

The detection, structure and uses of extended haplotype identity in population genetic data



Dionysia Kiara Xifara

Department of Statistics
University of Oxford

Supervisor: Professor Gil McVean

*Thesis Submitted in Partial Fulfilment of the Requirements for the
Degree of Doctor of Philosophy
Hilary 2014*

Acknowledgements

This work would not have been possible without the guidance and care of my supervisor, Gil McVean. Thank you Gil for your dedicated advice and constant support.

I would further like to thank Ele Zeginni and her lab for a fruitful collaboration and Chris Holmes, Jonathan Marchini, Simon Myers and Chris Spencer for invaluable comments on my research.

This thesis is dedicated to my parents and my family; their love and encouragement over the years underpins its pursuit and completion.

Abstract

In large-scale population genomic data sets, individual chromosomes are likely to share extended regions of haplotype identity with others in the sample. Patterns of local haplotype sharing can be highly informative about many processes including historical demography, selection and recombination. However, in outbred diploid populations, the identification of extended shared haplotypes is not straightforward, particularly in the presence of low levels of genotyping error. Here, we introduce a model-based method for accurately detecting extended haplotype sharing between sets of individuals from unphased data. We describe two implementations of the algorithm that can be applied to data sets consisting of thousands of samples. The method leads naturally to an approach for statistical haplotype estimation, which is shown to be comparable in accuracy to current methods. By applying the method to genome-wide SNP data from over 5,000 samples from the UK we show that the N50 maximal haplotype sharing between unrelated samples is typically 2cM, consistent with a population history of rapid exponential growth that started ~ 125 generations ago. In contrast, within two Greek population isolates of ~ 700 individuals the N50 for maximal haplotype sharing is 12.5cM, while for an unrelated Greek sample of the same size the N50 is 1.3cM. By assessing the size and geographical distribution of maximal haplotype sharing within and between all Greek cohorts, we make inference on the extent of isolatedness of each cohort and on recent migration. We additionally date recent coancestry to about 10 generations for the isolates and 90 generations for the unrelated sample, and finally attempt to date the time of divergence between them.

Contents

List of Figures	ix
1 Introduction	1
1.1 What is relatedness and how do we measure it?	2
1.2 Identity by descent	7
1.2.1 Methods: using a length threshold	10
1.2.2 Methods: using haplotype frequency thresholds	11
1.2.3 Probabilistic methods	11
1.3 Applications of relatedness: Statistical haplotype estimation . . .	13
1.3.1 Family data	14
1.3.2 Stochastic estimation	16
1.3.3 Using IBD	18
1.3.4 The coalescent and HMMs	19
1.3.5 Influential variables	23
1.4 Applications of relatedness: Demographic inference	25
1.5 Applications of relatedness: Genome-wide association studies . . .	27
1.6 Long Range Phasing	31
1.7 Nearest Neighbours	33
1.7.1 The Li and Stephens model	36
1.8 This thesis	38

2	Model and Methodology	42
2.1	Method Motivation	42
2.2	The Hidden Markov Model	44
2.2.1	Transition Probabilities	44
2.2.2	Emission Scores	45
2.2.3	The Viterbi Algorithm - Identifying and Using Pairs of In- dividuals	50
2.3	Application 1: Haplotype Estimation	51
2.3.1	The Second best and other paths	51
2.3.2	Phasing with a path of pseudo-parents	56
2.3.3	Combining path predictions	60
2.3.3.1	Other attempted methods	61
2.4	Application 2: Nearest Neighbours and haplotype sharing	65
2.4.1	The Forward/Backward Algorithms	66
2.4.2	The Graph	67
2.4.2.1	Constructing a genotype graph	68
2.4.2.2	Using a genotype graph to select pseudo-parents	69
2.4.2.3	Subsampling	76
2.4.3	An HMM with Subsampling	82
2.5	Model Summary	89
3	Statistical haplotype estimation	93
3.1	Overview	93
3.2	Measuring accuracy	94
3.3	Building a phaser	97
3.3.1	Simulated Data	97
3.3.2	Real Parents	98
3.3.3	Picking the emission scores	100
3.3.4	The effect of suboptimality	103
3.3.5	Managing suboptimality	106
3.3.5.1	Other schemes for combining information across paths	107
3.3.6	Phasing all sites	113

3.3.7	PACE	115
3.3.8	Remarks	116
3.4	Phaser competition	116
3.4.1	Data	117
3.4.2	Results	119
3.4.3	Conclusions	120
4	Finding Nearest Neighbours: detection and haplotype sharing	123
4.1	Overview	123
4.2	Limits on the ability to detect Nearest Neighbours by means of haplotype sharing	126
4.2.1	Simulations under the coalescent	127
4.2.2	Simulations under an exponential growth model	130
4.2.3	Results	132
4.3	A statistical method for detecting NNs and extended haplotype sharing from unphased genotype data	139
4.3.1	Measuring accuracy	141
4.3.2	Simulated Data	143
4.3.3	Results	145
4.4	Discussion	156
5	Applications to real data: HELIC	158
5.1	Objectives	160
5.2	Methods	162
5.2.1	The decay of coancestry using carrier status	166
5.2.2	Another approach for estimating the effective population size and the TMRCA between NNs	168
5.2.2.1	Estimating the TMRCA from Simulated Data	169
5.3	Data and quality control	173
5.4	Results: <i>APOC3</i>	173
5.5	Results: A genome-wide analysis of relatedness, demography and isolatedness	178

5.6	Dating the time of divergence	186
6	Applications to real data: WTCCC	189
6.1	A prequel: 1000 Genomes Omni SNP genotypes	191
6.1.1	Results	193
6.2	The structure of haplotype identity in a sample of over 5000 UK individuals	198
6.2.1	WTCCC: data, aims and methods	198
6.2.2	Sample size, thinning, error and a first look at haplotype sharing	201
6.2.3	Genome-wide analysis: results using the graph and HMMsub	206
7	Discussion	219
7.1	A few final remarks on phasing	222
7.2	A few final remarks on relatedness in populations	224
7.3	Outlook: Disease Risk	228
	Bibliography	230

List of Figures

1.1	A Pseudo Family	33
1.2	Nearest neighbours	35
1.3	An illustration of the Li and Stephens imperfect mosaic model . .	37
2.1	Emission probabilities	48
2.2	Building a genotype graph	70
2.3	Using the graph to select pseudo-parents.	73
2.4	Pruning the graph to select pseudo-parents	75
2.5	Outline of the graph method	81
2.6	Subsampling for an HMM	88
2.7	The HMM	91
3.1	Measuring phasing accuracy	96
3.2	The effect of path Likelihood.	104
3.3	Calibration plots for SM	108
3.4	Calibration plots for all other path-combining algorithms	109
3.5	Calibration counts for TTB	111
3.6	Phasing all sites: Calibration	114
3.7	PACE	115
3.8	Method comparisons.	121
4.1	Adding recombination distances to a coalescent tree	130
4.2	Constant population size: NN vs NN_p	133
4.3	Exponential growth: NN vs NN_p	134

LIST OF FIGURES

4.4	Identifying NNs under the coalescent with exponential growth . . .	135
4.5	Differences in TMRCA between NN and NN_p	137
4.6	Differences in shared haplotype lengths between NN and NN_p for haplotype lengths	138
4.7	Measuring average sharing	143
4.8	Haplotype sharing and finding NNs	151
4.9	Detected shared haplotype chunks	152
4.10	TMRCA with a NN vs NN_p for $\rho = 2.5\text{cM/Mb}$	153
4.11	TMRCA with a NN vs NN_p for $\rho = 2.5\text{cM/Mb}$ and 1% genotyping error.	154
5.1	Estimating the TMRCA from simulated data	172
5.2	R19X N50	174
5.3	Excess relatedness with carriers of R19X	176
5.4	Excess relatedness with carriers of another variant	177
5.5	Haplotype sharing in the isolates and TEENAGE	180
5.6	An 80-20 study of the isolates and TEENAGE	181
5.7	Choosing NNs across cohorts	182
5.8	Haplotype sharing among cohorts	183
5.9	A local TMRCA	184
5.10	TMRCA with a NN	185
5.11	Estimating time of divergence	187
6.1	Comparing the three methods on 10 individuals	195
6.2	Comparing H300 and H1000 on 100 individuals	196
6.3	Where do the NNs change?	197
6.4	Haplotype Sharing on Chr12	204
6.5	Maximal Haplotype Sharing.	205
6.6	Observed and expected haplotype chunks for different sample sizes. .	206
6.7	Haplotype Sharing across chromosomes.	209
6.8	Comparing the per chromosome N50 for H300 and the graph method. .	210
6.9	The TMRCA to a NN.	211
6.10	Physical haplotype chunk sizes point-wise across chromosomes us- ing the graph method.	213

LIST OF FIGURES

6.11 Physical haplotype chunk sizes point-wise across chromosomes using HMMsub with 300 SNPs.	214
6.12 Physical haplotype chunk sizes point-wise across chromosomes using HMMsub with 1000 SNPs.	215

Chapter 1

Introduction

In the early 20th century, humans' increased understanding of natural selection, mutation and recombination revolutionised our concepts of genetic relatedness. Moving far beyond basic models of Mendelian transmission, theoretical tools such as the Wright-Fisher model (Fisher [1930], Wright [1931]) and its extension the coalescent (Kingman [1982b]) have empowered us to characterize and capture properties of genetic variation.

As it became possible to genotype individuals experimentally, population genetics necessarily evolved from a theoretical field into one incorporating data science. The development of sequencing platforms, such as next generation sequencing, has only enhanced the need for computational approaches to data analysis. The majority of these approaches implicitly incorporate a model that captures the relationship among individuals in a sample.

Aside from such applications that focus on data analysis, more generally, understanding relatedness is important for two main reasons. First, it provides insight into a population's demographic history; it determines whether a population has experienced migration or changes in its effective population size (Palamara et al. [2012], Palamara and Pe'er [2013], Harris and Nielsen [2013]), for instance. Second, relatedness patterns within a sample are a confounder in disease association studies and can lead to spurious findings if not incorporated in the model

testing candidate mutations (Price et al. [2006], Wellcome Trust Case Control Consortium et al. [2007], International Multiple Sclerosis Genetics Consortium et al. [2011], Mathieson and McVean [2012]). A large part of the literature is therefore dedicated to investigating methods that detect relatedness in a way that allows for it to be accommodated in association tests.

1.1 What is relatedness and how do we measure it?

So, what is relatedness? Perhaps unsurprisingly, it does not appear like there is an obvious consensus on how best to define and measure it. Between two individuals, an intuitive approach to defining it is via the concept of identity by descent (IBD). Two genes at a locus are IBD if they have been inherited from a common ancestor and are identical. More closely related individuals will tend to have a higher proportion of their genome in IBD. IBD contrasts with identity by state (IBS), which arises when two genes are the same by chance: when their common ancestor lived too long ago. We can use the IBD approach to define relatedness between two individuals (A and B , say) as the kinship coefficient: the probability that a randomly selected gene they carry is IBD. This concept was pioneered in the early work of Cotterman [1940] and Malécot [1948]. If $A = B$, it follows that the kinship coefficient is $\frac{1}{2} + f(A)$, where $f(A)$ is the inbreeding coefficient for individual A (the probability that the two copies of A , inherited from the mother and the father, are IBD). Kinship coefficients can be explicitly defined for close relationships, such as for siblings, uncles and nieces and others (for examples see Balding et al. [2008], Volume 2, Part 5, Chapter 23). For more distant pairwise relationships, it is possible to obtain explicit expressions for kinship by recursively considering possible pedigrees describing their relationship; this becomes an increasingly harder problem as it is often easy to ignore certain routes up a pedigree back to the common ancestor. Kinship coefficients are therefore typically obtained via simulation. For two diploid individuals, there are 4 versions of the gene in question (2 for each person) and there are therefore 15 identity states. When we do not distinguish between copies of the gene carried

1.1 What is relatedness and how do we measure it?

by the same person (for example, when only genotype data are available) the number of identity states collapses to 9 (refer to Balding et al. [2008] for more details).

Of course, it is useful to consider the notion of kinship in the context of a pedigree or an unrelated sample, rather than in a pairwise framework. Cockerham [1971] extended kinship to up to four individuals, and Thompson [1974] up to five. Karigl [1981] generalised the concept to n individuals by considering the possible IBD-state partitions of the n genes, defining a recurrence relationship for the calculation.

Kinship is a useful metric (Kang et al. [2008], Zhang et al. [2009], Rakovski and Stram [2009], Vilhjalmsen [2012], Hoffman [2013], Pirinen et al. [2013]) that follows directly from Mendel's laws of inheritance, but it is by no means the only summary statistic that can describe the relationship between individuals. Hamming distance (Hamming [1950]), which looks at the proportion of sites in IBS between individuals in a region, is another measure for population diversity (Morrison and De Jong [2002]) and is often used to detect more closely related individuals (Deza and Deza [2009], Blackburne and Whelan [2012], Howie et al. [2011], Howie et al. [2012], Zhang et al. [2013]).

An alternative approach to capturing relatedness patterns is Principle Component Analysis (PCA). PCA uses genome-wide information to project samples into orthogonal axes and the distance between samples in PCA space is directly linked with the TMRCA between them (McVean [2009]). PCA is therefore a summary of genotype covariance and a measure of relatedness. The first few principle components are typically used to visually identify structure in a population and are also used as covariates in disease association studies in order to account for any cryptic relatedness present; we discuss this further in Section 1.5.

These metrics are useful in many settings but are not informative of the full genealogy of a sample under study. Nor do they capture the changing relationship between individuals along the genome. Further, they often fail to distinguish between various evolutionary processes that give rise to the variation that they are summarising: there are multiple demographic histories that plausibly give rise to the same summary statistics and the same distribution of TMRCAs and therefore also the same distances between individuals in PCA space.

1.1 What is relatedness and how do we measure it?

In contrast, the most widely adopted modelling framework for describing relatedness within any sample of individuals is the coalescent (reviewed for example in Balding et al. [2008]) and its extensions. First proposed by Kingman (Kingman [1982a,b,c]), the coalescent extends the Wright Fisher model (Fisher [1930], Wright [1931]) to reconstruct a sample's genealogy. The model is a continuous-time Markov process typically going backwards in time, tracing the coalescence events between individuals. In each step, two randomly selected individuals (from the k that are present, say, at any one time) coalesce after an exponential waiting time; the parameter of the waiting time is $\binom{k}{2}$. The process thus builds the genealogical tree of the sample, timing each of the events, and continues until there is a single ancestor left in the sample. For a full derivation of the coalescent from the Wright-Fisher model, refer to Balding et al. [2008]. Mutations can then be added onto the genealogy using a Poisson Process, with the number of mutations on each branch of the tree being Poisson distributed with parameter depending on the branch length (b_L) and the mutation rate, θ ($\frac{\theta}{2} \cdot b_L$).

The model can be further extended to incorporate variable population size: coalescent time is derived by rescaling the discrete time of the Wright-Fisher model by the effective population size (N_e). Therefore incorporating a changing population size is straightforward, simply by rescaling according to the population size at any given time. Intuitively, this follows from the fact that the process of tracing ancestry in a genealogy does not change when sample size changes. Instead, individuals are more likely to coalesce when the population size is small, and therefore only the rate of coalescence is affected over time. As an example, the effect of exponential growth would be an increasing coalescence rate going backwards in time, with the individuals nearer the present taking the longest to find an ancestor. A related generalisation of the model is allowing for structure, which can be accommodated by controlling the probability with which different combinations of individuals may coalesce; this is called the structured coalescent. More details on the generalisation of the coalescent for variable population size can be found in Balding et al. [2008], Donnelly and Tavaré [1995], Marjoram and Donnelly [1997] and for structure in Herbots [1994], Notohara [1990], Wilkinson-Herbots [1998], Rousset [1999].

1.1 What is relatedness and how do we measure it?

The coalescent can be further extended to incorporate recombination (Hudson [1983], Griffiths and Marjoram [1996, 1997]), using the concept of the ‘ancestral recombination graph’ (ARG). Recombination splits the ancestry of a lineage into two and so at any given time in the coalescent process, the number of lineages present either decreases by 1 (if coalescence occurs) or increases by 1 (if recombination occurs). The process is still a Markov process, but is now a birth-death process. Overall, multiple recombination events split the ancestry of a sample over multiple chromosome lineages. At any given time, each lineage present is considered as before, regardless of whether it has risen from recombination. At each position there is a single unambiguous marginal genealogical tree, with the coalescent modelling the relationship between them. These marginal trees are embedded in the ARG.

In all, the coalescent and its extensions allow us to investigate the factors that influence the relationship between individuals, such as mutation, recombination and selection. Crucially, and more relevant to the discussion in this thesis, it provides a probabilistic framework that allows us to use simulations to make likelihood-based inference on these relationships and how they change along the genome. The ability to use simulations is key; in practice little inference is exact since it is either difficult to derive closed-form expressions for the metrics under investigation or it is computationally intractable to calculate them. The coalescent can more generally be used to understand the historical relationship between subpopulations. We can similarly make inference on ancestry, for example in admixed populations.

This provides us with alternative definitions of relatedness (using summary statistics from the coalescent trees of a sample), such as the time to the most recent common ancestor (TMRCA) between pairs of individuals at a SNP. The effect of recombination means that there is no single TMRCA between any two individuals; rather, this is a relationship that changes at each position along the genome. Nevertheless, we can distinguish individuals that are closer on any given marginal genealogical tree, or that are Nearest Neighbours (NNs; individuals between which the TMRCA is minimal compared to the TMRCA they share with other individuals. This is more formally discussed in Section 1.6). Therefore, the coalescent allows us (in theory; practical implementations are often a concern)

1.1 What is relatedness and how do we measure it?

to unambiguously identify and date the topological relationship between all individuals in a sample, while tracing how it changes along the genome. In other words, it provides a framework for measuring and characterizing relatedness.

However, these concepts do not directly contain information about the amount of allele sharing between individuals. An alternative approach to defining relatedness, which merges the concept of kinship with the coalescent is via haplotype sharing. Haplotype sharing occurs when two individuals carry the same alleles on the same single chromosome at consecutive SNPs. More closely related individuals in a sample will share greater proportions of their genome. Moreover, this sharing will occur in chunks. Average sharing across the entire sample is a measure of relatedness and can therefore be informative of recent demographic history (Lawson et al. [2012], Palamara et al. [2012]). It can also be useful in the imputation of variants from a reference panel (imputation is the statistical estimation of missing variants in a data set; Howie et al. [2009]) and (by identifying where the breaks in extended haplotype sharing occur) for the study of recombination rate variation (Hinch et al. [2011], Wegmann et al. [2011]). Measures across sub-populations may be a measure of divergence (Ralph and Coop [2013]), though it would be necessary to calibrate this as a metric so that absolute values of sharing can be interpretable with respect to divergence. Moreover, locally shared haplotype chunks across cases of a particular disease compared to controls can help narrow down regions containing causal loci for the disease, though this would typically be effective for highly penetrative variants with large effects that explain a substantial proportion of the phenotypic variance.

In an idealised situation, where haplotypes are known perfectly, extended haplotype sharing can be identified through model-based frameworks, for example using the hidden Markov model (HMM) approaches underlying BEAGLE (Browning and Browning [2011a]; we discuss this model more formally in Section 1.3. We also briefly introduce HMMs in 1.3.4). However, in outbred diploid populations, detection of extended haplotype identity is complicated by the unphased nature of the data and the presence of genotype errors. Currently, many algorithms that use extended haplotype sharing, such as the analysis of fine-scale population structure (Lawson et al. [2012]) and statistical imputation (Howie

et al. [2012]), typically use a two-stage approach of statistical haplotype estimation (with imputation of missing data) followed by analysis.

1.2 Identity by descent

The classic approach to detecting haplotype sharing is using the concept of IBD which asks whether regions that are shared between individuals do so because they have been inherited from a common ancestor rather than by chance. IBD has been extensively used in the literature for the aforementioned applications of haplotype sharing (Browning and Browning [2012] and others) and it has been shown to capture relatedness patterns more accurately (for example by detecting shorter shared haplotype segments) than classic IBS-based approaches (Browning and Browning [2011a]). Of course, individuals share an ancestor at every position and it is therefore required to define a historical threshold before which common inheritance is no longer assumed to be IBD.

We briefly defined IBD when introducing kinship coefficients, and here we extend the discussion to describe IBD over a region rather than at a single locus. The probabilities of two individuals carrying the same allele at a locus by descent which we discussed previously were introduced by Wright (Wright [1921]) but IBD for a genomic segment was first quantified by Malécot [1948] by averaging such probabilities across the genome. This is an oversimplification of inheritance patterns, however, as linkage and linkage disequilibrium dictate that whenever one locus is in IBD it is highly likely that its surrounding segment is in IBD as well. IBD is therefore typically considered in chunks and two haplotypes are said to be in IBD if they carry the same set of alleles that have been inherited from a common ancestor. If alleles are the same but have not been inherited from a common ancestor then we have IBS. The length of the shared IBD tract depends on the number of meioses that have occurred since the common ancestor and therefore the proportion of the genome shared decreases exponentially (on average) as the relationship becomes more distant.

The driving assumption underlying IBD detection is that, given that the haplotypes are relatively rare, it is unlikely that two individuals will be carrying

them by chance. When this is true, IBD can be identified in three different ways (Browning and Browning [2012]):

1. Using the length of the segment that is shared: if two haplotypes share a long region at which they are (almost, allowing for some error) identical, it is unlikely that this will have occurred by chance and therefore IBD can be deduced,
2. Using haplotype frequency: when haplotypes with frequency below some predetermined threshold are shared they are said to be in IBD, again because the probability that this occurs by chance is negligible,
3. Using a population genetics model to directly calculate the probability that two haplotypes are in IBD.

When might we come across the first two conditions? Under the standard neutral coalescent model, individuals share an ancestor at every position along the genome and therefore every position can be said to be in IBD (if the individuals carry the same alleles) so long as there is no time constraint to the allowed common ancestry. This is ancient IBD and is typically a harder property to detect in light of the large number of mutations that will have aggregated over such time scales and the larger number of recombination events that will have disrupted any shared IBD tracts.

Consequently, for the case of humans, between completely unrelated samples IBD is typically expected to be detectable around the loci at which the time to their common ancestor is as recent as 25-30 generations ago (Browning and Browning [2011a, 2012]), in which case the length of their shared tracts is ~ 1.5 - 2 cM (Browning and Browning [2010]). In practice the false positive rate for tracts below 1 - 1.5 cM is too high (Browning and Browning [2011a]), and the expected sharing should also be adjusted for a variable effective population size. It is often the case therefore that for samples where recent coalescence events are rare, there is too little information for IBD detection to be possible unless the sample size is very large (Palamara et al. [2012]). This latter scenario can only be handled by methods that are computationally very efficient, like PLINK (Purcell et al.

[2007]), GERMLINE (Gusev et al. [2011]) and RefineIBD (Browning and Browning [2013a]). Moreover, though IBD tracts occur rarely in unrelated samples, and while they might not be very long, across large samples their sum might lead to considerable aggregate IBD. As such it is in principle possible to therefore use it as a metric of overall relatedness within and between samples and populations, hence enabling its applicability to detecting ancestry and making inference on demographic history (Ralph and Coop [2013]).

In general, the proportion of the genome shared by two individuals separated by m meioses is $2^{-(m-1)}$, but the length of an IBD tract is exponentially distributed with mean $\frac{100}{m}$ cM. Therefore between completely unrelated samples very little IBD is expected to exist; when it does, however, it will typically consist of long IBD chunks. For example, between fifth cousins there is usually no detectable IBD (the proportion of their genomes shared is 0.00049), but when IBD is detected it will consist of single IBD tracts of mean length 8.3cM (Browning and Browning [2012]).

The exact time scale at which IBD is detectable is defined by the sizes of the chunks that are identified by the method being used. If a method can detect up to 2cM tracts, then from the above it follows that its limit for IBD detection is ~ 25 generations. IBD segments are therefore a special case of shared haplotype chunks: they do not capture the full length distribution of tracts that are shared in a population. They further assume a theoretical, historical cut-off point past which coancestry is not captured (Browning and Browning [2012], Harris and Nielsen [2013]), whereas haplotype sharing may still exist between completely unrelated individuals and may be indicative of their TMRCA (Harris and Nielsen [2013]). Nevertheless, they are currently the most popular equivalent of haplotype sharing in the literature and have numerous useful applications, that are reviewed in Browning and Browning [2012] and which we present briefly in a subsequent subsection.

IBD detection methods are often haplotype-based requiring the statistical estimation of haplotypes prior to analysis. The presence of phasing errors significantly decreases power to detect IBD segments (Browning and Browning [2012], Palamara and Pe'er [2013]). Power is critically linked to genotyping errors as well (Browning and Browning [2012]), and therefore stringent data quality control is

essential even for methods that allow some discordance to exist within shared IBD chunks.

1.2.1 Methods: using a length threshold

IBD consists of genomic regions which are identical. A straightforward approach to detecting such candidate regions is by searching for IBS in regions that are too long for IBS to be occurring by chance. IBS in haplotypes occurs when the alleles carried are identical and for haplotype-based methods identification of IBD is simply a question of defining a threshold beyond which IBS haplotypes are called IBD. One method that takes this route is GERMLINE (Gusev et al. [2009, 2011]). GERMLINE uses a sliding window with a hash table to find matching haplotypes. The width of the window is inversely proportional to the phase and genotyping errors (in other words, the mismatches) that the method is allowed to tolerate. For one mismatch per window the width will be ≈ 100 SNPs. GERMLINE is very computationally efficient, yet it is a heuristic approach and more successful approaches have used it as a starting method before using probabilistic models (Browning and Browning [2013a]). Price et al. [2011] exploit the long-range phased data from the Icelandic population in order to search for IBD segments. They split the genome in 2cM chunks and call the segments IBD if they are identical at $>95\%$ of sites (from a 4-way comparison of haplotypes of two individuals), non-IBD if they match at $<85\%$ and unknown otherwise.

Methods following this approach need not be haplotype based. In fact, several methods have been proposed that consider length thresholds on genotypes (Nelson et al. [2006], Miyazawa et al. [2007], Kong et al. [2008], Leibon et al. [2008], Thomas et al. [2008], Thomas [2010]) all of which set these thresholds heuristically. For a genotype, a locus between two individuals is always IBS unless they carry discordant homozygotes, i.e. if they are incompatible. We discuss incompatibility in Chapter 2. For common SNPs, the frequency of incompatible homozygotes is rare and therefore a very large number of SNPs in IBS are required for IBD to be called confidently. This is further enhanced by LD, which

increases the probability of seeing segments that are consistently IBS or consistently not in IBS. The heuristic threshold for length therefore depends on the trade-off between sensitivity and specificity to IBD tracts: an equal weight is equivalent to about 3cM (Miyazawa et al. [2007]), while 10cM will achieve high specificity rates (Kong et al. [2008]).

1.2.2 Methods: using haplotype frequency thresholds

Similarly to imposing length thresholds, it is equivalent to consider a threshold on haplotype frequency: when rare haplotypes are in IBS it is unlikely that they are shared by chance and may therefore be determined to be IBD. One method following this approach is BEAGLE fastIBD (Browning and Browning [2011a]), which uses the BEAGLE LD structure and is an extension of the probabilistic method BEAGLE IBD. FastIBD uses a sliding window of hash tables, using multiple estimates of an individual's haplotypes in order to accommodate the possibility of phasing errors. The threshold for the haplotype frequency for calling IBD is user-specified.

Both of these approaches for calling IBD are heuristic and will thus suffer with identifying short IBD segments ($<2\text{cM}$). It is therefore preferable to consider probabilistic approaches that are model-based.

1.2.3 Probabilistic methods

Model-based methods for detecting IBD typically employ Hidden Markov Models (HMM; reviewed in Rabiner [1989] and briefly introduced in Section 2.2). The hidden states are the IBD status and the number of states depends on whether the method is haplotype- or genotype-based. Haplotype-based methods will have two states: either a locus is IBD or it is not. Equivalently, genotype-based methods have two states (IBD or non-IBD) or three states, one for each of the possible the number of alleles shared by the two individuals (0, 1 or 2). The exceptions to this setup are models that may have more hidden states for Jacquard's nine or fifteen

IBD states for phased and unphased data respectively. Moreover, probabilistic methods may or may not incorporate LD in their modelling framework.

For methods that do not consider LD (for example Brown et al. [2012], Huff et al. [2011], Purcell et al. [2007], Thompson [2008]), variable recombination rates are accommodated in their modelling framework by measure distances on the genetic scale (cM). However, accuracy can be improved by considering LD and therefore numerous methods include it (for instance Genovese et al. [2010]). One accurate approach is BEAGLE IBD (Browning and Browning [2010]) which is the extension of the BEAGLE software (Browning [2006], Browning and Browning [2007], Browning [2008a]) to incorporate LD. Although this is a very accurate method it is not computationally efficient enough to be applicable to thousands of samples. In contrast RefineIBD (Browning and Browning [2013a], which combines an initial pass using GERMLINE (Gusev et al. [2009, 2011]) followed by a similar approach to fastIBD (Browning and Browning [2011a]), is a computationally efficient and highly accurate algorithm (Browning and Browning [2013a]). Alternatively, LD can be incorporated for example by conditioning on the IBD status not just of the current locus but of nearby SNPs as well (e.g. RELATE, Albrechtsen et al. [2009]). IBDLD (Han and Abney [2011]) achieves greater accuracy by extending this approach and conditioning on twenty neighbouring SNPs.

Hardy-Weinberg equilibrium is typically assumed in all the aforementioned methods, so that allele frequencies can be used to obtain genotype probabilities, which in turn are used to calculate the probability of a site-specific IBD status. To the author's knowledge, none of the IBD approaches incorporate selection in their models. This can be problematic under certain conditions as sites that are selected can appear to be in IBD when in fact they are not (Albrechtsen et al. [2010]).

So far, the approaches that have been discussed consider IBD between pairs of samples. In a probabilistic framework it is possible to search for IBD by considering multiple samples simultaneously. This increases power to detect short IBD segments as, for a query pair of individuals, we can use information from other individuals that are IBD with it. One approach in this direction is to consider IBS across all the individuals in the sample, which can be considered both at the haplotype level (for example Houwen et al. [1994]) and the genotype level (for

1.3 Applications of relatedness: Statistical haplotype estimation

example Leibon et al. [2008], Thomas et al. [2008], Hansen et al. [2009], Moltke et al. [2011]). These methods are typically slower, being computationally intractable for large samples. Indicatively, MCMC-IBDfinder (Moltke et al. [2011]) has only been run on 15 individuals. In situations where pedigree information for the sample is known, familial relationships can be used to detect multi-individual IBD status and thus the process may speed up considerably (for example ALADIN, Albers et al. [2008]). In ALADIN, a clustering approach (Abecasis and Wigginton [2005], Browning et al. [2004]) allows for LD to be incorporated within haplotype blocks but not between them.

As previously discussed, IBD tract sharing is a measure of relatedness and the IBD segments can be used to make inference on numerous applications of population genetics for which relatedness is relevant. We review these applications here (following Browning and Browning [2012]).

1.3 Applications of relatedness: Statistical haplotype estimation

Haplotype-based inference is typically more powerful than genotype-based analysis; knowledge of haplotypes allows for more robust inference on selection and recombination patterns (see Sabeti et al. [2007] and Fearnhead and Donnelly [2001], respectively), for example. Haplotype-based analyses as a result can sometimes be more informative than individual markers in association studies (for example Akey et al. [2001]) and, in particular, as shown by Schaid [2004], consideration of clustering of haplotype data is even more powerful regarding low frequency variants. Similarly, multilocus analyses suitable for whole-genome association studies should effectively account for the LD structure present in the data, further highlighting the necessity for use of haplotypes. However, experimental determination of haplotypes is expensive and time consuming. Instead, haplotypes are typically estimated statistically from genotypes (Marchini et al. [2006], Browning and Browning [2011b]).

Haplotype estimation involves the determination of a sequence of alleles that are carried on the same chromosome from genotype data. In this thesis we will

1.3 Applications of relatedness: Statistical haplotype estimation

assume that all loci are bi-allelic and that therefore the genotypic make-up of an individual at each locus counts the number of derived alleles that it carries there: 0 if it has two copies of the ancestral allele, 1 if it is heterozygous and 2 if it carries two derived copies. It follows that phasing is only an issue at heterozygous sites where it is not obvious which chromosome the derived allele belongs to.

For a genotype sequence that comprises of only one heterozygous site haplotype reconstruction is possible unambiguously, up to a permutation of the two derivatives. As the number of heterozygous sites in the sequence increases however, the complexity grows exponentially; in particular, for a sequence consisting of $n > 1$ heterozygous sites, there are 2^{n-1} distinct possible haplotype pairs, often deeming the analytic determination of the phase to be computationally prohibitive (for example, by calculating the likelihoods of all the possible arrangements), even if it were known how to carry out the computation. This section is dedicated to exploring statistical methods for the estimation of haplotypes from genotype data.

1.3.1 Family data

Family-based analyses have several advantages, and many haplotype estimation methods (for example reviewed in Gao et al. [2009]) have been developed to accommodate them. Consider first the specific case of trio data, when the the genotypes of the parents of a query genotype are also known. In this case, phase can be resolved deterministically by use of the Mendelian inheritance laws (we revisit this in Sections 2.2.2 and 2.3.2). For example, if the genotype of the father is 2 and that of the mother is 1, then the ancestral allele must have been inherited from the mother and therefore belongs to the maternal haplotype. However, if both parents are heterozygous it is not straightforward to resolve phase in this way as both parents could have transmitted the mutation. In such cases, knowledge of other relatives' genotypes, such as siblings and grandparents (or even more distant relatives if the genealogy is known) can be used to fill in these gaps by tracing haplotype sharing across the family.

1.3 Applications of relatedness: Statistical haplotype estimation

This is one example where knowledge of the full pedigree may be useful. It is often the case that samples consist of pedigrees (and even extended pedigrees) may also be present amongst a larger sample of unrelated individuals. When the pedigree of the sample is known, it is possible to use the same Mendelian constraints and linkage in order to make inference on phase: we extend the Mendelian rules that govern inheritance from parents to govern inheritance for grandparents and so on. This is the concept exploited by the IBD-based methods on unrelated samples, which estimate pedigrees and linkage via IBD. For example, recently Kong et al. [2008] (discussed in detail in Section 1.6) developed a method that phases deterministically using very distant relatives identified by IBD. There are other similar IBD-based methods presented in Section 1.3.3.

Returning to our discussion of family-based approaches, the set of consistent haplotype configurations within a pedigree is often very large, particularly when there are high levels of missing data; developing more sophisticated methods for estimating phase is therefore necessary. With related individuals statistical methods can make use of the greater levels of linkage that can be found in the data: there will not just be IBD present but longer shared chunks that are linked. As a result, most methods (such as Merlin; Abecasis et al. [2001]) are based on the Lander-Green algorithm (Lander and Green [1987]), which constructs a linkage map. Briefly, the Lander-Green algorithm constructs a linkage map using the Expectation-Maximization (EM) algorithm (Dempster et al. [1977]) to calculate the recombination fraction between adjacent loci. These recombination fractions are then used to find the expected number of recombinant meioses for each meiosis.

Family data, however, are not typically available in most studies and therefore numerous methods have been developed to estimate haplotypes for unrelated samples (reviewed in detail in Browning and Browning [2011b] and Marchini et al. [2006]). Here we will focus on some of these approaches for stochastic haplotype estimation, all of which rely on some assessment of relatedness within a sample. Identifying regions where individuals are more closely related enables haplotype estimation, in that, limited to that environment, these individuals can be considered as (distant) relatives and would thus be informative for phasing

1.3 Applications of relatedness: Statistical haplotype estimation

either the way pedigree data are or by allowing the calculation of population-level variables (for instance, allele frequencies) for their specific, local genetic background. This is why phasing methods are typically excellent for imputation (for example Browning [2008b]) and methods are generally developed jointly for both purposes.

Although here we distinguish between family-based studies and those catering to unrelated samples, we should note that trends in the literature suggest that methods are either adapting to accommodate both types (for example Delaneau et al. [2012]) or incorporate methods that cater to both types, for instance by exploiting familial relationships within unrelated samples (Kong et al. [2008], Browning and Browning [2011b]). For example, cryptic relatedness within unrelated samples can be accommodated and exploited with IBD-based methods leading to more accurate phase calls. The notion of merging methods for pedigrees and for unrelated individuals is one of the main themes explored in our work. Studies which have incorporated family data have achieved better phase performance (Marchini et al. [2006], O’Connell et al. [2014]); even when treated as unrelated individuals, relatives will be typically phased more accurately. In the same spirit, factors such as haplotype diversity and LD levels will typically influence a method’s performance. Although consistent patterns of performance on samples of different ethnicities have not been observed (Browning and Browning [2009], Marchini et al. [2006], Andrés et al. [2007]), imputation results have shown that including more closely related individuals in the reference sample (for example individuals of the same population or sub-population) improves imputation accuracy, particularly for low frequency variants (Huang et al. [2009], Jostins et al. [2011]). Similarly, for admixed samples, individuals from the ancestral populations should be included as well (Andrés et al. [2007]).

1.3.2 Stochastic estimation

The first ever phaser was Clark’s algorithm in 1990 (Clark [1990]), which lists all possible haplotypes that must be present in a sample for phase to be determined unambiguously. The method asks whether the genotypes can be resolved using

1.3 Applications of relatedness: Statistical haplotype estimation

any of the haplotypes + a complementary haplotype. The order in which the initial set of haplotypes is put together obviously affects outcome and the method is only suitable for tightly linked SNPs.

When SNPs are not tightly linked there may be multiple haplotype assignments consistent with the genotypes, and it was necessary to develop additional methodology to distinguish the correct ones. For this, the Expectation Maximisation (EM; Dempster et al. [1977]) was employed (e.g. Excoffier and Slatkin [1995], Hawley and Kidd [1995], Long et al. [1995]), which conditioned on haplotypes to maximise the likelihood of the observed genotype. This approach is effective for a small number of markers but breaks down for a large set of SNPs since the number of observed haplotypes increases, making the computational time required intractable and leading to their frequencies becoming too small for direct estimation. As the number of SNPs, n , increases, the number of possible haplotypes increases and the number of differences between them increases as well. They all have frequency $\frac{1}{n}$ and so all are equally likely. Therefore these methods perform more poorly since they rely on assessing the similarities between them.

It is possible to extend these methods so that they are applicable to a large number of markers, employing a sliding window approach, with size varying according to LD structure (Excoffier et al. [2003]) or using the partition-ligation technique introduced by Qin et al. [2002] in HAPLOTYPER. Another approach for accommodating models to a large set of markers and the correlations between them is using haplotype blocks (Halperin and Eskin [2004]). This however does not adequately capture recombination patterns since LD is not limited within block boundaries and can also vary within them (International HapMap Consortium et al. [2005]). In particular, McVean et al. [2004] and Myers and Griffiths [2003] have demonstrated that recombination rates vary greatly across the genome and between regions (with high recombination regions named hotspots and regions of low recombination called coldspots), a characteristic that must be particularly modelled for in order to achieve accurate haplotype estimates (Browning and Browning [2011b]). If the correlation between markers is not considered over an extended region, sampling variation will produce noise, leading to false correlations between markers and thus inaccurate phasing estimates.

1.3.3 Using IBD

We briefly mentioned the relevance of relatedness on imputation and inference of selection. Methods that are developed for these applications (particularly IBD-based methods), are typically used for phasing as well. When two individuals share an IBD segment, it follows that they are IBD at every position in that segment and typically these IBD alleles will be carried by the same chromosome, assuming that recombination has not occurred. Therefore if one of the individuals is homozygous where the other is heterozygous, this can be informative on the phase of the heterozygous site using haplotype sharing in the adjacent positions. That site is then phased relative to these neighbouring positions.

Its effectiveness in imputation follows from the same principle: for both processes family data are most powerful and most accurate (Marchini et al. [2006], Browning and Browning [2009]) and IBD mimics this relationship such that individuals in IBD can be treated locally like families. If an individual has been phased using IBD, then imputation can be carried out simply by copying from the haplotype with which s/he is in IBD. This concept was pioneered by Kong et al. [2008] with their method Long Range Phasing (LRP), which we discuss in detail in Section 1.6. Holm et al. [2011] showed the effectiveness of using Kong et al.’s estimated haplotypes by using them to impute putative rare causal variants, discovering stronger association signals.

Other methods that use IBD for phasing and/or imputation are BEAGLE and its derivative, fastIBD (Browning and Browning [2011a]). Palin et al. [2011]’s graphical model approach is inspired by Kong et al. [2008] but uses an HMM for detecting IBD introduced by Genovese et al. [2010]. Phasing is then achieved by using the Min-Sum algorithm (Kschischang et al. [2001]; a similar algorithm to Viterbi, which we discuss in Section 2.2.3) to identify the most probable haplotype phase solutions given the IBD relationships.

Finally, since the presence of selection leads to an increase in IBD sharing (Albrechtsen et al. [2010]), this can be exploited to search for excess sharing as a signature of selection. Albrechtsen et al. [2010] and Cai et al. [2011] both pursued this relationship with promising results. The problem with such an approach is that extended LD is itself a signature of natural selection and might thus lead to

1.3 Applications of relatedness: Statistical haplotype estimation

false-positives: it may not be clear whether IBD is detected because of selection or high levels of LD in a region.

1.3.4 The coalescent and HMMs

Before continuing our discussion on phasing methods, we briefly review HMMs (Rabiner [1989]), owing to the variety of population genetics and phasing methods that incorporate them, including those relevant to this thesis. HMMs are probabilistic models that are widely used (Yoon [2009]) in biological sequence analysis. In bioinformatics, they have been used for gene prediction (Munch and Krogh [2006]), pairwise- and multi- sequence alignment of DNA (Durbin [1998], Pachter et al. [2002], Lunter et al. [2008]) and RNA (Yoon and Vaidyanathan [2008]), base calling and modelling DNA sequencing errors (Liang et al. [2007], Lottaz et al. [2003]) and for prediction of protein structures (Won et al. [2007]). They have also been used extensively in population genetics, for instance making inference on selection (Gay et al. [2007], Kern and Haussler [2010], Mathieson and McVean [2013]), recombination and linkage (Felsenstein and Churchill [1996], Li and Stephens [2003], Lander and Green [1987], Kruglyak et al. [1996]). For applications that are more similar to the work presented in this thesis, HMMs have also been widely used for studying demographic history (Li and Durbin [2011]) and for making inference on speciation between humans and other primates (Hobolth et al. [2007], Mailund et al. [2011]). They have also been used extensively for phasing (Scheet and Stephens [2006], Delaneau et al. [2012]) and in the study of ancestry in a population (Price et al. [2009], Tang et al. [2006], Patterson et al. [2004], Hoggart et al. [2004], Montana and Pritchard [2004], Dutheil et al. [2009]). Some of these latter methods are reviewed in Liu et al. [2013]. Moreover, there is a variety of types of HMM, with profile-HMMs and pair-HMMs being some of the more popular examples for biological sequence analysis.

In their basic form, HMMs consist of a set of discrete, unobserved, ‘hidden’ states. In biological sequence analyses this state space is typically finite. The hidden states form a Markov chain (Norris [1998]) and changes between these

1.3 Applications of relatedness: Statistical haplotype estimation

states are called transitions. In simple Markov chains a first-order Markov property applies, and the probability of transitioning to a certain state in the next step depends only on the current position. HMMs have the additional feature that these unobserved states are associated with a set of events, called emissions. Therefore, in an HMM there are two separate stochastic processes that are taking place: the hidden transitions and the emission process of observed states. This is why HMMs are often called doubly-embedded stochastic processes (Rabiner [1989]). Since only the emitted states are visible to the observer, the challenge in the analyses using HMMs is to make inference on what the underlying sequence of hidden states may have been, given a set of observed states.

The main appeal behind HMMs is the fact that they are very effective in modelling the correlation between the hidden states and are further associated with a set of algorithms that are both computationally efficient (in general) but are also such that their parameters have useful real-life interpretations. These key algorithms are the Viterbi algorithm and the Forward/Backward algorithms. Full details on these algorithms are presented in Chapter 2. Using HMMs we can calculate the likelihood of the sequence of observed states while further making inference on the underlying sequence of hidden states that may have emitted the observations. For example, using the Forward/Backward algorithms for every hidden state we can calculate the posterior probability of a state at a given position. The Viterbi algorithm allows us to find the most likely sequence of hidden states, given our set of observations. When the transition and emission probabilities are unknown, the Expectation-Maximization (Baum-Welch) algorithm can be employed to estimate them. This is not relevant to this work and therefore the Baum-Welch algorithm is not explained further in this thesis.

Returning to our discussion on phasing, phasing techniques achieved a breakthrough in accuracy when they incorporated population haplotype frequencies in their models, allowing models to use similarity and not identity of samples as well as develop efficient Markov Chain Monte Carlo (MCMC) techniques. This was possible from the development of methods that approximated the coalescent, such as McVean and Cardin’s sequentially Markov coalescent (SMC; McVean and Cardin [2005]), and methods allowing for the approximation of genotype likelihoods given other individuals in the sample like Stephens and Donnelly [2000],

1.3 Applications of relatedness: Statistical haplotype estimation

Fearnhead and Donnelly [2001] and the Li and Stephens [2003] method (which we discuss more at the end of this chapter). The approximate coalescent allows for HMMs to be used for haplotype estimation, where the hidden states are typically some representation of the unknown haplotypes that are to be determined, and the observed data are the genotypes in the sample. These HMMs capitalise on the fact that mutation and recombination events are infrequent over short genomic regions and therefore in these regions haplotypes are expected to be similar to each other. This concept is discussed further in Li and Stephens [2003] and in the review of haplotype methods by Browning & Browning 2011 (Browning and Browning [2011b]).

One of the earliest and most prominent and accurate models (Browning and Browning [2011b], Marchini et al. [2006], Stephens and Donnelly [2003a]) is PHASE and its extensions (Stephens et al. [2001], Stephens and Donnelly [2003b], Stephens and Scheet [2005]). PHASE adopts the coalescent in introducing the concept that haplotypes that we have already seen are more likely than completely different ones. It employs an MCMC/Gibbs sampling approach achieving haplotype reconstruction by sampling estimates from the equilibrium distribution of the samples' haplotypes. Although it achieves high accuracies with an error rate for simulated data being around 2%, the software is computationally demanding that would be completely intractable for modern samples. It is furthermore only appropriate for a limited number of SNPs and regions covering few LD blocks.

The first computationally efficient approach was fastPHASE (Scheet and Stephens [2006]). FastPHASE employs a haplotype clustering approach which is more flexible than the traditional LD block-based methods described earlier, in that cluster membership varies along the genome. This clustering mechanism forms the basis for an HMM where pairs of individuals comprise of the hidden states, representing the haplotypes of the genotype in question. fastPHASE is computationally much more efficient than PHASE, but this comes with the penalty of lower accuracy for larger sample sizes.

In 2007, BEAGLE was developed (Browning and Browning [2007], Browning and Yu [2009]) which accounts heuristically for the variable recombination patterns in a similar approach to fastPHASE: by applying a localised haplotype-cluster model that was developed for association testing (Browning [2006]). It

1.3 Applications of relatedness: Statistical haplotype estimation

clusters individuals locally, according to the Markov property and demands that the haplotypes in the same cluster predict the same alleles in the neighbourhood of the current site. The number of clusters increases globally with sample size and locally with increasing LD. Their clustering therefore models the empirical pattern of LD and sets up an HMM (where the hidden states are clusters of pairs of individuals from the sample and the observed states the alleles of the genotype). Along the genotype it is possible to sample from the HMM, using the Forward/Backwards algorithms (a review of which can be found in Rabiner [1989] and Section 2.4.1). The best pair of haplotypes, the combination of which reproduces the genotype in question, can then be identified iteratively. BEAGLE is an order of magnitude faster than fastPHASE and it is known to be accurate for large sample sizes but less effective for smaller sample sizes or regions containing a small number of markers (Browning and Browning [2011b]).

IMPUTE (and its follow-up IMPUTE2; Howie et al. [2009] and Howie et al. [2011]) was developed as a phasing technique in order to achieve more accurate genotype imputation. The model starts by guessing phase and then adopts an MCMC iterative approach to improve its guesses by pooling information across the sample. During the updating procedure in the MCMC step, a pair of haplotypes is sampled from their conditional distribution, given their genotype, the haplotype guesses for the other individuals and the haplotypes in a reference panel. In the iterations, IMPUTE2 uses Hamming distance to select haplotypes that are similar to the haplotypes of the individual being phased. A very similar approach to IMPUTE2 is MACH (Li et al. [2010]). In contrast to IMPUTE2, MACH uses a random subset of sample haplotypes in the iterations; it internally calculates recombination and mutation rates (whereas IMPUTE2 requires input from the user). Both methods use the Li and Stephens [2003] imperfect mosaic idea to approximate the calculations. In general, IMPUTE2 is considered to be a slightly more accurate phaser and more computationally efficient than MACH, though overall BEAGLE is the fastest and also competitively accurate at large sample sizes (Browning and Browning [2011b]).

In this thesis we will be presenting a phasing approach. Though the methods described here are not directly comparable to ours, they are efficient phasers and

1.3 Applications of relatedness: Statistical haplotype estimation

we must therefore consider them as alternatives to our model. In our phaser comparisons we consider IMPUTE2 and BEAGLE. At the time when the analyses were carried out they were considered the gold-standard phasing methods, highlighting the trade-off between computational efficiency and high accuracies. Since then however, a new phaser has risen to prominence: Delaneau et al.’s SHAPEIT (Delaneau et al. [2008], Delaneau et al. [2012], Delaneau et al. [2013]). SHAPEIT extends the HMMs of IMPUTE2 and MACH to produce a method that scales linearly with sample size, thus making it computationally very efficient (second after BEAGLE, Delaneau et al. [2012]). The model splits the region into short segments (typically including 3 or 4 heterozygous sites) and represents the reference haplotype data for each segment as a haplotype graph. Its HMM then uses this collapsed graph as its hidden state space. SHAPEIT also has an efficient subsampling scheme for obtaining compatible genotypes from the full genotype sample, which further enhances computational performance. Finally, it has been extended to incorporate knowledge of family data and of read information and has been shown to achieve better haplotype estimates than other methods (Delaneau et al. [2012], Delaneau et al. [2013]).

In terms of phasing performance, SHAPEIT has not been considered here. More recently however, SHAPEIT has been incorporated in the phasing stage of IMPUTE2 and it is this version of IMPUTE2 that we have used in Chapter 4. By assessing the ability of the Li and Stephen’s method to detect recent ancestry using haplotypes estimated by IMPUTE2 compared to other methods we indirectly infer SHAPEIT’s superiority as a phaser. Refer to Chapters 2 and 4 for more details on this analysis.

1.3.5 Influential variables

There are several factors that may influence the performance of phasers. The most critical one is sample size: performance of all phasers improves for larger sample sizes, though relative improvement varies across methods (Browning and Browning [2011b]). This is true also for methods that are family based (Browning and Browning [2009]). Some approaches (for example MACH) exhibit less

1.3 Applications of relatedness: Statistical haplotype estimation

dramatic improvements than others (for example, BEAGLE), with sample size being most impactful for methods that incorporate population haplotype frequencies (Browning and Browning [2007]). It is not obvious how our own phaser uses population level information. We incorporate knowledge of LD but it is unclear how our performance will improve with sample size relative to BEAGLE and IMPUTE2.

Another factor that affects phase quality is marker density. Its effect on accuracy depends on what measure of accuracy is being used. Typically, phaser comparisons consider the switch rate (an explanation of the switch rate can be found in Stephens and Donnelly [2003a] and in Chapter 3 of this thesis) which measures error rate on a per-marker basis. In this case, increasing marker density is known to improve performance (Browning and Browning [2011b]). An alternative approach to assess haplotype estimation, which we also discuss but do not use here, asks what proportion of the haplotype has been correctly estimated. In this case, marker density can decrease performance as it increases the number of positions within a haplotype where an error might occur.

Genotyping error decreases phaser performance as it directly implies that at least one of the inferred haplotypes will carry an error where the original genotype was wrong. One way for methods to hedge against genotyping errors is by using genotype likelihoods rather than hard genotype calls. Both EM and HMM approaches can be extended to incorporate likelihoods (Kang et al. [2004], Browning and Yu [2009]), which has often led to higher accuracies (Browning and Yu [2009], 1000 Genomes Project Consortium et al. [2010], Nielsen et al. [2011]).

Rare variants are often considered harder to phase (Browning and Browning [2011b]). IBD-based methods (or similar) should, however, correctly capture their haplotype backgrounds as they are expected to live on longer (and therefore more easily detectable) shared haplotypes. Similarly, de novo mutations cannot be statistically estimated, unless the offspring of the individuals carrying them are available.

1.4 Applications of relatedness: Demographic inference

The relationship between individuals in a sample may inform on the historical relationship among individuals in the population. We already discussed, for instance, how models that use the coalescent can be used to shed light on possible clustering of individuals, thus informing on population structure. For example, if certain individuals seem to coalesce with each other before finding common ancestors with other individuals, this might reveal subgroups in a sample. Similarly, by modelling a changing population size we can detect exponential growth or historical bottlenecks. For the purposes of this thesis, we will focus on applications of relatedness on demographic inference by means of haplotype sharing. The length of IBD segments is inversely proportional to the TMRCA between the individuals at the region that they share, exploited by methods that attempt to date historical events (for example Ralph and Coop [2013], Lawson et al. [2012]). Average pair-wise IBD is a natural metric for relatedness, as we have already discussed. The main inhibiting factor is that IBD methods are underpowered for detecting shared segments among unrelated individuals whose common ancestors typically lived hundreds of generations ago (Browning and Browning [2012]), even though it has been shown that such shared haplotypes exist in abundance (Gusev et al. [2012]). Still, they may be used to pick up on the segments deriving from recent ancestry among these samples which is useful for making inference on recent demographic events.

In fact, Browning and Browning [2011a] showed that the proportion of the genome that two individuals share in IBD is a more accurate estimate of the true proportion shared than classic IBS-based methods and Huff et al. [2011] show that the inferred relationship between individuals is accurate within one meiotic step if they are up to seventh degree relatives (as for third cousins, for example).

IBD can also be used to infer population structure by showing that there is excess sharing within sub-populations compared to sharing between them. Browning and Browning [2011a] showed this with Welsh individuals against other individuals from the United Kingdom. Palamara et al. [2012] use GERMLINE

1.4 Applications of relatedness: Demographic inference

(Gusev et al. [2009]) to obtain IBD estimates, and then express these IBD estimates as a function of population demography, thus making inference on the population structure and demographic history of the Ashkenazi Jews and the Masai. Palamara and Pe’er [2013] extend this method and explicitly incorporate migration in their model to analyse the structure in the Masai samples. Soi et al. [2011] similarly compared African hunting-gathering populations and neighbouring agriculturalist and pastoralist populations.

Ralph and Coop [2013] use long shared IBD chunks to show that even seemingly unrelated individuals across Europe share numerous common ancestors that are less than 3,000 years old (~ 120 generations). To obtain IBD chunks within and between populations they used fastIBD (Browning and Browning [2011a]), and were able to identify chunks as small as 0.4cM; dating, however was restricted to chunks longer than 2cM due to the high error rates expected for chunks of shorter lengths.

Other examples include Campbell et al. [2012] who used IBD calls (using GERMLINE; Gusev et al. [2009] and PCA) to distinguish between North African Jewish and non-Jewish populations, and Gusev et al. [2012] who also extend GERMLINE (Gusev et al. [2009]) to demonstrate that IBD is more common than expected by naive models and that it can be used to cluster individuals into populations and to infer bottlenecks and founder effects.

In parallel to IBD methods, but also in light of their inability to detect shared segments that are very old, alternative methodologies have been developed for making inference on demography. Harris and Nielsen [2013] use IBS estimates from sequence data from Africans and Europeans to estimate the time and magnitude of admixture events, population divergence times and changes in effective population size. This latter aim highlights another application of IBD detection: estimating historical changes of the effective population size. In their method, they use McVean and Cardin’s SMC (McVean and Cardin [2005]) and Marjoram and Wall’s SMC’ (Marjoram and Wall [2006]) to obtain an expression for the expected IBS tract length distribution in a two-haplotype sample, that accommodates a complex model of demography between two diverged populations. They can therefore also obtain the entire length spectrum for shared tract lengths, rather than those which are detectable by IBD. To handle sample sizes greater

1.5 Applications of relatedness: Genome-wide association studies

than two, they subsample all pairs from a sample and regard them as independent, which may decrease power to detect more recent events. Finally, in the words of Harris and Nielsen, this approach ‘can distinguish short, extreme population bottlenecks from longer, less extreme ones that produce similar reductions in total genetic diversity’.

However, methods characterising relatedness and ancestry need not go via the IBS/IBD route at all. PCAdmix (Brisbin [2010], Brisbin et al. [2012]) divides the genome in small chunks of 10-50kb and within each probabilities admixture are estimated for individuals in a sample according to their positions in PCA space. These probabilities are then incorporated in a HMM framework with Viterbi in order to finalise admixture inference. More methods that detect admixture are reviewed by Seldin et al. [2011], focusing on their applications in disease mapping.

Other methods are HMM-based, where the hidden states, rather than IBD status, are alternative indicators of extent of relatedness. Price et al.’s 2009 haplotype-based HAPMIX (Price et al. [2009]), for instance, studies admixture from two ancestral populations which are set as the hidden states in the model. Price et al. also highlight the effect of phase errors, which result in spurious ancestry switches, particularly within regions where a query individual might be a heterozygote for ancestry. This in turn may lead to an overestimation of the time since admixture. Lawson et al. [2012] is another haplotype-based HMM that infers local ancestry patterns; we focus on this model in more detail in Section 1.6.

1.5 Applications of relatedness: Genome-wide association studies

One of the main applications of human genetics is to identify genetic risk factors for disease. Many methods have been developed to this end, such as family-based approaches, but the advent of genotyping and sequencing technologies that allow large-scale (both in terms of sample sizes and number of loci considered)

1.5 Applications of relatedness: Genome-wide association studies

analyses to take place. Genome-wide association studies (GWAS) search for putative mutations across the genome, both for continuous traits (such as eQTLs, cardiometabolic, biochemical information) and binary (case/control status) phenotypes. Several methods have been developed for carrying out the association tests and characterizing the relationship between an individual's genetic background and a phenotype and there have been numerous successful detections of putative loci (Klein et al. [2005], Wellcome Trust Case Control Consortium et al. [2007], Stranger et al. [2011] and many others).

Some of these approaches make use of the presence of IBD between individuals. The concept of using IBD in disease association studies is not new (Houwen et al. [1994], Meerman et al. [1995]). One of the most straightforward ways of incorporating IBD knowledge in this context is via IBD mapping: gene mapping using IBD. IBD mapping lies between linkage analysis with family data and association analysis. Causal loci can be identified by observing whether there is excess sharing of IBD tracts among cases (Purcell et al. [2007], Bercovici et al. [2010]). An alternative approach, employed by DASH (Gusev et al. [2011]), clusters individuals according to IBD status and then tests each cluster for association.

There are two key issues that often overshadow GWAS findings. First, it is not always possible to make functional interpretations for the mutations that have been shown to be associated; we do not discuss this further here. Second, though this has not directly been shown, it is assumed that common variation (which is typically considered in GWAS) so far only explains a fraction of the known heritability of most traits (Manolio et al. [2009], Eichler et al. [2010], Zuk et al. [2012]), and rare variation is therefore considered a significant contributor to increasing disease risk (Panoutsopoulou et al. [2013]).

Historically less attention was paid to rare variants as by definition only common variation was reliably genotyped and thus considered in population studies. New sequencing platforms have now though pushed the boundaries of our ability to detect genetic variation (e.g. 1000 Genomes Project Consortium et al. [2012]) and more rare variants are being discovered. What is more, de novo assembly and sequence alignment methods have become more accurate and we are thus better powered to distinguish rare variants from genotyping error. In the absence of iso-

1.5 Applications of relatedness: Genome-wide association studies

lated population data, large sample sizes are necessary for a significant proportion of the rare variation in a population to be captured by these new technologies.

Most variants are rare (1000 Genomes Project Consortium et al. [2012], Nelson et al. [2012]) and their introduction and frequency in a population is influenced by mutation rates, selection and demographic history - with exponential growth, for example, resulting in a significant increase in their frequency. They have also been shown to be population specific (for example 1000 Genomes Project Consortium et al. [2012]), which follows from the fact that they are typically the result of recent mutations and have thus not had enough time to distribute uniformly across populations. Most rare variants in coding sequences are also believed to be deleterious and as a result to be under purifying selection as per the “common-disease rare-variant” hypothesis, which is why it has been suggested that they may account to the missing heritability mentioned earlier. Even with current large sample sizes, traditional statistical association tests (such as those in SNPTEST; Marchini et al. [2007], Wellcome Trust Case Control Consortium et al. [2007], Marchini and Howie [2010] and PLINK, (Purcell et al. [2007]) are underpowered in the case of rare variants. This has led to the development of specialised statistical techniques (Asimit and Zeggini [2010], Bansal et al. [2010], Ladouceur et al. [2012]), though their power is also put into question under different association scenarios (Ladouceur et al. [2012]).

Characterizing stratification patterns is key when performing genome-wide association studies, as unaccounted for excess relatedness within a sample is known to be a principal cause for false positives, even within largely homogeneous populations like the United Kingdom. However, most of the focus for developing methods that will account for this has also revolved around common variation ($MAF > 5\%$), driven by the “common-disease common-variant” hypothesis. These methods also neglect the fact that rare variants have been shown to contribute to numerous diseases (e.g. Bodmer and Bonilla [2008], Manolio et al. [2009]) and are plausibly exhibiting different stratification patterns than common variants (as a result of them being under purifying selection and being population specific). Recent studies (for example Babron et al. [2012], Moore et al. [2013a,b]) have confirmed that their stratification patterns are different, and crucially, that

1.5 Applications of relatedness: Genome-wide association studies

these patterns are not always well accounted for by the classic correction methods (such as Principle Component Analysis (PCA) and mixed models; (Mathieson and McVean [2012], Jiang [2011], Jiang et al. [2013])). Mixed models capture background polygenic effects using the standard kinship matrix which is effective for common variation but underpowered for detecting local relationships – relationships between individuals that are more prominent in a particular region of the genome but not strong genome-wide.

The main advantage of IBD methods versus classical association tests is that they focus on rare variants. In fact, Browning and Thompson [2012] showed that IBD mapping is more powerful in detecting association when multiple rare variants are causal within the same gene. Finally, the boundaries of the IBD segments provide clear indicators for where sequencing could take place and which individuals are most likely the carriers, for example for follow-up functional studies following an association test. A thorough comparison between IBD testing methods and the classic haplotypic tests can be found in Browning and Thompson [2012].

There have been some studies that used IBD mapping in unrelated samples (Albrechtsen et al. [2009], Moltke et al. [2011], Gusev et al. [2011], Francks et al. [2010]). Although their results have been promising in identifying IBD segments around known causal loci, IBD mapping has mostly been used with pedigree data to avoid the computational expense of linkage-based methods (Wijsman et al. [2006]). These latter methods are also less reliable when individuals early in the pedigree have not been genotyped (Brown et al. [2012], Thompson [2000]), which is typically the case. There are numerous examples of IBD mapping with pedigrees (for example Glazner and Thompson [2012], Roach et al. [2010]) and a more formal account of their findings as well as information on other studies can be found in Browning and Browning [2012].

IBD methods can decrease the effect of confounding by shared environments among distantly related samples, by allowing the estimation of heritability using pair-wise genome-wide IBD sharing (Browning and Browning [2013b], Price et al. [2011], Yang et al. [2010], Zuk et al. [2012]). A caveat here is that these IBD estimates can sometimes be biased due to population structure (Browning and Browning [2011c], Goddard et al. [2011]). Finally, by observing which samples

share IBD segments and assuming that their ancestry is known, it is possible to distinguish between recurrent mutation and shared ancestry (Girirajan et al. [2010], Hansen et al. [2009], Hansen et al. [2010]).

1.6 Long Range Phasing

The main bridge between family study designs for phasing and methods applicable to unrelated samples was developed by Kong et al. [2008]. Kong et al. took advantage of the demographic history of Iceland and a data set that represents a substantial fraction of the entire population (approximately 10%) in order to construct an IBD-based approach for phasing and imputation, Long Range Phasing (LRP). Their approach is heuristic and capitalises from the fact that the kinship coefficient between the genotyped samples is high enough for considerable amounts of IBD to be expected ($\approx 10\text{cM}$). In fact, given that the sample size is large relative to the size of the underlying population, one would expect multiple individuals to be in IBD at any given locus or region.

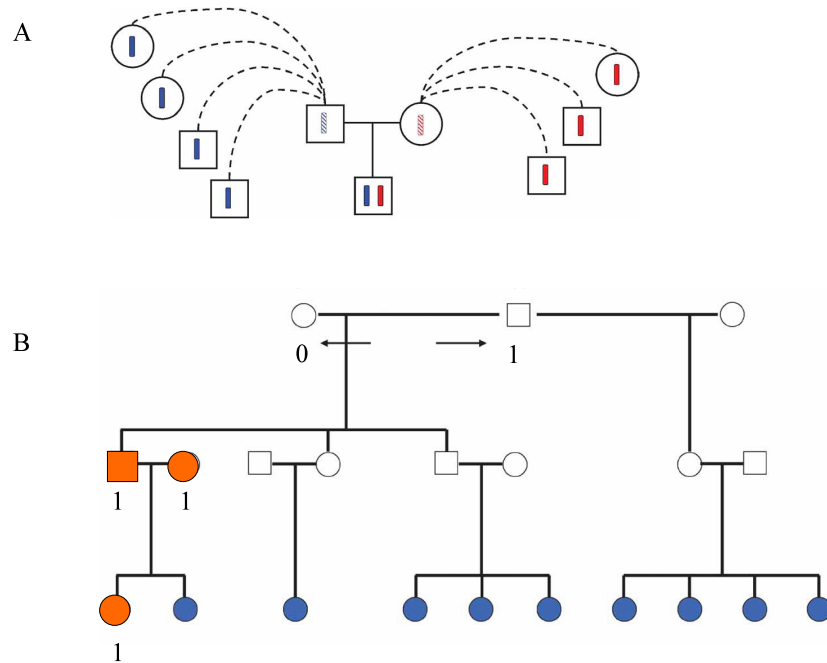
To identify IBD, Kong et al. used the IBS length threshold approach and identified segments as IBD if they were IBS for at least 1000 consecutive SNPs. For every query genotype, individuals in IBS (call them $\mathcal{J}$, say) with it were identified in this way. An individual in IBS with a genotype, will have genotype IBS and will thus share just one of the two haplotypes of the query individual. The innovation of Kong et al. was then to split the individuals in $\mathcal{J}$ into two groups, depending on which of the two haplotypes they share with the query genotype in the region.

These individuals are locally close relatives of the query genotype, Kong et al. coined the terms ‘surrogate parents’ and ‘pseudo-parents’ to describe them (an illustration can be found in Figure 1.1). Kong et al.’s innovation is that these two groups can be randomly assigned to the pseudo-father and pseudo-mother categories, thus constructing a pseudo-trio. Then, the pseudo-mothers and the pseudo-fathers can be used to phase the query genotype (the ‘child’) by transmission, in the same way that real parents would. Having more than one pseudo-parent in each set therefore increases the power of the approach since only

one parent in each need be homozygous in order for phase to be assigned. Note that pseudo-parents can also be of either gender and may further swap groups along different regions of the child. Moreover, if an individual X is a parent of individual Y , Y is also a parent of X .

In the absence of recombination (or for regions where it is known that a very low recombination rate applies), while phasing the ‘child’, the parents can be phased as well. In light of the entire ‘family’ this extends to phasing the more distant ‘pseudo-relatives’. This can be useful for instance in cases where the child and the parents are all heterozygous, while the genotypes of the grandparents are informative and can then be considered to distinguish phase (see Figure 1.1). For example, in Figure 1.1B, the fact that one grandparent is homozygous solves the problem of the triple heterozygote at the bottom left part of the diagram. Part A of the figure illustrates how several individuals may be assigned the pseudo-parent role. Individuals in blue are in IBS with the child’s first haplotype, while individuals in red share the other haplotype.

However, Kong et al.’s justification for relatedness being identified by means of IBD and IBS stems from the particularly high kinship coefficient that holds for the Icelandic population that they considered. In particular, they suggest that for the method to be applicable on a different sample, it must include 1% of the population under study, which in real terms is still a prohibitively large number of individuals for most populations. Moreover, the LRP method requires very high quality genotype data in order for IBD tracts to be identified with high accuracy. In fact, within a surrogate parent group it is possible that the pseudo-mothers and pseudo-fathers are such there are inconsistencies with the query individual: the two groups may either give multiple phase calls at a site or they may simply be incompatible with the ‘child’. This could suggest that they have been falsely identified as IBD individuals, or they may be genotyping errors for instance. Kong et al. do not discuss how they deal with such cases. In fact, dividing the IBS surrogate-parents into the two groups is in itself not a straightforward process. Finally, LRP ignores recombination - incorporating recombination would prevent the phasing of parents together with the child which leaves open the question of phasing triple heterozygotes.



Kong et al. (2008); Figures 1 and 4 (edited).

Figure 1.1 **A Pseudo Family** The figures are from Kong et al. [2008]; Figure A is Figure 1 in the publication, and Figure B is Figure 4 and has been slightly modified here. A: Individuals which can be considered as relatives are identified using IBS, and can be split into pseudo-fathers (blue) and pseudo-mothers (red). They may then be used for phasing like a real trio. B: A family; the numbers under some individuals are their genotypes. Each individual can be phased as in A but assuming no recombination, haplotype estimation of the ‘child’ directly leads to the phasing of the ‘parents’ as well, and similarly phasing of the ‘grandparents’, since we know which parental haplotype the child is copying from. In the triple heterozygous case, indicated in the diagram (shown in orange) the grandparents can be used to resolve ambiguity by tracing down the haplotypes along which the mutation is transmitted across the 3 generations. In this example the arrows point to the grand-parental genotypes that may be useful for phasing the first leaf on the left (the ‘child’ in the trio in question).

1.7 Nearest Neighbours

As an alternative approach for considering genealogical relationships in the sample that capture patterns of coancestry, we next turn to the work of Lawson et al.

[2012]. In a sample of unrelated individuals from a population, each haplotype has one or more genealogical nearest-neighbours (NNs) at a given position along a chromosome (Figure 1.2). While NNs cannot be directly detected from genotype data, pairs of chromosomes with recent common ancestors are likely to also share extended haplotypes around the position; we illustrate this in Figure 1.2. In a large sample of unrelated individuals, at each location in the genome, the two chromosomes in an individual have NN sets that are typically distinct and a chromosome may have more than one NNs. If it is assumed that over a short interval (for example, approximately 1Mb in the case of LRP) there is no recombination, in which case the NNs do not change, then extended genetic similarity or haplotype sharing can be used to identify them.

NN knowledge and the extent of haplotype sharing with them fully captures the recent relationships between samples. In addition to being informative of recent demographic events in the same way as IBD, this recent coancestry is also relevant to rare variants, whose full characterization is often crucial in disease association studies, as we previously discussed. The concept of NNs was pioneered by Lawson et al. [2012] in the development of ChromoPainter and fineSTRUCTURE for the detection of shared ancestry and structure within a sample. The methods assume knowledge of haplotypes and an initial step in their implementation is statistical haplotype estimation using fastPHASE (Scheet and Stephens [2006]). The authors warn of the problems that phase errors can lead to, though this is not quantified.

For each query haplotype, ChromoPainter uses the Li and Stephens [2003] model (LS; reviewed briefly below) in order to express the haplotype as an imperfect mosaic of the other haplotypes in the sample (we explain this terminology in the next subsection). Averaging over the proportion of sites that the haplotype is copying from each of the other samples gives a measure of coancestry between this haplotype and the others. Repeating this for every haplotype produces a coancestry matrix which contains all the pairwise coancestry values. Lawson et al. show that this captures the true genealogical relationships well, though an exact success rate for NN discovery is not provided. FineStructure involves a Bayesian model that analyses the coancestry matrix, which is a dimensionality reduction step, in order to determine whether there is structure in the data. In

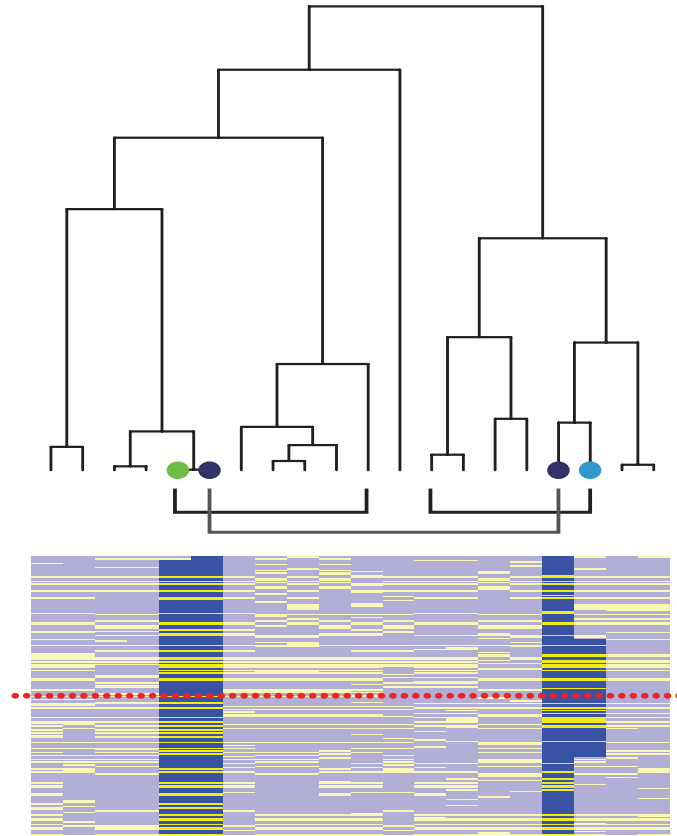


Figure 1.2 **Nearest neighbours** The figure shows a marginal genealogical tree for a sample of haplotypes and an illustration of the genetic variation that has given rise to it. Rows are different genomic positions. Going along a position, the alternating purple and yellow indicate the different alleles (assumed 2) that individuals might be carrying. The horizontal red line goes along the site at which the tree shown is the marginal tree of the sample. Individuals consist of two haplotypes and are therefore defined by two leaves on this tree. The dark blue dots are the two haplotypes of one individual in this sample. Each of the haplotypes is the genealogical sibling of another haplotype in this tree. These siblings are the Nearest Neighbours (NN). For example, the first dark blue haplotype has the leaf marked in green as its NN, while the second haplotype has the light blue leaf as its NN. When the most recent common ancestor of NNs is recent, as depicted here, NNs are expected to be genetically similar over extended regions around the current position. This is illustrated by the regions highlighted in dark blue.

in addition to identifying clusters and potential admixture, the Bayesian framework allows for the estimation of coalescence times between groups.

1.7.1 The Li and Stephens model

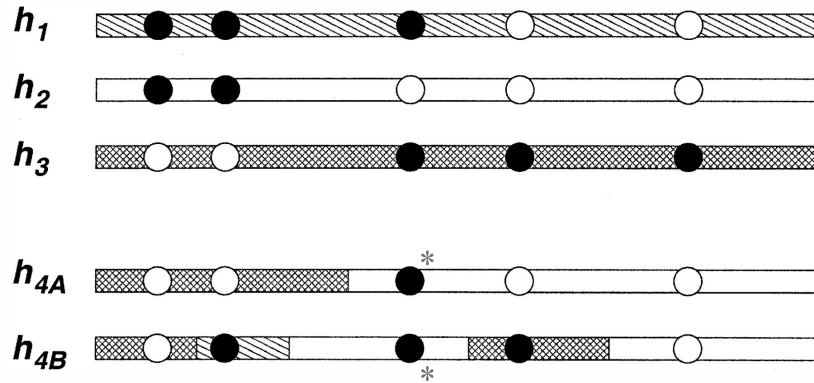
We end our discussion on ChromoPainter with a brief overview of the Li and Stephens (LS) model, both to better explain how ChromoPainter works but also because we will be employing the LS method later in this thesis. Li and Stephens [2003] developed a model to approximate the probability of observing a new haplotype, given a set of haplotypes. The motivation behind their model was to estimate recombination rates between adjacent SNPs, using haplotype data. Their approach coins the ‘imperfect mosaic’ idea (imperfect due to mutation; mosaic due to recombination), which describes a query haplotype as a sequence of genetic chunks copied from the other haplotypes in the sample. The assumption here is that the query haplotype and the haplotypes it copies from have inherited their shared genomic material together form a common ancestor.

The recombination rates in the region determine the variety of the mosaic: regions with hotspots should contain numerous small segments, while regions of low recombination activity lead to longer inherited chunks. The extent of imperfection in the mosaic depends on both the mutation rate and the genotyping error present in the data. Li and Stephens demonstrated that new genotypes can be simulated using imperfect mosaics of existing existing haplotypes in a sample (see Figure 1.3). Conversely, it follows that any genotype in a sample can be expressed as an imperfect mosaic of the remaining genotypes, as h4 in Figure 1.3.

Briefly, LS construct an HMM with hidden states the individual haplotypes in a sample, and observed states the alleles of the query haplotype. For a query haplotype, the aim is to select an individual from the sample at each SNP that the query haplotype ‘copies’ from. Transitions model changes in ancestry and therefore depend on the recombination rate, ρ_j at some site j . Following Li and Stephens [2003], the transition probability from individual i to j from a sample of size N are as follows:

$$t_{i,j} = \begin{cases} \exp(-\frac{\rho_j d_j}{N}) + (1 - \exp(-\frac{\rho_j d_j}{N})) \cdot \frac{1}{N} & \text{if } i = j \\ (1 - \exp(-\frac{\rho_j d_j}{N})) \cdot \frac{1}{N} & \text{otherwise} \end{cases}$$

where d_j is the physical distance between sites j and $j + 1$.



Li N, and Stephens M *Genetics* 2003;165:2213-2233

Figure 1.3 **An illustration of the Li and Stephens [2003] imperfect mosaic model.** There are three toy haplotypes (h_1 - h_3) present in this example which are used to simulate two new haplotypes for an individual. This is achieved by combining different segments from each of the three haplotypes and putting them together to form the two new ones (h_{4A} , h_{4B}). The shading shows which haplotypes were selected for each chunk with the change in segment, illustrating recombination events. In this diagram, columns represent SNPs, with the white and black dots representing the two possible alleles for each SNP (assuming a biallelic setup). The copying of the segments to make h_{4A} and h_{4B} is imperfect as can be seen in the case of the mutations marked with a star which are not the same colour as on the haplotypes that they have been ‘copied’ from.

Emissions look at whether the query haplotype carries the same allele as the hidden state at each SNP. In the original paper, the authors use Watterson’s estimator (Watterson [1975]) for the mutation rate and emission probabilities are the probabilities that a mutation takes place, using the geometric distribution for the number of mutations in the genealogy and given the sample size.

The LS method has been used extensively as the basis for more sophisticated models that characterize ancestry in a sample, such ChromoPainter (Lawson et al. [2012]), PCAdmix (Brisbin et al. [2012]), HAPGEN (Spencer et al. [2009]), HAP-SAMPLE (Wright et al. [2007]). It has also been used directly on genotypes and haplotypes, for example Wang et al. [2013]. It has been shown to approximate

the probabilities of the full coalescent well and is therefore what is considered here as the gold-standard for stochastic models that analyse population genetic data.

1.8 This thesis

Phasing was originally the main motivation behind the work undertaken for this thesis. We developed a model-based method that captures many of the ideas of the Kong et al. pseudo-parent approach, but accommodates smaller samples that are not a significant proportion of the underlying population. Our aim is therefore to analyse completely unrelated samples. Specifically, we aim to detect sets of pseudo-trios (mother-father-offspring) using genotype data, allowing for multiple pseudo-trios for each individual at each position. Our implementation uses a HMM with hidden states the ordered pairs of individuals, representing pseudo-parents. Pseudo-parents are allowed to change along the sequence through a recombination-like process. Unlike Kong et al., we therefore incorporate LD information in our model.

Our pseudo-parents can be used for phasing following the same principles introduced by Kong et al. We can take this a step further though and additionally investigate the extent of relatedness between the ‘children’ and the pseudo-parents that we identify. Similar to Kong et al., we are effectively searching for recent ancestry, expecting therefore that the individuals that we identify are close relatives that share extended haplotypes with the query individuals. Identifying local, and therefore recent, relatedness is important as it is relevant to rare variation.

A natural question therefore is how close this relationship is between an individual and the pseudo-parents that we identify and, further, whether we in fact identify nearest neighbours (NNs) and maximally shared haplotype chunks. We do not explore the implications of NN knowledge for identifying structure in this thesis, in the same way as *fineStructure*. Rather, our approach is limited just to the initial step of NN identification and is in effect the genotype equivalent of *ChromoPainter*, though we do not consider averaged ‘time spent’ with pseudo-parent pairs. In addition to phasing therefore, in this project we consider

whether regions of extended haplotype identity and NNs, can be detected from genotype data without prior phasing. To assess our performance, rather than comparing against ChromoPainter it suffices to consider an implementation of the LS method and this is the approach that we pursue.

Our method detects extended shared haplotypes with recent ancestors and by definition these shared haplotypes are IBD. However, we do not model IBD nor do we threshold the allowed TMRCA or haplotype lengths for our chunks to be defined. As a result we can obtain the full length distribution of haplotype sharing among recent relatives. Our method is therefore not directly comparable to IBD approaches and such a comparison is not considered in this thesis. IBD accuracy is measured from simulated data in terms of false positive (specificity) and false negative (sensitivity) IBD detection rates. In contrast, our performance is assessed via our ability to detect Nearest Neighbours and the sharing that is expected to exist between them.

The caveat to our HMM-based model is that it suffers computationally both in terms of speed and memory requirements, making it intractable for sample sizes greater than even a few hundred individuals. In addition to our HMM, we therefore develop two related but alternative implementations of our algorithm that enable the analysis of thousands of samples: a graph-based approach and an HMM with subsampling (referred to as HMMsub).

In Chapter 2 we introduce the details of our HMM and its two approximations (graph and HMMsub). We describe how our model was developed and how it is implemented for the purposes of haplotype estimation. We further explain how we pursue NN identification and detection of extended shared haplotypes; we call the joint pursuit of these latter goals as ‘coancestry characterization’.

Chapters 3 and 4 present the results of our performance on phasing and coancestry characterization respectively, using simulated data. In its current form our model does not revolutionise the field of statistical haplotype estimation. It does, however, compete with cutting-edge approaches and we will argue that it is possible to extend it to achieve improved results. We further show that our model leads to more accurate estimates of extended haplotype sharing than approaches using pre-phasing and that it can detect NNs at a rate close to the

theoretical maximum. Our genotype-based HMM has three main advantages over haplotype-based methods:

1. It is unaffected by phasing errors. Although phasing in large population data sets can be highly accurate (Browning and Yu [2009], Delaneau et al. [2012], Kong et al. [2008], Palin et al. [2011], Williams et al. [2012]), some errors will occur which will inevitably tend to underestimate the true extent of haplotype identity as phasing errors will interrupt shared genomic chunks.
2. It is better insulated against genotyping error.
3. Even in the absence of errors, haplotype-based methods will prefer to detect less sharing in favour of finding more NNs when there are several in any given region. In contrast, our method will prefer to detect long shared chunks, sometimes selecting genealogical cousins rather than siblings in order to do so.

In fact, our simulation-based investigation reveals that there is a trade-off between detecting extended sharing and the most recent ancestry, in particular for populations that have experienced exponential growth.

Chapter 5 explores the application of our graph method on an unrelated and two isolated samples from Greece – the HELIC data (www.helic.org), in which there are about 700 individuals per cohort. Chapter 6 examines the results from the graph method and HMMsub on 5,144 WTCCC (Wellcome Trust Case Control Consortium et al. [2007]) British individuals. For both data sets we use the lengths of the shared haplotype chunks to estimate the TMRCA between an individual and its NNs. Our work on the Greek isolates illustrates how haplotype sharing can be used to characterize isolatedness, to explore whether there has been any migration between the isolates and the unrelated sample and to attempt to date the time of isolation. By studying the British sample we discover that haplotype sharing is typically 1.7cM-2cM suggesting that the population must have undergone rapid exponential growth. Finally, we end with an assessment of our findings and a discussion on how the current implementation of our method can be useful in the second main application of relatedness: disease association studies.

Aside to the work presented in this thesis, the author has participated in projects that have contributed to an additional three publications. First, I imputed Chromosome 12 variants from the 1000 Genomes data into the ~ 7000 WTCCC UK Multiple Sclerosis (MS) cases and controls. The aim was to fine-map the signal around rs1800693 in the TNFRSF1A gene which was shown to be associated with MS. Lars Fugger's lab then worked on understanding the pathways by which rs1800693 influenced MS in British individuals as well as the clinical implications of the discovery. They confirmed that the rs1800693 variant was indeed causal, they characterised the biochemical changes that it leads to and demonstrated that these changes mimic the effect of TNF-blocking drugs. The results are published in Gregory et al. [2012]. Following this project, I was involved in the MS part of the Immunochip project (Cortes and Brown [2011]), where some 200,000 SNPs were carefully selected and genotyped for a large-scale association study on autoimmune diseases. My involvement saw the genome-wide imputation of 1000 Genomes variants into $\sim 45,000$ individuals from different populations, followed by Bayesian and frequentist association testing using SNPTEST (Marchini et al. [2007]). By studying the population-specific signals and the meta-analysis results around the typed hits, we investigated which signals could be fine-mapped to just a few variants. Our findings were incorporated in International Multiple Sclerosis Genetics Consortium et al. [2013]. Finally, I contributed to the second publication of the 1000 Genomes project (1000 Genomes Project Consortium et al. [2012]). For each population separately, for each variant I examined the length of the haplotype shared by the individuals that carried that variant. It was then possible to see how this length depended on variant frequency. As expected, more common variants lived on shorter shared haplotype chunks, while rare variants lived on long haplotypes, indicative of the more recent ancestry that must exist between the individuals that carried them.

Chapter 2

Model and Methodology

2.1 Method Motivation

We aim to generalise the LRP method (Kong et al. [2008]) through a model-based approach that identifies pseudo-parents along the genome. Our model is an HMM that uses genotypes and, by not making assumptions about the levels of kinship present in the sample, is constructed to accommodate unrelated individuals. By identifying pseudo-parents we are constructing pseudo-trios which we may then use for phasing in the same way as Kong et al. We further wish to investigate whether the pseudo-parents that we identify are in fact local NNs of the query genotypes. In this way we build the genotype equivalent of ChromoPainter (Lawson et al. [2012]), seeking out to characterize coancestry levels in the sample: both identify most closely related individuals and detect extended haplotype sharing. In this chapter we will motivate and introduce the model that we have constructed, its approximations that accommodate large samples consisting of thousands of individuals and the modifications to our approach that specialise it either for haplotype estimation or coancestry characterization.

We wish to explore more local relationships and we therefore begin our construction by considering familial relationships. Within a family, a mosaic of

siblings' genomes can imperfectly recover their parents' transmitted haplotypes. The LS model introduced the concept of identifying a mosaic of individuals in the sample that collectively reconstruct the genome of a query individual. Here, we are searching for pseudo-parents (which are pseudo-ancestors, simply more recent ones) and we can therefore use this imperfect mosaic concept: two imperfect mosaics of the individuals in a sample can reconstruct the transmitted haplotypes of the query individual's parents. In other words, our search can be thought of as an extension of the LS method, except we are now searching for two individuals at any given site. This motivates our modelling framework to be an HMM where the hidden states at a site are the pairs of individuals in the sample that act as pseudo-parents rather than a single haplotype acting as the ancestor. The emitted states constitute the query individual, similar to the LS approach. The questions of interest then are:

1. Are these pseudo-parents effective for haplotype estimation in the same way that true parents are?
2. Does each of the pseudo-parents that we have detected in fact contain one of the two haplotypes that are genealogical NNs of the two haplotypes that make up our query genotype?
3. Can we use NN identification to estimate haplotype sharing between pairs of individuals in the sample?

It is not obvious that the same method can effectively address the first two questions: perhaps searching for representatives (imperfect mosaics) of parents and representatives of NNs are two different problems. In this thesis we show that our model can suitably address both, albeit with a change in parameterisation.

Section 2.2 contains full details on the core part of the model that is relevant to both, namely the details of our HMM and the method by which we identify a sequence of pseudo-parents along the genome. Section 2.3 presents the details of the model specific to haplotype estimation. Finally, Section 2.4 focuses on methods used for finding NNs and estimating haplotype sharing. We also introduce two ways for approximating our model in order to gain computational efficiency:

the graph-based method (Section 2.4.2) and an HMM with subsampling (HMM-sub; Section 2.4.3). We end with a summary of our HMM approach that outlines our full algorithm, including its application to phasing and coancestry characterization. Figure 2.7 at the end of this Chapter provides an illustration of the entire approach.

2.2 The Hidden Markov Model

Given a query individual, we are searching for one pair of individuals (or more) which we set as the hidden states in our HMM. Our strategy is to use the standard HMM algorithms to select the most suitable pair at every SNP, depending on their genotype at that SNP and the genotype carried by the query individual. The genotypes at each SNP of the query individual therefore constitute the observed states. In the next few sections we explain our choice for transition probabilities as well as the different strategies we considered for assessing emissions. We will also present the main HMM algorithms that we will be considering: the Viterbi and the Forward/Backward algorithms.

2.2.1 Transition Probabilities

In our model, transitions between hidden states model recombination events: when one of the parents that the query genotype (the ‘child’) inherits from changes, this represents a change in ancestry, or alternatively a change in the genealogical tree. Similarly to the ideas proposed by Li and Stephens [2003], who describe a sample as a ‘mosaic’ of the ancestral genetic material, here we want to capture this mosaicism but within the framework of Mendelian inheritance. Therefore, while in Li and Stephens [2003] a recombination event meant a change in ‘ancestor’, here a recombination event translates to a change in ‘parent’.

More formally, suppose we have n sequences consisting of L SNPs. These can be arranged into $\binom{n}{2} = m$ possible pairs of ‘parents’ for the ‘child’. Between two SNPs, let $P(\text{a recomb. event occurs}) = r$, suppose r to be quite small, and let

the transition probability from state i to state j at a particular site be denoted by $t_{i,j}$. Note that the probability of recombination depends on which pair of SNPs is being considered, but here we just consider r for simplicity. Suppose further that when a recombination event does occur, for individuals a and b , $P(a \text{ goes to } b) = \frac{1}{n}$, for $a, b \in \{1, \dots, n\}$. Note that $P(a \text{ goes to } a) = \frac{1}{n} > 0$. It follows that, for distinct individuals a, b, c and d :

$$t_{i,j} = \begin{cases} (1-r)^2 + O\left(\frac{r}{n}\right) & \text{if } i = \{a, b\}, j = \{a, b\} \\ \frac{r(1-r)}{n} + O\left(\frac{r^2}{n^2}\right) & \text{if } i = \{a, b\}, j = \{a, d\} \\ \frac{2r^2}{n^2} & \text{if } i = \{a, b\}, j = \{c, d\} \end{cases} \quad (2.1)$$

where the first probability is that of no observable recombination occurring, the second is for one parent in the pair changing and the third for both parents changing. The error terms arise from the unobservable recombination events: from the cases where a parent does recombine but the individual it chooses to go to is itself, or both parents recombine but choose themselves or each other. In practice the latter scenario of double jumps is disallowed in our implementation for computational efficiency, but also to reduce the error rate in the case of phasing. In other words, between any two SNPs we only allow one candidate parent to change.

2.2.2 Emission Scores

One natural approach for representing the local parent-offspring relationship in a trio is using the Mendelian transmission (MT) probabilities independently at every site. The first order Markov property and the conditional independence of sites imposed by the simple HMM that we are using in fact imposes such an approach. In practice, transmission from parents to offspring does not occur independently site-wise but is governed by linkage and LD patterns. In other words, what the child inherits at a site is correlated with what was inherited in neighbouring positions due to allelic associations in the population. However, these

dependence properties although strong are, as we discover, not critical to incorporate in our model and therefore using this emission process to describe the fit of two candidate parents with the query genotype is a reasonable approximation. Our first choice for emission probabilities are therefore the MT probabilities.

Mendelian transmission (and its extensions, which we will discuss shortly) however, is not the only way to quantify the fit between two parents and their offspring. Instead, we could consider a simpler scenario which looks just at compatibility: could the offspring's genotype plausibly have been inherited from the particular pair of candidate parents? Compatibility scores have been indirectly used in numerous population genetics models. Compatibility is a similar concept to IBS in that the pseudo-parents are required to be in IBS with the 'child' at the locus at which they are selected, but also such that the child's genotype at that locus has been plausibly inherited from them as a pair. Compatibility is also similar to Hamming distance, as are kinship coefficients. We discussed how all these metrics are relatedness metrics in Chapter 1, and there was therefore strong motivation to consider compatibility emissions (CE) as an alternative to MT in our model.

More formally, let the score of emitting genotype $x \in \{0, 1, 2\}$ from state i at site j be denoted as $e_{i,j}(x)$ and suppose the child's genotype at site j to be x_j , the father's to be f_j and the mother's m_j . Further, let the probability that a genotyping error or that a mutation has occurred at SNP j be equal to ϵ_j . Considering first the Mendelian Transmission Probabilities, we obtain Table 2.1. In Mendelian transmission, the probability of the triple heterozygote, ζ , is 0.5. When this is the case, the emission probabilities are such that when the query genotype is heterozygous (which is the situation in which haplotype estimation is non-trivial), the HMM places equal weight on selecting parents that are both also heterozygous as to some of the other compatible genotype combinations (see Table 2.1). While this is biologically justifiable, it does not deter the construction of triple heterozygotes, which do not help resolve phase.

As a result, in order to ensure more informative paths are selected, we could adopt a weighted Mendelian approach, with ζ being significantly reduced to, say, 0.01. By an informative path we mean one where the pseudo-parent pairs that it consists of are such that they can resolve phase via Mendelian transmission;

for more details see Section 2.3.2. This weighted Mendelian approach is not biologically justified, but is a plausible heuristic solution to our problem.

Compatibility emissions (Table 2.2) would suffer from the same problem as the simple MT case and are therefore also not expected to yield informative paths for haplotype estimation. They may be efficient nevertheless at capturing local relatedness structures in the same way that IBS-based metrics are believed to, enabling our model to identify true genealogical NNs and extended shared haplotype chunks. Figure 2.1 is an illustration of the emission scores that are considered, highlighting the changes in the weight associated to different combinations of pseudo-parental and offspring genotypes. In subsequent Chapters we explore which choices best match the various applications of our method.

Note that in our method the mutation/error probability is kept fixed at 0.01. This is in fact an error model. Our approach does not incorporate a mutation model, in that we do not use population- or sample-based estimates to calculate this error probability. For example, we could have used the more sophisticated estimate for the mutation rate which is used by ChromoPainter and the original LS method: Watterson’s estimate (Watterson [1975]). We do not explore such a strategy here and the reason for this is two-fold. First, simulations (data not shown) and other work (for instance Moutsianas [2011]; Section 6.9.1, Figure 6.7) have shown that variable mutation rates have a negligible effect on the fit of LS-type models (at least in the context of imputation and phasing). Recombination is what determines the mosaicism, which is the object of interest here, and we therefore only explore how our results change when the difference in the order of magnitude between mutation and recombination changes. The second reason for incorporating an error model is that genotyping errors in contrast do have real effects in a method’s performance. An error model can guard against them.

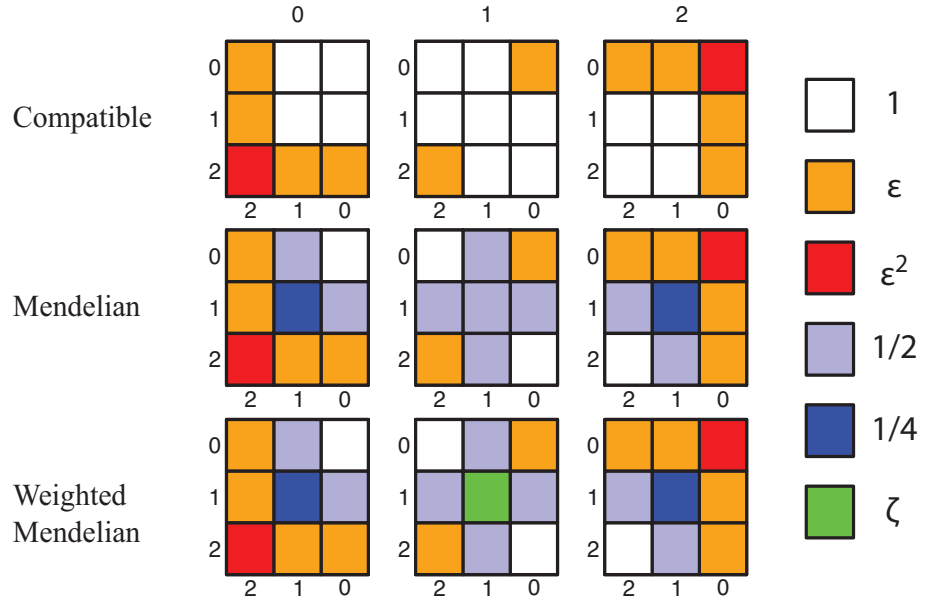


Figure 2.1 **Emission probabilities.** The diagram shows a 3 by 3 arrangement of squares. Each of the three rows represents one of the three main emission scores that we consider: Compatibility emissions, Mendelian probabilities and Weighted Mendelian probabilities. The 3 main columns are for the 3 possible genotypes ($\{0, 1, 2\}$) that a query individual may carry at a particular SNP. Each individual square comprises of 9 (3 by 3) smaller squares. Without loss of generality, rows and columns represent the genotype of the pseudo-father and pseudo-mother respectively. Therefore each smaller square represents a particular pseudo-trio genotype configuration and is coloured according to its equivalent emission score. ϵ = the probability of a genotyping error; ζ = the probability assigned to the triple heterozygote.

2.2 The Hidden Markov Model

Table 2.1 **Weighted Mendelian emissions.** For the classic Mendelian transmissions, $\zeta = 0.5$ but for the purpose of phasing we would typically use a weighted approach with $\zeta = 0.01$, reducing the frequency of triple heterozygotes. Formally, all equivalent entries over the 3 blocks should sum to 1. Since, however, ϵ is typically small ($\sim O(10^{-2})$) we adopt these approximations. f_j, m_j, x_j = paternal, maternal, child's genotype at a site j , respectively.

$x_j = 0$				$x_j = 1$				$x_j = 2$			
f_j, m_j	0	1	2	f_j, m_j	0	1	2	f_j, m_j	0	1	2
0	1	$\frac{1}{2}$	ϵ	0	ϵ	$\frac{1}{2}$	1	0	ϵ^2	$\frac{\epsilon}{2}$	ϵ
1	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{\epsilon}{2}$	1	$\frac{1}{2}$	ζ	$\frac{1}{2}$	1	$\frac{\epsilon}{2}$	$\frac{1}{4}$	$\frac{1}{2}$
2	ϵ	$\frac{\epsilon}{2}$	ϵ^2	2	1	$\frac{1}{2}$	ϵ	2	ϵ	$\frac{1}{2}$	1

Table 2.2 **Compatibility emissions.** These emissions mark whether the query individual's genotype at a site j (x_j) could have been inherited from a father and mother with genotypes f_j and m_j respectively at that site. The maximum score is obtained when both parents are compatible.

$x_j = 0$				$x_j = 1$				$x_j = 2$			
f_j, m_j	0	1	2	f_j, m_j	0	1	2	f_j, m_j	0	1	2
0	1	1	ϵ	0	ϵ	1	1	0	ϵ^2	ϵ	ϵ
1	1	1	ϵ	1	1	1	1	1	ϵ	1	1
2	ϵ	ϵ	ϵ^2	2	1	1	ϵ	2	ϵ	1	1

2.2.3 The Viterbi Algorithm - Identifying and Using Pairs of Individuals

The Viterbi algorithm (an overview of which can be found in Foreman [1992]) identifies the optimal path of hidden states (here the sequence of parent-pairs along the genome), given the observed sequence. This is the path of hidden states along sites which, compared to all other combinations of states, will give our observed sequence the highest achievable likelihood. If x_j is the genotype of the ‘child’ at site j and m is the total number of pairs of candidate parents ($\binom{n}{2}$), the algorithm is as follows:

$$\begin{aligned} \text{Initialisation: } V_{i,1} &= \frac{1}{m} \cdot e_{i,1}(x_1) \quad \forall i \in [1, m] \\ \text{Recursion: } V_{i,j} &= e_{i,j}(x_j) \times \max_k \{V_{k,j-1} \cdot t_{k,i}\} \\ &\quad \forall k, i \in [1, m], \quad j \in [2, L] \end{aligned} \tag{2.2}$$

Using Equation 2.2 we build the Viterbi matrix $V = (V_{i,j})$, $i \in \{1, \dots, m\}$, $j \in \{1, \dots, L\}$, normalising each column with $c_j = \max_i \{V_{i,j}\}$, and storing the traceback at each step: the best pair $(k, j-1)$ that should precede each current pair (i, j) in the recursion step of Equation 2.2. We have that $V_{i,j}$ is the probability of the observed sequence until site j given state i and that we arrived there following its traceback. Note here that this traceback for state i is the best possible path that can bring us there. In other words, no other path that ends at i will give the child a greater likelihood. Equivalently, each entry in the last column of V is the likelihood of the ‘child’ given the path that leads to that entry. Without loss of generality, suppose each entry is unique, i.e. $\nexists i, k$ s.t $V_{i,L} = V_{k,L} \quad \forall i, k \in \{1, \dots, m\}$. We may thus identify the optimal path of hidden states by looking at the traceback of state f for which $V_{f,L} > V_{i,L} \quad \forall i \in [1, m] / f$.

The Viterbi algorithm is the backbone of our model: at every SNP it gives us a single pair of individuals which we hope contains the true NN haplotypes of the query genotype and/or a set of pseudo-parents which we can use for haplotype estimation. It is also the computationally expensive part of our model: at every SNP, to compute the likelihood for a given pair of pseudo-parents requires an

2.3 Application 1: Haplotype Estimation

iteration that scales quadratically with the sample size. We improve on this by allowing only single transitions (as previously described), yet the algorithm remains slow even for modest sample sizes.

The remainder of this chapter now splits into two parts. First, we examine the remaining steps of the algorithm for the purposes of phasing: how do we collect more paths? How do we use each to phase? How do we combine the predictions across all of them? Next, we turn to the identification of NNs and detection of extended haplotype sharing. We explore the F/B algorithms as an alternative to the Viterbi algorithm for obtaining the posterior distribution on ‘relatedness’ thus capturing the full relatedness patterns present in the data. Finally, we discuss two approximations to the Viterbi algorithm: the graph method and the HMM with subsampling.

2.3 Application 1: Haplotype Estimation

2.3.1 The Second best and other paths

For haplotype estimation we require pairs of candidate parents that are informative. As discussed earlier, for all of our choices of emission scores we expect some proportion of SNPs to be unphasable either due to triple heterozygotes or because the optimal choice of pseudo-parents is incompatible with the query genotype at some sites. What is more, it is often the case (data not shown) that the difference in likelihood between the higher ranking paths is small, suggesting that information across paths might yield more reliable haplotype estimates. We thus continue with a discussion on the best strategy for collecting more paths. In principle, a formal approach to this problem would be one that incorporates a mechanism for identifying a set of paths with decreasing likelihood. We therefore present the full approach that we developed (independently of Foreman [1992]) here and more particularly, we begin by exploring how the second best path can be identified (the path with second highest likelihood).

2.3 Application 1: Haplotype Estimation

Suppose the optimal path is called A and the second best path (a suboptimal path) is called B . B could be the traceback obtained from state k at site L s.t $V_{i,L} > V_{k,L} > V_{j,L} \forall j \in [1, L]/\{i, k\}$, i.e. the traceback from the state with the second highest value in $V_{.,L}$. However, this is not the only way that the second best path could have been derived. Recalling the construction of V , it is possible to find the second best path by considering deviations away from the optimal path. In the recursion step of Equation 2.2, for each state i , we consider only the maximal state k , in the previous column as the state we could have come from. If instead, at a particular site, j , we considered the second best hidden state, l , for our traceback (rather than the best possible pair at the previous site), then we could obtain a path which from states L to j follows the traceback of A and from j backwards follows the traceback of state l . Call this path U . $P(P(U) > P(B)) > 0$ and we claim that B and U are the only two candidates for the second best path.

Consider a path S and let S_j be the hidden state of S at site j . Now, the second best path either will or will not end at A_L . If it does not, then by construction of the Viterbi algorithm it must be B since B_L has the second highest value in $V_{.,L}$ and B is the best possibly achievable path that ends in B_L (recall that we have assumed that values in $V_{.,L}$ are unique).

If the second best path does end in A_L then it must be the best path deviating away from A at some site j (as proposed by Foreman [1992]). The fact that at $V_{i,j}$ we will now have this second best value will have no influence on our considerations of other paths, other than U , which overlapped with the optimal path at this entry. This is because any other path, U' , that will be affected by the change must, by construction of the Viterbi algorithm, have a likelihood less than that of B , i.e $P(U') < P(B)$, and so is not a candidate for the second best path.

There remains therefore only the question of which site $j \in [2, L]$ is the best site at which the deviation should occur. Let U^j be the path that is derived from deviating at site j . The path that should be chosen is the one with the highest relative likelihood. One approach would be to calculate the full relative likelihood for each and then observe which one is highest after renormalising. Since, however, each U^j is identical to A from sites j to L , instead of comparing

2.3 Application 1: Haplotype Estimation

the suboptimal paths to each other, we may compare their relative likelihood to that of the optimal path until the deviation site. In other words, we look for the suboptimal path U^j which at site j has relative likelihood that is ‘closest’ to that of the optimal path. $\forall j \in [2, L]$, if $A_j = i$, we record:

$$V'_{i,j} = e_{i,j}(x_j) \times \text{submax}_k \{V_{k,j-1} \cdot t_{k,i}\} / c_j \quad (2.3)$$

(where submax refers to choosing the second highest value), obtaining thus the new Viterbi score for site (i, j) . Let $|A_j| = V_{A_j,j}$. We choose the path U^j that maximises $\frac{V'_{i,j}}{|A_j|}$. Recall that $|A_j|$ is the likelihood of the query genotype until site j given that we have followed the path ending at A_j . Taking the ratio of the two values cancels the effect of the normalising constants and emission probabilities (which are the same for both paths) and thus allows direct comparison of the different suboptimal paths. For simplicity we complete Table 2.3, which records all the above information.

Table 2.3 Searching for U . The table contains a row for every SNP and row j contains the relevant information for the optimal and suboptimal paths, if the deviation was to occur at SNP j . L^{max} is the likelihood of the optimal path A until the site in question, $|A_j|$; T^{max} is the state in the previous site (the traceback) that the optimal path comes from, A_{j-1} ; L^{submax} is the likelihood of the suboptimal path U until that site, $|U_j|$; T^{submax} is the state with the second highest value in the previous column, U_{j-1} . $U_j = A_j$.

Site	L^{max}	T^{max}	L^{submax}	T^{submax}	$\frac{L^{submax}}{L^{max}}$
L	$ A_L $	A_{L-1}	...		
$L-1$	$ A_{L-1} $	...			
$L-2$	$\vdots$				
$\vdots$					
2					

We choose U^j by identifying the row which has the highest value in the last column of Table 2.3. In practice in this particular setup, we disallow deviations

2.3 Application 1: Haplotype Estimation

from site 2 as otherwise, due to similar values of Viterbi scores being present along $V_{.,1}$, most deviations would take place there. For our choice of U , we calculate the full relative likelihood of this path (by multiplying with the appropriate transition probabilities and emission probabilities and dividing by the normalising constants, all of which are already known) and by comparing to the relative likelihood of B , we identify the second best path for our ‘child’.

It is straightforward to extend this method for finding the third best path by looking at the second or third highest entries of $V_{.,L}$ and at the paths deviating away from A , B and U . We can then extend further for the fourth best path and so on. By “likelihood order” we mean the path’s rank with respect to likelihood: first order refers to the optimal path, second order refers to the second best path and so on. We have also been abusing the term “path likelihood” to mean the observed sequence’s likelihood given the particular path of hidden states. The number of paths to consider increases with the the order of likelihood that we are interested in. We will be discussing the methods for using the various suboptimal paths that we will collect in Section 2.3.3.

Actually, Foreman [1992] independently presented the generalisation of this strategy for directly obtaining the top K (possibly overlapping) paths by considering the best K states to have come from in each recursion of the Viterbi algorithm. We do not adopt this approach as it is considerably more computationally expensive than the classic Viterbi and would require prohibitively greater memory requirements during the iterations in Equation 2.2. Even with our current setup, vast amounts of RAM ($\sim 30\text{G}$ for a single individual in a sample of ~ 350 individuals and for $\sim 30,000$ SNPs) are required for storing the Viterbi and Traceback arrays. Moreover, for the purposes of phasing we would only typically use $K < 10$ (see Chapter 3; Figure 3.7), and what is more, we know (Chapter 3; Figure 3.2) that using a set of less likely paths does not compromise overall phasing performance. A strict approach to the level of suboptimality incorporated in the model is unnecessary, at least in terms of maintaining high accuracy. Foreman’s strategy (and therefore the strategy we also presented here) explicitly does not guard against paths overlapping, which in some sense in this context defeats the purpose of collecting them in the first place: more paths are necessary in order to ensure that we have at least one informative path at each heterozygous

2.3 Application 1: Haplotype Estimation

site. Finally, it is in fact possible to extend our more ad-hoc approach in order to recover these paths after running Equation 2.2. We may in fact iterate back to recover as many suboptimal paths as necessary, while Foreman’s approach demands in-advance knowledge of the value of K .

We initially implemented the ‘full’ approach to recovering paths in the way just described. However, our initial results (presented in Chapter 3) do not support this strategy for phasing performance to be maintained as path likelihood does not correlate with phase accuracy. In practice, therefore, we only pick the best K entries from the last column of V . These paths are more informative for phasing as there is less overlap between them (data not shown). Moreover, since path likelihood is not correlated with performance it is unnecessary to perform these more computationally expensive iterations and, what is more, B is much more often the 2nd best path compared to U (see Chapter 3 for more details). When there are non-unique entries in $V_{.,L}$ we sample one path uniformly and at random from the set of paths with the same likelihood order and therefore the K best paths are each random picks from the top K likelihood sets. It was, however, worth discussing the ‘correct’, likelihood-based approach for picking paths as it is not obvious that under complicated demographic histories it will always be the case that performance will not be affected when using less likely paths.

A heterozygous site is phasable if there is at least one parent-pair in the sample which is informative at that position. Consequently, for each site that has not been phased by the top K paths it is possible to identify the best path that will phase it. To achieve this for each such site j , we select the best informative parent pair from $V_{.,j}$ and obtain the traceback. To complete the sequence forwards we rerun the Viterbi algorithm once more with the order of the SNPs reversed and repeat the procedure. All in all therefore, to ensure a fully phased sequence the Viterbi algorithm would need to be run twice, collecting the set of K paths and enriching it as necessary such that there is at least one informative path for every heterozygous site.

The caveat here is that this procedure requires fully stored Viterbi and traceback matrices which has prohibitive memory requirements for even modest sample sizes when entire chromosomes are being analysed. As a result our present algo-

2.3 Application 1: Haplotype Estimation

rithm will leave some sites unphased. In our implementation, the default value for K is 10 and any remaining unphased sites are phased randomly.

2.3.2 Phasing with a path of pseudo-parents

We revisit here some of the core concepts of our model for the purpose of phasing. We have aimed to identify the two individuals from the sample that may represent the true parents of a query genotype at a SNP. If we have successfully selected them then it follows that the pseudo-parental genotypes at that site are the same as the parental genotypes. Phasing a heterozygous site is then achieved by transmission. For example, if the genotype of the pseudo-mother is 0 and that of the pseudo-father is 2 then the ‘child’ must have inherited the alternative allele from the pseudo-paternal haplotype.

More generally, suppose we have an optimal path of pseudo-parents, $S = (M, F)$, where M is the sequence of pseudo-mothers and F is the sequence of pseudo-fathers. Then (M_1, F_1) is the optimal pair of individuals selected at the first SNP. Given the genotypes of M and F at each position, we determine the child’s two haplotypes, (H_1, H_2) , inherited from M and F respectively as demonstrated in the toy example below (Example 0, Table 2.4).

The example illustrates the two cases that we have already discussed, in which phasing by transmission is impossible. When both parents carry the same alleles (first pair of ‘?’s), we have an incompatibility, indicative of a de novo mutation or genotyping error. If both parents are heterozygous (second pair of ‘?’s) it is similarly not possible to determine from which parent the child has inherited the derived allele.

In this example we are assuming that it is straightforward to assign the two individuals selected at each SNP to the pseudo-mother or pseudo-father group. This is trivial here because there is a single pair of individuals that is optimal for the entire sequence. If only one of the parents changes along the sequence at some SNP j then we can assign the individuals in the two groups unambiguously in this case as well. To illustrate this, suppose that the optimal pair of individuals is (A, B) from site 1 to j and (A, C) from j to L . Then without loss of generality

2.3 Application 1: Haplotype Estimation

Table 2.4 **Example 0: Phasing by transmission.** Each site is considered separately by determining which parent has transmitted the ancestral allele and which the derived allele. In cases where this is impossible to distinguish, a prediction cannot be made and this is marked by a ‘?’ in the diagram. (H_1, H_2) are the haplotypes of the child, inherited from M and F respectively.

$M:$	0	0	1	1
$F:$	2	0	0	1
child:	1	1	1	1
			↓	
$H_1:$	0	?	1	?
$H_2:$	1	?	0	?

$M_i = A$ everywhere (i.e. $\forall i \in \{1, \dots, L\}$) and $F_i = B$ for $i \in \{1, \dots, j-1\}$ and $F_i = C$ for the remainder of the sequence (i.e. $i \in \{j, \dots, L\}$).

Problems arise if double jumps are allowed to occur: when both parents change between two adjacent SNPs. In such instances it is impossible to know how best to align the four sets of transmitted haplotypes into two estimates for the haplotypes of the ‘child’. Misalignment may then lead to phasing errors. To illustrate this, consider the toy data from Example 1 below (Table 2.5).

Let the optimal pair of individuals is (A, B) at the first two positions and (C, D) at the next two. Without loss of generality, there are two possibilities for aligning the two groups. In the first possibility, $M = \{A, A, C, C\}$ and therefore $F = \{B, B, D, D\}$ and the phasing results are demonstrated in Case 1 of Table 2.6. However, it is equally possible that $M = \{A, A, D, D\}$ and $F = \{B, B, C, C\}$ in which case we get the diagram configurations in Case 2 of Table 2.6. In Case 2, H_1 and H_2 are as in Case 1 for the first two sites, but then flip over for the next two. Using our model, it is impossible to infer which of the two is the correct

2.3 Application 1: Haplotype Estimation

Table 2.5 **Example 1: A toy data set of genotypes.** The data here consists of four individuals $A - D$ and the query individual, the ‘child’. Genotypes are shown for four SNPs.

A:	0	0	1	1
B:	2	1	0	1
C:	1	0	2	0
D:	1	0	1	1
child:	1	1	1	1

orientation. This is the reason for which restricting to single jumps is advantageous. When only single transitions are allowed, the optimal path could instead be (A, B) at the first two sites, then (A, C) at site 3 and (D, C) at the last site. Then, we unambiguously infer that $M = \{A, A, A, D\}$ and $F = \{B, B, C, C\}$ and the prediction for the child’s haplotypes is now also unambiguous, as illustrated in Case 3 (Table 2.6).

Note again that this restriction has little effect on the choice of pseudo-parents, as what previously may have been a double jump (which was a rare event; data not shown) is typically translated into two (near) consecutive single transitions (data not shown). In this way, the overall choice of pseudo-parents is negligibly affected, while allowing for the transmitted haplotypes’ orientation to be resolved. Finally, recall the second major advantage of single transitions: the restriction of the hidden states considered during the iterations of the Viterbi algorithm and the great improvement that follows in computational efficiency.

2.3 Application 1: Haplotype Estimation

Table 2.6 Aligning pseudo-parents and phasing by transmission as demonstrated in Example 0 (Table 2.4). The parental genotypes for M and F are obtained using the the data from Example 1 (Table 2.5). Case 1: $M = \{A, A, C, C\}$; $F = \{B, B, D, D\}$; Case 2: $M = \{A, A, D, D\}$; $F = \{B, B, C, C\}$; Case 3: $M = \{A, A, A, D\}$; $F = \{B, B, C, C\}$

	<u>Case 1</u>			
M:	0	0	2	0
F:	2	1	1	1
	↓			
H_1 :	0	0	1	0
H_2 :	1	1	0	1
	<u>Case 2</u>			
M:	0	0	1	1
F:	2	1	2	0
	↓			
H_1 :	0	0	0	1
H_2 :	1	1	1	0
	<u>Case 3</u>			
M :	0	0	1	1
F :	2	1	2	0
	↓			
H_1 :	0	0	0	1
H_2 :	1	1	1	0

2.3.3 Combining path predictions

We have presented how haplotype estimation is achieved when a single path of pseudo-parents is being used. We must now consider how to phase the query genotype when multiple pseudo-parental paths have been collected. Each path when considered in isolation should produce its own pair of haplotype estimates, though one would hope that overall there will be great consistency across the different predictions (up to a permutation of the two haplotypes). There are two different strategies that we may adopt: either via ‘Left to Right’ (LTR) approach by which a prediction is made at each SNP in turn using all the available paths, or via a ‘Top to Bottom Scaffolding’ (TTB) approach, in which we initialise the process with an estimate of the two haplotypes across the whole region and then update this estimate using the full, separate path predictions.

Both the ‘Left to Right’ and the ‘Top to Bottom Scaffolding’ strategies may be implemented in various ways and we explored a range of approaches for each. A detailed description for all the attempted methods follows in the next sub-subsection, though first we present the method that yielded the most accurate results and that was used for analysing real data (see Chapter 3). We discovered that a ‘Left to Right’ approach that puts equal weight on the predictions of each path at each site was the most accurate and best calibrated method among all attempted strategies. We call this method the SM (Simple Means) approach, and also describe it in more detail in the next sub-subsection as a specific instance of a general Left to Right strategy. In our phaser, therefore, to integrate the predictions of the optimal and suboptimal surrogate parents into a final pair of haplotypes we simply choose the orientation of the heterozygote that is favoured by the greater number of paths at each SNP. The full SM algorithm is given in Algorithm 1.

Algorithm 1 Combining path predictions: Simple Means

1. Obtain separate haplotype estimates from each path
2. Orientate the path predictions such that all paths agree at the first site and fix the prediction at that site

2.3 Application 1: Haplotype Estimation

3. For each subsequent site, predict the orientation of the heterozygote by choosing the one that agrees with most paths
4. For the paths that disagreed with our selection, flip over their predictions from the current site onwards
5. Record the proportion of informative paths that had agreed, setting it to be the prediction probability for that site

We have found (see Chapter 3) that the accuracy of a path's predictions is largely uncorrelated with the path likelihood and therefore a more sophisticated weighting scheme for combining path predictions would offer no additional predictive power. It is therefore no surprise that this is the strategy that performed best.

2.3.3.1 Other attempted methods

Left to Right Approach

This approach is the generalisation of the SM method, in which we build on our confidence for the phase of every SNP progressively, using the predictions of all paths at each location and considering only the SNPs that we have previously phased. To illustrate this procedure, let $H_i = \{H_{i1}, H_{i2}\}$ be the pair of haplotypes predicted by path i , suppose we have K paths and L SNPs (at which the child is heterozygous) and let a bullet represent the predicted sequence until the current site. Overleaf is an illustration of how the method works.

2.3 Application 1: Haplotype Estimation

An illustration of the left to right approach

site 1	→	site 2	→	site 3	→	...	→	site j	→	...	→	site L
0 or 1		●0 or ●1		●0 or ●1		...		●0 or ●1		...		●0 or ●1
$H_{11,1}$		$H_{11,2}$		$H_{11,3}$		...		$H_{11,j}$		...		$H_{11,L}$
$H_{21,1}$		$H_{21,2}$		$H_{21,3}$		...		$H_{21,j}$		...		$H_{21,L}$
⋮												
$H_{K1,1}$		$H_{K1,2}$		$H_{K1,3}$		...		$H_{K1,j}$		...		$H_{K1,L}$

The diagram shows the first haplotype predicted from a set of K paths (rows; $\{H_{11}, \dots, H_{K1}\}$) along L sites (columns). Starting from left to right, at each position the paths vote on whether the first haplotype of the child should carry the 0 or the 1 allele. After each prediction and before the next one, the paths are re-aligned such that they agree on all preceding sites.

We examine each site separately, considering just the sites preceding it. At every SNP we decide whether the first haplotype should have a 0 or 1 allele and set that to be our prediction for that site. After the prediction is made, we flip over the predictions (interchange H_{i1} and H_{i2}) from that site onwards of all the paths in our collection whose configuration disagreed with the decision made. As a result, all paths will have the same haplotypes until the current site j , which accounts for the fact that the predictions of the various paths are unordered and so could be permutations of each other: what is the ‘first haplotype’ (H_{i1}) of one path could initially be the ‘second’ one for another path ($H_{j2}, i \neq j$). Re-orienting the predictions of the paths whenever they disagree with the call made at a SNP also guards against further phasing errors. This is because assuming our call is correct, when a path disagrees it suggests that at that position its haplotypes are arranged with the opposite orientation to our estimated haplotypes. Therefore, swapping them over ensures that they re-align with our overall prediction. Following this approach ensures that during the iteration all path predictions agree on phase for all sites until the current one.

To count the votes on the prediction at any one SNP we may:

2.3 Application 1: Haplotype Estimation

1. *Simple Means (SM)*: obtain the simple mean m over all predictions at each SNP and set this to be the final prediction. This is the best strategy that we have already described – we are only putting it in the context of similar approaches here. The probability of making the prediction is $\max\{m, 1 - m\}$, i.e. the proportion of the paths that agreed with it. This method performed best and is the one implemented in our phaser.
2. *Weighted Means (WM)*: use the mean over all predictions at each SNP, weighted by their relative likelihoods. To obtain a probability for each prediction, consider the following. Call a particular phase at a SNP R . Then one way to score R is by $Q_2 = \sum_{i \in A_R} L_i$, where A_R is the set of paths that favour orientation R and L_i is the likelihood of path i . We can similarly calculate Q_2^{-1} for the opposite prediction at that site and then the probability for our prediction is $\frac{\max\{Q_2, Q_2^{-1}\}}{Q_2 + Q_2^{-1}}$.
3. *Best Candidate (BP)*: at each SNP choose the path with the highest likelihood that is informative of phase at that SNP and use just this pair of parents to phase from. If path j is selected, we set the probability of our predictions to be $\frac{L_j}{\sum_k L_k}$.

Top to bottom Scaffolding Approach

Suppose that it is possible to “guess” a consensus pair of haplotypes in advance, before using all the individual path predictions. In the Top to Bottom Scaffolding (TTB) approach we revisit our “guess” at each SNP, then modify it by flipping the orientation of the prediction at each SNP in turn if necessary. To decide on whether to flip we consider whether our collection of paths support the current orientation over the previous one for that particular SNP, *given the entire sequence*. We use the same annotation as before, with $H_{i1,j}$ referring to the first haplotype of path i until site j (so that $H_{i1,L} = H_{i1}$). Considering just the first haplotype (H_{i1}), at the top of page 65 is an illustration of this scheme. In practice, the consensus is initialised by setting it as the prediction of the optimal path. If there were any unphased sites which can be phased by other paths in our set, we initialise these by phasing them at random.

2.3 Application 1: Haplotype Estimation

To assess the consensus we consider the following: assuming we knew the true haplotypes, we would like to add more weight to paths that have predictions closer to the truth and less weight on paths whose predictions carry more errors. We define a new score that will incorporate both of these criteria, Q_1 , and compute it for each path. Q_1 considers the likelihood and the number of switch errors (a type of phasing error; see Chapter 3, Section 3.2 for the full definition) realised when comparing a path's prediction to our prediction until the current site. If K is the number of paths, L_i the likelihood of path i , and s_i the number of switches that it leads to when comparing its pair of predictions against our estimated haplotypes until the current SNP, at each SNP we consider:

$$Q_1 = \frac{\sum_{i=1}^{i=K} L_i s_i}{\sum_{i=1}^{i=K} L_i} \quad (2.4)$$

by evaluating Q_1 at the predictions for the whole sequence. We then recalculate this score considering the whole sequence, but with the orientation of the predictions of the current SNP reversed (call this second score value Q_1^{-1}). The orientation that minimises our score is set as the new current prediction, and an ad-hoc score $\in [0,1]$ that can be used for calibration is $\frac{\max\{Q_1, Q_1^{-1}\}}{Q_1 + Q_1^{-1}}$. We then proceed to the next site.

Following the illustration overleaf, at a site j we are trying to decide whether the first haplotype should have the 0 or 1 allele by considering the predictions of all the paths over the entire sequence. After deciding for that site and before examining the next one, we flip over the predictions from that site onwards for the paths that had the opposite allele. In the toy example the predictions for path H_1 would be flipped from sites j to L . In the end, the derived, modified consensus is our final estimate for the haplotypes of the child. We set the calibration score at each heterozygous site to be $\frac{\max\{Q_1, Q_1^{-1}\}}{Q_1 + Q_1^{-1}}$.

2.4 Application 2: Nearest Neighbours and haplotype sharing

An illustration of the Top to Bottom Scaffolding approach

Sites:	1	2	3	...	j	...	...	L	
Original Consensus:	0	0	1	...	1	...	0	0	1
Alternative Consensus:	0	0	1	...	0	...	0	0	1
H_{11} :	0	0	1	...	1	...	0	0	1
H_{21} :	0	0	1	...	0	...	0	0	0
					⋮				
H_{K1} :	0	0	0	...	0	...	0	0	1

Obtain first a consensus prediction. Then revisit each SNP and consider an alternative consensus which has the orientation at the current SNP reversed. The diagram shows the first haplotype predicted from a set of K paths (rows; $\{H_{11}, \dots, H_{K1}\}$) along L sites (columns). Revisiting SNP j : here the set of paths $\{H_1, \dots, H_k\}$ would most likely support the alternative orientation.

2.4 Application 2: Nearest Neighbours and haplotype sharing

So far we have discussed the application of our model to haplotype estimation. We now consider the details for the alternative scenario where we investigate whether we identify the query genotype's NNs at every position along the genome. By observing the regions over which a NN does not change, we have effectively identified the region over which the query genotype shares a haplotype with that NN (for an illustration you may refer back to Figure 1.2). These shared haplotypes are not necessarily IBD chunks, in that they could be shared with NNs with which the TMRCA precedes the theoretical threshold assumed by IBD approaches. They are nevertheless shared chunks in their own right and may be used to characterize relatedness levels present in a sample.

In this context, collecting suboptimal paths is not useful. Rather, we will discuss the Forward/Backward algorithms as alternatives to the Viterbi algorithm

2.4 Application 2: Nearest Neighbours and haplotype sharing

for searching for NNs. Finally, we present two separate methods that were developed to approximate the full HMM, such that identification of NNs and shared haplotypes is still accurate, but significantly more computationally efficient. The two methods are a graph approximation to our HMM and the original HMM but with a subsampling strategy that massively reduces the state space at each position.

2.4.1 The Forward/Backward Algorithms

Identifying an individual's nearest neighbours in a sample can be very useful, as we have previously discussed. For certain applications, obtaining a single pair of individuals at every SNP is sufficient for capturing local structure present in a data set. The Viterbi algorithm achieves precisely this: it identifies a single best path of individual pairs along the genome and provides a straightforward method for identifying shared haplotypes between them and the query genotype. This approach, however, offers no insight on the extent of suboptimality that the true NNs might have under our model. Finally, it also does not explore the full posterior distribution of NN-candidacy of all the individuals in the sample. The posterior of an individual being a NN of the query genotype at a certain position is a metric for the relatedness between them at that position. Generalising therefore over the entire sample, this approach could offer a new measure for local relatedness. It is therefore of interest to establish whether the Forward/Backward (F/B) (Rabiner [1989]) characterizes local coancestry at least as effectively as the Viterbi algorithm.

We employed the F/B algorithms to investigate whether in the cases that the Viterbi algorithm did not identify the true NNs this was due to chance (that is, the NNs could have been picked during the iteration but were not due to non-uniqueness of likelihoods) or whether in fact the posterior of the true NN pair was suboptimal, suggesting that sampling a path of NNs is in fact unlikely using our HMM. The F/B algorithms each calculate the likelihood of a path of hidden states until a site, conditional on the observed sequence from start/end until that

2.4 Application 2: Nearest Neighbours and haplotype sharing

site. Using the same notation as in Section 2.2.3, the Forward algorithm is:

$$\text{Initialisation: } F_{i,1} = \frac{1}{m} \cdot e_{i,1}(x_1) \quad \forall i \in [1, m]$$

$$\text{Recursion: } F_{i,j} = e_{i,j}(x_j) \times \sum_k \{F_{k,j-1} \cdot t_{k,i}\} \quad \forall k, i \in [1, m], j \in [2, L]$$

$$P(\text{observed sequence}) = \sum_{k=1}^{k=m} F_{k,L}.$$

Similarly, the Backward algorithm is:

$$\text{Initialisation: } B_{i,1} = 1 \quad \forall i \in [1, m]$$

$$\text{Recursion: } B_{i,j-1} = \sum_k e_{k,j}(x_j) \times \{B_{k,j} \cdot t_{i,k}\} \quad \forall k, i \in [1, m], j \in [2, L]$$

$$P(\text{observed sequence}) = \sum_{k=1}^{k=m} e_{k,1} \cdot B_{k,1} \cdot \frac{1}{m}$$

Then, the posterior probability of being at hidden state i at site j given the sequence is $P_{i,j} = \frac{F_{i,j} \cdot B_{i,j}}{\sum_k F_{k,j} \cdot B_{k,j}}$. It further follows that the posterior of a particular individual, A say, being a NN at site j is $P_j^A = \sum_B P_{B,j} \quad \forall B$ in the remaining individuals.

The F/B algorithms assess all the choices of pseudo-parents site-wise, providing the full distribution of posterior probabilities for all individual pairs. In principle it is possible to look at the set of pairs with maximal posterior at each site and decode the minimum number of transitions required along the sequence in order to identify maximum haplotype sharing, in the way that is achievable using the results from Viterbi. We do not explore this further here.

As with the Viterbi algorithm, in their current implementation the F/B algorithms cannot be implemented on large data sets, both due to speed and due to memory requirements. A sparse approach would be more tractable, as is our HMM with subsampling (HMMsub).

2.4.2 The Graph

The algorithm for identifying pseudo-parents has a state-space that scales quadratically with the number of samples. Consequently, for moderate sample sizes (e.g.

2.4 Application 2: Nearest Neighbours and haplotype sharing

>1,000 individuals) it becomes impractical to implement, both in terms of speed but also in terms of memory requirements. For example, for a sample of 400 individuals for the whole of chromosome 20 ($\approx 30,000$ SNPs), the algorithm requires 12Gb of RAM and about two weeks to obtain results for a single individual. We have therefore explored strategies for improving efficiency that enable analysis of large samples with very little loss of accuracy. This subsection presents the first such method that was attempted. The method gains computationally by:

1. Using a graph-based method to summarise variation in a data set
2. Prioritising certain pairs of candidate pseudo-parents

Given a starting position a set of genotype sequences can be represented by a tree structure, T , specifically a suffix tree. Such representations have been used for haplotype data, for example, in the haplotype-graph approach of BEAGLE (Browning and Yu [2009]), in imputation of HLA alleles (Dilthey et al. [2013]) and in the representation of extended haplotype structures arising from selective sweeps (Sabeti et al. [2007]). In this project, we have adopted an equivalent approach which is genotype-based.

2.4.2.1 Constructing a genotype graph

The genotype tree T is constructed from left to right, starting at the first SNP: initialise the tree with as many nodes as there are unique genotypes (coded as 0, 1 or 2 copies of the alternative allele) present in the sample at that first SNP. At every subsequent position j , T is extended by adding one new node for each unique genotype that follows each of the nodes at $j - 1$. These nodes are labeled as $T_{j,i}$, $i \in [1, N]$, where N is the total number nodes present at that SNP. The edges join nodes in adjacent positions such that they represent the observed combinations of genotypes at these positions. Therefore, nodes $T_{j-1,*}$ and $T_{j,*}$ are joined by an edge, if in the data there exists an individual that at $j - 1$ and j carries the genotypes represented by these nodes respectively. It follows therefore that each node will have:

1. One incoming edge, from its parental node,

2.4 Application 2: Nearest Neighbours and haplotype sharing

2. At least one and up to three outgoing edges.

For an illustration of the construction process refer to Figure 2.2. In the example in the figure, there are 4 individuals with genotypes at 3 SNPs and a single tree can be constructed to represent the data (depicted in the middle of the diagram). There are two unique genotypes present in the sample at SNP 1, namely 0 and 2, and therefore the graph representation of the data is initialised with 2 nodes, $T_{1,1}$ and $T_{1,2}$, each of size 2 (since both carry two individuals, (A, B) and (C, D) respectively). Looking next at SNP 2, all individuals that had 0 at SNP 1 have genotype 0 at SNP 2 and therefore node $T_{1,2}$ is extended, by an edge, to just one node $T_{2,1}$, of size 2. In contrast, the 2 individuals that had genotype 2 at SNP 1, now have genotypes 0 and 2. Therefore $T_{1,2}$ is extended into two nodes, $T_{2,2}$ and $T_{2,3}$. Following the same logic, we extend again obtaining 4 nodes for SNP 3, at which point we have identified each individual unambiguously and the iteration ends. The purpose of the bottom tree in the figure is to illustrate how the node labelling system works.

We see, therefore, that as the suffix tree grows from left to right away from the starting position, the number of nodes present at each position increases until the number of leaves equals the number of samples. Then, the genotype of every individual may be recovered unambiguously by following up the tree from the relevant leaf back to the root. In the example in Figure 2.2 a single tree is sufficient to represent the data for the whole sequence. In general, however, this is typically not the case; tree size depends on the levels of variation present in the sample, with lower levels leading to trees that span greater genomic distances. If a tree ends before the end of the sequence is reached, then a new tree can be constructed from the ending position (the breakpoint) onwards.

2.4.2.2 Using a genotype graph to select pseudo-parents

The purpose of representing the data in the form of a genotype graph is to redefine the hidden state-space of candidate pseudo-parents. In the original representation of the data, the variation present at a SNP j is captured by the full list of genotypes. Given a tree structure T , however, this variation can be summarised

2.4 Application 2: Nearest Neighbours and haplotype sharing

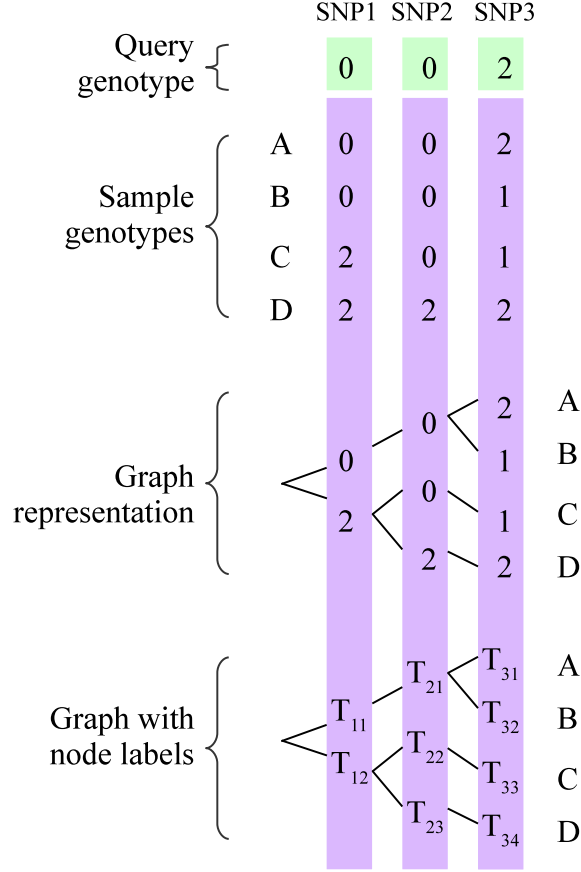


Figure 2.2 **Building a genotype graph.** The figure is split into 3 main parts. At the top is a toy sample of genotypes, consisting of 4 individuals, A-D over 3 SNPs. The middle part contains the graph representation of these genotypes while the bottom is an illustration of the same graph, consisting of node labels: $T_{i,j}$ is the name of node j on SNP i . The graph is constructed left to right: at the first SNP there are as many nodes as there are unique genotypes in the data at that SNP. Each node is then extended to the right: it is joined by edges to nodes on the next SNP for all the possible genotypes that can follow the current node, given the data. This is continued until the end of the sequence is reached (as is the case here) or until all the individuals can be identified uniquely (as is also the case here).

by the set of nodes $T_{j,\star}$ present at that SNP. In the original HMM we selected pseudo-parents by considering pairs of individuals; now it is sufficient to consider pairs of nodes instead. Suppose the number of nodes at j is N_j and that two

2.4 Application 2: Nearest Neighbours and haplotype sharing

unique nodes are $T_{j,a}$ and $T_{j,b}$ (i.e. $a, b \in \{1, \dots, N_j\}, a \neq b$). Then, the pair of nodes $(T_{j,a}, T_{j,b})$ represents all the pairings between the individuals that live on node $T_{j,a}$ and those that live on $T_{j,b}$. A node can pair with itself if it has size ≥ 2 .

Moving along a sequence results in an increase in the number of nodes in the tree such that the number of states approaches the number of samples. Therefore, while the number of pairs considered earlier (further to the left) in the tree is significantly smaller than $\binom{n}{2}$, this number increases to $\binom{n}{2}$ by the end of the tree. Nevertheless, even from this change in representation in large parts of the sequence the state space has been greatly reduced making this a more computationally tractable approach.

Each pair here is assessed in a similar fashion to assessing individual pairs, namely by considering an emission score. This score will depend on the genotypes represented by the two nodes, and by the genotype carried by the ‘child’. Here, it is not enough to assign an emission score to each node pair separately for each SNP. Instead, at any given position the score must reflect both the compatibility at the current position but also the compatibility of the entire traceback of the two nodes along T leading up to it. Recall from Table 2.2 that the emission score of a pair of genotypes u, v at a site is $e_{u,v}(x)$, where x is the genotype of the query individual at that site ($u, v, x \in \{0, 1, 2\}$). Here, rather than considering the emission score, we introduce a penalty score, S . At site j , for a pair of nodes, i say, representing genotypes u and v , we have that $p_{j,i} := -\log_{10}(e_{u,v}(x))$. Then, if pair i contains nodes whose parental nodes formed pair k at $j - 1$, the penalty score for that pair i is $S_{j,i}$ such that:

$$S_{j,i} = \begin{cases} p_{j,i} & \text{if } j = 1 \\ S_{j-1,k} + p_{j,i} & \text{otherwise} \end{cases} \quad (2.5)$$

Progressing along the tree, nodes whose traceback up the tree has been incompatible with the query genotype accumulate higher scores. Note that for pairs of nodes where no incompatibilities have been encountered $S = 0$. Higher S values may reflect multiple single incompatibilities (where one of the nodes in a pair represents a genotype that is incompatible with the query genotype) and/or double incompatibilities (where both nodes in a pair are incompatible with the query genotype).

2.4 Application 2: Nearest Neighbours and haplotype sharing

An illustration of the how node pairs are used in order to select pseudo-parents is shown in Figure 2.3, which uses the data and T from the example introduced in Figure 2.2. At the first SNP there are two nodes, $T_{1,1}$ and $T_{1,2}$, each of size 2. Therefore, there are 3 possible node pairings: the two nodes with themselves ($(T_{1,1}, T_{1,1})$ and $(T_{1,2}, T_{1,2})$) and a pair consisting of both of them, $(T_{1,1}, T_{1,2})$. This is illustrated in the first column of Figure 2.3. Each pair is assigned its respective penalty score $S_{1,*}$. Proceeding to SNP 2, there are 3 nodes in the graph of Figure 2.2, $T_{2,1}$ - $T_{2,3}$. Therefore, there are 4 possible pairings, illustrated in the second column of Figure 2.3: $T_{2,1}$ can pair with itself since it has size > 1 and the other pairs are of all the nodes pairing with each other. The edges between columns 1 and 2 in Figure 2.3 join each pair of nodes at SNP 2 with the pair of nodes at SNP 1 that contains its parental nodes. For example, following the bottom part of Figure 2.2, the parent node of $T_{2,1}$ is $T_{1,1}$ and of $T_{2,3}$ is $T_{1,2}$. Therefore, in Figure 2.3, the pair $(T_{2,1}, T_{2,3})$ is joined by an edge to the pair $(T_{1,1}, T_{1,2})$. Iterating in this manner, we consider and score all possible pairs of nodes at each SNP. We may then identify the optimal pair of pseudo-parents for this genomic region by identifying the pair of leaf nodes (which are unique individuals) with minimum penalty score S .

In the process described here, when we consider all the pairs of nodes at a SNP j , we have already computed the penalty scores for their parental pairs of nodes. We can thus identify pairs of nodes which are unlikely pseudo-parent candidates, that is, the pairs of nodes whose traceback up the tree has been accumulating incompatibilities. As we are iterating along, we can therefore improve on efficiency even further by pruning those parts of the tree that have low likelihood, thus controlling the growth of the state space. In other words, when we consider candidate pairs of nodes at every position we can choose to discard some of them if until our current position they have not demonstrated that they are a good fit with the query genotype. This heavily reduces the number of pairs of nodes that need be considered at each SNP. For example, we may keep track of all of node pairs whose history has no more than X incompatibilities with the query genotype (X being typically 0 or 1). Once the minimum number of mismatches exceeds X for all remaining pairs, we reach a breakpoint and the iteration stops at that position rather than at the end of the tree. The pseudo-parents are identified

2.4 Application 2: Nearest Neighbours and haplotype sharing

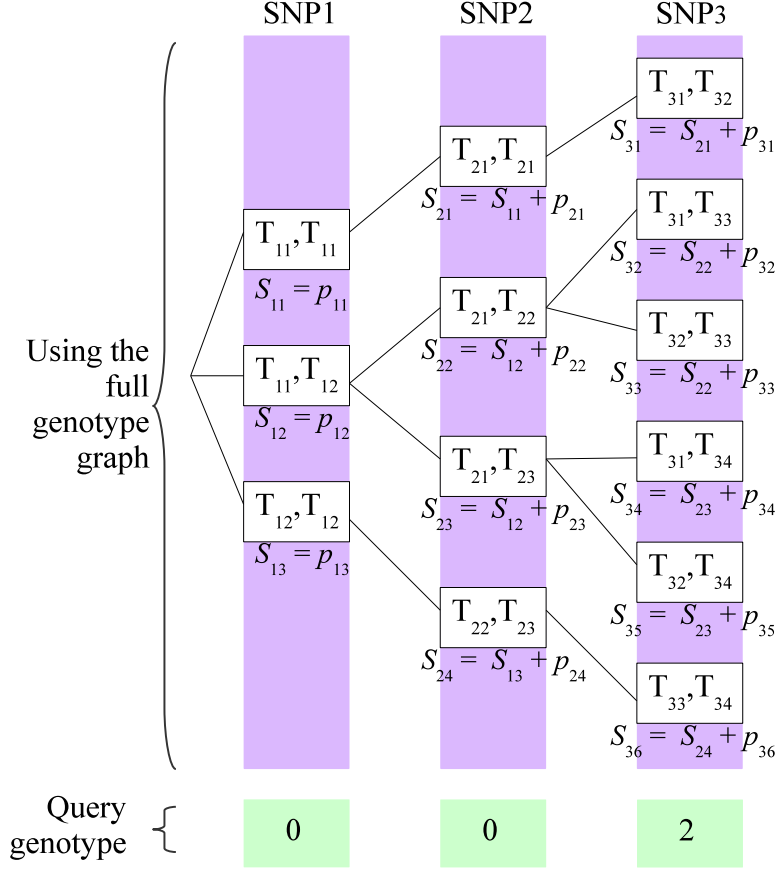


Figure 2.3 **Using the graph to select pseudo-parents.** The diagram demonstrates the process for selecting optimal pseudo-parents, continuing from the example presented in Figure 2.2. At each SNP, nodes here represent the possible pairs of nodes from the graph in Figure 2.2. Each node here, node i say, at each SNP, j , is assigned a penalty, $S_{j,i}$. $p_{j,i}$ is $-\log_{10}$ of the compatibility emission score (Table 2.2) of pair i , given the query genotype at j . Edges in this illustration represent the parent-daughter relationship for the members of the node pairs. Refer to the text for a full explanation of the example being presented here.

only until that position by sampling uniformly across all optimal pairs of nodes, and uniformly across all individuals carried by these nodes.

Continuing with the example of Figures 2.2 and 2.3, this pruning process is illustrated in Figure 2.4. Let pruning take place if a double incompatibility or two single incompatibilities occur. At SNP 1, the pair $(T_{1,2}, T_{1,2})$ represents the

2.4 Application 2: Nearest Neighbours and haplotype sharing

pair of genotypes (2,2) both of which are incompatible with the query genotype at that site (refer to Figure 2.2 for the toy data). In Figure 2.4 therefore this pair is not extended at the next site. Similarly, the pair $(T_{1,1}, T_{1,2})$ has a single incompatibility, and the pair $T_{2,1}, T_{2,3}$ has another and therefore the latter is also not extended at the next site. Overall, therefore, Figure 2.4 depicts the significant reduction in the state space compared to Figure 2.3, which leads to a great improvement in computational efficiency.

So far we have described the process for obtaining a pair of pseudo-parents for the region captured by a single genotype tree, until its breakpoint. As previously mentioned, however, the breakpoint (or the end of the tree) is unlikely to occur at the end of the sequence and we must therefore discuss the way for extending the method forwards, in order to obtain a full path of pseudo-parents. When we reach the end of the iteration, the process restarts by building a new genotype tree from the breakpoint (or an earlier position) onwards. The new tree can then be considered in the same way in order to identify the optimal pair of individuals for this new region. This new pair could be a different pair to the previous one, indicating a recombination event. To ensure that only single recombination events occur, the initialisation of the new tree at the new start site must differ in the following way:

- There must be 2 nodes, each for the optimal pair of individuals identified using the preceding tree,
- Then there may be as many more nodes as there are unique genotypes among the remaining individuals at the start site.

What is more, during the recursion only pairs of nodes including at least one of the previously-optimal individuals should be considered. This way, we guarantee that at least one of them ends up in the new optimal pair derived from the current genotype tree. This approach ensures only single parent transitions. While this is not necessary here for computational gains, it contributes to our ability to detect extended shared haplotype chunks: if we did not impose this condition then a double jump (a transition where both pseudo-parents change between adjacent SNPs) might occur by chance (there are multiple NNs and multiple

2.4 Application 2: Nearest Neighbours and haplotype sharing

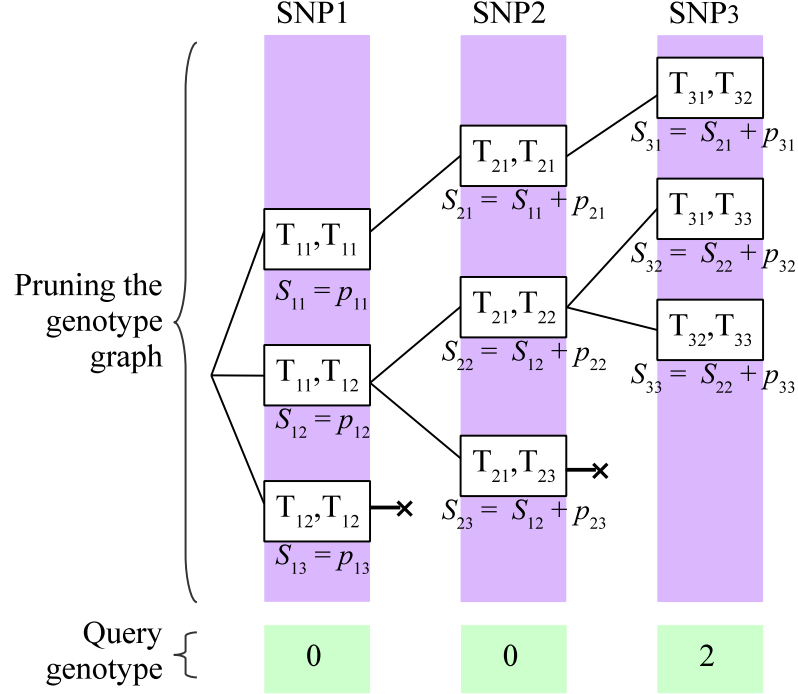


Figure 2.4 **Pruning the graph to select pseudo-parents.** A tree is shown here to demonstrate how the graph structure representing the toy data from the toy example in Figure 2.2 is used to select pseudo-parents. The diagram shown here is the pruned version of the illustration in Figure 2.3. Nodes here represent the pair of nodes from the graph in Figure 2.2 but pairs are included at each site only if they have not accumulated too many incompatibilities at previous sites. We indicate the nodes that would have been extended by a cross. Each node here, node i say, at each SNP, j , is assigned a score. $S_{j,i}$ = the score of pair i at site j ; $p_{j,i}$ is $-\log_{10}$ of the compatibility emission score (Table 2.2) of pair i , given the query genotype at j . Edges in this illustration represent this parent-daughter node relationship. Refer to the text for a full explanation.

optimal pairings and a different one is equally likely to be selected) and therefore a longer shared chunk will be missed. Moreover, if both parents really do need to change, just as in the HMM this may be reflected by another breakpoint occurring soon afterwards; that is, by two nearby single transitions.

We are now in a position to briefly outline the full algorithm for the graph method:

2.4 Application 2: Nearest Neighbours and haplotype sharing

Algorithm 2 The graph method

Initialisation:

- 1: Construct a suffix tree starting at SNP 1
- 2: Iterate over all pairs of nodes, scoring and pruning until a breakpoint, at SNP j say, is reached.
- 3: Identify the optimal pair of individuals, I_1 for the region.

Recursion:

- 1: Set $J = j$
- 2: Set $K = 1$
- 3: **while** $J \neq L$ **do**
 - 4: Construct a suffix tree, starting at J , initialising with separate nodes for the individuals in I_K .
 - 5: Iterate over all pairs of nodes, scoring and pruning until a breakpoint, at SNP j say, is reached.
 - 6: Set $K \rightarrow K + 1$
 - 7: Identify the optimal pair of individuals, I_{K+1} for the region.
 - 8: Set $J = j$
- 9: **end while**

Termination:

- 1: Obtain the set of pseudo-mothers and pseudo-fathers using $\{I_1, I_2, \dots, I_K\}$, such that only one parent changes at any one time.
-

2.4.2.3 Subsampling

In our graph implementation, for each genotype tree that is constructed pairs of nodes are pruned when they accumulate too many incompatibilities with respect to the target genotype. If it were possible to exclude these nodes in advance - by removing the appropriate individuals - this would heavily reduce run-time. This motivates a subsampling approach: if individuals are selected such that they are compatible with the query genotype over longer regions, then they are most

2.4 Application 2: Nearest Neighbours and haplotype sharing

likely the ones to end up in the optimal pairs, allowing for the remaining samples to be discarded safely. This would speed up the algorithm by several orders of magnitude since:

1. T would now represent fewer individuals and could thus be built more quickly (for a reminder on how the tree is constructed refer to Figure 2.2).
2. T would now carry fewer nodes and therefore significantly fewer pairs of nodes would need to be considered at each SNP.

To devise the best subsampling algorithm for the graph method that will have minimal impact on accuracy, we must first consider how the graph method works. For each genotype tree, we score pairs of nodes according to some emission matrix (the approach was demonstrated in Figures 2.3 and 2.4). For NN detection and for making inference on haplotype sharing we have discovered that the IBS-inspired compatibility emissions (Table 2.2) perform best (see Chapter 4). It follows, therefore, that any subsampling approach should be one where the discarded individuals are ones which would have been hugely disfavoured by the compatibility scoring system. It also follows that were we to use the graph method for a different application, such as haplotype estimation, where a different scoring system is more effective, the current subsampling scheme would not be suitable.

Adopting therefore a compatibility-based subsampling approach, in order to prevent loss of accuracy with the graph method, we would need to ensure that:

1. The subsampled individuals are such that they will maximise the distance until the next breakpoint. Recall that a breakpoint occurs at the SNP along the genotype tree when all pairs of nodes have accumulated enough incompatibilities such that they are deemed unsuitable by that point. Moreover, recall that one of our aims is to detect maximal haplotype sharing as a proxy for local relatedness and maximising the distance until the next point is key for achieving this.
2. All the alleles present in the query genotype until the breakpoint (*which is unknown*) are represented in the subsampled set of individuals

2.4 Application 2: Nearest Neighbours and haplotype sharing

Suppose we have a sample of n individuals, $\{G^1, \dots, G^n\}$. At SNP j , let the genotype of G^i be denoted by G_j^i and let the genotype of the query individual, Q , be denoted by Q_j . The alleles of Q_j are the two haplotypes of Q at j , and are denoted by Q_j^1 and Q_j^2 . For example, if $Q_j = 1$, then without loss of generality $Q_j^1 = 0$ and $Q_j^2 = 1$. Then, following Table 2.2, Q_j and G_j^i are compatible if $|Q_j - G_j^i| < 2$. Define the indicator function:

$$\delta_{i,Q}^j = \begin{cases} 1 & \text{if } G_j^i \text{ and } Q_j \text{ are compatible} \\ 0 & \text{otherwise} \end{cases} \quad (2.6)$$

Finally, let $\Delta_j = \{G_j^1, \dots, G_j^n\}$, the set of genotypes carried by the sample at SNP j . We say the alleles of Q_j are accounted for by Δ_j if both alleles of Q_j exist in Δ_j (i.e. both Q_j^1 and Q_j^2 can be found in Δ_j). Note that we exclude singletons from our analyses in this thesis so that every allele carried by a query genotype is always carried by at least another individual in the sample. For some subset of individuals however, it is not obvious that this will always be true. For example, if $Q_j = 1$, then if all the genotypes in Δ_j are 0, they are all compatible with Q_j but the mutation carried by the query individual is not accounted for in that set. Define a second indicator function:

$$f_{Q_j}(\Delta_j) = \begin{cases} 1 & \text{if } Q_j \text{ is accounted for by } \Delta_j \\ 0 & \text{otherwise} \end{cases} \quad (2.7)$$

Suppose we would like to subsample K individuals (or $x\%$ of n), then subsampling occurs before initialisation and before Step 4 in the recursion step of Algorithm 2. Before each new genotype tree is constructed, subsampling occurs such that the new tree is built to represent just these selected individuals, with Algorithm 2 then continuing as before. Before a new tree at SNP $l \in [1, L]$, the outline for our algorithm for subsampling a set of individuals, I_l is:

1. Order individuals according to their distance to the next incompatibility with the target
2. Keep only the top K individuals or the top $x\%$.
3. Ensure that all alleles in the target are accounted for until the site of incompatibility with the last individual in the subsampled set. For every allele

2.4 Application 2: Nearest Neighbours and haplotype sharing

that might be absent in the subsample, pick the best (defined by distance to an incompatibility) individual that carries it from the remaining individuals and add it to the set.

4. Return the final subsample, I_l .

More formally, the pseudo-code for the algorithm is:

Algorithm 3 Subsampling for the graph method

```

for  $i \in [1, n]$  do
  Set  $D_i = \max_j \{j : \delta_{i,Q}^l = 0, j \in [1, l]\}$ 
end for
1. Set  $\hat{D} = \{D_{(n)}, D_{(n-1)}, \dots, D_{(1)}\}$  such that  $\hat{D}_k = D_{(n-k)}$ , the  $k$ th largest distance to an incompatibility
2. Order the individuals according to  $\hat{D}$ 
3. Subsample: Keep the first  $K$  individuals and call this set  $G$ 
4. Set  $Z = \{1, \dots, n\} / G$ 
5. Define the current chunk  $\in [l, l + \hat{D}_K]$ 
6. Ensure that all alleles in the current chunk are accounted for:
for  $j \in [l, l + \hat{D}_K]$  do
   $\Delta_j = \{G_j^1, \dots, G_j^K\}$ 
  if  $f_{Q_j}(\Delta_j) = 0$  then
    1. Order  $Z$  in decreasing order according to  $\hat{D}$ 
    2. Set  $U \subset Z$  such that  $\forall u \in U$ , the missing allele from  $Q_j \in u$ .
    3. Set  $G = \{G, U_1\}$ 
    4. Set  $Z = \{1, \dots, n\} / G$ 
    5. Set  $K \rightarrow K + 1$ 
  end if
end for
Set  $I_l = G$ 

```

The current chunk that is constructed defines the region over which the current tree can be used for selecting pseudo-parents from the subsampled set. Therefore,

2.4 Application 2: Nearest Neighbours and haplotype sharing

the current tree is considered either until the end of this chunk, or potentially earlier, if all the pairs of nodes on the tree have accumulated too many incompatibilities.

As already mentioned, we subsample an initial set of individuals at the first SNP, and then repeat after every breakpoint. Overall, therefore, there may be at least K subsampled individuals for the iteration until the first breakpoint, and at least $K+2$ thereafter. K is a lower bound since it might be necessary to add more individuals in order to ensure that all relevant alleles of Q are represented in each chunk. An illustration of the full graph method with subsampling is shown in Figure 2.5, where K is assumed to be 3.

Assuming that enough of the local variation is captured, removing incompatible genotypes will have little effect on the output of the graph approach as it does not interfere with the parent-selection process. However, one can imagine the following scenario as an example of how accuracy could be affected. Suppose an individual is incompatible with the query genotype close to the starting point but nowhere else in the previous chunk and nowhere else until the theoretically most distant breakpoint. This person will have a very low D distance value, despite being compatible with the query genotype overall in that local environment. In other words, individuals with incompatibilities with the query genotype early on in the sequence covered by the current T but not elsewhere could be excluded initially. Had they not however, it is possible that they would ultimately have been selected as a NN.

This is an inherent limitation of this graph scheme: it is directed in that it ranks individuals according to their distance to an incompatibility to the right of the starting position. The effect of this however could be small for two reasons. Firstly, if an individual is wrongly excluded early on in the sequence, they may typically be reinserted after the next breakpoint. Secondly, the graph method is itself directed as well, with results being different when moving left to right than right to left and sensitive to the start site from which each tree is built. Therefore our subsampling method is in sync with the way the algorithm operates and directedness is rather an inherent limitation of the graph method: we may expect there to be some loss of accuracy because of it, but little or *no further loss* due to the subsampling.

2.4 Application 2: Nearest Neighbours and haplotype sharing

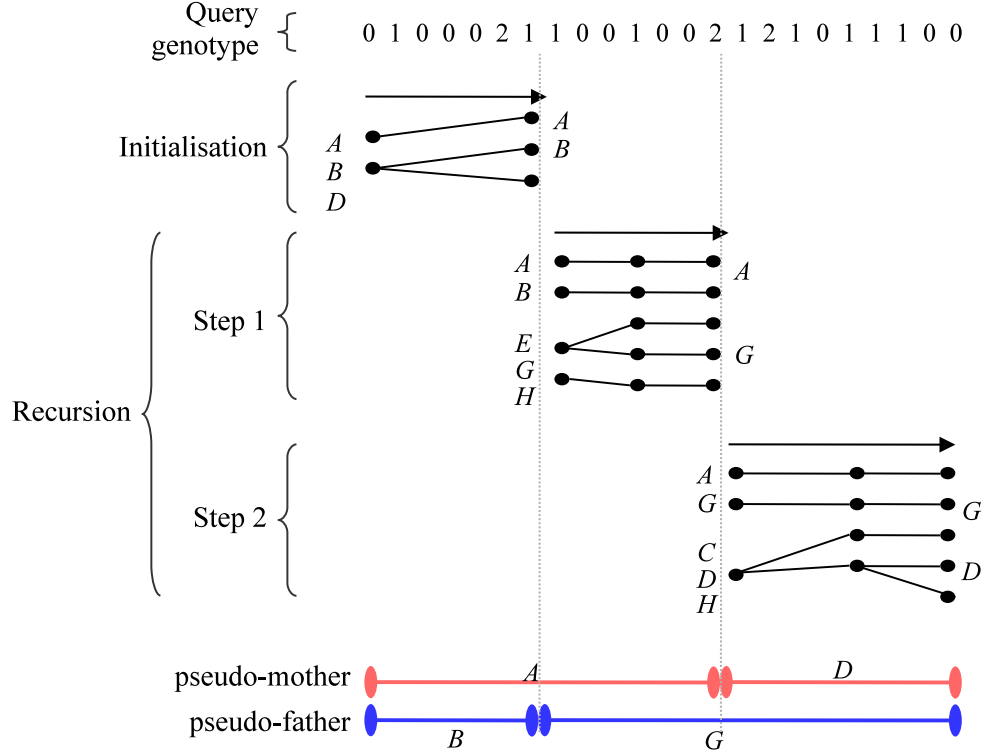


Figure 2.5 **Outline of the graph method.** The figure illustrates a toy version of the algorithm for the graph method. For a query genotype and individuals $A - H$, the algorithm here is split into three parts: initialisation, recursion step 1 and recursion step 2. The three parts are a result of breakpoints occurring at the end of each preceding part, indicated by the vertical dotted grey lines. Each part involves subsampling K (here $K = 3$) individuals and then building a genotype tree for these individuals, as outlined in Figure 2.2. The data used to construct the trees illustrated here are not shown. Each tree is then iterated over in order to select two pseudo-parents and the choice of pseudo-parents in each case is indicated by labelling the relevant leaves at the end of the trees. In this example, individuals (A, B) were picked at the end of Initialisation, (A, G) at the end of Recursions: Step 1 and (D, G) at the end of Recursion: Step 2. We can therefore unambiguously obtain a sequence of pseudo-mothers and pseudo-fathers, as illustrated at the bottom of the figure by the red and blue segments. In this example, there were two recombination events and four shared haplotype chunks. Note that the trees are initialised differently during the Initialisation and Recursion stages such that just a single parent can change between breakpoints - refer to text for details.

2.4 Application 2: Nearest Neighbours and haplotype sharing

The combined application of our graph method and subsampling strategy enables efficient analysis of large data sets, which would have been impossible using our original HMM. For example, to infer pseudo-trios for an individual in a sample of over 5,000 (see Chapter 5) across data on 27,000 SNPs takes c. 200 seconds on a 32Gb, 1.6 Ghz machine. Moreover, the graph implementation, like the HMM, is fully parallelisable over individuals.

Finally, it should be noted that in its current implementation our graph method ignores LD information. Recombination affects the initialisation of trees in the iteration. For example, in cold spots we could choose to initialise the next tree after a breakpoint such that no change in pseudo-parents is allowed. Moreover, recombination is the driving force that influences changes in haplotype sharing between individuals, and is a main contributor to the variation that defines local relatedness between individuals. Furthermore, the number of incompatibilities, X , that must be accumulated in order to prune nodes and in order to reach a breakpoint is determined heuristically and is fixed across the genome. These features suggest that the graph method is a rather crude approximation of the HMM and hint at the graph method being more susceptible to errors and less accurate at finding recent coancestry and haplotype sharing than the original HMM.

2.4.3 An HMM with Subsampling

In light of the graph method’s weaknesses, namely that it is directed and does not model LD and genotype error rate, in more recent work we developed an alternative approach that maintains the advantageous properties of the original HMM but also incorporates some of the computational tricks that were effective in previous efforts. Rather than using our subsampling approach for the graph method, we considered how we might apply it within the framework of an HMM. We call this new method HMMsub.

The subsampling technique ranks individuals according to their distance to an incompatibility with the query genotype to the right of a given position. In order to eliminate the directedness of this approach we modified our query and for each

2.4 Application 2: Nearest Neighbours and haplotype sharing

individual we now consider the sum of its distances to an incompatibility to the left and to the right of the given position. In other words, we rank individuals according to their total distance to an incompatibility in both directions around the position rather than away from it in one direction. In order to apply the Viterbi algorithm over the whole sequence, the subsampling must have occurred in advance. We therefore split the region into $M + 1$ parts by selecting M evenly spaced SNPs along the sequence, and subsampling individuals around them. The top K individuals can then be kept around each of these points and the genomic positions over which they are optimal (i.e. the chunk for these individuals) are the ones over which these K individuals are ‘all more or less in IBS’. The algorithm should ensure that there is at least one chunk covering each part of the sequence and we therefore check that chunks overlap. If this does not occur, we subsample further locally, creating a new chunk for each such region. To ensure that single transitions from one subsampled set to the next is possible, around each breakpoint we consider the K individuals that we selected, as well as the ones from the chunk to the left and to the right of the current chunk.

Let j represent SNP j from a total of L SNPs, and i individual i from a sample of size n . The process, is summarised in Figure 2.6. In the figure, the query genotype is labelled by Q and there are 8 individuals in the sample, $A - H$. The figure illustrates the 3 main stages of the subsampling process:

1. Subsampling around an initial set of SNPs and defining the chunks around them. In the example in the figure this is shown by the part labelled ‘Split and subsample’. Three initial SNPs were selected (marked by the black arrows), resulting in the construction of 3 (some overlapping in this case) chunks, C_1 - C_3 , each of which has a set of $X \approx \frac{K}{3}$ individuals. In this example we have assumed $X = 2$. C_3 contains 3 individuals (F, G, H) as more subsampling must have been needed in order for the alleles of Q in the region of C_3 to be accounted for. It is not required that the sets of individuals are distinct nor that there are no common individuals between them.
2. Identifying regions of the sequence not covered by these chunks and subsampling at these regions. This is illustrated by the part labelled ‘Adding

2.4 Application 2: Nearest Neighbours and haplotype sharing

more chunks'. The region of the query genotype in bold was not spanned by an existing chunk and so a new one (coloured in red) was obtained by subsampling further at the end of chunk C_1 . This now gives a total of 4 chunks, C_1 - C_4 , with some chunk labels changing to accommodating the new chunk that has been defined.

3. Merging sets of individuals from adjacent chunks, thus obtaining the 'Final State Space' at each position. Our strategy does not ensure that there will be overlap between the individuals subsampled in each chunk. We thus merge adjacent chunks in order to ensure that single transitions are possible during the Viterbi iteration.

Having subsampled, only individuals marked by crosses (bottom of Figure 2.6) at any given SNP may be considered. The HMM is then used just as in the original, full case on this constrained hidden state space. Following the notation and principles in algorithm 3, the pseudo-code for the subsampling follows below:

Algorithm 4 Subsampling for an HMM

Partition the region

Select M equally spaced SNPs along the sequence.

Obtaining initial chunks and subsampling

for $j \in [1, M]$ **do**

for $i \in [1, n]$ **do**

$$D_i^l = \max_j \{j : \delta_{i,Q}^l = 0, j \in [1, j]\}$$

$$D_i^r = \min_j \{j : \delta_{i,Q}^l = 0, j \in [j, L]\}$$

$$D_i = D_i^l + D_i^r$$

end for

Set $\hat{D} = \{D_{(n)}, D_{(n-1)}, \dots, D_{(1)}\}$ such that $\hat{D}_k = D_{(n-k)}$, the k th largest distance to an incompatibility

Order the individuals according to $\hat{D}$

Subsample: Keep the first X individuals and call this set G , such that:

if $j \in 1, L$ **then**

2.4 Application 2: Nearest Neighbours and haplotype sharing

```

     $X = \lfloor 2\frac{K}{3} \rfloor$ 
else
     $X = \lfloor \frac{K}{3} \rfloor$ 
end if
Set  $Z = \{1, \dots, n\} / G$ 
Create a chunk  $C_j$  in the range  $[C_j^l, C_j^r]$  where:
 $C_j^l = \min \{D_{(n)}^l, \dots, D_{(n-X+1)}^l\}$ 
 $C_j^r = \min \{D_{(n)}^r, \dots, D_{(n-X+1)}^r\}$ 
Ensure that all alleles in the current chunk are accounted for:
for  $j \in [C_j^l, C_j^r]$  do
     $\Delta_j = \{G_j^1, \dots, G_j^K\}$ 
    if  $f_{Q_j}(\Delta_j) = 0$  then
        Order  $Z$  in decreasing order according to  $\hat{D}$ 
        Set  $U \subset Z$  such that  $\forall u \in U$ , the missing allele from  $Q_j \in u$ .
        Set  $G = \{G, U_1\}$ 
        Set  $Z = \{1, \dots, n\} / G$ 
        Set  $K \rightarrow K + 1$ 
    end if
end for
Set  $I_{C_j} = G$ , the collection of all subsampled individuals for SNP  $j$  and
chunk  $C_j$ .
end for
Ensure that every point in the sequence is covered by at least one chunk
Set  $P = M$ , the number of chunks
for  $C \in \{1, \dots, P\}$  do
     $j = -1$ 
    if  $C = 1$  then
        if  $C_1^l \neq 0$  then
             $j = 1$ 
        end if
    else if  $C = M$  then
        if  $C_M^r \neq L$  then

```

2.4 Application 2: Nearest Neighbours and haplotype sharing

```
     $j = L$ 
  end if
else
  if  $C^r < (C + 1)^l$  then
     $j = C^r$ 
  end if
end if
if  $j > 0$  then
  if  $j \in [1, L]$  then
     $X = \lfloor 2\frac{K}{3} \rfloor$ 
  else
     $X = \lfloor \frac{K}{3} \rfloor$ 
  end if
  Subsample  $X$  individuals around  $j$  and create a new chunk
  Set  $P \rightarrow P + 1$ 
end if
end for
Merge adjacent subsampled sets to define the set of individuals for each chunk:
for  $C \in \{1, \dots, P\}$  do
  if  $C = 0$  then
    Set  $K_0 = I_0 \cup I_1$ 
  else if  $C = P$  then
    Set  $K_P = I_{P-1} \cup I_P$ 
  else
    Set  $K_C = I_{(C-1)} \cup I_C \cup I_{(C+1)}$ 
  end if
end for
```

In practice we select M such that the chosen SNPs are approximately 200kb apart (though this might change for example with sequencing data), and K , as with the graph method can safely be set to about 20 individuals. Note that due to pooling of individuals from adjacent chunks and extra individuals being selected to account for all alleles present in the child in any given chunk the total

2.4 Application 2: Nearest Neighbours and haplotype sharing

number of individuals that will be considered at any one part of the sequence is unknown. Assuming this number is X_j , which is variable for $j \in \{1, \dots, L\}$, then the total number of hidden states to iterate over is $\binom{X_j}{2}$ which is still manageable, independently of the original total sample size. For comparison purposes, to run over a whole chromosome for 5,000 individuals with the graph method requires about 3 minutes while HMMsub takes between 15 minutes to 45, depending on the variation in the size of X_j . HMMsub is also parallelisable over individuals.

2.4 Application 2: Nearest Neighbours and haplotype sharing

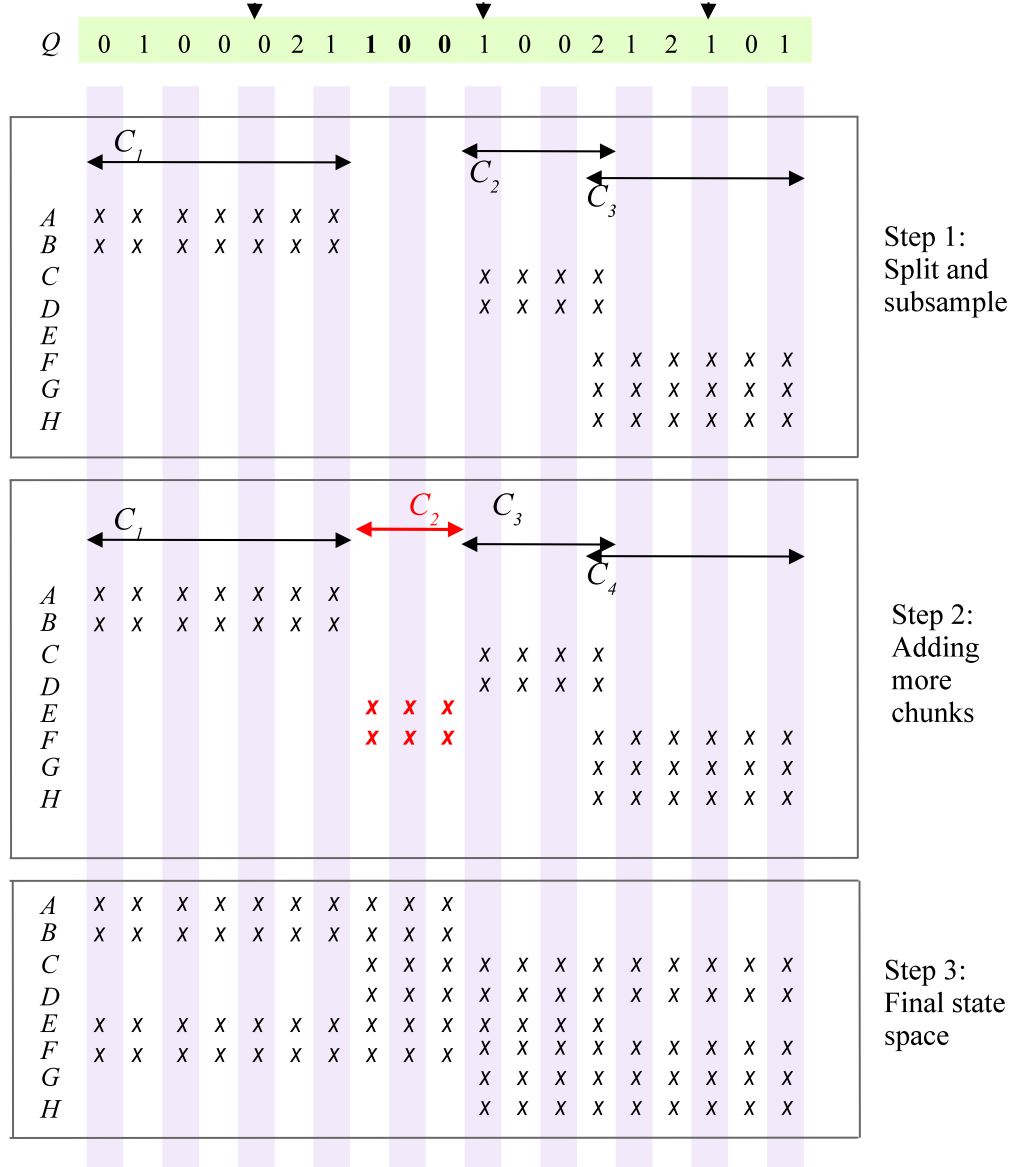


Figure 2.6 **Subsampling for an HMM**. From top to bottom the figure demonstrates the subsampling process. Q is the query genotype and there are 8 individuals in the sample, $A - H$. The crosses show which individuals are being considered at each site. The region is initially split into 4 parts, around 3 SNPs marked by the black arrows. Subsampling occurs around these SNPs, obtaining 3 sets of individuals (marked by crosses) for 3 chunks $C_1 - C_3$ (marked by the black segments; Step 1). The grey dashed lines show the chunk boundaries. A part of the sequence (marked in bold on Q) remains uncovered by $C_1 - C_3$ and so more local subsampling occurs creating an extra chunk (red; Step 2). The final set of individuals is obtained by merging of adjacent sets (Step 3; refer to text).

2.5 Model Summary

To construct local pseudo-trios, we use a Hidden Markov Model (HMM; Rabiner [1989] and Section 2.2), where the emissions are the target’s genotypes and the hidden states are all the $\binom{n}{2}$ pairs of ‘parents’ at each position. The Viterbi algorithm (Foreman [1992], see Section 2.2.3), which identifies the optimal path of hidden states, is used to obtain the optimal parent pairs at each site along the sequence. Transitions between the hidden states (see Section 2.2.1) model recombination events, where the parent being copied for one haplotype changes, and emissions at each site are considered as scores of the “goodness-of-fit” of the candidate parent pair’s genotypes to that of the ‘child’s’.

For identifying NNs, we assess whether the parent pairs in a single maximum likelihood path (sampled uniformly from all optimal paths) contains the true NNs at every position. To further explore the full posterior distribution of candidate parent-pairs, we also consider the Forward/Backward algorithms (Rabiner [1989]; Section 2.4.1), asking whether the pairs of parents with maximum posterior at each site contained the true NNs. The Viterbi and F/B algorithms by construction are not expected to yield identical results and it is therefore of interest to investigate whether one is more effective at identifying NNs than the other.

To measure shared haplotype lengths we record the genetic and physical chunks over which our optimal choice of pseudo-father and pseudo-mother does not change: a change in parent signals a break to the haplotype that is being inherited by transmission from that individual. Therefore, this parent change marks the end of the haplotype that is being shared between the pseudo-parent that changed and the query genotype. Further details on our assessment methods can be found in Chapter 4, while an illustration of this core part of the model is shown in Figure 2.7. These pseudo-parents and the query genotype also form pseudo-trios along the sequence, which can then be used to phase by transmission.

Unlike Kong et al., our approach takes recombination into account and does not phase pseudo-parents while phasing the query genotypes. We also do not construct pseudo-families. For phasing we must therefore develop a strategy for making inference at triple heterozygous sites (see the positions marked with ‘?’ in Figure 2.7) or sites where there are incompatibilities between the parents and the

‘child’. Our solution for phasing every position along the genome by transmission is to additionally sample suboptimal pseudo-parent paths. In this way, we may have multiple pseudo-mother and pseudo-fathers like in Kong et al.’s LRP. In the absence of singletons in the sample, for every SNP there should exist a candidate individual in the sample that can be used to phase by transmission.

For each query genotype, suppose we have n individuals remaining in the sample. Suppose further that the number of SNPs is L . Below we provide an informal version of our full algorithm.

Algorithm 5 Our Model

1. Run the Viterbi algorithm on pairs of individuals, filling in the Viterbi matrix. The transition step allows only single jumps to occur such that a single parent may change between two adjacent SNPs.
2. Identify the best K paths of pseudo-parents by determining the traceback of the largest K entries in the last column of the Viterbi matrix. For obtaining a pair of parents at each site K is 1, while for phasing K would typically be 10.

In the context of haplotype estimation the algorithm continues:

3. Phase the query genotype with each path separately in the same way that phasing is achieved with true trio data (see Section 2.3.2).
 4. Combine the separate path predictions into a final pair of haplotypes.
-

In the first step, only one parent is allowed to change between any two adjacent SNPs, decreasing the computational burden of considering all possible parent pairs to $O(N)$ from $O(N^2)$. We found that these double jumps occurred extremely rarely, and could often be replaced by single transitions at adjacent sites (data not shown). The restriction of the transitions therefore does not influence the

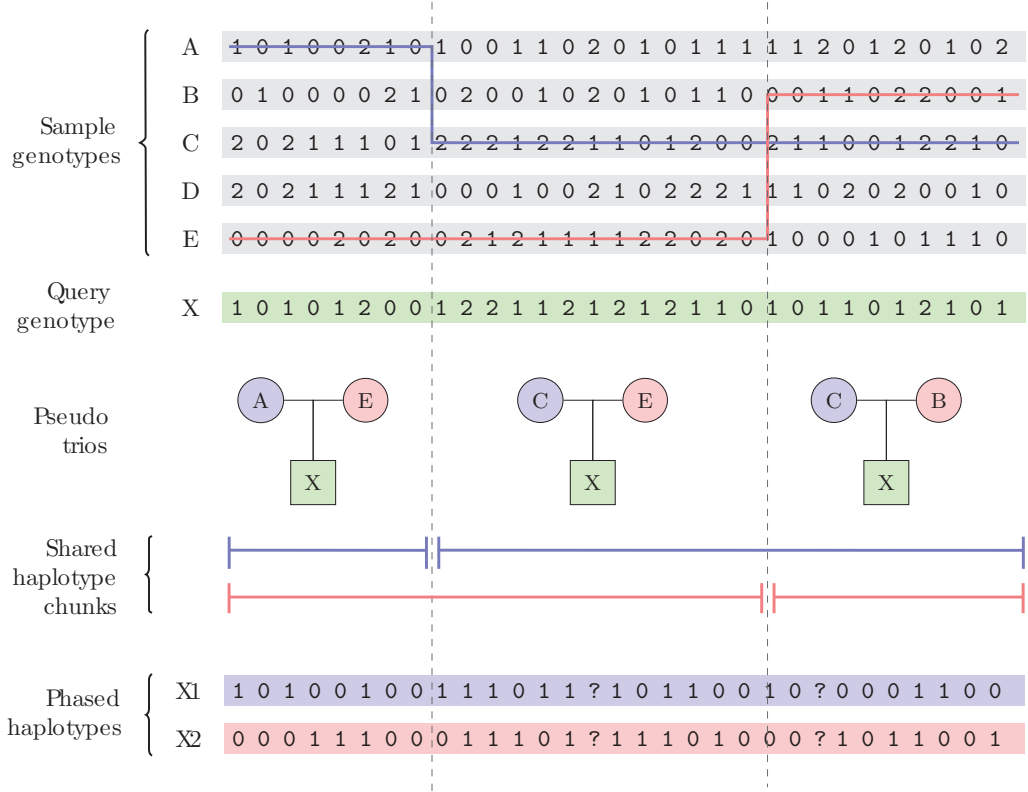


Figure 2.7 **The HMM.** Running the Viterbi algorithm on the population sample of genotypes A-E, we determine at each site which are the two most likely parents for the query genotype, X. By construction of the HMM, emissions occur independently site-wise which is an approximation, albeit reasonable, to the true parents-to-offspring transmission process. Changes in these most likely parents are interpreted as recombination events, of which there are two in this example (marked by the vertical dotted grey lines). The chunks over which the pseudo-parents are constant define our shared haplotype chunks (4 chunks here). Further, we ask whether our choice of pseudo-parents contain the true nearest neighbours for the haplotypes contained in X. We then construct the pseudo-trios for each region, and phase as with true trio-data. Some of the sites are unphasable (marked with “?”) because the most likely parents are both heterozygous or because they are incompatible with the genotype of the child (not shown). To discourage this and find informative pairs, we find the most likely parents using weighted Mendelian emission probabilities, which weight against the triple heterozygote case.

pseudo-parent selection process. To our knowledge, limiting the transitions in this way has not been used in other work.

Another advantage of single-parent transitions is that it enables groups of pseudo-mothers and pseudo-fathers to be aligned relative to each other by considering which parent they have in common. This avoids unnecessary phasing errors from occurring at the transition sites. We explain this in full detail in Section 2.3.2. While this is less relevant for the case of NN detection and haplotype sharing, we maintain the single-transition setup for computational purposes and to ensure that shared haplotype chunks are not unnecessarily disrupted.

Our algorithm and its two computationally efficient extensions run independently for each genotype. The choice of pseudo-parents along the sequence for one individual is deterministic (given the genotypes and assuming uniqueness in likelihood values at $V_{.,L}$) and is independent of the pseudo-parents selected for other individuals. We remind the reader that when there is non-uniqueness of values, either in the Viterbi matrix or during the traceback process, pairs individuals are selected at random from all pairs with equal likelihood. The process of phasing by transmission is also deterministic and is carried out in one iteration. Therefore our method does not update its prediction, as is the case for example for MCMC based approaches (e.g. Howie et al. [2009]). It thus gains computational efficiency by allowing its implementation over the entire sample to occur in parallel. The total run time is therefore comparable to that required for a single genotype (assuming availability of a large number of processors and linear memory scaling). Running time is $O(N^2L)$, where L is the number of SNPs, for the entire sample, and $O(NL)$ for a single individual. Memory usage for the algorithm with storage of the entire Traceback matrix would generally be in the order of N^2L , though it is possible to store it sparsely, reducing this to N^2 (or even further with thinning), allowing for larger datasets.

The consequence of allowing pseudo-parents to be detected independently for each individual is that surrogate parenthood is no longer a non-directional relationship like in LRP. This is because the candidacy of an individual for being a pseudo-parent is judged relative to the other individuals in the sample and not just on its absolute IBS status with the query genotype.

Chapter 3

Statistical haplotype estimation

3.1 Overview

The original motivation behind the project described in this thesis was to develop an algorithm for haplotype estimation from unphased SNP genotype data. The aim was to generalise the LRP method of Kong et al. [2008] so that smaller sample sizes of unrelated individuals could be considered, with a method that takes into account recombination. Our model, an HMM where the hidden states are all the possible pairs of individuals in the sample (they are all candidate pseudo-parent pairings), constructs pseudo-trios at each SNP using the Viterbi algorithm to obtain paths of hidden states along the query individual’s genomic sequence. A detailed overview of our model for phasing can be found in Chapter 2 but we revisit some of the main concepts here for the reader’s convenience.

With SNP genotype data on a mother-father-offspring trio, it is possible to assign haplotype phase (of transmitted alleles) by segregation, unless all three individuals are heterozygous, or if a de-novo mutation has occurred. Within a pseudo-trio, we can take an identical approach, with the caveat that where one of the pseudo-parents changes (by recombination, which is modelled as a hidden state transition in our HMM), the ‘transmitted’ haplotypes must be linked to-

gether, which is achieved by restricting only a single parent to change between any two markers. The importance of this restriction is demonstrated in Section 2.3.2. Where all three individuals are heterozygous, or where there is a Mendelian inconsistency (due to mutation or genotyping error), the target genotype cannot be phased (see Figure 2.7 and Section 2.3.2). For this reason, it is advantageous to combine information from multiple pseudo-parent paths. We therefore obtain multiple sets of pseudo-parent paths by sampling a single parent-pair path uniformly from the set of maximum likelihood paths, then sampling another path from all paths with the second highest likelihood and so on (Section 2.3.1).

The paths are orientated at the first heterozygous site in the target. At the next heterozygous site the paths ‘vote’ on the phase of this site relative to the first. Phase is assigned by a majority vote and the confidence is determined by the fraction of paths in agreement. Paths are re-orientated by the current SNP and the process restarts (Section 2.3.3). Our implementation of our model for phasing is referred to as PACE (available upon request).

In this chapter we describe the analysis that was carried out to investigate the effectiveness of our method. It consists of two parts:

1. A preliminary investigation using smaller simulated data sets, whose purpose was to distinguish between different versions of our model in order to identify the best for phasing.
2. A larger scale analysis comparing the final version of our phaser with other software, both on a variety of simulated data sets and real genotype data.

We first discuss our method for assessing phasing performance.

3.2 Measuring accuracy

In order to be able to carry out comparisons between different versions of our model and between our model and other methods, we must first define what we mean by phasing accuracy. When phasing, we predict the two haplotypes, $(\hat{H}^1, \hat{H}^2)$ say, of the query individual at the sites where they are heterozygous.

At these (biallelic) positions the two haplotypes must be opposites of each other, i.e. $\hat{H}_j^1 = 1 - \hat{H}_j^2$ (where $\hat{H}_j^1 \in \{0, 1\}$) is the allele carried by $\hat{H}^1$ at a site j) since their sum must be 1. In order to determine accuracy, $\mathcal{A}$, without loss of generality it is therefore sufficient to compare (in some way) the first haplotype of our prediction ($\hat{H}^1$) with the first (H^1) of the true pair of haplotypes (H^1, H^2).

To best illustrate how we could measure accuracy, it is useful to consider the toy example in Figure 3.1 which compares H^1 with $\hat{H}^1$. In this example, the two haplotypes are the same for the first part of the sequence (except for site 3) and then they are opposites for the remaining sequence. Therefore, there are 2 regions where the haplotypes differ, indicated by the light purple shading. A naive approach, which counts the proportion of sites where $\hat{H}^1 = H^1$, is clearly incorrect as it significantly underestimates phasing performance. Instead, phasers have traditionally been measuring *switch errors* (described in Stephens and Donnelly [2003a]). Switch errors count the ‘cuts and flips’ needed to align a predicted haplotype with the correct orientations: they occur when the orientation of our predictions is one way around until a particular site but then is reversed. A ‘cut’ in the diagram is illustrated each time there is a change in the background colour.

As we can see in this example however, two consecutive switches (for instance before and after site 3) are in fact a single, local error. We would like to distinguish such cases (when the orientation of all sites but one is correct) because the traditional error-counting approach in essence double-counts them. The effect of this is illustrated in Table 3.1. In this work therefore, rather than counting the switch error rate, we split this into two error types: the flip error rate ($\mathcal{F}$), which counts the local, single-position errors, and the phase error rate ($\mathcal{P}$) which counts errors where the orientation of the haplotypes is reversed at more than one positions. For our final measure for accuracy, we have $\mathcal{A} = 1 - \mathcal{E}$, where $\mathcal{E} = \mathcal{F} + \mathcal{P}$. In the initial part of our analysis (Section 3.3) we also look at each error type ($\mathcal{F}$ and $\mathcal{P}$) on its own to examine whether it is worth reporting them separately and indeed distinguish between them when we compare different phasers.

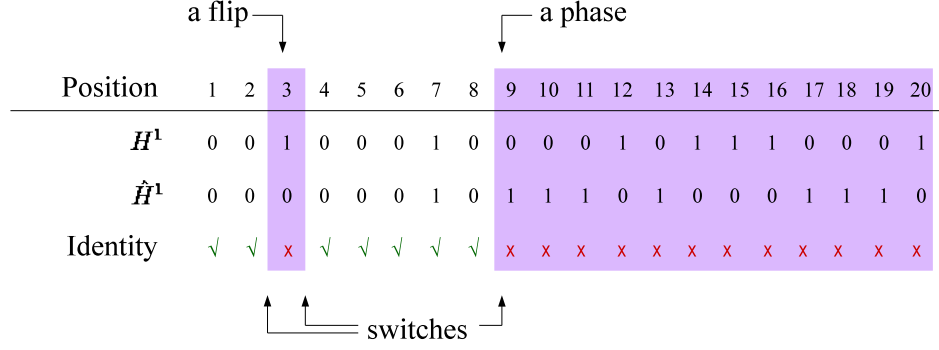


Figure 3.1 **Measuring phasing accuracy.** The figure shows the first predicted haplotype, $\hat{H}^1$, of a query individual, and its first true haplotype, H^1 , at 20 positions where the individual is heterozygous. The green ticks illustrate the positions at which the two haplotypes are the same, while the red crosses mark the positions where they are different. The background shading indicates the changes between regions where there is agreement and disagreement between the haplotypes. Traditional measures of accuracy would count an error (a switch error) every time the shading changes. In this example there would therefore be 3 errors: before position 3, after position 3 and before position 9. In this thesis, we distinguish the different types of switch errors: local errors at one site which we call flips, and errors where orientation changes for more than one site, which we call phase errors.

	No. Errors	Error Rate
$\mathcal{N}$	7	$\frac{7}{20} = 0.35$
$\mathcal{S}$	3	$\frac{3}{20} = 0.15$
$\mathcal{E}$	2	$\frac{2}{20} = 0.1$

Table 3.1 **Defining phasing accuracy.** In this table we compare 3 different methods for measure phasing accuracy: a naive approach ($\mathcal{N}$) which counts the proportion of sites where two haplotypes are the same (or $1 - \mathcal{N}$ if $\mathcal{N} < 0.5$); the traditional switch error rate, $\mathcal{S}$; our measure, $\mathcal{E}$, which looks at the sum of flips and phase errors. The values in the table reflect the data from the toy example on Figure 3.1.

3.3 Building a phaser

With a measure for accuracy, we are now in a position to compare methods and begin our inference. As described in Chapter 2, while the overall architecture of the HMM for phasing was well described from the outset of this project there were several key elements that needed to be determined:

1. How do pseudo-parents compare to real parents when it comes to phasing by transmission?
2. Which emission matrix is best suited for haplotype estimation?
3. How many sites are left unphased when a single optimal path is used?
4. How effective are suboptimal paths at phasing?
5. How many paths could we consider such that the proportion of unphased sites and accuracy are jointly minimised and maximised respectively?
6. What is the best method for combining predictions across multiple pseudo-parent paths?
7. Can we always phase every single heterozygous site?

In what follows we explore each of these questions in turn. First we provide an overview of the data that were simulated for carrying out these investigations.

3.3.1 Simulated Data

In order to carry out rapid testing of our method, we performed a series of basic analyses on a series of data sets. We considered relatively small datasets that are sufficient for distinguishing the strengths and weaknesses of the different parameterisations and versions of our model. With this in mind, we simulated 100 different data sets, each consisting of 100 haplotypes (50 individuals) over 500 SNPs ($\sim 100\text{kb}$; using a recombination rate of 1.5cM/Mb). Given the small size of the data, it was unnecessary to use published software for simulating data from the coalescent. Using such software, like Hudson’s *ms* (Hudson [2002]) and MaCS

(Chen et al. [2009]) is necessary for larger samples (and/or more complicated demographic histories) where data simulation is a much more time-consuming process and sophisticated approximation techniques need to be employed.

For our purposes here, it was sufficient to use R code from McVean (www.stats.ox.ac.uk/~mcvean/R/coalescent.r), which implements the standard coalescent model with recombination and mutation (Hein et al. [2004]) directly. The model here assumes a constant population size, where coalescent and recombination events compete for the next event (going backwards in time) which takes place after an exponentially distributed waiting time. Mutations are then added onto the simulated Ancestral Recombination Graph (ARG) using a Poisson Process.

This process generates sets of haplotypes. By pairing up each pair of haplotypes we obtain genotype data. We can therefore phase each individual in turn. In fact, individuals can be phased in parallel as our model runs independently for every person. By comparing our predicted haplotypes with the originally simulated data we can assess each of the phasing methods that will be considered.

3.3.2 Real Parents

In order to assess the performance of pseudo-parents for phasing, we must first benchmark optimal phasing performance by investigating how true parents phase by Mendelian transmission. Let individual k in a data set be denoted by G_k , consisting of the pair of haplotypes $(H_{G_k}^1, H_{G_k}^2)$. Using the simulated data described in Section 3.3.1, for each data set (of 50 individuals) we constructed real trios, consisting of a mother, father and child, $\{M, F, C\}$, as follows:

Algorithm 6 Generating real trios

```

for  $i \in [1, 25]$  do
  Set  $M_i = G_{2i-1}$ 
  Set  $F_i = G_{2i}$ 
  Set  $C_i = H_{M_i}^1 + H_{F_i}^1$ 
  Phase  $C_i$  by transmission, using  $M_i$  and  $F_i$ .
  Record the proportion of unphased sites.

```

end for

In the absence of mutation, as is the case here, there can obviously be no phasing errors when phasing by transmission. With real data any phasing errors when using trios and Mendelian transmission will have occurred due to genotyping error, or de novo mutation whose incidence rate depends on the local, genomic mutation rate. Instead, it was of interest to obtain a benchmark for the proportion of unphased sites, $\mathcal{U}$, (in other words, the triple heterozygote incidence rate). Carrying out this experiment we discover that $\mathcal{U}$ stands at 17.7% - this close to $\frac{1}{6}$ which is what is expected under the standard neutral model.

This is a very high fail rate for a phaser, suggesting that it was justified to consider an emission scoring system that will weigh against triple heterozygotes. Unless a collection of suboptimal paths proves to be particularly efficient at rectifying this issue, these results indicate that the pseudo-parents that our model selects should in fact *not* behave like real parents.

This observation begs the question whether our model would choose the real parents as pseudo-parents, were the former to exist in our sample. To explore this, for each of the simulated children, and for the Mendelian transmission (MT) probabilities and compatibility scores (CE) separately we performed the following:

Algorithm 7 Searching for the real parents.

1. Run the Viterbi algorithm
 2. Select all optimal paths
 3. For each path, record the proportion of sites at which a real parent is a pseudo-parent
 4. For each path, also recorded $\mathcal{S}$ and $\mathcal{U}$
-

We can see from Table 3.2 that neither emission scoring system favours the real parents - though the percentage is significantly higher for CE (88% vs 11%). When using CE, however, there are many more optimal paths, increasing the chances that the real parents are included in the full optimal set.

	%Parents	$\mathcal{S}$	$\mathcal{U}$
MT	0.11	0.02	0.164
CE	0.88	0.03	0.25

Table 3.2 **Finding real parents.** The table shows whether the real parents were included in all optimal pseudo-parent pairs averaged across sites, the switch error rate, $\mathcal{S}$ and the proportion of unphased sites $\mathcal{U}$ averaged over all optimal paths, using Mendelian Transmission (MT) and the Compatibility emissions (CE). The results were obtained by running our phaser on simulated children when their true parents were included in the data set as candidate pseudo-parents.

We also observe that when switching to CE from MT, the proportion of unphased sites $\mathcal{U}$ increase by 1.5 times, and even with Mendelian transmission (MT) it stands at a high $\sim 16\%$. This confirms that both these emission scoring schemes are not catered to dealing with triple heterozygotes, further motivating a weighted approach. Finally, from this initial analysis we can see that MT has a lower error rate than CE. This, coupled with CE’s significantly higher rate of leaving sites unphased, is our first piece of evidence against CE’s suitability as a scoring system for the application of phasing (while conversely suggesting that it may be better for finding NNs). We explore the suitability of our three emission scoring systems more formally in the next subsection.

3.3.3 Picking the emission scores

The choice of emission matrix plays a pivotal role in the construction of the Viterbi matrix and therefore in the selection of surrogate parents for the ‘child’. In this section we compare the Mendelian Transmission (MT), our Weighted Mendelian transmission (WMT) approach and the Compatibility scores (CE) in order to conclude which among them is best suited for our phaser. For WMT, we have used $\zeta = 0.01$ (refer to Table 2.1, Section 2.2.2 for more details). In the remainder of our analysis we no longer use the children that we generated in the previous section by haplotype transmission. Instead, we take each simulated genotype

from our data sets in turn, as described in Section 3.3.1 and use the remaining simulated genotypes from the same data set as candidate pseudo-parents.

In this section, in our analyses we randomly sample a single optimal path from all optimal paths obtained using the Viterbi algorithm. This was done for computational efficiency, as the data that has been simulated for this part of the investigation resulted in multiple paths attaining optimal likelihood, with all choices of emission matrices. We selected a single path uniformly from all optimal paths in the last column of the Viterbi matrix and during the traceback process (see description of Equation 2.2); we have observed that this does not affect performance (data not shown).

The analysis described here is the only one carried out where double recombination events are allowed (double changes where both pseudo-parents are allowed to change between two adjacent sites; note, these are generally disallowed to improve phasing accuracy and increase computational efficiency). This was to confirm that they are extremely rare events and that therefore disallowing them will have little influence on our performance.

The result here (shown in Table 3.3) allow us to make the following observations:

- The Mendelian transmission probabilities (MT) reduce the error rate the most, giving an accuracy, $\mathcal{A}$, of 99.3%.
- The difference in accuracy between MT and the weighted Mendelian model (WMT) is $< 0.5\%$, which is not a significant difference.
- WMT reduces the frequency of unphasable sites ($\mathcal{U}$) to about a fourth of the frequency for the simple Mendelian case, which stands at a high 12%.
- The Compatibility emissions (CE) leave a significantly higher proportion of sites unphased ($\mathcal{U}$), namely $\sim 23\%$.
- Both WMT and MT outcompete the Compatibility model (CE) which has a switch error rate, $\mathcal{S}$, almost two to three times greater than for the other models, primarily due to the considerably larger frequency of flip errors, $\mathcal{F}$.

	MT	WMT	CE
$\mathcal{A}$	99.3	99	97.7
$\mathcal{F}$	0.2	0.3	1
$\mathcal{P}$	0.5	0.7	1.3
$\mathcal{S}$	0.8	1.1	2.3
$\mathcal{U}$	12.4	2.9	23.3
DB	0	0	0
H	87	12.6	97.7

Table 3.3 **Comparing the emission scores.** The table shows phasing results (% values) for the three emission scoring schemes when applying our model to simulated data. $\mathcal{A} = 1 - \mathcal{E}$; $\mathcal{E} = \mathcal{F} + \mathcal{P}$; $\mathcal{F}$ = flip error rate; $\mathcal{P}$ = phase error rate; $\mathcal{S}$ = switch error rate; DB = the proportion of double recombination events that have occurred over all heterozygous sites; H = proportion of unphased sites due to triple heterozygotes (the alternative is double homozygote incompatibilities). MT = Mendelian Transmission; WMT = Weighted Mendelian Transmission, with $\zeta = 0.01$; CE = Compatibility emission scores.

- One can speculate that the high proportion of unphased sites is the catalyst for the flip errors which compromise the accuracy of the CE-based model. Unphased sites do not contribute to the assessment of the error rate. Recalling, however, the ‘flipping’ of the paths whenever they disagree with the prediction at a given position we can see that when there are multiple unphased sites in a row this flipping process might not take place sufficiently well to ensure that all paths are properly aligned.
- Triple heterozygotes were the principle driver for unphased sites in the MT and CE cases.
- With WMT, which specifically weighs against triple heterozygotes, most unphased sites were due to picking parents that were incompatible with the child at these sites; double minor homozygotes (where both parents were homozygous for the minor allele) accounted for 84% of unphased sites, and

major homozygotes (where both parents were homozygous for the major allele) for 3% of cases.

- In these data, there were no double recombination events.

The above results confirm that MT is the most suitable choice, with respect to phasing accuracy. However, WMT is a very close second. What is more, the proportion of extra sites that are phased with a single optimal using WMT compensates for the slightly decrease in accuracy. Both methods outcompete CE, which as suspected in Chapter 2, picks paths which are largely uninformative for haplotype estimation. Considering that our aim is to phase unambiguously all the heterozygous sites and that the compromise in $\mathcal{U}$ is negligibly small, we continued our investigation using WMT (Table 2.1; $\zeta=0.01$).

3.3.4 The effect of suboptimality

In order to decide how many suboptimal paths we could incorporate in our phaser, which is necessary for minimising the proportion of sites left unphased, we must first understand how $\mathcal{A}$ relates to path likelihood. As explained in Chapter 2, by path likelihood we refer to the query genotype's likelihood, given a certain path of hidden states (pseudo-parent pairs). By path order we refer to path rank according to likelihood: first order paths are those achieving maximal likelihood, second order paths are those with second highest likelihood, and so on.

In this part of our investigation, we continue using the same data sets described earlier in this chapter. Here, a series of optimal and suboptimal paths were considered in order to get a rich set of pseudo-parent sequences of a variety of likelihoods. For each individual, we proceeded as follows:

Algorithm 8 Investigating path likelihood

1. Collect all paths arising when tracing back from each of the possible parent pairs at the last SNP (i.e. trace back from each hidden state at the last SNP). When there were more than one optimal states in the traceback (while running the Viterbi algorithm), we selected states uniformly.

2. For each of the paths obtained in Step 1, deviate away from them once, as described in Section 2.3.1.
3. Use each path to obtain a separate pair of haplotype estimates for the query genotype.
4. For each path record its likelihood, $\mathcal{A}$ and $\mathcal{U}$

Using WMT, the mean accuracy, $\mathcal{A}$, over all paths was 0.978, and the proportion of unphasable sites stood at 0.048. In Figure 3.2 we stratify these observations according to path likelihood.

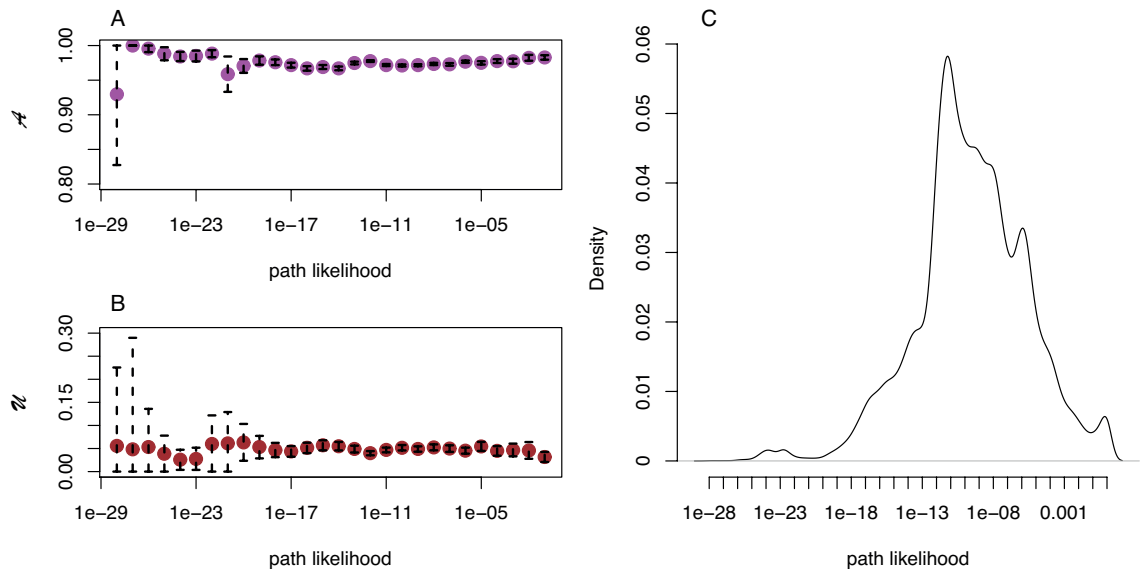


Figure 3.2 The effect of path Likelihood. A: the accuracy, $\mathcal{A}$ against path likelihood. B: the proportion of unphased sites, $\mathcal{U}$ against path likelihood. The paths used were either the paths when starting from each of the hidden states at the last site, or are the best paths deviating away from them. The bars represent 95% confidence intervals. C: density curve of path likelihoods.

Figure 3.2 illustrates a striking robustness of the HMM to a decreasing likelihood, with both accuracy, $\mathcal{A}$ (Figure 3.2A) and the proportion of unphased sites, $\mathcal{U}$ (Figure 3.2B), remaining constant over all orders of log likelihood. Characteristically, the figure shows that up to a relative likelihood of approximately 10^{-25}

there is no apparent trend in accuracy, which remains high at $\approx 98\%$, close to the overall mean across paths. The large drop in efficiency, in terms of phasing capacity, that can be seen for the lowest likelihood as well as the respective rapid increase in the error rate can be attributed to the fact that (as reflected in Figure 3.2C and the wider confidence intervals in A and B) there were very few paths obtained in that region (likelihood $< 10^{-18}$) and so the low performance might have occurred by chance. Note that this analysis does not represent the space of all possible paths (i.e. all possible combinations of parent-pairs along sites) but rather the maximum likelihood paths (and best paths deviating from them) starting from each parent-pair at the last SNP.

Moreover, if we collect paths sequentially, going down the likelihood orders until we have collected at least one path that can phase each heterozygous site we make the following observations:

- In these simulations, the first 25 paths will typically phase $\approx 98.7\%$ of heterozygous sites
- Although this statement implies that we need quite a large number of paths to account for all the heterozygous sites of an individual, in $> 80\%$ of cases, $> 98\%$ of sites were phased using the top 6 paths
- 72% of individuals were phased entirely using just a single optimal path.

It is more frequently the case, therefore, that most sites are phasable with only a small number of paths, and that on some occasions quite a large number is needed (i.e. some sites for some individuals are atypically difficult to phase). We additionally discover that the path deviating away from the optimal path is actually second best (has the second highest likelihood) only 1.2% of the time. In fact, the mean rank of paths deviating away from optimal paths is 35, showing that there is little support for this strategy. In order to enrich our collection of paths therefore, in the remaining analyses we only sample from the last column of the Viterbi matrix. If K paths are to be collected, these will be obtained by tracing back from the K hidden states associated with the highest K likelihoods at the last SNP. Whenever there are multiple paths with the same likelihood,

we sample one uniformly and at random. In all, these findings support employing as many paths as is necessary, constrained only by computational efficiency considerations.

3.3.5 Managing suboptimality

In this section we attempt to answer two questions:

1. How many paths should we be collecting?
2. What is the best strategy from those introduced in Section 2.3.3 for combining their predictions into one final pair of haplotypes? The schemes that we consider are the Left to Right approach, for which there were three options: SM, WM and BP, and the Top to Bottom Scaffolding Method (TTB).

To address these questions, for each scheme that combines path predictions we consider $\mathcal{A}$, $\mathcal{F}$, $\mathcal{P}$, $\mathcal{S}$, $\mathcal{U}$, the probability (Pr) of the phase calls (defined for each scheme in Section 2.3.3) and the real accuracy $\hat{\mathcal{A}}$, where $\hat{\mathcal{A}} = \mathcal{A} \cdot (1 - \mathcal{U})$. In all computations we use the WMT emissions (Table 2.1; $\zeta = 0.01$) and consider results when collecting 2, 5, 7 and 10 paths in total. We tabulate the results for each method and also produce calibration plots using the phase call probabilities. Calibration plots show the proportion of calls that were made correctly against the probability of the call. Calls were assumed to be made incorrectly if they were at a site at which a flip error occurred or where a phase error (change in orientation) started. A well-calibrated model will be one where the the gradient between calls and probabilities of a call is 1: this suggests that the probability of a call is perfectly informative about the accuracy that we should expect for that call. Finally, we also count the proportion of calls that were obtained within each bin of prediction probabilities.

As stated in Chapter 2, in the results below we discover that our Simple Means (SM) strategy from the Left to Right scheme performs best. The results for SM are shown first, followed by the results from the remaining methods in a subsequent subsection.

	2	5	7	10
$\mathcal{A}$	98.1	98.3	98.3	98.3
$\mathcal{F}$	0.7	0.7	0.6	0.6
$\mathcal{P}$	1.2	1	1.1	1.1
$\mathcal{S}$	2.3	2	2	2
$\mathcal{U}$	2.2	1.6	1.4	1.3
Pr	97.5	97.8	97.9	97.9
$\hat{\mathcal{A}}$	95.9	96.5	96.9	97

Table 3.4 **Simple Means scheme for combining path predictions.** The columns are results (%) when considering the 2, 5, 7 and 10 paths associated with the respectively highest likelihoods at the last SNP. $\mathcal{A} = 1 - \mathcal{E}$; $\mathcal{E} = \mathcal{F} + \mathcal{P}$; $\mathcal{F}$ = flip error rate; $\mathcal{P}$ = phase error rate; $\mathcal{S}$ = switch error rate; $\mathcal{U}$ = the proportion of unphased sites; Pr = probability of the phase calls; $\hat{\mathcal{A}} = \mathcal{A} \cdot (1 - \mathcal{U})$, the real accuracy.

The results in Table 3.4 show that as we increase the number of paths, raw accuracy, $\mathcal{A}$ remains constant while the proportion of unphased sites decreases significantly, particularly when increasing to 5 from 2 paths. This confirms that adding extra paths to our model will not affect performance, but will rather improve it by ensuring that a large proportion of heterozygous sites are phased. Figure 3.3 shows that the SM scheme is well calibrated, with the results lying on the $y = x$ line, in all of the four cases tested.

3.3.5.1 Other schemes for combining information across paths

Consider first the Top to Bottom Scaffolding method (TTB; Table 3.5). Even when 10 suboptimal paths are used, $\mathcal{A}$ does not go below 97.5%, with a switch rate, $\mathcal{S}$, maintained around 3%, offering us thus complete freedom on the number of paths that we can use. We also observe a reduction in proportion of unphasable sites ($\mathcal{U}$) as order increases, with the proportion decreasing by 50% for 10 paths, down to 1.3%. We also record ‘the real accuracy’ ($\hat{\mathcal{A}}$), which does increase, suggesting that there is an overall improvement as we include more paths and

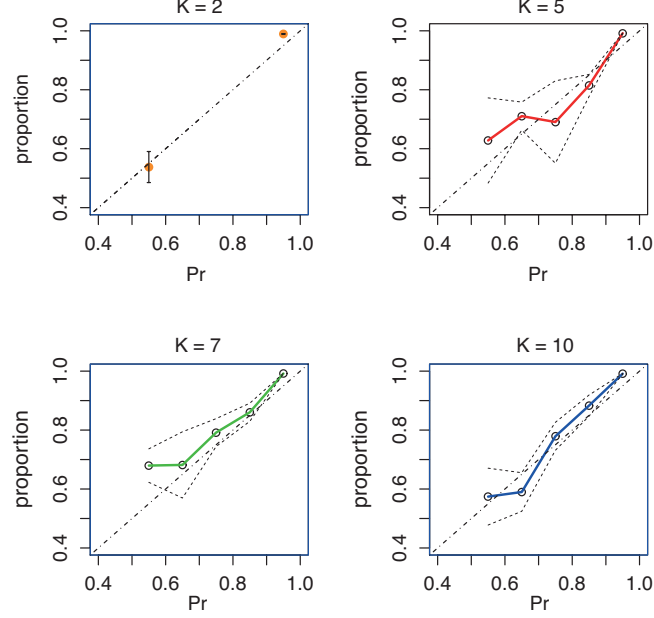


Figure 3.3 **Calibration plots for SM.** K = the number of paths that have been collected; Pr = the probability of a phase call. The y-axis shows the proportion of calls that were correct. The grey dotted lines represent 95% confidence intervals, except for the top left plot where error bars are depicted instead. In that particular case, observations were made only with probabilities between 0.5 and 0.6 and above 0.9 (i.e. either the paths agreed or they didn't).

phase more sites. The increase, however, is very small, $< 0.5\%$. Calibration is not optimal for any of the orders considered (Figure 3.4), with the plots being far from the $y = x$ line. We moreover observe that the distribution of counts among the intervals is highly skewed, as shown in Figure 3.5, with the overwhelming majority of predictions having probability $> 90\%$.

We consider now the other versions of the Left to Right approach WM (Table 3.6) and BP (Table 3.7) We also compare our observations here with those seen for the SM scheme, presented in the previous subsection.

For the BP scheme we must note that all predictions were made with a prob-

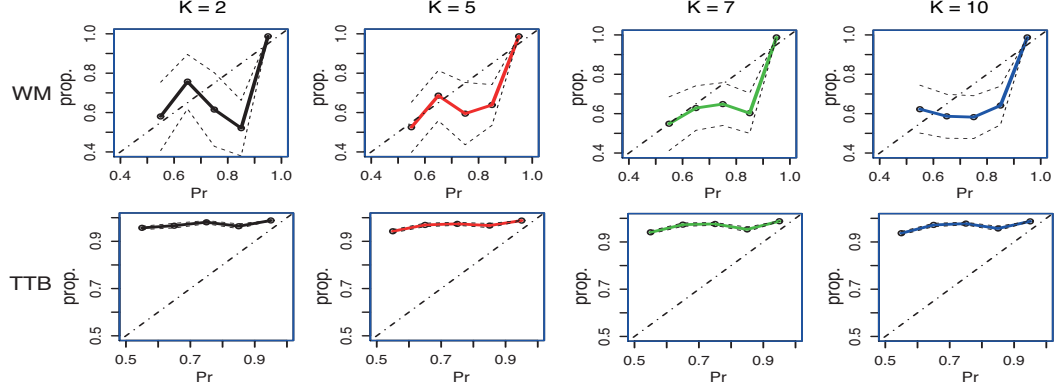


Figure 3.4 **Calibration plots for all other path-combining algorithms.** K = the number of paths that have been collected (black = 2; red = 5; green = 7; blue = 10); Pr = the probability of a phase call. The y-axis show the proportion of calls that were correct and the grey dotted lines represent 95% confidence intervals. The diagonal, dotted black line is the $y = x$ line. WM = left to right approach using weighted means; TTB = the Top to Bottom Scaffolding scheme.

ability of $> 90\%$, and consequently no calibration plots were produced. In particular, for 2 orders, the proportion of correct predictions was 0.97, for 5 orders it was 0.96, for 7 orders it was 0.97 and finally for 10 orders it was 0.96.

Evidently, the WM mechanism performs efficiently as well, with $\mathcal{A}$ not breaking down at all as order increases. However its performance does not match the SM scheme which achieved an accuracy of $> 98\%$ and a real accuracy reaching 97%, increasing the most ($> 1\%$) as order increases. The WM method also does not exhibit smooth calibration nor are the confidence intervals on its calibration plots as tight as for the SM method, which as order increases, moves close to the $y = x$ line. The BP method performs worst, with $\mathcal{A}$ at approximately 95% due to the high phase error rate; we can therefore easily eliminate this approach.

Overall, we observe that the error rate is maintained, as expected, as we

3.3 Building a phaser

	2	5	7	10
$\mathcal{A}$	98	97.6	97.5	97.5
$\mathcal{F}$	0.8	1.0	1.0	1.0
$\mathcal{P}$	1.2	1.4	1.4	1.5
$\mathcal{S}$	2.5	2.9	3.0	3.1
$\mathcal{U}$	2.2	1.6	1.4	1.3
Pr	92.6	89.7	89.3	89.3
$\hat{\mathcal{A}}$	95.8	96.0	96.1	96.2

Table 3.5 **The Top to Bottom Scaffolding Strategy** We consider the best 2,5,7, and 10 paths (tabulated in each column respectively) to obtain measures for assessing this approach. $\mathcal{A} = 1 - \mathcal{E}$; $\mathcal{E} = \mathcal{F} + \mathcal{P}$; $\mathcal{F}$ = flip error rate; $\mathcal{P}$ = phase error rate; $\mathcal{S}$ = switch error rate; $\mathcal{U}$ = the proportion of unphased sites; Pr = probability of the phase calls; $\hat{\mathcal{A}} = \mathcal{A} \cdot (1 - \mathcal{U})$, the real accuracy.

	2	5	7	10
$\mathcal{A}$	98.2	98.1	98.2	98.1
$\mathcal{F}$	0.6	0.7	0.7	0.7
$\mathcal{P}$	1.1	1.1	1.2	1.2
$\mathcal{S}$	2.1	2.2	2.2	2.2
$\mathcal{U}$	2.2	1.6	1.4	1.3
Pr	97.9	98.3	98.5	98.5
$\hat{\mathcal{A}}$	96	96.5	96.8	96.8

Table 3.6 **Weighted Mean.** Each column is for 2, 5, 7 and 10 paths respectively. $\mathcal{A} = 1 - \mathcal{E}$; $\mathcal{E} = \mathcal{F} + \mathcal{P}$; $\mathcal{F}$ = flip error rate; $\mathcal{P}$ = phase error rate; $\mathcal{S}$ = switch error rate; $\mathcal{U}$ = the proportion of unphased sites; Pr = probability of the phase calls; $\hat{\mathcal{A}} = \mathcal{A} \cdot (1 - \mathcal{U})$, the real accuracy.

consider more paths, while the proportion of unphased sites decreases, though at a slower rate progressively. The main differences between the methods are with

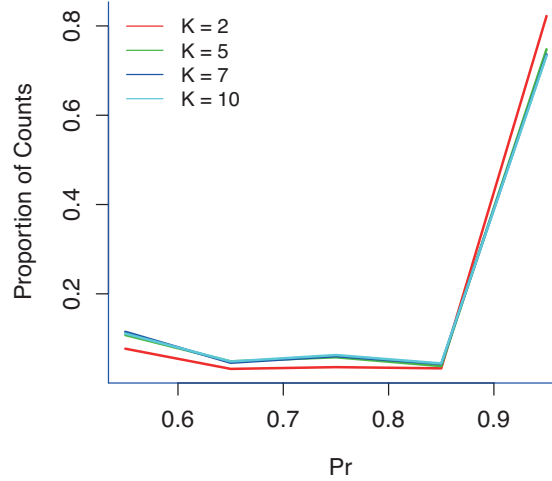


Figure 3.5 **Calibration counts for the Top to Bottom scaffolding approach.** The figure shows the proportion of counts per probability interval out of all calls made when combining information across path using the TTB scheme. Pr = probability of a phase call; K = the number of suboptimal paths used.

respect to distinguishing between flips and phase errors and for calibration:

- the TTB approach leads to a slightly higher flip error rate ($\mathcal{F}$)
- Calibration is generally poor, due to the highly skewed distribution of counts within probability intervals, with most predictions having a probability $\geq 90\%$. The exception here is the SM, where the model is, in fact, quite well calibrated.

We conclude that all our proposed methods, except the BP approach, are efficient at combining information over the selected paths. Considering just two paths does decrease the proportion of unphasable sites, yet the main reduction occurs when we use the best 5 paths. This reduction, however, is for all methods $< 0.5\%$ between the second and tenth likelihood orders.

	2	5	7	10
$\mathcal{A}$	89.4	96.7	97.3	97.3
$\mathcal{F}$	3.7	0.8	0.6	0.6
$\mathcal{P}$	6.9	2.53	2.2	2.2
$\mathcal{S}$	13.5	3.53	2.65	2.6
$\mathcal{U}$	2.2	1.6	1.4	1.3
Pr	30.6	21.9	16.9	12.1
$\hat{\mathcal{A}}$	87.4	95.2	95.9	96

Table 3.7 **Best Path**. Each column is for 2, 5, 7 and 10 paths respectively. $\mathcal{A} = 1 - \mathcal{E}$; $\mathcal{E} = \mathcal{F} + \mathcal{P}$; $\mathcal{F}$ = flip error rate; $\mathcal{P}$ = phase error rate; $\mathcal{S}$ = switch error rate; $\mathcal{U}$ = the proportion of unphased sites; Pr = probability of the phase calls; $\hat{\mathcal{A}} = \mathcal{A} \cdot (1 - \mathcal{U})$, the real accuracy

For all methods, the real accuracy increases with the number of paths, although the rate of increase is relatively poor, with the TTB approach exhibiting the lowest and (ignoring Best Paths) with SM exhibiting the greatest. We must further add, that these real accuracies also exceed the real accuracy obtained just from the optimal paths which is maximal for the WMT and stands at 96.1% (calculated using the values from Table 3.3). This accentuates the benefit of including suboptimal paths in our phaser, aside of the obvious advantage of phasing more heterozygous sites.

Finally, we note that the high performance of the SM method is a reasonable conclusion: since path accuracy does not deteriorate with decreasing likelihood (Figure 3.2), there is no obvious justification for introducing a scoring/weighting system that uses likelihood to combine predictions across paths. It should therefore be no surprise that the scheme that would work best is the one that assigns equal weight to path ‘votes’.

3.3.6 Phasing all sites

Finally, the phase of the few sites that remain ambiguous must be accounted for. To address this issue, for each such SNP one may simply identify the best path that is informative at that SNP. Given the removal of all the singletons from our data set, unless it is the case that all individuals are heterozygous at a particular SNP which should never happen under the assumption of Hardy-Weinberg equilibrium, there will always be a surrogate pair capable of phasing every site of the query genotype. To identify the appropriate path, the traceback from that site backwards can be obtained trivially, as has been performed so far. To determine the rest of the path forwards, from that SNP until the end of the sequence, the Viterbi algorithm must be run on the genotype of the child ordered backwards and with the data rearranged such that SNP order is reversed. We consider the SM, WM, LTR and TTB at order 2, and also at order 5 for the SM scheme, in order to test the performance of the proposed strategy.

	TTB	WM	SM 2	SM 5
$\mathcal{A}$	97	97.6	97.6	97.8
$\mathcal{F}$	1.3	1	1	0.9
$\mathcal{P}$	1.7	1.4	1.4	1.3
$\mathcal{S}$	3.5	2.8	2.8	2.3
$\mathcal{U}$	0	0	0	0
Paths	3.337	3.337	3.337	6.020

Table 3.8 **Phasing all sites.** We used the top 2 orders (and 5 orders for the SM method; SM5) searching for extra paths to phase all sites. TTB = the Top to Bottom Scaffolding method; WM = weighted means approach; SM2, SM5 = simple means approach with 2 and 5 paths respectively; Paths = the total number of paths required for all sites to be phased; $\mathcal{A} = 1 - \mathcal{E}$; $\mathcal{E} = \mathcal{F} + \mathcal{P}$; $\mathcal{F}$ = flip error rate; $\mathcal{P}$ = phase error rate; $\mathcal{S}$ = switch error rate; $\mathcal{U}$ = the proportion of unphased sites.

As can be seen from Table 3.8, this approach solves the problem of unphasable sites with a compromise in the error rate at $< 0.5\%$. The overall accuracy, $\mathcal{A}$,

which now is the real accuracy as well, for all methods is $\sim 97.5\%$, not far from the $\sim 98\%$ accuracy of the optimal path., with a mean switch error rate at around 3%. Turning to the calibration data in Figure 3.6, we observe that with the exception of the SM method, calibration remains poor, with the confidence intervals becoming wider as probability decreases especially for the WM scheme. We thus conclude that the probabilities of our predictions can generally not be used as indicators of performance, unless the SM approach is used, preferably with order 5 or above.

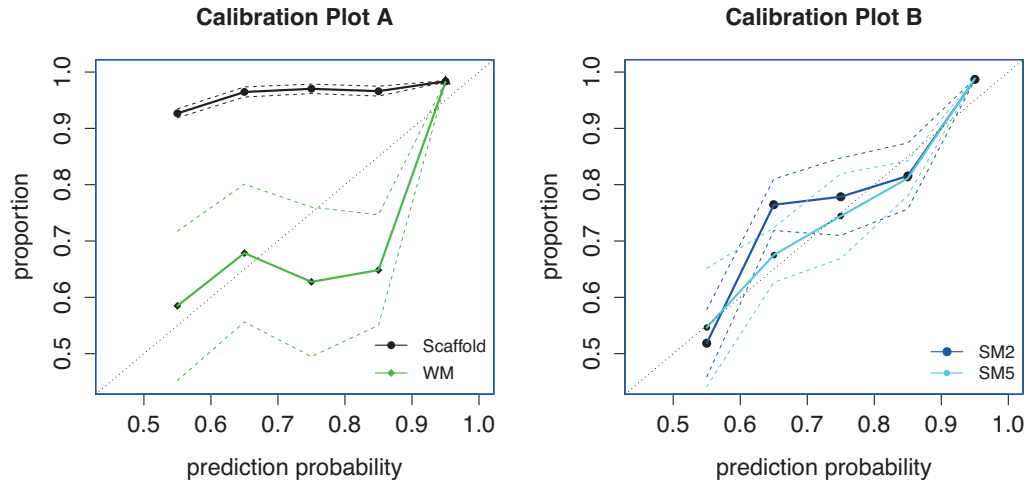


Figure 3.6 **Calibration Plots.** Plot A depicts the results for the Top to Bottom Scaffolding (TTB) and Weighted Means scheme, at order 2, while plot B illustrates the results for the Simple Means (SM) method at orders 2 and 5. The simulations carried out ensured that every site is phased. The black dotted diagonal lines are the $y = x$ lines. The coloured dotted lines are 95% confidence intervals for the results of each method.

The reader is reminded that in practice this approach for phasing all sites is intractable computationally as it requires full storage of two Viterbi and Trace-back matrices. This investigation rather serves as a proof of concept for how the method could in principle accommodate all heterozygous sites.

3.3.7 PACE

Having established the preferred emission matrix for phasing, namely the Weighted Mendelian Transmissions (WT) and the optimal strategy for combining information across paths, the SM approach, the final version of our model was re-implemented in order to create releasable software. The implementation is called PACE and was developed by Dr Iain Mathieson (<https://github.com/mathii/pace>). PACE stores the Viterbi matrix sparsely so that it is significantly more computationally and memory efficient compared to the naive HMM implementation. In all other respects, PACE functions as described in Chapter 2.

In order to confirm that PACE maintains the properties discovered earlier in this chapter, a final simulation investigation was carried out by Dr Mathieson. Data were simulated as in Section 3.3.1, for two different sample sizes of 50 and 500 genotypes and two different recombination rates: 0.5cM/Mb and 1.5cM/Mb. It was then of interest to observe how $\mathcal{A}$, our measure of accuracy, and $\mathcal{U}$, the proportion of unphased sites changed as we increased the number of suboptimal paths considered.

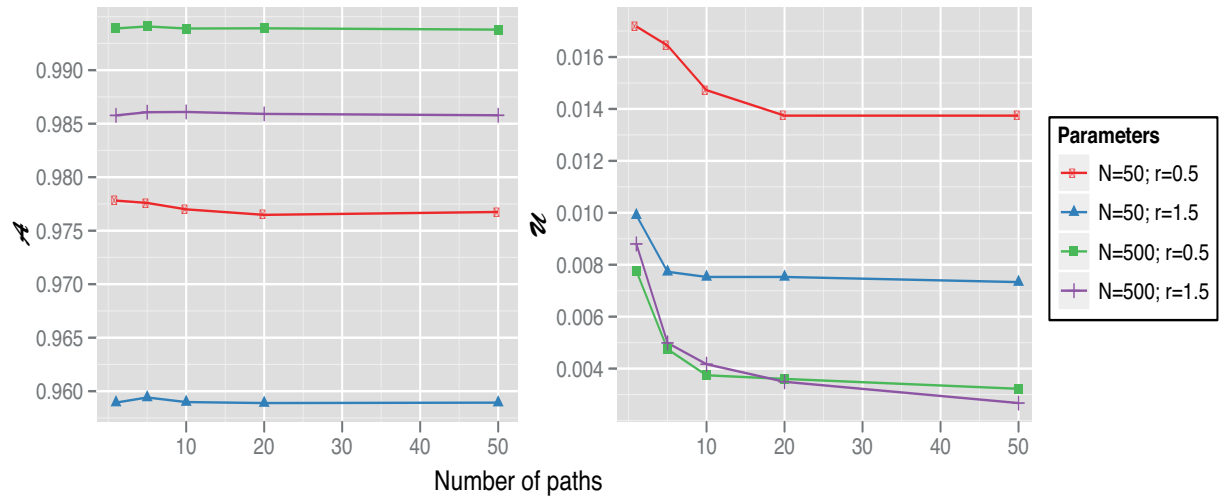


Figure 3.7 **PACE**: Performance as the number of paths increases. $\mathcal{A}$ is accuracy (refer to text in Section 3.2 for the full definition); $\mathcal{U}$ is the proportion of unphased sites; N =number of genotypes; r =recombination rate (cM/Mb).

Figure 3.7 confirms earlier observations: adding extra paths does not lead to a deterioration in accuracy, while there is a significant improvement in $\mathcal{U}$. PACE (and presumably the original implementation) has the expected and useful property of having an improved phasing performance for the larger sample sizes, with only a small compromise in accuracy due to a higher recombination rate.

3.3.8 Remarks

In the experiments presented here, only the Viterbi algorithm was used for identifying suitable pseudo-parents for haplotype estimation. It would have been possible to instead search for the individuals with maximal posterior at each site, using the Forward/Backward algorithms. In fact, given that posteriors are computed at every position, it would have been straightforward to obtain the best K pairs of individuals, and the best most informative pairs at every SNP, making it possible to phase each heterozygous site without memory restrictions.

However, the fact that path likelihood was shown to have no influence on phasing performance (see Figure 3.2) suggests that no advantage will be gained by considering a F/B approach for picking pseudo-parents. The Viterbi algorithm provides a straightforward way for obtaining sequences of hidden states with few parent transitions, simplifying pseudo-mother and pseudo-father alignment. Conversely, a F/B implementation would require a more involved deconvolution of hidden states with optimal posterior across sites. This is why the Viterbi-based scheme explored here was considered sufficient.

3.4 Phaser competition

Having finalised the parameters and the algorithm behind our phaser, it is possible to compare its performance against state-of-the-art algorithms that are available. At the time when these investigations were carried out, SHAPEIT (Delaneau et al. [2012]), one of the more competitive methods currently available (Delaneau et al. [2012], Browning and Browning [2011b]) was not established as a leading

phaser. As a result, our model was instead compared to the BEAGLE haplotype graph approach (Browning and Yu [2009]) and the LS-driven MCMC approach of IMPUTE2 (Howie et al. [2009]).

Both BEAGLE and IMPUTE2 were applied using their default settings. The comparisons explored here focus exclusively on phasing accuracy and not computational efficiency. Concerning run-time, BEAGLE has a clear advantage, requiring only a few minutes to phase a chromosome for thousands of samples. IMPUTE2, by comparison, would require a few days. In this thesis, haplotype estimation was only explored using our full HMM, with single hidden state space transitions as the only modification to the model for computational efficiency. In this form, our model cannot be run on more than a few hundred individuals; even in this case, run time for a single individual would be in the order of a few weeks. Recall that computations run in parallel across individuals and therefore, given availability of a cluster, run time for a single individual is indicative of run time for the whole sample.

For the small sample sizes simulated for the purposes of this investigation, run time for our phaser was the same order of magnitude as that of IMPUTE2. A direct comparison is not straightforward as IMPUTE2 is restricted to phasing all individuals simultaneously via an MCMC approach, whereas our phaser was run separately per person, with run time being restricted only by the number of processors available.

3.4.1 Data

We compared the three phasers (PACE, BEAGLE and IMPUTE2) on simulated data and on real data. Data sets were simulated (again using R code provided by McVean, <http://www.stats.ox.ac.uk/~mcvean/R/coalescent.r>) for a variety of sample sizes, N , and recombination rates, r in order to compare phaser performance as both of these changed. In particular, we considered $N \in \{10, 50, 75, 100, 200, 500, 1000\}$ and $r \in \{0, 0.25, 0.5, 1, 1.5, 2\}$, r measured in cM/Mb. We assumed an effective population size of 10,000.

The real data considered were chromosomes 6, 20 and X, separately for the CEU and YRI populations from the HapMap PHASE III (International HapMap 3 Consortium et al. [2010]) project. The HapMap data were selected as at the time that this investigation was carried out (prior to the 1000 Genomes (1000 Genomes Project Consortium et al. [2010, 2012]) release), it was the only open-source database of high quality haplotypes; the data underwent rigorous quality control described, for example, in International HapMap 3 Consortium et al. [2010] and Moutsianas [2011]. Moreover, the database consisted of trio and duo samples, allowing for the determination of high quality haplotype estimates that were mostly determined by transmission, rather than stochastic estimation. Using such high quality haplotypes is crucial, as they will be used as the true haplotypes that the phasers' results will be compared against.

For all chromosomes, we considered the duo, trio and unrelated samples. Using the duo and trio individuals, we looked at sites that were phased by transmission, thus obtaining the haplotypes of the children. For chromosomes 6 and 20, the phasers were then run on these children and on the unrelated samples, but results were only assessed for the children, for which the high-quality haplotypes were known. For chromosome 6, we also isolated the HLA region ($\sim 29\text{Mb}$ - 33Mb) and explored how the phasers behaved in that region specifically.

For chromosome X, the haploid male samples were randomly paired to create pseudo-diploids. The phasers were then run on the female samples and the male pseudo-genotypes, but were only assessed on the latter. All in all, for the CEU population, for chromosomes 6 and 20, 69 individuals were included, of which 52 were from duos or trios, and for the YRI population, 63 individuals were considered, of which 54 were assessed. For chromosome X, there were 15 pseudo-diploids, out of 40 individuals for the CEU population, while for the YRI population 19 pseudo-diploids were obtained, out of 45 individuals. The data are tabulated in Table 3.9). Just like with all the simulation-based analyses that we have performed, all singletons are excluded from the real data as well.

Pop	Chr	Candidate Parents	Children	Pseudo-males
CEU	aut.	69	52	0
CEU	X	40	0	15
YRI	aut.	63	54	0
YRI	X	45	0	19

Table 3.9 **HapMap III data.** The table summarises the sample sizes considered for each analysis. Pop = population; Chr = chromosome; Candidate parents = the total number of individuals considered as a candidate parents; Children = the number of individuals within the candidate parent set, which were obtained by transmission from trio and duo samples and which were the ones whose haplotype estimates were assessed; Pseudo-males = the number of pseudo-diploids within the candidate parent set, which were obtained by pairing male haploids and which were the ones whose haplotype estimates were assessed; aut = autosome, referring to chromosomes 6 and 20 that were considered in this analysis.

3.4.2 Results

Considering first the simulated data (Figure 3.8), we see that the phasers are comparable, with no method out-competing the other two for all combinations of sample size, N , and recombination rate, r . When sample size is small and the recombination rate is low, PACE performs significantly better than the other two methods. As sample size increases, the methods perform near-identically.

For the higher recombination rate, BEAGLE suffers for small sample sizes, and PACE suffers for the larger sample sizes, with accuracy not increasing as fast as for the other methods. When recombination rate increases, BEAGLE and PACE perform similarly (and worse than IMPUTE2) when sample size is small, but PACE is the weakest for the larger sample size, with BEAGLE restoring its accuracy at the level of IMPUTE2.

Overall, IMPUTE2 performs the best for larger samples and is the most robust to increasing recombination rate, while PACE performs best for smaller samples and lower recombination rates (Figure 3.8). Results are comparable, however, for BEAGLE, which is computationally the fastest algorithm. Finally, since only a

3.4 Phaser competition

Table 3.10. **HapMap III Phasing Results** C = CEU; Y = YRI; The values tabulated for each phaser (Pace, BEAGLE and IMPUTEv2) are the accuracy $\mathcal{A}$ (%). For the X chromosome results from the male individuals were considered, while for the autosomal chromosomes we considered the results for the parents from trio data for which true phase was known with greater confidence. Pace used the top 10 paths with WMT (with $\zeta = 0.05$, however) and the SM approach. The values in bold indicate the maximum $\mathcal{A}$ achieved for the respective column.

Method	C HLA	C6	C20	CX	Y HLA	Y6	Y20	YX
PACE	99.64	98.70	98.29	97.16	99.55	98.06	97.49	95.61
BEAGLE	99.63	98.71	98.28	97.12	99.59	98.04	97.46	95.30
IMPUTE2	99.50	98.88	98.55	96.30	99.55	98.96	98.59	96.15
$\mathcal{U}$	1.48	1.52	1.69	1.63	1.83	4.40	4.99	4.06

small fraction of heterozygous sites remains unphased, we see that phasing these sites at random makes no difference to PACE’s overall accuracy.

Turning to our results for real data (Table 3.10), we see near-identical performance for all methods. Encouragingly, we obtain near-perfect results over the HLA region for both populations. Results on other chromosomes are slightly better for the CEU population than for the YRI, consistent with the greater diversity present in the YRI population that makes phasing a more challenging statistical problem to solve, particularly considering the small sample sizes that are being considered here. While for the CEU population all methods perform equally well, IMPUTEv2 outperforms the rest in the case of the YRI population.

3.4.3 Conclusions

PACE does not achieve better results than BEAGLE and IMPUTE2, it is however a competitive phaser in its own right, particularly for smaller sample sizes and for

3.4 Phaser competition

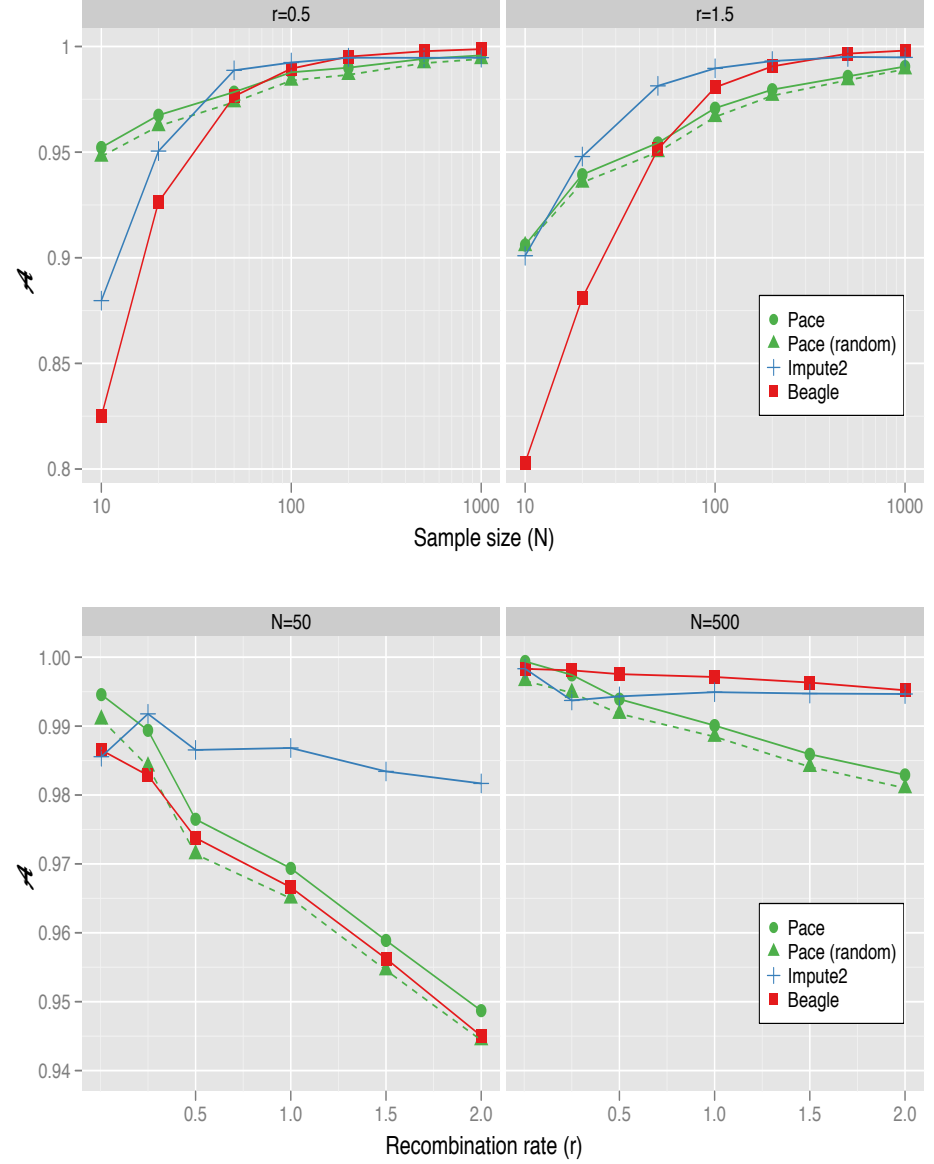


Figure 3.8 **Method comparisons.** N = number of genotypes in the sample; r = recombination rate, cM/Mb; $\mathcal{A}$ = our measure of accuracy. For Pace we consider error rate for all phasable sites (solid green) and after phasing any left-over sites at random (dashed green). Pace uses the top 10 paths with WMT ($\zeta = 0.01$) and the SM approach.

regions with lower recombination rate. We observe, however, that while PACE’s performance improves for larger data sets, it does so less effectively than the other methods (Figure 3.8). PACE does not directly incorporate the more abundant population level information that is available when there are more individuals in the sample. However, as sample size increases we do expect that more individuals will exist in a sample that will be suitable pseudo-parents. In fact, the TMRCA with a $NN \sim O(\frac{1}{n})$ and we therefore expect that with a greater sample size there will be more people with long-range compatibility. While this might be true, however, it is not obvious that these will be the individuals that are most suitable for phasing. In contrast, our results suggest that by using the weighted Mendelian emissions we are actually selecting individuals with which the MRCA is further up the genealogical tree of the sample. The number of such individuals therefore grows more slowly, consistent with our observations, thus explaining the slower rate of improvement. We explore this in full detail in the next chapter.

Chapter 4

Finding Nearest Neighbours: detection and haplotype sharing

4.1 Overview

This chapter explores the model that we have developed in the context of the detection of recent shared ancestry within an unrelated sample. The successful application of our HMM in the estimation of haplotypes from genotype data indirectly demonstrates its ability to identify individuals in the sample that are closely related with the query genotype locally. The question of characterizing and establishing this close and local relationship therefore naturally arises: how are these pseudo-parents related to the query genotype? Are they really close relatives, in the genealogical sense? What does ‘close’ mean in this context?

In the introduction we discuss how the relationship between individuals in a sample is fully captured by the ARG of that sample. With this in mind, in order to fully characterize the coancestry between the query genotype and the pseudo-parents selected by our model, it would suffice to record:

1. The number of nodes in the sample's genealogical tree separating them.
2. Their TMRCA; in other words, the ARG's branch lengths between them.
3. The length of the haplotypes that they share, which illustrates the distance over which their relationship in the ARG does not change.

The number of nodes measures relatedness discretely (if all events in a pedigree are detectable): haplotypes are siblings, cousins and so on, on the genealogical tree. Conversely, the other two characterize relatedness in a continuous fashion.

These metrics are clearly related and the coalescent model provides a framework for investigating them. However, they are influenced by the demographic history of a sample, which adds a layer of difficulty in inferring their relationship, particularly when considering all three simultaneously. Similar to the work of others (for example Gravel [2012] and Harris and Nielsen [2013]), we are interested in estimating these metrics and using them as summaries of structure in a sample. Our model promises to explore patterns of recent relatedness and so it is important to determine our pseudo-parent measurements with respect to all three metrics.

However, we can extend this discussion and ask whether the selected pseudo-parents are in fact the query genotype's closest relatives (NNs; illustrated in Figure 1.2) in the sample. Very recent coancestry is informative of relatedness patterns specific to rare variants and is further informative on recent demographic history. This does not have to be the case for phasing to be successful but it is of interest to establish whether our model could be calibrated to specifically aim to detect NNs in the place of pseudo-parents, if the latter are not NNs already. We should therefore determine whether each pseudo-parent contains a haplotype that is one of the query genotype's haplotypes' siblings on the marginal genealogical tree.

The genealogical relationship between individuals in an unrelated sample is typically unknown. In real data analysis what is instead straightforward to estimate is the extent of haplotype sharing between the individuals. The second question that therefore naturally follows is whether haplotype sharing is maximised between individuals that are NNs. Several methods (Browning and Browning

[2012], Harris and Nielsen [2013], Kong et al. [2008]), including the one introduced in this thesis attempt to detect local relatedness by searching for extensive sharing between individuals. It is crucial to understand whether this is actually a different pursuit than the one for NN detection.

This chapter is split into two main parts. First, we begin our investigation with a simulation study that explores this relationship between maximal haplotype sharing and NN status, assuming the standard coalescent framework. The coalescent was formalised by Kingman (Kingman [1982a,b,c]; for reviews see Hein et al. [2004] and Wakeley [2008], among others) and is the most widely used population genetics model. To assess whether a method is effective, we should assess whether it detects the amount of sharing that is expected and we should explore who this sharing occurs with. Coalescent simulations allow us to benchmark these expectations.

In the second part, we use simulated data to explore the behaviour of our model with respect to the following questions:

1. What is the extent of the relatedness between our selected pseudo-parents and the query genotype? Further, can we calibrate our method to identify NNs?
2. Can we detect maximal haplotype sharing? More specifically, is the amount of haplotype sharing between individuals what is expected on average under the standard coalescent?
3. How are these questions influenced by the presence of phasing and/or genotyping errors in the data?

Haplotype data are rarely obtained experimentally and will therefore carry errors from the phasing process. These only add to errors incurred from the genotype-calling stage of data collection. It is important to consider these questions in this context as haplotype sharing and identifying closer relatives are defined by the similarity between individuals and will thus be affected by erroneous variation (Browning and Browning [2012], Lawson et al. [2012]).

In our assessment, we considered both the Viterbi and F/B implementations of our HMM. For the Viterbi implementation, we explored the performance of all

4.2 Limits on the ability to detect Nearest Neighbours by means of haplotype sharing

three options of emission scores. The graph approximation (introduced in Section 2.4.2) and HMMsub (Section 2.4.3) are also incorporated in the investigation here.

4.2 Limits on the ability to detect Nearest Neighbours by means of haplotype sharing

We first consider the relationship between maximal haplotype identity and NN status. Formally, we say that chromosome j is a NN of chromosome i in the sample if the time to the most recent common ancestor (TMRCA) of i and j is less than or equal to the TMRCA of i and any other individual in the sample. NNs are expected to share extended haplotypes, since the expected shared haplotype length is inversely proportional to the TMRCA between any two individuals. In any genealogy, a particular haploid sample will have one or more NNs (the NN set) at a given position. However, which other sample it shares greatest haplotype identity with will depend on the rate of mutations (relative to recombination) and where recombination events occur in the genealogy. If mutation is rare relative to recombination, or rare variants are not assayed (as is typical of SNP genotype data sets), then, due to the stochastic nature of recombination, sample-pairs sharing maximal identity may not be genealogical NNs. For example, if no mutations occur on the branch separating two NNs with their genealogical cousin and if this branch is short, then it is possible that recombination events will be such that a longer haplotype will be shared between an individual and the cousin, than their NN. In general, there are of course other reasons why NNs may not be identifiable, for instance insufficient typing that does not reveal enough of the variation present in a sample and genotyping and phasing errors that disrupt IBD chunks.

In any case, the question therefore arises of whether maximal haplotype sharing can be used as a proxy for recent relatedness, and for NN status in particular. Is looking for one the same as looking for the other? Moreover, how much is the maximal sharing that we expect between a query individual and the person (whether it is the NN or not) with which they share the most?

4.2 Limits on the ability to detect Nearest Neighbours by means of haplotype sharing

4.2.1 Simulations under the coalescent

To assess the frequency of NN mis-assignment using maximal haplotype identity we performed coalescent simulations of haploid samples and asked whether the marginal tree for true NNs remains unchanged over greater distances than for pairs of more distantly related individuals. The assumption made in this strategy is that the region over which the TMRCA for two individuals does not change defines maximal haplotype identity between them. In other words, we ignore the effect of mutation.

For a sample of size n , we constructed a marginal coalescent tree (e.g. Figures 1.2 and 4.1) assuming the coalescent with constant population size using the standard approach (Hein et al. [2004]), which we outline in Algorithm 9.

Algorithm 9 Generating a coalescent tree

Set $k = n$ and $\mathcal{J} = \{1, \dots, k\}$

while $k \geq 2$ **do**

1. Generate $u \sim \mathcal{U}(0, 1)$
2. $T_k = -\frac{2}{k(k-1)} \log(u)$ {since $T_k \sim \mathcal{E}(\binom{k}{2})$ }
3. Sample 2 individuals, $i, j \in [1, k]$ uniformly and at random from $\mathcal{J}$
4. Set $N_k = \{i, j\}$, $\mathcal{J} = \{1, \dots, k\} / \{i, j\} \cup N_k$
5. $k \rightarrow k - 1$

end while

The coalescent tree, T , is then defined by the sets $N_k, k \in \{n, \dots, 2\}$. These are the sets of nodes of the tree which carry the individuals that coalesce at each time point k . The branch lengths are defined by the exponential random variables, T_k , which are the waiting times between coalescent events while there are k individuals present in the sample.

This simulated tree represents the marginal tree for the sample at a certain position. Ignoring mutations, shared haplotypes between chromosomes are defined by the genomic distance over which the topology of the tree that joins them does not change. For every pair of chromosomes we should therefore explore distances to recombination events to the left and to the right of the current marginal tree.

4.2 Limits on the ability to detect Nearest Neighbours by means of haplotype sharing

Alternatively, we may simulate recombination events in the genealogy first and then consider which events are relevant to every pair of chromosomes in the sample. This is the simulation strategy that was adopted here. Given a certain topology joining two chromosomes, the distance to a recombination event is exponentially distributed, with parameter half of the length of the relevant topology. The length of the topology is $2 \times \text{TMRCAs}$ between the two chromosomes. This distance is in units of ρ , the recombination rate at that position. Equivalently, we can add a distance separately to each branch b of length t_b ; this distance $\sim E(\frac{t_b}{2})$. Then haplotype length is determined by the minimum of each of these distances. Let $\mathcal{B}$ be the space of branches that make up T . The equivalence follows from a property of the exponential distribution: if L is the total tree length and $L = \sum_{b \in \mathcal{B}} t_b$, then $X \sim \min_{b \in \mathcal{B}} (Y_b)$ if $X \sim \text{Exp}(L)$ and $Y_b \sim \text{Exp}(t_b)$.

Given the marginal coalescent tree T , two distances to recombination events are simulated for each branch. For each pair of chromosomes, the only part of T that will influence the length of their shared haplotype is the set of branches that join them. Suppose this pair is $p = \{i, j\}, i \neq j$ and denote the set of branches that join them by $\mathcal{B}_p$. Their shared haplotype will be bounded on the left by the minimum of all the distances to the left simulated for each of the branches in $\mathcal{B}_p$. Similarly, the shared haplotype will be bounded on the right by the minimum of all the distances to the right from the set of those branches. Consequently, the overall length of this haplotype is the sum of these two distances. Call this length d_p . In this set-up, we are ignoring the possibility of back coalescence events: cases where two recombinant lineages coalesce back together before any other events take place. In the absence of mutation, this would make no difference to the genomic composition of the samples. We are also assuming that all the recombination events that we are simulating are detectable. The overall algorithm for the simulation, which is illustrated in Figure 4.1, is as follows:

Algorithm 10 Determining the length of shared haplotypes

Simulate distances to recombination events in the genealogy, $\mathcal{B}$, by considering each branch separately:

4.2 Limits on the ability to detect Nearest Neighbours by means of haplotype sharing

for $b \in \mathcal{B}$ **do**

 Generate $u_1, u_2 \sim \mathcal{U}(0, 1)$

 Set $r_l^b = -\frac{2}{t_b} \log(u_1)$ and $r_r^b = -\frac{2}{t_b} \log(u_2)$

end for

Next, obtain the length of the shared haplotype for each pair of chromosomes

for $i \in \{1, \dots, n\}$ **do**

for $j \in \{i + 1, \dots, n\} / i$ **do**

 Set $p = \{i, j\}$ and identify $\mathcal{B}_p$, the set of branches joining i and j

 Set $d_l = \min\{r_l^b : b \in \mathcal{B}_p\}$ {This is the left bound}

 Set $d_r = \min\{r_r^b : b \in \mathcal{B}_p\}$ {This is the left bound}

 Set $d_{i,j} = d_l + d_r$

end for

 Set $D_i = \max_j(d_{i,j})$ and $\text{NN}_p^i = j$ such that $D_i = d_{i,j}$ {Select an individual uniformly and at random if there are more than one.}

end for

We record maximal haplotype sharing for each individual so that we can obtain a benchmark for the expected value. This expected value cannot be derived using the standard exponential distribution as the individual with which the maximal haplotype is shared, and therefore their TMRCA, is unknown. Moreover, we may now ask whether each individual i shares its maximum haplotype chunk, of length D_i , with (any of) its true NNs. In other words we ask whether $\text{NN}_p^i = \text{NN}$.

In our simulations using Algorithm 10 it is possible that the the branch for which d_l and d_r for some chromosome i , is in fact the branch directly above i (a terminal branch). When this is the case, it is in fact impossible to distinguish between true NNs and NN_p^i . This illustrates how maximal haplotype sharing can fail in distinguishing NNs with more distantly related individuals and it is therefore also of interest to investigate how often this is the case.

4.2 Limits on the ability to detect Nearest Neighbours by means of haplotype sharing

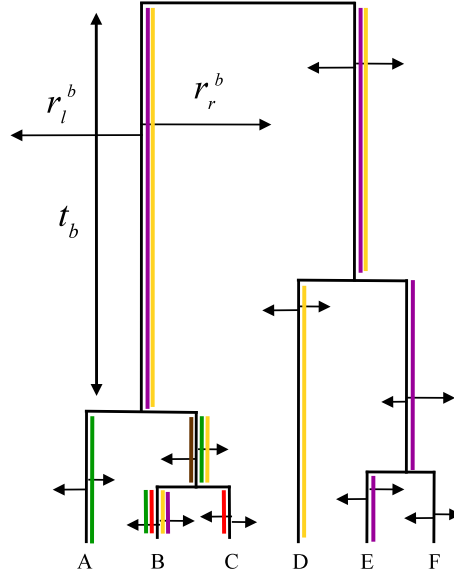


Figure 4.1 **Adding recombination distances to a coalescent tree.** The diagram shows a coalescent tree (black) for 6 chromosomes (A-F). Each branch has a distance to a recombination event to the left and to the right, marked by the black arrows. The top left branch is branch b , of length t_b . The distance to the left for this branch is $r_b^l \sim E(\frac{t_b}{2})$ and the distance to the right is $r_b^r \sim E(\frac{t_b}{2})$, in units of the recombination rate ρ . For a query genotype, B say, its NN is C and its cousin is A. For B and some other individual, we select the branches that join them. For example, for B and C, the branches that join them are highlighted in red; for B and A in green; B and D in yellow; B and E in purple. For each pair, the highlighted branches are the ones considered for determining the length of their shared haplotype, $d_{B,\star}$. Refer to text for a full description.

4.2.2 Simulations under an exponential growth model

If there is NN miss-assignment because of extended haplotype sharing, this will only be accentuated if there has been exponential growth in the genealogy. The branch length in a coalescent tree is scaled by the effective population size, N_e . Therefore, we can change the branch lengths of a coalescent tree that has been simulated under the constant size model (see Algorithm 9) simply by rescaling these lengths proportionally to the changes that have occurred to N_e . When there has been exponential growth, this manifests itself in the topology of the tree by

4.2 Limits on the ability to detect Nearest Neighbours by means of haplotype sharing

extending more recent branch lengths and contracting older ones, thus changing the tree into a ‘star-like’ shape (Hein et al. [2004]). The difference between siblings and cousins on the tree is therefore made smaller, and the terminal branches are longer allowing for more confusion to be possible between NN and NN_p .

Assuming an exponential growth model, the effective population size changes from N_S to N_B over a growth period G . We have that $N_S = N_B \cdot e^{-\beta \cdot G}$, with the growth rate β being $\frac{1}{G} \cdot \log(\frac{N_B}{N_S})$. Coalescent time in the constant size case is:

$$\hat{t} = \frac{\exp(\beta \cdot G) - 1}{2N_B \cdot \beta}$$

and for the case of exponential growth it is transformed to (Hein et al. [2004]):

$$\hat{t}_g = \frac{1}{2N_B \beta} \cdot \log(1 + 2N_B \beta \hat{t})$$

In our simulations we build the coalescent trees assuming constant population size and then rescale the branch lengths using these parameters. If t is the time we are simulating that measures time back in the constant size case, our rescaled time is t_g and:

```

if  $t < \hat{t}$  then
     $t^g = \frac{1}{2N_B \beta} \cdot \log(1 + 2N_B \beta t)$ 
else
     $t^g = \hat{t}_g + \frac{N_S}{N_B} (t - \hat{t})$ 
end if

```

where $\hat{t}$ and $\hat{t}_g$ are the times in the past when growth began for the constant-size and growth case respectively.

Let $S_j = \sum_{k=n}^{k=j} T_k$ be the total time until the j th coalescence event in the constant-size case, and $S_j^g = \sum_{k=n}^{k=j} T_k^g$ be the equivalent time under the growth model. We use the above transformation to update our algorithm (Algorithm 9 for constant N_e) for obtaining a coalescent tree.

Algorithm 11 Generating a coalescent tree under exponential growth

```

Set  $k = n$  and  $\mathcal{J} = \{1, \dots, k\}$ 
while  $k \geq 2$  do
    1. Generate  $u \sim \mathcal{U}(0, 1)$ 

```

4.2 Limits on the ability to detect Nearest Neighbours by means of haplotype sharing

```

2.  $T_k = -\frac{2}{k(k-1)} \log(u)$  and set  $S_k = \sum_{j=n}^{j=k} T_j$ .
if  $S_k < \hat{t}$  then
    3.  $S_k^g = \frac{1}{2N_B\beta} \cdot \log(1 + 2N_B\beta S_k)$ 
else
    3.  $S_k^g = \hat{t}^g + \frac{N_S}{N_B}(t - \hat{t})$ 
end if
4. Set  $T_k^g = S_k^g - S_{k+1}^g$  {The branch length for the current coalescence event}

5. Sample 2 individuals,  $i, j \in [1, k]$  uniformly and at random from  $\mathcal{J}$ 
6. Set  $N_k = \{i, j\}$ ,  $\mathcal{J} = \{1, \dots, k\} / \{i, j\} \cup N_k$ 
7.  $k \rightarrow k - 1$ 
end while

```

4.2.3 Results

The simulations described here were carried out for varying sample size $n \in \{10, 50, 100, 200, 500, 750, 1000\}$, with 1000 repeats for each size, both for the constant size and exponential growth cases. In the latter scenario we considered N_e increasing from 10,000 (N_S) to N_B over G generations, for $N_B \in \{1 \times 10^6, 2 \times 10^6, 3 \times 10^6, 4 \times 10^6\}$ and $G \in \{100, 200, 300, 400\}$, which reflect a range of parameters that are plausible for human population history.

Under the constant-size model, we find that in approximately 20% of cases the sample with maximal identity is not a genealogical NN (Figure 4.2). Moreover, in 5% of cases it is not possible to distinguish true NNs from others (Figure 4.2 inset). These proportions are largely independent of sample size, with the observed trends levelling off when there are more than ~ 50 chromosomes in a sample. The rate of convergence to these limits is also surprisingly fast; even for samples with 25 individuals, there is a significant probability that one shares their longest haplotype with an individual that is not a NN.

This constant rate of miss-assignment is in fact only constant for the constant-size case. Under exponential growth (Figure 4.3), detection of NNs becomes even harder for increasing sample size. The ability to distinguish NNs unambiguously

4.2 Limits on the ability to detect Nearest Neighbours by means of haplotype sharing

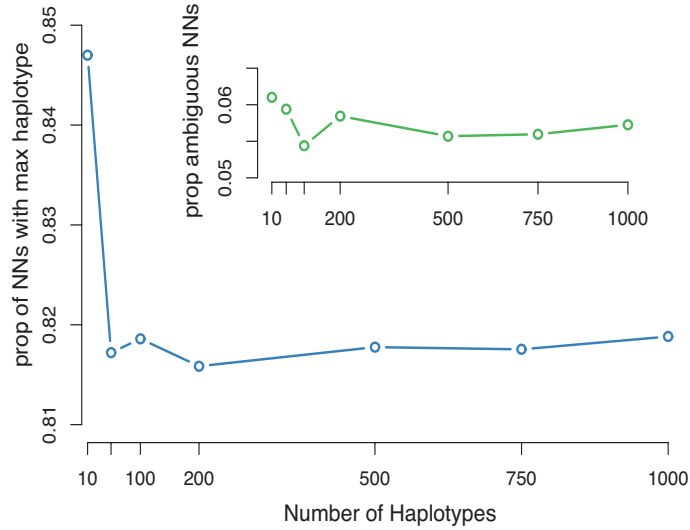


Figure 4.2 **Constant population size: NN vs NN_p** . The main figure shows the proportion of times that the individual for which maximal haplotype sharing is greatest is actually a NN. Inset: The proportion of times that we cannot distinguish between NNs and other individuals using maximal haplotype sharing.

also deteriorates under these conditions (Figure 4.3 inset) . Given a fixed sample size, both of these effects are further accentuated when N_B and G increase (Figure 4.4). A larger N_B increases the growth rate, β and since $G \gg 1$ when G goes further back in time there are longer times over which rescaling of branch lengths can occur. A more prolonged growth period or growth with a greater change in N_e both accentuate the effects seen in Figure 4.4. In both cases it is reasonable therefore that the difference between NNs and cousins or other very similar recent relatives becomes less distinguishable, in terms of the branch lengths that join them.

It follows that we should look at how different the selected NN_p s are compared to the NNs:

1. How much older is the TMRCA with NN_p compared to the NN.
2. How does the haplotype length shared with a NN compare to that shared with NN_p s?

4.2 Limits on the ability to detect Nearest Neighbours by means of haplotype sharing

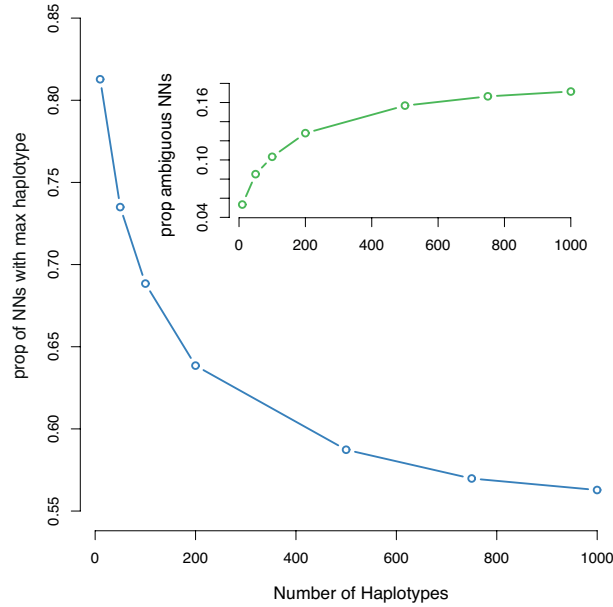


Figure 4.3 **Exponential growth: NN vs NN_p** . The main figure shows the proportion of times that the individual for which maximal haplotype sharing is greatest is actually a NN. Inset: The proportion of times that we cannot distinguish between NNs and other individuals using maximal haplotype sharing. The simulations here used $N_B = 4 \times 10^6$ and $G = 400$.

We discover that the difference between the TMRCA of the true and inferred NNs is typically small (Figure 4.5) both for the constant-size and the growth cases, with most observations lying close to the $y = x$ line. For both cases we do however still see differences, albeit rarely, that are different even by an order of magnitude. The larger differences in TMRCA peak around the mean. Zooming into these regions, the insets in Figure 4.5 show the density of the differences, confirming that in all cases its mass is at about two orders of magnitude smaller than that of the mean. These findings support the hypothesis that whenever the true NN is not identified, it is typically in favour of another individual with which coancestry is almost as recent as with the NN and the shorter haplotypes obtained with the real NNs is a consequence of the stochasticity of the recombination process.

Similarly, Figure 4.6 shows that haplotype lengths with NNs and NN_p s are also quite similar, with the median of their ratio being mostly at 90% (Figure

4.2 Limits on the ability to detect Nearest Neighbours by means of haplotype sharing

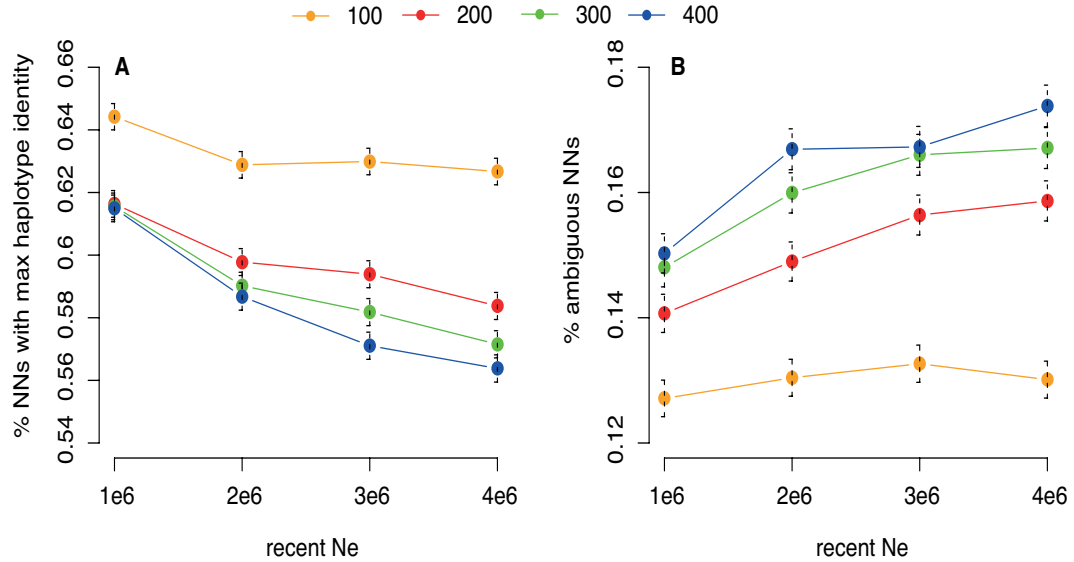


Figure 4.4 **Identifying NNs under the coalescent with exponential growth.** We are assuming a model of exponential growth from an effective population (N_e) that starts at 10,000 and grows over a period of G generations. The diagrams show the proportional of maximal shared chunks that are shared with a NN (left) and the proportion of NNs that are indistinguishable from other individuals (right) as a function of recent N_e and G . The different values of G that are considered are illustrated by the different colours: blue = 400; green = 300; red = 200; orange = 100. In our simulations we have used a sample of 1000 haplotypes. The error bars depict 95% confidence intervals.

4.6 insets). The difference between the chunks is, as expected, greater under exponential growth. In this case and for the larger sample size, we also observe that miss-assignment occurs more often when the chunks with the true NNs are shorter than the observed median (as we would expect).

In practice, consideration of mutations, and in particular rare variants, will provide additional resolution to relationships and so some improvement on these observations is expected. Nevertheless, these results demonstrate that there are fundamental limits on the ability to detect genealogical NNs through extended haplotype sharing - a result of the inherent stochasticity in the genealogical process. These limits are particularly relevant for populations which have experienced population explosion, and similarly bottlenecks, which we know is the case

4.2 Limits on the ability to detect Nearest Neighbours by means of haplotype sharing

for humans. The reduction in the levels of diversity that result from such demographic histories requires that we distinguish between these metrics of relatedness (NN relationship or maximal haplotype sharing), particularly when they are the ones used to make inference on demographic history, timing of historical events and on disease association studies. In other words, whether we aim to detect NNs or individuals with which there is maximal haplotype sharing should depend on the type of inference that we are trying to make.

4.2 Limits on the ability to detect Nearest Neighbours by means of haplotype sharing

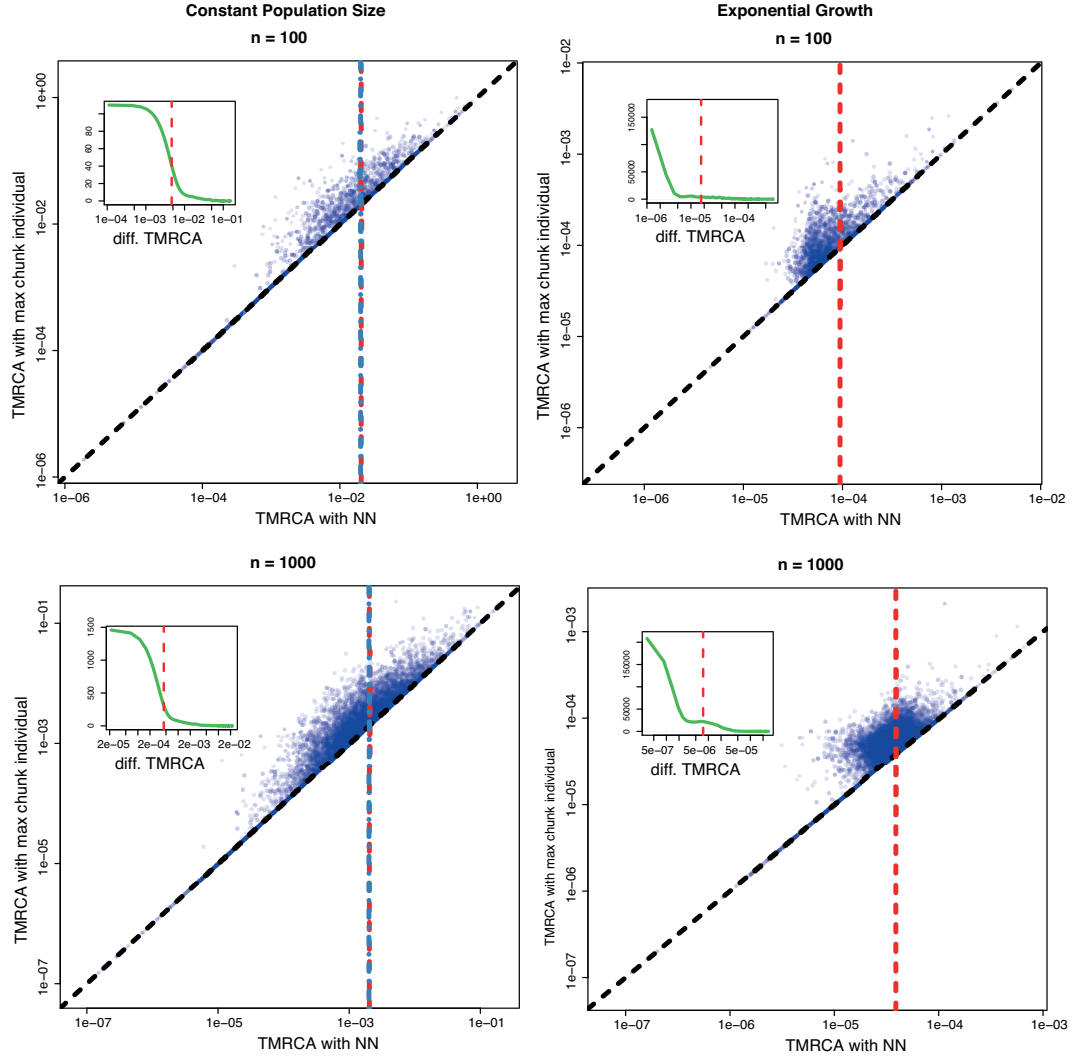


Figure 4.5 **Differences in TMRCA between NN and NN_p** Each main figure shows the TMRCA with an NN against the TMRCA with the NN_p for every haplotype from our simulations. The blue dashed line is the expected mean for the TMRCA with a NN under the standard coalescent with a constant population size. The red dashed line is the observed median. We are recording the median here because the variance in shared haplotype chunks is infinite. The insets show the density of the difference between the NN TMRCA and NN_p TMRCA around the observed mean. The left column shows the result for constant population size and the right column under a model of exponential growth ($N_B = 4 \times 10^6, G = 400$). NN_p is the maximum chunk individual.

4.2 Limits on the ability to detect Nearest Neighbours by means of haplotype sharing

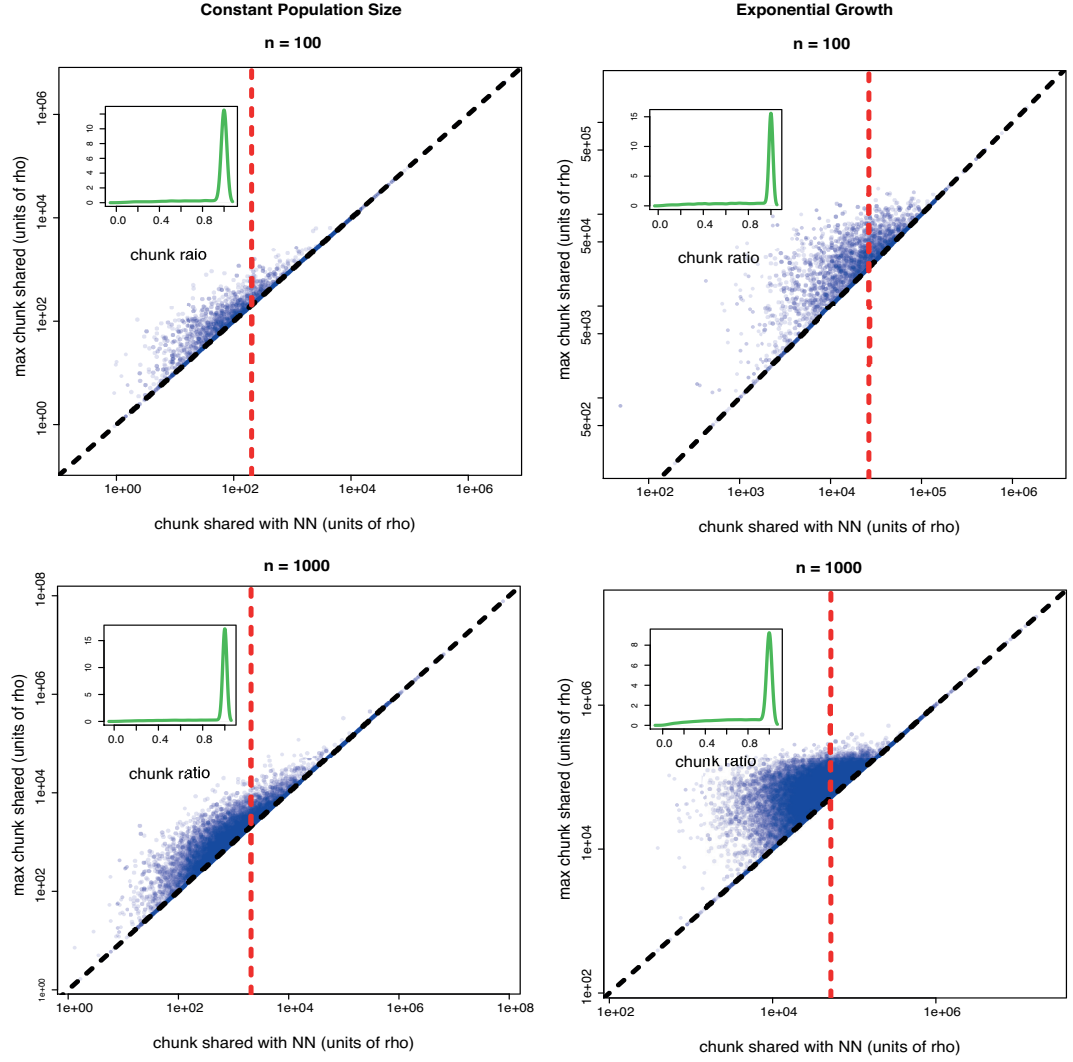


Figure 4.6 **Differences in shared haplotype lengths between NN and NN_p for haplotype lengths.** Each main figure shows the length of the shared haplotype with an NN against the length with the NN_p (which was the maximum attained) for every chromosome from our simulations. The red dashed line is the observed median of the length of the shared haplotype with an NN. The insets show the density of the ratio between the NN length and NN_p length around the observed median. The left column shows the result for constant population size and the right column under a model of exponential growth ($N_B = 4 \times 10^6$, $G = 400$). NN_p is the maximum chunk individual.

4.3 A statistical method for detecting NNs and extended haplotype sharing from unphased genotype data

As a brief reminder, for a query individual our methods' goal is to identify at every position along the genome two (or more) other individuals from the sample who contain chromosomes which are closely related to those in the target. In fact, we would further like these recent relatives to be genealogical NNs to the two chromosomes present in the target, or to otherwise explore how closely related they are.

Following the ideas introduced by Kong et al. [2008], we refer to them as pseudo-parents. However, unlike the LRP method, we wish to allow for the pseudo-parents to change along the sequence and for a low degree of divergence (arising through error or true variation) between the target and pseudo-parental genomes. We note that, particularly in small sample sizes, both NNs may be present in a single individual (and occasionally the two target chromosomes could be NNs of each other). However, this is rare and for most uses of NN information (for example, haplotype estimation) we wish to constrain the pseudo-parents to be distinct from each other and from the target.

In view of the results obtained earlier in this chapter, we should investigate how we fare in the balance between detecting the expected maximal haplotype sharing versus identifying true NNs. We consider the three different emission processes: Mendelian transmission (MT; Table 2.1 with $\zeta=0.5$), weighted Mendelian transmission (WMT; Table 2.1 with $\zeta = 0.01$) and compatibility (CE; Table 2.2). All three indirectly demonstrated their capacity to detect close relatives by allowing for the effective estimation of haplotypes (see Chapter 3).

We also allow for a low rate of Mendel-incompatible inheritance patterns (setting $\epsilon = 0.01$, Tables 2.1 and 2.2). We identify shared haplotype chunks, by studying the regions over which our choice of pseudo-parents remained constant. This was illustrated in Figures 2.5 and 2.7. A formal description of our algorithm and the HMM that we employ (the genotype-HMM), was given in Chapter 2.

4.3 A statistical method for detecting NNs and extended haplotype sharing from unphased genotype data

We applied our methods independently for each target individual in simulated data, excluding the target from the set of possible pseudo-parents. Our adaptation of the Viterbi algorithm (Section 2.2.3) was used to find the set of paths with the maximum likelihood. Alternatively, we also consider parent-pairs with maximum posterior probability at each site, by running the Forward/Backward algorithms (Section 2.4.1).

We also considered the graph approximation of the HMM (Section 2.4.2) and HMMsub (Section 2.4.3). The graph with its subsampling scheme heuristically approximates our HMM with compatibility emissions. It is directed in that whether we start choosing pseudo-parents from left-to-right or vice versa will influence the final choice of pseudo-parents. HMMsub has only been used for the implementation of the Viterbi algorithm with CE.

To assess our performance we must first obtain a threshold for optimal performance. Our methods are genotype-based but it is haplotypes that carry the most information on the (dis-)similarity of individuals, their genealogical relationship and the parts of their genomes that they share. The baseline stochastic method that uses haplotypes to address the issues discussed here is the Li and Stephens (LS) algorithm (Li and Stephens [2003]), which was introduced in Section 1.7.1. The LS algorithm is the basis of ChromoPainter (Lawson et al. [2012]) which has been successfully used to characterize ancestry and admixture (Lawson et al. [2012]) and in which the concept of NNs was introduced.

For a given sample of genotypes, the approaches that we consider and assess in this chapter are:

1. The Viterbi implementation of our HMM using each of the three emission matrices
2. The F/B implementation of our HMM with compatibility emissions
3. The graph method
4. HMMsub
5. The LS method (Viterbi and F/B) on the original, true haplotypes of the sample

4.3 A statistical method for detecting NNs and extended haplotype sharing from unphased genotype data

6. the LS method (Viterbi and F/B) on haplotypes obtained after phasing genotypes (obtained from the true haplotypes) using IMPUTEv2 (Howie et al. [2009]) and BEAGLE (Browning and Yu [2009])
7. All of the above on data that contain genotyping error.

4.3.1 Measuring accuracy

For each genotype in the sample, our methods (genotype-HMM, the graph and HMMsub) output a single path of genotype pairs at each site, while the LS approach runs on haplotypes and outputs a sequence of optimal haplotype states. To make the results comparable, we run the LS methods on each haplotype but record the genotype individual that the selected haplotype corresponds to.

For a sample of genotypes of size n , for each $i \in \{1, \dots, n\}$ and at each site $j \in \{1, \dots, L\}$, we will be using the following notation:

- G_{ij} : the genotype of individual i at site j
- H_{ij}^1 and H_{ij}^2 : the two haplotypes of G_{ij}
- NN_{ij} : the NN of some haplotype H_{ij} ($NN_{ij} = H_{kj}, k \in \{1, \dots, n\}/\{i\}$)
- $GN_{ij}(= G_{kj})$: the genotype containing NN_{ij}

Therefore, individual G_{ij} has genotype NNs GN_{ij}^1 and GN_{ij}^2 . At every site j , for each genotype G_{ij} , each method selects two parental genotypes, denoted by:

- $P_{ij}^1 (= G_{kj}, k \in \{1, \dots, n\}/i)$; this is the first parent at site j
- $P_{ij}^2 (= G_{mj}, m \in \{1, \dots, n\}/\{i, k\})$; this is the second parent at site j .

For our assessment we would like to check whether $\{P_{ij}^1, P_{ij}^2\}$ contains $\{GN_{ij}^1, GN_{ij}^2\}$. We therefore define the weighted sensitivity, WS_n , as the ability of the method to identify pseudo-parents containing correct NNs:

$$WS_n = \frac{0.5}{L} \times \sum_{j=0}^{j=L} \max\{\mathbf{1}_{\{P_{ij}^1 \in GN_{ij}^1\}} + \mathbf{1}_{\{P_{ij}^2 \in GN_{ij}^2\}}, [\mathbf{1}_{\{P_{ij}^2 \in GN_{ij}^1\}} + \mathbf{1}_{\{P_{ij}^1 \in GN_{ij}^2\}}]\} \quad (4.1)$$

4.3 A statistical method for detecting NNs and extended haplotype sharing from unphased genotype data

We would also like to assess the distance between the selected pseudo-parents and the query genotype, especially for the cases when they are not NNs. We therefore additionally measure:

1. The number of nodes (NMRCA) between the selected parents and the target individual, and
2. The TMRCA between them.

Between two true NNs, the number of nodes is 1 and the expected TMRCA is $\frac{2}{n}$ (Blum and François [2005]). The variance of this TMRCA is, however, infinite (Janson and Kersting [2011]; we discuss this further in Chapter 5, Section 5.2.2).

Finally, we estimated the shared chunk size between the query genotype and its surrogate parents. For each parent separately, breakpoints occur when a parent changes and a shared chunk is defined as the region between two such breaks (see Figures 2.5 and 2.7). The query individual shares a haplotype chunk with each of its pseudo-parents. We record the median such length per parent, and then the mean of these two median values to obtain average haplotype sharing per person. It is necessary to obtain medians first as the distribution of chunk sizes is such that variance is infinite (following from the infinite variance of the external branch lengths of coalescent trees). Similarly, for all haplotype chunks shared across the entire sample, we obtain the median and the N50. We illustrate the difference between these metrics in Figure 4.7. The N50 is the median weighted by physical length: the size of chunk that 50% of sites in a sequence would fall in. The N50 is equivalent to the median chunk size obtained from our marginal tree simulations, where we generated chunks around a given position. In this way we can compare our findings from each method from the simulated data here, with the expected values obtained from the marginal tree simulations. Note that we expect some discordance due to the effect of mutation. Mutation was ignored in the tree simulations but should now influence results in two ways:

1. Reduce the levels of haplotype sharing;
2. Increase power to distinguish NNs from cousins and more distant relatives in the genealogy.

4.3 A statistical method for detecting NNs and extended haplotype sharing from unphased genotype data

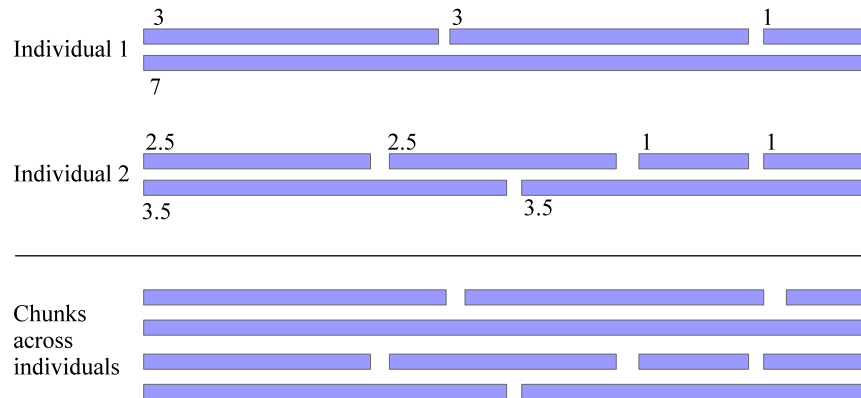


Figure 4.7 Measuring average sharing The diagram is a toy illustration of how we measure average sharing in a sample. The top panel shows the shared haplotype chunks for two individuals. There are two series of shared chunks for each person, representing the shared haplotypes of the query individuals with each of its NNs. On the left of each chunk is its physical length (in Mb, say). For individual 1, the median chunk shared with each pseudo-parent is 3 and 7 and therefore overall sharing for individual 1 is the mean of the two: 5. Similarly, overall sharing for individual 2 is 2.625. Across the sample therefore, the mean per person sharing is 3.8125. An alternative approach to calculating sharing across the sample is to collect all shared chunks, as illustrated in the bottom panel and calculating the overall median across them. Here, this is 2.75. The N50 is the weighted median, with the weights being the physical length of the chunks. Across all chunks from the sample the N50 is 3.25. In contrast, the per person N50 is 4 (the N50 is calculated for each pseudo-parent as above, then averaged across parents and samples).

4.3.2 Simulated Data

To assess the accuracy of the algorithm and to compare it to alternative strategies for identifying NNs from unphased genotype data we performed coalescent simulations across a variety of sample sizes and recombination rates. Our assessment requires full knowledge of the genealogy of a sample and our analysis here is therefore restricted to just simulated data. We therefore generated haplotypes using the MaCS software (Chen et al. [2009]), assuming the standard neutral model, with constant-population size. We generated 100 different data sets each consisting of 200 haplotypes (100 genotypes) spanning 1Mb. We used a pop-

4.3 A statistical method for detecting NNs and extended haplotype sharing from unphased genotype data

ulation scaled mutation rate of 0.01 with no recombination and recombination rates of 0.25cM/Mb, 2.5cM/Mb and 25cM/Mb. For the two latter cases we also modified the data by applying a 1% genotyping error. Overall, we considered the following data sets:

1. 100 data sets with $\rho = 0$
2. 100 data sets with $\rho = 0.25\text{cM/Mb}$
3. 100 data sets with $\rho = 2.5\text{cM/Mb}$
4. 100 data sets with $\rho = 2.5\text{cM/Mb}$ and 1% genotyping error
5. 100 data sets with $\rho = 25\text{cM/Mb}$
6. 100 data sets with $\rho = 25\text{cM/Mb}$ and 1% genotyping error

In all, we therefore obtained a collection of haplotype data sets that reflect a range of recombination and genotyping error rates relevant to human populations. Singletons are excluded from all our analyses. Just as with our phasing study, we obtained genotype data by combining each pair of haplotypes. For each individual in turn we may then run our analyses using the remaining 99 individuals from the sample. We record:

1. WS_n , NMRCA and TMRCA, taking the mean over sites for each metric.
2. The median chunk length and N50 across our samples. This may be measured either per person or the median across all chunks collected from all individuals (refer to Figure 4.7). We do not expect both measures to give the same results due to the skewed distribution of shared haplotype chunks: smaller chunks are far more numerous and will pull down the median to a greater extent when the entire sample is considered collectively. For a single individual, longer shared haplotype chunks with a pseudo-parent dominate more, and therefore the per person values are expected to be larger.
3. The proportion of times that (any of) the true NNs had maximum posterior probability, by using the F/B version of our HMM with CE and of the LS method. We report the mean over sites.

4.3 A statistical method for detecting NNs and extended haplotype sharing from unphased genotype data

4. The proportion of the posterior carried by the true NNs; we used the maximum posterior achieved by a NN when there were more than one NNs at a SNP. We report the mean over sites.

On the data sets containing genotyping error, we use the different methods with their standard parameters in the same way that we use them for the other data sets. Error is expected to compromise performance and changing parameters like the error rate might alleviate some of this deterioration. However, with real data error rates are typically not known, at least not with high accuracy and it is therefore of interest to explore how the methods cope in the absence of this knowledge.

For each of these data sets our graph method was used without subsampling. However, in order to investigate whether subsampling would have had any effect, for the data sets with $\rho = 2.5\text{cM/Mb}$ and no error we ran the graph method for different values of K (number of individuals subsampled before the construction of each genotype tree; refer to Figure 2.5 for an outline on how the method works).

Finally, we considered HMMsub aiming to consider about 20 individuals at each site. We do not investigate the effect of subsampling a different number of individuals here (we expect this to have a negligible effect). Instead, we consider two alternative scenarios: subsampling around 10 (H10) and around 300 (H300) randomly selected sites (refer to Figure 2.6). Our aim is to establish whether this choice might influence HMMsub’s ability to approximate the full HMM.

4.3.3 Results

With 200 haplotypes and assuming a constant effective population size of 10,000, from our tree simulations we estimate the expected per person N50 haplotype size to be at most 1.15cM (meaning that 50% of sites in the genome lie in extended haplotypes of at least 1.15cM). This is an upper limit as our tree simulations ignore the effect of mutation (which will disrupt haplotype sharing). In the absence of genotyping error, the genotype HMM with compatibility emissions, the graph-based approximation and HMMsub recover haplotype sharing lengths closest to this expectation across different recombination rates (Figures 4.8A and

4.3 A statistical method for detecting NNs and extended haplotype sharing from unphased genotype data

4.9). These genotype-based methods detect haplotype sharing at around $\sim 0.8cM$. Note that the performance of the graph method is not affected by our subsampling strategy (Table 4.1). This justifies both approaches and suggests that for K as low as 20 we can expect to detect exactly the same levels of coancestry and haplotype sharing as when there is no subsampling. There is also no difference in performance between H10 and H300 (subsampling around 10 and 300 SNPs respectively for HMMsub) at least for the constant population size scenario investigated here. The results also highlight that the Mendelian and weighted Mendelian emissions are unsuitable for detecting shared haplotype chunks, as they consistently significantly underestimate the extent of sharing and are unable to recover extended haplotype chunks. A full summary of the simulation results is presented in Table 4.2.

Interestingly, this is also true of the LS approach applied to the original haplotypes, though the effect is very small. In contrast, the pre-phasing approaches also consistently underestimate the extent of haplotype sharing, with the effect becoming more prominent as recombination rate increases (Figure 4.9 B and C). The presence of even one phasing error will prevent the methods from detecting fully shared haplotypes across the region that we simulated, while our HMM with CE, graph, HMMsub (both versions) and the LS method on the true haplotypes manage to recover these extended chunks more effectively.

The genotype-based models that use compatibility (and therefore a form of IBS) are designed to detect maximal haplotype chunk sizes. We would therefore expect them to potentially choose an individual's cousin, say, as a pseudo-parent rather than the actual NN, if the TMRCA with the cousin is close to the TMRCA with the true NN. This might compromise sensitivity in NN detection but will result in the recovery of the longer shared haplotype chunks. In fact, our tree simulations confirmed that this compromise is necessary as maximal sharing not with a NN occurs 20% of the time (in the absence of mutation, in a constant-size scenario, which is the case here). Conversely, the LS model may transition more liberally between NNs when there are many, which is often the case. It may therefore maintain a higher accuracy for NN detection. The frequent changing among NN sets however underestimates haplotype sharing and as a result it also underestimates the relatedness levels in the sample. The compatibility driven

4.3 A statistical method for detecting NNs and extended haplotype sharing from unphased genotype data

genotype-based methods, will not switch between NNs within a set, or across overlapping sets, unless a substantial amount of incompatibilities are accumulated. In fact, for the 2.5cM/Mb case with no errors the full HMM with CE picks a single pair of parents across the whole region for $\sim 12\%$ of individuals tested, while this never happens with LS on the original haplotypes.

With 1% genotyping error, the ability of all methods to detect extended haplotype sharing is reduced (Figure 4.9D), though the proportional reduction is greater at lower recombination rates (Figures 4.8A, 4.9D). At lower rates the physical length of shared haplotypes is longer, hence genotype errors are likely to be more disrupting. In contrast, at higher rates, recombination dominates as the driver of transitions, weakening the effect of genotyping error. Nevertheless, the genotype-HMM with CE, HMMsub (both versions) and the graph-approximation consistently provide the most accurate estimates of extended haplotype sharing compared to the haplotype-based approaches, though our graph method suffers substantially more than the HMM approach. Only our genotype-HMM with CE and HMMsub (both versions) manages to capture some of the longer haplotypes (Figure 4.9D). The LS method on true haplotypes exhibits some deterioration though this case highlights the effect of phasing errors even more: the results from phased haplotypes fair worse than the HMM using MT and WMT emissions. The latter detect even a few extended haplotypes which are entirely missed by the LS method on phased haplotypes. We should note however that overall the LS method using haplotypes obtained from IMPUTE2 consistently achieves measurements closest to those from the true haplotypes, illustrating the high phasing accuracy of the software.

Finally for the case of 0.25cM/Mb our expected value from our tree simulations does not apply. Here the effect of mutation in distinguishing extended sharing versus recombination may be much more critical. This effect is being ignored by our tree simulations, while, in contrast, our approaches (both genotype- and haplotype-based) do incorporate a mutation/error model. What is more, the total length of the simulated region is only 0.25cM and this is therefore an upper bound on the detectable lengths.

Turning to the detection of genealogical NNs (Figure 4.8B, Table 4.2), we find that the genotype-HMM performs best among approaches, reaching the maximal

4.3 A statistical method for detecting NNs and extended haplotype sharing from unphased genotype data

ρ	K	Gn.Dist.	Gn.Dist.N50	Ph.Dist.N50	WS_n	NMRCA	TMRCA
2.5	20	0.398	1.418	442.788	0.661	1.675	0.032
	40	0.398	1.458	451.293	0.658	1.687	0.016
	60	0.397	1.453	453.070	0.658	1.685	0.032
	80	0.397	1.444	453.070	0.659	1.685	0.032
	100	0.398	1.423	451.130	0.660	1.679	0.032

Table 4.1 **The graph method with subsampling.** Initial sample size was 100 individuals. ρ = recombination rate (cM/Mb); K = number of individuals subsampled. Gn.Dist = genetic distance, measured in cM; Ph.Dist = physical haplotype chunk size, measured in kb. The values shown are to 3 s.f.

value of 0.80 sensitivity (where the set of pairs with maximal posterior probability are considered from the F/B approach) in almost all cases. The success rate is even greater than 80% for very low recombination rates, where low-frequency variants give further resolution to the genealogy as mutation becomes a more prominent force compared to recombination.

Where a single best-pair is sampled (as is the case with our implementation of the Viterbi algorithm; Figure 4.8B), the performance of all methods decreases, reflecting the sometimes considerable uncertainty over which haplotype has the greatest extent of identity whenever there are multiple NNs at a site. We see that HMMsub (both versions) perform at least as well as the full HMM, and interestingly, even better for the 25cM/Mb scenario. Here, the graph-approximation performs less well, though comparable to pre-phasing by BEAGLE. Pre-phasing by IMPUTEv2 performs considerably better, typically close to the accuracy on original haplotypes, again reflecting its high quality phasing. There are several key observations to make here:

1. The advantage of our compatibility and genotype-based approach over the haplotype-based LS is less clear
2. There is still a strong effect from phasing errors, with NN miss-assignment occurring up to 30% more frequently (Figure 4.8B)
3. In the realistic scenario of 2.5cM/Mb, our HMM with CE and HMMsub (both versions) are distinctly most robust to errors. If we consider posteri-

4.3 A statistical method for detecting NNs and extended haplotype sharing from unphased genotype data

ors, there is in fact no effect on the HMM’s ability to detect NNs even in the higher recombination rate scenario.

4. There is no difference between the two versions of HMMsub.

Finally, apart from the high recombination rate with error scenario, we find that all approaches perform well in terms of choosing NNs that are genealogically close, if not true NNs (Figure 4.8 C and D). We further see that when NNs are not chosen, they will typically not be more distant than a genealogical cousin or second cousin. While this is representative of the expected behaviour when a single NN is to be sampled per site, note that these measurements of the TMRCA and NMRCA do not reflect their distribution of times were we to consider the full spectrum of posterior probabilities from the F/B algorithm. Since we observe an improvement in performance, for all methods, using the F/B algorithms we would expect that the TMRCA and NMRCA to achieve better results over all the individuals with optimal or near-optimal posteriors.

Indeed, Figure 4.10 which plots the TMRCA of the NN against the TMRCA of the selected pseudo-parents by each method, shows that our HMM (full and both versions of HMMsub) with compatibility finds pseudo-parents that are generally almost as recent as those detected by the LS on true haplotypes. Our graph method is less competitive in that respect, though still faring well compared to the results from phased haplotypes. In fact, our results show that when selecting individuals with a TMRCA that is much older than the TMRCA with a NN (Figures 4.10 and 4.11), this is due to the effect of phasing or genotyping errors. Moreover, Figure 4.11 confirms our compatibility-based HMM’s (full and HMMsub) robustness to errors. Our HMM (full and HMMsub) consistently stays close to the $y = x$ line, which further suggests that though our method may more often prefer cousins, this happens when the coancestry shared with them is close to that shared with the NNs; this enables our method to detect maximal haplotype sharing.

The results confirm that Mendelian-based emission schemes are unsuitable for this application. The shared haplotype chunks obtained using WMT and MT are underestimated. Nevertheless, there are still cases when even this scheme is better than using LS on haplotypes obtained from phasing with BEAGLE and

4.3 A statistical method for detecting NNs and extended haplotype sharing from unphased genotype data

our own phaser. This is not to say that MT and WMT do not detect some recent ancestry: the pseudo-parents from these methods are typically not more distant than third cousins. The caveat is that these ancestors are consistently more distant relatives than the ones detected by the other methods, even the LS on haplotypes obtained by phasing. This is perhaps unsurprising in that we have discovered (see Chapter 3, Section 3.3.4) that for a homogeneous population phasing can be effective using a very wide range of suboptimal paths and therefore very recent ancestry is not a requirement.

4.3 A statistical method for detecting NNs and extended haplotype sharing from unphased genotype data

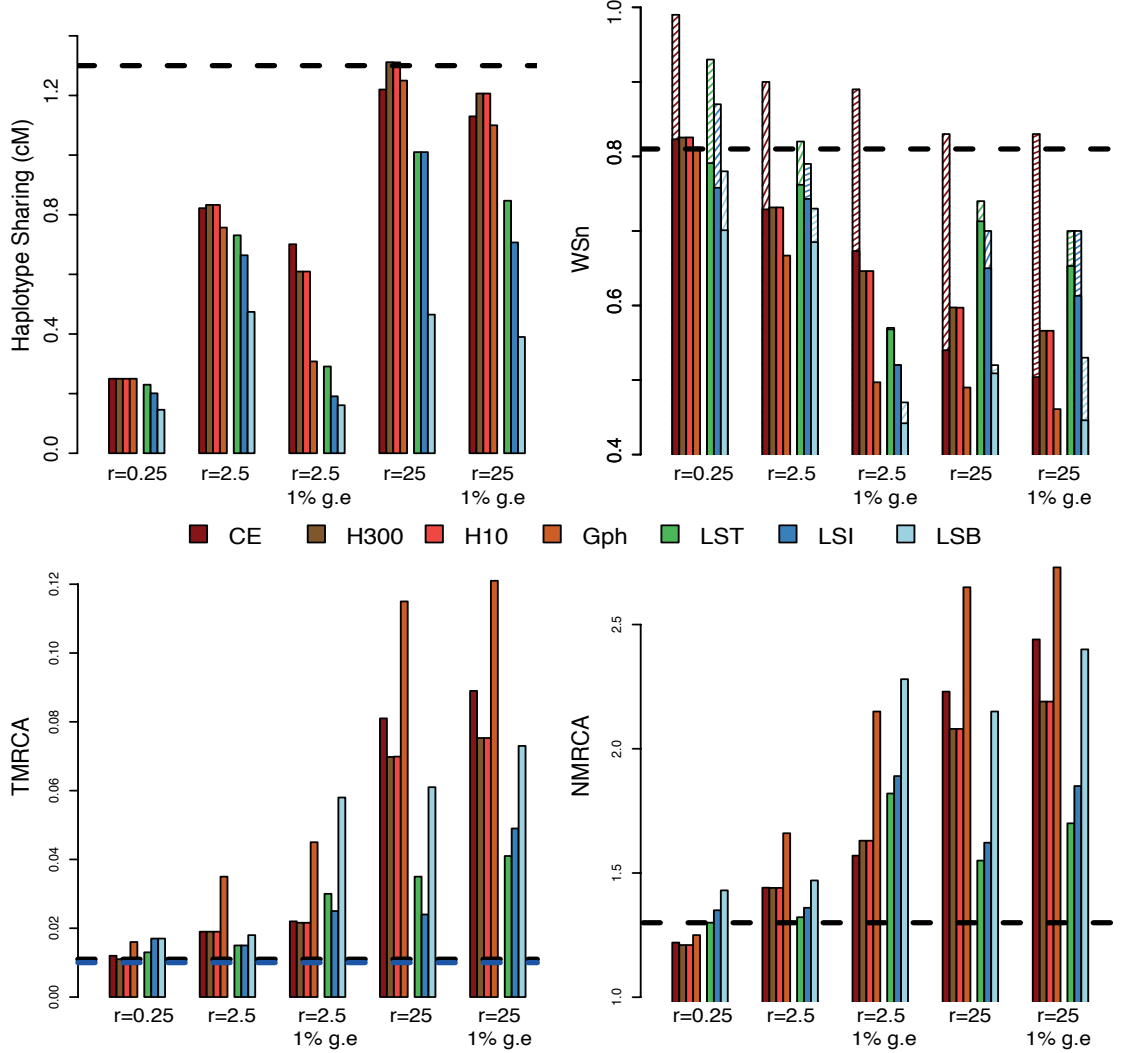


Figure 4.8 **Haplotype sharing and finding NNs.** We consider the genotype-HMM with compatibility emissions (CE), the graph method (Gph), HMMsub around 10 SNPs (H10) and around 300 SNPs (H300), and the Li and Stephens algorithm (Li and Stephens [2003]) applied on true haplotypes (LST) and haplotypes obtained after phasing with BEAGLE (LSB) or IMPUTE2 (LSI). A) N50 haplotype chunk sizes averaged across individual. These averages are expected to be slightly inflated compared to the tree simulations; refer to text for more details. B) Solid bar is WS_n as defined in Equation 4.1 (using the Viterbi optimal path), informally it measure how often the Viterbi path includes either both NNs or a single NN are found; striped bars are WS_n all candidate pairs with optimal posterior (from Forward/Backward algorithms). C) TMRCA with detected individuals. D) NMRCA with detected individuals. g.e. = genotyping error, r= recombination rate (cM/Mb). The black, dotted, horizontal lines are the values obtained from the tree simulations described in Section 4.2. The results shown here are from Table 4.2.

4.3 A statistical method for detecting NNs and extended haplotype sharing from unphased genotype data

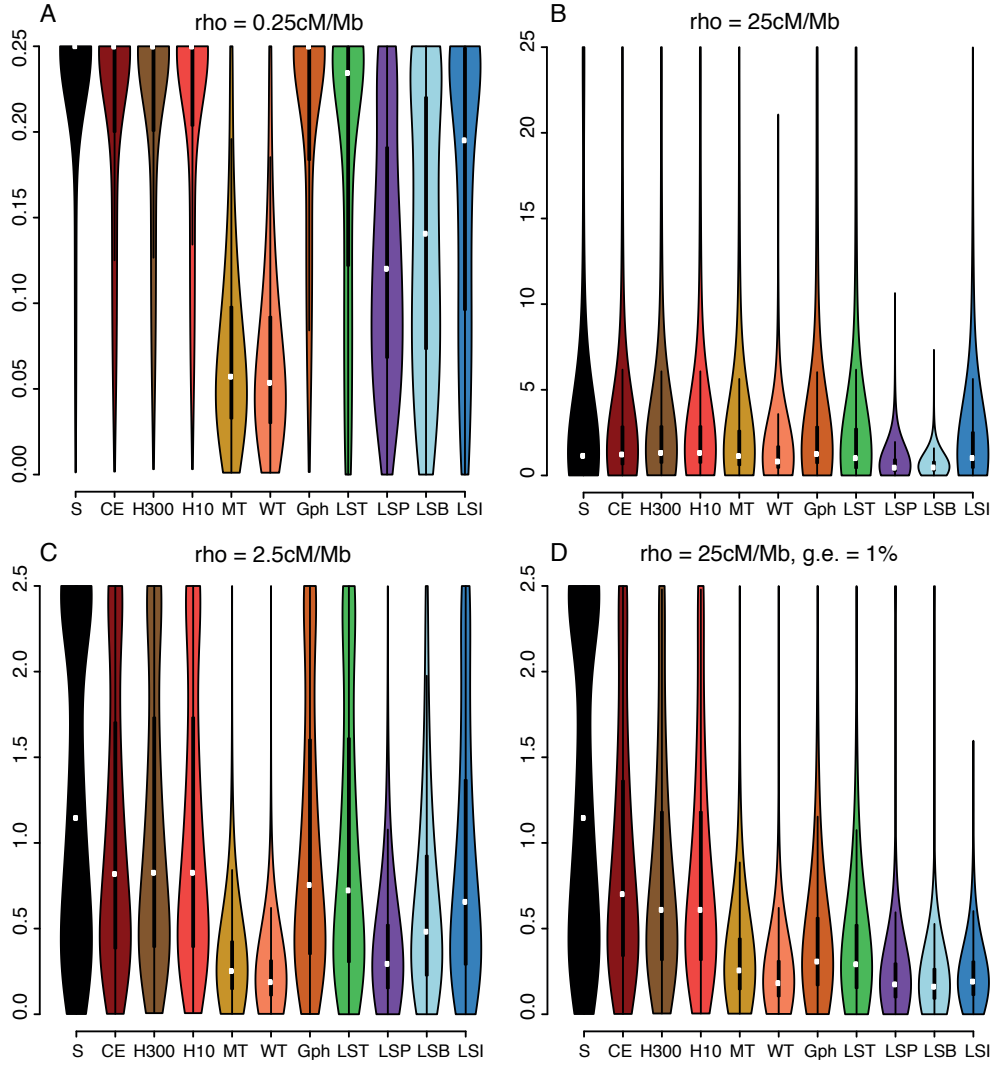


Figure 4.9 Detected shared haplotype chunks. 100 points were sampled at random from each of the simulated data sets and the length of the chunk shared with each pseudo-parent around each such point was recorded. Each panel shows the distribution of shared haplotype chunks obtained from each method. S = the theoretical results obtained from the tree simulations described in Section 4.2; CE = the HMM with compatibility emissions; H10 = HMMsub with subsampling around 10 SNPs; H300 = HMMsub with subsampling around 300 SNPs; MT = the HMM with Mendelian transmission; WMT = the HMM with weighted Mendelian transmission; Gph = the graph method; LST = the LS method on true haplotypes; LSP, LSB, LSI = the LS method on haplotypes estimated using PACE, BEAGLE and IMPUTE2 respectively; g.e. = genotyping error. Note that we do not expect the results from our tree simulations to match the results from the methods both because the tree simulations ignore mutation.

4.3 A statistical method for detecting NNs and extended haplotype sharing from unphased genotype data

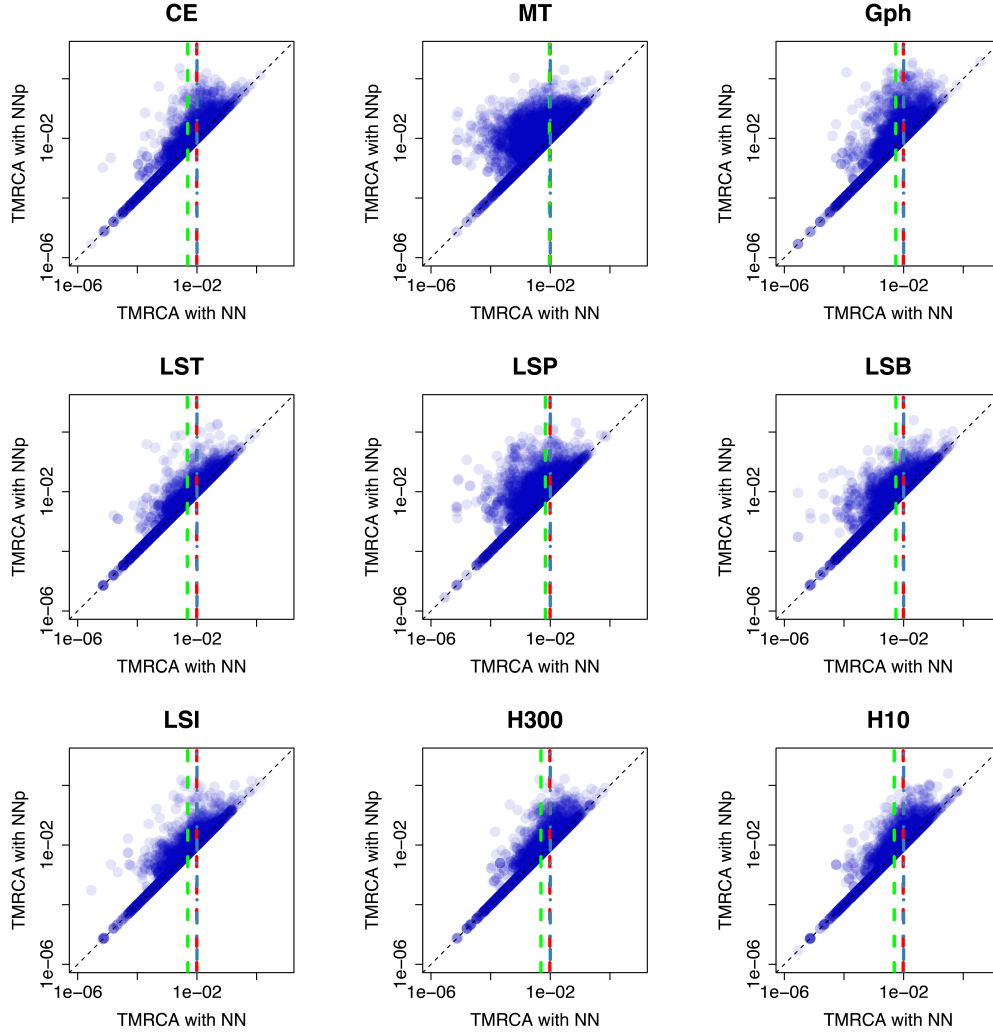


Figure 4.10 **TMRCA with a NN vs NN_p for $\rho = 2.5\text{cM/Mb}$.** 100 points were sampled at random from each of the simulated data sets and the TMRCA with the pseudo-parents (NN_p ; the mean TMRCA between both pseudo-parents) and the true NN were recorded. When there were more than one NNs, the minimum TMRCA was considered. Each main figure shows the TMRCA with an NN against the TMRCA with the NN_p for every haplotype from our simulations. The blue dashed line is the expected mean for the TMRCA with a NN under the standard coalescent with a constant population size. The red dashed line is the observed mean. The results shown in each panel are after sampling 10,000 out of all the observations made. CE = the HMM with compatibility emissions; Gph = the graph method; LST = the LS method on true haplotypes; LSP, LSB, LSI = the LS method on haplotypes estimated using PACE, BEAGLE and IMPUTE2 respectively; H10, H300 = HMMsub subsampling around 10 and 300 SNPs respectively.

4.3 A statistical method for detecting NNs and extended haplotype sharing from unphased genotype data

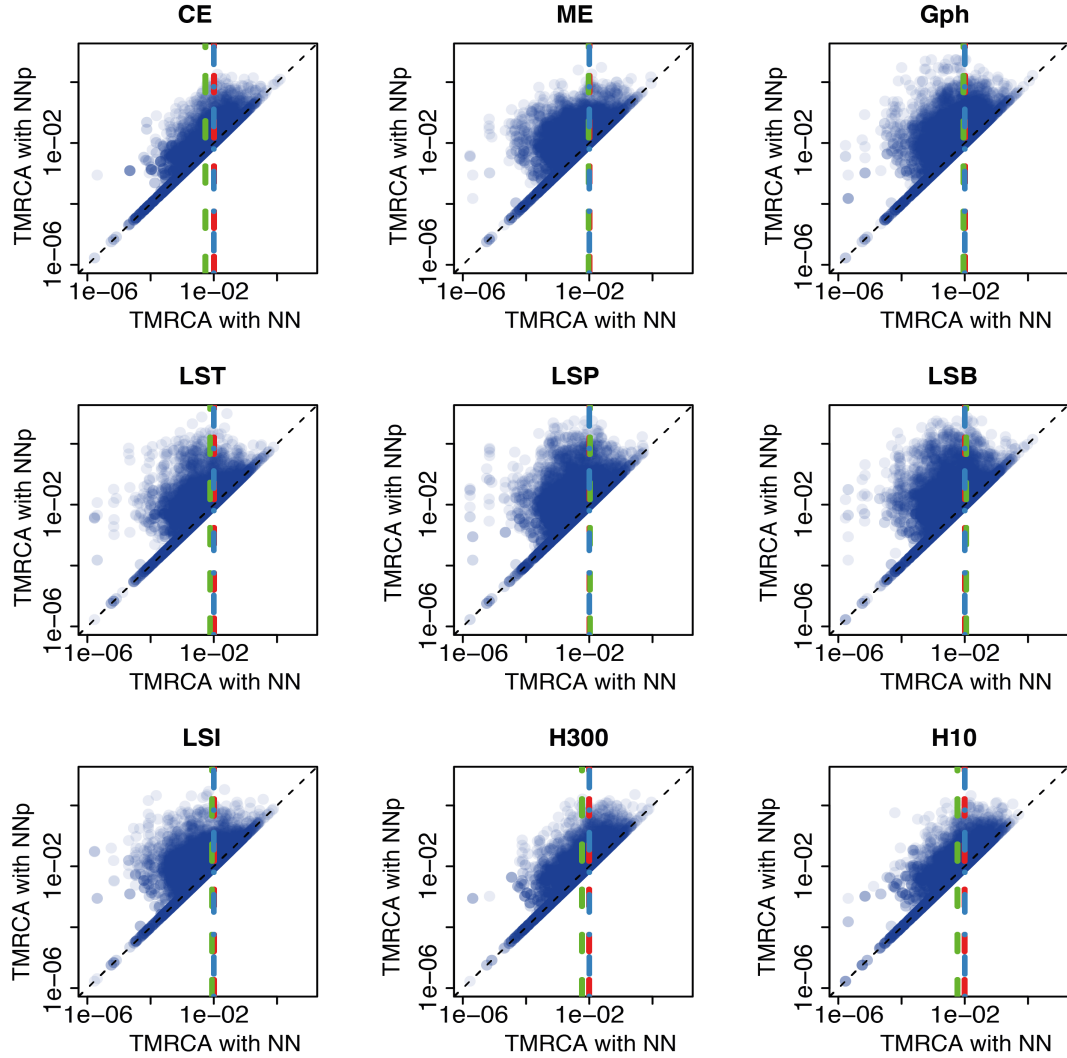


Figure 4.11 **TMRCA with a NN vs NN_p for $\rho = 2.5\text{cM/Mb}$ and 1% genotyping error.** 100 points were sampled at random from each of the simulated data sets and the TMRCA with the pseudo-parents (NN_p ; the mean TMRCA between both pseudo-parents) and the true NN were recorded. When there were more than one NNs, the minimum TMRCA was considered. Each main figure shows the TMRCA with an NN against the TMRCA with the NN_p for every haplotype from our simulations. The blue dashed line is the expected mean for the TMRCA with a NN under the standard coalescent with a constant population size. The red dashed line is the observed mean. The results shown in each panel are after sampling 10,000 out of all the observations made. CE = the HMM with compatibility emissions; Gph = the graph method; LST = the LS method on true haplotypes; LSP, LSB, LSI = the LS method on haplotypes estimated using PACE, BEAGLE and IMPUTE2 respectively; H10, H300 = HMMsub subsampling around 10 and 300 SNPs respectively.

4.3 A statistical method for detecting NNs and extended haplotype sharing from unphased genotype data

ρ	Method	GD	GD (N50)	PD (N50)	WS_n	P	$\frac{P_{NN}}{P_{opt}}$	NMRCA	TMRCA
0	CT	-	-	1000	0.894	1.000	1.00	1.11	0.010
	MT	-	-	1000	0.369	0.406	0.83	2.22	0.027
	WT	-	-	1000	0.359	0.375	0.77	2.27	0.028
	Graph	-	-	1000	0.894	-	-	1.11	0.010
	LST	-	-	999	0.710	1.000	1.00	1.35	0.013
	LSP	-	-	986	0.570	0.580	1.00	1.66	0.017
	LSB	-	-	971	0.65	0.837	1.00	1.47	0.014
	LSI	-	-	1000	0.700	1.00	1.00	1.38	0.013
	H300	-	-	1000	0.892	-	-	1.12	0.010
	H10	-	-	1000	0.892	-	-	1.12	0.010
0.25cM/Mb	CT	0.188	0.250	999	0.823	0.993	0.997	1.21	0.012
	MT	0.035	0.056	223	0.354	0.349	0.637	2.20	0.029
	WT	0.032	0.056	210	0.342	0.333	0.620	2.35	0.030
	Graph	0.188	0.250	999	0.812	-	-	1.25	0.013
	LST	0.068	0.234	937	0.791	0.933	0.980	1.30	0.016
	LSP	0.046	0.117	468	0.626	0.685	0.752	1.57	0.018
	LSB	0.046	0.146	584	0.701	0.785	0.848	1.43	0.017
	LSI	0.058	0.202	806	0.758	0.870	0.920	1.35	0.017
	H300	0.188	0.250	999	0.825	-	-	1.21	0.011
	H10	0.188	0.250	999	0.826	-	-	1.21	0.011
2.5cM/Mb	CT	0.486	0.822	329	0.729	0.900	0.829	1.441	0.019
	MT	0.147	0.253	191	0.483	0.492	0.790	2.02	0.026
	WT	0.109	0.187	75	0.433	0.446	0.749	2.21	0.031
	Graph	0.397	0.757	303	0.667	-	-	1.66	0.034
	LST	0.181	0.731	292	0.762	0.820	0.880	1.32	0.015
	LSP	0.103	0.292	117	0.580	0.620	0.700	1.76	0.026
	LSB	0.134	0.473	190	0.685	0.730	0.800	1.47	0.018
	LSI	0.159	0.664	266	0.744	0.790	0.800	1.35	0.015
	H300	0.159	0.833	333	0.732	-	-	1.44	0.019
	H10	0.159	0.833	333	0.732	-	-	1.44	0.019
2.5cM/Mb 1% g.e.	CT	0.372	0.701	280	0.673	0.900	0.820	1.57	0.021
	MT	0.143	0.256	102	0.485	0.523	0.828	2.06	0.028
	WT	0.103	0.117	72	0.426	0.469	0.774	2.28	0.034
	Graph	0.159	0.308	123	0.497	-	-	2.15	0.044
	LST	0.145	0.291	166	0.568	0.578	0.640	1.82	0.030
	LSP	0.081	0.174	70	0.442	0.473	0.540	2.21	0.053
	LSB	0.086	0.161	64	0.462	0.479	0.550	2.28	0.058
	LSI	0.100	0.191	76	0.520	0.529	0.600	1.89	0.025
	H300	0.345	0.610	244	0.646	-	-	1.63	0.022
	H10	0.345	0.610	244	0.646	-	-	1.63	0.022
25cM/Mb	CT	0.597	1.223	49	0.540	0.830	0.710	2.30	0.08
	MT	0.550	1.130	45	0.538	0.840	0.710	2.23	0.076
	WT	0.420	0.813	33	0.446	0.600	0.893	2.45	0.084
	Graph	0.685	1.253	50	0.490	-	-	2.64	0.113
	LST	0.302	1.007	40	0.713	0.740	0.800	1.55	0.025
	LSP	0.235	0.457	18	0.483	0.468	0.560	2.35	0.758
	LSB	0.271	0.475	19	0.509	0.527	0.640	2.15	0.613
	LSI	0.335	0.1014	41	0.367	0.700	0.770	3.68	0.238
	H300	0.695	1.312	53	0.597	-	-	2.08	0.070
	H10	0.695	1.311	52	0.597	-	-	2.08	0.070
25cM/Mb 1% g.e.	CT	0.559	1.148	46	0.505	0.830	0.700	2.44	0.089
	MT	0.530	1.072	43	0.503	0.663	0.925	2.37	0.083
	WT	0.409	0.774	31	0.444	0.576	0.877	2.60	0.094
	Graph	0.633	1.125	45	0.461	-	-	2.73	0.121
	LST	0.292	0.841	33	0.653	0.648	0.720	1.70	0.041
	LSP	0.214	0.372	15	0.424	0.415	0.490	2.61	0.088
	LSB	0.247	0.390	16	0.446	0.537	0.640	2.38	0.073
	LSI	0.280	0.707	28	0.613	0.696	0.770	1.85	0.049
	H300	0.670	1.207	48	0.566	-	-	2.19	0.075
	H10	0.670	1.207	48	0.566	-	-	2.19	0.075
Expectation		-	1.15	-	0.81	-	-	1.299	0.012

Table 4.2 **Full results on identifying NNs and shared haplotype chunks.** GD = genetic length of shared haplotype chunks (cM); GD (N50) = the N50 of genetic haplotype lengths (cM); PD = physical length of shared haplotype chunks (kb); WS_n is define in Eq. 4.1; P = the proportion of times the pair of genotype NNs had optimal posterior; P_{NN}/P_{Opt} = the maximum posterior of NNs over the optimal posterior; NMRCA = number of nodes to a common ancestor; TMRCA = time to a common ancestor; ρ = recombination rate (cM/Mb); CT = the HMM using compatibility emissions; MT = the HMM using Mendelian transmission; WT = the HMM using weighted Mendelian transmissions; graph = the graph method; H10, H300 = HMMsub with 10 and 300 SNPs for subsampling respectively; LST = Li and Stephens on the true haplotypes; LSP = Li and Stephens on haplotypes from Pace; LSB = Li and Stephens on haplotypes from Beagle; LSI = Li and Stephens on haplotypes from IMPUTEv2. All results except the chunk lengths are mean values over all sites; the chunk lengths are median values per person and then averaged over the sample. The expected values are from the tree simulations described in Section 4.2.

4.4 Discussion

In this chapter we have explored two different metrics for detecting and characterising local ancestry: NN status and maximal haplotype sharing. The main insight gained from this work is that they are not strongly correlated, especially when the demographic history of the sample under study involves exponential growth, which we know is the case for humans.

We have seen that haplotype-based methods may be more accurate in NN detection, but this is only true in the absence of genotyping and phasing errors. The latter significantly undermine the LS method’s ability to detect recent coancestry. Moreover, haplotype-based methods even without error systematically underestimate the extent of haplotype sharing, as is expected given that true NNs will not always contribute maximal shared haplotypes.

Haplotype sharing is used as the proxy for the TMRCA between individuals, with longer haplotypes suggesting a more recent TMRCA. Underestimating it will inevitably lead to an overestimation of the TMRCA, which in turn will

suggest that relationships in a sample are older than they really are. This has implications when studying demographic history such as in studies like the HELIC (www.helic.org; see Chapter 5) or POBI (www.peopleofthebritishisles.org), where inferring detailed ancestry and dating historical events is relevant.

Our genotype-based models will account for the expected amount of sharing in the sample, with the equivalent expected increase in NN-miss assignment. We have shown that the individuals we do identify are still very close relatives, with TMRCAs typically very close to those with the true NNs. Moreover, if we were to consider F/B results, the then full set of posterior distributions fully captures the most recent ancestry more effectively than the haplotype-based methods. Our HMM with CE consistently produces results similar to those obtained using the tree simulations. Moreover, HMMsub has near-identical (and in some cases better) performance to the full HMM, irrespectively of the number of SNPs chosen for subsampling. Though it is not guaranteed that this will be true under different demographic scenarios, we confirm that in this case subsampling within the HMM framework does not affect our effectiveness in characterizing coancestry. The graph method does not match the HMM's ability to find NNs, but does find cousins and recovers the expected maximal sharing (in the absence of error). It is therefore not obvious that these genotype-based NN-estimates are substantially suboptimal to the true NNs.

So, which is the better metric? This obviously depends on the application. Haplotype sharing is highly informative for demographic history. Alternatively, knowledge of NNs could be useful for inferring stratification. In our final two chapters we will continue with coancestry characterisation of two real data sets. We wish to make inference on their demography, to date the average TMRCA between any two NNs in the sample, to characterize isolatedness and to investigate migration. In Chapter 5 we explore the isolates and unrelated individuals from the HELIC project and in Chapter 6 the WTCCC data (Wellcome Trust Case Control Consortium et al. [2007]).

Chapter 5

Applications to real data: HELIC

Thus far in this thesis we have explored our method’s ability to detect recent ancestry in small homogeneous samples derived from the standard coalescent. We have discovered that we can cluster individuals according to a recent TMRCA, albeit not always most recent, while faithfully recovering maximally shared haplotypes. In this chapter we take advantage of these properties and use our graph method to explore relatedness in the context of two isolated populations that were genotyped for the HELIC project (HELlenic Isolated Cohorts; www.helic.org). We note that our HMM with subsampling approach (HMMsub) had not been developed at the time when the HELIC investigation was carried out.

Isolated populations arise following a founding event, such as a bottleneck, and are found in geographically and/or culturally isolated regions. In the absence of gene flow in and out of neighbouring populations, as well as their typical exposure to a homogeneous set of environmental variables like diet and lifestyle, isolated populations have unique characteristics that are interesting to study from a population genetics perspective. For example, the small effective populations size, that remains small over time, leads to increased levels of homozygosity and LD. In fact, isolated populations are characterised by reduced phenotypic, environmental and genetic heterogeneity (Peltonen et al. [2000], Heutink and Oostra [2002]) allowing for genetic drift to significantly reduce genetic variability. More

crucially, under such conditions, drift may lead variants that are otherwise rare to increase to higher frequencies.

Low-frequency and rare variants have been shown to play an important role in human phenotypes; we discussed this further in Chapter 1. If such variants have risen to higher frequencies in isolates due to drift, these isolates may then empower the detection of any associations that these variants may have with disease (Pollin et al. [2008], Stacey et al. [2011], Sulem et al. [2011], Zeggini [2011], Gudmundsson et al. [2012], Jonsson et al. [2012]). However, variants that are private to the isolate in question may have limited transferability to other populations and therefore the generalisability of novel associations identified in isolates is a subject of debate. Isolated populations may still though empower next generation association studies that focus on this rare variation (Zeggini [2011]).

The HELIC project entailed the collection of approximately 3,500 samples from two Greek population isolates. The first population includes 1267 individuals from the Pomak villages, which are a set of religiously and culturally isolated villages in North-east Greece. This isolate is referred to as Pomak in the study. The second population, which we call MANOLIS and which includes 1100 samples, is from the area of Anogia in Crete. The MANOLIS individuals are primarily geographically isolated from the wider Cretan population. In order to represent and compare against the general Greek population, the TEENAGE (TEENs of Attica: Genes & Environment) cohort was also considered. This sample was collected in a separate study (Ntalla et al. [2013]) and involved the genotyping of 857 adolescents (13-15 years) from the wider Athens area.

The purpose of the HELIC project was to fully characterise the isolated populations in terms of the extent of their isolatedness, changes in effective population size and F_{ST} and to inform on their demographic histories. The cohorts were also tested for enrichment of functional variants. Moreover, by additionally collecting anthropometric, cardiometabolic, biochemical, haematological and diet-related information for all samples it was possible to carry out medically-relevant association studies with complex traits, providing a proof of principle of the value of isolated populations in successfully detecting rare variants of high medical relevance.

The analyses were carried out in two main stages, resulting in two publications: Tachmazidou et al. [2013] and Panoutsopoulou et al. [in prep.]. Tachmazidou et al. focus on the MANOLIS cohort and identify an association between R19X (rs76353203) with increased HDL and decreased triglyceride levels (p-value $\approx 4.6 \times 10^{-9}$, explaining 2.9% of the trait variance). This association had previously been shown in the Amish founder populations (Pollin et al. [2008]) and was considered private to that population as its allele frequency in Europeans is 0.05%. In the MANOLIS individuals however it has reached a frequency of $\sim 2\%$ (being carried by $\sim 3.8\%$ of the population) due to drift, thus allowing for the association signal to be detectable. Using the UK10K data (www.uk10k.org), Tachmazidou et al. show that R19X has a single recent origin in Europe, pre-dating the establishment of the MANOLIS cohort and it must therefore have risen in frequency separately in the Amish and MANOLIS populations.

Panoutsopoulou et al. will report the genome-wide findings using the genotype data of both population isolates. This paper aims to characterise population structure and isolatedness, report on all association studies undertaken and medically relevant analyses performed and finally compare their allelic architecture with the TEENAGE cohort. Full details of all the analyses and results may be found in Tachmazidou et al. [2013] and Panoutsopoulou et al. [in prep.]. In this chapter we will discuss this author's contribution: we present the questions that we considered and the application of and results from our graph method that addressed them. Our work was also divided into two stages: an initial pass over Chromosome 11 for the investigation of *APOC3* and a subsequent genome-wide analysis.

5.1 Objectives

In the context of association with disease in the MANOLIS cohort we illustrate how local haplotype sharing around a putative locus can be informative on structure within a data set. Though this implementation of our approach does not directly propose a correction method for stratification that may be present, it demonstrates how such stratification may be detected even when the locus in

question is rare. We therefore highlight how such stratification may be captured and thus potentially corrected by further extensions of our method; we elaborate on this more extensively in the Discussion (Chapter 7). Identifying any local structure in this context is crucial as it may be a confounder in the association of R19X. We further attempt to use haplotype sharing around R19X in order to estimate a lower bound on the age of the variant. Equivalently to IBD mapping, we additionally explore whether there is excess shared coancestry among carriers of the R19X variant. Despite its effect size, even in the MANOLIS cohort R19X remains rare and it is therefore not obvious whether excess sharing will be detectable.

We generalise our approach in our work for the Panoutsopoulou et al. publication, considering both isolates and the TEENAGE cohort. By detecting recent coancestry we aim to characterise the extent of isolatedness for both the MANOLIS and the Pomak samples. We further attempt to date recent coancestry by obtaining estimates for the TMRCA between NNs. The distribution of haplotype sharing between them informs on any post-divergence gene flow that may have occurred. Using a heuristic approach, we estimate the TMRCA of NNs between each isolate and the TEENAGE cohort and attempt to make inference on the time of divergence.

In all, we explore our method’s ability to characterize stratification within and between samples and whether choice of NNs and haplotype sharing are meaningful measures for this characterization. Finally, understanding our method’s limitations, for instance in dating events or in dealing with genotyping error, will be critical in interpreting our findings in this study. We note that we will be using our graph method which is susceptible to genotyping error and therefore any time estimates may underestimate coancestry. Nevertheless, the stringent quality checks that were carried out, briefly described in Section 5.3 and more elaborately in Tachmazidou et al. [2013] and Panoutsopoulou et al. [in prep.], as well as the larger sample size here (compared to our simulations), promise rather accurate haplotype sharing and therefore time estimates.

5.2 Methods

We employed our graph method with subsampling. We set K , the number of individuals subsampled, to 20 and pruned our graph when either a double incompatibility or two single incompatibilities were accumulated. For more details on the graph method, pruning and subsampling refer back to Section 2.4.2. The method is run for each query genotype against a sample of candidate pseudo-parents (or NNs), identifying a single best pair of parents at each SNP. Since the true genealogical history of the three cohorts considered here is unknown, it is impossible to infer exactly how recent is the shared coancestry between the query genotype and the selected pseudo-parents. Instead, we infer haplotype sharing by identifying the genomic chunks over which each pseudo-parent does not change (haplotype sharing inference is illustrated in Figures 2.5 and 2.7). We explore the distribution of haplotype sharing along the genome for each cohort and how it varies between cohorts.

More specifically, for the *APOC3* study we analyse Chromosome 11 for the individuals in the MANOLIS cohort. The objective here is not the characterisation of haplotype sharing patterns generally but rather determining whether these patterns differ between carriers and non-carriers of the R19X variant. We record the choices of pseudo-parents and the length of haplotypes shared with them, while further noting the size of the specific haplotype chunks that carry R19X. We may thus infer:

1. The TMRCA between carrier NNs at R19X thus estimating a lower bound for the age of the variant. Our estimate is a lower bound as we are detecting the most recent ancestry between individuals and any alleles shared between them must therefore have pre-dated their TMRCA. Note that this is a very substantial underestimate.
2. The haplotype length around R19X, whether it is significantly different to sharing elsewhere on the chromosome, and whether it is moreover distinct between carriers and non-carriers.
3. Whether carriers typically pick other carriers as pseudo-parents more often than non-carriers at R19X and over the rest of the chromosome.

4. If there is any stratification within the sample at R19X; we investigate this by examining the decay of coancestry between carriers and between non-carriers away from R19X and compare against the expected decay. We may additionally compare against results at other SNPs on the same chromosome thus determining whether any stratification is private to R19X or whether it is a general feature of the MANOLIS sample. More details on the calculations carried out here are presented in Section 5.2.1.

Significance is assessed for the different cases using the Kolmogorov-Smirnoff test (Kolmogorov [1933], Smirnov [1948]). Any differences between carriers and non-carriers found using metrics 1 and 2 would suggest that our method may be useful in signal detection. We should note, however, that our method would only be powered to detect this if the coancestry in the region is recent. Though R19X is a rare variant, Tachmazidou et al. have shown that its age probably pre-dates the time of isolation from Europeans and therefore we do not in fact expect excess coancestry in the region. It follows that it is unlikely that we will be able to detect a significant difference in sharing between carriers and non-carriers unless there is genome-wide stratification between them. Points 3 and 4 demonstrates how our method may eliminate the possibility of any detected signal being due to the confounding effect of structure in the data.

We next consider the genome-wide analyses across all cohorts. We consider three separate cases:

1. Analysis A: each cohort is analysed separately such that for each individual the graph method may assess candidate pseudo-parents only from the host cohort. In this setup we may investigate coancestry within each cohort and answer questions like:
 - (a) What is the TMRCA between NNs? A recent TMRCA suggests high levels of relatedness, and the distribution of TMRCA's may be informative on the demographic history of the sample. For example, a strong enrichment for recent ancestry will be the signature of a bottleneck, whereas a more uniform distribution with older TMRCA's will indicate lack of isolatedness and a population that follows the standard neutral

coalescent; there may of course be alternative histories resulting in similar TMRCA distributions and we can only hope for a qualitative assessment of demographic history here, in collaboration with findings using alternative metrics.

- (b) What is the distribution of haplotype sharing and how does this inform on demographic history? The concerns here are similar to those for TMRCA, in terms of definitively identifying specific demographic events. Rather, we hope to identify excess sharing in the isolates as a qualitative signature of the founding event.
 - (c) By investigating patterns of haplotype sharing and whether there is clustering in NN choices, can we determine whether there is stratification present in the samples?
2. Analysis B: all cohorts are analysed together allowing for pseudo-parents to be selected from any of the three cohorts at each SNP. This study allows us to investigate shared coancestry across the cohorts, which further informs on the extent of isolatedness. We may measure:
- (a) How often NNs are selected from a different cohort to that of the query genotype and
 - (b) The length of the haplotypes shared with NNs from each cohort.

As above, different demographic histories may result in different combinations of observations for these two metrics. In the simple scenario of a founding event with no subsequent gene flow, we would expect a very low proportion of NNs to be outside the host cohort and any haplotype sharing with them to be limited. If there has been gene flow, the proportion of non-host NNs would not increase significantly but we would expect to see longer haplotypes shared with them. Alternatively, a high proportion of non-host pseudo-parents with which haplotype sharing has the same distribution as the host pseudo-parents would be evidence against isolation, while longer shared haplotypes might be indicative of a more complex demographic history. While there may be multiple possible more complicated demographies

consistent with different combinations of observations, the use of both NN choice in addition to haplotype sharing may help distinguish between some of them to a greater extent than an IBD approach, say.

3. Analysis C: we consider the MANOLIS and Pomak samples separately, allowing however pseudo-parents to be selected only from the TEENAGE cohort. The purpose of this study is:
 - (a) To obtain benchmark for distant haplotype sharing that is most likely shared prior to the founding event that gave rise to the isolates. This can be compared against the results from Analysis B in order to determine whether non-host NN choices in Analysis B indeed reflect older coancestry and not post-divergence gene flow events.
 - (b) To characterise the more distant coancestry that must exist between the isolates and the general Greek population; when pseudo-parents are not forcedly selected from the TEENAGE cohort, our method will never choose to select them since they are by construction more distantly related individuals with which haplotypes are not maximally shared.
 - (c) Estimate a bound on the time of divergence, using the TMRCA with the TEENAGE NNs. To carry out this analysis we will make assumptions on the demographic history of the isolates. We further note that use of our method for making inference on bottlenecks and isolation events has not been tested by simulation. Our strategy for this estimation is a heuristic approach, whose purpose is to illustrate potential applications of our model as well as to offer some initial insight on the demographic events that may have taken place. We revisit this part of our analysis and the results that we obtain at the very end of this Chapter (Section 5.6).

We note here that quantitative comparisons between the different analyses, such as the extent of haplotype sharing, is not straightforward as sample sizes differ, albeit not by very much. Nevertheless, it is still possible to compare the TMRCA

estimated under the different scenarios, as well as to make qualitative statements about isolatedness and structure.

The TMRCA between two NNs is a property of their relatedness that is site-specific, dependent on the topology of their marginal genealogy at that site. In order to fully characterise the distribution of these times it is therefore important to capture both any genome-wide variation and variation across individuals. We therefore sampled 100 randomly selected SNPs across the genome and calculated the TMRCA between each individual and their pseudo-parents at these SNPs thus obtaining a set of TMRCA observations for each Analysis A-C. The method for calculating the TMRCA at a locus is presented in section 5.2.2.

5.2.1 The decay of coancestry using carrier status

We are interested in examining the levels of coancestry among carriers of the R19X variant, compared to those of non-carriers. We would like to calculate the TMRCA between carrier NNs, thus obtaining a lower bound for the age of the variant, while further determining whether there is any local structure present in the sample. We would expect that individuals who carry a rare variant, like R19X, are more closely related locally than both elsewhere in the chromosome and with respect to non-carriers of that variant. As a result, at that locus carriers would typically identify carrier pseudo-parents, while moving away from the variant the rate of picking another carrier will decrease. More particularly, this rate of decrease away from the locus is captured by the distance to the left and to the right at which pseudo-parents change and therefore depends on the probability of recombination between R19X and each subsequent locus in either direction. In contrast, for non-carriers we expect no excess relatedness to carriers around R19X. The rate at which non-carriers therefore pick carriers as pseudo-parents should be uniform across the chromosome and equal to the background probability of picking any of the compatible candidate pseudo-parents (some of which will be carriers) by chance.

For each query genotype we record the number of carrier pseudo-parents ($\in \{0, 1, 2\}$) that it picks at each locus. We call this excess relatedness to carriers,

and for the carriers we call this R . Across the carrier individuals and the non-carrier individuals separately, we may thus obtain the mean excess relatedness to carriers across the chromosome. We may further obtain the expected values for R using standard coalescent theory.

For a carrier query haplotype, let R_{ij} be the decay of coancestry between sites i and j , assuming R19X is at j , say. R_{ij} is the probability that between positions i and j the query haplotype's NN remains a carrier which is the probability of there being 0 recombination events times the probability of there being a recombination event and selecting a carrier to transition to. Following the arguments of Li and Stephens (Li and Stephens [2003] and Section 1.7.1), it follows that:

$$R_{ij} = e^{-2 \cdot r_{ij} \cdot g} + \frac{c}{n}(1 - e^{-2 \cdot r_{ij} \cdot g}) \quad (5.1)$$

where r_{ij} is the genetic distance between sites i and j , c is the number of carriers, n is the sample size and g is the TMRCA in generations between the individual and its NN at j . At R19X and therefore at that position (where $i = j$) $R_{ij} = 1$. The first term in Equation 5.1 is the probability that there is no recombination event, using the fact that the number of events has a Poisson distribution with parameter $-2r_{ij}g$. The parameter follows from the fact that the number of recombination events that separate two individuals is twice the number of meioses since their TMRCA, i.e. twice the number of generations between them. We must also have that recombination occurs in the region between i and j which has probability proportional to the genetic distance between them. The second term in Equation 5.1 is the probability of having a recombination event, and then choosing uniformly over the carrier individuals present in the sample to transition to.

It is then possible to determine the value of g using the half life of the decay from our observations. We therefore solve:

$$\frac{1}{2} = e^{-2 \cdot r \cdot g} + \frac{c}{n}(1 - e^{-2 \cdot r \cdot g}) \quad (5.2)$$

with r being the genetic length of the half life. We thus obtain an estimate for the TMRCA between a query haplotype and its haplotype NN for the marginal genealogy at SNP j . In practice, we set R_{ij} to be the mean number over carriers

selected by the carriers in the sample divided by two, thus averaging over the two haplotype NNs at each SNP.

If there is no local structure for carriers, then we expect to see a smooth decay (on the genetic map) of coancestry away from R19X, following the relationship described by Equation 5.1. Conversely, any structure will be reflected by excess relatedness among carriers even further away from the R19X variant.

5.2.2 Another approach for estimating the effective population size and the TMRCA between NNs

Local coancestry around a locus for the purpose of estimating the marginal TMRCA does not need to be defined as excess relatedness with the carriers at that locus. A more general approach simply examines how this coancestry decays away from it, irrespectively of the status of the query individual and its genotype-NNs. Following the same arguments as above, for every individual under study we need only observe how far to the left and to the right we must go until the first pseudo-parent changes and then until the second pseudo-parent changes. This is near-equivalent to the approach above which counted the number of carriers selected at each SNP away from a locus: the distance at which this number changes informs of a recombination event which is the distance recorded here. Averaging across samples gives us a new observed decay of coancestry that may be used to date the average local TMRCA.

Our expectation of this decay follows the same distribution as above, with the caveat that any recombination event that changes the local pseudo-parent may now involve a transition to any individual in the sample and not a carrier specifically. The expected decay is therefore:

$$R_{ij} = e^{-2 \cdot r_{ij} \cdot g} + \frac{1}{n}(1 - e^{-2 \cdot r_{ij} \cdot g}) \quad (5.3)$$

using the same notation as in Equation 5.1. We follow the procedure described in Equation 5.2 to estimate the TMRCA using the half-life from our observations at a given locus.

By estimating the TMRCA between NNs we can derive a moment estimator for the effective population size (N_e) of the underlying population. If n is the

number of haplotypes in the sample, then the TMRCA between NNs (t say) is expected to be $\frac{2}{n}$. It follows that the number of generations to the TMRCA, T (which we can separately estimate using the aforementioned decay approach), is $t \cdot 2N_e$. Rearranging we obtain a moment estimator: $\hat{N}_e = \frac{T}{2t}$. This is of course a rather crude estimator, not least because we know (see Chapter 4) that the individuals that we will be identifying are not expected to always be true NNs but second or third cousins, or perhaps even more distant relatives if there has been some growth in the history of the population. In fact, this estimator assumes a constant population size and its accuracy has also not been tested via simulation. We expect to be slightly overestimating T and therefore N_e and seek only to make inference on the order of magnitude of its size.

5.2.2.1 Estimating the TMRCA from Simulated Data

To ensure that this newly proposed approach is a valid and reliable way for estimating the TMRCA between NNs in a sample we briefly return to our simulated data sets. We will consider the graph method, as this is the method that is used to analyse the real data, and the the full HMM with compatibility emissions (CE) and HMMsub (with 10 and 300 SNPs for subsampling) as a reference of optimal performance. We remind the reader that our simulated data sets consist of 100 individuals, simulated over 1Mb and for our purposes here we will consider the 2.5cM/Mb case without errors. For each of the data sets we sample 100 SNPs at random (excluding the first and last 500 SNPs to avoid edge effects).

At each SNP, we may infer the actual TMRCA between each query individual and its true NN and the actual TMRCA with the pseudo-parents that have been selected at that SNP. At each SNP we therefore have a distribution of real TMRCA values, which we know are slightly different for NNs and pseudo-parents (refer to Chapter 4, for example Figure 4.10). As previously mentioned, the expected branch length between true NNs is $\frac{2}{n}$ (Blum and François [2005]), however the variance is infinite (for example Janson and Kersting [2011]). As a result, we expect the distribution of TMRCAs between NNs (and pseudo-parents) to be highly skewed to the right (for example, as shown in Janson and Kersting [2011]). We additionally remind the reader that this is why the variance of

shared haplotype chunks is also infinite and we use the median and the N50 as summary statistics. In this case, we are interested in determining whether the point-estimates of the TMRCA using the decay approach described in Equation 5.3 capture the properties of this distribution well. We therefore consider:

- Where the point estimate falls in the full distribution of TMRCA (of true NNs and pseudo-parents) at each SNP, and
- Whether the point estimate is (almost) equal to the median of the true TMRCA.

Our results are illustrated in Figure 5.1. The top panel shows the full results at two of the randomly selected SNPs. We confirm the highly skewed distribution of the TMRCA, and further discover that while extremely old TMRCA are possible (though very rare; for example the top right figure where the actual TMRCA between true NNs at that SNP can go over 30,000 generations) in such cases the methods fail to recover these extremely short chunks. Encouragingly though, the point estimates of the TMRCA obtained using the decay approach (vertical dashed lines) fall just to the right of the spike of the distributions (of both NNs and pseudo-parents, which are themselves very similar) consistently across all of the sampled SNPs (results from other SNPs not shown). This suggests that our point estimator is accurately capturing the mass of this distribution.

Turning to the bottom panel of Figure 5.1, however, where we summarise the full distributions by their medians, we see that our point-estimator (from all methods) is poor at predicting the median TMRCA (both for true NNs and pseudo-parents) at any-one particular SNP. We therefore don't anticipate the estimator to produce accurate estimates at specific SNPs. The figures are in fact rather noisy and the bulk of the points suggest that the decay-based estimator will tend to be an overestimate compared to the median of the real TMRCA. These observations are a signature of the infinite variance property of the original distribution, suggesting that either the longer chunks have a smaller influence on the point-estimator or that the median of the true TMRCA is often too susceptible to the very long tail of the distribution resulting in them being inflated. These results further draw attention to the inherent difficulty in accurately summarising a distribution with infinite variance.

The stars in the figures mark the overall median values. Since true NNs are not always found, it is expected that there is a discrepancy between the TMRCA of NNs and the TMRCA with pseudo-parents (though, in agreement with our results from Chapter 4, this discrepancy is small). We can at best therefore hope that the ‘red’ star for NNp falls close to the $y = x$ line, which is indeed the case.

Overall, therefore, this experiment has shed light into the capabilities of this approach: while we do not expect to be able to make site-specific statements regarding the TMRCA between NNs, we are confident that the estimates will accurately capture the genome-wide values. In addition, we note the great similarity between the results from the graph method and those from the HMM, showing that using the graph method on the real data should recover the expected TMRCA values. We further note the similarities between the TMRCA values for true NNs and those of pseudo-parents, supporting our approach of assuming in our subsequent work in this chapter (and the next) that pseudo-parents are NNs. We finally end with a reminder of the fact that this simulation experiment was carried out using a constant population size model. Under a growth scenario, which is realistic for the real data explored in this thesis, we expect the difference between NNs and pseudo-parents to become smaller. Growth rescales the branch lengths of the genealogical tree of a sample, resulting in elongated terminal branches. Coalescence events start further back in time and the differences between genealogical siblings, cousins (etc) become less prominent, especially as the change in N_e increases or the growth period becomes larger. In this scenario, it is harder to distinguish between NNs and more distantly related individuals by means of haplotype sharing (refer to Chapter 4 for details). Under growth, we would also expect coancestry around a locus to decay more rapidly than in the constant-size case, indicative of the fact that we expect to see less sharing in the population (haplotype chunks would be disrupted for instance by the higher frequency of rare variants).

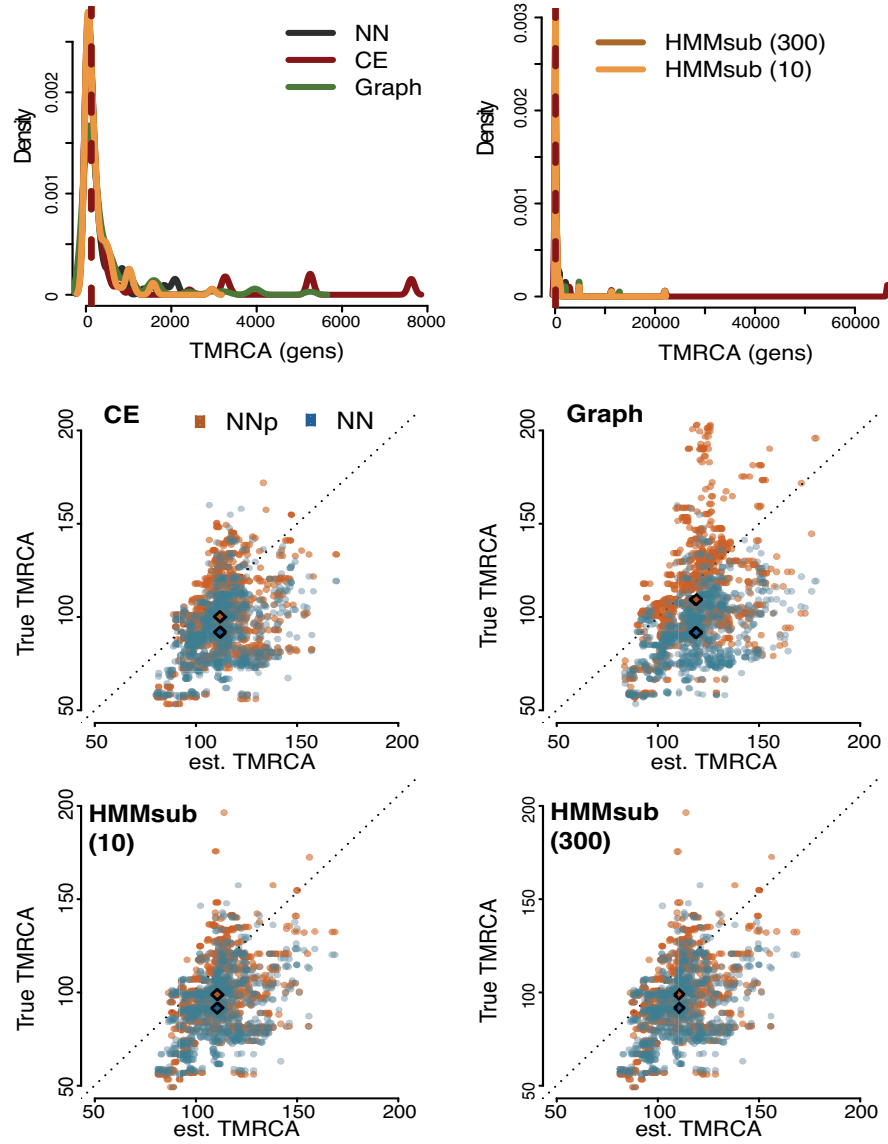


Figure 5.1 Estimating the TMRCA from simulated data. We sample SNPs across simulated data and at each SNP determine a) all TMRCA between all NNs, b) all TMRCA between pseudo-parents (selected by the full HMM with compatibility emissions (CE), the graph method (Graph) or HMMsub with 10 and 300 SNPs for subsampling, c) TMRCA point estimates using the decay approach (either from the CE, Graph or HMMsub results). The top row focuses at two SNPs (left and right). Each figure shows the density curve for the TMRCA between true NNs (black), between pseudo-parents from CE (dark red), from the Graph (green) and HMMsub (dark brown and orange for 300 and 10 SNPs respectively) at that SNP. The vertical dashed lines show the TMRCA point estimates using the decay approach. The bottom two rows panel shows the TMRCA decay point estimates (x-axis) against the median TMRCA across NNs (blue) and pseudo-parents (brown) for each method. There is a different point for the results at each SNP.

5.3 Data and quality control

Genotyping and quality control (QC) were performed separately for the individuals and SNPs in each cohort and full details for both stages may be found in Tachmazidou et al. [2013] and Panoutsopoulou et al. [in prep.]. SNPs with a $MAF < 1\%$ were excluded and therefore there were no singletons in the data. In addition to the QC steps carried out by the Zeggini group at the Sanger institute, I applied a more stringent filter on SNPs and excluded those that failed the Hardy-Weinberg equilibrium (HWE) test (log-likelihood ratio test, $p\text{-value} < 0.001$), as well as SNPs that contained too much missing information ($>5\%$). Overall only the SNPs that passed QC for all three cohorts were used in the analyses.

Given the nature of the data, the presence of cryptic relatedness in the samples is not unlikely. This would be picked up by our graph method, resulting in the detection of longer than expected shared haplotypes, which would inflate our estimates of coancestry. In order to avoid this effect, the analysis was restricted to the individuals in each of the three cohorts that exhibited low levels of relatedness (all pairwise $\hat{\pi} < 0.2$; $\hat{\pi}$ is the proportion of sites in IBD between pairs of individuals, as calculated by PLINK, Purcell et al. [2007]). The resulting datasets consisted of 754 unrelated individuals in MANOLIS, 567 in Pomak and 707 in TEENAGE. There were 555,403 SNPs overlapping among the three cohorts. In the MANOLIS cohort, only 32 individuals were carriers of the R19X variant.

5.4 Results: *APOC3*

As expected from an isolated population, haplotype sharing over Chromosome 11 is extensive among the MANOLIS individuals, such that the per person median size of maximal haplotype identity around a randomly sampled chromosomal location (the N50) is 14.8 Mb (15.3 cM). The per person median chunk size is 2.9Mb (3.8cM). Assuming, these haplotype chunks are IBD it follows that the number of meioses separating MANOLIS NNs is about 25.

What is more, the region around the variant does not appear to be different from other regions on the chromosome, with haplotype sharing ranging from $<1\text{cM}$ to $>30\text{cM}$ (Figure 5.3 inset). However, we also find that sharing is not greater among carriers than non-carriers (Kolmogorov-Smirnov test $p\text{-value}=0.8$; Figure 5.2). The limited sharing around the region, coupled with the fact that there are no differences between carriers and non carriers, suggest that the mutation is old in the sample's genealogy, consistent with the findings using the UK10K data. The results further conclude that a study of haplotype chunk sizes in the form of IBD mapping would not be successful in this case, as expected.

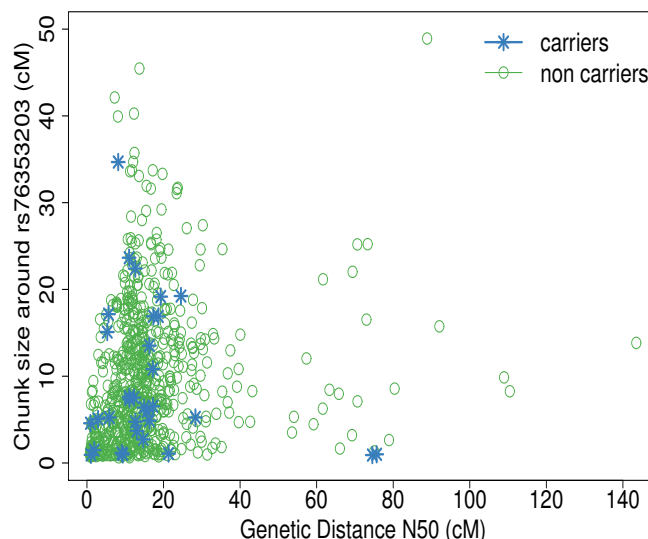


Figure 5.2 **R19X N50**. Scatterplot showing the relationship between the N50 of maximal haplotype sharing for each sample across the chromosome (x-axis) and at the risk variant (y-axis). Carriers (blue) and non-carriers (green) are shown separately. Note that carriers and non-carriers show similar patterns (no significant difference, $p>0.05$, by Kolmogorov-Smirnov test).

Interestingly, there does not appear to be any statistically significant stratification within this sample (Figure 5.3 inset) across the chromosome. There is also no evidence for cryptic relatedness, substructure or any extended correlation between genetic background and carrier status (Figure 5.3), as excess relatedness decays exponentially away from the risk locus as expected. This illustrates how

our method can be used to check for local stratification of both common and rare variants. We have already discussed that this is key in the context of disease association studies, where classic methods fail to account for stratification of rare variants. Our method illustrates the local structure of relatedness, confirming that in this case there is no confounding and that the strength of the association that was detected is reliable.

Further study of the profile of maximal haplotype identity around R19X reveals that it is well-approximated by a model in which the common ancestor to the genealogical nearest neighbour occurred an average of 4.4 generations ago (that is, g from Equation 5.1 is 4.4, with a half life of $\sim 7\text{Mb}$). In light of the results from simulation Section 5.2.2 we know we should have limited confidence in the absolute value of this estimate, though we would expect the true median TMRCA between carrier NNs at this locus to have the same order of magnitude. This provides a lower bound for the age of the mutation at c. 100 years ago, assuming a generation time of 25 years, though the evidence that we have collected suggests that it is likely to be considerably older. This lower bound therefore does not improve much on the confidence interval estimated by other methods (0-9ky), but this very recent TRMCA to a NN is informative of the very close relatedness among the individuals in this population. Interestingly, around other variants of the same frequency (such as rs61747005 shown in Figure 5.4, for which there were 30 carriers in the sample), we find greater levels of cryptic relatedness (results are similar at more loci - data not shown), suggesting that these are typically more recent. This further highlights how our method can be used to avoid confounding (for instance at this second locus we would expect deflated p-values).

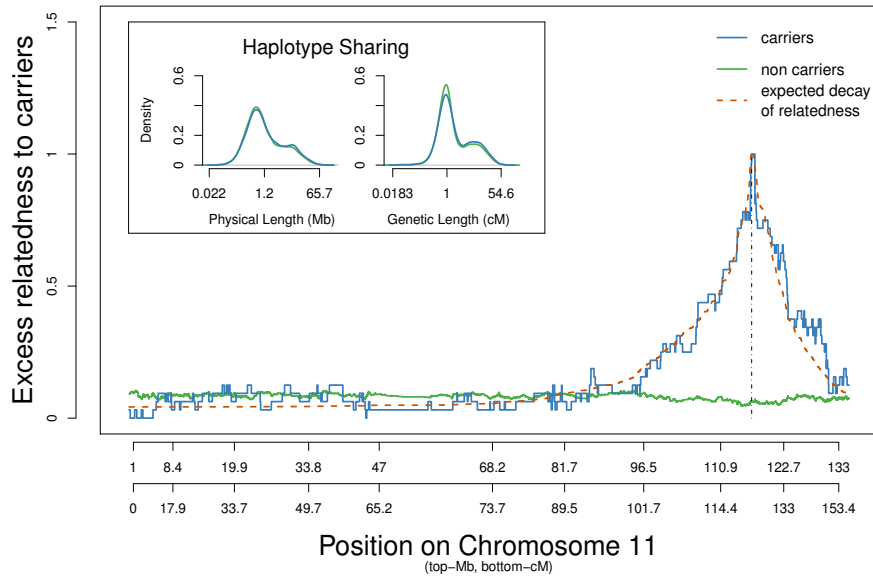


Figure 5.3 **Excess relatedness with carriers of R19X.** The main plot shows the average number of nearest neighbours that are carriers of the rare risk allele for carriers (blue) and non-carriers (green). Analysis only for unrelated samples ($\hat{\pi} < 0.2$, see Methods). The orange line shows the expectation from a model when all most recent common ancestor events to the nearest neighbour occurred 4.4 generations ago. y-axis is the average proportion (divided by 2) of parents who are carriers, for carriers and non-carriers. Inset shows the distributions of maximal shared haplotype length (on a log scale) in physical (left) and genetic (right) distance for carriers (blue) and non-carriers (green).

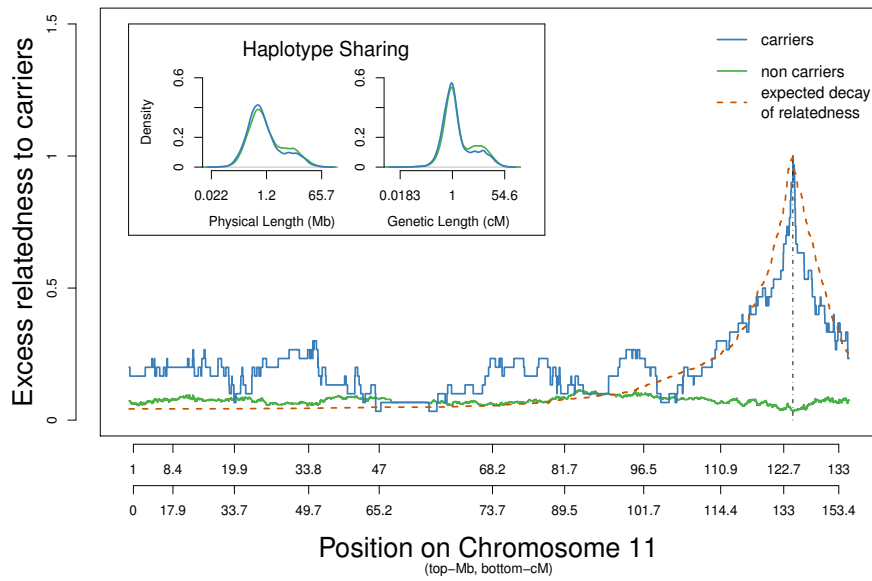


Figure 5.4 **Excess relatedness with carriers of another variant**. The figure shows maximal haplotype sharing around another variant on Chromosome 11 where the derived allele is at a similar frequency as the risk allele in *APOC3*: rs61747005, for which there were 30 carriers in the sample. The main plot shows the average number of nearest neighbours that are carriers of the rare risk allele for carriers (blue) and non-carriers (green). The orange line shows the expectation from a model when all most recent common ancestor events to the nearest neighbour occurred 4.1 generations ago. Inset shows the distributions of maximal shared haplotype length (on a log scale) in physical (left) and genetic (right) distance for carriers (blue) and non-carriers (green).

5.5 Results: A genome-wide analysis of relatedness, demography and isolatedness

We now turn to the genome-wide analysis for all three cohorts: MANOLIS, Pomak and TEENAGE, using just the unrelated individuals. We remind the reader that we carried out three separate analyses: Analysis A in which our graph method was run separately on each cohort, Analysis B in which all samples were considered together and Analysis C where the MANOLIS and Pomak individuals were asked to select NNs solely from the TEENAGE cohort.

Considering first Analysis A, we find, as expected, a much greater extent of haplotype sharing in the isolated populations, with the median length of haplotypes being 1.78cM, 1.42cM and 1.07cM in the Pomak, MANOLIS and TEENAGE samples respectively (Figure 5.5 lower panel; sampling across all chunks of all individuals). Consequently, 50% of randomly sampled locations within the genome lie in shared haplotype lengths of at least 12.4cM, 12.5cM and 1.3cM respectively (N50 haplotype length; Figure 5.5, lower panel; using all chunks from all individuals). Interestingly, the distribution of shared haplotype lengths in the isolates is bimodal (Figure 5.5, top panel), with the two peaks corresponding to very recent co-ancestry (within the last few generations) and much older-co-ancestry (as seen in the TEENAGE population). Since only unrelated samples have been used in this analysis, it is likely that this difference reflects old coancestry dating prior to the population founding event.

The high levels of relatedness within the isolates are also reflected by the fact that they need only ~ 3 NNs (with which there is maximal haplotype sharing), such that their haplotype sharing covers at least 50% of their genomes (Figure 5.6, left panel). By comparison, the TEENAGE samples require ~ 20 NNs for the same length. We discover the greater extent of isolatedness among the Pomak samples (which is confirmed by other methods carried out in Panoutsopoulou et al. [in prep.]) when considering 90% coverage (Figure 5.6, right panel): the Pomak samples require only ~ 12 NNs, while results for the MANOLIS are starting to resemble those of the TEENAGE, with the samples in each case requiring approximately 45 and 55 NNs respectively.

5.5 Results: A genome-wide analysis of relatedness, demography and isolatedness

To investigate whether there is clustering in the data when we remove cohort labels from the individuals, we proceed with Analysis B. The extent of isolation of the Pomak and MANOLIS samples is again reflected in that NNs are typically chosen from within their cohort: 88.4% and 90.5% respectively (Figure 5.7; mean values are shown in yellow), versus 48.5% for the TEENAGE samples. While about 10% of the isolates have NNs from the TEENAGE cohort, the shared haplotype chunks specific to these NNs are significantly shorter, 3-fold and 4-fold for the MANOLIS and Pomak respectively (Figure 5.8), again suggesting that typically the inferred coancestry pre-dates the isolation events. The limited number of larger haplotype chunks (1% of chunks in MANOLIS copied from TEENAGE are $>5\text{cM}$, 0.5% for the Pomak; Figure 5.8) is indicative of very limited migration and gene flow events, though we cannot eliminate the possibility of isolation being established in more than one wave. Our method is only powered to detect the most recent ancestry and we thus only infer a lower limit on the time when gene-flow ended.

We finally attempt to date NN coancestry within and between cohorts. To estimate these ages we fit a model of exponential decay of haplotype sharing around 100 randomly sampled locations across the genome in Analyses A and C (Figure 5.9). By indicatively examining one of these SNPs (results are equivalent at other loci; data not shown) we observe that, as expected, decay in coancestry is much more rapid between the TEENAGE individuals (panel ‘T’) than for the isolates (top two panels). In all cases, the observed decay (blue curve) away from the locus is as expected under the theoretical model (red curve). We do not attempt any further SNP-specific assessments as we anticipate results to not be particularly reliable (as per our findings and discussion in Section 5.2.2), but we do proceed to consider the genome-wide findings.

We find that the median age of NN common ancestors is 8.56 and 8.65 generations in Pomak and MANOLIS respectively, which is very recent (again as expected given the extent of isolatedness), but 89.7 generations in the TEENAGE cohort (Figure 5.10), which is comparable to that observed in other European populations (data not shown). Using our simple moment-estimator for N_e , these values imply recent effective population sizes of 2,400, 3,200 and 32,000 for the three populations respectively. Our values are similar to separate estimates that

5.5 Results: A genome-wide analysis of relatedness, demography and isolatedness

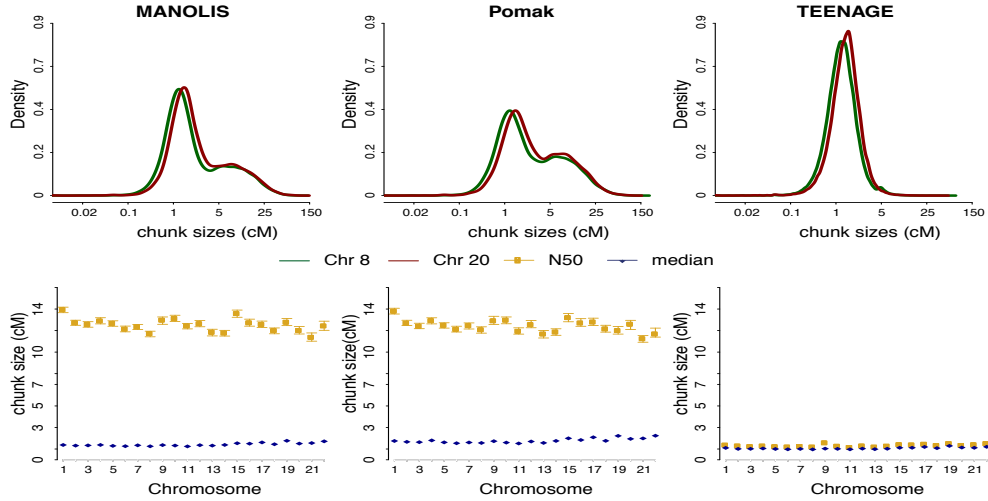


Figure 5.5 **Haplotype sharing in the isolates and TEENAGE**. Maximal haplotype sharing between NNs for each cohort. The top panel shows the density curves for the shared haplotype chunks for Chromosomes 8 (dark green) and 20 (dark red). Results were similar for the other chromosomes (data not shown). The bottom panel shows the per chromosome median (dark blue) and N50 (gold), in cM, of all shared haplotype chunks obtained from all the unrelated individuals in each cohort. The results shown here are each from Analysis A (refer to text).

were calculated in Panoutsopoulou et al. [in prep.] using a pairwise-LD based method (the McEvoy approach; McEvoy et al. [2011], Tenesa et al. [2007], Hayes et al. [2003]). This latter approach estimated recent N_e to be $\sim 8,000$ for the isolates and $\sim 17,000$ for TEENAGE (refer to Supp. Figure S7 in the paper for the full historical results). Overall, we see that the rate of genetic drift in the isolated populations is approximately 10 times that of the TEENAGE population.

We turn to the coancestry between the isolates and TEENAGE, using the findings from Analysis C and considering recent common ancestors between Pomak-TEENAGE and MANOLIS-TEENAGE. The median TMRCA is estimated to be 109 and 106 generations respectively, which is significantly older than the 90 generations of the TMRCA between TEENAGE NNs. The density of the TMRCA for each case is shifted to the right (Figure 5.10), mostly resembles the density of the TEENAGE unrelated individuals. The curves illustrate the older coancestry between these samples, compatible with the hypothesis that their coancestry

5.5 Results: A genome-wide analysis of relatedness, demography and isolatedness

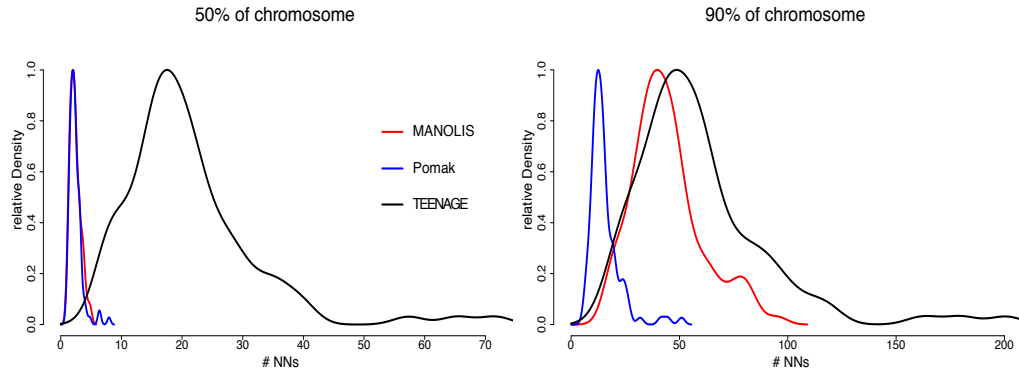


Figure 5.6 **An 80-20 study of the isolates and TEENAGE.** For each person we measure the number of pseudo-parents which collectively share 50% (left) of a chromosome and 90% (right) of a chromosome. We normalise by chromosome length such that values represent 50% and 90% respectively of the whole genome. Aggregating across individuals per cohort, the figures show the density curves for each case. Density values have been normalised such that the density curves across cohorts are comparable. The results shown here are each from Analysis A (see Methods).

mostly predates the time of isolation.

5.5 Results: A genome-wide analysis of relatedness, demography and isolatedness

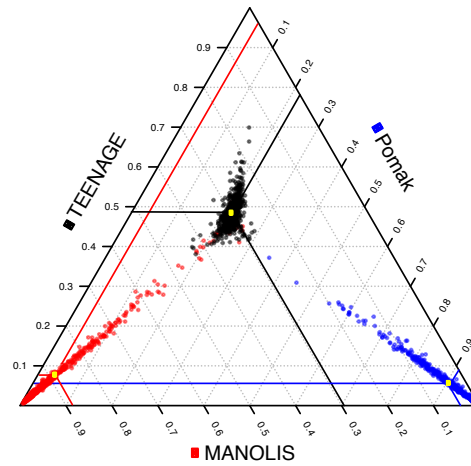


Figure 5.7 **Choosing NNs across cohorts.** The triangle plot shows how often individuals in each cohort (MANOLIS, red; Pomak, blue; TEENAGE, black) select NNs from each of the three cohorts when the algorithm was run with all three cohorts lumped together (Analysis B). The yellow points show the mean values for each cohort. For example, a MANOLIS individual has NNs that are also MANOLIS 88.3% of the time, NNs from the Pomak samples 3.8% and TEENAGE 7.8% of the time. The results are genome-wide means per person. The figure reveals clear clustering, highlighting the divergence of the two isolates from the general Greek population and from each other. Moreover, the figure highlights the greater extent of isolatedness for the Pomak samples, compared to MANOLIS.

5.5 Results: A genome-wide analysis of relatedness, demography and isolatedness

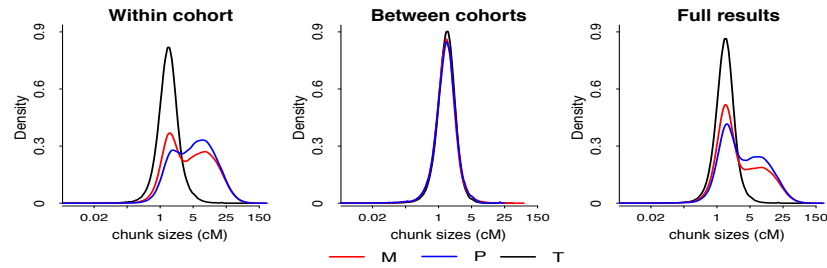


Figure 5.8 **Haplotype sharing among cohorts**. The figures show density curves for the haplotype chunks when our graph method is run with all samples lumped together (Analysis B). Panel A shows the density of chunks shared with NNs that belong to the same cohort as the query individual (red, M = MANOLIS; blue, P = Pomak; black, T = TEENAGE). Panel B looks at chunks for the cases where an individual's NN is from a different cohort: red = between MANOLIS and TEENAGE; blue = between Pomak and TEENAGE; black = between TEENAGE and either MANOLIS or Pomak. Panel C looks at the full density of shared chunk sizes with all of an individual's NNs, independently of the NNs' cohort. The results have been combined across chromosomes as there was minimal variation among them (data not shown).

5.5 Results: A genome-wide analysis of relatedness, demography and isolatedness

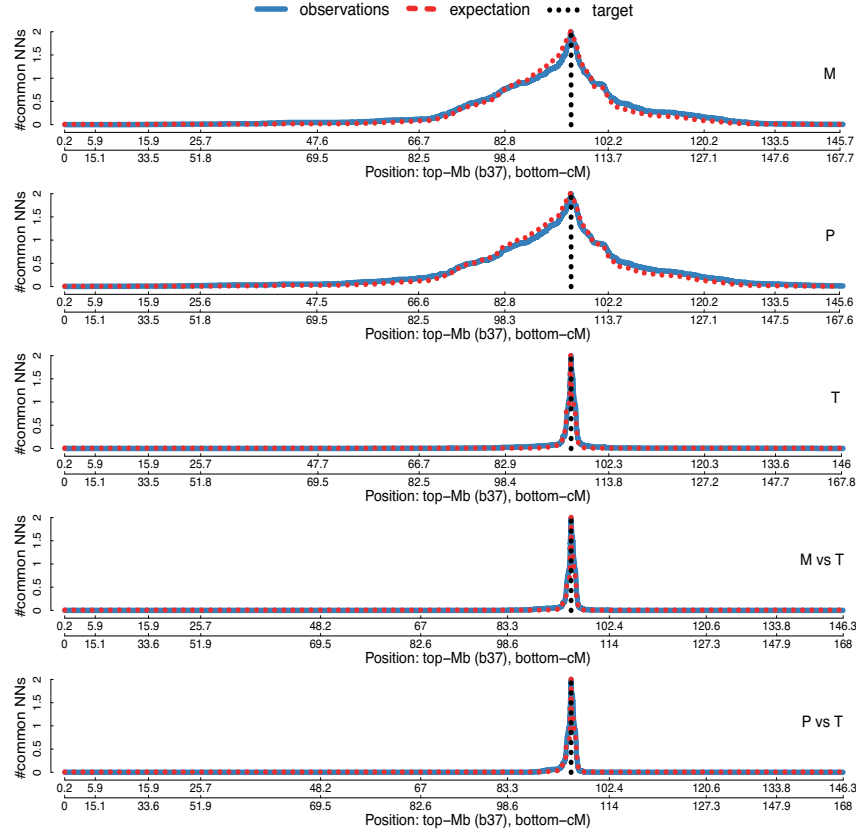


Figure 5.9 A local TMRCA. The decay of haplotype sharing with an individuals NNs at rs6989467 (randomly selected) on Chromosome 8. The x-axis is position on Chromosome 8 (physical, Mb and genetic, cM) and the y axis is the number of NNs that are unchanged (compared against the NN choice at rs6989467), averaged over all the individuals in the sample. In blue is the observation for each cohort and in red the expected curve of NN sharing decay (see Methods). The top 3 plots show results when the algorithm was run on each cohort (MANOLIS = M; Pomak = P; TEENAGE = T) separately (Analysis A). Here, the mean TMRCA with a NN in MANOLIS is 7.3 generations, among the Pomak individuals it is 6.4 generations, while in the TEENAGE cohort it is 77.2 generations. The bottom two plots show results for MANOLIS (M vs T) and for Pomak (P vs T) when forced to pick NNs from the TEENAGE cohort (Analysis C). In the former case the mean TMRCA to a TEENAGE NN is 100.4 generations (at this SNP) and in the latter it is 89.2 generations. We have repeated this at another 99 randomly sampled SNPs across the genome obtaining equivalent results; the figures are omitted.

5.5 Results: A genome-wide analysis of relatedness, demography and isolatedness

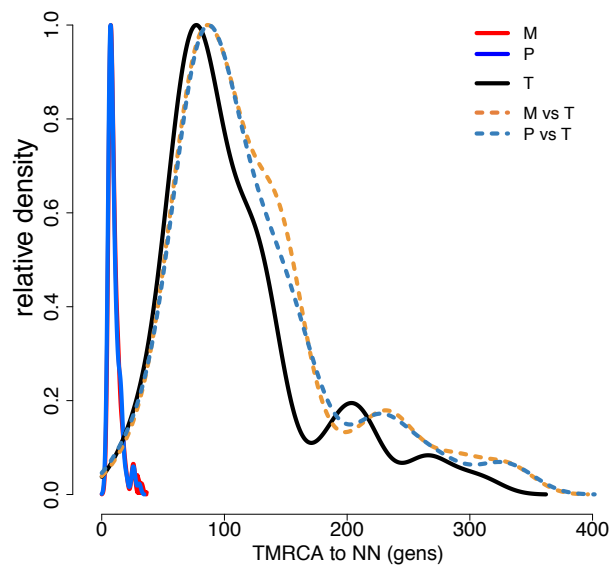


Figure 5.10 **TMRCA with a NN**. Relative density curves showing the estimated number of generations to the common ancestor between an individual and their nearest-neighbour. The results are from 100 randomly sampled SNPs from across the genome for the MANOLIS (M), the Pomak (P), the TEENAGE (T) individuals, each considered separately (Analysis A), and a scenario where the MANOLIS (M vs T) and the Pomak (P vs T) individuals each are forced to select NNs from the TEENAGE cohort (Analysis C).

5.6 Dating the time of divergence

To end this chapter, we revisit the issue of estimating the time of divergence. Although we have not tested this approach, we can use our estimates of the TMRCA between NNs to make some initial heuristic inference on the time of divergence (T_{div}) between the general Greek population and the isolates from an ancestral population. Our aim is simply to illustrate the potential application of our findings in exploring demography by dating events.

We suppose that the TEENAGE cohort has not experienced recent rapid exponential growth, and further, that T_{div} is not very old. We may then assume a constant population size model for the TEENAGE cohort from the present until that time of divergence. The TEENAGE samples represent the general Greek population and are therefore largely unrelated. We thus expect the TMRCA between NNs (T_{Teen}) to be significantly older than the time of divergence; more specifically we will explicitly assume that no coalescence events have occurred in the TEENAGE cohort since divergence occurred, i.e. $T_{Teen} > T_{div}$. From this it follows that the sample size of the TEENAGE cohort has not changed by T_{div} . Under the aforementioned assumption of a constant-population size, it follows that the time to a coalescence event between NNs at the present is the same as the time to a coalescence event between NNs starting from the time of divergence, T_{old} . In other words, we will be assuming that $T_{old} \sim T_{Teen}$.

Analysis A provides us with an estimate of T_{Teen} , using the technique presented in the previous section. Using the results from Analysis C we may obtain the TMRCA between NNs in which one individual is from an isolate and the other is from the general Greek population, $T_{between}$. Assuming there has been no gene flow since divergence, which is reasonable (see the Results in Section 5.5), it follows that $T_{between} > T_{div}$. More particularly, we have that $T_{between} = T_{old} + T_{div}$ which gives us an estimator for the time of divergence:

$$T_{div} = T_{between} - T_{Teen}$$

The overall framework for this approach is illustrated in Figure 5.11.

The critical step in this derivation of this estimator is the approximation $T_{old} \sim T_{Teen}$. If we lift our assumptions about a constant-population sized history

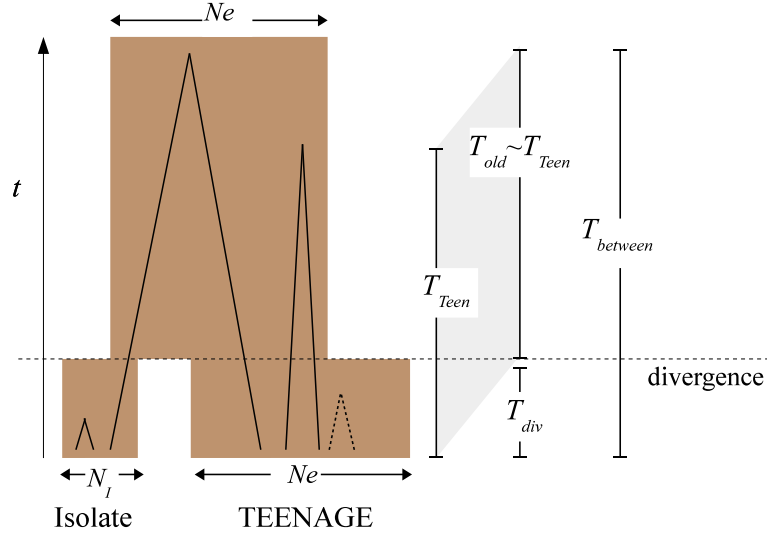


Figure 5.11 **Estimating time of divergence** The figure demonstrates the event of divergence (horizontal dashed line) between an isolate and the TEENAGE cohort from an ancestral population. T_{div} is the time at which divergence took place. The brown shaded region represents the genealogy, and includes possible coalescence events marked by the black lines. The dotted black lines showing a recent coalescence between TEENAGE NNs before the time of divergence highlight the type of events that we are assuming have not taken place. The width of each part of the genealogy is proportional to the effective population size, which is N_I for the isolate and N_e for the TEENAGE cohort and ancestral population. In other words, we are further assuming here that the effective population size for the TEENAGE cohort does not change going backwards in time. In this set up we may calculate the TMRCA between TEENAGE NNs. T_{Teen} (using our Analysis A on the TEENAGE cohort; refer to text) and the TMRCA between NNs when these involve an individual from an isolate and an individual from the TEENAGE samples, $T_{between}$ (Analysis C; refer to text). T_{old} is $T_{Teen} - T_{div}$, the time to the TMRCA between TEENAGE individuals since divergence. We approximate T_{old} by T_{Teen} ; refer to text for a detailed explanation of this approximation.

then this relationship will not hold. For example, in the case of a bottleneck for the overall founding population, followed by rapid exponential growth we would have that $T_{old} \gg T_{Teen}$ (and therefore if this is the case here we expect our assumptions to break down). Given the small deviations from the constant size model that we expect here, we are probably only slightly underestimating T_{old}

5.6 Dating the time of divergence

and our T_{div} is therefore an upper bound, though reasonably close to the true value. As these assertions have not been confirmed by simulation, we remind the reader that the purpose of this Section is simply to illustrate potential extensions and applications of our approach.

Turning to the results, as already mentioned the median TMRCA between MANOLIS-TEENAGE and Pomak-TEENAGE is estimated to be 109 and 106 generations respectively, suggesting that an upper bound for the time of isolation is 19 generations and 16 generations respectively. Note again that these estimates assume a simple model in which separation is a single event, the effective population size of the TEENAGE population has remained constant over time and that there has been no gene flow between the two isolates and the general Greek population since isolation began. For example, the data may also be compatible with a model of older divergence with a low level of more recent migration or a situation where there were more than one historical waves of migration. A more formal simulation study built on the assumptions that we have made should in principle be carried out to test whether the expected values of haplotype sharing, TMRCA between NNs and time of divergence match the findings that we are discovering here. We finally note that the McEvoy-based approach estimates divergence to have occurred 39 and 52 generations ago for the MANOLIS and Pomak samples, which are older than our estimates by slightly more than a factor of two. Given the high uncertainty in both approaches, this is not considered to be a large difference.

Chapter 6

Applications to real data:

WTCCC

The HELIC study demonstrated the graph method’s capabilities in characterizing recent coancestry within unrelated and isolated samples. Although we have established the merits of such an investigation, it is nevertheless the case that the sample sizes of only a few hundred individuals involved in this project are rather small by modern standards. Genome-wide association studies are now frequently carried out on cohorts consisting of many thousands of individuals (for example Wellcome Trust Case Control Consortium et al. [2007], International Multiple Sclerosis Genetics Consortium et al. [2011], 1000 Genomes Project Consortium et al. [2012]).

Just like with the HELIC data, even within large data sets, it is not unusual to identify pairs of individuals who are related and these are typically removed from analysis. However, all ‘unrelated’ individuals have genealogical nearest neighbours and, within a large cohort, the TMRCA between NNs may be relatively recent. Identifying these individuals will help understand patterns of local relatedness and latent stratification that may be present in a sample, even if the extent of this stratification is weak. In fact, numerous studies have

shown that extensive population substructure may exist within samples (Li et al. [2008], 1000 Genomes Project Consortium et al. [2012]). The identification of NNs and the determination of how they change captures an important aspect of the genome-wide relationships between pairs of individuals, thus reconstructing the most recent part of their genealogical tree which is relevant for recent demographic events. Local relatedness is also relevant for rare variants ($\text{MAF} < 5\%$), whose importance we have already discussed. Therefore knowledge of NNs and/or haplotype sharing could be useful in association studies, with haplotype sharing around rare variants offering insights that could be incorporated in a mixed model framework.

In light of the above considerations, in this chapter we perform a more extensive exploration of our methods' performance using the WTCCC genotype data which involves ≈ 5000 unrelated British samples (Wellcome Trust Case Control Consortium et al. [2007]; more details are presented in section 6.2.1). This genotype data is mostly limited to common variation and was not selected to investigate our performance with respect to rare variants; this is one potentially future application of our model. Rather, the WTCCC data was a suitable choice for investigating our methods' behaviour with large sample sizes and for understanding the insights that we can derive from this well-studied sample. Here we further simply wish to illustrate that our methods behave as expected in large sample sizes, are informative and will therefore potentially accommodate the aforementioned applications (among others).

Traditionally, an IBD-based approach would be employed to identify haplotype sharing. The WTCCC samples that we use here have been shown to have very little structure (Wellcome Trust Case Control Consortium et al. [2007]) and is therefore precisely the type of data set where IBD detection is by definition not possible. Common ancestry, and recent coancestry, do of course exist and we explore our method as an alternative for detecting it.

The analyses carried out in this chapter were initially to be conducted exclusively using the graph method (see Chapter 2, section 2.4.2). More recently, however, we developed our HMM with subsampling approach (HMMsub; see Chapter 2, section 2.4.3), and therefore some of our analyses have been carried out with this method as well. We initially make use of the WTCCC samples to

further explore our methods’ properties in a ‘real-data’-based environment. More specifically, with the graph method we explore the changes in detected haplotype sharing when we increase the number of incompatibilities required for a breakpoint (see Chapter 2, section 2.4.2 for details). Using both the graph method and the HMMsub we investigate the effects of thinning and error rate on haplotype sharing. We then consider both methods as alternative approaches to making inference on the full WTCCC data.

The assumption that we are making is that our baseline full-HMM is capable of detecting the true levels of sharing present in this sample, and that these two alternative methods are appropriate approximations of the full model that can accommodate larger sample sizes. We use them to investigate the levels of haplotype sharing and the TMRCA between NNs, comparing it to our findings from the HELIC project. We then use our insights to make inference on the demographic history of the British population by estimating exponential growth parameters.

It is however crucial to gain some understanding of how both our approximation methods actually compare against the full HMM. Running the full HMM on the WTCCC samples is computationally impossible both due to memory requirements and time required to complete the analysis. What is more, such a comparison should use very high quality genotypes to control for the effects of genotyping error. To this end, prior to our extensive WTCCC investigation, we first consider a small-scale analysis using the 1000 Genomes Omni SNP genotype data (1000 Genomes Project Consortium et al. [2012]).

6.1 A prequel: 1000 Genomes Omni SNP genotypes

We consider the European samples from the 1000 Genomes (1000G) Omni SNP genotype data (1000 Genomes Project Consortium et al. [2012]), consisting of 388 individuals. These are high quality samples that have undergone rigorous quality controls, described in detail in 1000 Genomes Project Consortium et al. [2010] and 1000 Genomes Project Consortium et al. [2012], and are thus suitable

6.1 A prequel: 1000 Genomes Omni SNP genotypes

for comparing the performance of our methods: the full HMM against HMMsub (subsampling around 300 and 1000 randomly selected SNPs for each chromosome) and the graph method. It is essential to use data with minimal error rate as the different methods are not expected to be equally robust to genotyping error. As errors are expected to lead to an underestimation of haplotype sharing, it is crucial to ensure that any differences in the coancestry estimated by the three methods is solely due to their respective properties.

Being restricted by the memory requirements and the very low speed of the full HMM, we only consider results for Chromosome 20 of 10 individuals. We excluded singletons and any SNPs that fail our HWE log-likelihood ratio test (thresholding with a pvalue of 10^{-3}) and were left with a total of 38510 SNPs and indels. We ran our graph method, subsampling 20 individuals after 2 single incompatibilities or 1 double incompatibility (these are the same settings as those used in the HELIC study). We used the compatibility emissions for both HMM implementations. In our HMM with subsampling, subsampling was such that within each chunk there were approximately 20 individuals.

Given the limited number of individuals that were investigated and the fact that only a single chromosome was used, we do not aim for this investigation to be a comprehensive comparison of the two methods. Rather, we hope to gain insight on the bias that results from the WTCCC study are expected to exhibit. More specifically, here we measure:

1. The extent of haplotype sharing, as described previously in Chapters 2, 4 and 5.
2. The proportion of sites at which the selected pseudo-parents are incompatible with the query individuals.
3. The number of individuals that are selected as pseudo-parents over the chromosome.

The current implementation of our full HMM required 30G and 2 weeks to run for a single individual. Although it is possible to reprogram the algorithm such that the memory requirements are reduced, it is not possible to further speed up the algorithm; most of the computational time was spent constructing the Viterbi

matrix. The run time for the other two methods is slightly variable (depending on the number of samples considered at every position) but around 2 minutes for the graph method and half an hour for HMMsub.

6.1.1 Results

The results show that although neither of the two methods replicate the performance of the full HMM, both versions of HMMsub approximate it very well (Figure 6.1). We will refer to HMMsub with subsampling around 300 SNPs and 1000 SNPs as H300 and H1000 respectively. Across all chunks from all individuals, we find that the graph method detects an N50 of 1.05cM, H300 finds 1.15cM and H1000 finds 1.18. The full HMM, which we assume to be most accurate in detecting chunk sizes, estimates the N50 to be 1.19cM. The graph method consistently underestimates sharing across all samples (Figure 6.1 D and E), probably being very sensitive to even the very low levels of genotyping error that inevitably must be present in the data. It is interesting that the graph method changes pseudo-parents more often than the other two methods (Figure 6.1 A and B), while maintaining the highest proportion of incompatible sites (Figure 6.1C). Conversely, HMMsub tolerates incompatibilities at a slightly higher rate than the full HMM (Figure 6.1C). It swaps NNs at a rate between the graph method and the full HMM (Figure 6.1 A and B), estimating equivalent levels of sharing as the full HMM (Figure 6.1D). The overall distribution of the shared chunk sizes detected by the HMM with subsampling closely resembles that of the full HMM (Figure 6.1E), suggesting that it might still be a reasonable proxy. We notice that H1000 has not recovered the higher levels of sharing present in the third individual (Figure 6.1 D) and it therefore appears to slightly underestimate sharing (Figure 6.1 E). Overall, the graph method and HMMsub can be used, if not to identify true levels of coancestry, at least to reliably define a lower bound for the extent of haplotype sharing. Though HMMsub appears to be more accurate, the graph method remains an order of magnitude faster.

To determine whether indeed H300 and H1000 perform different or whether the differences observed here are an artefact of the small number of individuals

6.1 A prequel: 1000 Genomes Omni SNP genotypes

selected, we analysed an additional 100 individuals using these two methods. The results, shown in Figure 6.2 confirm that, like in our results from simulation (see Chapter 4) there is no difference between the two approaches.

Finally, an inspection of where the NN switches take place across the 10 individuals along the chromosome, reveals that at the 10kb- and 100kb-level the graph method often changes at different positions for each person (Figure 6.3). Conversely, the full HMM and HMMsub (both versions) will both frequently choose to change at the same positions for many individuals, in a way that correlates with the genetic map in the region (data not shown). These observations confirm that it is mostly local recombination rates that dictate pseudo-parent switching (Figure 6.3). The effect is weaker for the graph method, which is expected since it does not incorporate LD information in its implementation.

6.1 A prequel: 1000 Genomes Omni SNP genotypes

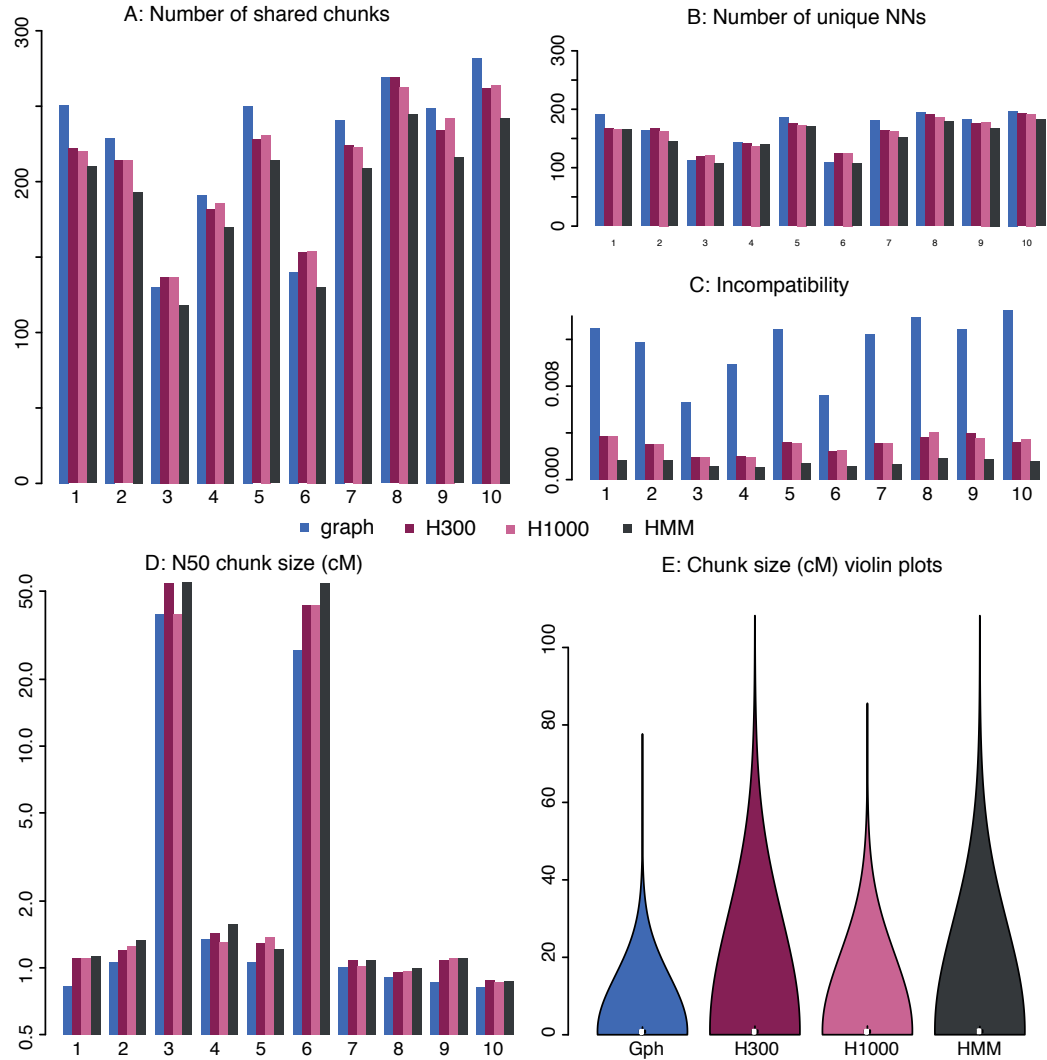


Figure 6.1 **Comparing the three methods on 10 individuals.** The figures show the results obtained by each method (graph, HMMsub – with 300 and 1000 SNPs for subsampling, labelled H300 and H1000 respectively – and the full HMM) on the 10 Omni genotype individuals that were considered, from Chromosome 20 with 38,509 SNPs. Each time there is a NN change along one of the chromosomes of each person a new shared chunk is defined. Figure A shows the total such chunks per person. It is possible that more than one chunks are shared with the same person in the sample, at different parts of the chromosome. Figure B shows the total number of unique individuals identified as pseudo-parents by the 3 methods. Figure C shows the mean proportion of incompatible sites between the query individuals and the selected pseudo-parents. Figure D shows the N50 (median shared genetic chunk size, weighted by physical distance) per person; note the log scale. Finally, Figure E contains violin plots illustrating the distribution of all the shared chunk sizes across all individuals for each method.

6.1 A prequel: 1000 Genomes Omni SNP genotypes

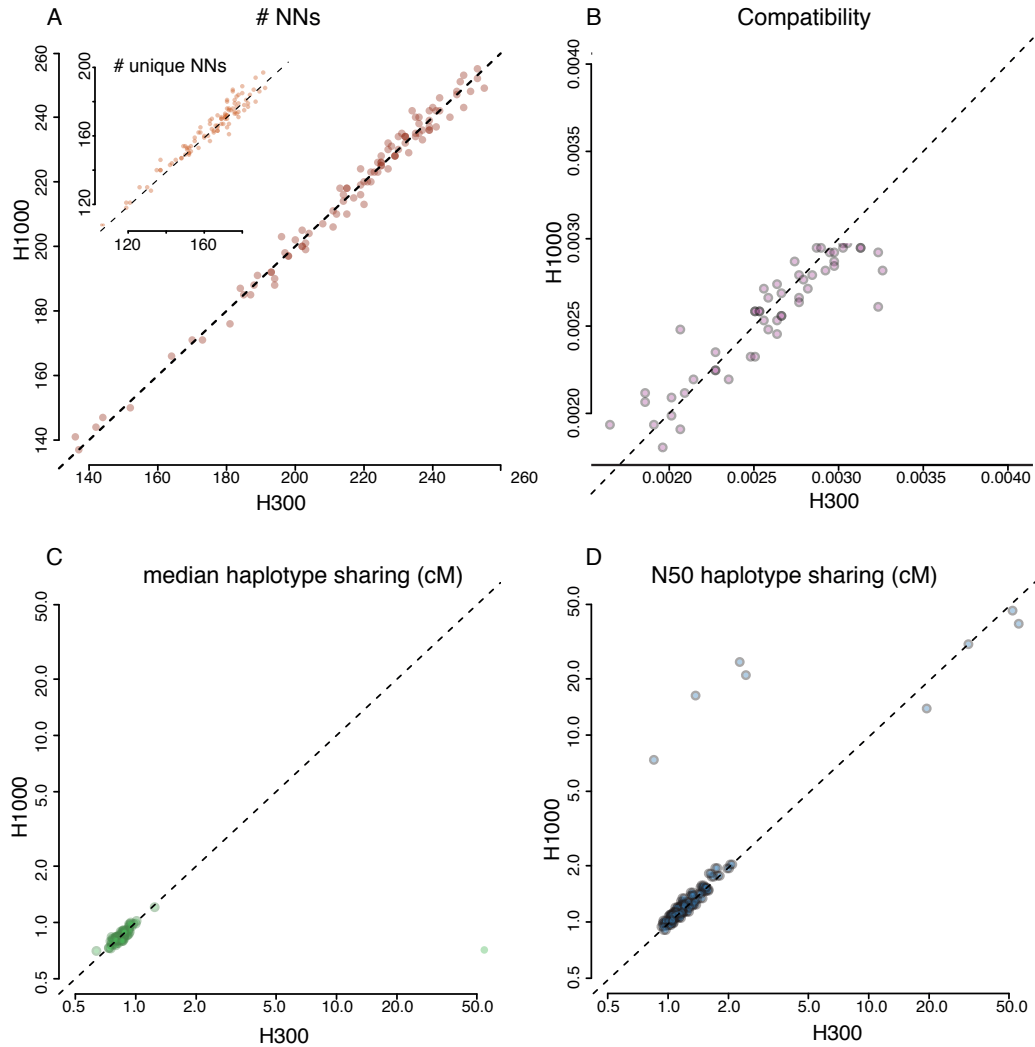


Figure 6.2 **Comparing H300 and H1000 on 100 individuals.** The figures show the results obtained by the two versions of HMMsub – with 300 and 1000 SNPs for subsampling, labelled H300 and H1000 respectively – on 100 Omni genotype individuals, from Chromosome 20. Each time there is a NN change along one of the chromosomes of each person a new shared chunk is defined. Figure A shows the total such chunks per person, while the inset shows the number of unique pseudo-parents selected. Figure B shows the mean proportion of incompatible sites between the query individuals and the selected pseudo-parents. Figures C and D respectively show the median and the N50 (median shared genetic chunk size, weighted by physical distance) per person.

6.1 A prequel: 1000 Genomes Omni SNP genotypes

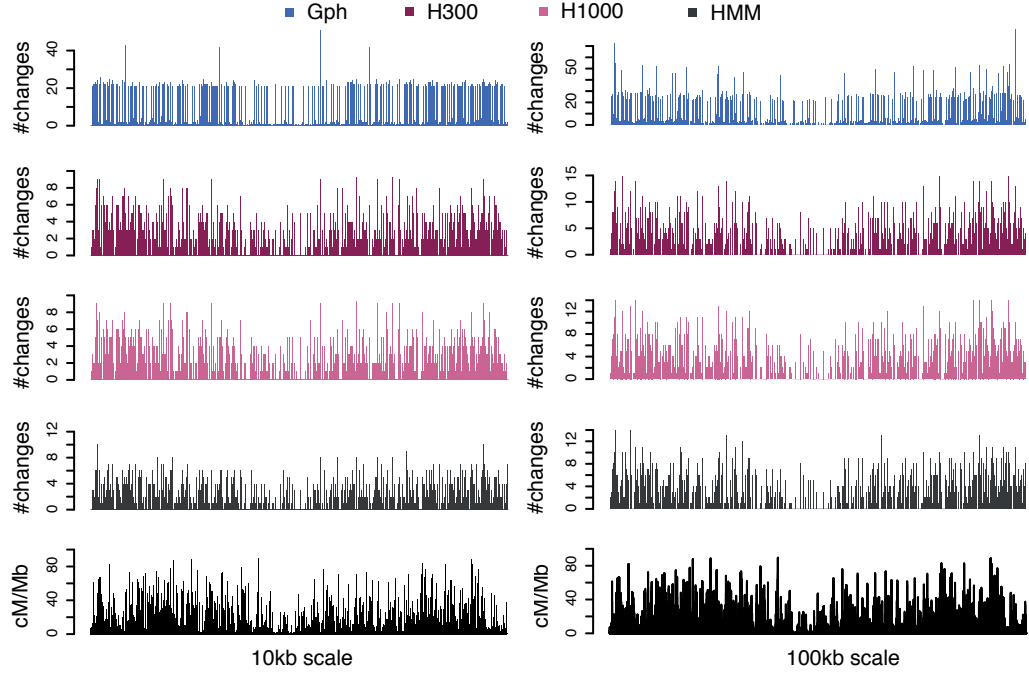


Figure 6.3 Where do the NNs change? We monitor the physical positions at which an individual changes NN and summing across individuals we get the total number of changes occurring at any given site; this is therefore at most 10 in this case, given that we disallow double-jumps. Summing across all changes in consecutive 10kb (left panel) and 100kb (right panel) windows we obtain the total number of changes for each region. Since it is possible that a person changes NN more than once within each region, the number of changes may now exceed 10. Gph = results from the graph method; H300, H1000 = results from our HMM with subsampling with 300 and 1000 SNPs respectively; HMM = results from the full HMM. Finally, the bottom figures illustrate the recombination rate at each physical position along Chromosome 20.

6.2 The structure of haplotype identity in a sample of over 5000 UK individuals

6.2.1 WTCCC: data, aims and methods

The WTCCC data consists of 5144 individuals from the UK, collected as control samples within the Wellcome Trust Case Control Consortium (www.wtccc.org.uk) and genotyped on high density SNP arrays. The data has previously undergone rigorous quality checks, including:

- Checks for missingness, with samples being excluded if they contained more than 3% missing information
- Exclusion of samples with excess heterozygosity
- Concordance between different sets of genotypes for the same samples, when these were available
- PCA analysis to confirm uniformity of ancestry and lack of large-scale stratification
- Visual inspection of cluster plots of SNPs and consideration of quality metrics from the genotype-caller CHIAMO (Marchini et al. [2007])
- Checks for Hardy-Weinberg equilibrium (using a p-value threshold of 5.7×10^{-7})

We applied an additional test for HWE using a p-value threshold of 10^{-3} and are therefore confident that the levels of genotyping error in the data are low. In additions, we ensured that there was no cryptic relatedness in the sample by checking whether individuals shared whole chromosomes. This was achieved by measuring genotype-IBS status across entire chromosomes; any individuals that were compatible across the whole region (for example because of duplications in the data) were excluded.

We first consider different subsets of the data to better understand how our graph method behaves, identifying the factors that influence its performance.

6.2 The structure of haplotype identity in a sample of over 5000 UK individuals

More specifically, in this real-data setting we explore the effects of sample size, SNP density and genotyping error. We remind the reader that our subsampling scheme for the HMM (HMMsub) was developed more recently and has therefore been included in only a subset of this study. We perform the following analyses:

1. Employing the graph method we consider just Chromosome 12 and only use the 26,715 SNPs that were genotyped in cases for Multiple Sclerosis from the 2011 study (International Multiple Sclerosis Genetics Consortium et al. [2011]). This eliminated approximately 50% of the typed variants. We then ran the graph method on the entire sample, and on random subsets of 89, 100, 500 and 1000 individuals, recording how haplotype chunk size (median and N50) changes for each case.
 - Using all 5144 individuals for this data, we also explore the effect of changing our stopping rule for the graph method, which has been heuristically set to 2 incompatibilities. We investigate the changes in haplotype sharing when this threshold increases from 1 incompatibility to 5.
 - We considered 89 individuals, as this is the number of British individuals that were sequenced in the Phase 1 of the 1000 Genomes project. We wish to compare our results from the 89 randomly subsampled WTCCC individuals to those obtained by running our method on the whole Chromosome 12 (369,038 SNPs) of the 89 UK individuals from the 1000 Genomes. Our preparation of the 1000 Genomes data was the same as that for the WTCCC data, though a repeat analysis with more rigorous quality checks on the 1000 Genomes data is recommended, owing to the higher genotyping error that is known to be present in those samples. We therefore expect a decrease in the extent of haplotype sharing when moving to the 1000 Genomes data.
2. Using Chromosome 22 (16,032 SNPs), we more formally explore the effect of thinning and genotyping error on haplotype chunk sizes (median and N50). More specifically, using all 5,144 individuals, we apply a genotyping error of 0.5%, 1% and 2%. In this investigation we consider both the graph method's

6.2 The structure of haplotype identity in a sample of over 5000 UK individuals

and HMMsub’s behaviour. For the graph method we maintain a breakpoint threshold of 2 single incompatibilities or a double incompatibility. For the case of HMMsub, again we consider subsampling both around 300 SNPs (H300) and 1000 SNPs (H1000). With both methods we aim to subsample around 20 individuals at each position. The purpose of our experiment is to explore how either method performs when the underlying error rate in the data is unknown. By subsampling SNPs uniformly from the original data set without error, we further consider thinning by keeping only 90%, 60%, 50% and 30% of SNPs. In principle, some thinning should not influence the extent of haplotype sharing detected between individuals. However, it is unclear how well powered the method will be in detecting short chunks and with thinned genotype data the detection of small chunks might be even harder. In the case of real data though, an increase in the number of SNPs increases the number of genotyping errors that may be present (errors are not necessarily distributed evenly across the chromosomes), leading to smaller chunk sizes being detected.

In the second part of our investigation, we perform an exploratory genome-wide analysis of the original WTCCC data using both the graph method and HMMsub (H300 and H1000), with about 20 individuals per locus for each method. Lacking knowledge of true NN relationships, our aim is to make inference on coancestry and demographic history using our detected shared haplotype chunks. More specifically:

1. We obtain the distribution of maximal haplotype sharing across the genome and wish to address the following questions:
 - (a) Can we identify small shared haplotype chunks, those that IBD-based approaches would be unable to detect?
 - (b) Are there patterns of excess relatedness present in the sample, genome-wide or chromosome specific, that might indicate stratification?
 - (c) What coalescent-based demographic model could plausibly explain the observed distribution of haplotype sharing? Is it possible to calibrate the tree-simulations that we describe in Chapter 4, Section 4.2.2 such

6.2 The structure of haplotype identity in a sample of over 5000 UK individuals

that the simulated haplotype chunks match those observed here? This would allow us to estimate the exponential growth parameters (current N_e and number of generations since growth started) relevant to the British population, thus making inference on its demographic history. We assume that the effective population size before growth started is 10,000.

2. We map our shared haplotype chunks on the genome, in an approach similar to IBD mapping: are there regions of excess haplotype sharing, for example?
3. We repeat the approach performed in the HELIC study, and estimate the TMRCA between NN; by sampling 100 SNPs randomly across the genome, we use Equation 5.3 to get the full distribution of this TMRCA. How does the distribution discovered here compare with that of the TEENAGE samples, which are also unrelated?

6.2.2 Sample size, thinning, error and a first look at haplotype sharing

We consider first the thinned version of Chromosome 12 and explore the size of the haplotype chunks detected by the graph method, when the threshold of allowed incompatibilities between the query genotype and candidate pseudo-parents is raised. Initial experiments had suggested that a stopping-rule of two Mendelian inconsistencies was sufficient to overcome the low-level of genotyping error in the data set: results did not change substantially if more (or less) than 2 mismatches were allowed, as shown in Table 6.1, with the relative increase remaining small and chunk sizes within the same order of magnitude. Our heuristic choice of two therefore leads to a reasonable estimation of chunk sizes, while it seems as an intuitive choice in that it is stringent enough to disallow too much inconsistency, but flexible to allow some in order to account for mutations and potentially some of the errors in the data. What is more, we can identify larger chunks for sequences with many breakpoints, where one parent remains constant while the other changes much more frequently, suggesting that the choice of when to ‘break’ will not prevent highly related individuals from being discovered.

6.2 The structure of haplotype identity in a sample of over 5000 UK individuals

Threshold:	1	2	3	4	5
Phys.	0.969	1.090	1.181	1.267	1.347
Phys. N50	1.716	1.812	1.903	1.966	2.047
Gen.	1.482	1.655	1.804	1.914	2.011
Gen. N50	2.04	2.115	2.218	2.306	2.381

Table 6.1 **Changing the breakpoint threshold.** We use the graph method on the thinned version of Chromosome 12 from the WTCCC data and consider the size of detected haplotype chunks as the number of allowed mismatches increases. The reported values are either medians or N50s. Phys. = physical distances (Mb); Gen. = genetic distance (cM).

Using our preferred stopping rule of 2 for the graph method, the median length of maximal haplotype sharing for sample is estimated to be 1.66 cM and the N50 to be 2.1cM (or 1.1 Mb and 1.6Mb respectively). The distribution of maximal shared haplotype length is highly skewed (Figure 6.4) and the 95th percentile is c. 4 cM. This indicates that some latent relatedness may be present in the sample. These results indicate that NNs within a cohort of even this size are likely to have a common ancestor over 100 generations ago. Note, however, over the entire Chromosome 12 it is highly likely that any one individual will have much closer common ancestors. Indeed, the distribution of the largest NN haplotype length across the genome, which has a median of 8 cM (Figure 6.4 and 6.5) suggests that the closest MRCA is typically approximately 57.5 generations ago. It will be interesting to compare this figure with estimates of TMRCA with a NN obtained by sampling positions randomly across the genome, which we consider in the next subsection.

The extent of haplotype sharing is much less than that expected under a model of constant population size (using the tree simulations described in Chapter 4, median maximal haplotype sharing is expected to be ~ 60 cM; Fig 6.6), reflecting explosive population growth. The strong deviation of our observed values for varying sample size, against our expectation from a constant-size model is further evidence of the fact that this a very strongly growing population. As

6.2 The structure of haplotype identity in a sample of over 5000 UK individuals

an aside, we also therefore witness a limited effect of sample size in the extent of haplotype sharing. Interestingly, the parameters of growth estimated from allele frequency spectrum data (namely a recent N_e of 4×10^{-6} with growth over 400 generations (Nelson et al. [2012])) also do not agree well. Rather, simulations from a demographic model in which the current effective population size (N_e) is approximately 4 million, having increased exponentially over the last 10,000 years from a population size of 10,000 indicate that the median NN haplotype length in a sample of 10,000 chromosomes should be 0.8cM, just over a third of the N50 of our observations. Instead, we obtain a better fit using a model which assumes that growth started about 110 generations ago (Figure 6.4).

The caveat when assessing these findings is that our graph method is in fact greatly susceptible to genotyping error and thinning. Indeed, when we compare our findings from 89 subsampled WTCCC individuals to those from the British 1000 Genomes individuals on Chromosome 12 we discover that there is a significant difference in the results. The median haplotype chunk N50 for the WTCCC samples is 0.7cM, while for the 1000 Genomes individuals is 0.23cM - a third of the size, indicating that of our algorithm breaks and restarts more frequently with the 1000 Genomes data. Of course the 1000G data is both 14-fold more dense, but further, given that it is sequence data, it is expected to contain a higher error rate than the WTCCC genotyped samples. Crucially, it is also richer in rare variation which refines the patterns of relatedness in the data.

We therefore formally investigate the effects of error and thinning separately. In this analysis we additionally consider the performance of HMMsub (H300 and H1000). Table 6.2 presents our findings from Chromosome 22 of the WTCCC samples, indicating a susceptibility of both the graph method and HMMsub to both parameters: haplotype sharing decreases with an increasing error rate, while thinning the data underpowers the methods from distinguishing smaller haplotype chunks, leading to an overestimation of the sharing that is present. The effects are more prominent for the N50, which is largely influenced by the longer haplotypes. Error rate disrupts these longer chunks, thus leading to a greater underestimation of the N50. Conversely, the lack of power to detect small chunks when the data is thinned results in the merging of adjacent chunks which inflates the N50.

6.2 The structure of haplotype identity in a sample of over 5000 UK individuals

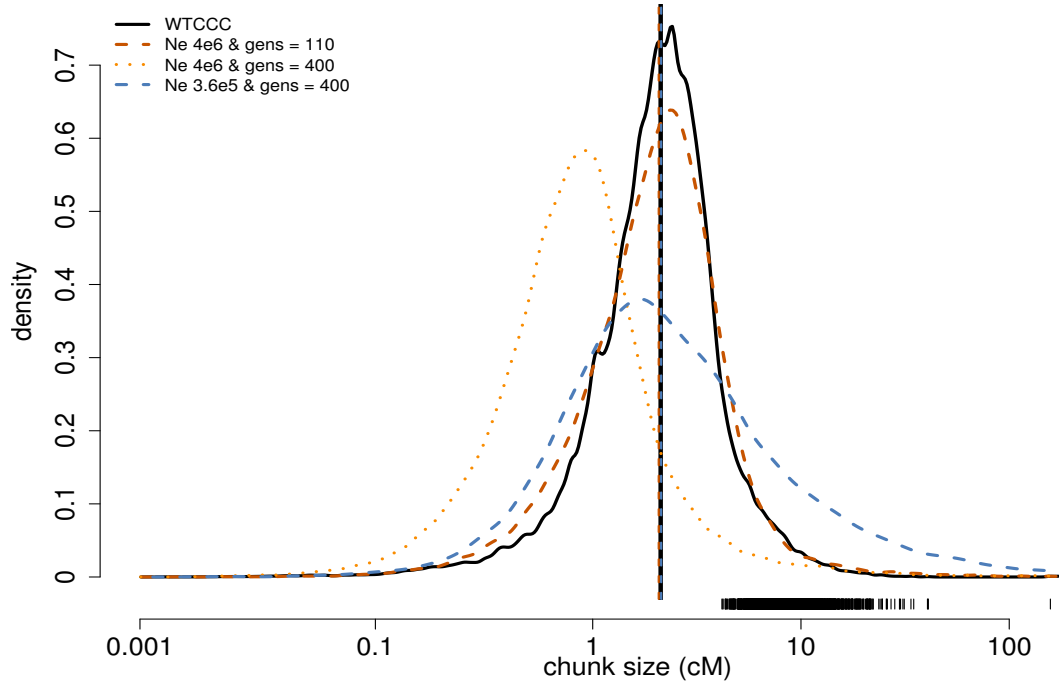


Figure 6.4 **Haplotype Sharing**. The distribution of the haplotype chunks from the WTCCC data (black). In dark orange: curve obtained from tree simulations, assuming exponential growth from $N_e = 10,000$ to $N_e = 4 \text{ million}$ over 110 generations. In light orange: curve obtained from tree simulations, assuming exponential growth from $N_e = 10,000$ to $N_e = 4 \text{ million}$ over 400 generations. In blue: curve obtained from tree simulations, assuming exponential growth from $N_e = 10,000$ to $N_e = 350,000$ over 400 generations. The vertical lines are median values for the tree simulations and the N50 from the WTCCC. The black tick marks on the bottom right mark the size of the largest haplotype chunk estimated for each individual.

Unsurprisingly, although both methods are influenced by error and thinning, HMMsub (both versions) is significantly more robust to errors than the graph method, thus confirming that it behaves more closely to what we would expect from the original, full HMM. Interestingly however, the effect of thinning has a similar impact on both methods, with a slightly greater inflation seen for the results from HMMsub. With an average density of a SNP every 2Kb, keeping only 30% of SNPs results in a vastly sparse data set and it is no surprise that there

6.2 The structure of haplotype identity in a sample of over 5000 UK individuals

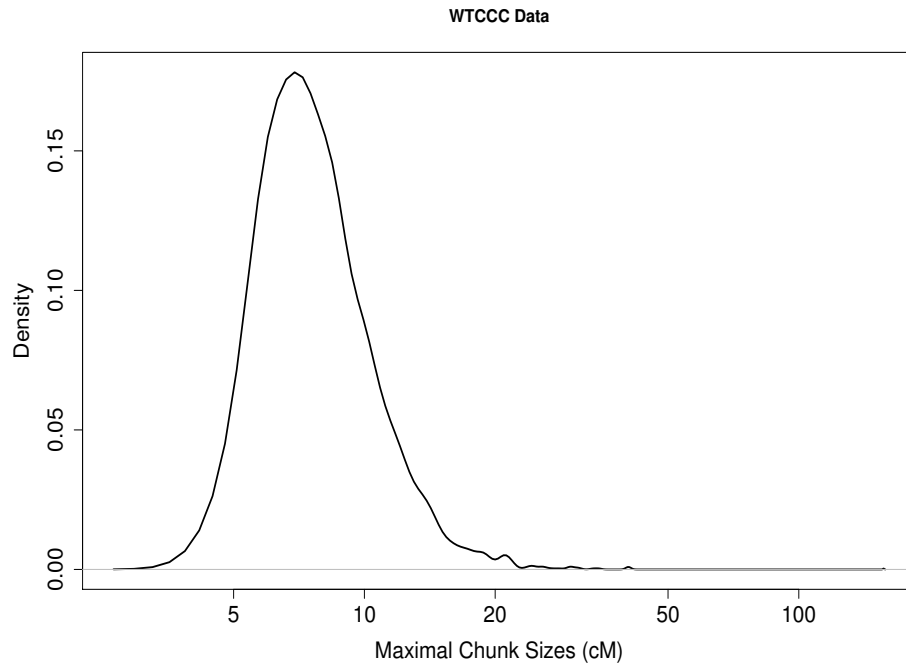


Figure 6.5 **Maximal Haplotype Sharing.** The distribution of the maximum haplotype chunks from the WTCCC data achieved for each individual. The median is 8 cM.

is a loss in power to discover shorter shared chunks with either method. Still, the fact that there is no improvement with HMMsub compared to the graph method further highlights the effect of our subsampling approach, which is very similar in both cases. Note that as in previous analyses, H1000 and H300 obtain near-identical estimates in all scenarios.

Finally, the trend observed from our thinning investigation suggest that the observations with the 1000 Genomes sequence data are solely the by-product of the genotyping error present in sequence data. The WTCCC contained only 10% of SNPs (though a much higher proportion of common variants), yet rather than increasing, the N50 was decreased by $\approx 75.5\%$, highlighting that error rate is most likely a greater contributing factor to shared haplotype estimation.

6.2 The structure of haplotype identity in a sample of over 5000 UK individuals

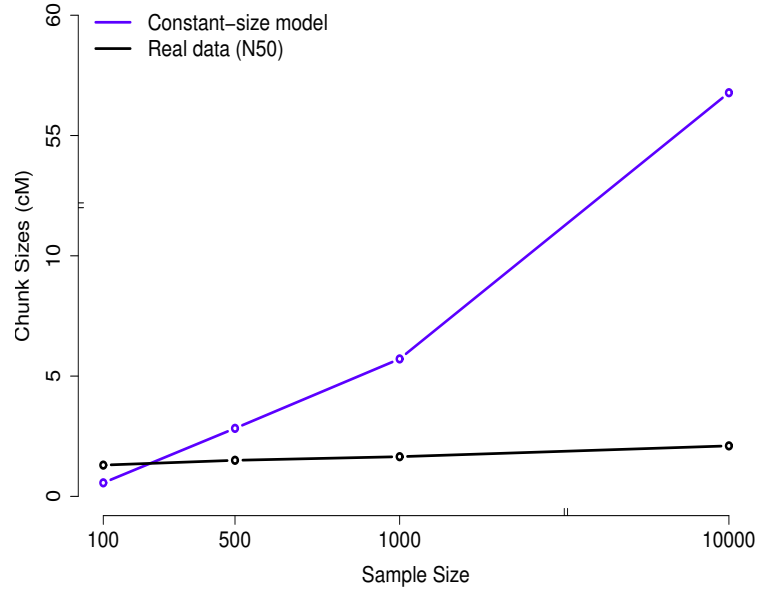


Figure 6.6 **Observed and expected haplotype chunks.** The expected values were obtained using our tree simulations for various sample sizes. The observed values were obtained by randomly subsampling individuals from the WTCCC (chromosome 12; thinned) data and running the graph method, with subsampling. We report the observed N50s as these correspond to the values obtained by simulation.

6.2.3 Genome-wide analysis: results using the graph and HMMsub

In light of these preliminary insights, we turn to our genome-wide analysis using both the graph method and the HMM with subsampling on the full data set. Studying the full distribution of haplotype chunks across all chromosomes (Figures 6.7 and 6.8; the full details for the findings for each method are tabulated in Tables 6.3 and 6.4, we make several key observations:

1. All methods exhibit consistency across chromosomes on the amount of haplotype sharing among individuals (see Figure 6.7 inset: the distribution is similar across all chromosomes, overlapping for both methods). The shape of the distribution is as the one observed for the thinned version of Chromosome 12 (Figure 6.4) and that of the TEENAGE individuals in the HELIC

6.2 The structure of haplotype identity in a sample of over 5000 UK individuals

Method	Graph		H300		H1000	
Metric	median	N50	median	N50	median	N50
Original	1.12	1.98	1.28	2.56	1.28	2.54
0.5% error	0.99	1.57	1.22	2.26	1.22	2.26
1% error	0.89	1.34	1.17	2.03	1.17	2.03
2% error	0.76	1.04	1.08	1.69	1.08	1.69
90% thinned	1.13	2.05	1.34	2.64	1.33	2.63
60% thinned	1.40	2.38	1.63	3.00	1.62	3.01
50% thinned	1.50	2.54	1.82	3.17	1.82	3.16
30% thinning	2.02	3.19	2.54	3.92	2.53	3.91

Table 6.2 **The effect of genotyping error and thinning on the graph method and HMMsub.** The table contains results when an error rate of $y\%$ ($y \in \{0.5, 1, 2\}$) is randomly applied in the data. Then, using the original full Chromosome 22, we randomly subsample SNPs in order to thin the data by leaving $x\%$ of SNPs ($x \in \{30, 50, 60, 90\}$). In each case we use the graph method and HMMsub with 300 (H300) and 1000 (H1000) SNPs for subsampling to examine the changes in the detected haplotype chunks. The reported values are either medians or N50s of the genetic chunk size across all chunks from all individuals, measured in cM.

study (Figure 5.5, top panel, right). We therefore observe the signature of a homogeneous, unrelated sample that has undergone exponential growth.

2. We observe a decrease in the physical length of haplotype chunks with decreasing chromosomal length (Tables 6.3 and 6.4). This is expected as there is an inverse relationship between recombination rate and chromosome length. There is one crossover event per chromosome per meiosis (Kauppi et al. [2004]) meaning that for larger chromosomes the candidate region is spread over greater distances, decreasing the per base recombination rate. Since our haplotype chunks are defined by a recombination driven process we expect to see a reduction in the chunk sizes for shorter chromosomes, but no changes in genetic length.

6.2 The structure of haplotype identity in a sample of over 5000 UK individuals

3. The genome-wide N50 obtained using the graph method is 1.64cM (across all chunks identified for all individuals; 1.79 obtaining medians per person and then a sample mean), down from the 2.1cM that was found using the thinned Chromosome 12 (which was thinned by $\sim 50\%$). As before, this highlights the susceptibility of the graph method to thinning. There is a 20% reduction here, which is what we expect (Table 6.2). Despite the higher SNP density, which enriches variation, thus affecting chunk size, the strong growth rate of the population would suggest that the NN changes should happen at a larger scale than the increase in density here. It is more likely therefore that this reduction is a by-product of our method's sensitivity to errors, which would now be more numerous.
4. The results from HMMsub are consistently greater than those from the graph method (Figures 6.7 and 6.8), as expected in light of our findings from the 1000 Genomes omni genotype data. This suggests that we may use these results as a closer estimate of the true levels of sharing present in the sample. HMMsub's greater robustness to error (Table 6.2) and its tendency to tolerate incompatibilities when it selects pseudo-parents suggest (Figure 6.1) that this is a better estimate.
5. Both versions of HMMsub achieve very similar results. The N50 calculated from all chunks obtained by either method is 2.1cM (or 2.27cM per person).

To match an N50 of 1.64cM (estimated using the graph method), our simulations suggest that the growth parameters that would now be most consistent are a current N_e of 4 million, over 140 generations (Figure 6.7) — only 40 generations earlier than our previous estimate (Figure 6.4). Conversely, the N50 of 2.1cM from the HMM with subsampling suggests that the growth period should be 110 generations, with the current N_e being consistently kept at 4 million. We may thus estimate the growth period to be at most 140 generations, though probably closer to 110 generations.

We should finally note that in our current simulations we are considering the effect of growth alone, ignoring bottlenecks. The events that we are examining are recent and it is therefore not obvious that this is a significant omission. Still,

6.2 The structure of haplotype identity in a sample of over 5000 UK individuals

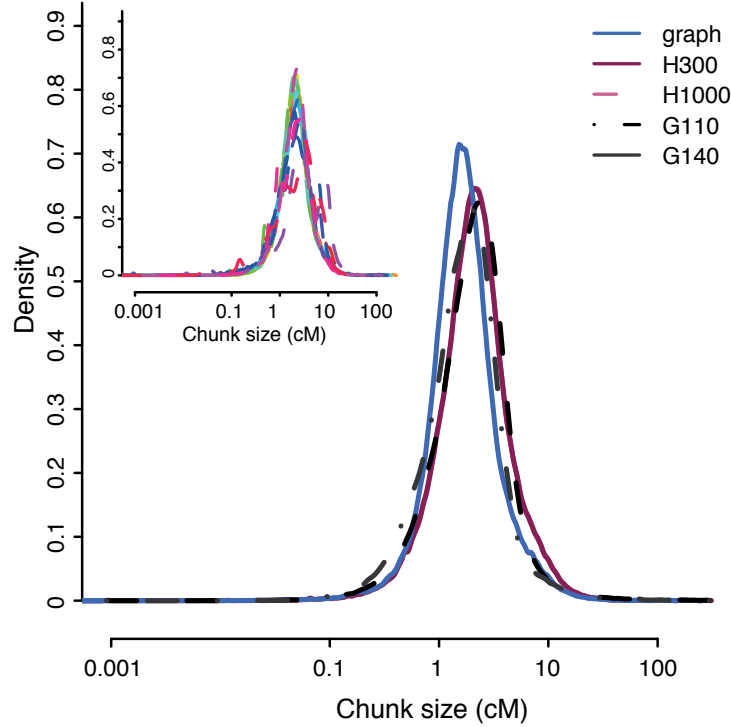


Figure 6.7 **Haplotype Sharing across chromosomes.** The distribution of the haplotype chunks from the WTCCC data. The main figure shows the results by aggregating all chunks detected across the genome, both for the graph method (blue; labelled graph) and HMM with subsampling (purple; labelled HMMsub). In dashed black and dark grey are the curves obtained from our tree simulations with a model that assumed exponential growth from an effective population size of 10,000 to 4 million over a period of G generations. The inset contains the individual density curves for each chromosome for both methods (graph and HMMsub); they are solely intended to illustrate that there is in general genome-wide consensus from both methods on the distribution of chunk sizes and that aggregating results across chromosomes is reasonable. Equivalent chromosomes have been coloured in the same way for both methods, using rainbow colours: red for Chromosome 1.

perhaps a model that accounts for the out of Africa bottleneck and the Eurasian split, both of which are believed to have significantly reduced variation in the population, would require the more extended growth period of 400 generations suggested by Nelson et al. [2012] to restore variation to today's observed levels.

6.2 The structure of haplotype identity in a sample of over 5000 UK individuals

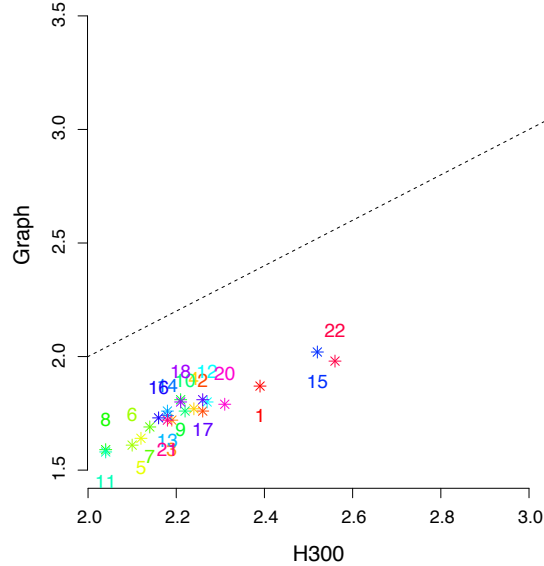


Figure 6.8 **Comparing the per chromosome N50 for H300 and the graph method.** The figure shows the N50 for each chromosome (each point is illustrated by a star that is coloured differently and is labelled with the same colour) for the HMM with subsampling with 300 SNPs (H300; x-axis) against the N50 obtained using the graph method (y-axis). The N50s were calculated by aggregating all chunks across all individuals. In dotted black is the $y = x$ line. HMMsub with 1000 SNPs produced near-identical results as H300.

Nevertheless, our assertion of recent and rapid exponential growth is in agreement with our findings for the TMRCA between NNs (Figure 6.9). By sampling 100 randomly selected SNPs across the genome and following the approach adopted in the HELIC project, we obtain a mean TMRCA of 79.7 (median 53.8) generations using the graph method and 65.4 generations using the HMMsub versions (the medians are ~ 45.5), which are slightly more recent than the estimated growth periods, consistent with the rapid growth hypothesis. The values here are means across all the SNPs used. The distribution of the TMRCA for both cases also matches that observed for the TEENAGE sample, though with a significantly longer tail (the 99% quantile is 296 generations for the graph method and 208.5 for the HMMsub versions), further confirming that this is a homogeneous, completely unrelated collection of individuals. The mean values here are only up

6.2 The structure of haplotype identity in a sample of over 5000 UK individuals

to a factor of 2 less than that calculated for the TEENAGE cohort, despite the fact that we now have just under a 10-fold increase in sample size. This further supports the argument of rapid exponential growth. Finally, it is still interesting to observe that for a sample size as large as 5000 samples, one needs to go back $\sim 1,500$ years in order to find a common ancestor between NNs (or more likely, genealogical cousins).

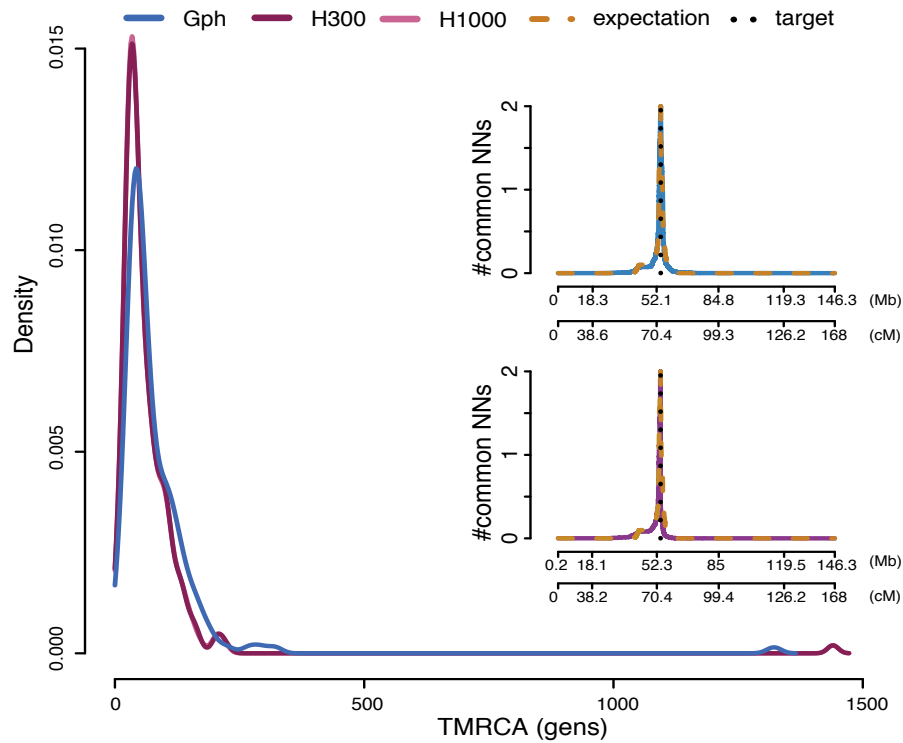


Figure 6.9 **The TMRCA to a NN.** The main figure shows the distribution of the TMRCA to a NN by sampling 100 points randomly across the genome and measuring the decay of haplotype sharing around them (refer to text for details); this is the same approach as that adopted in the HELIC project in Chapter 5. The three density curves are for the graph method (blue), HMMsub with 300 SNPs (H300; magenta) and HMMsub with 1000 SNPs (pink; does not show). The two insets illustrate the process for one such SNP, rs1425894 on chromosome 8; top = results from the graph method; bottom = results from H300.

Finally, we investigated how haplotype chunk size (physical and genetic) changed across chromosomes for our two methods. The chunks are generally

6.2 The structure of haplotype identity in a sample of over 5000 UK individuals

uniform across sites (Figures 6.10, 6.11 and 6.12; only physical lengths shown here), with major spikes occurring for physical length, as expected, primarily around the centromeres. Still, there is local variation, for instance an increase in physical lengths around the HLA on Chromosome 6. Encouragingly, this variation in sharing is consistently observed with both of the methods that we have used, with results being remarkably similar genome-wide. There are other ‘suspicious’ spikes, for example at 50Mb along Chromosome 3, around 120Mb for Chromosome 7 and 70Mb on Chromosome 16. Overall, there are non-uniform fluctuations in haplotype sharing across the genome suggesting that it is worth investigating whether small increases occur over ‘interesting’ regions and therefore promoting this approach as an alternative to IBD mapping.

6.2 The structure of haplotype identity in a sample of over 5000 UK individuals

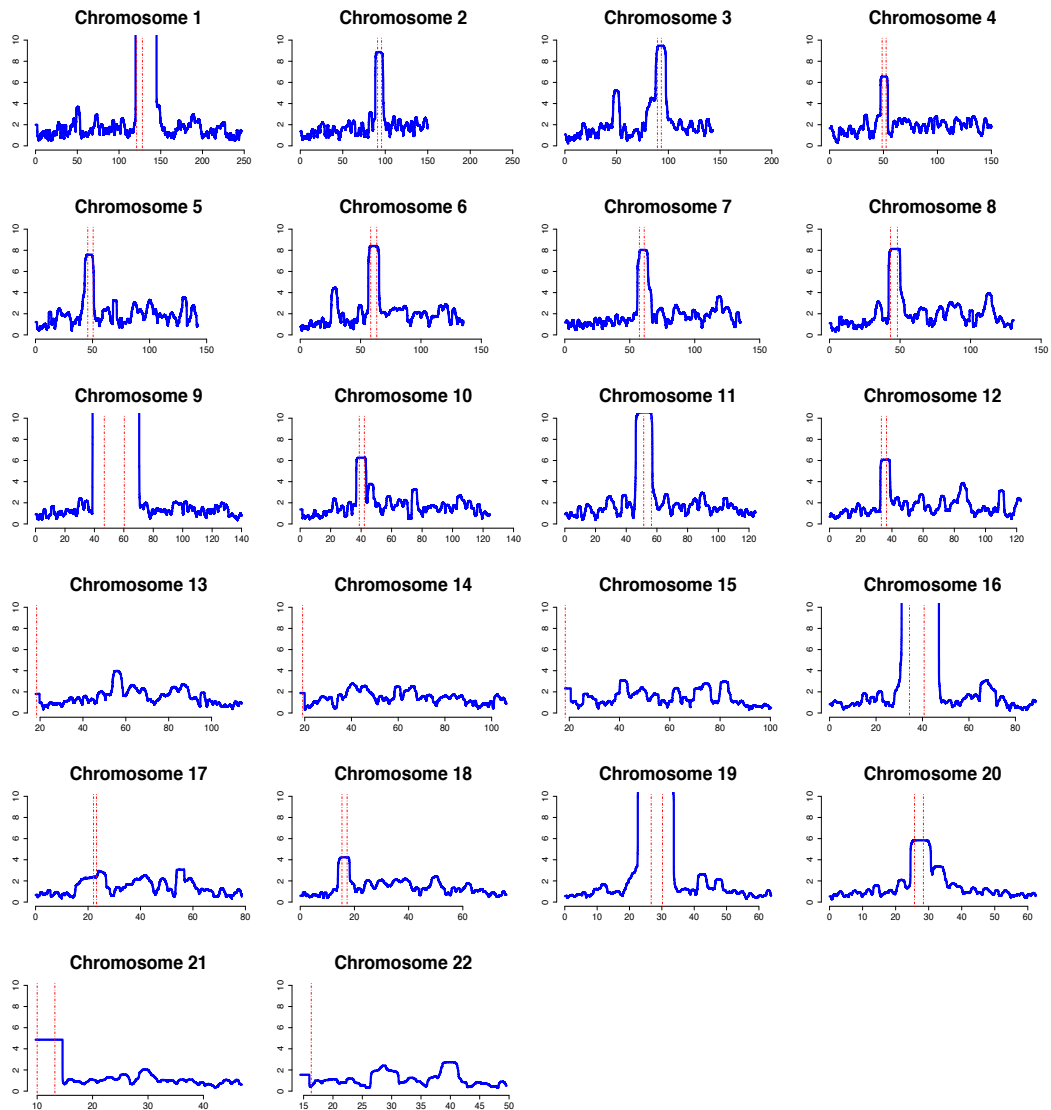


Figure 6.10 **Physical haplotype chunk sizes point-wise across chromosomes using the graph method.** The red dotted lines outline the centromere. For each individual at each SNP we calculate the size of the chunk that the SNP falls in. The values shown are per-base medians across individuals. The y-axis is chunk size in Mb; we have imposed an upper limit of 10Mb and only the chunks around the centromeres surpass it (which is an artefact). The x-axis shows position along each chromosome (Mb).

6.2 The structure of haplotype identity in a sample of over 5000 UK individuals

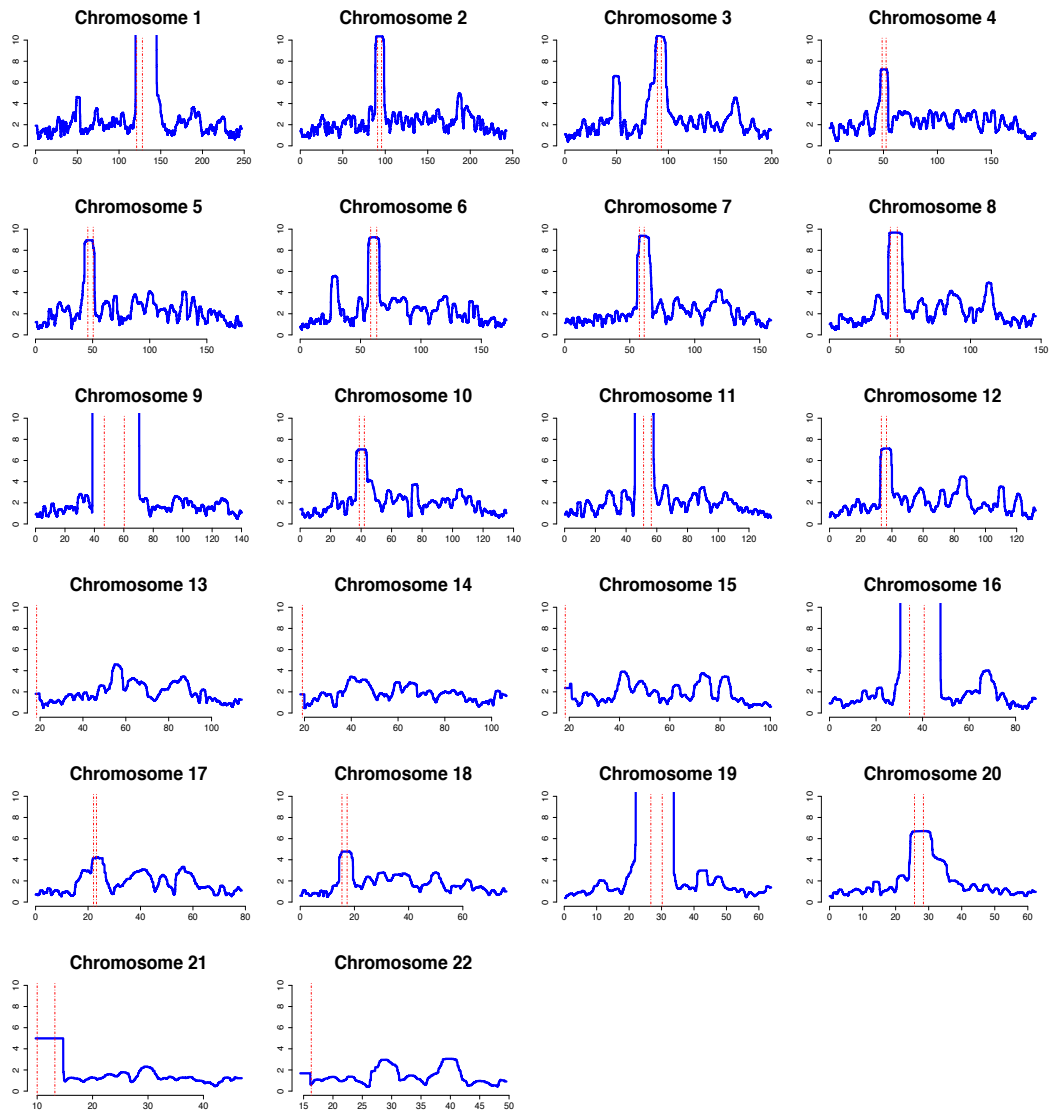


Figure 6.11 **Physical haplotype chunk sizes point-wise across chromosomes using HMMsub with 300 SNPs.** The red dotted lines outline the centromere. For each individual at each SNP we calculate the size of the chunk that the SNP falls in. The values shown are per-base medians across individuals. The y-axis is chunk size in Mb; we have imposed an upper limit of 10Mb and only the chunks around the centromeres surpass it (which is an artefact). The x-axis shows position along each chromosome (Mb).

6.2 The structure of haplotype identity in a sample of over 5000 UK individuals

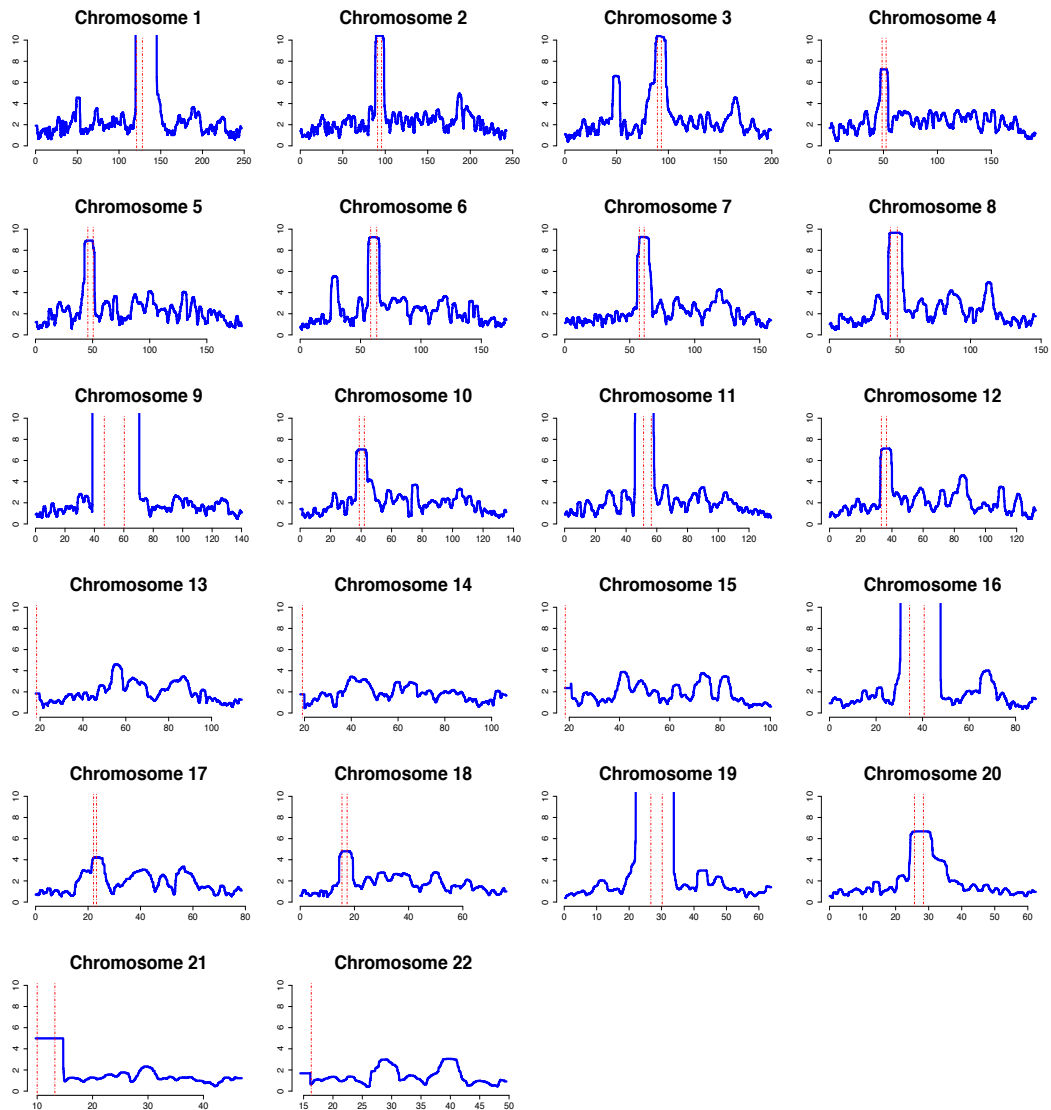


Figure 6.12 **Physical haplotype chunk sizes point-wise across chromosomes using HMMsub with 1000 SNPs.** The red dotted lines outline the centromere. For each individual at each SNP we calculate the size of the chunk that the SNP falls in. The values shown are per-base medians across individuals. The y-axis is chunk size in Mb; we have imposed an upper limit of 10Mb and only the chunks around the centromeres surpass it (which is an artefact). The x-axis shows position along each chromosome (Mb).

6.2 The structure of haplotype identity in a sample of over 5000 UK individuals

Chr	SNPs	L(Mb)	L(cM)	Ph.(N50)	Gen.(N50)
1	84135	249.22	286.28	1.73	1.87
2	82978	243.09	268.84	1.83	1.76
3	68397	197.87	223.36	1.74	1.72
4	60317	191.03	214.69	1.89	1.77
5	61892	180.72	204.09	1.76	1.64
6	64435	171.05	192.04	1.71	1.61
7	54926	159.13	187.22	1.65	1.69
8	52997	146.3	168	1.70	1.59
9	45288	141.11	166.36	1.72	1.81
10	51623	135.5	181.14	1.59	1.76
11	50206	134.95	158.22	1.61	1.58
12	49753	133.78	174.68	1.55	1.80
13	36272	115.11	125.71	1.52	1.76
14	33164	107.29	120.2	1.50	1.74
15	30538	102.51	141.86	1.43	2.02
16	31980	90.16	134.04	1.47	1.73
17	29826	81.05	128.49	1.39	1.81
18	28751	78.02	117.71	1.37	1.80
19	22729	59.1	107.73	1.27	2.33
20	25346	62.95	108.27	1.17	1.79
21	13840	48.1	62.79	1.17	1.72
22	16031	51.23	74.11	1.09	1.98

Table 6.3 **Full results obtained using the graph method.** Chr = chromosome; SNPs = the number of SNPs on the chromosomes; L(cM) = the genetic length of the chromosome; L(Mb) = the physical length of the chromosome; Ph. = physical length of the detected shared haplotype chunks; Gen. = the genetic length of the haplotype chunks. The results for the chunk sizes shown here are means over all the values calculated per person. These (median and N50) are expected to be different to the median and N50 values obtained by aggregating all chunks across all individuals for each chromosome.

6.2 The structure of haplotype identity in a sample of over 5000 UK individuals

Chr	SNPs	L(Mb)	L(cM)	Ph.(N50)	Gen.(N50)
1	84135	249.22	286.28	2.18	2.39
2	82978	243.09	268.84	2.32	2.26
3	68397	197.87	223.36	2.22	2.19
4	60317	191.03	214.69	2.34	2.24
5	61892	180.72	204.09	2.24	2.12
6	64435	171.05	192.04	2.18	2.10
7	54926	159.13	187.22	2.07	2.14
8	52997	146.3	168	2.13	2.04
9	45288	141.11	166.36	2.16	2.21
10	51623	135.5	181.14	1.96	2.22
11	50206	134.95	158.22	2.04	2.04
12	49753	133.78	174.68	1.97	2.27
13	36272	115.11	125.71	1.84	2.18
14	33164	107.29	120.2	1.83	2.18
15	30538	102.51	141.86	1.79	2.52
16	31980	90.16	134.04	1.86	2.16
17	29826	81.05	128.49	1.70	2.26
18	28751	78.02	117.71	1.69	2.21
19	22729	59.1	107.73	1.72	3.07
20	25346	62.95	108.27	1.51	2.31
21	13840	48.1	62.79	1.49	2.18
22	16031	51.23	74.11	1.38	2.56

Table 6.4 **Full results obtained using HMMsub with 300 SNPs.** Chr = chromosome; SNPs = the number of SNPs on the chromosomes; L(cM) = the genetic length of the chromosome; L(Mb) = the physical length of the chromosome; Ph. = physical length of the detected shared haplotype chunks; Gen. = the genetic length of the haplotype chunks. The results for the chunk sizes shown here are means over all the values calculated per person. These (median and N50) are expected to be different to the median and N50 values obtained by aggregating all chunks across all individuals for each chromosome.

6.2 The structure of haplotype identity in a sample of over 5000 UK individuals

Chr	SNPs	L(Mb)	L(cM)	Ph.(N50)	Gen.(N50)
1	84135	249.22	286.28	2.19	2.41
2	82978	243.09	268.84	2.32	2.26
3	68397	197.87	223.36	2.22	2.19
4	60317	191.03	214.69	2.35	2.24
5	61892	180.72	204.09	2.24	2.13
6	64435	171.05	192.04	2.20	2.12
7	54926	159.13	187.22	2.06	2.14
8	52997	146.3	168	2.16	2.07
9	45288	141.11	166.36	2.16	2.23
10	51623	135.5	181.14	1.99	2.24
11	50206	134.95	158.22	2.04	2.04
12	49753	133.78	174.68	1.97	2.28
13	36272	115.11	125.71	1.84	2.18
14	33164	107.29	120.2	1.84	2.19
15	30538	102.51	141.86	1.79	2.52
16	31980	90.16	134.04	1.85	2.15
17	29826	81.05	128.49	1.72	2.28
18	28751	78.02	117.71	1.71	2.23
19	22729	59.1	107.73	1.73	3.11
20	25346	62.95	108.27	1.50	2.29
21	13840	48.1	62.79	1.49	2.18
22	16031	51.23	74.11	1.37	2.54

Table 6.5 **Full results obtained using HMMsub with 1000 SNPs.** Chr = chromosome; SNPs = the number of SNPs on the chromosomes; L(cM) = the genetic length of the chromosome; L(Mb) = the physical length of the chromosome; Ph. = physical length of the detected shared haplotype chunks; Gen. = the genetic length of the haplotype chunks. The results for the chunk sizes shown here are means over all the values calculated per person. These (median and N50) are expected to be different to the median and N50 values obtained by aggregating all chunks across all individuals for each chromosome.

Chapter 7

Discussion

In this thesis we have presented the method that we have developed for detecting and measuring coancestry among unrelated individuals. Our method, inspired by the work of Kong et al. [2008] and ChromoPainter (Lawson et al. [2012]), uses a Hidden Markov Model and aims to characterize recent relatedness by searching for pairs of individuals at each SNP that are most closely related to a query individual. Following Kong et al. and Lawson et al. we call these individuals Nearest Neighbours (NNs) or pseudo-parents. Our HMM is genotype-based and therefore avoids the need for initial haplotype estimation. The hidden states are all the pairs of individuals present in the sample, and each query individual constitutes the observed states. Transition probabilities model recombination between adjacent SNPs and for the purposes of detecting coancestry we adopt an emission model that looks at the compatibility between each candidate pair of individuals and the query genotype at each SNP. Our emissions do not incorporate a mutation model but rather have a fixed probability for the occurrence of incompatibilities that reflects the genotyping error that might be present in the data. By studying the NN choices made by individuals in a sample we may make inference on stratification patterns and on the level of coancestry that is present. Using NN knowledge we can additionally calculate the extent of haplotype sharing between them. Haplotype sharing offers insight on the levels of relatedness

within a sample, as it is correlated with the TMRCA between NNs and may further provide clues regarding the genomic location of mutations associated with disease risk.

Aside from the practical applications of our approach, we hope that our model will add to the discussion of how best to define and quantify relatedness. We have used our method to explore the various ways in which individuals relate to each other. Our simulations demonstrated that there is a difference between Nearest Neighbours (NNs) and individuals between which there is maximal haplotype sharing; in fact, we discovered that in a population where N_e is constant, maximal haplotype sharing occurs with a NN only at $\approx 80\%$ of sites. This upper limit is significantly reduced when the sample under study has experienced exponential growth. In light of these observations, we showed that our method detects close to the expected maximal haplotype sharing by identifying NNs (though lower estimates are expected due to the presence of mutations), contrasting the performance of a haplotype-based approach, regardless of whether the haplotypes were real or estimated. When the true NNs were not detected, the alternative individuals that were selected had TMRCA that were similar to what we would expect. What is more, our genotype-based method was significantly more robust to moderate levels of genotyping errors. We have seen how phasing errors lead to an underestimation of relatedness. With real data switch errors on average occur every Mb, which is shorter than the haplotype chunks that should be detected to infer recent coancestry, an effect that has been observed in the literature (for example Browning and Browning [2012]), but which has not previously been quantified. This is not to say, of course, that with the evident improvement of phasing methods this issue will persist in the future.

The principle limitation of our HMM is the fact that its computational speed scales quadratically with the sample size, making it intractable even for a moderate size of a few hundred individuals. With modern data sets containing a few thousand individuals and with an outlook of sequencing hundreds of thousands more (if not millions), our HMM is an inadequate approach for analysing samples. For this reason we developed two alternative methods for approximating the original model: the graph method and HMMsub. Both are parallelisable over

individuals, similar to the HMM, therefore allowing for the computation time for the entire sample to be comparable to that of a single individual.

The graph method was developed first. In this approach, a genotype sample is represented in the form of a suffix tree. By considering pairs of nodes on the tree rather than pairs of individuals, we limit the size of the state space considered at each SNP. We further reduce the number of calculations by introducing a sub-sampling scheme in which samples are excluded in advance from the suffix tree if they are incompatible with the query genotype at too many positions. The graph method is faster than the original HMM by several orders of magnitude, making it an appropriate algorithm for analysing large samples. For a chromosome of size 30,000 SNPs the graph method requires ≈ 3 minutes per person using a sample of 5,000 individuals. While the method recovers the expected amount of sharing within a sample, it is less successful at detecting NNs. Rather, it detects slightly more distantly related individuals. Moreover, the graph method is susceptible to genotyping errors, though its performance is generally at least as efficient as haplotype-based approaches. Finally, the graph method has the inherent limitation of being directed, with results expected to be slightly different were the method to be applied right-to-left than left-to-right.

We hoped to bridge some of the graph method's computational advantages with the more effective HMM framework. As an alternative to the graph method, we therefore more recently developed HMMsub: our original HMM with a sub-sampling approach very similar to that adopted in the graph method. HMMsub is not as fast as the graph method, nor very memory efficient, but it remains computationally efficient so that it is tractable on large sample sizes. For a chromosome of size 30,000 SNPs, HMMsub requires about half an hour per person for the same sample of 5,000 individuals. This method proved to be more robust to errors than the graph method and in simulations and the small 1000 Genomes omni genotype sample the approach obtains haplotype estimates very close to those obtained by the full HMM. Moreover, these observations stand regardless (within reason) of the number of SNPs used for subsampling (though this might not hold under different demographic scenarios).

It is clear, therefore, that while the original HMM has very useful properties and achieves the desirable accuracy both in terms of finding closely related

individuals and recovering the sharing between them, the question of developing a more computationally and memory efficient version of the method remains open. Nevertheless, from the results that we have collected using our current methods on both simulated and real data we remain confident about the validity of our findings. Looking ahead, however, we would recommend the development of alternative methodologies that capture the properties of the original HMM. These might be in the form of extensions of the currently suggested approaches (for example by improving the subsampling and implementation of HMMsub or experimenting with the graph method’s stopping rule). In our view the critical criteria are that a new method be robust to errors in order to accurately estimate sharing. It is also advantageous to be able to detect more closely or more distantly related pseudo-parents by a change in parameterisation, similar to our full HMM.

Interestingly, we discovered that our full HMM achieves this with a change the emission matrix. By changing the emission matrix in our HMM from compatibility-based to a weighted Mendelian scheme, rather than detecting NNs we can climb higher up in the genealogical tree of the sample and search for individuals that are suitable for phasing.

The remainder of this discussion is dedicated to consolidating our insights on phasing and coancestry characterisation, as well as addressing some implications of our work. We end with possible applications of our approach in investigating disease risk.

7.1 A few final remarks on phasing

Using the weighted Mendelian emissions, our HMM (which we call PACE) proved to be a competitive phaser, with similar performance to other cutting-edge methods. Because of timing, we did not compare PACE to SHAPEIT Delaneau et al. [2012], which we anticipate is a superior phasing algorithm. However, we could make a case that an approach like ours could be used to construct the Scaffold that SHAPEIT requires to initialise. For example, when sample sizes are small we have demonstrated that our method performs more competitively than BEAGLE.

7.1 A few final remarks on phasing

We have also not explored additional ways of implementing our method to phasing. For example, one might use knowledge of NNs to resolve ambiguity around triple heterozygote sites. We could alternatively adopt Kong et al.’s approach of phasing pseudo-parents as well and using knowledge of their haplotypes to resolve triple heterozygotes. Both of these suggestions imply incorporating a post-analysis strategy that ignores unphased sites and then uses new knowledge to resolve ambiguity at those sites. An alternative implementation of our HMM to phasing could have been via the Forward/Backward algorithms, which we have shown are effective at detecting coancestry. In this scenario we would need to develop a mechanism for joining the phase estimates that the most likely pseudo-parents make at each individual site.

Finally, we have not explored the graph method and HMMsub as potential haplotype estimation methods. These latter methods were developed more recently and are designed to detect recent relatedness, using the compatibility emission system. In their current implementation they are therefore unsuitable for phasing, as we expect them to leave a very large number of sites unphased.

Overall, our findings revealed that accuracy is not correlated with path likelihood. We should clarify, however, that our results and conclusions for phasing apply in the setting where there is a single homogeneous population, without any structure or admixture. If the latter scenarios were the case, one can imagine that suboptimality might in fact influence performance. To illustrate this, consider the simpler case of a population which is a mixture of two sub-populations, A and B . If a query individual comes from population A , and the number of pairs of pseudo-parents that are purely from A is $< K$ (the number of suboptimal paths that we are collecting), then the algorithm will be forced to select paths that include individuals from population B . It is not obvious that in this case these latter suboptimal paths would be as suitable for phasing as our findings suggest. In fact, if the query genotype is the only individual of type A , there is no way of determining in advance whether even a single optimal path is in fact a good fit.

This highlights a weakness of the method: we do not have a measure of ‘goodness-of-fit’ for phasing for each of the Viterbi paths. That is to say that we have not yet discovered a metric that indicates how well a certain path of pseudo-parents is expected to phase a query genotype. More generally, we do not

have a way to objectively measure the suitability of a pair of individuals as NNs or pseudo-parents outside the F/B framework. In studies involving less homogeneous populations, this would be essential in order to ensure that expectations regarding accuracy and ‘goodness-of-fit’ are well managed.

7.2 A few final remarks on relatedness in populations

Returning to the detection of coancestry, we have discussed how NN information might be used in the spatial analysis of genetic variation. Even if common variation shows little differentiation across a region, both theoretical (Mathieson and McVean [2012]) and empirical (1000 Genomes Project Consortium et al. [2012]) analyses show that rare variants can have a highly localised distribution. If so, it should follow that the NNs detected using rare variant information are also highly localised. Whether there is a substantial advantage in avoiding phasing remains to be seen. It is clear, however, that where estimates of NN haplotype length are important in the analysis of population history (for example, in dating events), there is a clear advantage in avoiding even low levels of switch errors.

We have employed the approximations to our full HMM to determine the distribution of extended haplotype identity within a large sample of unrelated individuals and used our findings to characterise coancestry for two real data studies. In the HELIC project, we used the graph method to compare two different isolated populations to their ancestral, general sample of unrelated individuals. Our inference on haplotype sharing highlighted the special nature of the two isolates, revealing their high levels of relatedness and very recent coancestry. For example, haplotype sharing within the isolates is 10-fold greater than within the TEENAGE cohort. We further discovered that the TMRCA between NNs from the unrelated sample is approximately 90 generations ago, while within either of the isolates it is as recent as 10. By detecting NN clustering, with each cohort overwhelmingly selecting NNs from within the same cohort, we illustrated this additional feature of our method that can be used to understand relatedness and

7.2 A few final remarks on relatedness in populations

stratification patterns in a cohort. Finally, we used our results to make inference on migration and on the TMRCA between NNs. We showed that while our method does not predict the TMRCA between NNs at specific loci accurately, it does effectively capture the overall properties of the distribution of the TMRCA. We further consistently discovered greater levels of isolatedness among the Pomak individuals, compared to MANOLIS. We finally proposed a novel method for dating separation, though the accuracy of this approach has not been validated by simulation. A more sophisticated investigation on demographic history, migration and ancestry would require a more bespoke modelling approach which we believe falls outside the scope of the work necessary for the completion of this thesis.

Nevertheless, we can be confident about the main conclusions emerging from our analysis. Our findings were consistent with the results from all other methods used in the HELIC publications, with the only deviations occurring in the dating of events. Our time estimates are consistently lower than those obtained from other methods (for example using McEvoy et al. [2011]). Aside from the inherent differences between the approaches and the fact that estimates of the time of demographic events always involve large uncertainties (which we think might justify the differences seen here), we finally note that our estimates are not point estimates of the time of the event. Rather, since we are detecting the most recent coancestry, we are only estimating bounds on those times.

More generally, the results obtained in the HELIC project illustrate how a method like ours could be applied in a setting where ancestry or stratification patterns are unknown. It may equivalently be used to study how the joint inference of NN status and the extent of haplotype sharing correlate with varying levels of isolatedness or certain types of demographic history.

Contrasting the HELIC analysis, we used the graph method and HMMsub to quantify the distribution of extended haplotype identity within the sample of $\sim 5,000$ unrelated UK individuals from the WTCCC study (Wellcome Trust Case Control Consortium et al. [2007]). We demonstrated our methods' applicability to a large, modern data set in which the classic IBD approach could fail. We recover extended sharing of up to $\sim 10\text{cM}$ but also detect small chunks of $< 0.5\text{cM}$ and estimate an average TMRCA between NNs to be as recent as ~ 50 generations

7.2 A few final remarks on relatedness in populations

ago. Though not discussed here, we did look for NN clustering in the WTCCC data and found none, supporting further the evidence in the literature for lack of stratification in this cohort.

In addition, we used tree simulations from the coalescent to estimate the extent of exponential growth experienced by the UK population. In light of the limitations of the graph method and HMMsub, we bounded the growth period between 110 and 140 generations, rather than obtaining a single point estimate. Still, this remains a rather tight bound, clearly contrasting the alternative estimate of 400 generations suggested by Nelson et al. [2012]. Whether our haplotype-based approach is more accurate over their approach that uses a site frequency spectrum remains to be seen. It would also be interesting, though this is not attempted here, to compare our approach to detecting haplotype sharing to one which identifies shared haplotype chunks using sharing of rare variants, both from genotypes or from haplotypes. Such work has been recently undertaken by Mathieson and McVean [2014]. For European samples they date recent coancestry between 50-160 generations (TMRCA between individuals carrying doubletons), in agreement with our observations here. Our results are also consistent with the average TMRCA of ~ 120 generations between unrelated Europeans that was calculated by Ralph and Coop [2013].

In all, we hope to have demonstrated the usefulness of our approach for making inference both on recent coancestry and on historical demographic events. In our real data analyses, we recovered the expected properties and characteristics of the populations under study, highlighting the advantages of our method in cases where ancestry is unknown or where haplotype errors might be especially damaging. Our results indicate our method's applicability to studies where geographical stratification will play a greater role, such as the POBI project (Winney et al. [2012]). In the same spirit, it would be of interest to explore changes in haplotype sharing between populations, for instance by comparing findings for the different populations in the 1000G data. Would we, for example, observe less sharing for the Yoruba population as expected? What is the TMRCA between NNs in each cohort and how could we use the TMRCA across cohorts to date migration? Similarly, in such projects where geographical information is known,

7.2 A few final remarks on relatedness in populations

one can also investigate how shared chunk size correlates with geographical distance. Could our NN estimates be used to estimate time since divergence between different (sub-)clusters? Could inference of haplotype sharing be similarly employed to distinguish between cluster-specific demographic histories, perhaps as an alternative to methods like Harris and Nielsen [2013], or as a test of whether these shared chunks can indeed be used as an indicator of relatedness? These approaches are inspired by our most recent work with the HELIC data, in which we employed a crude, first-principles method for estimating the time of divergence between the isolates and the TEENAGE cohort. As mentioned, a more sophisticated modelling strategy should be developed, or a more dedicated simulation study should be undertaken to understand the efficacy and the limitations of this implementation of our method. We therefore do not expect our method to be able to fully address the questions raised here, but suggest that it could be used to complement or bound estimates obtained by other methods (for example Harris and Nielsen [2013], Mathieson and McVean [2014]).

A more immediate extension of our current implementation is the construction of a coancestry matrix—the genotype equivalent of the matrix obtained by ChromoPainter (Lawson et al. [2012]). It would be interesting to determine whether our genotype-based matrix is a good summary of global relatedness between individuals. One would be tempted to apply the post-analysis of fineStructure (Lawson et al. [2012]) to infer whether such a matrix correlates well with stratification patterns in a sample. Seeing as our method focuses on local patterns of coancestry, it is not obvious whether it is best constructed by averaging the ‘time-spent copying’ individuals (in the same fashion as ChromoPainter). Perhaps an approach which highlights locally shared chunks between individuals (who might not copy each other elsewhere) is more suitable. We note that the non-symmetrical nature of choosing NNs/pseudo-parents using our algorithm might add a layer of complication to the interpretability of this matrix.

We further highlight the fact that a more intuitive approach for building a coancestry matrix is to make use of posterior probabilities obtained using the Forward/Backward algorithms, rather than those obtained by the Viterbi approach. We remind the reader that our initial results using the Forward/Backward approach were promising, capturing coancestry between multiple individuals (not

just pairs) and thus offering a better representation of relatedness among samples. Assuming we had at hand an efficient implementation of our HMM that allowed for the Forward/Backward algorithms to be computed for large samples, posterior probabilities would be obtained at each SNP and we would still need to investigate the most effective way for summarising them over a region or a whole chromosome.

7.3 Outlook: Disease Risk

We end this discussion with a few remarks on the potential applications of our method to the analysis of disease risk. We have already discussed what is well known regarding the importance of understanding stratification patterns within cohorts, particularly in the case of rare variation. We envisage multiple ways by which our method directly lends itself to association studies, making use of the two main features that we calculate: pseudo-parent/NN status and haplotype sharing.

We have demonstrated already the use of haplotype sharing as an alternative to IBD mapping in our analysis of the HELIC and WTCCC data. Excess sharing among cases is likely to be detectable only in cases of highly penetrative variants or with variants with very large effects (for instance failing in such cases as the APOC3). However, there is a solid case to be made in favour of an approach that looks at haplotype sharing approach rather than IBD methods, for example when unrelated individuals are considered and levels of IBD are expected to be very low. In fact, most successful uses of IBD mapping for making inference on disease risk have involved pedigrees (as discuss in Chapter 1).

A second application might make use of the genotype-based coancestry matrix introduced previously. Mixed model approaches to discovering associations make use of a variance-covariance matrix that accounts for global relatedness patterns among the individuals under study. This variance-covariance matrix is typically constructed using genome-wide kinship estimates which capture large-scale environmental effects but which do not account for any local patterns due to

stratification of rare variation. A coancestry matrix constructed using just local information might solve the issue of differential stratification of rare variants.

Finally, we might make use of the NN estimates directly. Mother-father-offspring trios provide the key structure in family-based association tests, which are robust to population stratification. Similar ideas using pseudo-parents could potentially be used to provide robustness against fine-scale stratification in the analysis of case control data. In the absence of trio data, tests like the TDT (Spielman et al. [1993]) could be applied using pseudo-trios. For example, cases could select pseudo-parents from among controls, allowing thus the ‘non-transmitted’ alleles to be used as a control against the ‘transmitted’ ones.

Interestingly, a relatedness variance-covariance matrix that uses our pseudo-parent estimates could bridge the TDT with logistic regression models. Whether such an approach has good power and robustness remains to be seen. Nonetheless, there may be multiple ways of exploiting NN structures in the analysis of genetic risk to provide a more fine-grained view of geographical differentiation than methods based on patterns of allele sharing (including PCA and current mixed models).

Bibliography

- 1000 Genomes Project Consortium et al. A map of human genome variation from population-scale sequencing. *Nature*, 467(7319):1061–1073, 2010. 24, 118, 191
- 1000 Genomes Project Consortium et al. An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491(7422):56–65, 2012. 28, 29, 41, 118, 189, 190, 191, 224
- Gonçalo R Abecasis and Janis E Wigginton. Handling marker-marker linkage disequilibrium: pedigree analysis with clustered markers. *The American Journal of Human Genetics*, 77(5):754–767, 2005. 13
- Gonçalo R Abecasis, Stacey S Cherny, William O Cookson, and Lon R Cardon. Merlinrapid analysis of dense genetic maps using sparse gene flow trees. *Nature genetics*, 30(1):97–101, 2001. 15
- Joshua Akey, Li Jin, and Momiao Xiong. Haplotypes vs single marker linkage disequilibrium tests: what do we gain? *European Journal of Human Genetics*, 9(4), 2001. 13
- Cornelis A Albers, Jim Stankovich, Russell Thomson, Melanie Bahlo, and Hilbert J Kappen. Multipoint approximations of identity-by-descent probabilities for accurate linkage analysis of distantly related individuals. *The American Journal of Human Genetics*, 82(3):607–622, 2008. 13
- Anders Albrechtsen, Thorfinn Sand Korneliussen, Ida Moltke, Thomas van Overseem Hansen, Finn Cilius Nielsen, and Rasmus Nielsen. Relatedness mapping

BIBLIOGRAPHY

- and tracts of relatedness for genome-wide data in the presence of linkage disequilibrium. *Genetic epidemiology*, 33(3):266–274, 2009. 12, 30
- Anders Albrechtsen, Ida Moltke, and Rasmus Nielsen. Natural selection and the distribution of identity-by-descent in the human genome. *Genetics*, 186(1):295–308, 2010. 12, 18
- Aida M Andrés, Andrew G Clark, Lawrence Shimmin, Eric Boerwinkle, Charles F Sing, and James E Hixson. Understanding the accuracy of statistical haplotype inference with sequence data of known phase. *Genetic epidemiology*, 31(7):659–671, 2007. 16
- Jennifer Asimit and Eleftheria Zeggini. Rare variant association analysis methods for complex traits. *Annual review of genetics*, 44:293–308, 2010. 29
- Marie-Claude Babron, Marie de Tayrac, Douglas N Rutledge, Eleftheria Zeggini, and Emmanuelle Génin. Rare and low frequency variant stratification in the UK population: description and impact on association tests. *PloS one*, 7(10):e46519, 2012. 29
- David J Balding, Martin Bishop, and Chris Cannings. *Handbook of statistical genetics*, volume 1. John Wiley & Sons, 2008. 2, 3, 4
- Vikas Bansal, Ondrej Libiger, Ali Torkamani, and Nicholas J Schork. Statistical analysis strategies for association studies involving rare variants. *Nature Reviews Genetics*, 11(11):773–785, 2010. 29
- Sivan Bercovici, Christopher Meek, Ydo Wexler, and Dan Geiger. Estimating genome-wide IBD sharing from SNP data via an efficient hidden Markov model of LD with application to gene mapping. *Bioinformatics*, 26(12):i175–i182, 2010. 28
- Benjamin P Blackburne and Simon Whelan. Measuring the distance between multiple sequence alignments. *Bioinformatics*, 28(4):495–502, 2012. 3
- Michael GB Blum and Olivier François. Minimal clade size and external branch length under the neutral coalescent. *Advances in applied probability*, pages 647–662, 2005. 142, 169

BIBLIOGRAPHY

- Walter Bodmer and Carolina Bonilla. Common and rare variants in multifactorial susceptibility to common diseases. *Nature genetics*, 40(6):695–701, 2008. 29
- Abra Brisbin. *Linkage analysis for categorical traits and ancestry assignment in admixed individuals*. PhD thesis, Ph.D. thesis, Cornell University, 2010. 27
- Abra Brisbin, Katarzyna Bryc, Jake Byrnes, Fouad Zakharia, Larsson Omberg, Jeremiah Degenhardt, Andrew Reynolds, Harry Ostrer, Jason G Mezey, and Carlos D Bustamante. PCAdmix: principal components-based assignment of ancestry along each chromosome in individuals with admixed ancestry from two or more populations. *Human biology*, 84(4):343, 2012. 27, 37
- Marshall D Brown, Chris G Glazner, Chaozhi Zheng, and Elizabeth A Thompson. Inferring coancestry in population samples in the presence of linkage disequilibrium. *Genetics*, 190(4):1447–1460, 2012. 12, 30
- Brian L Browning and Sharon R Browning. A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *The American Journal of Human Genetics*, 84(2):210–223, 2009. 16, 18, 23
- Brian L Browning and Sharon R Browning. A fast, powerful method for detecting identity by descent. *The American Journal of Human Genetics*, 88(2):173–182, 2011a. 6, 7, 8, 11, 12, 18, 25, 26
- Brian L Browning and Sharon R Browning. Improving the accuracy and efficiency of identity-by-descent detection in population data. *Genetics*, 194(2):459–471, 2013a. 9, 10, 12
- Brian L Browning and Zhaoxia Yu. Simultaneous genotype calling and haplotype phasing improves genotype accuracy and reduces false-positive associations for genome-wide association studies. *The American Journal of Human Genetics*, 85(6):847–861, 2009. 21, 24, 40, 68, 117, 141

BIBLIOGRAPHY

- Brian L Browning, Donna L Brashear, Andy A Butler, Devon D Cyr, Elizabeth C Harris, Anita J Nelsen, David P Yarnall, Margaret G Ehm, and Michael J Wagner. Linkage analysis using single nucleotide polymorphisms. *Human heredity*, 57(4):220–227, 2004. 13
- Sharon R Browning. Multilocus association mapping using variable-length markov chains. *The American Journal of Human Genetics*, 78(6):903–913, 2006. 12, 21
- Sharon R Browning. Estimation of pairwise identity by descent from dense genetic marker data in a population sample of haplotypes. *Genetics*, 178(4):2123–2132, 2008a. 12
- Sharon R Browning. Missing data imputation and haplotype phase inference for genome-wide association studies. *Human genetics*, 124(5):439–450, 2008b. 16
- Sharon R Browning and Brian L Browning. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *The American Journal of Human Genetics*, 81(5):1084–1097, 2007. 12, 21, 24
- Sharon R Browning and Brian L Browning. High-resolution detection of identity by descent in unrelated individuals. *The American Journal of Human Genetics*, 86(4):526–539, 2010. 8, 12
- Sharon R Browning and Brian L Browning. Haplotype phasing: existing methods and new developments. *Nature Reviews Genetics*, 12(10):703–714, 2011b. 13, 15, 16, 17, 21, 22, 23, 24, 116
- Sharon R Browning and Brian L Browning. Population structure can inflate SNP-based heritability estimates. *American journal of human genetics*, 89(1):191, 2011c. 30
- Sharon R Browning and Brian L Browning. Identity by descent between distant relatives: detection and applications. *Annual review of genetics*, 46:617–633, 2012. 7, 8, 9, 13, 25, 30, 124, 125, 220

BIBLIOGRAPHY

- Sharon R Browning and Brian L Browning. Identity-by-descent-based heritability analysis in the northern finland birth cohort. *Human genetics*, 132(2):129–138, 2013b. 30
- Sharon R Browning and Elizabeth A Thompson. Detecting rare variant associations by identity-by-descent mapping in case-control studies. *Genetics*, 190(4):1521–1531, 2012. 30
- Zheng Cai, Nicola J Camp, Lisa Cannon-Albright, and Alun Thomas. Identification of regions of positive selection using shared genomic segment analysis. *European Journal of Human Genetics*, 19(6):667–671, 2011. 18
- Christopher L Campbell, Pier F Palamara, Maya Dubrovsky, Laura R Botigué, Marc Fellous, Gil Atzmon, Carole Oddoux, Alexander Pearlman, Li Hao, Brenna M Henn, et al. North African Jewish and non-Jewish populations form distinctive, orthogonal clusters. *Proceedings of the National Academy of Sciences*, 109(34):13865–13870, 2012. 26
- Gary K Chen, Paul Marjoram, and Jeffrey D Wall. Fast and flexible simulation of dna sequence data. *Genome research*, 19(1):136–142, 2009. 98, 143
- Andrew G Clark. Inference of haplotypes from pcr-amplified samples of diploid populations. *Molecular biology and evolution*, 7(2):111–122, 1990. 16
- C Clark Cockerham. Higher order probability functions of identity of alleles by descent. *Genetics*, 69(2):235, 1971. 3
- Adrian Cortes and Matthew A Brown. Promise and pitfalls of the immunochip. *Arthritis Res Ther*, 13(1):101, 2011. 41
- Charles William Cotterman. *A calculus for statistico-genetics*. PhD thesis, The Ohio State University, 1940. 2
- Olivier Delaneau, Cédric Coulonges, and Jean-François Zagury. Shape-it: new rapid and accurate algorithm for haplotype inference. *BMC bioinformatics*, 9(1):540, 2008. 23

BIBLIOGRAPHY

- Olivier Delaneau, Jonathan Marchini, and Jean-François Zagury. A linear complexity phasing method for thousands of genomes. *Nature methods*, 9(2):179–181, 2012. 16, 19, 23, 40, 116, 222
- Olivier Delaneau, Jean-François Zagury, and Jonathan Marchini. Improved whole-chromosome phasing for disease and population genetic studies. *Nature methods*, 10(1):5–6, 2013. 23
- Arthur P Dempster, Nan M Laird, Donald B Rubin, et al. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal statistical Society*, 39(1):1–38, 1977. 15, 17
- Michel Marie Deza and Elena Deza. *Encyclopedia of distances*. Springer, 2009. 3
- Alexander Dilthey, Stephen Leslie, Loukas Moutsianas, Judong Shen, Charles Cox, Matthew R Nelson, and Gil McVean. Multi-population classical hla type imputation. *PLoS computational biology*, 9(2):e1002877, 2013. 68
- Peter Donnelly and Simon Tavaré. Coalescents and genealogical structure under neutrality. *Annual review of genetics*, 29(1):401–421, 1995. 4
- Richard Durbin. *Biological sequence analysis: probabilistic models of proteins and nucleic acids*. Cambridge university press, 1998. 19
- Julien Y Dutheil, Ganesh Ganapathy, Asger Hobolth, Thomas Mailund, Marcy K Uyenoyama, and Mikkel H Schierup. Ancestral population genomics: the coalescent hidden markov model approach. *Genetics*, 183(1):259–274, 2009. 19
- Evan E Eichler, Jonathan Flint, Greg Gibson, Augustine Kong, Suzanne M Leal, Jason H Moore, and Joseph H Nadeau. Missing heritability and strategies for finding the underlying causes of complex disease. *Nature Reviews Genetics*, 11(6):446–450, 2010. 28
- Laurent Excoffier and Montgomery Slatkin. Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. *Molecular biology and evolution*, 12(5):921–927, 1995. 17

BIBLIOGRAPHY

- Laurent Excoffier, Guillaume Laval, and David Balding. Gametic phase estimation over large genomic regions using an adaptive window approach. *Human genomics*, 1(1), 2003. 17
- Paul Fearnhead and Peter Donnelly. Estimating recombination rates from population genetic data. *Genetics*, 159(3):1299–1318, 2001. 13, 21
- Joseph Felsenstein and Gary A Churchill. A hidden markov model approach to variation among sites in rate of evolution. *Molecular Biology and Evolution*, 13(1):93–104, 1996. 19
- Ronald Aylmer Fisher. *The genetical theory of natural selection.*, volume 1. Oxford Clarendon Press, 1930. 1, 4
- Lindsey A Foreman. Generalisation of the viterbi algorithm. *IMA Journal of Management Mathematics*, 4(4):351–367, 1992. 50, 51, 52, 54, 89
- C Francks, F Tozzi, A Farmer, JB Vincent, D Rujescu, D St Clair, and P Muglia. Population-based linkage analysis of schizophrenia and bipolar case-control cohorts identifies a potential susceptibility locus on 19q13. *Molecular psychiatry*, 15(3):319–325, 2010. 30
- Guimin Gao, David B Allison, and Ina Hoeschele. Haplotyping methods for pedigrees. *Human heredity*, 67(4):248–266, 2009. 14
- J Gay, Simon Myers, and Gilean McVean. Estimating meiotic gene conversion rates from population genetic data. *Genetics*, 177(2):881–894, 2007. 19
- Giulio Genovese, Gregory Leibon, Martin R Pollak, and Daniel N Rockmore. Improved IBD detection using incomplete haplotype information. *BMC genetics*, 11(1):58, 2010. 12, 18
- Santhosh Girirajan, Jill A Rosenfeld, Gregory M Cooper, Francesca Antonacci, Priscillia Siswara, Andy Itsara, Laura Vives, Tom Walsh, Shane E McCarthy, Carl Baker, et al. A recurrent 16p12. 1 microdeletion supports a two-hit model for severe developmental delay. *Nature genetics*, 42(3):203–209, 2010. 31

BIBLIOGRAPHY

- Chris Glazner and Alison Elizabeth Thompson. Improving pedigree-based linkage analysis by estimating coancestry among families. *Statistical Applications in Genetics and Molecular Biology*, 11(2):1–18, 2012. 30
- Michael E Goddard, S Hong Lee, Jian Yang, Naomi R Wray, and Peter M Visscher. Response to browning and browning. *American journal of human genetics*, 89(1):193, 2011. 30
- Simon Gravel. Population genetics models of local ancestry. *Genetics*, 191(2):607–619, 2012. 124
- Adam P Gregory, Calliope A Dendrou, Kathrine E Attfield, Aiden Haghikia, Dionysia K Xifara, Falk Butter, Gereon Poschmann, Gurman Kaur, Lydia Lambert, Oliver A Leach, et al. Tnf receptor 1 genetic risk mirrors outcome of anti-tnf therapy in multiple sclerosis. *Nature*, 2012. 41
- Robert C Griffiths and Paul Marjoram. Ancestral inference from samples of dna sequences with recombination. *Journal of Computational Biology*, 3(4):479–502, 1996. 5
- Robert C Griffiths and Paul Marjoram. . In Peter Donnelly and Simon Tavaré, editors, *Progress in Population Genetics and Human Evolution*, pages 257–270. Springer-Verlag, 1997. 5
- Julius Gudmundsson, Patrick Sulem, Daniel F Gudbjartsson, Gisli Masson, Bjarni A Agnarsson, Kristrun R Benediktsdottir, Asgeir Sigurdsson, Olafur Th Magnusson, Sigurjon A Gudjonsson, Droplaug N Magnusdottir, et al. A study based on whole-genome sequencing yields a rare variant at 8q24 associated with prostate cancer. *Nature genetics*, 44(12):1326–1329, 2012. 159
- Alexander Gusev, Jennifer K Lowe, Markus Stoffel, Mark J Daly, David Altshuler, Jan L Breslow, Jeffrey M Friedman, and Itsik Pe’er. Whole population, genome-wide mapping of hidden relatedness. *Genome Research*, 19(2):318–326, 2009. 10, 12, 26

BIBLIOGRAPHY

- Alexander Gusev, Eimear E Kenny, Jennifer K Lowe, Jaqueline Salit, Richa Saxena, Sekar Kathiresan, David M Altshuler, Jeffrey M Friedman, Jan L Breslow, and Itsik Pe'er. DASH: a method for identical-by-descent haplotype mapping uncovers association with recent variation. *The American Journal of Human Genetics*, 88(6):706–717, 2011. 9, 10, 12, 28, 30
- Alexander Gusev, Pier Francesco Palamara, Gregory Aponte, Zhong Zhuang, Ariel Darvasi, Peter Gregersen, and Itsik Pe'er. The architecture of long-range haplotypes shared within and across populations. *Molecular biology and evolution*, 29(2):473–486, 2012. 25, 26
- Eran Halperin and Eleazar Eskin. Haplotype reconstruction from genotype data using imperfect phylogeny. *Bioinformatics*, 20(12):1842–1849, 2004. 17
- Richard W Hamming. Error detecting and error correcting codes. *Bell System technical journal*, 29(2):147–160, 1950. 3
- Lide Han and Mark Abney. Identity by descent estimation with dense genome-wide genotype data. *Genetic epidemiology*, 35(6):557–567, 2011. 12
- Thomas v O Hansen, Bent Ejlersen, Anders Albrechtsen, Eva Bergsten, Peter Bjerregaard, Torben Hansen, Torben Myrhøj, Peter B Nielsen, Vera Timmermans-Wielenga, Mette K Andersen, et al. A common greenlandic inuit brca1 ring domain founder mutation. *Breast cancer research and treatment*, 115(1):69–76, 2009. 13, 31
- Thomas v O Hansen, Lars Jønson, Anders Albrechtsen, Ane Y Steffensen, Eva Bergsten, Torben Myrhøj, Bent Ejlersen, and Finn C Nielsen. Identification of a novel brca1 nucleotide 4803delcc/c. 4684delcc mutation and a nucleotide 249t_c a/c. 130t_c a (p. cys44ser) mutation in two greenlandic inuit families: implications for genetic screening of greenlandic inuit families with high risk for breast and/or ovarian cancer. *Breast cancer research and treatment*, 124(1):259–264, 2010. 31
- Kelley Harris and Rasmus Nielsen. Inferring demographic history from a spectrum of shared haplotype lengths. *PLoS genetics*, 9(6):e1003521, 2013. 1, 9, 26, 124, 125, 227

BIBLIOGRAPHY

- ME Hawley and KK Kidd. Haplo: a program using the em algorithm to estimate the frequencies of multi-site haplotypes. *Journal of Heredity*, 86(5):409–411, 1995. 17
- Ben J Hayes, Peter M Visscher, Helen C McPartlan, and Mike E Goddard. Novel multilocus measure of linkage disequilibrium to estimate past effective population size. *Genome Research*, 13(4):635–643, 2003. 180
- Jotun Hein, Mikkel Schierup, and Carsten Wiuf. *Gene genealogies, variation and evolution: a primer in coalescent theory*. Oxford university press, 2004. 98, 125, 127, 131
- Hilde Maria Jozefa Dominiek Herbots. *Stochastic models in population genetics: genealogy and genetic differentiation in structured populations*. PhD thesis, University of London, 1994. 4
- Peter Heutink and Ben A Oostra. Gene finding in genetically isolated populations. *Human molecular genetics*, 11(20):2507–2515, 2002. 158
- Anjali G Hinch, Arti Tandon, Nick Patterson, Yunli Song, Nadin Rohland, Cameron D Palmer, Gary K Chen, Kai Wang, Sarah G Buxbaum, Ermeg L Akylbekova, et al. The landscape of recombination in African Americans. *Nature*, 476(7359):170–175, 2011. 6
- Asger Hobolth, Ole F Christensen, Thomas Mailund, and Mikkel H Schierup. Genomic relationships and speciation times of human, chimpanzee, and gorilla inferred from a coalescent hidden markov model. *PLoS genetics*, 3(2):e7, 2007. 19
- Gabriel E Hoffman. Correcting for population structure and kinship using the linear mixed model: Theory and extensions. *PloS one*, 8(10):e75707, 2013. 3
- Clive J Hoggart, MARK D Shriver, Rick A Kittles, David G Clayton, and Paul M McKeigue. Design and analysis of admixture mapping studies. *The American Journal of Human Genetics*, 74(5):965–978, 2004. 19

BIBLIOGRAPHY

- Hilma Holm, Daniel F Gudbjartsson, Patrick Sulem, Gisli Masson, Hafdis Th Helgadóttir, Carlo Zanon, Olafur Th Magnusson, Agnar Helgason, Jona Sæmundsdóttir, Arnaldur Gylfason, et al. A rare variant in MYH6 is associated with high risk of sick sinus syndrome. *Nature genetics*, 43(4):316–320, 2011. 18
- Roderick HJ Houwen, Siamak Baharloo, Kathleen Blankenship, Peter Raeymaekers, Jenneke Juyn, Lodewijk A Sandkuijl, and Nelson B Freimer. Genome screening by searching for shared segments: mapping a gene for benign recurrent intrahepatic cholestasis. *Nature genetics*, 8(4):380–386, 1994. 12, 28
- Bryan Howie, Jonathan Marchini, and Matthew Stephens. Genotype imputation with thousands of genomes. *G3: Genes, Genomes, Genetics*, 1(6):457–470, 2011. 3, 22
- Bryan Howie, Christian Fuchsberger, Matthew Stephens, Jonathan Marchini, and Gonçalo R Abecasis. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nature genetics*, 44(8):955–959, 2012. 3, 6
- Bryan N Howie, Peter Donnelly, and Jonathan Marchini. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS genetics*, 5(6):e1000529, 2009. 6, 22, 92, 117, 141
- Lucy Huang, Yun Li, Andrew B Singleton, John A Hardy, Gonçalo Abecasis, Noah A Rosenberg, and Paul Scheet. Genotype-imputation accuracy across worldwide human populations. *The American Journal of Human Genetics*, 84(2):235–250, 2009. 16
- Richard R Hudson. Properties of a neutral allele model with intragenic recombination. *Theoretical population biology*, 23(2):183–201, 1983. 5
- Richard R Hudson. Generating samples under a wright–fisher neutral model of genetic variation. *Bioinformatics*, 18(2):337–338, 2002. 97
- Chad D Huff, David J Witherspoon, Tatum S Simonson, Jinchuan Xing, W Scott Watkins, Yuhua Zhang, Therese M Tuohy, Deborah W Neklason, Randall W

BIBLIOGRAPHY

- Burt, Stephen L Guthery, et al. Maximum-likelihood estimation of recent shared ancestry (ERSA). *Genome research*, 21(5):768–774, 2011. 12, 25
- International HapMap 3 Consortium et al. Integrating common and rare genetic variation in diverse human populations. *Nature*, 467(7311):52–58, 2010. 118
- International HapMap Consortium et al. A haplotype map of the human genome. *Nature*, 437(7063):1299–1320, 2005. 17
- International Multiple Sclerosis Genetics Consortium, Wellcome Trust Case Control Consortium 2, et al. Genetic risk and a primary role for cell-mediated immune mechanisms in multiple sclerosis. *Nature*, 476(7359):214–219, 2011. 2, 189, 199
- International Multiple Sclerosis Genetics Consortium et al. Analysis of immune-related loci identifies 48 new susceptibility variants for multiple sclerosis. *Nature genetics*, 2013. 41
- Svante Janson and Götz Kersting. On the total external length of the kingman coalescent. *Electron. J. Probab*, 16:2203–2218, 2011. 142, 169
- Yunxuan Jiang. Identification of the effect of population stratification on association studies of rare variants. Dissertation for master of science in public health biostatistics, Emory University, 2011. 30
- Yunxuan Jiang, Michael P Epstein, and Karen N Conneely. Assessing the impact of population stratification on association studies of rare variation. *Human heredity*, 76(1):28–35, 2013. 30
- Thorlakur Jonsson, Jasvinder K Atwal, Stacy Steinberg, Jon Snaedal, Palmi V Jonsson, Sigurbjorn Bjornsson, Hreinn Stefansson, Patrick Sulem, Daniel Gudbjartsson, Janice Maloney, et al. A mutation in app protects against alzheimer/’s disease and age-related cognitive decline. *Nature*, 488(7409):96–99, 2012. 159
- Luke Jostins, Katherine I Morley, and Jeffrey C Barrett. Imputation of low-frequency variants using the HapMap3 benefits from large, diverse reference sets. *European Journal of Human Genetics*, 19(6):662–666, 2011. 16

BIBLIOGRAPHY

- Hosung Kang, Zhaohui S Qin, Tianhua Niu, and Jun S Liu. Incorporating genotyping uncertainty in haplotype inference for single-nucleotide polymorphisms. *The American Journal of Human Genetics*, 74(3):495–510, 2004. 24
- Hyun Min Kang, Noah A Zaitlen, Claire M Wade, Andrew Kirby, David Heckerman, Mark J Daly, and Eleazar Eskin. Efficient control of population structure in model organism association mapping. *Genetics*, 178(3):1709–1723, 2008. 3
- G Karigl. A recursive algorithm for the calculation of identity coefficients. *Annals of human genetics*, 45(3):299–305, 1981. 3
- Liisa Kauppi, Alec J Jeffreys, and Scott Keeney. Where the crossovers are: recombination distributions in mammals. *Nature Reviews Genetics*, 5(6):413–424, 2004. 207
- Andrew D Kern and David Haussler. A population genetic hidden markov model for detecting genomic regions under selection. *Molecular biology and evolution*, 27(7):1673–1685, 2010. 19
- John FC Kingman. On the genealogy of large populations. *Journal of Applied Probability*, pages 27–43, 1982a. 4, 125
- John Frank Charles Kingman. The coalescent. *Stochastic processes and their applications*, 13(3):235–248, 1982b. 1, 4, 125
- John Frank Charles Kingman. Exchangeability and the evolution of large populations. In G Koch and F Spizzichino, editors, *Exchangeability in Probability and Statistics*, pages 97–112. North-Holland Publishing Company, 1982c. 4, 125
- Robert J Klein, Caroline Zeiss, Emily Y Chew, Jen-Yue Tsai, Richard S Sackler, Chad Haynes, Alice K Henning, John Paul SanGiovanni, Shrikant M Mane, Susan T Mayne, et al. Complement factor H polymorphism in age-related macular degeneration. *Science*, 308(5720):385–389, 2005. 28
- Andrey Nikolaevich Kolmogorov. Sulla determinazione empirica di una legge di distribuzione. *Giornale dell’Istituto Italiano degli Attuari*, 4:83–91, 1933. 163

BIBLIOGRAPHY

- Augustine Kong, Gisli Masson, Michael L Frigge, Arnaldur Gylfason, Pasha Zushmanovich, Gudmar Thorleifsson, Pall I Olason, Andres Ingason, Stacy Steinberg, Thorunn Rafnar, et al. Detection of sharing by descent, long-range phasing and haplotype imputation. *Nature genetics*, 40(9):1068–1075, 2008. 10, 11, 15, 16, 18, 31, 33, 40, 42, 93, 125, 139, 219
- Leonid Kruglyak, Mark J Daly, Mary Pat Reeve-Daly, and Eric S Lander. Parametric and nonparametric linkage analysis: a unified multipoint approach. *American journal of human genetics*, 58(6):1347, 1996. 19
- Frank R Kschischang, Brendan J Frey, and H-A Loeliger. Factor graphs and the sum-product algorithm. *Information Theory, IEEE Transactions on*, 47(2):498–519, 2001. 18
- Martin Ladouceur, Zari Dastani, Yurii S Aulchenko, Celia MT Greenwood, and J Brent Richards. The empirical power of rare variant association methods: Results from sanger sequencing in 1,998 individuals. *PLoS genetics*, 8(2):e1002496, 2012. 29
- Eric S Lander and Philip Green. Construction of multilocus genetic linkage maps in humans. *Proceedings of the National Academy of Sciences*, 84(8):2363–2367, 1987. 15, 19
- Daniel John Lawson, Garrett Hellenthal, Simon Myers, and Daniel Falush. Inference of population structure using dense haplotype data. *PLoS genetics*, 8(1):e1002453, 2012. 6, 25, 27, 33, 34, 37, 42, 125, 140, 219, 227
- Gregory Leibon, Daniel N Rockmore, and Martin R Pollak. A SNP streak model for the identification of genetic regions identical-by-descent. *Statistical applications in genetics and molecular biology*, 7(1), 2008. 10, 13
- Heng Li and Richard Durbin. Inference of human population history from individual whole-genome sequences. *Nature*, 475(7357):493–496, 2011. 19
- Jun Z Li, Devin M Absher, Hua Tang, Audrey M Southwick, Amanda M Casto, Sohini Ramachandran, Howard M Cann, Gregory S Barsh, Marcus Feldman,

BIBLIOGRAPHY

- Luigi L Cavalli-Sforza, et al. Worldwide human relationships inferred from genome-wide patterns of variation. *science*, 319(5866):1100–1104, 2008. 190
- Na Li and Matthew Stephens. Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics*, 165(4):2213–2233, 2003. 19, 21, 22, 34, 36, 37, 44, 140, 151, 167
- Yun Li, Cristen J Willer, Jun Ding, Paul Scheet, and Gonalo R Abecasis. MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genetic epidemiology*, 34(8):816–834, 2010. 22
- Kuo-ching Liang, Xiaodong Wang, and Dimitris Anastassiou. Bayesian basecalling for dna sequence analysis using hidden markov models. *Computational Biology and Bioinformatics, IEEE/ACM Transactions on*, 4(3):430–440, 2007. 19
- Yushi Liu, Toru Nyunoya, Shuguang Leng, Steven A Belinsky, Yohannes Tesfaigzi, Shannon Bruse, et al. Softwares and methods for estimating genetic ancestry in human populations. *Hum. Genomics*, 7(1), 2013. 19
- Jeffrey C Long, Robert C Williams, and Margrit Urbanek. An em algorithm and testing strategy for multiple-locus haplotypes. *American journal of human genetics*, 56(3):799, 1995. 17
- Claudio Lottaz, Christian Iseli, C Victor Jongeneel, and Philipp Bucher. Modeling sequencing errors by combining hidden markov models. *Bioinformatics*, 19(suppl 2):ii103–ii112, 2003. 19
- Gerton Lunter, Andrea Rocco, Naila Mimouni, Andreas Heger, Alexandre Caldeira, and Jotun Hein. Uncertainty in homology inferences: assessing and improving genomic sequence alignment. *Genome research*, 18(2):298–309, 2008. 19
- Thomas Mailund, Julien Y Dutheil, Asger Hobolth, Gerton Lunter, and Mikkel H Schierup. Estimating divergence time and ancestral effective population size of bornean and sumatran orangutan subspecies using a coalescent hidden markov model. *PLoS genetics*, 7(3):e1001319, 2011. 19

BIBLIOGRAPHY

- Gustav Malécot. *Les mathématiques de l'hérédité*. Masson et Cie: Paris, 1948. 2, 7
- Teri A Manolio, Francis S Collins, Nancy J Cox, David B Goldstein, Lucia A Hindorff, David J Hunter, Mark I McCarthy, Erin M Ramos, Lon R Cardon, Aravinda Chakravarti, et al. Finding the missing heritability of complex diseases. *Nature*, 461(7265):747–753, 2009. 28, 29
- Jonathan Marchini and Bryan Howie. Genotype imputation for genome-wide association studies. *Nature Reviews Genetics*, 11(7):499–511, 2010. 29
- Jonathan Marchini, David Cutler, Nick Patterson, Matthew Stephens, Eleazar Eskin, Eran Halperin, Shin Lin, Zhaohui S Qin, Heather M Munro, Gonçalo R Abecasis, et al. A comparison of phasing algorithms for trios and unrelated individuals. *The American Journal of Human Genetics*, 78(3):437–450, 2006. 13, 15, 16, 18, 21
- Jonathan Marchini, Bryan Howie, Simon Myers, Gil McVean, and Peter Donnelly. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nature genetics*, 39(7):906–913, 2007. 29, 41, 198
- Paul Marjoram and Peter Donnelly. . In Peter Donnelly and Simon Tavaré, editors, *Progress in Population Genetics and Human Evolution*, pages 107–131. Springer-Verlag, 1997. 4
- Paul Marjoram and Jeff D Wall. Fast “coalescent” simulation. *BMC genetics*, 7(1):16, 2006. 26
- Iain Mathieson and Gil McVean. Differential confounding of rare and common variants in spatially structured populations. *Nature genetics*, 44(3):243–246, 2012. 2, 30, 224
- Iain Mathieson and Gil McVean. Estimating selection coefficients in spatially structured populations from time series data of allele frequencies. *Genetics*, 193(3):973–984, 2013. 19

BIBLIOGRAPHY

- Iain Mathieson and Gil McVean. Demography and the age of rare variants. submitted and under review by PLoS Genetics, 2014. URL <http://arxiv.org/abs/1401.4181>. 226, 227
- Brian P McEvoy, Joseph E Powell, Michael E Goddard, and Peter M Visscher. Human population dispersal “Out of Africa” estimated from linkage disequilibrium and allele frequencies of SNPs. *Genome research*, 21(6):821–829, 2011. 180, 225
- Gilean AT McVean. URL <http://www.stats.ox.ac.uk/~mcvean/R/coalescent.r>. R code. 98, 117
- Gilean AT McVean. A genealogical interpretation of principal components analysis. *PLoS Genetics*, 5(10):e1000686, 2009. 3
- Gilean AT McVean and Niall J Cardin. Approximating the coalescent with recombination. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 360(1459):1387–1393, 2005. 20, 26
- Gilean AT McVean, Simon R Myers, Sarah Hunt, Panos Deloukas, David R Bentley, and Peter Donnelly. The fine-scale structure of recombination rate variation in the human genome. *Science*, 304(5670):581–584, 2004. 17
- GJ Meerman, MA Meulen, and LA Sandkuijl. Perspectives of identity by descent (IBD) mapping in founder populations. *Clinical & Experimental Allergy*, 25(s2):97–102, 1995. 28
- Hitoshi Miyazawa, Masaaki Kato, Takuya Awata, Masakazu Kohda, Hiroyasu Iwasa, Nobuyuki Koyama, Tomoaki Tanaka, Shunei Kyo, Yasushi Okazaki, Koichi Hagiwara, et al. Homozygosity haplotype allows a genomewide search for the autosomal segments shared among patients. *The American Journal of Human Genetics*, 80(6):1090–1102, 2007. 10, 11
- Ida Moltke, Anders Albrechtsen, Thomas vO Hansen, Finn C Nielsen, and Rasmus Nielsen. A method for detecting IBD regions simultaneously in multiple individuals with applications to disease genetics. *Genome research*, 21(7):1168–1180, 2011. 13, 30

BIBLIOGRAPHY

- Giovanni Montana and Jonathan K Pritchard. Statistical tests for admixture mapping with case-control and cases-only data. *The American Journal of Human Genetics*, 75(5):771–789, 2004. 19
- Carrie B Moore, John R Wallace, Alex T Frase, Sarah A Pendergrass, and Marylyn D Ritchie. Using biobin to explore rare variant population stratification. In *Pacific Symposium on Biocomputing*. World Scientific, 2013a. 29
- Carrie B Moore, John R Wallace, Daniel J Wolfe, Alex T Frase, Sarah A Pendergrass, Kenneth M Weiss, and Marylyn D Ritchie. Low frequency variants, collapsed based on biological knowledge, uncover complexity of population stratification in 1000 genomes project data. *PLoS genetics*, 9(12):e1003959, 2013b. 29
- Ronald W Morrison and Kenneth A De Jong. Measurement of population diversity. In *Artificial Evolution*, pages 31–41. Springer, 2002. 3
- Loukas Moutsianas. *Imputation aided analysis of the association between autoimmune diseases and the MHC*. PhD thesis, Oxford University, 2011. 47, 118
- Kasper Munch and Anders Krogh. Automatic generation of gene finders for eukaryotic species. *BMC bioinformatics*, 7(1):263, 2006. 19
- Simon R Myers and Robert C Griffiths. Bounds on the minimum number of recombination events in a sample history. *Genetics*, 163(1):375–394, 2003. 17
- Matthew R Nelson, Daniel Wegmann, Margaret G Ehm, Darren Kessner, Pamela St Jean, Claudio Verzilli, Judong Shen, Zhengzheng Tang, Silviu-Alin Bacanu, Dana Fraser, et al. An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. *Science*, 337(6090):100–104, 2012. 29, 203, 209, 226
- S Nelson, B Merriman, Z Chen, M Ogdie, J Stone, and S Strom. Applications of pedigree-free identity-by-descent mapping to localizing disease genes. American Society of Human Genetics Annual Conference, 2006. URL <http://www.ashg.org/genetics/ashg06s/f20776.htm>. 10

BIBLIOGRAPHY

- Rasmus Nielsen, Joshua S Paul, Anders Albrechtsen, and Yun S Song. Genotype and SNP calling from next-generation sequencing data. *Nature Reviews Genetics*, 12(6):443–451, 2011. 24
- James R Norris. *Markov chains*. Cambridge university press, 1998. 19
- M Notohara. The coalescent and the genealogical process in geographically structured population. *Journal of mathematical biology*, 29(1):59–75, 1990. 4
- Ioanna Ntalla, Margarita Giannakopoulou, Panagiota Vlachou, Kalliope Giannitsopoulou, Vasiliki Gkesou, Chrysoula Makridi, Maria Marougka, Georgia Mikou, Kyriaki Ntaoutidou, Eirini Prountzou, et al. Body composition and eating behaviours in relation to dieting involvement in a sample of urban greek adolescents from the teenage (teens of attica: Genes & environment) study. *Public health nutrition*, pages 1–8, 2013. 159
- Jared O’Connell, Deepti Gurdasani, Olivier Delaneau, Nicola Pirastu, Sheila Ulivi, Massimiliano Cocca, Michela Traglia, Jie Huang, Jennifer E Huffman, Igor Rudan, et al. A general approach for haplotype phasing across the full spectrum of relatedness. *PLoS genetics*, 10(4):e1004234, 2014. 16
- Lior Pachter, Marina Alexandersson, and Simon Cawley. Applications of generalized pair hidden markov models to alignment and gene finding problems. *Journal of Computational Biology*, 9(2):389–399, 2002. 19
- Pier Francesco Palamara and Itsik Pe’er. Inference of historical migration rates via haplotype sharing. *Bioinformatics*, 29(13):i180–i188, 2013. 1, 9, 26
- Pier Francesco Palamara, Todd Lencz, Ariel Darvasi, and Itsik Pe’er. Length distributions of identity by descent reveal fine-scale demographic history. *The American Journal of Human Genetics*, 91(5):809–822, 2012. 1, 6, 8, 25
- Kimmo Palin, Harry Campbell, Alan F Wright, James F Wilson, and Richard Durbin. Identity-by-descent-based phasing and imputation in founder populations using graphical models. *Genetic epidemiology*, 35(8):853–860, 2011. 18, 40

BIBLIOGRAPHY

- Kalliope Panoutsopoulou, Ioanna Tachmazidou, and Eleftheria Zeggini. In search of low-frequency and rare variants affecting complex traits. *Human molecular genetics*, 22(R1):R16–R21, 2013. 28
- Kalliope Panoutsopoulou, Konstantinos Hatzikotoulas, Dionysia Kiara Xifara, Vincenza Colonna, Aliko-Eleni Farmaki, Graham RS Ritchie, Lorraine Southam, Arthur Gilly, Ioanna Tachmazidou, Segun Fatumo, Angela Matchan, Nigel W Rayner, J Huang, Ioanna Ntalla, Chrysoula Kiagiadaki, Manjinder Sandhu, Gilean AT McVean, Emmanouil Tsafantakis, Maria Karaleftheri, George Dedoussis, and Eleftheria Zeggini. Genetic characterisation of two greek population isolates. in prep. 160, 161, 173, 178, 180
- Nick Patterson, Neil Hattangadi, Barton Lane, Kirk E Lohmueller, David A Hafler, Jorge R Oksenberg, Stephen L Hauser, Michael W Smith, Stephen J OBrien, David Altshuler, et al. Methods for high-density admixture mapping of disease genes. *The American Journal of Human Genetics*, 74(5):979–1000, 2004. 19
- Leena Peltonen, Aarno Palotie, and Kenneth Lange. Use of population isolates for mapping complex traits. *Nature Reviews Genetics*, 1(3):182–190, 2000. 158
- Matti Pirinen, Peter Donnelly, Chris CA Spencer, et al. Efficient computation with a linear mixed model on large-scale data sets with applications to genetic studies. *The Annals of Applied Statistics*, 7(1):369–390, 2013. 3
- Toni I Pollin, Coleen M Damcott, Haiqing Shen, Sandra H Ott, John Shelton, Richard B Horenstein, Wendy Post, John C McLenithan, Lawrence F Bielak, Patricia A Peyser, et al. A null mutation in human APOC3 confers a favorable plasma lipid profile and apparent cardioprotection. *Science*, 322(5908):1702–1705, 2008. 159, 160
- Alkes L Price, Nick J Patterson, Robert M Plenge, Michael E Weinblatt, Nancy A Shadick, and David Reich. Principal components analysis corrects for stratification in genome-wide association studies. *Nature genetics*, 38(8):904–909, 2006. 2

BIBLIOGRAPHY

- Alkes L Price, Arti Tandon, Nick Patterson, Kathleen C Barnes, Nicholas Rafaels, Ingo Ruczinski, Terri H Beaty, Rasika Mathias, David Reich, and Simon Myers. Sensitive detection of chromosomal segments of distinct ancestry in admixed populations. *PLoS genetics*, 5(6):e1000519, 2009. 19, 27
- Alkes L Price, Agnar Helgason, Gudmar Thorleifsson, Steven A McCarroll, Augustine Kong, and Kari Stefansson. Single-tissue and cross-tissue heritability of gene expression via identity-by-descent in related or unrelated individuals. *PLoS genetics*, 7(2):e1001317, 2011. 10, 30
- Shaun Purcell, Benjamin Neale, Kathe Todd-Brown, Lori Thomas, Manuel AR Ferreira, David Bender, Julian Maller, Pamela Sklar, Paul IW De Bakker, Mark J Daly, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics*, 81(3):559–575, 2007. 8, 12, 28, 29, 173
- Zhaohui S Qin, Tianhua Niu, and Jun S Liu. Partition-ligation–expectation-maximization algorithm for haplotype inference with single-nucleotide polymorphisms. *American journal of human genetics*, 71(5):1242, 2002. 17
- Lawrence Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989. 11, 19, 20, 22, 66, 89
- Cyril S Rakovski and Daniel O Stram. A kinship-based modification of the armitage trend test to address hidden population structure and small differential genotyping errors. *PLoS one*, 4(6):e5825, 2009. 3
- Peter Ralph and Graham Coop. The geography of recent genetic ancestry across Europe. *PLoS biology*, 11(5):e1001555, 2013. 6, 9, 25, 26, 226
- Jared C Roach, Gustavo Glusman, Arian FA Smit, Chad D Huff, Robert Hubley, Paul T Shannon, Lee Rowen, Krishna P Pant, Nathan Goodman, Michael Bamshad, et al. Analysis of genetic inheritance in a family quartet by whole-genome sequencing. *Science*, 328(5978):636–639, 2010. 30

BIBLIOGRAPHY

- François Rousset. Genetic differentiation in populations with different classes of individuals. *Theoretical population biology*, 55(3):297–308, 1999. 4
- Pardis C Sabeti, Patrick Varilly, Ben Fry, Jason Lohmueller, Elizabeth Hostetter, Chris Cotsapas, Xiaohui Xie, Elizabeth H Byrne, Steven A McCarroll, Rachelle Gaudet, et al. Genome-wide detection and characterization of positive selection in human populations. *Nature*, 449(7164):913–918, 2007. 13, 68
- Daniel J Schaid. Evaluating associations of haplotypes with traits. *Genetic epidemiology*, 27(4):348–364, 2004. 13
- Paul Scheet and Matthew Stephens. 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(4):629–644, 2006. 19, 21, 34
- Michael F Seldin, Bogdan Pasaniuc, and Alkes L Price. New approaches to disease mapping in admixed populations. *Nature Reviews Genetics*, 12(8):523–528, 2011. 27
- Nikolay Smirnov. Table for estimating the goodness of fit of empirical distributions. *The Annals of Mathematical Statistics*, pages 279–281, 1948. 163
- S Soi, L Scheinfeldt, C Lambert, J Hirbo, A Ranciaro, S Thompson, J M Bodo, A Froment, M Ibrahim, A. Juma, T Nyambo, S Omar, and C Wambebe. Demographic histories of African hunting-gathering populations inferred from genome-wide SNP variation. American Society of Human Genetics Annual Conference, 2011. URL <http://www.ichg2011.org/cgi-bin/ichg11s?author=Soi&sort=ptimes&sbutton=Detail&absno=21450&sid=564811>. 26
- Chris CA Spencer, Zhan Su, Peter Donnelly, and Jonathan Marchini. Designing genome-wide association studies: sample size, power, imputation, and the choice of genotyping chip. *PLoS genetics*, 5(5):e1000477, 2009. 37
- Richard S Spielman, Ralph E McGinnis, and Warren J Ewens. Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent

BIBLIOGRAPHY

- diabetes mellitus (iddm). *American journal of human genetics*, 52(3):506, 1993. 229
- Simon N Stacey, Patrick Sulem, Aslaug Jonasdottir, Gisli Masson, Julius Gudmundsson, Daniel F Gudbjartsson, Olafur T Magnusson, Sigurjon A Gudjonsson, Bardur Sigurgeirsson, Kristin Thorisdottir, et al. A germline variant in the tp53 polyadenylation signal confers cancer susceptibility. *Nature genetics*, 43(11):1098–1103, 2011. 159
- Matthew Stephens and Peter Donnelly. Inference in molecular population genetics. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62(4):605–635, 2000. 20
- Matthew Stephens and Peter Donnelly. A comparison of bayesian methods for haplotype reconstruction from population genotype data. *The American Journal of Human Genetics*, 73(5):1162–1169, 2003a. 21, 24, 95
- Matthew Stephens and Peter Donnelly. A comparison of bayesian methods for haplotype reconstruction from population genotype data. *The American Journal of Human Genetics*, 73(5):1162–1169, 2003b. 21
- Matthew Stephens and Paul Scheet. Accounting for decay of linkage disequilibrium in haplotype inference and missing-data imputation. *The American Journal of Human Genetics*, 76(3):449–462, 2005. 21
- Matthew Stephens, Nicholas J Smith, and Peter Donnelly. A new statistical method for haplotype reconstruction from population data. *The American Journal of Human Genetics*, 68(4):978–989, 2001. 21
- Barbara E Stranger, Eli A Stahl, and Towfique Raj. Progress and promise of genome-wide association studies for human complex trait genetics. *Genetics*, 187(2):367–383, 2011. 28
- Patrick Sulem, Daniel F Gudbjartsson, G Bragi Walters, Hafdis T Helgadóttir, Agnar Helgason, Sigurjon A Gudjonsson, Carlo Zanon, Soren Besenbacher, Gyda Bjornsdottir, Olafur T Magnusson, et al. Identification of low-frequency

BIBLIOGRAPHY

- variants associated with gout and serum uric acid levels. *Nature genetics*, 43(11):1127–1130, 2011. 159
- Ioanna Tachmazidou, George Dedoussis, Lorraine Southam, Aliko-Eleni Farmaki, Graham RS Ritchie, Dionysia Kiara Xifara, Angela Matchan, Konstantinos Hatzikotoulas, Nigel W Rayner, Yuan Chen, et al. A rare functional cardio-protective apoc3 variant has risen in frequency in distinct population isolates. *Nature communications*, 4, 2013. 160, 161, 173
- Hua Tang, Marc Coram, Pei Wang, Xiaofeng Zhu, and Neil Risch. Reconstructing genetic ancestry blocks in admixed individuals. *The American Journal of Human Genetics*, 79(1):1–12, 2006. 19
- Albert Tenesa, Pau Navarro, Ben J Hayes, David L Duffy, Geraldine M Clarke, Mike E Goddard, and Peter M Visscher. Recent human effective population size estimated from linkage disequilibrium. *Genome research*, 17(4):520–526, 2007. 180
- Alun Thomas. Assessment of SNP streak statistics using gene drop simulation with linkage disequilibrium. *Genetic epidemiology*, 34(2):119–124, 2010. 10
- Alun Thomas, Nicola J Camp, James M Farnham, Kristina Allen-Brady, and Lisa A Cannon-Albright. Shared genomic segment analysis. Mapping disease predisposition genes in extended pedigrees using SNP genotype assays. *Annals of human genetics*, 72(2):279–287, 2008. 10, 13
- EA Thompson. Gene identities and multiple relationships. *Biometrics*, 30(4):667, 1974. 3
- Elizabeth A Thompson. Statistical inference from genetic data on pedigrees. In *NSF-CBMS regional conference series in probability and statistics*, pages i–169. JSTOR, 2000. 30
- Elizabeth A Thompson. The IBD process along four chromosomes. *Theoretical population biology*, 73(3):369–373, 2008. 12

BIBLIOGRAPHY

- Bjarni J Vilhjalmsen, 2012. URL <https://github.com/bvilhjal/mixmogam>. mixmogam, Commit a40f3e2c95. 3
- John Wakeley. Coalescent theory. *Roberts & Company*, 2008. 125
- Yi Wang, James Lu, Jin Yu, Richard A Gibbs, and Fuli Yu. An integrative variant analysis pipeline for accurate genotype/haplotype inference in population NGS data. *Genome research*, 23(5):833–842, 2013. 37
- GA Watterson. On the number of segregating sites in genetical models without recombination. *Theoretical population biology*, 7(2):256–276, 1975. 37, 47
- Daniel Wegmann, Darren E Kessner, Krishna R Veeramah, Rasika A Mathias, Dan L Nicolae, Lisa R Yanek, Yan V Sun, Dara G Torgerson, Nicholas Rafaels, Thomas Mosley, et al. Recombination rates in admixed individuals identified by ancestry-based inference. *Nature genetics*, 43(9):847–853, 2011. 6
- Wellcome Trust Case Control Consortium, Paul R Burton, David G Clayton, Lon R Cardon, Nick Craddock, Panos Deloukas, Audrey Duncanson, Dominic P Kwiatkowski, Mark I McCarthy, Willem H Ouwehand, Nilesh J Samani, et al. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447(7145):661–678, 2007. 2, 28, 29, 40, 157, 189, 190, 225
- Ellen M Wijsman, Joseph H Rothstein, and Elizabeth A Thompson. Multi-point linkage analysis with many multiallelic or dense diallelic markers: Markov chain–monte carlo provides practical approaches for genome scans on general pedigrees. *The American Journal of Human Genetics*, 79(5):846–858, 2006. 30
- Hilde M Wilkinson-Herbots. Genealogy and subpopulation differentiation under various models of population structure. *Journal of Mathematical Biology*, 37(6):535–585, 1998. 4
- Amy L Williams, Nick Patterson, Joseph Glessner, Hakon Hakonarson, and David Reich. Phasing of many thousands of genotyped samples. *The American Journal of Human Genetics*, 91(2):238–251, 2012. 40

- Bruce Winney, Abdelhamid Boumertit, Tammy Day, Dan Davison, Chikodi Echeta, Irina Evseeva, Katarzyna Hutnik, Stephen Leslie, Kristin Nicodemus, Ellen C Royrvik, et al. People of the british isles: preliminary analysis of genotypes and surnames in a UK-control population. *European Journal of Human Genetics*, 20(2):203–210, 2012. 226
- Kyoung-Jae Won, Thomas Hamelryck, Adam Prügél-Bennett, and Anders Krogh. An evolutionary method for learning hmm structure: prediction of protein secondary structure. *BMC bioinformatics*, 8(1):357, 2007. 19
- Fred A Wright, Hanwen Huang, Xiaojun Guan, Kevin Gamiel, Clark Jeffries, William T Barry, Fernando Pardo-Manuel de Villena, Patrick F Sullivan, Kirk C Wilhelmsen, and Fei Zou. Simulating association studies: a data-based resampling method for candidate regions or whole genome scans. *Bioinformatics*, 23(19):2581–2588, 2007. 37
- Sewall Wright. Systems of mating. i. the biometric relations between parent and offspring. *Genetics*, 6(2):111, 1921. 7
- Sewall Wright. Evolution in mendelian populations. *Genetics*, 16(2):97, 1931. 1, 4
- Jian Yang, Beben Benyamin, Brian P McEvoy, Scott Gordon, Anjali K Henders, Dale R Nyholt, Pamela A Madden, Andrew C Heath, Nicholas G Martin, Grant W Montgomery, et al. Common SNPs explain a large proportion of the heritability for human height. *Nature genetics*, 42(7):565–569, 2010. 30
- Byung-Jun Yoon. Hidden markov models and their applications in biological sequence analysis. *Current genomics*, 10(6):402, 2009. 19
- Byung-Jun Yoon and Palghat P Vaidyanathan. Structural alignment of rnas using profile-cshmms and its application to rna homology search: Overview and new results. *Automatic Control, IEEE Transactions on*, 53(Special Issue): 10–25, 2008. 19
- Eleftheria Zeggini. Next-generation association studies for complex traits. *Nature genetics*, 43(4):287, 2011. 159

BIBLIOGRAPHY

- Peng Zhang, Xiaowei Zhan, Noah A Rosenberg, and Sebastian Zöllner. Genotype imputation reference panel selection using maximal phylogenetic diversity. *Genetics*, 195(2):319–330, 2013. 3
- Zhiwu Zhang, Edward S Buckler, Terry M Casstevens, and Peter J Bradbury. Software engineering the mixed model for genome-wide association studies on large samples. *Briefings in Bioinformatics*, 10(6):664–675, 2009. 3
- Or Zuk, Eliana Hechter, Shamil R Sunyaev, and Eric S Lander. The mystery of missing heritability: Genetic interactions create phantom heritability. *Proceedings of the National Academy of Sciences*, 109(4):1193–1198, 2012. 28, 30