.43	0.43	0.43
	(0.23,0.74,,0.74)	(0.21,0.74,0.71)	(0.20,0.76,0.68)	(0.19,0.78,0.67)	(0.18,0.80,0.66)
30	0.43	0.41	0.40	0.41	0.40
	(0.22,0.70,0.74)	(0.21,0.74,0.71)	(0.19,0.68,0.69)	(0.18,0.71,0.68)	(0.17,0.73,0.66)
40	0.43	0.41	0.40	0.40	0.40
	(0.22,0.67,0.74)	(0.20,0.66,0.70)	(0.19,0.67,0.69)	(0.18,0.68,0.68)	(0.17,0.70,0.66)
50	0.43	0.41	**0.39**	**0.39**	**0.39**
	(0.22,0.67,0.74)	(0.19,0.65,0.68)	(0.19,0.65,0.67)	(0.18,0.66,0.66)	(0.17,0.68,0.64)

Table C.2: Genotype discordance comparison under different combinations of tuning parameters (λ and number of flanking sites), for k=**100**. We report the percent overall discordance and in brackets we break down the overall discordance by genotype type (homozygous reference, heterozygous and homozygous non-reference). Results are based on 2,935 SNPs present in HapMap3 but not in OMNI, for 167 Asian individuals present in HapMap3 and 1KGP.

# SNPs / λ	0.04	0.05	0.06	0.07	0.08
20	0.46	0.44	0.43	0.43	0.44
	(0.24,0.76,0.76)	(0.21,0.75,0.72)	(0.19,0.78,0.69)	(0.18,0.81,0.68)	(0.17,0.85,0.67)
30	0.45	0.42	0.41	0.41	0.41
	(0.23,0.70,0.76)	(0.20,0.69,0.71)	(0.19,0.71,0.69)	(0.18,0.73,0.67)	(0.17,0.77,0.65)
40	0.43	0.41	0.40	0.39	0.39
	(0.22,0.67,0.75)	(0.20,0.67,0.71)	(0.19,0.67,0.69)	(0.17,0.70,0.67)	(0.16,0.73,0.64)
50	0.42	0.40	**0.39**	**0.39**	**0.39**
	(0.22,0.66,0.73)	(0.19,0.65,0.69)	(0.18,0.66,0.66)	(0.17,0.68,0.66)	(0.16,0.71,0.65)

Table C.3: Genotype discordance comparison under different combinations of tuning parameters (λ and number of flanking sites), for k=**200**. We report the percent overall discordance and in brackets we break down the overall discordance by genotype type (homozygous reference, heterozygous and homozygous non-reference). Results are based on 2,935 SNPs present in HapMap3 but not in OMNI, for 167 Asian individuals present in HapMap3 and 1KGP.

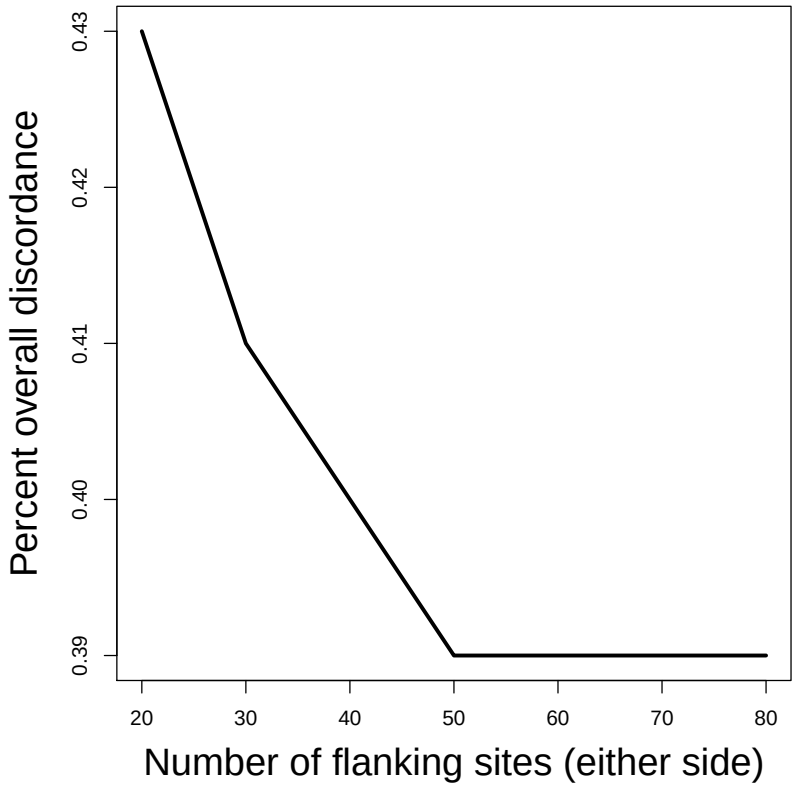

Figure C.1: Number of flanking sites at either side versus percent overall discordance.

C.2 Genotype comparison on validation sites

From the Phase 1 of the 1KGP, a subset of 382 individuals were validated on a set of SNPs with Sequenom[2]. Quality control was applied to the raw set and a curated validation set of SNPs is created. 119 SNPs pass the quality control and are present in all four datasets. We compare the genotype accuracy on the curated set. The data can be found in ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/technical/working/20110524_phase1_chr20_snp_validation/.

In Table C.4 we present the results obtained from the different methods. Each

[2]http://www.sequenom.com/home

Val/Est	HomR	Het	HomA
HomR	59,518	108	2
Het	43	1,552	4
HomA	2	5	1,178

(a) MVNcall

Val/Est	HomR	Het	HomA
HomR	59,446	175	7
Het	36	1,559	4
HomA	2	7	1,176

(b) Thunder

Val/Est	HomR	Het	HomA
HomR	59,428	198	2
Het	54	1,540	5
HomA	2	10	1,173

(c) SNPTools

Val/Est	HomR	Het	HomA
HomR	59,517	109	2
Het	53	1,541	5
HomA	3	5	1,177

(d) BEAGLE

Table C.4: Genotype call comparison on the filtered curated validation set. Each row represents the genotype from the validation set (Val), and on each column what was the genotype called by the different methods (Est). The diagonals of the tables are the correct calls made by the methods. Four tables are presented: MVNcall (top left), Thunder (top right), SNPTools (bottom left), and BEAGLE (bottom right).

row represents the genotypes called by Sequenom and each column represents the estimated genotype from each method. The diagonals denote the number of correct calls. MVNcall is in good agreement with the Sequenom genotypes, with the highest concordance for homozygous genotypes, followed by BEAGLE. In terms of heterozygous genotype concordance, Thunder has the highest concordance, followed by MVNcall, BEAGLE, and SNPTools.

C.3 Additional plots from imputation analysis

In this section we present additional plots from the imputation accuracy analysis presented in Section 5.3.2. In the main text, the imputation comparison between the different methods was presented using the Affy500k as an imputation scaffold. Here, we present the results regarding the imputation scaffolds (i) Affy6.0, (ii) Illumina1M and (iii) Omni2.5M. For all plots, on the x-axis is the

non-reference allele frequency of imputed SNPs on log scale, and on the y-axis is the non-reference genotype concordance. Some remarks from the plots below, focusing mostly on the rare SNPs spectrum

- For the European individuals, SNPTools and Thunder are indistinguishable and provide the highest non-reference concordance for rare SNPs. MVNcall follows with marginal differences from SNPTools and Thunder, with the discrepancy on the rare SNPs interval to increase as the imputation scaffold is more densed. Apart from the case where the OMNI2.5M imputation scaffold is used, MVNcall outperforms BEAGLE,
- For the African individuas, MVNcall has the highest non-reference concordance for all different imputation scaffolds, followed by Thunder, whereas SNPTools and BEAGLE have marginal differences,
- For the Admixed individuals, SNPTools has a slight advantage over the other methods, followed by Thunder, MVNcall and BEAGLE.

In addition, in Tables C.5, C.6, C.7, and C.8, we report the non-reference concordance for the difference reference panels for each imputation scaffold / populatio group.

Imputation scaffold / Population	European	African	Admixed
Affy500k	0.9158	0.8245	0.8962
Affy6.0	0.9516	0.8973	0.9390
Illumina1M	0.9655	0.9114	0.9553
OMNI2.5M	0.9729	0.9371	0.9623

Table C.5: Non-reference genotype concordance using the reference panel from **Thunder** for different imputation scaffolds and population groups

Imputation scaffold / Population	European	African	Admixed
Affy500k	0.9170	0.8316	0.9022
Affy6.0	0.9540	0.8979	0.9398
Illumina1M	0.9676	0.9117	0.9573
OMNI2.5M	0.9746	0.9366	0.9629

Table C.6: Non-reference genotype concordance using the reference panel from **SNPTools** for different imputation scaffolds and population groups

Imputation scaffold / Population	European	African	Admixed
Affy500k	0.9191	0.8408	0.9031
Affy6.0	0.9539	0.9067	0.9417
Illumina1M	0.9665	0.9182	0.9577
OMNI2.5M	0.9720	0.9429	0.9608

Table C.7: Non-reference genotype concordance using the reference panel from **MVNcall** for different imputation scaffolds and population groups

Imputation scaffold / Population	European	African	Admixed
Affy500k	0.9119	0.8210	0.8943
Affy6.0	0.9481	0.8916	0.9352
Illumina1M	0.9632	0.9071	0.9527
OMNI2.5M	0.9716	0.9348	0.9590

Table C.8: Non-reference genotype concordance using the reference panel from **BEAGLE** for different imputation scaffolds and population groups

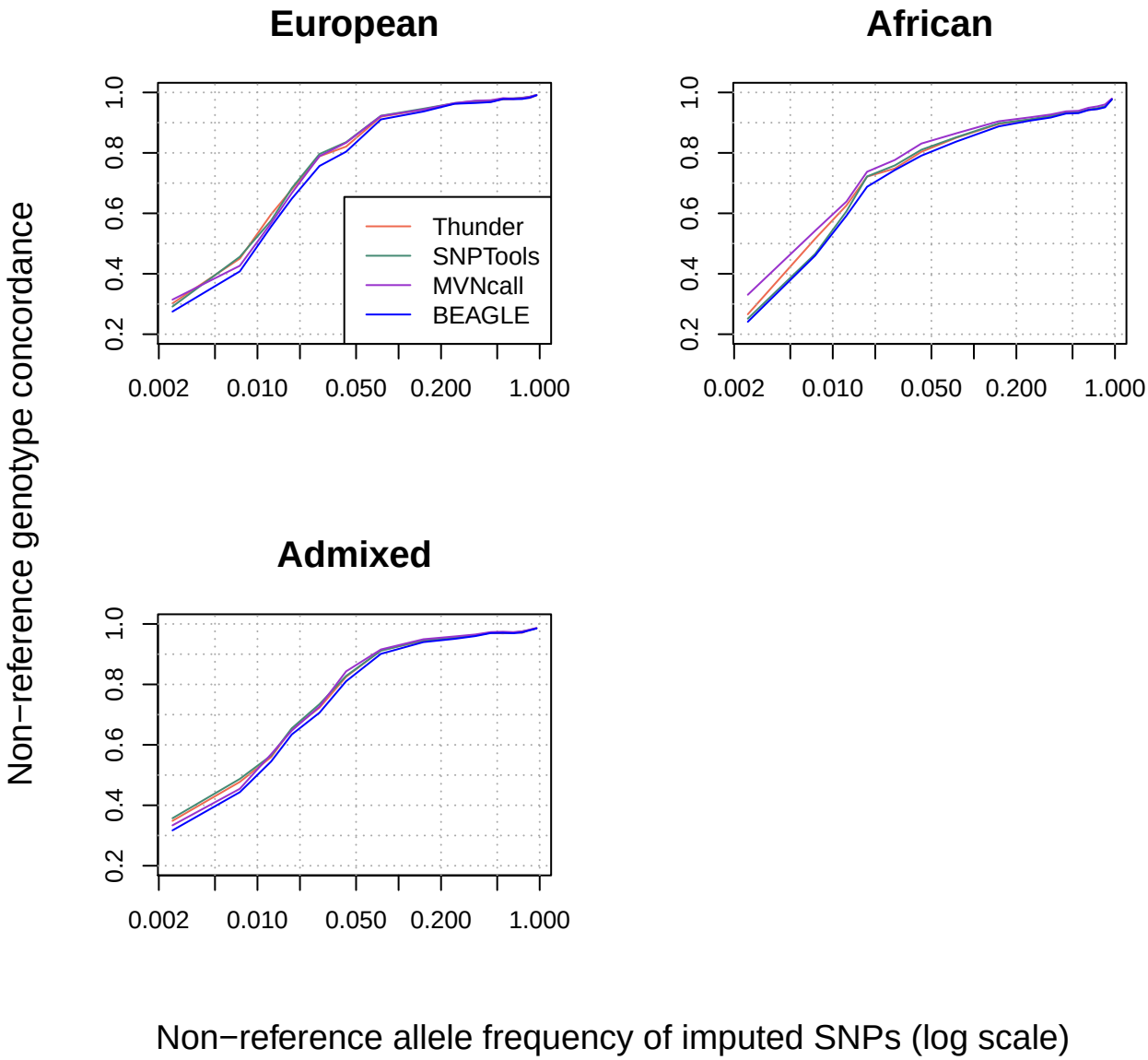

Figure C.2: Allele frequency at imputed SNPs (log scale) versus the non-reference genotype concordance. Different lines correspond to different reference panels. The results are stratified by population group: European (top left), African (top right), and Admixed (bottom left). Imputation scaffold used is the **Affy6.0**.

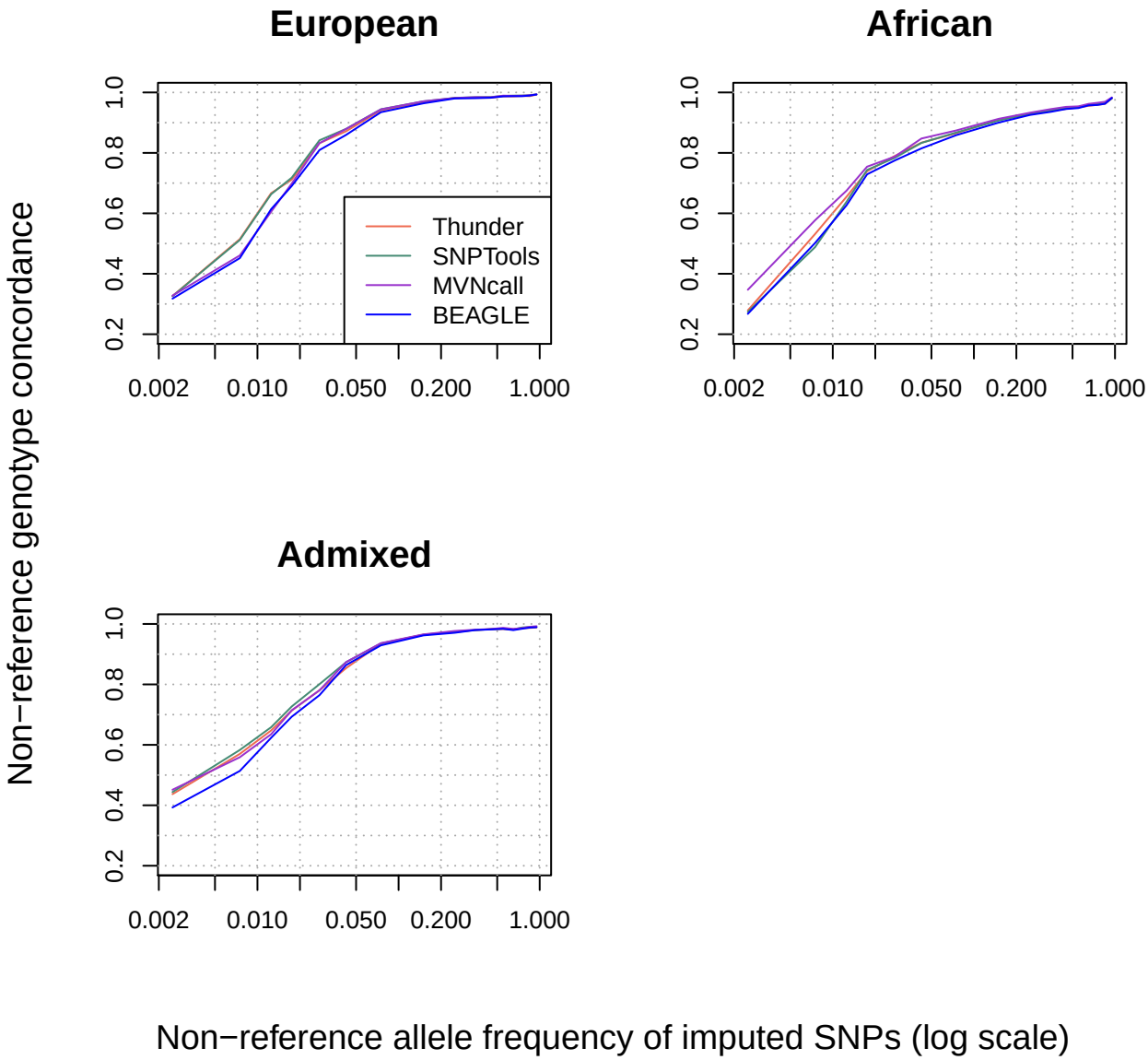

Figure C.3: Allele frequency at imputed SNPs (log scale) versus the non-reference genotype concordance. Different lines correspond to different reference panels. The results are stratified by population group: European (top left), African (top right), and Admixed (bottom left). Imputation scaffold used is the **Illumina1M**.

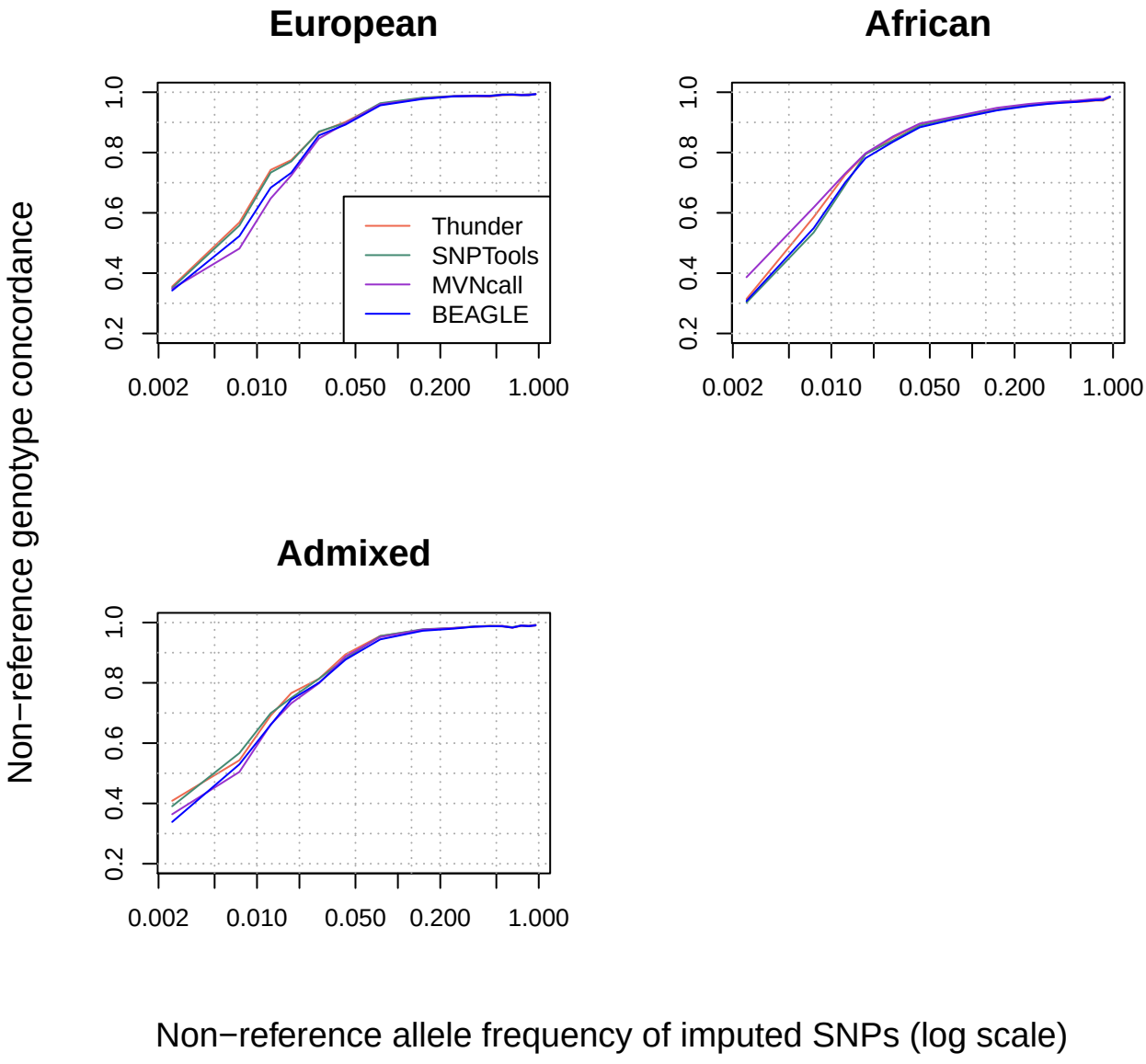

Figure C.4: Allele frequency at imputed SNPs (log scale) versus the non-reference genotype concordance. Different lines correspond to different reference panels. The results are stratified by population group: European (top left), African (top right), and Admixed (bottom left). Imputation scaffold used is the **Omni2.5M**.

Bibliography

Alkan, C., Sajjadian, S. and Eichler, E. E. (2011) Limitations of next-generation genome sequence assembly. *Nature methods*, **8**, 61–65.

Beckmann, L. (2010) Haplotype sharing methods. *Encyclopedia of Life Sciences, John Wiley & Sons, Ltd*.

Bentley, D. R., Balasubramanian, S., Swerdlow, H. P., Smith, G. P., Milton, J., Brown, C. G., Hall, K. P., Evers, D. J., Barnes, C. L., Bignell, H. R., Boutell, J. M., Bryant, J., Carter, R. J., Keira Cheetham, R., Cox, A. J., Ellis, D. J., Flatbush, M. R., Gormley, N. A., Humphray, S. J., Irving, L. J., Karbelashvili, M. S., Kirk, S. M., Li, H., Liu, X., Maisinger, K. S., Murray, L. J., Obradovic, B., Ost, T., Parkinson, M. L., Pratt, M. R., Rasolonjatovo, I. M. J., Reed, M. T., Rigatti, R., Rodighiero, C., Ross, M. T., Sabot, A., Sankar, S. V., Scally, A., Schroth, G. P., Smith, M. E., Smith, V. P., Spiridou, A., Torrance, P. E., Tzonev, S. S., Vermaas, E. H., Walter, K., Wu, X., Zhang, L., Alam, M. D., Anastasi, C., Aniebo, I. C., Bailey, D. M. D., Bancarz, I. R., Banerjee, S., Barbour, S. G., Baybayan, P. A., Benoit, V. A., Benson, K. F., Bevis, C., Black, P. J., Boodhun, A., Brennan, J. S., Bridgham, J. A., Brown, R. C., Brown, A. A., Buermann, D. H., Bundu, A. A.,

Burrows, J. C., Carter, N. P., Castillo, N., Chiara, Chang, S., Neil Cooley, R., Crake, N. R., Dada, O. O., Diakoumakos, K. D., Dominguez-Fernandez, B., Earnshaw, D. J., Egbujor, U. C., Elmore, D. W., Etchin, S. S., Ewan, M. R., Fedurco, M., Fraser, L. J., Fuentes Fajardo, K. V., Scott Furey, W., George, D., Gietzen, K. J., Goddard, C. P., Golda, G. S., Granieri, P. A., Green, D. E., Gustafson, D. L., Hansen, N. F., Harnish, K., Haudenschild, C. D., Heyer, N. I., Hims, M. M., Ho, J. T., Horgan, A. M., Hoschler, K., Hurwitz, S., Ivanov, D. V., Johnson, M. Q., James, T., Huw Jones, T. A., Kang, G.-D., Kerelska, T. H., Kersey, A. D., Khrebtukova, I., Kindwall, A. P., Kingsbury, Z., Kokko-Gonzales, P. I., Kumar, A., Laurent, M. A., Lawley, C. T., Lee, S. E., Lee, X., Liao, A. K., Loch, J. A., Lok, M., Luo, S., Mammen, R. M., Martin, J. W., McCauley, P. G., McNitt, P., Mehta, P., Moon, K. W., Mullens, J. W., Newington, T., Ning, Z., Ling Ng, B., Novo, S. M., O/'Neill, M. J., Osborne, M. A., Osnowski, A., Ostadan, O., Paraschos, L. L., Pickering, L., Pike, A. C., Pike, A. C., Chris Pinkard, D., Pliskin, D. P., Podhasky, J., Quijano, V. J., Raczy, C., Rae, V. H., Rawlings, S. R., Chiva Rodriguez, A., Roe, P. M., Rogers, J., Rogert Bacigalupo, M. C., Romanov, N., Romieu, A., Roth, R. K., Rourke, N. J., Ruediger, S. T., Rusman, E., Sanches-Kuiper, R. M., Schenker, M. R., Seoane, J. M., Shaw, R. J., Shiver, M. K., Short, S. W., Sizto, N. L., Sluis, J. P., Smith, M. A., Ernest Sohna Sohna, J., Spence, E. J., Stevens, K., Sutton, N., Szajkowski, L., Tregidgo, C. L., Turcatti, G., vandeVondele, S., Verhovsky, Y., Virk, S. M., Wakelin, S., Walcott, G. C., Wang, J., Worsley, G. J., Yan, J., Yau, L., Zuerlein, M., Rogers, J., Mullikin, J. C., Hurles, M. E., McCooke, N. J., West, J. S., Oaks, F. L., Lundberg, P. L., Klen-

erman, D., Durbin, R. and Smith, A. J. (2008) Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, **456**, 53–59.

Browning, B. L. and Browning, S. R. (2007a) Efficient multilocus association testing for whole genome association studies using localized haplotype clustering. *Genetic epidemiology*, **31**, 365–375.

Browning, S. R. (2006) Multilocus association mapping using variable-length markov chains. *American journal of human genetics*, **78**, 903–913.

Browning, S. R. and Browning, B. L. (2007b) Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *American journal of human genetics*, **81**, 1084–1097.

Chen, G., Marjoram, P. and Wall, J. D. (2009) Fast and flexible simulation of DNA sequence data. *Genome Research*, **advance access**, x. URL `http://genome.cshlp.org/content/early/2008/11/24/gr.083634.108.short?rss=1`.

Claeskens, G. and Hjort, N. L. (2008) *Model Selection and Model Averaging*. Cambridge University Press.

Collins, F. S., Morgan, M. and Patrinos, A. (2003) The human genome project: lessons from large-scale biology. *Science*, **300**, 286–290.

Cox, A. (2007) Eland: Efficient large-scale alignment of nucleotide bases.

Delaneau, O., Marchini, J. and Zagury, J. (2011) A linear complexity phasing method for thousands of genomes. *Nature Methods*, **9**, 179181.

DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., Philippakis, A. A., del Angel, G., Rivas, M. A., Hanna, M., McKenna, A., Fennell, T. J., Kernytsky, A. M., Sivachenko, A. Y., Cibulskis, K., Gabriel, S. B., Altshuler, D. and Daly, M. J. (2011) A framework for variation discovery and genotyping using next-generation dna sequencing data. *Nature Genetics*, **43**, 491–498.

Draper, D. (1995) Assessment and Propagation of Model Uncertainty. *Journal of the Royal Statistical Society. Series B (Methodological)*, **57**, 45–97.

Ewing, B. and Green, P. (1998) Base-calling of automated sequencer traces using phred ii. error probabilities. *Genome Research*, **8**, 186–194.

Ewing, B., Hillier, L., Wendl, M. C. and Green, P. (1998) Base-calling of automated sequencer traces using phred. i. accuracy assessment. *Genome Research*, **8**, 175–185.

Fearnhead, P. and Donnelly, P. (2001) Estimating recombination rates from population genetic data. *Genetics*, **159**, 1299–1318. URL `http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1461855/`.

Fisher, R. (1930) *The genetical theory of natural selection*. Oxford: Clarendon Press.

Flicek, P. and Birney, E. (2009) Sense from sequence reads: methods for alignment and assembly. *Nature Methods*, **6**, S6–S12.

Fu, Y.-X. (1995) Statistical properties of segregating sites. *Theoretical Population Biology*, **48**, 172–197.

Golub, G. H. and Van Loan, C. F. (1996) *Matrix computations*. Baltimore, MD: John Hopkins University Press, 3rd edn.

Greenspan, G. and Geiger, D. (2004) High density linkage disequilibrium mapping using models of haplotype block variation. *Bioinformatics*, **20**, i137–i144. URL `http://bioinformatics.oxfordjournals.org/content/20/suppl_1/i137.abstract`.

Griffiths, R. C. and Marjoram, P. (1996) Ancestral inference from samples of dna sequences with recombination. *Journal of computational biology : a journal of computational molecular cell biology*, **3**, 479–502.

Hallsworth, C. (2010) *Approximate Genealogical Inference in the Presence of Recombination*. Ph.D. thesis, Department of Statistics, University of Oxford.

Harris, T. D., Buzby, P. R., Babcock, H., Beer, E., Bowers, J., Braslavsky, I., Causey, M., Colonell, J., DiMeo, J., Efcavitch, J. W., Giladi, E., Gill, J., Healy, J., Jarosz, M., Lapen, D., Moulton, K., Quake, S. R., Steinmann, K., Thayer, E., Tyurina, A., Ward, R., Weiss, H. and Xie, Z. (2008) Single-molecule dna sequencing of a viral genome. *Science*, **320**, 106–109.

Hindorff, L., Junkins, H., Hall, P., Mehta, J. and Manolio, T. (2010) A catalog of published genome-wide association studies. URL `http://www.genome.gov/gwastudies/`.

Hoeting, J. A., Madigan, D., Raftery, A. E. and Volinsky, C. T. (1999) Bayesian Model Averaging: A Tutorial. *Statistical Science*, **14**, 382–401.

Homer, N., Merriman, B. and Nelson, S. F. (2009) Bfast: An alignment tool for large scale genome resequencing. *PLoS ONE*, **4**, e7767+.

Howie, B., Marchini, J. and Stephens, M. (2011) Genotype imputation with thousands of genomes. *G3 (Bethesda, Md.)*, **1**, 457–470.

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

Huang, L., Li, Y., Singleton, A. B., Hardy, J. A., Abecasis, G., Rosenberg, N. A. and Scheet, P. (2009) Genotype-imputation accuracy across worldwide human populations. *American journal of human genetics*, **84**, 235–250.

Huang, W., Li, L., Myers, J. R. and Marth, G. T. (2012) Art: a next-generation sequencing read simulator. *Bioinformatics*, **28**, 593–594. URL `http://dx.doi.org/10.1093/bioinformatics/btr708`.

Hudson, R. R. (1990) Gene genealogies and the coalescent process. *Oxf. Surv. Evol. Biol.*, **7**, 1–44.

Human Genome Project (2010) URL `http://www.ornl.gov/sci/techresources/Human_Genome/home.shtml`.

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

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

Kim, J.-I., Ju, Y. S., Park, H., Kim, S., Lee, S., Yi, J.-H., Mudge, J., Miller, N. A., Hong, D., Bell, C. J., Kim, H.-S., Chung, I.-S., Lee, W.-C., Lee, J.-S., Seo, S.-H., Yun, J.-Y., Woo, H. N., Lee, H., Suh, D., Lee, S., Kim, H.-J., Yavartanoo, M., Kwak, M., Zheng, Y., Lee, M. K., Park, H., Kim, J. Y., Gokcumen, O., Mills, R. E., Zaranek, A. W., Thakuria, J., Wu, X., Kim, R. W., Huntley, J. J., Luo, S., Schroth, G. P., Wu, T. D., Kim, H., Yang, K.-S., Park, W.-Y., Kim, H., Church, G. M., Lee, C., Kingsmore, S. F. and Seo, J.-S. (2009) A highly annotated whole-genome sequence of a korean individual. *Nature*, **460**, 1011–1015.

Kimmel, G. and Shamir, R. (2005) GERBIL: Genotype resolution and block identification using likelihood. *Proc Natl Acad Sci U S A*, **102**, 158–62. URL `http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?cmd=prlinks&dbfrom=pubmed&retmode=ref&id=15615859`.

Kingman, J. (1982) The coalescent. *Stochastic Processes and their Applications*, **13**, 235–248.

Ku, C. S., Loy, E. Y., Pawitan, Y. and Chia, K. S. (2010) The pursuit of genome-wide association studies: where are we now & quest;. *Journal of Human Genetics*, **55**, 195–206.

Langmead, B., Trapnell, C., Pop, M. and Salzberg, S. (2009) Ultrafast and

memory-efficient alignment of short dna sequences to the human genome. *Genome Biology*, **10**, R25+.

Le, S. Q. and Durbin, R. (2011) Snp detection and genotyping from low-coverage sequencing data on multiple diploid samples. *Genome Research*, **21**, 952–960.

Levy, S., Sutton, G., Ng, P. C., Feuk, L., Halpern, A. L., Walenz, B. P., Axelrod, N., Huang, J., Kirkness, E. F., Denisov, G., Lin, Y., Macdonald, J. R., Pang, A. W. C., Shago, M., Stockwell, T. B., Tsiamouri, A., Bafna, V., Bansal, V., Kravitz, S. A., Busam, D. A., Beeson, K. Y., Mcintosh, T. C., Remington, K. A., Abril, J. F., Gill, J., Borman, J., Rogers, Y. H., Frazier, M. E., Scherer, S. W., Strausberg, R. L. and Venter, J. C. (2007) The diploid genome sequence of an individual human. *PLoS Biol*, **5**, e254.

Li, H. (2011a) A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*, **27**, 2987–2993.

— (2011b) Improving snp discovery by base alignment quality. *Bioinformatics*, **27**, 1157–1158.

Li, H. and Durbin, R. (2009) Fast and accurate short read alignment with burrows–wheeler transform. *Bioinformatics*, **25**, 1754–1760.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R. and Subgroup, . G. P. D. P. (2009a) The sequence alignment/map format and samtools. *Bioinformatics*, **25**, 2078–2079.

Li, H. and Homer, N. (2010) A survey of sequence alignment algorithms for next-generation sequencing. *Briefings in Bioinformatics*, **11**, 473–483.

Li, H., Ruan, J. and Durbin, R. (2008a) Mapping short dna sequencing reads and calling variants using mapping quality scores. *Genome research*, **18**, 1851–1858.

Li, J. and Jiang, T. (2005) Haplotype-based linkage disequilibrium mapping via direct data mining. *Bioinformatics*, **21**, 4384–4393.

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

Li, R., Li, Y., Kristiansen, K. and Wang, J. (2008b) Soap: short oligonucleotide alignment program. *Bioinformatics*, **24**, 713–714.

Li, R., Yu, C., Li, Y., Lam, T.-W., Yiu, S.-M., Kristiansen, K. and Wang, J. (2009b) Soap2: an improved ultrafast tool for short read alignment. *Bioinformatics*, **25**, 1966–1967.

Li, Y., Ding, G. and Abecasis, G. (2006) Mach 0.1: Rapid haplotype reconstruction and missing genotype inference. *American Journal of Human Genetics*, **79**, S2290.

Li, Y., Sidore, C., Kang, H. M., Boehnke, M. and Abecasis, G. R. (2011) Low-coverage sequencing: Implications for design of complex trait association studies. *Genome Research*, **21**, 940–951.

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

Lunter, G. and Goodson, M. (2011) Stampy: A statistical algorithm for sensitive and fast mapping of illumina sequence reads. *Genome research*, **21**, 936–939.

Magi, A., Benelli, M., Gozzini, A., Girolami, F., Torricelli, F. and Brandi, M. (2010) Bioinformatics for next generation sequencing data. *Genes*, **1**, 294–307.

Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., McCarthy, M. I., Ramos, E. M., Cardon, L. R., Chakravarti, A., Cho, J. H., Guttmacher, A. E., Kong, A., Kruglyak, L., Mardis, E., Rotimi, C. N., Slatkin, M., Valle, D., Whittemore, A. S., Boehnke, M., Clark, A. G., Eichler, E. E., Gibson, G., Haines, J. L., Mackay, T. F., McCarroll, S. A. and Visscher, P. M. (2009) Finding the missing heritability of complex diseases. *Nature*, **461**, 747–753.

Marchini, J., Cardon, L. R., Phillips, M. S. and Donnelly, P. (2004) The effects of human population structure on large genetic association studies. *Nature genetics*, **36**, 512–517.

Marchini, J., Howie, B., Myers, S., McVean, G. and Donnelly, P. (2007) A new multipoint method for genome-wide association studies by imputation of genotypes. *Nature Genetics*, **39**, 906–913.

Mardis, E. R. (2008) The impact of next-generation sequencing technology on genetics. *Trends Genet.*, **24**, 133–41.

Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bemben, L. A., Berka, J., Braverman, M. S., Chen, Y.-J. J., Chen, Z., Dewell, S. B., Du, L., Fierro, J. M., Gomes, X. V., Godwin, B. C., He, W., Helgesen, S., Ho, C. H. H., Ho, C. H. H., Irzyk, G. P., Jando, S. C., Alenquer, M. L., Jarvie, T. P., Jirage, K. B., Kim, J.-B. B., Knight, J. R., Lanza, J. R., Leamon, J. H., Lefkowitz, S. M., Lei, M., Li, J., Lohman, K. L., Lu, H., Makhijani, V. B., McDade, K. E., McKenna, M. P., Myers, E. W., Nickerson, E., Nobile, J. R., Plant, R., Puc, B. P., Ronan, M. T., Roth, G. T., Sarkis, G. J., Simons, J. F. F., Simpson, J. W., Srinivasan, M., Tartaro, K. R., Tomasz, A., Vogt, K. A., Volkmer, G. A., Wang, S. H., Wang, Y., Weiner, M. P., Yu, P., Begley, R. F. and Rothberg, J. M. (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, **437**, 376–380.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M. and DePristo, M. A. (2010) The genome analysis toolkit: A mapreduce framework for analyzing next-generation dna sequencing data. *Genome Research*, **20**, 1297–1303.

Miller, J. R., Koren, S. and Sutton, G. (2010) Assembly algorithms for next-generation sequencing data. *Genomics*, **95**, 315–327.

Minichiello, M. J. and Durbin, R. (2006) Mapping trait loci by use of inferred ancestral recombination graphs. *Am J Hum Genet*, **79**, 910–922.

Morton, N., Zhang, W., Taillon-Miller, P., Ennis, S., Kwok, P. and Collins, A. (2001) The optimal measure of allelic association. *Proceedings of the National Academy of Sciences*, **98**, 5217.

Nielsen, R., Paul, J. S., Albrechtsen, A. and Song, Y. S. (2011) Genotype and snp calling from next-generation sequencing data. *Nature Reviews Genetics*, **12**, 443–451.

Ning, Z., Cox, A. J. and Mullikin, J. C. (2001) Ssaha: A fast search method for large dna databases. *Genome Research*, **11**, 1725–1729.

Nordborg, M. and Tavaré, S. (2002) Linkage disequilibrium: what history has to tell us. *Trends Genet*, **18**, 83–90.

Novocraft (2009) Novoalign.

Pasaniuc, B., Avinery, R., Gur, T., Skibola, C. F., Bracci, P. M. and Halperin, E. (2010) A generic coalescent-based framework for the selection of a reference panel for imputation. *Genet. Epidemiol.*, **34**, 773–782.

Quinlan, A. R., Stewart, D. A., Stromberg, M. P. and Marth, G. T. (2008) Pyrobayes: an improved base caller for snp discovery in pyrosequences. *Nat Meth*, **5**, 179–181.

Sanger, F., Nicklen, S. and Coulson, A. R. (1977) Dna sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America*, **74**, 5463–5467.

Schäfer, J. and Strimmer, K. (2005) A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical applications in genetics and molecular biology*, **4**, 32.

Scheet, P. and Stephens, M. (2006) A Fast and Flexible Statistical Model for Large-Scale Population Genotype Data: Applications to Inferring Missing Genotypes and Haplotypic Phase. *The American Journal of Human Genetics*, **78**, 629–644. URL `http://dx.doi.org/10.1086/502802`.

Schork, N. J., Murray, S. S., Frazer, K. A. and Topol, E. J. (2009) Common vs. rare allele hypotheses for complex diseases. *Current opinion in genetics & development*, **19**, 212–219.

Shendure, J. and Ji, H. (2008) Next-generation dna sequencing. *Nature Biotechnology*, **26**, 1135–1145.

Shendure, J., Porreca, G. J., Reppas, N. B., Lin, X., McCutcheon, J. P., Rosenbaum, A. M., Wang, M. D., Zhang, K., Mitra, R. D. and Church, G. M. (2005) Accurate multiplex polony sequencing of an evolved bacterial genome. *Science*, **309**, 1728–1732.

Simpson, J. T., Wong, K., Jackman, S. D., Schein, J. E., Jones, S. J. and Birol, I. (2009) Abyss: a parallel assembler for short read sequence data. *Genome research*, **19**, 1117–1123.

Stephens, M. and Scheet, P. (2005) Accounting for decay of linkage disequilib-

rium in haplotype inference and missing-data imputation. *American journal of human genetics*, **76**, 449–462.

Stephens, M., Smith, N. J. and Donnelly, P. (2001) A new statistical method for haplotype reconstruction from population data. *American journal of human genetics*, **68**, 978–989.

The 1000 Genomes Project Consortium (2010) A map of human genome variation from population-scale sequencing. *Nature*, **467**, 1061–1073.

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

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

The Wellcome Trust Case Control Consortium (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, **447**, 661–678.

Tzeng, J.-Y., Devlin, B., Wasserman, L. and Roeder, K. (2003) On the identification of disease mutations by the analysis of haplotype similarity and goodness of fit. *Am. J. Hum. Genet.*, **72**, 891–902.

Wakeley, J. (2008) *Coalescent Theory: An Introduction*. Roberts & Company Publishers, 1 edn.

Walsh, T., McClellan, J. M., McCarthy, S. E., Addington, A. M., Pierce, S. B., Cooper, G. M., Nord, A. S., Kusenda, M., Malhotra, D., Bhandari, A., Stray, S. M., Rippey, C. F., Roccanova, P., Makarov, V., Lakshmi, B., Findling, R. L., Sikich, L., Stromberg, T., Merriman, B., Gogtay, N., Butler, P., Eckstrand, K., Noory, L., Gochman, P., Long, R., Chen, Z., Davis, S., Baker, C., Eichler, E. E., Meltzer, P. S., Nelson, S. F., Singleton, A. B., Lee, M. K., Rapoport, J. L., King, M.-C. and Sebat, J. (2008) Rare structural variants disrupt multiple genes in neurodevelopmental pathways in schizophrenia. *Science*, **320**, 539–543.

Wang, J., Wang, W., Li, R., Li, Y., Tian, G., Goodman, L., Fan, W., Zhang, J., Li, J., Zhang, J., Guo, Y., Feng, B., Li, H., Lu, Y., Fang, X., Liang, H., Du, Z., Li, D., Zhao, Y., Hu, Y., Yang, Z., Zheng, H., Hellmann, I., Inouye, M., Pool, J., Yi, X., Zhao, J., Duan, J., Zhou, Y., Qin, J., Ma, L., Li, G., Yang, Z., Zhang, G., Yang, B., Yu, C., Liang, F., Li, W., Li, S., Li, D., Ni, P., Ruan, J., Li, Q., Zhu, H., Liu, D., Lu, Z., Li, N., Guo, G., Zhang, J., Ye, J., Fang, L., Hao, Q., Chen, Q., Liang, Y., Su, Y., San, A., Ping, C., Yang, S., Chen, F., Li, L., Zhou, K., Zheng, H., Ren, Y., Yang, L., Gao, Y., Yang, G., Li, Z., Feng, X., Kristiansen, K., Wong, G. K., Nielsen, R., Durbin, R., Bolund, L., Zhang, X., Li, S., Yang, H. and Wang, J. (2008) The diploid genome sequence of an asian individual. *Nature*, **456**, 60–65.

Warren, R. L., Sutton, G. G., Jones, S. J. M. and Holt, R. A. (2007) Assembling millions of short dna sequences using ssake. *Bioinformatics*, **23**, 500–501.

Wen, X. and Stephens, M. (2010) Using linear predictors to impute allele fre-

quencies from summary or pooled genotype data. *Annals of Applied Statistics 2010, Vol. 4, No. 3, 1158-1182*.

Wheeler, D. A., Srinivasan, M., Egholm, M., Shen, Y., Chen, L., McGuire, A., He, W., Chen, Y.-J., Makhijani, V., Roth, G. T., Gomes, X., Tartaro, K., Niazi, F., Turcotte, C. L., Irzyk, G. P., Lupski, J. R., Chinault, C., Song, X.-z., Liu, Y., Yuan, Y., Nazareth, L., Qin, X., Muzny, D. M., Margulies, M., Weinstock, G. M., Gibbs, R. A. and Rothberg, J. M. (2008) The complete genome of an individual by massively parallel dna sequencing. *Nature*, **452**, 872–876.

Wright, S. (1931) Evolution in mendelian populations. *Genetics*, **16**, 97–159.

Yu, K., Gu, C., Province, M., Xiong, C. and Rao, D. (2004) Genetic association mapping under founder heterogeneity via weighted haplotype similarity analysis in candidate genes. *Genetic epidemiology*, **27**, 182–191.