Analysis of genetic variation data using Ancestral Recombination Graphs

Mark J. Minichiello,
Gonville and Caius College,
University of Cambridge.

This dissertation is submitted for
the degree of Doctor of Philosophy

04 July 2007

Declaration

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except where specifically indicated in the text and Acknowledgements. No part of this dissertation has been submitted for a degree or diploma or other qualification at this or any other university.

Mark J. Minichiello
Cambridge
04 July 2007

Summary

Analysis of Genetic Variation Data using Ancestral Recombination Graphs

Mark J. Minichiello

Large-scale association studies are being undertaken with the goal of uncovering the genetic determinants of complex disease. In this thesis I describe a computationally efficient method for inferring genealogies from population genotype data and show how these genealogies can be used to fine map disease loci and interpret association signals.

These genealogies take the form of the ancestral recombination graph (ARG). The ARG defines a genealogical tree for each locus, and, as one moves along the chromosome, the topologies of consecutive trees shift according to the impact of historical recombination events.

There are two stages to the analysis. First, I infer plausible ARGs using a heuristic algorithm, which can handle unphased and missing data, and is fast enough to be applied to large-scale studies. Second, I test the genealogical tree at each locus for a clustering of the disease cases beneath a branch, suggesting that a causative mutation occurred on that branch. Since the true ARG is unknown, I average this analysis over an ensemble of inferred ARGs.

I characterise the performance of the method across a wide range of simulated disease models. Compared with simpler tests, the method gives increased accuracy in positioning untyped causative loci and can also be used to estimate the frequencies of untyped causative alleles.

I apply the method to Ueda et al.'s association study of CTLA4 and Graves disease, and Yeager et al.'s association study of 8q24 and prostate cancer, showing how it can be used to dissect association signals. With the CTLA4 data, the method suggests a possible signal of allelic heterogeneity and interaction, not identified in the original analysis. With the 8q24 data, the method demonstrates the genealogical independence of two nearby association signals.

I also use inferred ARGs to impute missing data. The performance of the method is compared to a standard method by using genotype data with held-out values, and is shown to be competitive. I evaluate the utility of an approach where case-control studies are merged with more densely typed control sets, such as the HapMap, and the additional loci imputed, allowing them to be tested directly for association.

The thesis concludes with a discussion of further population genetic questions which may be addressable by use of inferred ARGs.

Acknowledgements

Firstly, I would like to thank my PhD supervisor, Richard Durbin, and my PhD adviser Simon Tavaré. I would also like to thank Ralph McGinnis and the members of the Durbin Research Group for many helpful discussions. I also thank: Lachlan Coin for providing code to calculate Extreme Value Distributions. David Balding and Clive Hoggart for providing the FREGENE forward simulator and simulated populations. John Todd and Neil Walker for providing the CTLA4 data, and David Clayton, Chris Lowe and Joanna Howson for helpful discussions on my analysis of that data. Gilles Thomas for providing the 8q24 data and jointly performing the ARG analysis of that data. Inês Barroso and Eleanor Wheeler for providing the Chromosome 20 Ashkenazi data. David Carter for extending the MARGARITA system so that it can be used for nucleotide imputation with resequencing data, and for providing the *S. cerevisiae* Chromosome 1 data. The Wellcome Trust Sanger Institute for my PhD studentship and Gonville and Caius College for funding conference travel.

Mark J. Minichiello,
Wellcome Trust Sanger Institute,
04 July 2007.

Contents

Chapter 1

Introduction

1.1 Complex Diseases

Identifying the genetic basis of disease is a primary goal of human genetics. In particular, identifying the polymorphisms that underly disease susceptibility may help contribute to the development of preventative and disease-modifying therapies.

So far, significant progress has been made on highly heritable Mendelian disorders, such as cystic fibrosis and Huntington's disease. By 2006, the OMIM data base contained details for 1822 genes affecting such monogenic disorders (Antonarakis and Beckmann, 2006). For these diseases, the presence or absence of disease alleles, mainly within a single gene, strongly determine presence of disease.

Progress, however, has been less certain for common, multifactorial diseases, such as heart disease, diabetes and cancer. Such diseases are believed to have a genetic component, for example, Narod et al. (1995) estimated the sibling recurrence risk (the ratio of disease presence given an affected sibling, compared with disease prevalence in the general population) for prostate cancer to be 2.62. However, by 2003, only 6 to 9 genetic variants had been identified with significant and replicated association to complex disease (Hirschhorn et al., 2002; Ioannidis et al., 2003; Lohmueller et al., 2003). Historically, many published associations with complex diseases have failed to replicate in follow-up studies (Nature Genetics Editorial, 1999). There are a number of reasons for this. Complex diseases are believed to be caused

by a combination of possibly hundreds of genetic and enviromental factors, with potentially small and interacting effects (Wang et al., 2005), and the use of sample sizes that do not have sufficient power to detect these is one possible reason for failure of replication. Other reasons include the difficulty in defining a suitable significance threshold (with a multiple testing problem confounded by publication bias), as well as the possibility that studies are complicated by sporadic cases and poorly defined phenotypes.

However, two recent advances now bring within reach the potential to identify robust associations to complex disease. First, there is a more complete description of genetic variation in the human population; and second, there have been substantial developments in large-scale genotyping technology. These advances allow us to undertake well powered genome-wide scans for complex disease alleles. And it is within the context of these developments that this thesis works: seeking further improvements in power and in our ability to dissect association signals, using large data sets.

1.2 Human Genetic Variation

The most common genetic variants in humans are single nucleotide polymorphisms (SNPs), of which there are estimated to be at least 10 million with minor allele frequency greater than 1% in the human population (Kruglyak and Nickerson, 2001). A great deal of effort has gone into discovering these, for example by the SNP Consortium (Sachidanandam et al., 2001), the HapMap Project (The International HapMap Consortium, 2005), and Perlegen Sciences (Hinds et al., 2005).

In the HapMap Project, more than 3 million SNPs have been identified in 269 individuals from four different populations: 30 parent-child trios from U.S. residents with northern and western European ancestry (called the "CEU" sample); 30 trios from the Yoruba people in Nigeria (YRI); 44 unrelated Japanese individuals (JPT); and 45 unrelated Chinese individuals (CHB). The primary goal of the HapMap Project is to aid association studies by discovering SNPs and by determining the correlation between alleles at nearby loci, called Linkage Disequilibrium (LD).

LD means that individuals who carry a particular SNP allele at one site often predictably carry specific alleles at other nearby sites. LD is crucial to the current design of association studies because it means that not all loci need to be assayed directly; rather, many can be tested indirectly for association by testing a correlated SNP. This results in a substantial reduction in genotyping effort. Using the HapMap data, markers, called tagSNPs (Kruglyak, 1999), have been selected for populations that efficiently capture the majority of polymorphism via LD. Hence, one approach is first to test the tagSNPs, and then follow up significant signals by genotyping correlated SNPs in order to further fine map the causative variants (Johnson et al., 2001).

The pattern of LD in a population is determined by the co-inheritance of haplotypes (that is, the sequence of alleles on a single copy of a chromosome). Specifically, when a mutation occurs, it does so on a particular haplotypic background, to which it is fully correlated. As the population evolves, recombinations between that marker and other markers cause the LD to be broken down. Since the frequency of recombinations between markers increases with distance between markers, LD is expected to decay with genetic distance and also with the age of the allele.

A variety of population genetic forces influence the patterns of co-inheritance, meaning that information about these forces can also be inferred from LD data. These forces include recombination hotspots, selection and demographic history such as bottlenecks, expansions and population subdivisions. However, because these can leave similar traces in the LD pattern, it is not always straightforward to separate out the effects (Reed and Tishkoff, 2006).

Recombination hotspots are regions of approximately 1-2 kb and occur roughly every 100kb (Jeffreys et al., 2001; Crawford et al., 2004; McVean et al., 2004; Myers et al., 2005). They have recombination rates orders of magnitude greater than the background recombination rate. These result in LD which does not show continuous decay but which is rather split into blocks of strong LD (Daly et al., 2001; Patil et al., 2001; Gabriel et al., 2002). Having knowledge of the LD block structure is helpful in the dissection of disease loci because association signals that arise in different LD blocks are likely to correspond to different causative variants (Yeager et al., 2007).

Selection can also leave strong signals in SNP data. Population specific selection causes potentially detectable differences in allele frequencies between populations (Akey et al., 2002), and selective sweeps cause a distinctive reduction in haplotypic diversity around the selected allele (Sabeti et al., 2002; Hanchard et al., 2006; Voight et al., 2006). Detection of selected regions may help in identifying disease causing regions of the genome, as selection can act to enhance disease resistance, as has been observed with malaria (Hamblin et al., 2002).

Both recombination and selective sweeps can aid the identification of susceptibility alleles by creating LD patterns which are amenable to tagging. However, population admixture and stratification can lead to spurious associations unless they are controlled for (Price et al., 2006) or exploited (Patterson et al., 2004).

The second advance is the development of large-scale genotyping technologies (Matsuzaki et al., 2004; Gunderson et al., 2005). It is now feasible to genotype thousands of individuals. This is crucial because in Wang et al. (2005) it was shown that in order to have sufficient power to detect complex disease alleles with relative risk effects of 1.5 or less, many thousands of affected case individuals and unaffected controls are required.

These developments mean that in the following years, the number of robustly replicated associations is likely to increase dramatically, as has already been seen with, for example, age-related macular degeneration (Edwards et al., 2005; Haines et al., 2005; Klein et al., 2005), prostate cancer (Gudmundsson et al., 2007; Haiman et al., 2007; Yeager et al., 2007), and others (The Wellcome Trust Case Control Consortium, 2007). All these studies follow the case-control association design, which I now describe.

1.3 Mapping Complex Traits via Case-Control Association Studies

In case-control association studies (Risch and Merikangas, 1996; Kruglyak, 1999; Risch, 2000; Cardon and Bell, 2001; Hirschhorn and Daly, 2005), the frequencies of alleles at sites of interest are compared in populations of cases (individuals affected by the disease) and controls (unaffected individuals). A higher frequency in cases is taken as evidence of that allele being

associated with increased disease risk. For a single SNP with alleles A and a, the data generated from a case-control study can be organised into a 2×2 or 3×2 contingency table. The 2×2 table counts the number of times each allele is observed in the cases and in the controls, while the 3×2 counts the genotypes, aa, aA and AA, in the cases and controls. A standard test, such as the chi-square test, can then be applied to either contingency table, to test for departure from the null model, where the alleles are evenly distributed between the phenotypes.

Suppose the allele counts in the cases and controls are represented in a contingency table as:

	A	a
cases	n_1	n_2
controls	n_3	n_4

.

Then the Pearson's chi-square test statistic is

$$X^2 = \sum_{c=1}^{4} \frac{(O_c - E_c)^2}{E_c},$$

where the O_c are the observed counts and the E_c are the expected number of counts under the model of no disease association. For example, E_1 the expected number of cases with the A allele, $E_1 = (n_1 + n_3) * \frac{(n_1+n_2)}{(n_1+n_2+n_3+n_4)}$, that is, the number of A alleles times the proportion of haplotypes that belong to case individuals.

In general, however, tests such as the Armitage test for trend and the genotypic (3×2) test are preferred over the allelic (2×2) chi-square test just described. This is because the allelic chi-square test does not in general give easily interpretable risk estimates, and can give inaccurate results when Hardy-Weinberg Equilibrium (HWE) does not hold in the population from which the cases and controls are sampled (Sasieni, 1997; Balding, 2006).

HWE holds when the alleles are independent: if the frequency of the A allele is p, then the frequency of the A homozygote is p^2, and the frequency of the heterozygote is $2p\,(1-p)$, and the frequency of the a homozygote is $(1-p)^2$.

Under the null hypothesis of no association, the allelic chi-square test statistic is distributed as χ^2 provided that the population is in HWE, and thus HWE must hold in order to

obtain the correct false-positive rate (Schaid and Jacobsen, 1999). The allelic test can give false-positive associations (it is anticonservative) when the homozygotes are more common in the general population than expected under HWE; and the test can be conservative when there are fewer homozygotes relative to HWE. In contrast, tests such as the Armitage test for trend are correctly calibrated even when HWE does not hold.

HWE holds under random mating and no selection, but the assumption of no selection in the cases tends to imply that the locus is not associated with the disease. Nevertheless, HWE can hold in the cases when the effect of the risk allele is multiplicative. The multiplicative risk model is as follows: if there is a d-fold increase in risk for the Aa genotype compared to the aa genotype, then there is an d^2 increased risk for AA. If HWE holds in the case population and the risk effect is multiplicative, then the allelic odds ratio is equal to the heterozygous odds ratio from the genotypic contingency table, giving the allelic odds ratio a straightforward interpretation. Otherwise, the allelic odds ratio is hard to interpret.

When using the allelic test to detect an association signal, it is implicitly assumed that single copies of the risk allele can confer disease and therefore have higher frequency in cases than in controls. However, the chi-square test applied to the genotypic 3×2 table has greater power to detect associations which depart from models where risk increases with number of alleles. In order to test specific risk models, the Armitage test for trend can be used, as can a logistic regression model, which also has the advantage of being able to jointly analyse multiple SNPs and other factors such as age at onset.

For these reasons, the Armitage trend test or the genotypic chi-square test tend to be recommended over the allelic chi-square test (Sasieni, 1997; Balding, 2006).

All these tests result in an assessment of departure from what is expected under the null model, which can be quantified as a P-value: the probability that such a significant departure from the null would be observed by chance. When the P-value passes a certain threshold, the marker is called significant, and disease associated. For association studies, this significance threshold is often set at $P < 10^{-6}$, which corresponds approximately to a genome-wide 5% type I error rate, that is, roughly corrected for the large number of correlated tests that are involved in scanning the whole genome.

When study designs and statistical methods are evaluated, their power to detect an association signal is usually reported. Power is the probability of observing a significant signal at a disease marker, given some disease model.

Affecting the power to detect a disease allele are its frequency, the genotype relative risk (GRR) and the disease prevalence. GRR describes the relative susceptibility to disease for each of the three genotypes: if $P(D|aa)$, $P(D|Aa)$ and $P(D|AA)$ are the probabilities of being affected for individuals with 0, 1, or 2 copies of the risk allele, then $\text{GRR}(Aa) = P(D|Aa)/P(D|aa)$ and $\text{GRR}(AA) = P(D|AA)/P(D|aa)$. Under the multiplicative model, $P(D|Aa)/P(D|aa) = P(D|AA)/P(D|Aa)$ and thus $\text{GRR}(AA) = \text{GRR}(Aa)^2$.

Estimating GRRs from data is typically a complex task, and in practice odds ratios (ORs) are reported instead. The OR is the ratio of those with the allele to those without the allele in cases compared with the controls. For the 2×2 contingency table, $OR = \frac{n_1 n_4}{n_2 n_3}$. An OR significantly different from 1 corresponds to a variant associated with the disease.

It is possible to determine, for a disease allele with parameters above, the number of samples required in order to achieve a significant signal (Purcell et al., 2003). For example, Wang et al. (2005) show that if the disease allele frequency is less than 0.01 and the odds ratio is less than 1.3, then over 10,000 cases and 10,000 controls are required to give 80% power at a significance threshold of $P < 10^{-6}$, when the causative polymorphism is typed. When the causative polymorphism is untyped, the power to detect that polymorphism is also dependent on the LD between it and the typed markers.

LD is often measured by a pairwise statistic, r^2, the square of the correlation coefficient (Devlin and Risch, 1995). Consider two SNPs, with alleles A and a at the first locus, and with alleles B and b at the second locus. The allele frequencies are written as π_A, π_a, π_B, π_b, and the frequency of the A-B haplotype is written as π_{AB}, then

$$r^2 = \frac{\left(\pi_{AB} - \pi_A \pi_B\right)^2}{\pi_A \pi_a \pi_B \pi_b}.$$

(It is worth noting that r^2 is the standard chi-square test statistic divided by the number of haplotype sequences in the sample.)

Suppose N_1 samples are required to achieve power when the causative polymorphism is typed. If we instead type a SNP in LD with the causative polymorphism, then N_1/r^2 samples are required in order to achieve approximately the same power (with the chi-square test) as if the causative polymorphism were typed (Pritchard and Przeworski, 2001). This means that markers around a causative SNP will tend to show a disease association. The strength of association will depend on their LD with the causative SNP, and the association signal will therefore decay with genetic distance. Consequently, a general guide is that tagSNPs are chosen so that every known common SNP has $r^2 \geq 0.8$ with a tag, although it is often possible to tag SNPs with greater power by using multiple SNPs as tags, called haplotype tags (Johnson et al., 2001).

1.4 The Ancestral Recombination Graph

From the description of the link between LD and association above, it starts to become clear that recombination is the key to mapping.

First consider linkage studies, where a small part of the history is known exactly in the form of a pedigree for closely related individuals. From the pedigree, the positions of recombinations can be inferred, and recombination distances between typed markers and unknown causative variants calculated. The number of recombinations in the history determine the resolution at which it is possible to map. Since only a few generations are considered in a pedigree, the number of recombinations is small and the ability to localise the causative gene is limited to a couple hundred kb.

Meanwhile, in association studies the genotyped individuals are more distantly related and the historical relationships stretch much deeper in time (Rohde et al., 2004). Consequently, there has been more opportunity for recombination to occur and mapping can take place at a finer scale. In population data, the recombination history is viewed via its effect on the co-inheritance of alleles on haplotypes, that is, via the LD patterns in the data. Uncertainty in the (unknown) ancestral history means that LD is relied upon as a proxy for the recombination history.

r^2 LD is, however, not a pure measure of recombination distance; as discussed earlier, it is affected by other factors such as the relative timing of mutation events and non-panmictic mating patterns, which may confound our ability to map disease loci.

To understand the pattern of variation in a population more fully, and to potentially improve mapping power and efficiency, the variation should be interpreted in terms of the evolutionary processes that produced it (Nordborg and Tavaré, 2002; McVean, 2002). And specifically for disease mapping, this means modelling the recombination history.

A formalism for describing these recombination histories is the Ancestral Recombination Graph (ARG). For a population of chromosome sequences, the ARG describes how they are related to each other, through mutation, recombination and coalescence, back to a common ancestor.

Note that there are two ways in which the term "Ancestral Recombination Graph" is used in the literature. The first, original use, in Griffiths and Marjoram (1997), uses the term to describe the stochastic process which gives the distribution of genealogies under the Wright-Fisher model with recombination (it is the analogue of the coalescent process when recombination is possible; which is also known as the "coalescent-with-recombination", described in the next section). The second use (as in, for example, Song and Hein (2005)), uses the term to refer the graph structure which describes the genealogy of a sample of sequences, where nodes in the graph correspond to mutation, recombination and coalescence events. Since the term Ancestral Recombination Graph is used in both ways in the literature, it seems sensible clarify the terminology. In this thesis I use the term coalescent-with-recombination to describe the stochastic process, and ARG refer to the graph structure which describes a genealogical history with recombinations, coalescences and mutations.

An ARG, under the definition I adopt, is illustrated in Figure 1.1. There are four chromosome sequences, which label the leaves of the ARG, and are written as strings of 0s and 1s (coding SNP alleles). Moving back in time (up the ARG), the first event we encounter is a mutation. A mutation is denoted by a black dot and a number specifying its marker position. The second event is a recombination between markers 2 and 3. Working back in time, this corresponds to splitting the lineage into two, with the alleles at positions 1 and 2 following

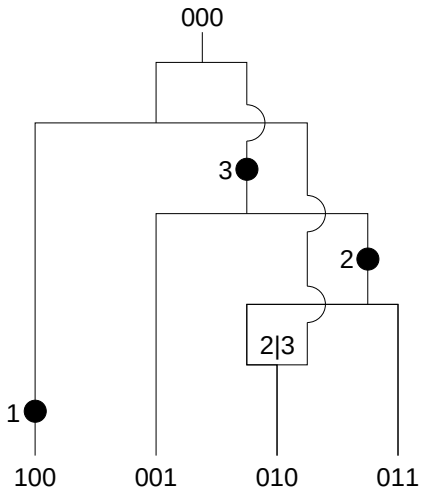


Figure 1.1: An Ancestral Recombination Graph

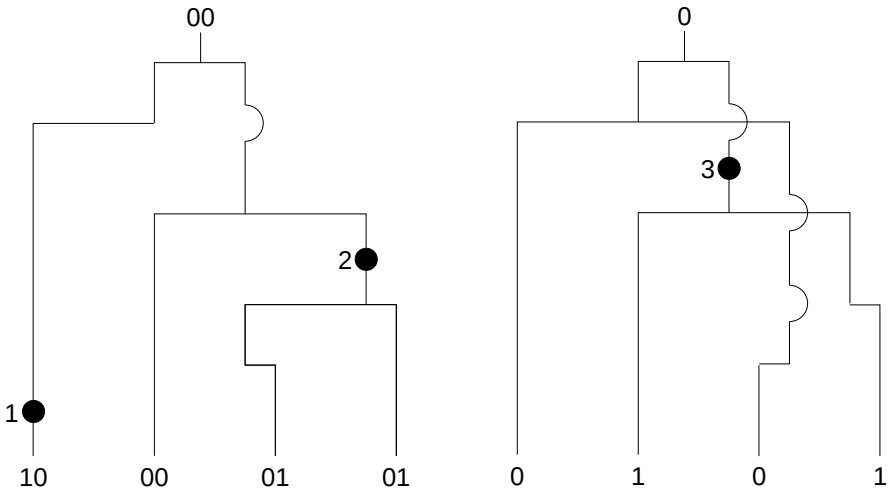


Figure 1.2: Marginal trees for the ARG in Figure 1.1.

the left lineage, and the allele at position 3 following the right lineage. Following this is a coalescence, merging two lineages into one, and so on, to the grand common ancestor.

For each marker, there is a coalescent tree embedded in the ARG—called a marginal tree. Moving along the chromosome, the topologies of consecutive marginal trees shift according to the impact of historical recombination events. The recombination events define the chromosomal region that each marginal tree spans, and since many recombination events have occurred in population history, the resolution is very fine. Figure 1.2 illustrates this. In fact, shifts in tree topology are entirely dependent on the positions of observable recombinations;

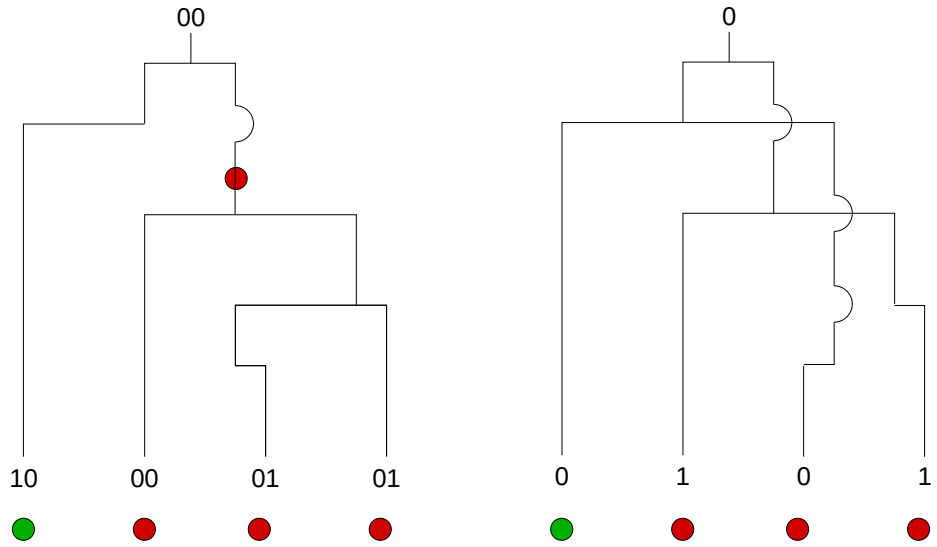


Figure 1.3: Disease mapping using the ARG. Suppose chromosomes denoted by a red dot are from disease individuals, then the branch with the red mutation shows the strongest clustering of cases beneath it.

so unlike the pairwise r^2 measure of LD, a measure of marginal tree correlation is a pure measure of (observable) recombination distance under the infinite sites model. This hints towards that idea that an association test that relies on marginal trees, rather than r^2 LD, might give more accurate positioning of causative polymorphisms.

The marginal trees for the ARG in Figure 1.1 are given in Figure 1.2. On the left is the marginal tree for the SNPs at positions 1 and 2; and on the right is the marginal tree for the SNP at position 3. For a given position, the marginal tree can be extracted from the ARG by tracing the genealogy of that position back in time from the leaves. When a recombination is encountered, the genealogy follows the path of the left recombination parent if the breakpoint is to the right of the position in question, and otherwise it follows the right parent.

If there is a disease-predisposing mutation at a particular chromosomal location, it would have occurred on some internal branch of the marginal tree at that location. So one way to find disease associations is to scan across the marginal trees looking for those with branches that discriminate well between cases and controls, i.e., have a large number of cases beneath them and significantly fewer controls. Such a clustering of the cases underneath a branch suggests that a causative mutation arose on that branch (see Figure 1.3).

If the true ARG were known, it would provide the optimal amount of information for mapping because it would fully describe the locations of recombinations and co-inheritance of genetic material—no extra information would be available from the genotypes. While performing a chi-square test on case-control data will only identify an association if a typed marker is in strong LD with the causative mutation, ideal ARG based mapping has no such requirement. Not only could disease-associated regions be identified, additionally the ARG would give the ages of the causative mutations, would specify the haplotypic background of those mutations, and so forth. It would also be possible to optimally impute missing data. But unfortunately, the true ARG is unknowable; there are infinitely many ARGs compatible with any set of genotype data, and although some are more likely than others, there are very many ARGs of comparable likelihood (McVean and Cardin, 2005; Song et al., 2006).

1.5 The Coalescent-with-Recombination

The distribution of ARG topologies under the Wright-Fisher model with recombination is described by a stochastic process called the coalescent-with-recombination (Hudson, 1983; Griffiths and Marjoram, 1997) (although note, as discussed earlier, that in Griffiths and Marjoram (1997) and others, the term Ancestral Recombination Graph is used to describe the stochastic process, rather than, as I use it, the graph structure describing a genealogical history with recombinations, coalescences and mutations).

The Wright-Fisher model (Wright, 1931) describes the transmission of genetic material in an idealised population as follows: Consider a population of N haploid individuals, and a single locus. Each individual in the next generation comes from sampling an individual, with replacement, from the previous generation. This is repeated until N individuals are sampled. This process assumes:

1. That the population size remains constant;

2. That the individuals are haploid;

3. That the generations are discrete and non-overlapping;

4. That all individuals are equally fit;

5. That there is no population substructure; and

6. That there is no recombination.

Mutations may be added into the process as follows: the locus is transmitted with a mutation with probability u, and is copied without mutation with probability $1 - u$. There are a number of mutation models in the population genetics literature, and the one used throughout this thesis is the infinite sites model, which specifies that every mutation event occurs at a unique position, meaning there are never back or recurrent mutations or SNPs with more than two alleles.

As described, it is conceptually straightforward to simulate a population by running this model forward in time. During the process, many of the lineages die out, failing to contribute genetic material to subsequent generations. This means that the genealogy of the population can be described by a tree, with a most recent common ancestor (MRCA) and the most recent generation forming the leaves. The distribution of these trees, including their branch lengths, is described by the coalescent model (Kingman, 1982).

While the Wright-Fisher process works forward in time, the coalescent model takes a backward in time approach. To simulate a population genealogy, a tree is sampled, starting with the leaves and coalescing lineages. Because there is (assumed to be) no selection, lineages choose their parents at random from the previous generation. When two lineages choose the same parent, the lineages coalesce, which under the Wright-Fisher model has probability $1/N$. In the coalescent tree, the number of observed lineages reduces with coalescence events, meaning that the rate of coalescence slows further back in time. This defines the distribution of branch lengths for coalescent trees. Coalescence events are sampled until the MRCA is reached.

Since selectively neutral mutations do not affect the probability of transmission from one generation to the next, the mutation process is independent of the genealogical process, and hence mutations can be added on the branches of the tree after it has been simulated. This gives a very efficient way to simulate populations, as, unlike with the forward in time

Wright-Fisher approach, it is not necessary to simulate those lineages which do not ultimately contribute genetic material to the sampled population.

The coalescent model can be extended to include recombination, and this is called the coalescent-with-recombination (Hudson, 1983; Griffiths and Marjoram, 1997). This describes the distribution of ARGs under the Wright-Fisher model with recombination. In addition to coalescence events, recombination events are also simulated, which result in splitting a lineage—a coalescence event reduces the number of lineages by one, and a recombination event increases the number by one. Nevertheless, the process does terminate with a single MRCA because the rate of coalescence exceeds that of recombination.

In theory, coalescent models can be used in a full likelihood framework to perform inference of population genetic parameters, such as mutation and recombination rates, as well as to fine map disease alleles (Larribe et al., 2002). In full likelihood methods, the probabilities of observing the data given the coalescent model and parameters are estimated, and maximum-likelihood estimates of the parameters are taken as those that maximise the probability of observing the data.

The likelihood surface is described as

$$L\left(\theta|D\right) = P\left(D|\theta\right) = \int P\left(D|G,\theta\right) P\left(G|\,\theta\right)\mathrm{d}G,$$

where θ are the coalescent model parameters, D is the data and G is the unknown genealogy. In order to evaluate this integral, simulation methods are required for all but the smallest of data sets (Griffiths and Marjoram, 1996). Genealogies $G^{(i)}$ are simulated from the coalescent $P\left(G|\theta\right)$, and Monte Carlo integration can be applied:

$$\int P\left(D|G,\theta\right) P\left(G|\,\theta\right)\mathrm{d}G \approx \frac{1}{M}\sum_{i=1}^{M} P\left(D|G^{(i)},\theta\right).$$

However, many of the genealogies will not contribute significantly to the sum in the Monte Carlo integration; many sampled genealogies will not fit the data, and there are infinitely many ARGs which do fit the data, very many of which are of comparable, small, likelihood (McVean and Cardin, 2005; Song et al., 2006). Therefore it is typical to focus the sampling of genealogies

using Importance Sampling (Griffiths and Tavaré, 1994a,b,c; Griffiths and Marjoram, 1996; Tavaré et al., 1997) and Markov Chain Monte Carlo methods (Wilson and Balding, 1998; Beaumont, 1999; Kuhner et al., 2000; Nielsen, 2000). Nevertheless, this approach to inference under the coalescence-with-recombination remains computationally prohibitive for all but the smallest of data sets, rendering such methods impractical for analysis of large scale association study data.

The computational challenges involved in coalescent-based inference have partly motivated the development of faster methods that approximate the coalescent-with-recombination.

1.6 Approximations to the Coalescent-with-Recombination

One approach to speed the calculation of an approximate likelihood is to consider small subsets of the data in turn. Specifically, methods have been developed that use two-locus systems (Hudson, 2001; McVean et al., 2002). For each pair of loci, a likelihood surface is calculated and a composite likelihood is obtained by multiplying all pairwise likelihoods. This approach is fast, in part because two-locus systems can be fully enumerated, and results stored in a look-up table.

Another strategy is to discard the genealogy but maintain important properties of the coalescent-with-recombination model, for example, Fearnhead and Donnelly (2001) and Li and Stephens (2003). The Li and Stephens (2003) model relates the distribution of sampled haplotypes $h_1, \ldots, h_n$ to the recombination rate ρ as

$$P\left(h_1, \ldots, h_n|\rho\right) = P\left(h_1|\rho\right) P\left(h_2|h_1, \rho\right) \cdots P\left(h_n|h_1, \ldots, h_{n-1}, \rho\right).$$

This expresses the likelihood in terms of a product of conditional probabilities. These conditional probabilities are amenable to approximation, which in turn gives an approximation for the distribution of the data given the recombination rate and model. This is called a product of approximate conditional likelihoods (PAC) model.

Fearnhead and Donnelly (2001) and Li and Stephens (2003) propose approximations for the conditional distributions that capture important population genetic features. As listed in

Li and Stephens (2003), these are:

1. The next haplotype is more likely to match a frequently observed haplotype than a rare one;

2. The probability of observing a novel haplotype decreases with the number of observed haplotypes;

3. The probability of seeing a novel haplotype increases with the mutation rate parameter;

4. Novel haplotypes will tend to be imperfect copies of previously seen haplotypes, rather than entirely novel; and

5. Because of recombination, the next haplotype will look like previous haplotypes over contiguous regions.

The basic process for generating h_{k+1} from previously observed haplotypes $h_1, \ldots, h_k$ is as follows: h_{k+1} is assumed to be related to the previously observed haplotypes by some shared ancestry, and so can be constructed by copying, with mutation, parts of the $h_1, \ldots, h_k$. This represents h_{k+1} as a mosaic of $h_1, \ldots, h_k$. The copying process works along the chromosome, and jumps between the $h_1, \ldots, h_k$ according to the recombination rate ρ. Thus computing $P(h_{k+1}|h_1, \ldots, h_k, \rho)$ is achieved by summing over all possible mosaics of $h_1, \ldots, h_k$. Since the copying process along the chromosome is Markov, this calculation can be done using standard theory for Markov models.

Many of these methods have been developed in order to estimate recombination rates and detect recombination hotspots (McVean, 2002; Fearnhead et al., 2004), or to resolve the genotype sequences of diploid individuals into their two haplotype sequences, a process known as phasing (Stephens et al., 2001).

Another approach is to keep the genealogies, but discard much of probabilistic model (Templeton et al., 1987; Molitor et al., 2003; Durrant et al., 2004; Halperin and Eskin, 2004; Templeton et al., 2005; Zollner and Pritchard, 2005; Waldron et al., 2006). These methods often work by estimating the local genealogical tree within a haplotype block or for a single locus by clustering the haplotype sequences into a cladogram. This approach is typically used

for fine scale mapping rather than estimating population genetic parameters, which brings us back to our original application.

A cladogram defines nested partitions of haplotypes, with each subsequent partition bringing together increasingly diverse haplotype sequences. In Durrant et al. (2004), for example, the cladogram is constructed using hierarchical group averaging. Initially, each haplotype is its own singleton cluster. Successive clusters are merged so that the mean pairwise haplotype diversity within the new cluster is minimised. Durrant et al. (2004) measure haplotype similarity in such a way that haplotypes sharing rarer alleles are treated as more similar, the motivation for this being that rarer alleles are likely to indicate a more recent common ancestor. Using such a similarity metric, averaged over the markers, does not account for the fact that recombination will mean that at different positions along the chromosome, the genealogical distance between haplotypes, and thus the clustering, will be different. Therefore, the cladogram is not constructed for the whole chromosome. Rather, Durrant et al. (2004) take a sliding window approach, calculating the cladogram for small windows of SNPs, corresponding to tens of kb at a time.

Such cladograms can then be used for disease mapping by testing each cluster for disease association, that is, testing the hypothesis that the haplotypes within that cluster harbour a causative allele. If a cladogram has an associated cluster, then this may lead us to conclude that a causative allele resides in the region spanned by that cladogram.

However, compared to the ARG, cladograms are a coarse approximation of population evolution, and there is often difficulty in modelling the relationships between similar haplotypes and handling rare haplotypes. Additionally, it is often assumed that haplotypes are observed directly (that is, the data is phased) and that one can define non-recombining haplotype blocks, which is in general not the case.

1.7 Contributions of this Thesis

In Chapter 2 I describe an algorithm for constructing ARGs from population genotype data, which may be unphased and have missing genotypes. It has computational efficiency nearing

that of haplotype clustering methods, and can be applied to thousands of individuals typed for SNPs across regions up to 1 Mb.

In Chapter 3 I show how inferred ARGs can be used to analyse case-control association study data. In particular, how they can be applied to fine mapping and interpretation of a signal at a potentially associated locus. The algorithm is compared to the single marker chi-square test and a haplotype clustering method (Durrant et al., 2004). Compared to these methods, the new ARG-based approach achieves significant increases in:

1. Power (ability to correctly say whether there is a causative allele in a given region);
2. Localisation (ability to finely map that causative allele); and
3. Interpretation (inferring properties of the causative allele, such as its frequency, which can guide further investigation).

In Chapters 4 and 5 I describe applications of the method to two disease data sets: one is a case-control study of 1306 individuals densely typed over a 300 kb region for association with Graves Disease; the second is a case-control study for Prostate Cancer, involving 1329 individuals typed over an 800 kb region. In both cases, the method is able to draw interesting observations from the data that may be missed using other methods. In the Graves disease data set, a potential epistatic interaction is identified, and in the Prostate Cancer data, the association peak is separated into two independent effects by identifying a recombination hotspot.

In Chapter 6 the method is extended to tackle a related problem, that of inferring missing data. I describe an approach where case-control association studies are merged with more densely typed data (such as the HapMap), allowing the SNPs typed in the dense data set to be imputed in the cases and controls and tested directly for association.

Chapter 7 extends beyond the question of disease mapping, and I describe other population genetic problems to which inferred ARGs could be applied. The thesis concludes with a brief summary in Chapter 8.

Chapter 2

ARG Inference

2.1 Discrete ARG Methods

In this chapter I describe a novel method for explicitly reconstructing ARGs from population genotype data. There are broadly two approaches to ARG based inference from population genetic data: methods based on the coalescent-with-recombination, and methods based on a discrete representation of the ARG. Discrete methods define the ARG as the genealogical topology that relates a set of sequences, and do not necessarily define a statistical model for those topologies. The method described here falls into that category.

The majority of work from the discrete perspective relates to finding the minimum number of recombination events required to derive a sample of sequences (Hudson and Kaplan, 1985; Myers and Griffiths, 2003; Wiuf, 2004; Song and Hein, 2004, 2005; Lyngsø et al., 2005; Bafna and Bansal, 2006). It is impossible to find the true number of recombination events as many recombinations are silent and leave no trace in the sample, but the minimum can be provably found when sufficient, and generally prohibitive (Wang et al., 2001), computational effort is applied. It is useful to know the minimum number of obligate recombination events, and their positions, when, say, exploring whether there has been mitochondrial recombination (Zsurka et al., 2005). The motivation often given in papers on this topic is towards detecting recombination hotspots; in Fearnhead et al. (2004) the obligate recombination inference method of Myers and Griffiths (2003) is shown to give results consistent with the presence of hotspots,

as inferred by a likelihood method and detected by sperm typing.

In Hudson and Kaplan (1985); Myers and Griffiths (2003); Wiuf (2004); Song and Hein (2004); Bafna and Bansal (2006) and Gusfield et al. (2007) lower bounds are developed for the number of recombination events required in the genealogy of a sample of sequences.

Hudson's method (Hudson and Kaplan, 1985) is based around what is known as the four gamete test. The method infers a recombination event between a pair of SNPs when all four possible gametic types are present (00, 01, 10 and 11) in the population. Under the infinite sites model, a recombination must have occurred; there is no tree that can describe such a configuration of haplotypes. The method tests all pair of SNPs in the region to construct a collection of intervals, within each of which there is an obligate recombination event. Under the conservative assumption that overlapping intervals correspond to the same recombination event, the algorithm finds the largest subset of non-overlapping intervals from the collection, and the lower bound is given as the number of intervals.

The four gamete test becomes less accurate as the recombination rate and sample size increase. The bound can be improved by considering haplotypes rather than pairs of SNPs, and by moving towards a genealogical framework, as in Myers and Griffiths (2003), where two methods are presented. The first counts the number of haplotypes within a local window and relates this to the number of recombination events, and then combines local estimates via dynamical programming in order to achieve a lower bound for the whole region. The second method derives local estimates by taking the sequences and performing coalescences and mutations, until no such further operations are possible. A sequence is then removed from the sample, and this corresponds to at least one recombination event. The algorithm proceeds in this way until only one sequence remains. The algorithm performs a search over these "histories". Again, local estimates are combined into a lower bound for the whole region by dynamical programming.

A more sophisticated and computationally intensive approach is to attempt explicit construction of minimal ARGs (Song and Hein, 2005; Lyngsø et al., 2005).

Song and Hein (2005) finds a sequence of trees, with a tree for each marker, and the recombination events required to shift the tree at one marker into the tree at the next.

To do this, all possible trees are constructed for each marker. Then, the recombination events required to move from every tree at one marker, to every tree at the next marker are computed. The minimum number of recombination events, and the minimal ARG(s) can then be constructed by following the path through the markers that minimises the number of recombinations. This method can handle at most nine sequences with current computing power.

Lyngsø et al. (2005) takes a branch and bound approach and offers a significant improvement in speed and memory requirements, although is still limited to tens of sequences. Rather than working left-to-right along the sequence, this method works backwards in time, applying mutation, coalescence and recombination events until a single grand common ancestor is reached. A search is performed over the possible sequences of events, attempting to find a history with a given number of recombination events. If this does not exist, the number of permitted recombination events is increased by one, and so on, until an ARG is found.

2.2 Motivation for the Method

In order to be applicable to the current cohort of case-control association studies, any method must be computationally feasible when applied to data involving thousands of individuals typed for hundreds of thousands of markers, and where the data is unphased and has missing genotypes. However, the minimal ARG methods described above are limited in that they can only be applied to the smallest of datasets; underlying this is the fact that finding the minimum number of recombination events is NP-Hard (Wang et al., 2001). Additionally, such methods often require the sequence data to be phased, which is potentially a problem because the way in which the data is phased will affect the minimum number of recombination events.

The method described here is able to explicitly construct ARGs for thousands of individuals, typed at hundreds of SNPs. This is achieved by removing the requirement that minimal ARGs are inferred. The algorithm can also handle unphased and missing data.

As discussed in Chapter 1, there are computationally intensive statistical methods based on approximations to the coalescent-with-recombination, which can often be applied to large

data sets. However, these approximate methods do not explicitly construct ARG topologies. The algorithm described here is heuristic, and I do not claim to sample from the coalescent-with-recombination, but rather, I attempt to show that the algorithm infers plausible ARGs.

My contribution in this chapter is therefore a method that explicitly constructs ARGs while remaining applicable to large-scale population genotype data. While the method falls into the line of discrete ARG construction methods, its use is not to construct minimal ARGs. Rather, in subsequent chapters, it is used to tackle population genetics questions that are often the domain of statistical methods.

The algorithm described here has been implemented in a program called MARGARITA (Minichiello and Durbin, 2006), which is available for download from
http://www.sanger.ac.uk/Software/analysis/margarita/

2.3 ARG Inference

In order to help develop an intuition for how the ARG inference algorithm works, I first give an informal description.

The goal of the algorithm is to coalesce all sequences into a grand common ancestor. Since a coalescence event corresponds to an ancestral sequence transmitting two copies of itself, a coalescence can only be made when two sequences are identical.

In order to make two sequences identical, mutations can be added, flipping any conflicting alleles into agreement. However, under the infinite sites model, a mutation can only occur once in the history of a SNP, meaning that there will be exactly one mutation in the ARG for every polymorphic site. Consequently, a mutation can only be applied at a position when there is only one copy of the mutant allele in the sequences. By applying coalescence events, the number of copies of an allele is gradually reduced to one, allowing this.

However, coalescence and mutation alone will not always be sufficient to arrive at a grand common ancestor. By applying a recombination event to a sequence, the sequence is split onto two separate lineages. On the left parent of the recombination event, the genetic material to the right of the recombination breakpoint is undefined because it was not successfully

transmitted into the sampled sequences. This means that while a recombination event increases the number of lineages by one, the actual amount of known genetic material remains constant. Furthermore, since there is no information on what was contained in the untransmitted regions, we let those regions coalesce with anything. This means that it may then be possible to coalesce a recombination parent with another sequence, where there was in the child a mismatching allele to the other side of the recombination.

Figure 2.1 illustrates this logic:

- A. Four sequences sampled from the contemporary population.
- B. It is possible to remove singleton alleles with a mutation event back to the ancestral allele.
- C. Since none of the sequences are identical, no coalescences are possible; in addition, no further mutations are possible. However, the second and third sequence are identical at the first two positions. Putting a recombination on the third sequence between the second and third positions will subsequently allow a coalescence.
- D. Undefined genetic material can coalesce with anything, allowing 01**.** and 011 to coalesce. Note that **..**0 and 110 can also be coalesced, but the ordering is chosen at random.
- E. After the coalescence event, there is only one copy of the 1 allele at the second position, allowing this to be mutated to 0.
- F. Two sequences are identical, allowing a coalescence.
- G-I. Further mutations and coalescences are performed until a grand common ancestor is reached.

Compared to other discrete ARG methods, the methods works from the bottom up, that is, backward in time, as in Myers and Griffiths (2003) and Lyngsø et al. (2005), rather from left to right across the sequences, as in Song and Hein (2004). The independently developed approach of Lyngsø et al. (2005) is the most similar, the fundamental difference being that

their method performs a branch and bound search over the ARG construction operations in order to minimise the number of recombinations.

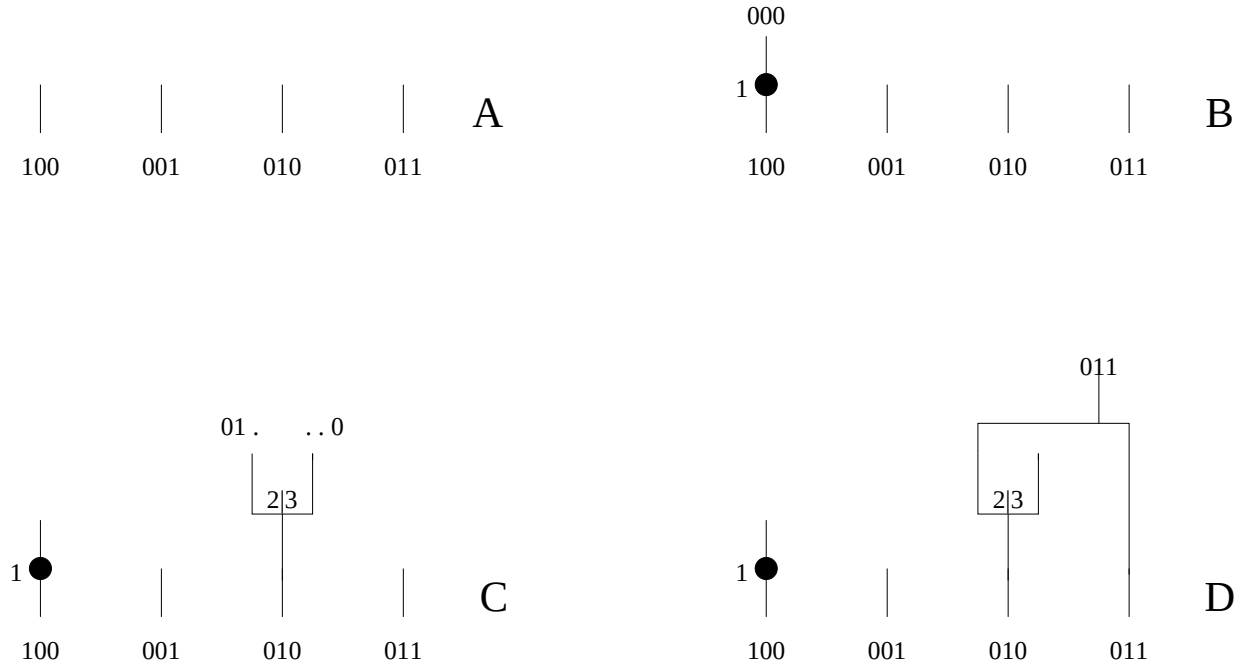
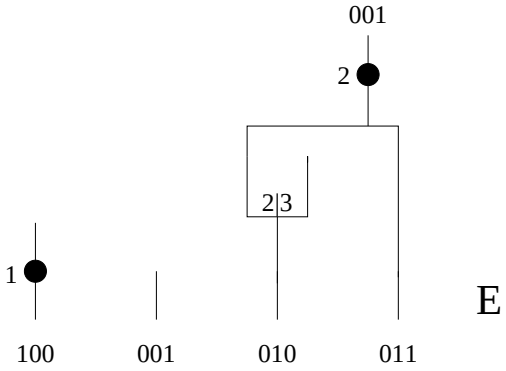
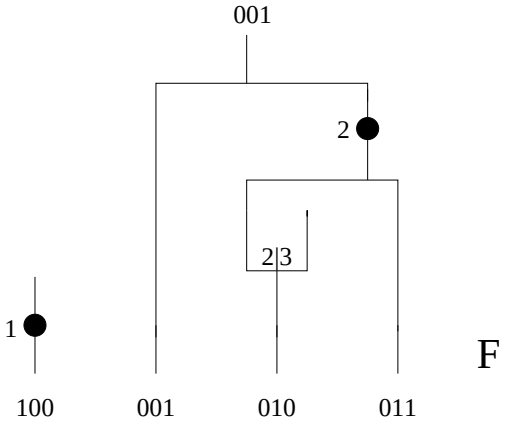


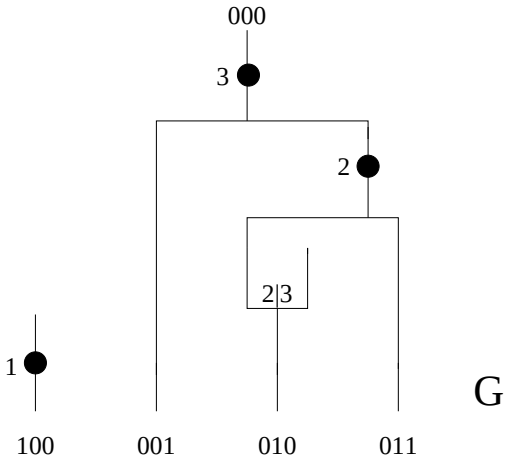
Figure 2.1: Constructing an ARG, see the text for a description. Figure continued on next page.


001
2
2|3
1
100
001
010
011

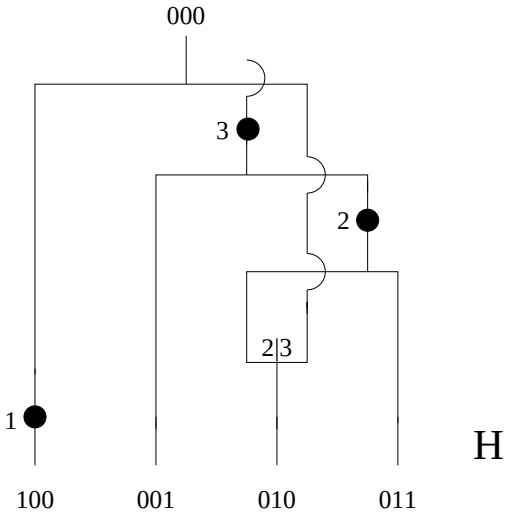
E


001
2
2|3
1
100
001
010
011

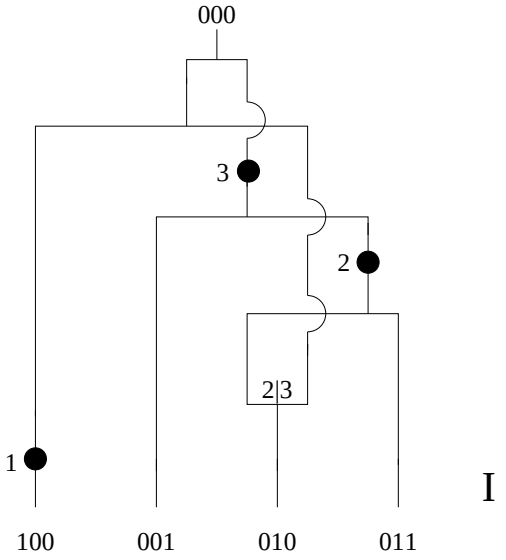
F


000
3
2
2|3
1
100
001
010
011

G


000
3
2
2|3
1
100
001
010
011

H


000
3
2
2|3
1
100
001
010
011

I

2.4 ARG Inference Algorithm

I now describe the algorithm precisely.

The algorithm works backwards in time from the contemporary, typed population of chromosome sequences to a single ancestor sequence. Each step back in time, accomplished with a recombination, mutation or coalescence, defines an ancestral population of sequences. We denote the set of sequences at time T as S_T, and the sequences are, in the phase-known case, strings of length m from the alphabet $\{0, 1, .\}$, where m is the number of markers, 0 is one of the SNP alleles, 1 is the other allele and . denotes an undefined copy of the allele—undefined because it was not inherited by any sequences in the contemporary, typed population.

The allelic state of a SNP on sequence C is denoted $C[i]$, where i is the marker position, numbered from 1, so $1 \leq i \leq m$. We define $C_1[i] \sim C_2[i]$ if and only if $C_1[i] = C_2[i]$ or $C_1[i] = .$ or $C_2[i] = .$. We define a complement operator $\neg$ such that if $C[i] = 0$ then $\neg C[i] = 1$ and vice versa, and . is its own complement.

There is a shared tract between sequences C_1 and C_2, over the contiguous set of markers $a, \ldots, b$, if :

1. $C_1[i] \sim C_2[i]$ for all $a \leq i \leq b$;
2. there is at least one i for which $C_1[i] = C_2[i] \neq .$;
3. if $a > 1$ then $C_1[a-1] \neq C_2[a-1]$ and neither are .;
4. if $b < m$ then $C_1[b+1] \neq C_2[b+1]$ and neither are ..

(1) requires that the two sequences have the same allelic state over the shared tract; (2) requires that for at least one position in the tract, both sequences are defined; and (3) and (4) require that the shared tract is maximal. We denote such a shared tract as $\{C_1, C_2\}[a, b]$.

The algorithm is initialised at time $T = 1$ (T is incremented as we move back in time) by setting S_1 to be the set of contemporary, typed sequences. The algorithm proceeds by finding which coalescences, mutations and recombinations can be performed, determining this according to the rules below. Applying one of these operations defines an ancestral population

S_{T+1}, which is constructed from S_T using the transitions also described below. The algorithm continues in this way until it arrives at a population with only one sequence.

- **Coalescence.**

  Rule. If there exist two sequences C_1 and C_2 in S_T such that for all i, $C_1[i] \sim C_2[i]$, then C_1 and C_2 can be coalesced into an ancestor.

  Transition. $S_{T+1} = (S_T \setminus \{C_1, C_2\}) \cup \{C'\}$ where $C'[i] = C_1[i]$ when $C_1[i] \neq \blacksquare$ and $C'[i] = C_2[i]$ otherwise. (By $(S_T \setminus \{C_1, C_2\}) \cup \{C'\}$ we mean S_T with the sequences C_1 and C_2 removed and the sequence C' added in.)

- **Mutation.**

  Rule. If there exists a sequence C_1 in S_T and a marker i, where for all C_2 in $S_T \setminus \{C_1\}$ we have $C_2[i]$ is $\neg C_1[i]$ or $\blacksquare$, then we can remove the derived allele ($C_1[i]$) from the population.

  Transition. $S_{T+1} = (S_T \setminus \{C_1\}) \cup \{C'\}$ where $C'[i] = \neg C_1[i]$ and $C'[j] = C_1[j]$ for all $j \neq i$.

- **Recombination.**

  Rule. When the rules for coalescence and mutation are not satisfied, we must perform a recombination (or a pair of recombinations) instead. We denote a recombination breakpoint as (α, β) meaning that it occurs between markers α and β. Picking a shared tract $\{C_1, C_2\}[a, b]$ from those available in S_T, the aim becomes putting recombinations on the lineages of C_1 and C_2 such that one recombination parent of C_1 and one recombination parent of C_2 satisfy the rule for coalescence. To do this, we must put a breakpoint at $(a-1, a)$ if $a \neq 1$, and a breakpoint at $(b, b+1)$ if $b \neq m$.

  Transition. From the tract $\{C_1, C_2\}[a, b]$, pick (1) a valid breakpoint (α, β), either $(\alpha, \beta) = (a-1, a)$ or $(\alpha, \beta) = (b, b+1)$; and (2) a recombinant sequence C_R, either $C_R = C_1$ or $C_R = C_2$. Then $S_{T+1} = (S_T \setminus \{C_R\}) \cup \{C'_1, C'_2\}$ where $C'_1[i] = C_R[i]$ for $i \leq \alpha$ and $C'_1[i] = \blacksquare$ otherwise; and $C'_2[i] = C_R[i]$ for all $i \geq \beta$ and $C'_2[i] = \blacksquare$ otherwise. If both $(a-1, a)$ and $(b, b+1)$ are valid breakpoints (that is, $a \neq 1$ and $b \neq m$), we must put the second recombination (taking us to state S_{T+2}) on an appropriate ancestor of

C_1 or C_2. See Figure 2.1 for an example.

These rules define the constraints on the algorithm that must be enforced if it is to produce legal ARGs. However, at any stage of the algorithm there may be a number of different coalescences, mutations or recombinations that satisfy the rules. We choose between these using the heuristics below, and the stochastic elements result in novel ARGs being generated each time the algorithm is run.

1. **Only perform a recombination if no mutations or coalescences are possible.**

2. **If it is possible to add multiple mutations and/or multiple coalescences at the same time, the order in which these are done is chosen arbitrarily.**

3. **Only coalesce sequences if they have an overlapping region of defined material,** i.e. the two sequences must match for at least one position that is not **.**. This restriction reflects ideas in the sequentially-Markovian coalescent-with-recombination model (McVean and Cardin, 2005).

4. **Recombinations are added at the ends of longer shared tracts first.** During the recombination step, I choose a shared tract $\{C_1, C_2\}[a, b]$ such that the base pair distance between markers a and b is maximised. I only use this heuristic a user-specified proportion of the time $plong$, and at other times $(1 - plong)$ a randomly selected shared tract is used.

5. **The first coalescence after a recombination is based on the shared tract that was used to decide the location of that recombination.**

Heuristic 1 was chosen in order to produce more parsimonious ARGs (fewer recombination events). Heuristic 2 was chosen in order to increase the stochastic element of the algorithm and thus the space of ARGs explored. Heuristic 3 was chosen in order to simplify the system. In McVean and Cardin (2005) it was shown that simulation using this restriction does not cause significant departure from simulations resulting from the standard coalescent-with-recombination. Heuristic 4 reflects the fact that longer shared tracts will tend to arise from

more recent recombination and coalescence events. However, because this is only a tendency, not absolute, I only use this heuristic a certain proportion of the time *plong*, and break this heuristic with probability $1 - plong$, when I instead use a randomly selected shared tract to position the recombination breakpoint(s). Throughout this thesis $plong = 0.9$, except as discussed in the next section. Heuristic 5 follows on from Heuristic 4 and again ensures that longer shared tracts are coalesced earlier.

There are a number of other heuristics which could be used. For example, basing the shared tract selection on genetic distance rather than physical distance. Another option would be to select shared tracts which terminate at the positions where recombinations have already been inferred in the ARG. This could be done iteratively, thereby making a preference to place recombinations in hotspots. Additionally, the ancestral sequence at the root of the ARG could be fixed to that suggested by, for example, Chimp data. Additionally, the heuristics could be modified to allow errors in genotyping, or multiple mutations at the same site.

In order to help guide the selection of suitable heuristics and their parameters, I used the set of small case-control studies simulated in Zollner and Pritchard (2005). I evaluated the fine mapping performance for ARGs inferred using different heuristics and adopted those which are (1) genetically sensible and (2) result in discernibly better mapping performance (mapping using inferred ARGs is described in the next chapter). In the next section I show the effect of varying the *plong* heuristic parameter on ARG inference.

2.5 Comparison to the Coalescent-with-Recombination

In this section I compare the ARGs inferred using the above algorithm, which I implemented in a program called MARGARITA, to those generated under the neutral coalescent-with-recombination.

I simulated samples of haplotype sequences using Hudson's MS program (Hudson, 2002), and compared:

- The true marginal trees from MS, which describe the true genealogical history underlying the sample, with the marginal trees inferred by MARGARITA; and

- The number of non-recombining segments as reported by MS (giving a lower bound on the true number of recombinations), with the number of recombinations inferred by MARGARITA.

Note that I compared marginal trees because MS does not output the full ARG. Also, MS does not output the total number of recombination events in the sample history, it only outputs the number of non-recombining segments, giving a lower bound on the true number of recombinations (number of segments - 1).

Using MS I simulated 100 samples of 100 haplotypes, over a 10kb region with a recombination rate of 2.2×10^{-9} per generation per nucleotide and a mutation rate of 1.1×10^{-9} per generation per nucleotide, and an effective population size of 4×10^6, giving a population scaled recombination rate for region of $\rho = 88$, and a scaled mutation rate of $\theta = 44$, and on average 230 segregating sites.

In order to compare MARGARITA's inferred marginal trees with those from MS, I calculated a tree correlation score for each true tree with the inferred trees at that position. Let $\mathcal{B}(T)$ be the set of all non-unary, inequivalent bipartitions of the leaves for a binary tree T. Each bipartition is obtained by cutting an internal branch of the tree and seeing which leaves fall under the cut, and which do not. There are $n - 3$ such bipartitions. The tree correlation score between trees T_1 and T_2 is then:

$$\frac{|\mathcal{B}(T_1) \cap \mathcal{B}(T_2)|}{n-3},$$

which is the proportion of bipartitions that are shared between the two trees. I use this metric because it is directly related to disease mapping using inferred ARGs. In my approach to disease mapping, described in the next chapter, hypothetical disease causing mutations are placed on marginal trees, and these mutations partition the leaves into two non-overlapping sets: those chromosomes with or without the putative causative mutation. The accuracy in partitioning reflects the accuracy of causative mutation inference.

For one simulation, Figure 2.2 shows the tree correlation of the inferred marginal tree at each segregating site with the true marginal tree for that position, averaged over 100 ARG

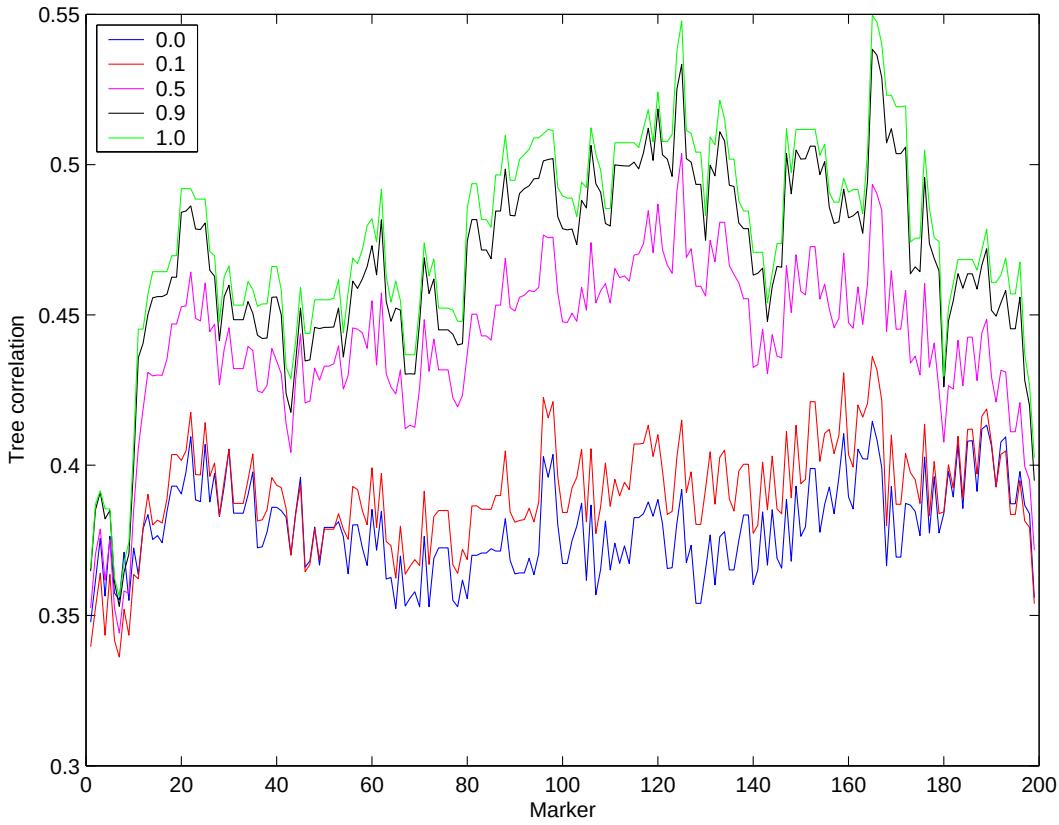


Figure 2.2: Tree correlation of the inferred marginal tree with the true tree simulated by MS at each segregating site, averaged over 100 inferred ARGs. This is for one simulated population, with 100 haplotypes over 10kb, with $\rho = 88$ and $\theta = 44$, with 199 segregating sites. Each line corresponds to a different value of the heuristic parameter *plong*.

inferences. Five lines are plotted, each one corresponding to a different value of *plong*, the heuristic parameter which specifies the frequency with which a longest shared tract is used to position recombination events, rather than a randomly selected shared tract.

Marginal tree reconstruction is more accurate (higher tree correlation) for greater values of *plong*, indicating that the longest shared tract heuristic appropriately captures features from the neutral coalescent-with-recombination. Also note that the accuracy of marginal tree reconstruction is better away from the edges of the "typed" region. This is expected because there is less information towards the edges about the long range sharing of haplotypes.

Figure 2.3 shows the quality of marginal tree reconstruction for samples of 100 sequences over a 10kb region, with a scaled mutation rate of $\theta = 44$ as above, but now with $\rho \in \{8, 88, 880\}$. The number of haplotypes in each sample was also varied, to be either 30 or 100,

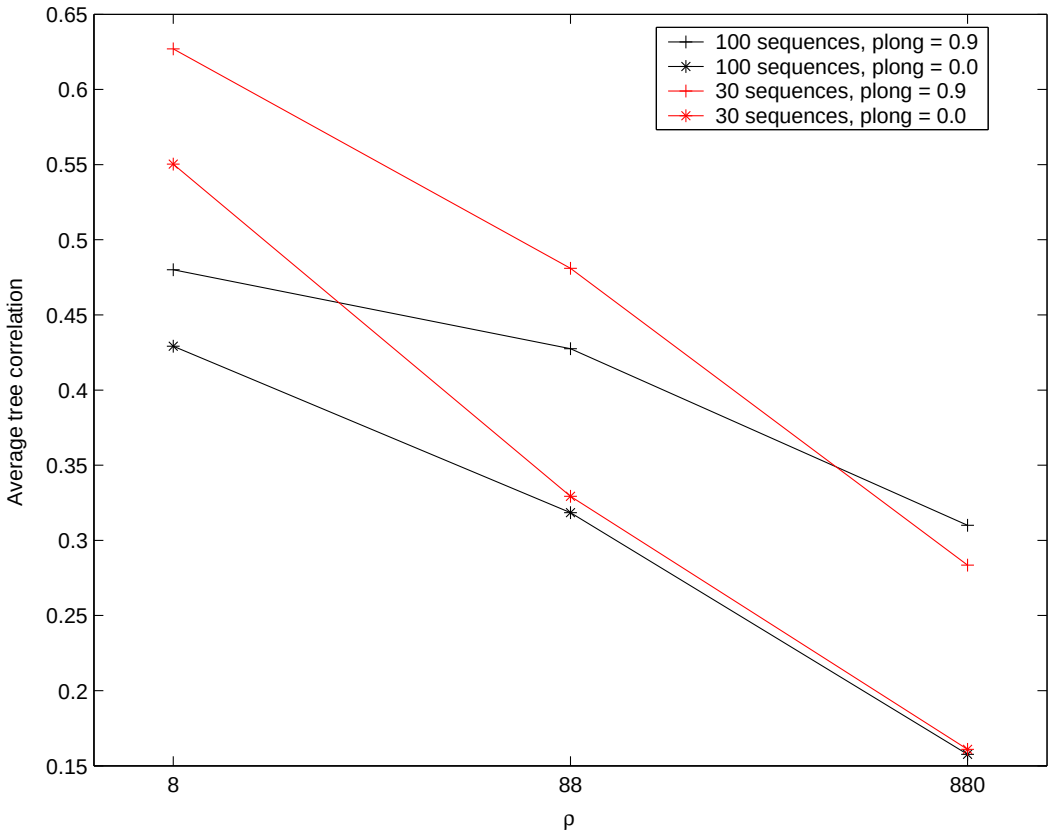


Figure 2.3: Tree correlation of inferred marginal trees with the true trees for a range of population scaled recombination rates ρ, heuristic parameters *plong* and number of sequences. Each point corresponds to the average over all segregating sites in 100 simulations, over 100 ARG inferences for each simulation.

and the heuristic parameter *plong* was set to either 0.9 or 0.0

As the recombination rate increases, the quality of marginal tree reconstruction decreases. This is because there become many more shifts in MS tree topology than segregating sites (θ remains set at 44), and it is harder to observe those recombination events (and their impact on the marginal tree topology) from the data.

Marginal tree reconstruction tends to be more accurate for samples with fewer sequences. This is expected because when more sequences are sampled, which may be very similar, the number of possible tree topologies within each non-recombining region increases, as does the uncertainty in the ordering of coalescence events.

As already observed, $plong = 0.9$ gives more accurate reconstruction than $plong = 0.0$. This helps justify the choice of setting $plong = 0.9$ for the disease mapping experiments in

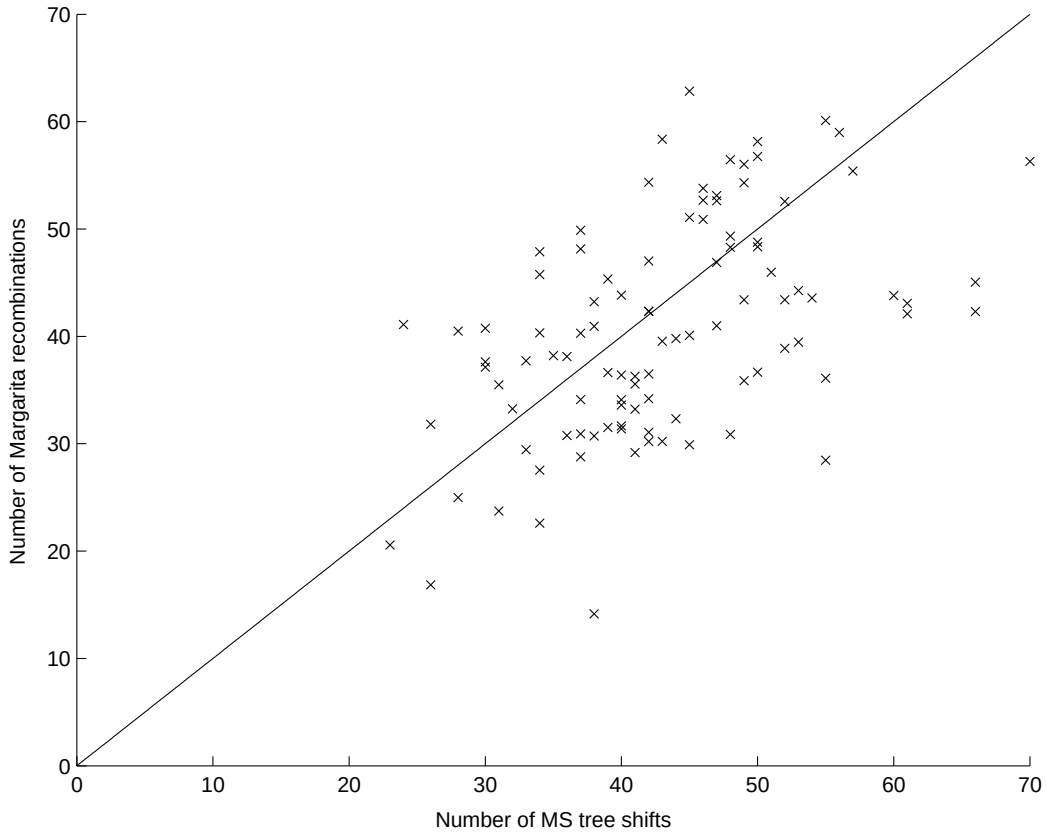


Figure 2.4: The number of recombination events in inferred ARGs from MARGARITA, compared with the number of tree shifts in the true history as simulated by MS. For 100 simulations, each with 100 haplotypes, with $\rho = 8$, $\theta = 44$ and $plong = 0.9$. The line $y = x$ is plotted.

latter chapters.

Figure 2.4 gives the number of recombination events in MARGARITA's inferred ARGs, compared with the number of shifts in tree topology in the true ARGs simulated by MS (the number of non-recombining segments - 1). The number of tree shifts gives a lower bound on the true number of recombination events. The number of inferred recombination events matches the number of tree shifts well (for these simulations with parameters $\rho = 8$, $\theta = 44$). However, the slope of the linear regression line for the data points in Figure 2.4 is 0.54. Part of this underestimation in the number of recombinations may be due to MARGARITA only inferring recombination events when no coalescences or mutations are possible. Also, there will be some recombination events for which there is no evidence in the data, and the proportion of these will increase as the recombination rate increases.

In the following chapters I discuss the application of inferred ARGs to disease mapping. The good mapping performance achieved when using inferred ARGs gives further evidence that they capture correctly many important properties of the true genealogy.

2.6 Existing Methods for Handling Unphased Data

So far I have only considered phase known haplotype sequences, but by extending the algorithm it is possible to resolve haplotype phase and impute missing data while constructing an ARG, and this is described in the next section. First I define more clearly the phase resolution problem and discuss other algorithms for handling unphased data.

In diploid organisms, there are two copies of each autosome. The sequence of a single chromosome is the haplotype sequence, and the genotype sequence is the conflation of those two haplotype sequences. Consider a SNP where there are two alleles, 0 and 1. If both copies of the haplotypes have the 0 (respectively 1) allele, then the genotype sequence will register homozygous 0 (respectively 1). However, if one of the haplotypes has the 0 allele, and the other haplotype the 1 allele, the genotype sequence will register a heterozygote. When genotype sequences are obtained, heterozygote calls do not indicate which chromosome has the 0 allele, and which has the 1 allele; and without further information it is impossible to determine the haplotype sequences.

Although haplotypes can be determined experimentally (Patil et al., 2001), current large-scale sequencing and genotyping technologies only provide genotype sequences, and it is more common to determine haplotype sequences computationally by using information from other individuals in the population (LD).

We write a genotype sequence as a string of 0,1,U characters, where 0 indicates that the individual is homozygous for the 0 allele; 1 indicates the individual is homozygous for the 1 allele; and U indicates that the individual is heterozygous. The haplotype inference problem is to correctly resolve the haplotype sequences from genotype sequences.

There are a number of existing phasing algorithms. An early one was Clark's algorithm (Clark, 1990). This algorithm starts by identifying any genotype sequences that are com-

pletely resolved, having no unphased U characters; this genotype sequence yields two identical haplotype sequences. The algorithm then looks for a match between an unphased genotype sequence and a phased haplotype sequence, that is, where the homozygous positions on the unphased sequence have the same alleles as the phased sequence. When such an instance is found, the unphased sequence is resolved into two haplotype sequences: one is the same as the phased sequence, and the other is set to the complement in the U positions. For example, if the genotyped sequence is 0UU0, and there is a phased haplotype sequence, 0100, then 0UU0 becomes the two haplotype sequences: 0100 and 0010. When a genotype sequence is resolved it is removed from the set of genotype sequences, and the resolved haplotype sequences are added to the set of sequences available for matching to the remaining genotype sequences. The algorithm stops when all genotype sequences have been resolved or when no further phase resolution is possible. With this approach, the order in which sequences are resolved affects the end phased result.

The underlying population genetic model for Clark's algorithm is that sequences in a population are likely to be similar to each other. This approach can also be viewed as an attempt to minimise the total number of haplotypes observed in the sample and, hence, is a parsimony approach.

A more common approach is to use Expectation-Maximisation (EM) to find maximum likelihood estimates for the haplotype phases (Excoffier and Slakin, 1995; Long et al., 1995; Chiano and Clayton, 1998). This overcomes the problems associated with Clark's algorithm, that the haplotype resolution depends on the order in which the algorithm is run, and that the algorithm will not be able to start if all the individuals have ambiguous haplotypes.

The likelihood of a possible phase resolution, assuming Hardy-Weinberg equilibrium, is

$$L = \prod_{h=1}^{n_d} \frac{\pi\left(h_1\right)\pi\left(h_2\right)}{\left(2n_d\right)^2},$$

where n_d is the number of diploid individuals, and the two haplotypes for individual h are h_1 and h_2, and the frequencies of those haplotypes in the population are $\pi\left(h_1\right)$ and $\pi\left(h_2\right)$.

The EM algorithm works iteratively to compute the unobserved haplotype frequencies

$\pi(h_i)$ and phase resolutions, starting from arbitrary initial values $\pi(h_i)^{(0)}$, for example, where all haplotypes have equal frequency. The initial values are taken as though they are the true values and used to estimate the probability of each possible phase resolution for each individual, and any missing genotype data are jointly estimated. This is called the expectation step.

These estimated phase resolutions are then used to estimate the haplotype frequencies in the next iteration $\pi(h_i)^{(1)}$. By summing the probabilities for each distinct haplotype and scaling the sum to 1, we obtain the updated haplotype frequencies (the maximisation step). This continues until convergence: when the changes in haplotype frequencies in consecutive iterations are small.

Another approach to finding maximum likelihood estimates is to generate possible phase resolutions using a tree model, and then to select the resolution that has the maximum likelihood (Halperin and Eskin, 2004).

The methods of Gusfield (2002) and Halperin and Eskin (2004) utilises the observation that haplotypic variation is organised into blocks of limited diversity (Daly et al., 2001), which result from population substructure, bottlenecks and recombination hotspots. Daly et al. (2001) and Rioux et al. (2001) genotyped 103 common SNPs in an 500 kb region for 129 parents-offspring trios. They found that the region can be divided into blocks (referred as haplotype blocks) spanning from tens up to about one hundred kb.

In these blocks, although there may be m SNPs, there are far fewer than 2^m haplotypes in the population: typically up to 5 different haplotype sequences (Daly et al., 2001; Rioux et al., 2001). Within haplotype blocks, the haplotypic variation may be organised on a tree, explaining how the haplotypes are related to each other by mutation. In Halperin and Eskin (2004), haplotype blocks are defined using an approach similar to the dynamic programming method of Zhang et al. (2002), which jointly maximises the lengths of blocks while minimising the number of SNPs within those blocks required to unambiguously tag the haplotypes. The algorithms of Gusfield (2002) and Halperin and Eskin (2004) differ in how strictly sequences within a block must fit a tree under the infinite sites model. Trees are then constructed for the sequences within a block, and these yield haplotype sequences at the leaves. From amongst

the candidate solutions that fit the block and tree model, the phase resolution is chosen that maximises the likelihood of observing the genotypes given the haplotype frequencies.

There are methods, such as PHASE (Stephens et al., 2001; Stephens and Donnelly, 2003; Stephens and Scheet, 2005) and fastPHASE (Scheet and Stephens, 2006) which do not require haplotypes to be organised into blocks.

PHASE (Stephens et al., 2001; Stephens and Donnelly, 2003; Stephens and Scheet, 2005) is partly based on the Li and Stephens (2003) model, described in Chapter 1. PHASE uses Markov chain-Monte Carlo to approximately sample form the distribution of haplotypes given the genotypes. It starts with an initial guess of the haplotype phases, and then repeatedly selects an individual at random, and assuming that all other individuals are correctly phased, estimates that individual's pair of haplotype sequences. The probability of a particular haplotype pair given the haplotypes of the other individuals is approximated using a model similar to Li and Stephens (2003). A large number of haplotype reconstructions are sampled in this way, discarding the first, say, 100,000 and keeping 200,000, sampled every 100 iterations (Stephens et al., 2001). These can then be used to estimate a single best haplotype phase resolution for the data.

FastPHASE (Scheet and Stephens, 2006) is again based on the observation that over short distances, haplotypes tend to cluster into groups of similar haplotypes, however, it does not force cluster membership to change at haplotype block boundaries. Rather, membership is allowed to change continuously along the chromosome, and this is modelled by a hidden Markov model. This ensures that at adjacent loci, a haplotype is more likely to remain in the same cluster. This results in haplotype sequences which are made up of mosaics of similarity with other haplotypes. This is a more realistic model because while haplotype blocks are believed to arise in part from recombination hotspots, not all recombination occurs in hotspots, hence cluster membership does change continually. The probability that part of a haplotype belongs to a particular cluster is dependent on the relative frequency of that cluster in the population and the allele frequencies in that cluster. The probability that a haplotype sequence changes cluster membership between loci is dependent on the recombination frequency between them and the frequencies of the clusters.

The parameters of the FASTPHASE model are estimated using expectation-maximisation. This typically finds parameter estimates corresponding to local likelihood maxima; however the goal of FASTPHASE is not to explicitly estimate these parameters, rather it is to find phase resolutions. Therefore, the expectation-maximisation algorithm is run multiple times, and results averaged across the different runs.

Phased haplotype sequences can then be sampled from the fitted model. In order to estimate the phase resolution for an individual, pairs of haplotypes which are consistent with the genotype sequence are sampled, and the most frequent pair can be taken as the best estimate. FASTPHASE can also be used for missing genotype imputation, which is described in Chapter 6.

Of the current haplotype phase resolution methods, comparative studies appear to show that PHASE provides the most accurate performance (Marchini et al., 2006). However, computationally, FASTPHASE is significantly more efficient than PHASE.

One solution to handling missing and unphased data is to apply one of the algorithms described above as a preprocessing step, and then construct ARGs for the resolved data. However, doing so with association studies can result in a loss of statistical power, because the uncertainty in the phase resolution is not taken into account (Morris et al., 2004). Hence, I have modified the algorithm to operate directly on unphased and missing data.

2.7 Unphased and Missing Data with Inferred ARGs

Handling missing data is the simpler of the two cases. A missing character is allowed to coalesce with any other character (0,1,**.**, or another missing character), and when it coalesces with a known allele (0 or 1), the missing character becomes fixed to that allele and this assignment is propagated down the ARG to the leaves.

Phasing data is similar, except a record of the diploid pairings of chromosomes is kept. A phase-unknown character may not coalesce with the corresponding phase-unknown character on its sister chromosome (because the individual is heterozygous at that position). When a phase-unknown character coalesces with a known allele, its phase becomes fixed, as does

the character on its sister chromosome, although to the complement allele. When phase-unknown characters from two chromosomes coalesce, these chromosomes and their sisters become dependent on each other. Neither of those chromosomes may coalesce with the sister of the other one. And when one of the chromosomes has a position phase-resolved, that position is also resolved on the other chromosome, and the two sister chromosomes are set to the complement allele. Of course, many more than four chromosomes can become involved in such interdependencies.

Figure 2.5 gives an example of this logic:

- A. Four sequences. The first two are from the same individual and are heterozygous for the third position; the fourth sequence has a missing character in the second position.
- B. An unphased character can match any other character as long as it is not from the sister chromosome (or a dependent), hence the coalescence shown is possible.
- C. The unphased character becomes fixed by inheritance from the coalescence parent.
- D. The dependency that the sister chromosome has the complement allele is propagated through the ARG.
- E. A missing character can coalesce with anything.
- F. The missing character is imputed by inheritance from the coalescence parent.

This approach is similar to Clark (1990), which performs the equivalent of coalescences between phase unknown genotype sequences and haplotype sequences. A key difference being that the method described here also allows mutation events and coalescences over subregions (via recombination). Nor does my method require any of the input sequences to be completely phased to begin with, and hence always finds a phase resolution. Like Gusfield (2002) and Halperin and Eskin (2004), phasing happens on a tree structure, however, my approach does not assume haplotype blocks and instead uses the inferred recombination events to define the span over which haplotypes are compared.

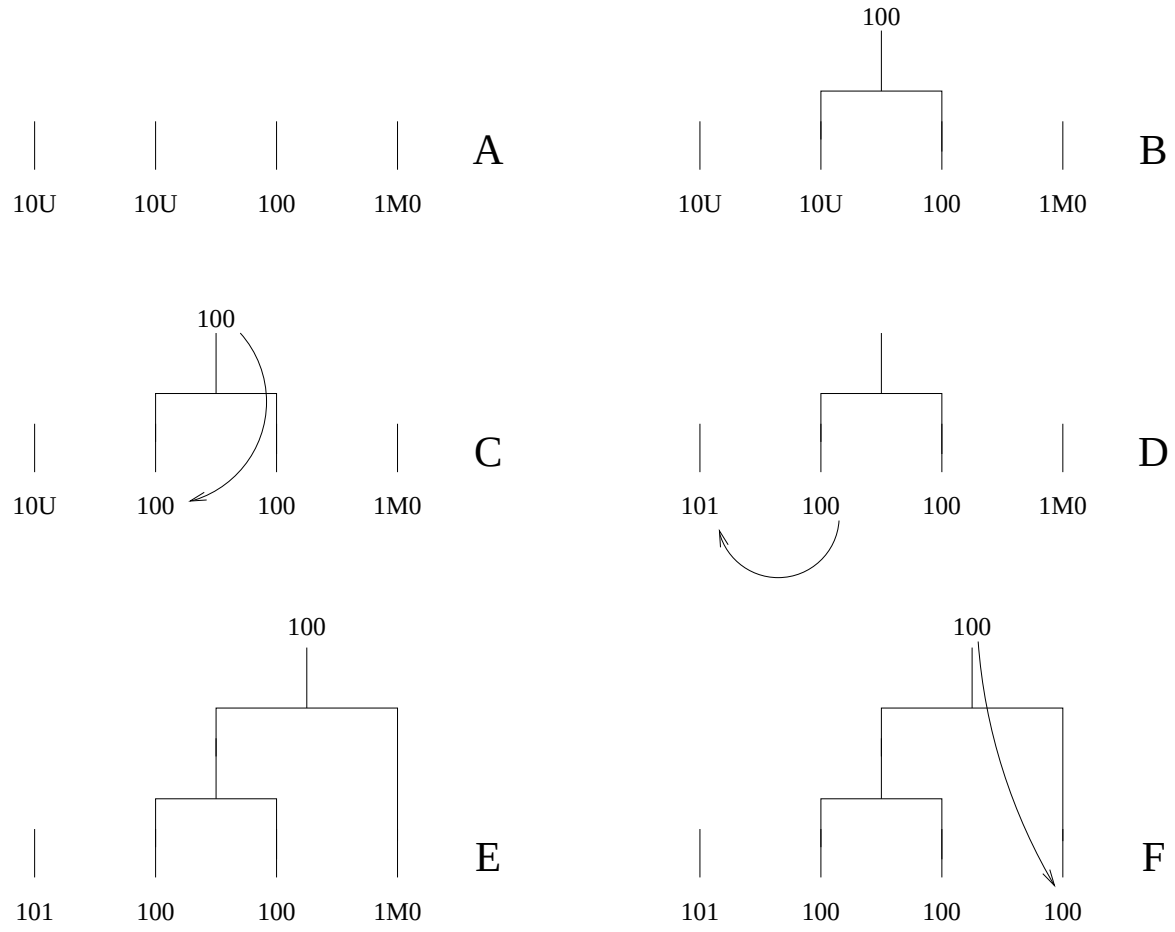


Figure 2.5: Inferring haplotype phase and missing data. See the text for a description.

Chapter 3

Fine Scale Mapping using Ancestral Recombination Graphs

3.1 Approaches to Fine Mapping

As described in the introductory chapter, population-based case-control studies involve individuals genotyped for a panel of SNPs that capture most, but not necessarily all, of the genetic variation in a population (Cordell and Clayton, 2005; Palmer and Cardon, 2005). The individuals are either disease affected cases or unaffected controls, and by analysing the segregation of SNP alleles between these subpopulations it is possible to identify loci with statistical association to disease.

One of the simplest tests for association is Pearson's chi-square test applied to each marker in turn. When causative polymorphisms are typed or are in strong r^2 LD with typed markers, the chi-square test is likely to successfully signal disease association (Devlin and Risch, 1995; Pritchard and Przeworski, 2001). However by testing each marker independently, information about the population history, and in particular, the co-inheritance of alleles, is discarded that can potentially yield a substantial increase in detective and interpretative power. Indeed, it is better in principle to model the genealogical forces that produced the pattern of genetic variation than rely on summary statistics (Nordborg and Tavaré, 2002; McVean, 2002).

There are a number of methods that attempt to model the genealogies of loci in order to

map diseases; the main idea was first described by Templeton et al. (1987):

> "If an undetected mutation causing a phenotypic effect occurred at some point in the evolutionary history of the population, it would be embedded within the same historical structure represented by the cladogram."

That is, if a disease causing allele is harboured at a particular locus, it will be possible to place a mutation on the genealogical tree at that position, giving a clustering of the cases to one side of the tree bipartition which that mutation defines.

The power of this approach was argued by Zollner and Pritchard (2005):

> "Unless we have the actual disease variants in our marker set, the best information that we could possibly get about association is to know the full coalescent genealogy of our sample at that position. If we knew this, the marker genotypes would provide no extra information; all the information about association is contained in the genealogy."

Indeed, if the genealogy were known, not only could disease-associated regions be identified, but the genealogy would give the ages of the causative mutations, would specify the haplotypic background of those mutations, help detect allelic and phenotypic heterogeneity and so on. It would also be possible to optimally impute missing data.

However, the true genealogy is nearly always unknown, and mapping methods must instead use models such as the coalescent-with-recombination. A number of coalescent based methods have been developed (Graham and Thompson, 1998; Rannala and Reeve, 2001; Larribe et al., 2002; Morris et al., 2002; Zollner and Pritchard, 2005), but none are computationally feasible for practical data sets.

In Morris et al. (2002, 2004) a Bayesian, Markov-chain Monte Carlo method is developed for disease mapping. The genealogical histories of disease causing mutations are modelled by trees, with the prior distribution of these being based on the coalescent. These are modelled only for case chromosomes. Another distinctive feature of this method is that it models multiple disease mutations and sporadic cases, representing each as a separate tree—hence this is called the "shattered coalescent". The Markov-chain Monte Carlo algorithm performs

a random walk in the parameter space, which includes the topology of the shattered tree, mutation and recombination rates, and the location of the disease locus.

In the method of Zollner and Pritchard (2005) a Markov-chain Monte Carlo method is used to sample from the distribution of coalescent genealogies for all case and control chromosomes, distinguishing this method from Morris et al. (2002). Furthermore, the genealogies are sampled from a model that approximates the coalescent-with-recombination. A coalescent tree is constructed at each position, called a focal point. On the leaves of the tree, the full extent of the sequences is represented. However, moving up the tree, recombinations occur, and part of a sequence may split off and follow an unmodelled path. Hence the extent of the sequence around the focal point reduces when there is a recombination event, and only the portion of the sequence remaining on that branch is traced further back on the tree. Averaging over the genealogies, the likelihood of the phenotype data under various models of mutation and penetrance is estimated.

The computational limits encountered when applying methods based on the coalescent-with-recombination have partly motivated the development of faster haplotype clustering methods (Templeton et al., 1987; Molitor et al., 2003; Durrant et al., 2004; Templeton et al., 2005; Waldron et al., 2006). These cluster the haplotype sequences (for small non-recombining regions) and perform statistical tests on these clusters. The clustering hierarchy is fast to calculate, and is often organised as a cladogram (Felsenstein, 1985), which is assumed to approximate the marginal tree for that region.

The first such method was that of Templeton et al. (1987) (see also Templeton et al., 1988, 1992; Templeton and Sing, 1993; Templeton, 1995; Templeton et al., 2005). Identical haplotype sequences are grouped together into a clade, and each clade is linked to the clades genetically closest to it by mutation events—those events that are required to convert the haplotypes in one clade into those in another. Here, the region under analysis is assumed to be effectively non-recombining, and hence can have its population history described in such a way, although recombinant haplotypes can potentially be detected by searching for branches of the cladogram with an unexpectedly high number of recurrent mutations.

In Templeton et al. (2005) the clades are grouped together into larger clades, which are

then tested for correlation with the phenotype using a nested ANOVA: the cladogram is systematically split into two or more mutually exclusive and exhaustive clades, and each clade is treated as an allele in an association test.

In Durrant et al. (2004), a hierarchical clustering algorithm is applied to the haplotype sequences, as described in Chapter 1. A cladogram is constructed for a window of SNPs, of user-defined width, and the analysis is windowed across the typed region. Correlations between phenotype and clusters in the cladograms are tested as follows: The cladogram partitions the haplotypes into clusters, with the first partitioning putting each haplotype into its own singleton cluster. The next partitioning, moving up the cladogram, merges the two most similar clusters, and so on. For each partitioning, a logistic regression is performed, where the model is parameterised in terms of the log-odds of disease for each cluster. A likelihood ratio test is performed, where the null model is defined by the null partitioning: the one in which all haplotypes belong to the same cluster, which corresponds to each haplotype having equal odds of being carried by a case or a control. The partitioning which gives the strongest signal of association is found, and this gives the association score for that window of SNPs. P-values can then be calculated by permuting the case/control labels. This method is implemented as a program called CLADH, to which I compare my method.

However, compared to the ARG, cladograms are a coarse approximation of population evolution, and there is often difficulty in modelling the relationships between similar haplotypes and handling rare haplotypes. Additionally, it is often assumed that haplotypes are observed directly and that one can define non-recombining haplotype blocks, which is in general not the case.

I have developed an ARG based mapping method that has computational efficiency nearing that of haplotype clustering methods. I achieve this by using the heuristic approach for ARG inference described in Chapter 2, and can thereby construct ARGs for thousands of individuals typed for hundreds of SNPs. This is sufficiently fast that the analysis can be windowed over the whole genome, fitting the scale of proposed large-scale case-control studies. The simulated experiments described later in this chapter correspond to 800Mb typed at a density of 1 SNP per 3.3kb, for 1000 cases and 1000 controls.

In this way, the proposed method fills the gap between methods that are based on more sophisticated coalescent models but require prohibitive computation, and haplotype based methods that model less precisely the structure and genealogy of a disease locus.

In addition, compared to the methods of Morris et al. (2002) and Zollner and Pritchard (2005), which only work locally on marginal trees, this method constructs full ARGs.

I now describe how inferred ARGs can be used for mapping. This mapping approach is implemented with the ARG inference algorithm in the MARGARITA program (Minichiello and Durbin, 2006). I evaluate the power of MARGARITA on simulated case-control studies and compare the new method to the chi-square test and CLADH. I also show how MARGARITA can be used to infer properties of untyped causative polymorphisms in addition to their genomic positions, which is perhaps the most novel contribution of this chapter.

3.2 Using Inferred ARGs for Mapping

An ARG generated as described in the previous chapter defines a marginal tree for each chromosome position (Figure 1.2). For a given position the marginal tree can be extracted from the ARG by tracing the genealogy of that position back in time from the leaves. When a recombination is encountered, the genealogy follows the path of the left recombination parent if the breakpoint is to the right of the position in question, and otherwise it follows the right parent.

A position can be tested by for association by seeing whether its marginal tree has a branch on which a hypothetical causative mutation can be placed that suitably explains the observed disease states of the genotyped individuals—as illustrated in Figure 1.3. (Note that although such a branch extends over an interval of markers in the ARG, localisation is refined by recombination events lower down the ARG—these change the number of case and control chromosomes under the branch at each position.)

The test is as follows: Since the true ARG is unknown, I infer an ensemble of 100 plausible ARGs. These are generated by running the ARG inference algorithm 100 times, and stochastic choices made during ARG construction mean that in practice these are all different. For each

marker, the corresponding 100 marginal trees are extracted from the ARGs. For each marginal tree, hypothetical disease-predisposing mutations are dropped on each branch in turn. These cause the case-control individuals (the leaves of the tree) to be bipartitioned into those with the mutant allele and those with the ancestral allele. A chi-square test can then be used to detect non-independence between inferred allelic state and disease state. If there are n leaves then there are $n-3$ nonequivalent, non-unary bipartitions of a tree, and hence $n-3$ chi-square test statistics for a tree. Assuming that the region spanned by one tree harbours at most one causative mutation, I take the maximum of these $n-3$ test statistics, calling this the "best cut" score. After finding the best cut score for each of the 100 trees, I take the mean, giving an association score for the marker (this assumes that all the inferred ARGs are equally likely).

Although I test for non-independence between alleles and disease, the test could easily be modified to test for association between genotype and disease (see Chapter 5). Similarly, a regression could be performed, rather than a chi-square test, allowing the method to be applied to quantitative phenotype data. Alternatively, the likelihood of the data given the tree could be calculated, although this would require an explicit disease and mutation model. Also, it is not necessary to assume that there is only one causative mutation on a tree (Zollner and Pritchard, 2005).

In Chapter 1, a number of important limitations of the allelic chi-square test are reviewed (Sasieni, 1997). In particular, under the null hypothesis of no disease association, the allelic chi-square test statistic is asymptotically χ^2 distributed only if the population from which the cases and controls are sampled is in HWE; the test statistic will be inflated if there is an excess of homozygotes relative to HWE. This means that the algorithm may select "best cut" branches that induce the greatest deviation from HWE, without regard to disease association. However, the appeal of the allelic chi-square test is that it can be calculated very fast, while determining the genotypes of individuals which result from bipartitioning the haplotypes requires additional book-keeping.

I calculate the statistical significance of the mapping score at each marker—the marker-wise P-value—by permuting the assignments of case and control labels of the individuals and

repeating the test above. By performing multiple permutations an empirical null distribution is generated from which the P-value can be calculated (Churchill and Doerge, 1994). For P-values exceeding the precision of the permutations, I fit an extreme value distribution to the empirical distribution (Dudbridge and Koeleman, 2004), although I do not rely on this for many of the analyses in this thesis.

Since multiple markers are being tested for association there is a multiple testing issue, which I correct for by calculating for each marker an experiment-wise P-value: the probability that any of the typed markers show such a strong association signal by chance. Again, this is done by permutation: after shuffling the case/control labels, the maximum association score of all the markers is recorded, so defining an empirical experiment-wise null distribution. Once again, an extreme value distribution can be fitted in order to estimate small P-values.

3.3 Simulation of Case-Control Studies

To evaluate the performance of the method under a variety of disease models, I simulated suites of case-control studies. Each suite contained 50 studies simulated under the same model, which was parameterised according to:

- Recombination model of the population from which the cases and controls were sampled.
- TagSNP ascertainment scheme.
- Whether the sequences were phased or unphased, and the amount of missing data.
- Disease model parameters: genotype relative risk, disease allele frequency and also size of study.

The case-control studies were sampled from one of two populations, called "constant" and "hot", depending on the recombination model. Both populations contained 20,000 1Mb chromosome sequences, which were simulated using the FREGENE forward simulator by the authors of that program (Hoggart et al., 2005), and are available for download from the BARGEN website http://www.ebi.ac.uk/projects/BARGEN.

- The **"constant" population** was simulated using the simple (no population expansion or complex demography) Wright-Fisher model with constant recombination rate. The mutation rate was 1.1×10^{-8} per base pair per generation and the recombination crossover rate was 2.2×10^{-8} per base pair per generation.

- The **"hot" population** was simulated with recombination hotspots. These were of length 2kb and accounted for 1% of the length of the region but 60% of all recombinations. The average recombination crossover rate was the same as before, resulting in recombination crossover rates within and between hotspots of 6.56×10^{-7} and 4.44×10^{-9} per base pair per generation respectively. Gene conversions were also included with a constant tract length of 50 base pairs and average rate across the genome of 1.1×10^{-7}. Gene conversions were assigned the same hotspots as crossovers and their rates within and between hotspots were 6.56×10^{-6} and 4.44×10^{-8} per base pair per generation respectively

For both populations, all SNPs with minor allele frequency ≥ 0.005 were recorded (giving 4621 SNPs in the "constant" population and 4825 SNPs in the "hot" population). I then selected tagSNPs using three schemes:

- **"Full" ascertainment.** 120 chromosomes were sampled without replacement from the population and presented to the tagging program TAGGER (de Bakker et al., 2005). (For the "constant" population, 4235 of the 4621 SNPs were polymorphic in this sample and thus considered for tagging, for the "hot" population, 4389 out of 4825 were polymorphic). I set TAGGER to use a maximum tagging distance of 100kb and specified that the tags be optimised for single marker, rather than haplotype-based, tests.

- **5% ascertainment.** As "full", but only SNPs with minor allele frequency $\geq 5\%$ were considered in the tagging process.

- **Random.** Tags were evenly spaced but otherwise selected at random from the SNPs with minor allele frequency $\geq 5\%$ in the population.

In all three cases, 300 tagSNPs were chosen for the 1Mb region. For "full" and 5% ascertainment, these were the best 300 tags as ranked by TAGGER.

The disease model for each suite of 50 case-control studies was specified by parameters q, $GRR(Aa)$, $GRR(AA)$, and n_{cc}.

- q is the frequency of the disease-predisposing allele;
- $GRR(Aa)$ is the genotype relative risk of the heterozygote;
- $GRR(AA)$ is the genotype relative risk of the mutant homozygote; and
- n_{cc} is the number of case chromosome sequences (which in my simulations is the same as the number of control sequences).

$GRR(Aa)$ was varied between 1.4 and 2.4; $GRR(AA)$ was set to $2 * GRR(Aa) - 1$ (an additive effect); q was varied between 0.02 and 0.20; and n_{cc} was varied between 500 and 3000. In order to calculate the penetrances of each genotype at a disease locus, it was also necessary to specify the population prevalence of the disease; this was set to 1% for all simulated studies.

To simulate a case-control study, I used the following process:

1. From one of the FREGENE populations (all SNPs with minor allele frequency ≥ 0.005), a SNP with minor allele frequency between $q-0.005$ and $q+0.005$ was picked at random to be causative.
2. Two sequences (a diploid individual) were picked at random (with replacement) from the population.
3. The individual was assigned to the case set or control set according to the probability of them having the disease given their genotype at the causative SNP.
4. Steps 2 and 3 were repeated until n_{cc} case sequences and n_{cc} control sequences were sampled.
5. Only the 300 tagSNPs were output.

Resampling from the population is not ideal, but I was limited by the size of population which it is computationally feasible to simulate. The resampling may be thought of as performing an additional round of the Wright-Fisher process with a sudden increase in population size, or as there being unidentified consanguinity in the study. This approach has been used elsewhere (de Bakker et al., 2005).

3.4 Evaluating the Performances of Mapping Methods

I implemented the algorithm as a Java program called MARGARITA, and assessed it on both simulated and real data sets involving thousands of individuals typed for hundreds of markers across megabase scale regions.

The performance of a mapping method can be measured according to three criteria:

- Power—the probability of obtaining a significant association signal in a region around a causative polymorphism;
- Localisation—how accurately the methods can estimate the position of a causative polymorphism; and
- Interpretation—the ability to estimate properties of an untyped causative polymorphism (in addition to its position), such as its frequency, which can then guide further investigation.

The power and localisation of MARGARITA was compared across a range of disease models to two other methods: the single marker chi-square test and the CLADH haplotype clustering method (Durrant et al., 2004). Single marker and haplotype-based tests are those most commonly used in practice; coalescent methods such as LATAG (Zollner and Pritchard, 2005) are not computationally feasible for the scale of data I consider here. The single marker chi-square test is often used, and I have selected tagSNPs that capture much of the population variation, meaning that this test is not as "naive" as it may be when markers are chosen at random. From the many available haplotype based methods, I chose to compare the method to CLADH because CLADH is designed to be applied to megabase scale regions, does not

require excessive computation, and has been shown to perform well against similar methods (Bardel et al., 2005).

3.5 Results on a Simulated Suite of Case-Control Studies

As described above, I simulated case-control studies typed for 300 markers across a 1Mb region. These correspond to fine mapping studies, where a causative polymorphism has been detected, or is otherwise suspected to exist in a region, and the next step is to finely localise and interpret that signal.

First, I compare MARGARITA, CLADH and the chi-square test on a suite of 50 case-control studies with parameters $GRR(Aa) = 2$, $GRR(AA) = 3$, $q = 0.04$ and $n_{cc} = 2000$, sampled from the "constant" population with the "full" ascertainment tag set, and using the true phased haplotype sequences with no missing data. The association structure for one of those studies is shown in Figure 3.1.

All MARGARITA P-values for the simulated studies were calculated by performing 10,000 permutations, and P-values < 0.0001 were estimated by fitting extreme value distributions. To analyse one case-control study (4,000 haplotype sequences of 300 SNPs) using MARGARITA on a 2.8 GHz Pentium IV processor required 3-4 minutes to construct 1 ARG, and 6 hours to perform the mapping test with 10,000 permutations on 100 ARGs. The mapping test for MARGARITA is on marginal trees, which potentially change at each marker and therefore I took the location of the typed marker to be the point location of the test. However, the branch that best segregates the cases and controls will be linked to that marker and may not correspond to it.

When using CLADH, the user is required to specify the number of SNPs in each haplotype window. I tried the range of window widths used in the CLADH paper (Durrant et al., 2004), and below I report the best results obtained (using windows of size 5). All CLADH P-values were calculated with 10,000 permutations and I took the location of the typed SNP closest to the centre of the window as the point location of the test.

Below I describe the performances of the methods according to the measures of power,

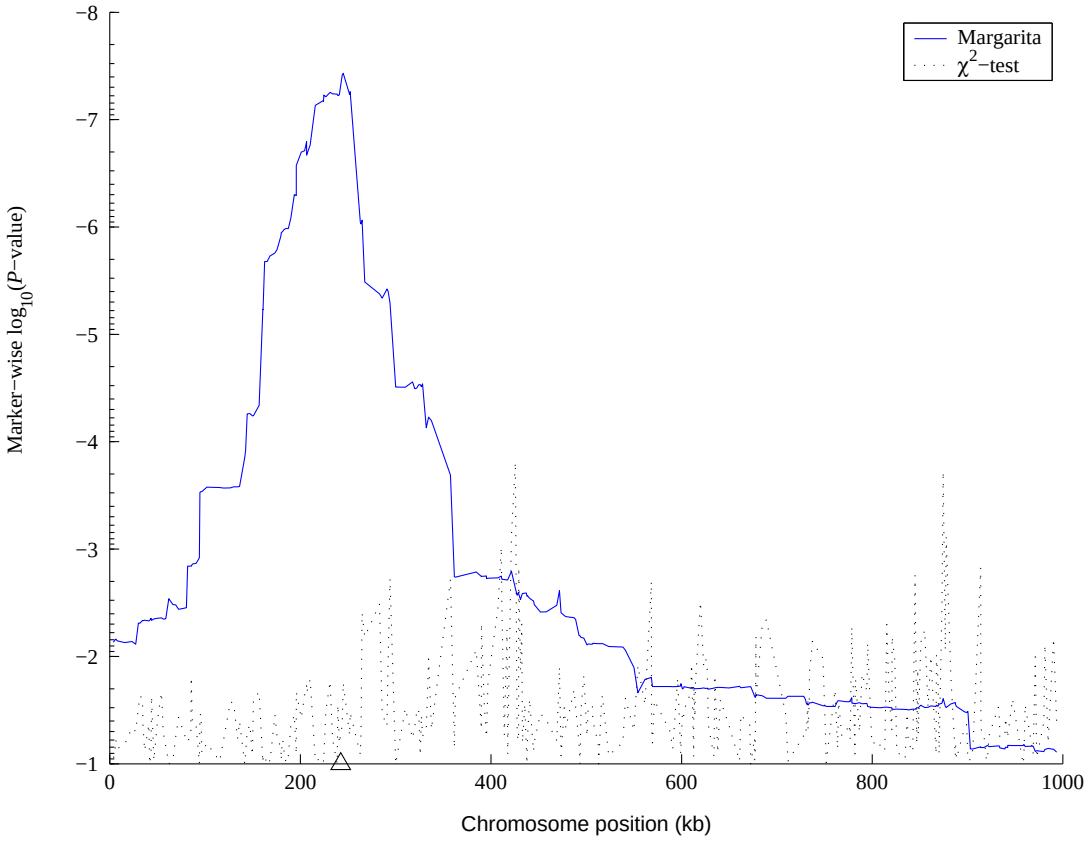


Figure 3.1: The association structure for a simulated case control study with disease parameters $GRR(Aa) = 2$, $GRR(AA) = 3$, $q = 0.04$ and $n_{cc} = 2000$, sampled from the "constant" population with the "full" ascertainment tag set. △ denotes the position of the (untyped) causative SNP.

localisation and interpretation.

Power. To determine power, I defined a window around the causative SNP and calculated the proportion of case-control studies with a significant signal (P$\leq$ 0.05) within that window. Figure 3.2 shows the probability of detecting a marker-wise and experiment-wise significant association within a window around the untyped causative SNP. I am unable to report the experiment-wise significances for CLADH because it does not calculate these. When considering marker-wise significance (top three lines in Figure 3.2), the chi-square test and CLADH have greater power than MARGARITA for windows of $>$ 25kb around the causative SNP. However, when correcting for multiple testing, MARGARITA has greater power than the chi-square test (lower two lines). This difference arises because MARGARITA's tests at adjacent SNPs are more strongly correlated through shared ancestry than the chi-square test's (see Figure

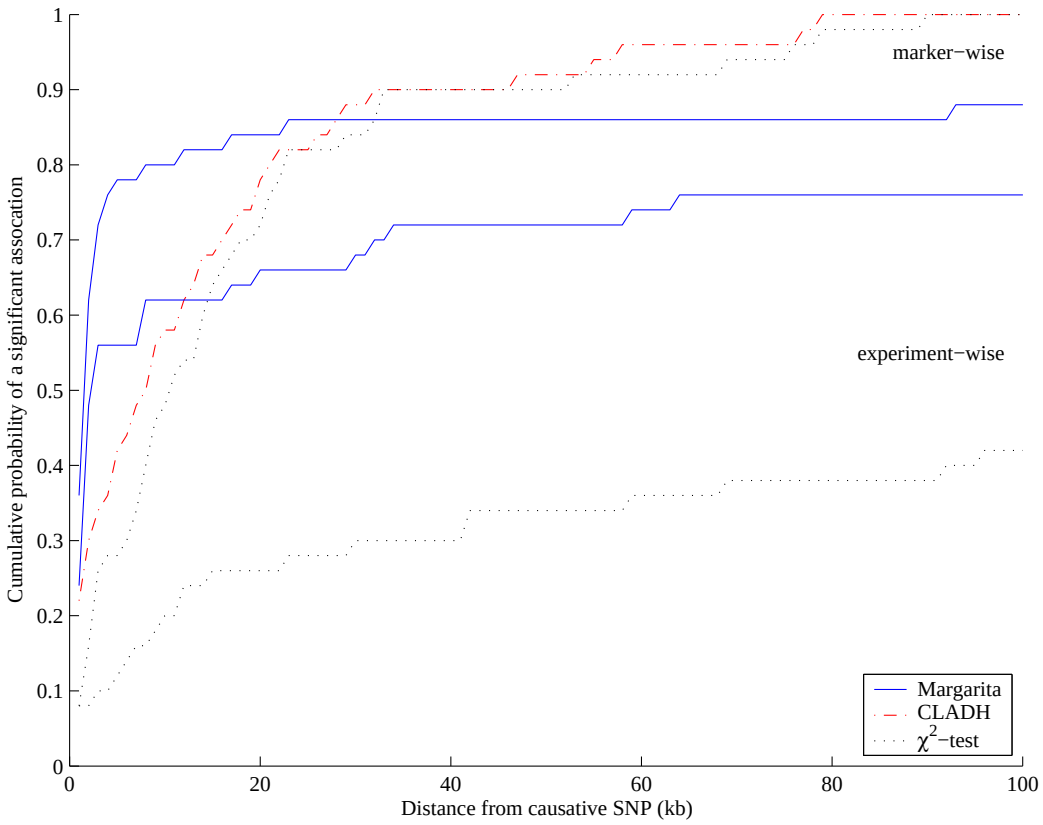


Figure 3.2: The probability of there being a significant association within an interval around the causative SNP. For a suite of case-control studies with disease parameters $GRR(Aa) = 2$, $GRR(AA) = 3$, $q = 0.04$ and $n_{cc} = 2000$, sampled from the "constant" population with the "full" ascertainment tag set.

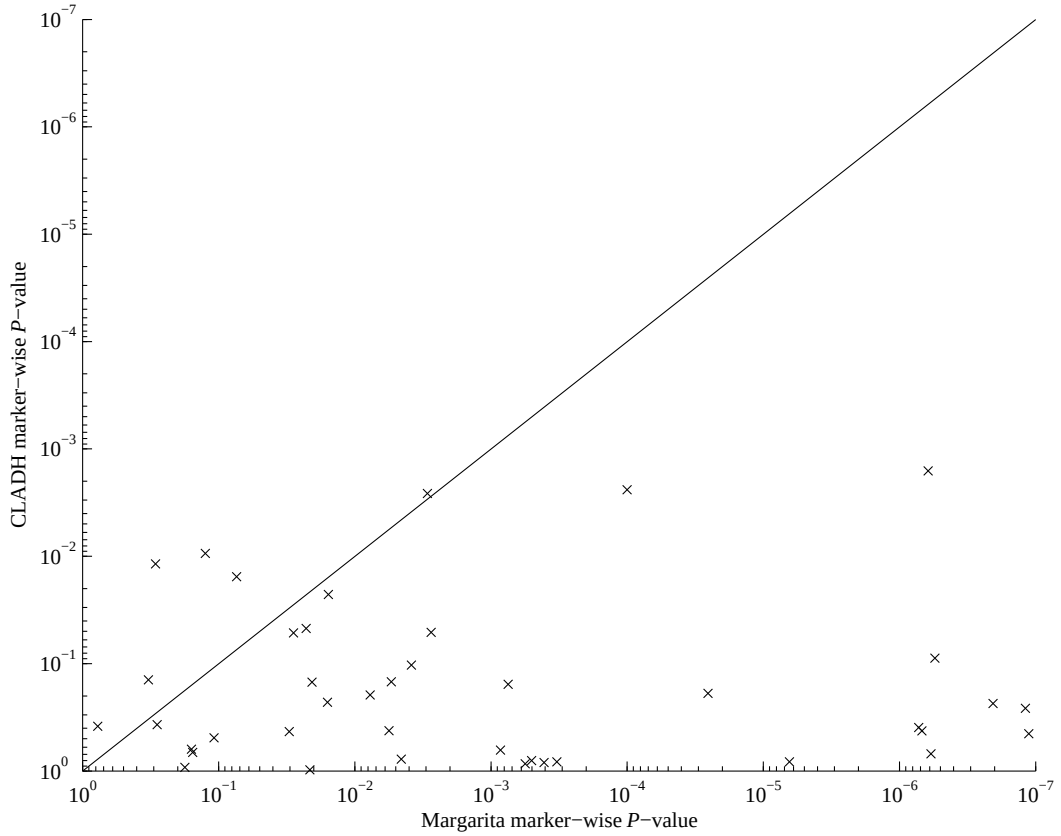


Figure 3.3: Marker-wise P-values at the marker closest to the causative SNP for 50 studies in a suite of case-control studies with disease parameters $GRR(Aa) = 2$, $GRR(AA) = 3$, $q = 0.04$ and $n_{cc} = 2000$, sampled from the "constant" population with the "full" ascertainment tag set.

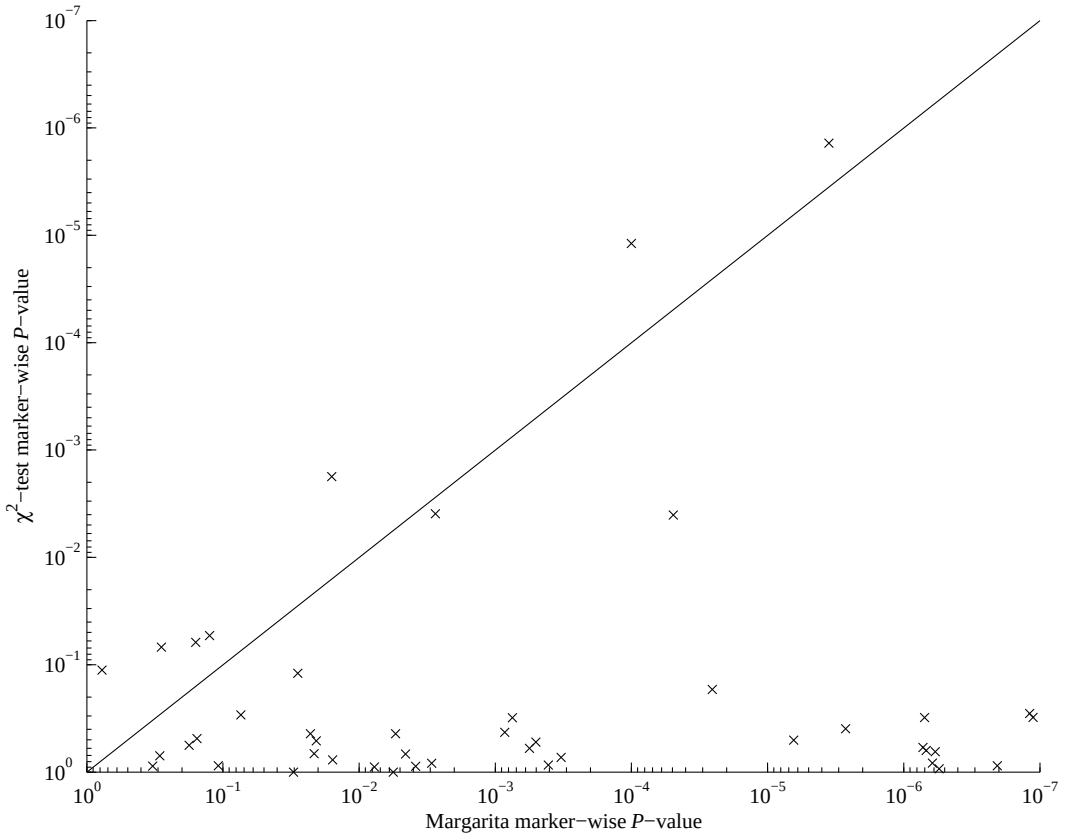


Figure 3.4: Marker-wise P-values at the marker closest to the causative SNP for 50 studies in a suite of case-control studies with disease parameters $GRR(Aa) = 2$, $GRR(AA) = 3$, $q = 0.04$ and $n_{cc} = 2000$, sampled from the "constant" population with the "full" ascertainment tag set.

3.1), reducing the effective number of independent tests across the region.

Figures 3.3 and 3.4 show the marker-wise P-values for the test that is closest (according to its point location) to the untyped causative SNP in each of the 50 case-control studies. The P-values attained by MARGARITA are typically stronger than for the other methods.

I compared the false positive rates of the three methods by counting the number of associations with marker-wise P-value ≤ 0.05 at a distance of greater than 250kb from the untyped causative SNP. An association is counted when the signal breaks below the 0.05 cutoff and then returns above it. The mean number of such false positives for a case-control study from this suite is 0.70 for MARGARITA, 6.16 for CLADH and 10.48 for the chi-square test. This may explain in part the apparent difference in marker-wise power at longer distances (in Figure 3.2).

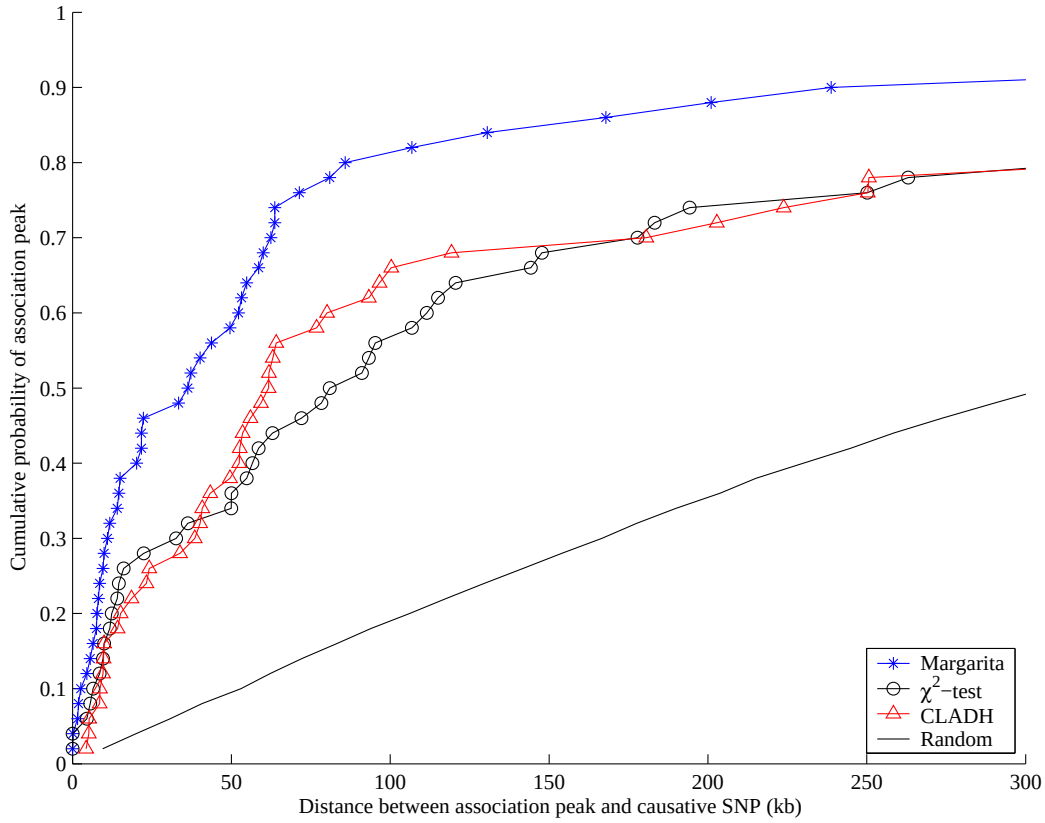


Figure 3.5: Cumulative distribution of distances between the association peak and the causative SNP. For a suite of case-control studies with disease parameters $GRR(Aa) = 2$, $GRR(AA) = 3$, $q = 0.04$ and $n_{cc} = 2000$, sampled from the "constant" population with the "full" ascertainment tag set.

Localisation. This means how accurately a method can estimate the position of the causative SNP. For each of the methods, I took the point location of the test with the strongest marker-wise P-value as the estimate of causative SNP location. Figure 3.5 shows that MARGARITA gives better localisation than CLADH and the chi-square test for this suite of studies.

Interpretation. In studies where the causative SNPs are untyped, it is useful to estimate properties of those SNPs, thus guiding the design of subsequent studies. For example, an estimate of causative allele frequency (which can also be obtained with haplotype clustering methods such as Waldron et al. (2006)) can be used to calculate the sample size required in order to achieve significance. To estimate this, I took the ensemble of marginal trees at the marker closest to the causative SNP, and recorded the branch (bipartition) of each tree that showed the strongest disease association—called the best cut.

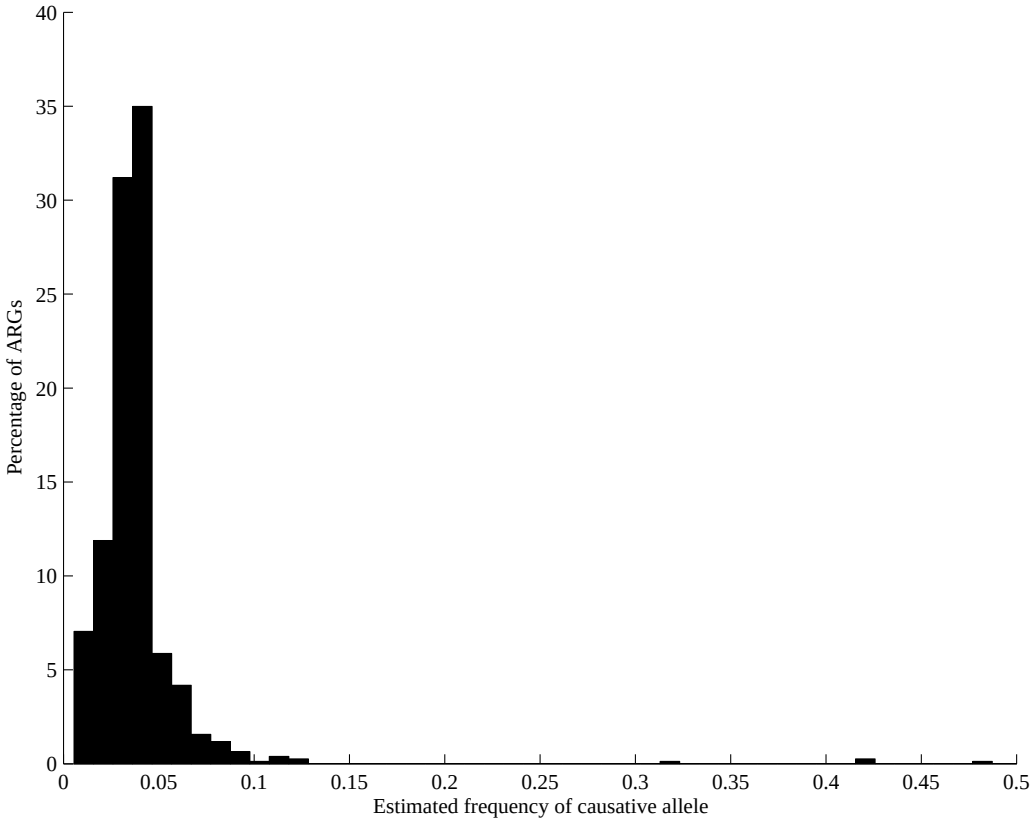


Figure 3.6: The distribution of estimated allele frequency in the general population. For a suite of case-control studies with disease parameters $GRR(Aa) = 2$, $GRR(AA) = 3$, $q = 0.04$ and $n_{cc} = 2000$, sampled from the "constant" population with the "full" ascertainment tag set.

For each tree, an estimate of causative allele frequency was obtained by calculating the fraction of chromosomes that fall under the best cut branch. The frequency of the causative allele in the controls is taken as the fraction of control chromosomes in the sample which fall under the best cut, and the frequency of the risk allele in the cases is taken as the fraction of the case chromosomes in the sample which fall under the best cut. If the population prevalence of the disease is known, then the frequency of the risk allele in the general population can be estimated as follows: Let P be the prevalence of the disease, and f_U be the fraction of the control chromosomes under the best cut, and f_A the fraction of the case chromosomes under the best cut, then the estimated frequency of the causative allele $\hat{q}$ in the general population is:

$$\hat{q} = P f_A + (1 - P) f_U$$

Figure 3.6 shows the distribution of causative allele frequencies as estimated by the ARGs constructed for this suite (causative allele frequency 0.04). The median estimate is 0.036. Note that I only report frequency estimates from studies with a significant association signal. Additionally, a sample of estimated ancestral haplotypes on which the causative allele may have arose can be obtained.

3.6 Results Across a Range of Simulated Disease Models

So far, the performances of the three methods have only been evaluated on one suite of case-control studies, that is, under one disease model. In this section I explore a range of models by varying each parameter (either the genotype relative risk $GRR(Aa)$, the causative allele frequency q, or the study size n_{cc}) in turn while fixing the others at "default" values of $GRR(Aa) = 2$, $GRR(AA) = 2 * GRR(Aa) - 1$, $q = 0.04$ and $n_{cc} = 2000$. In all these simulations I used the "constant" population with the "full" tag ascertainment scheme.

Figure 3.7 compares the power of MARGARITA and the chi-square test to detect an experiment-wise significant ($P \leq 0.05$) association within 100kb of the untyped causative SNP. CLADH is excluded from this comparison because it does not calculate experiment-wise P-values. When comparing experiment-wise P-values, MARGARITA outperforms the chi-square test.

Figure 3.8 shows the localisation performance of the three methods. For the majority of disease models, MARGARITA outperforms both the chi-square test and CLADH.

Finally, Figure 3.9 shows the median estimated causative allele frequency in the general population for a range of suites with varying causative allele frequency (I only report estimates from studies with a significant association signal). I compared the performance of MARGARITA to a simple haplotype approach. For this, I considered all windows of length up to 10 SNPs around the causative polymorphism. I tested each haplotype allele for association with the disease and used the frequency of the most strongly associated haplotype allele to estimate the frequency of the causative polymorphism. MARGARITA has a slight downward bias in its estimate, but it is, nevertheless, reasonable and outperforms the simple haplotype

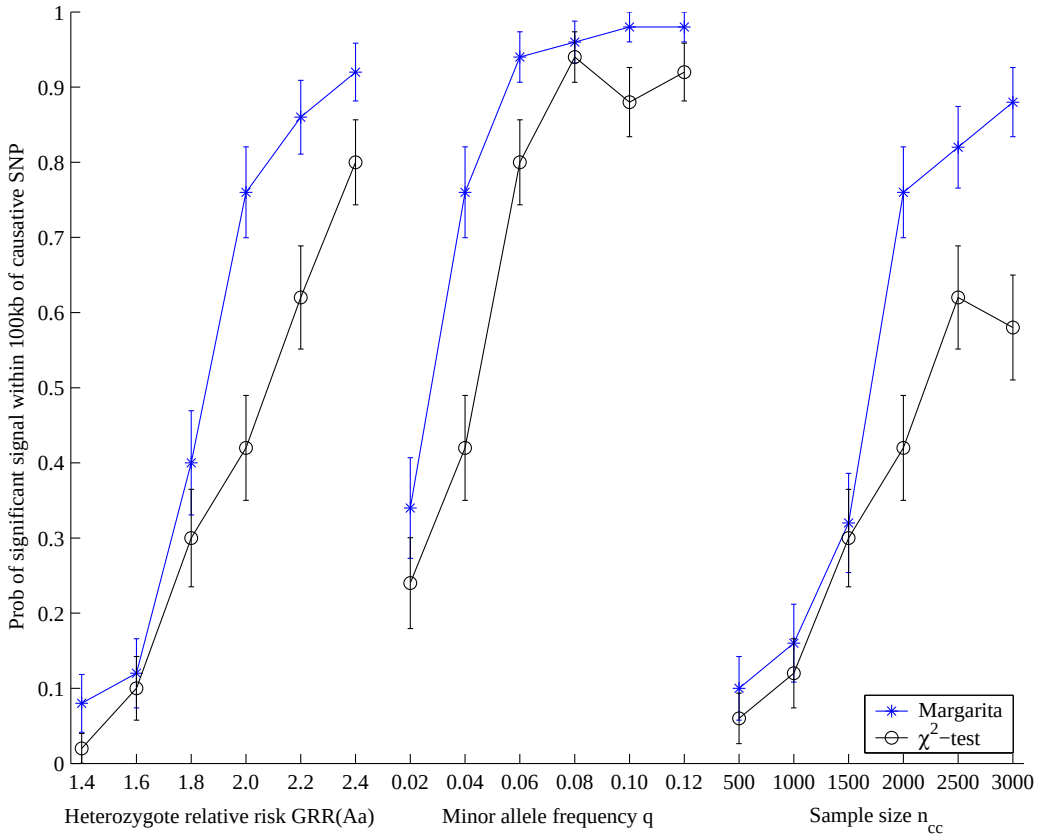


Figure 3.7: Probability of an experiment-wise significant signal within 100kb of the causative SNP (calculated as the proportion of studies in each suite that meet this criterion). Each point on the x-axis corresponds to a suite of 50 studies. Each of the disease parameters is varied between suites, while the other parameters are held at "default" values of $GRR(Aa) = 2$, $GRR(AA) = 2 * GRR(Aa) - 1$, $q = 0.04$ and $n_{cc} = 2000$. All studies are sampled from the "constant" population with the "full" ascertainment tag set.

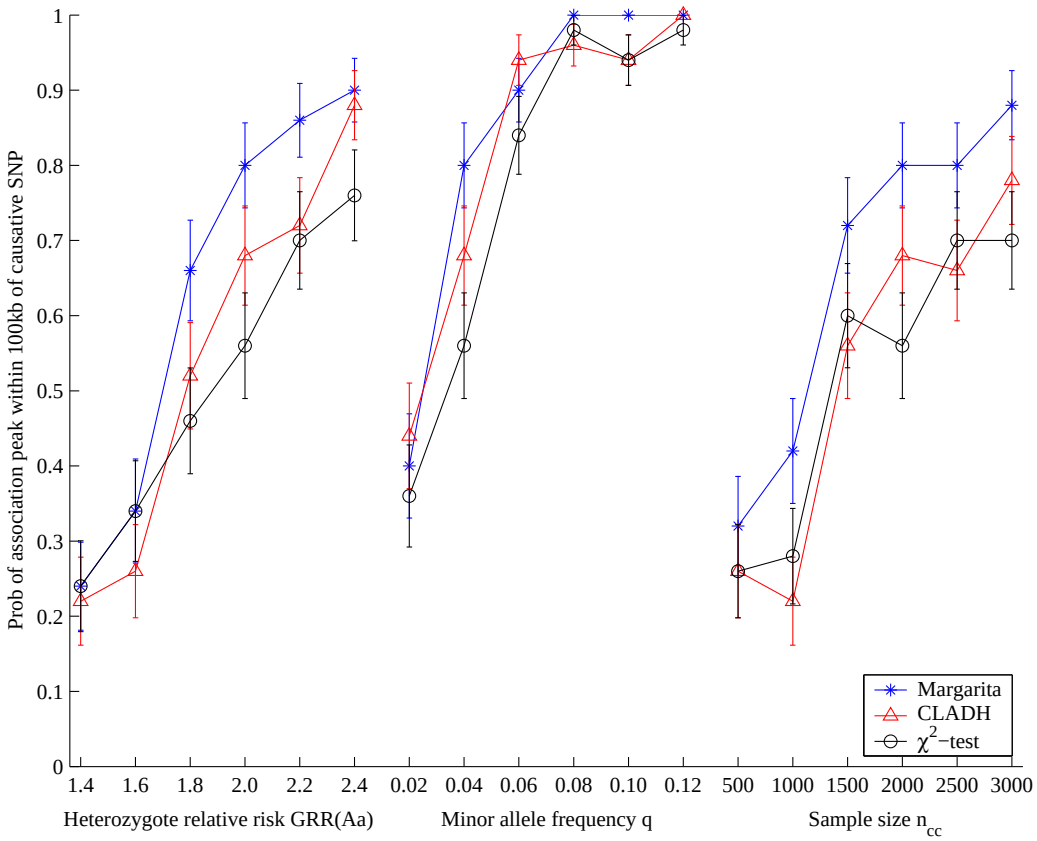


Figure 3.8: Probability that the association peak is within 100kb of the causative SNP. Each point on the x-axis corresponds to a suite of 50 studies. Each of the disease parameters is varied between suites, while the other parameters are held at "default" values of $GRR(Aa) = 2$, $GRR(AA) = 2 * GRR(Aa) - 1$, $q = 0.04$ and $n_{cc} = 2000$. All studies are sampled from the "constant" population with the "full" ascertainment tag set.

approach just described, which has a significant upward bias and a higher variance.

3.7 Results Across a Range of Simulated Population Models and Ascertainment Schemes

For the final set of simulations, the disease model was fixed to $GRR(Aa) = 2$, $GRR(AA) = 3$, $q = 0.04$ and $n_{cc} = 2000$, while the data quality, population model and tagSNP ascertainment scheme were varied.

Figure 3.10 shows the effect of having missing and unphased data on the performance of the method. For this figure, the same suite of case-control studies (sampled from the "constant" population) were used but with the samples output either as phased haplotype sequences, unphased genotype sequences or as phased sequences with 10% missing data. These results show that MARGARITA is robust against both these complications. I do not compare to CLADH because it requires phased haplotypes with no missing data.

Figure 3.11 shows the performance of MARGARITA on case-control studies sampled from a population simulated using a recombination hotspot model (the "hot" population). Under this scenario we see a performance increase for the chi-square test compared to when the "constant" population is used (compare to Figure 3.10). However, it still performs worse than MARGARITA. The chi-square test has increased performance because recombination hotspots give rise to blocks of strong linkage disequilibrium, resulting in tags that capture more of the population variation.

Figure 3.11 also compares the effect of tag ascertainment scheme on mapping performance. The same suite of case-control studies was used, but the samples were "typed" using each of the three tagSNP selection schemes. Tag selection based on less complete data (specifically, when the causative polymorphism is not included in the data used to select tags) results in significantly reduced performance of the chi-square test but has less effect on MARGARITA. Furthermore, the SNP ascertainment scheme which is best for the chi-square test ("full" ascertainment) is not necessarily the best for MARGARITA (which seems to prefer markers with frequency $\geq 5\%$). Consistent with the previous studies, the performance of CLADH

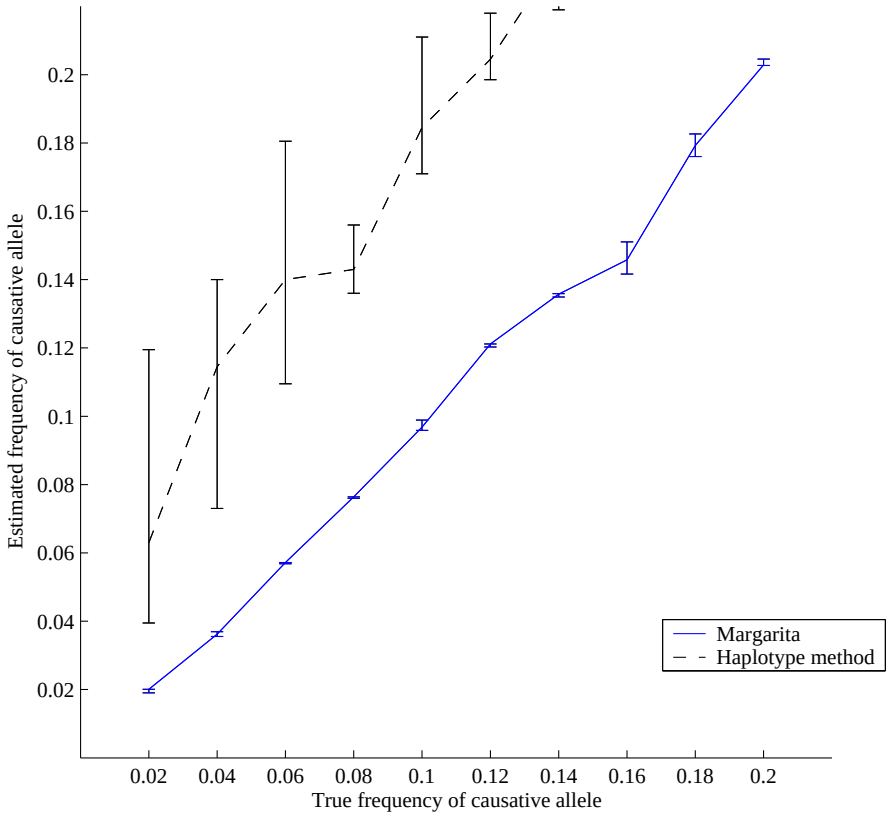


Figure 3.9: Estimated causative allele frequency versus true frequency q. Each point on the x-axis corresponds to a suite of 50 studies. q is varied while the other parameters are held at "default" values of $GRR(Aa) = 2$, $GRR(AA) = 2 * GRR(Aa) - 1$, $q = 0.04$ and $n_{cc} = 2000$. All studies are sampled from the "constant" population with the "full" ascertainment tag set.

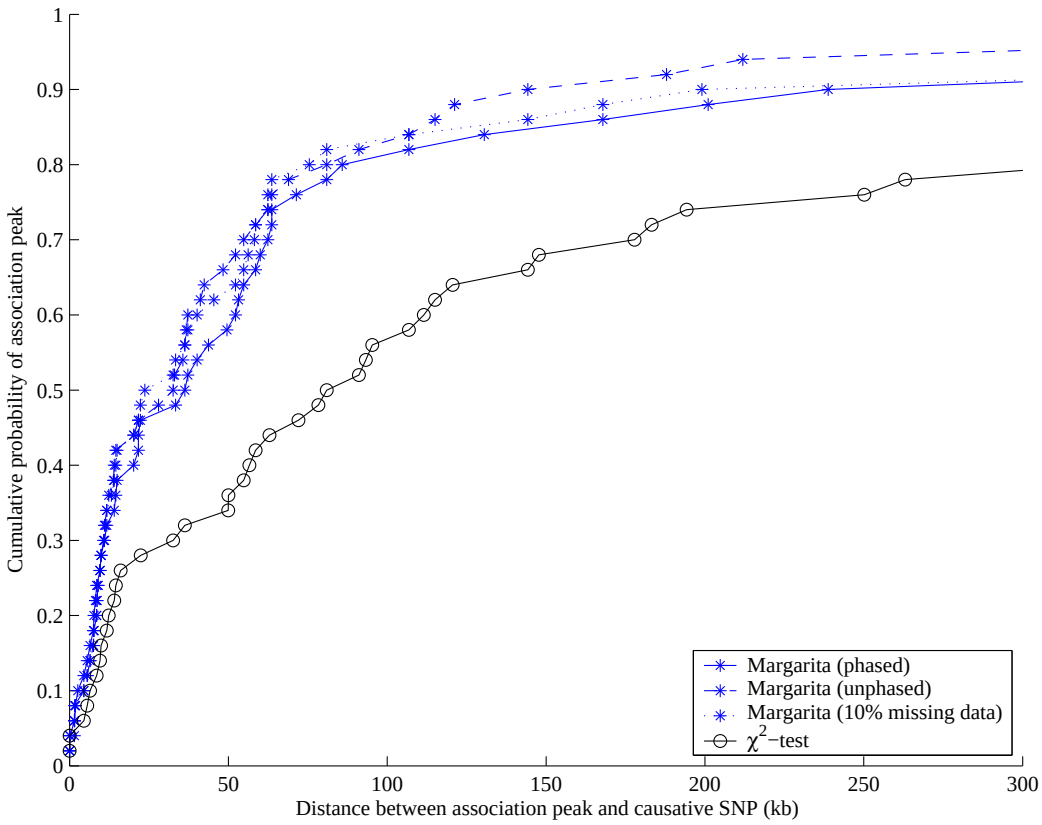


Figure 3.10: Performance on a suite of case-control studies with $GRR(Aa) = 2$, $GRR(AA) = 3$, $q = 0.04$, $n_{cc} = 2000$, sampled from the "constant" population and with the "full" ascertainment tag set. MARGARITA is applied to this suite under three scenarios: when the data is phased, when it is unphased and when it is phased but has 10% missing data.

tends to fall between MARGARITA and the chi-square test.

3.8 Summary

Compared with simpler tests, MARGARITA gives increased accuracy in positioning untyped causative loci and can also be used to estimate the frequencies of untyped causative alleles. MARGARITA also has greater power after correcting for multiple testing, and this is particularly dramatic for low frequency causative alleles.

In the next two chapters, MARGARITA is applied to real case-control association studies, demonstrating how association signals can be dissected using the inferred ARGs.

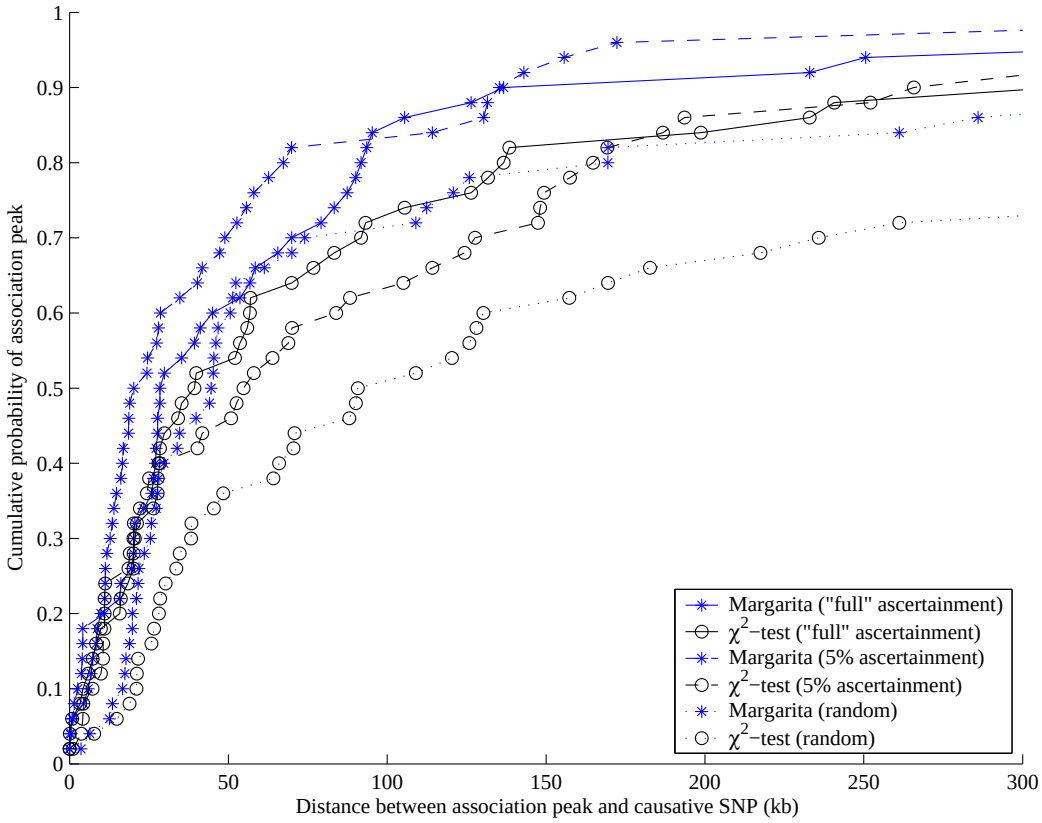


Figure 3.11: Localisation for different data, population and tag models. Performance on a suite of case-control studies sampled from the "hot" population (and with $GRR(Aa) = 2$, $GRR(AA) = 3$, $q = 0.04$, $n_{cc} = 2000$). Performance is compared using three different tagSNP ascertainment schemes.

Chapter 4

Analysis of Graves Disease Association Study Data

4.1 CTLA4 and Autoimmune Disease

Autoimmune disorders, such as Type 1 Diabetes and Graves Disease, together affect up to 3% of the UK population (Vaidya and Pearce, 2004). These diseases result from malfunctions in the immune system, causing intolerance to self. Autoimmune disorders are known to cluster in families, although the specific disease often varies between individuals in the family (Lesage and Goodnow, 2001). This clustering of autoimmune disorders along genetic lines suggests that they share common factors. It is expected that there are multiple genetic factors interacting with enviromental factors that affect susceptibility.

Along with the MHC, the CTLA4 gene has emerged as a convincing susceptibility locus for these diseases (Vaidya and Pearce, 2004). Initially, disease association was mapped to chromosome 2q33, which contains a number of T-lymphocyte regulatory genes, which are logical candidates for disease risk. A subsequent fine-mapping study of association between polymorphisms in 2q33 and Graves Disease (Ueda et al., 2003) showed a strong association between SNPs in the CTLA4 gene and Graves Disease.

In the Ueda et al. (2003) study, a 300kb region (CD28-CTLA4-ICOS) was genotyped for 108 SNPs in 652 control individuals and 384 Graves disease cases. In order to identify novel

SNPs, 32 individuals were resequenced, hence it is reasonable to assume that all common polymorphisms in the region, in the UK population, have been identified. In their analysis, three association peaks were identified; moving from left to right in Figure 4.1, these peaks are at SNPs MH30, CT60 and CTBC217_1. By performing a regression analysis, they concluded that the causative variant is more likely around the CT60 peak than the others. Around the CT60 peak there are three other SNPs, JO31, JO30 and JO27_1, which are also strongly associated, but their analysis was unable to further dissect the signal.

They performed functional studies to follow up the finding, and showed that the associated haplotype at CTLA4 is correlated with lower mRNA levels. In the non-obese diabetic mouse model, it was also shown that there was reduced production of CTLA4.

I applied MARGARITA to the data in order to see whether the CTLA4 signal could be further dissected.

4.2 ARG Analysis of the CTLA4 data

Since the data is unphased and has missing genotypes, I used MARGARITA to infer these (in inferring 100 ARGs, 100 different phase resolutions are obtained, thus marginalising over phase uncertainty when performing the mapping test). However, unphased data will present a hurdle for mapping methods that require phased haplotype sequences. One way to overcome this is to run a phasing algorithm (Marchini et al., 2006) on the data and then pass the result to the mapping method as though it is the true phase resolution (Morris et al., 2004). To examine the effect of this, I also ran MARGARITA on the "best" phase resolution of the data after applying one run of the program PHASE (Stephens et al., 2001), where "best" is defined as the most likely phase resolution found.

Figure 4.1 shows that CT60 has the strongest disease association in my analysis (both when using the PHASEd and unphased data), agreeing with Ueda et al. (2003)'s analysis. All MARGARITA P-values in this chapter were calculated by performing up to 1 million permutations.

MARGARITA on the unphased data gives a stronger association signal at CT60 ($P \approx$

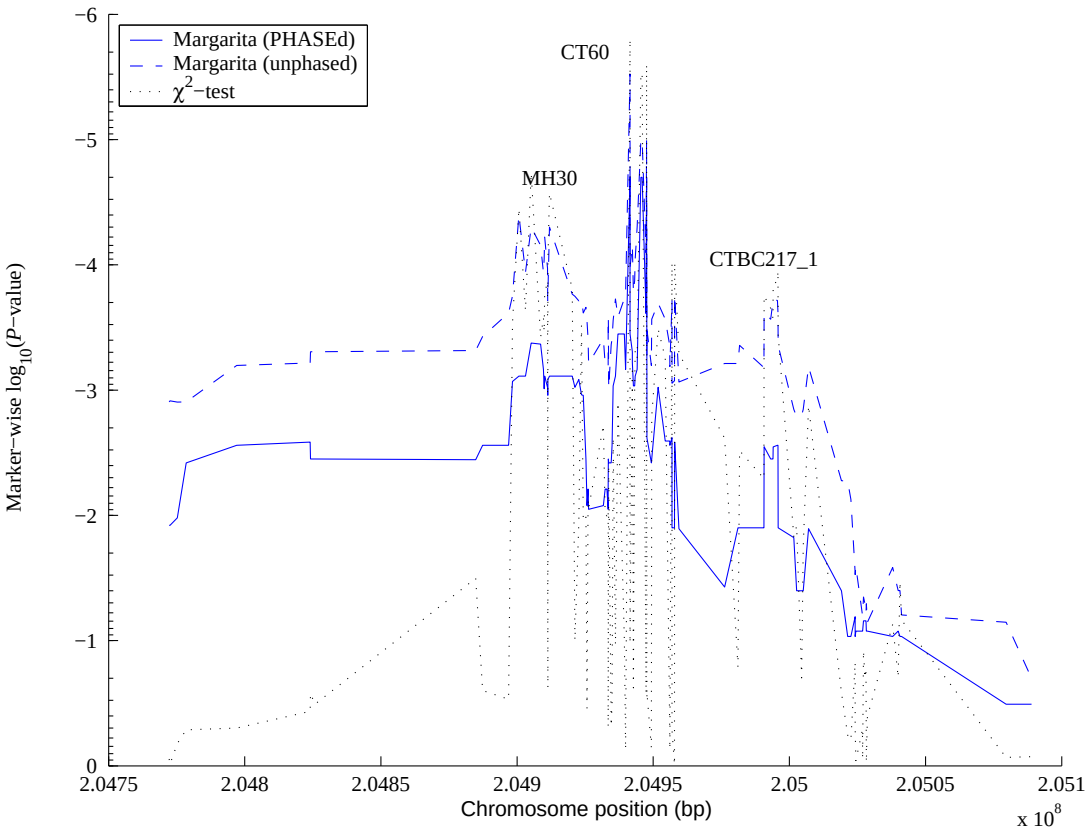


Figure 4.1: Analysis of the CTLA4 data. Association structure of the region.

2×10^{-6}) than does MARGARITA on the PHASEd sequences ($P \approx 2\times10^{-5}$). This result agrees with that of Morris et al. (2004), who similarly show that a two stage approach results in a loss of power compared to handling genotypes directly and marginalising over unknown phase. Both MARGARITA analyses have weaker significance than the chi-square test ($P \approx 1.6\times10^{-6}$) at CT60, which would be expected if CT60 is indeed the causative polymorphism, a hypothesis that can be explored by using the ARGs to further analyse the association signal.

Figure 4.2 gives the distribution of the estimated susceptibility allele frequency in the general population (calculated using the observation that Graves disease has population prevalence 0.5%). The mean estimate for the causative allele in the cases and controls is 65% and 54% respectively, corresponding to the G allele of CT60 (%63 and %52 in cases and controls respectively). This suggests that the bulk of the association signal at CT60 is due to susceptibility caused by CT60, or something extremely tightly linked to it.

However, in 43% of the inferred ARGs for the unphased data, MARGARITA is able to find

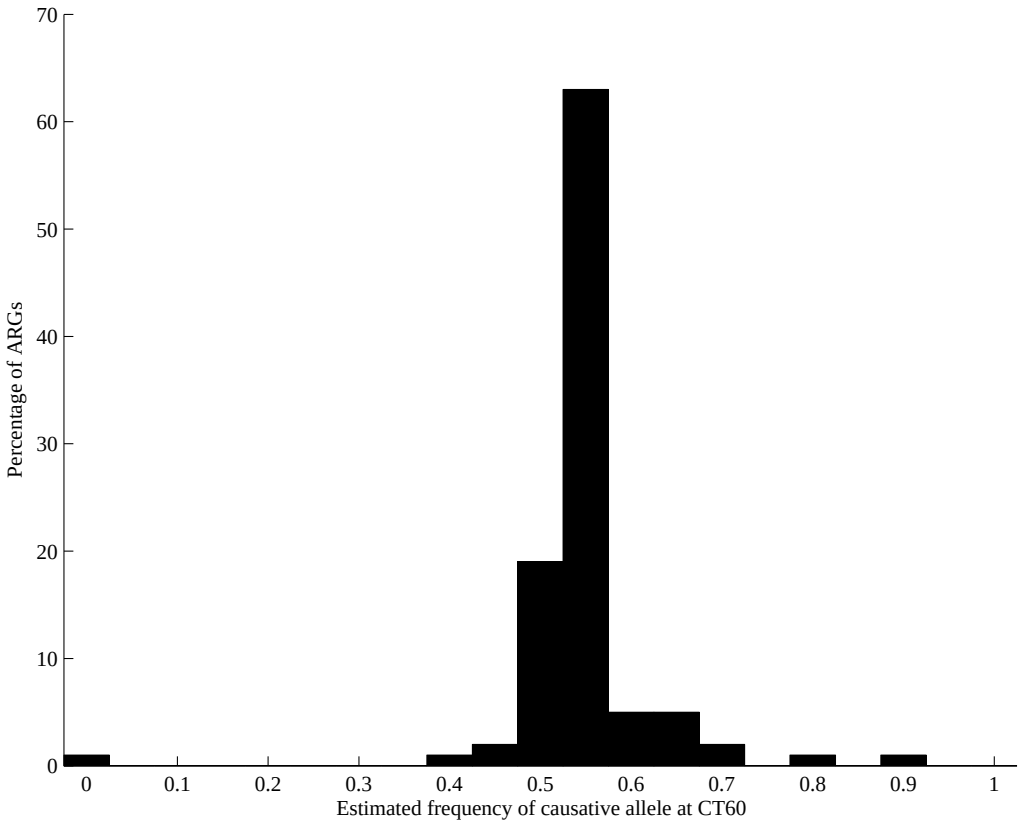


Figure 4.2: Distribution of estimated causative allele frequency in the general population, using marginal trees at CT60.

internal branches of the marginal tree at CT60 that segregate the cases and controls with chi-square test P-values of the order of 10^{-7} or less, with the strongest being of the order of 10^{-9}; this may suggest a second causative polymorphism. I therefore used the inferred ARGs to test explicitly for allelic heterogeneity. I took the 100 marginal trees inferred for each marker and counted the number of times each chromosome appeared under the branch corresponding to the best partitioning of cases and controls—the "best cut" branch. When a chromosome is under the best cut branch it means that if there is a disease causing allele at that position, then it is likely that the chromosome possesses it. Figure 4.3 shows this analysis for an illustrative sample of 167 case chromosomes (with phase inferred on the ARGs). For each marker and chromosome, the intensity of the plot represents the proportion of trees for which the chromosome is under the best cut. Case chromosomes 131-167 show a different pattern to the others. They occur less frequently under the best cut at CT60, and more frequently under

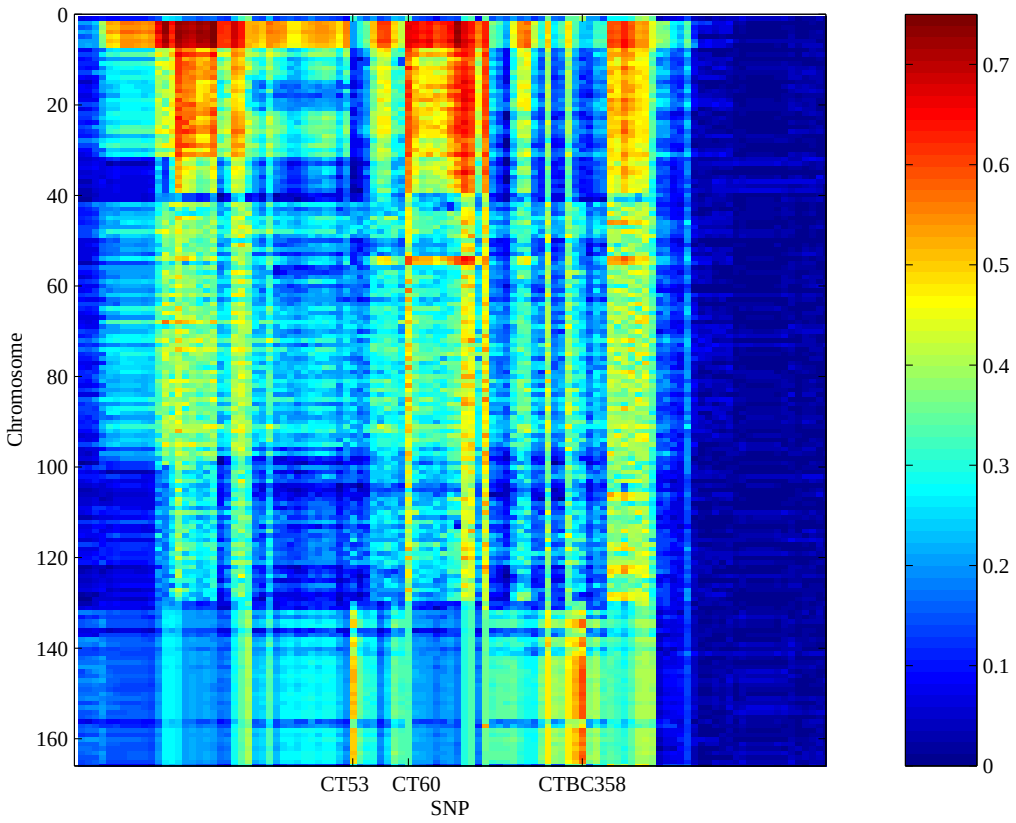


Figure 4.3: Test for allelic heterogeneity. The proportion of inferred marginal trees at each position for which a chromosome appears under the branch that best segregates the cases and controls.

the best cut at CT53 and CTBC358, whereas case chromosomes 1-130 appear frequently under the best cut at CT60, but infrequently under the best cut at CT53 and CTBC358. Although not shown in the figure, there are other case chromosomes not associated with any of these loci.

To test whether CT53 or CTBC358 are also susceptibility loci (or linked to susceptibility loci), I stratified the case-control population in three ways:

Only those chromosomes with the protective allele at CT60.

I took the PHASEd chromosomes and removed all those with the CT60 susceptibility allele, running the analysis on the remaining 282 case chromosomes and 620 controls with the protective allele (Figure 4.4). When the population is stratified in this way, the association signals at MH30 and CTBC217_1 collapse into the background, suggesting that the association signals at those locations are due to LD with CT60. Furthermore, there is an association

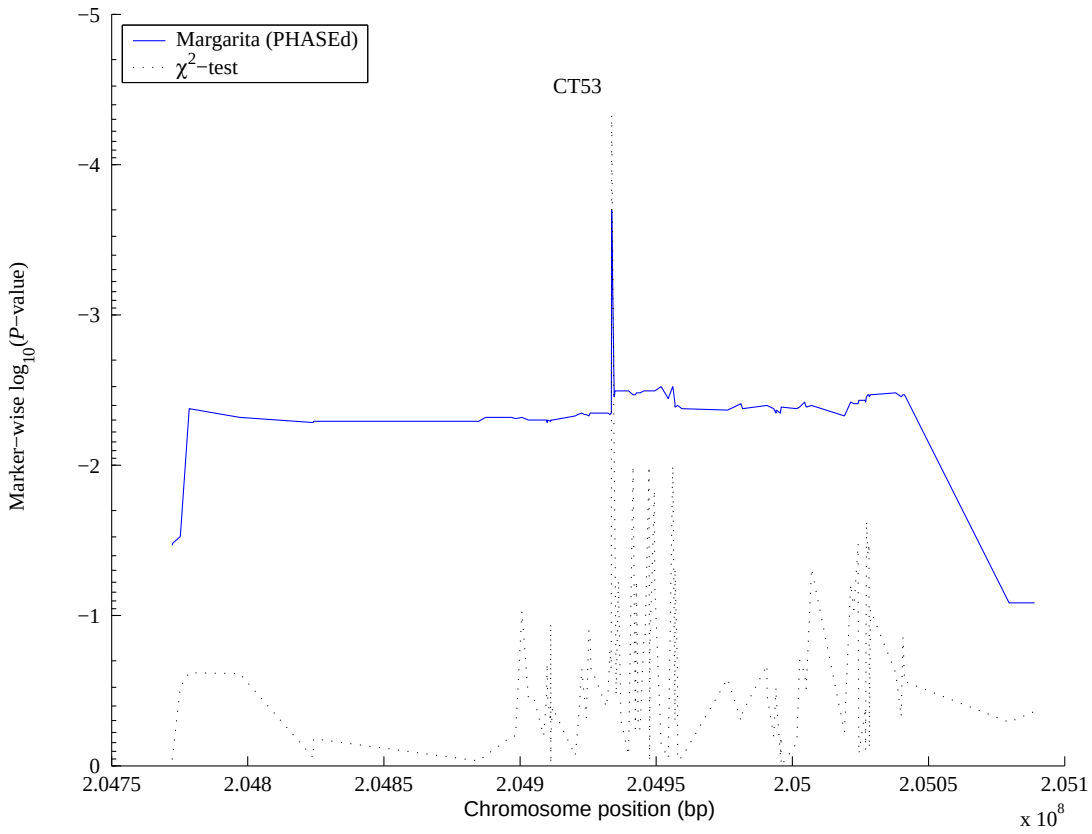


Figure 4.4: Association structure for a subset of the CTLA4 data—only those chromosomes with the protective CT60 allele.

peak at CT53 (marker-wise chi-square test $P \approx 5 \times 10^{-5}$; MARGARITA $P \approx 2 \times 10^{-4}$). Using MARGARITA, the estimated frequency of the causative allele (92% in the cases and 82% in the controls) matches that of the A allele at CT53 in this subpopulation (93% in the cases and 83% in the controls), suggesting that the A allele confers susceptibility on this CT60 background.

Only those chromosomes with the susceptibility allele at CT60.

After conditioning on the CT60 susceptibility allele there are 486 case chromosomes and 684 control chromosomes. In this subpopulation, CT53 has a weak signal of association with the disease (marker-wise chi-square test $P \approx 0.023$; MARGARITA $P \approx 0.016$). In contrast to the previous stratification, the A allele at CT53 is less frequent in the cases (2%) than in the controls (5%), suggesting that A may be protective on this haplotypic background. This reversal of the effect of CT53 dependent on CT60 status may explain why CT53 is not

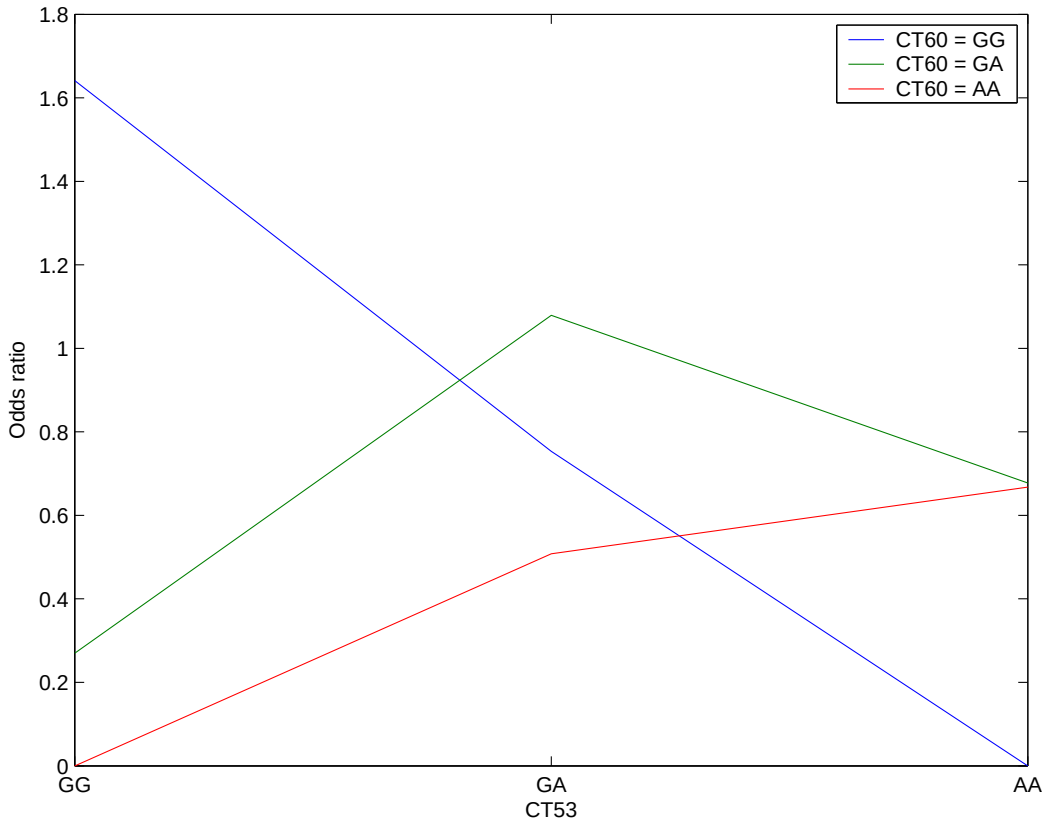


Figure 4.5: Epistasis between CT60 and CT53. The chart gives the odds ratios for genotypes at CT60-CT53.

detected in analyses using the full data.

Only those individuals that are homozygous for the CT60 protective allele.

To check that the CT53 association is not due to some spurious signal resulting from selecting chromosomes on the basis of their inferred haplotype phase, I took the genotype sequences homozygous for the CT60 protective allele and ran MARGARITA on these unphased sequences. There are 102 case chromosomes and 300 controls, giving a weaker but still significant signal of association (marker-wise chi-square test $P \approx 0.012$; MARGARITA $P \approx 0.013$). As expected, on this background the A allele of CT53 is the susceptibility allele.

These results suggest epistasis between CT60 and CT53, with the A allele at CT53 conferring susceptibility on a CT60 protective background, but being protective on a CT60 susceptibility background. Figure 4.5 shows how the disease effect of one locus is modified by the genotype at the other.

To test explicitly for epistasis between CT60 and CT53 I performed a logistic regression test for interaction (Cordell, 2002; Macgregor and Khan, 2006) and obtained $P \approx 0.004$ for interaction effects over and above single marker effects.

The reversal of the effect of alleles at CT53, conditional on the allele at CT60, would tend to reduce any significance in a logistic regression test for additional effects, as used in Ueda et al. (2003).

4.3 Replication of Result

Given the small samples sizes of the data after stratification into CT60 allele subpopulations, further genotyping in more samples is required in order to determine whether the observed signal at CT53 is a true positive or an artefact of the data.

In response to this ARG analysis, further genotyping was undertaken by my collaborators in the Diabetes and Inflammation Laboratory at the Cambridge Institute for Medical Research. They typed an additional 1,593 Graves cases and 4,055 controls at CT53 and CT60. In this larger sample, the previously reported effect of CT53 did not achieve significance when performing the same subgroup analysis.

However, an independently developed method (Dawy et al., 2006) applied to the original data (Ueda et al., 2003) did show the same signal as the ARG analysis. In Dawy et al. (2006) an information theoretic method was developed for disease mapping. Their method has the feature that it is entirely general regarding epistatic and risk model (for example, it does not assume a multiplicative model of genotype risk). It is designed to identify disease associated markers and to cluster them according to their pattern of variability, with the motivation being that markers with similar variability (i.e. in strong LD) are likely to have the same genealogical history and should be interpreted together. With this method, the degree of association between a marker and disease is measured as the quantity of information contained in the marker about the disease. They then search groups of jointly associated markers by using a "relevance chain" technique, where the reduction in uncertainty on the disease state given a genotype observation is measured conditional on observing the genotypes at other

positions.

Their approach identified the same peak at CT60 as in Ueda et al. (2003) and in the analysis above, and they also found an additional experiment-wise significant signal at CT53, in agreement with my analysis. However, it should be noted that since this is on the same data set, it is not a proper replication of the association signal; it merely shows that two independently developed methods identify the same signal (albeit not previously identified) in the same data set.

Dawy et al. (2006) give the following reason for why the CT53 signal was not detected in the original Ueda et al. (2003) analysis:

> "In the original article, the effect of secondary loci in addition to the main associated loci was tested, assuming a multiplicative model for the allele effects. Such trend regression approaches, however, imply a continually increasing or decreasing causality scheme across genotypes, which is possibly not always an accurate assumption. The slight difference between the original and our results might be attributed to the fact that the use of mutual information does not assume any particular mode of allelic risk."

In conclusion, since the signal has not been replicated in an independent population, the results suggesting epistatic interaction should be treated with caution; nevertheless, the analyses in this chapter show how the ARG approach can be used to dissect disease association signals, arriving at potentially interesting additional conclusions.

Chapter 5

Analysis of Prostate Cancer Association Study Data

5.1 Risk factors for Prostate Cancer

Globally, about 9.7% of cancers in men are prostate cancers, and the risk of developing the disease has been correlated with age, family history and ethnicity. The highest rates are reported in the Caribbean and Scandinavia and the lowest rates in China and Japan (Crawford, 2003). Within North America, the incidence of prostate cancer is 1.6 times greater among African American men than European American men (Freedman et al., 2006), suggesting a genetic component to the disease. However, it has also been shown that Japanese men who migrate to North America have an increased incidence of prostate cancer, suggesting that enviromental factors also have a role (Crawford, 2003).

Five recent association studies have sought to investigate the genetic basis of the disease (Amundadottir et al., 2006; Freedman et al., 2006; Gudmundsson et al., 2007; Haiman et al., 2007; Yeager et al., 2007), and all of them show strong association between variants in the 8q24 region and disease risk.

The first study, Amundadottir et al. (2006), identified a SNP and a microsatellite in 8q24 associated with risk in an Icelandic population, a signal which they found to be replicated in European American and Swedish populations. In the second study, Freedman et al. (2006),

admixture mapping (Patterson et al., 2004; Smith et al., 2004) was applied to a data set of 1,597 African Americans, identifying a 3.8Mb region of 8q24. The two polymorphisms identified in Amundadottir et al. (2006) were found to explain very little of the admixture signal in the African American population, suggesting additional causative variants in the region. This suggestion partly motivated further studies (Gudmundsson et al., 2007; Haiman et al., 2007; Yeager et al., 2007), although two of these (Gudmundsson et al., 2007; Yeager et al., 2007) were whole genome scans, looking beyond 8q24. In these subsequent studies independent risk alleles were identified to those in Amundadottir et al. (2006), however, these are all in 8q24. Of particular interest to us is Yeager et al. (2007), in which MARGARITA was applied to dissect the 8q24 association signal, aiding the identification of two independent association peaks. I now discuss results from these five studies in more detail, including a description of the MARGARITA analysis.

5.2 First Identification of the 8q24 Association Signal

In the first study (Amundadottir et al., 2006), a genome-wide linkage scan was performed, using 1,068 microsatellite markers typed in 871 Icelandic men with prostate cancer from 323 extended families. A second stage case-control association study in 8q24 identified the strongest associations to be a microsatellite DG8S737 ($P = 2.3 \times 10^{-8}$) and a SNP rs1447295 ($P = 1.7 \times 10^{-9}$). They replicated these signals in two additional case-control studies: one Swedish (1,435 cases and 779 controls) and one European American (458 cases and 247 controls).

They then investigated the independence of the two association signals, because in the Icelandic sample, allele -8 of DG8S737 and the A allele of rs1447295 have LD of $r^2 \approx 0.5$, and could therefore correspond to the same causative variant. They found that individuals with both risk alleles have greater risk than those with only one; and those with one have greater risk than those with neither. This suggests that neither of the alleles by themselves explains the risk, so either there are multiple causative variants in the region, or the two risk alleles are in strong, but imperfect, LD with an unknown risk variant.

They then undertook a study in African Americans in order to map more finely the risk variants; the basis for this being that African populations tend to have greater genetic diversity and weaker LD. Specifically, consider the 92kb LD block around DG8S737. In the CEU HapMap sample, there are 19 SNPs, including rs1447295, that have $r^2 = 1$ with each other; whereas in the YRI HapMap sample, only 2 SNPs have $r^2 = 1$ with rs1447295. In the African American sample, DG8S737 showed association of $P = 0.0022$ and rs1447295 was nonsignificant.

In the second study (Freedman et al., 2006), a technique known as admixture mapping was applied genome-wide to a data set of 1,597 African Americans. Admixture mapping (Patterson et al., 2004; Smith et al., 2004) is based on the observations that:

1. Some disease causing variants have significantly different frequencies in different populations; and

2. There are some diseases where the incidence of disease is also significantly different between populations. For example, in Americans, autoimmune diseases are more common in those of European descent (Patterson et al., 2004); whereas prostate cancer is more common in those of African decent (Amundadottir et al., 2006).

The technique involves scanning case individuals from populations of mixed ancestry. When a chromosomal region contains causative variants, it may show an over-representation of ancestry from the population with more risk alleles at that locus.

The advantage of admixture mapping is that it requires around 1% of the markers required in a LD based scan (Patterson et al., 2004); for example, in African Americans, admixture has occurred within the past 15 generations (Smith et al., 2004), hence there has been little time for recombination to break up the tracts of ancestral material, which means that the regions of excess ancestry around causative variants are likely to extend for tens of Mb, requiring far fewer markers to tag. In Freedman et al. (2006) only 1,365 SNPs were used for a genome-wide scan.

From their study of 1,597 cases, they located the admixture peak to a 3.8Mb region of 8q24. By also genotyping 873 African-American controls (controls are not required for

admixture mapping, but are useful for subsequent analyses) they estimated that the fraction of all prostate cancer incidence for African Americans below 72 years of age that could be explained by ancestry at this locus is 49%. This suggests that if the region of 8q24 were replaced with that from European ancestors, the rate of prostate cancer in African Americans would decrease by approximately 49%. However, it should be noted that such population attributable risk estimates should be treated with caution; they often sum to greater than 100%.

Freedman et al. (2006) also compared their results to Amundadottir et al. (2006). Because of the systematic differences in ancestry between cases and controls across 8q24, Freedman et al. (2006) tested whether the association at the DG8S737 microsatellite, detected in Amundadottir et al. (2006), corresponds to a fine mapping signal or the admixture signal of the larger region. (Amundadottir et al. (2006), tested for mismatching of cases and controls in overall ancestry, but not for a local rise in African ancestry at 8q24 in the cases.) Freedman et al. (2006) corrected for this local effect in their African-American cases and controls by testing whether the differences in allele frequencies between cases and controls could be explained just from the enrichment of African ancestry in the cases. After correction, they found that the contribution of the microsatellite to disease risk was nonsignificant. However, when typing rs1447295 in 1,614 cases and 1,547 controls from four non-African populations (Japanese Americans, Native Hawaiians, Latino Americans and European Americans) they replicated a strong association signal ($P < 4.2 \times 10^{-9}$).

Together, these two studies support the hypothesis of rs1447295 being associated with prostate cancer risk in non-Africans, while also suggesting a higher proportion of as yet unidentified risk alleles at 8q24 in the African American population.

5.3 ARG Analysis of 8q24 data

In the study of Yeager et al. (2007), 550,000 SNPs were genotyped in 1,172 cases and 1,157 controls of European origin. The association signal at rs1447295 was replicated with $P = 9.75 \times 10^{-5}$ (using a four degree of freedom logistic regression test), with seven SNPs near to

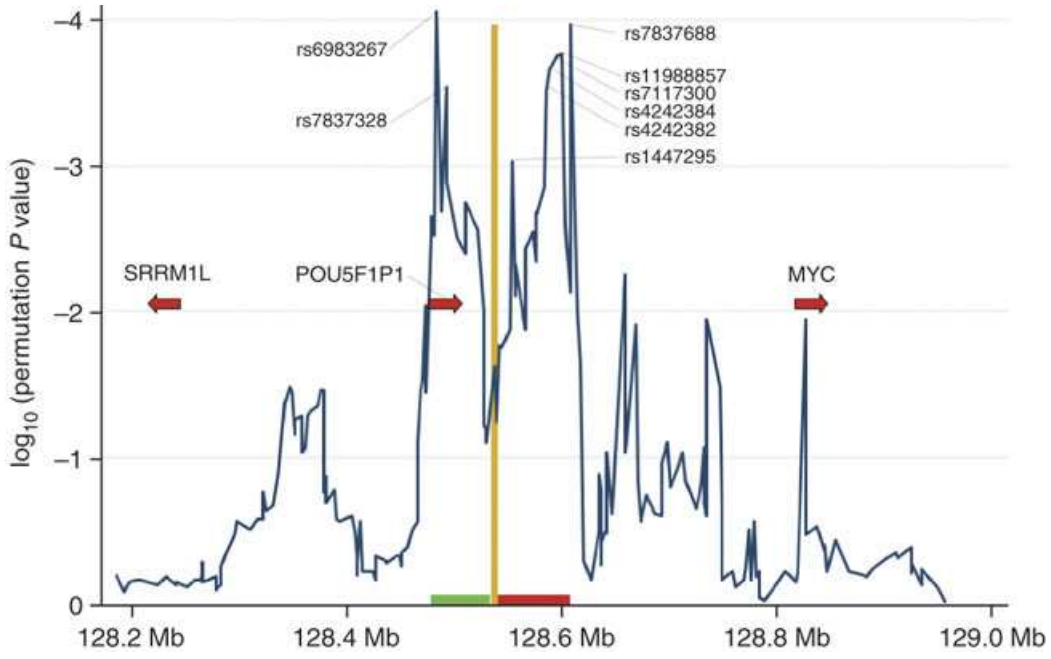


Figure 5.1: MARGARITA associations structure for the 8q24 region of Yeager et al. (2007).

rs1447295 showing still greater association.

In order to dissect the association signal, 197 SNPs in an 800kb region around rs1147295 were analysed using an adapted version of MARGARITA. I made modifications to MARGARITA so as to handle two departures from the standard model:

1. There are three phenotypes: case, non-aggressive prostate cancer and aggressive prostate cancer, with aggression defined by standard clinical phenotypes (Gleason index and disease stage).

2. Genotypes, rather than alleles were used in the chi-square test; which together with the three phenotypes gives a 3×3 contingency table with four degrees of freedom.

The results of the initial MARGARITA analysis are shown in Figure 5.1. MARGARITA gives two association signals: at a "centromeric" region around the association peak of rs6983267 and at a "telomeric" region around the peak of rs7837688.

When MARGARITA was first applied, the run time was significantly longer than expected for a similarly sized dataset with constant recombination rate; from experience this suggests a

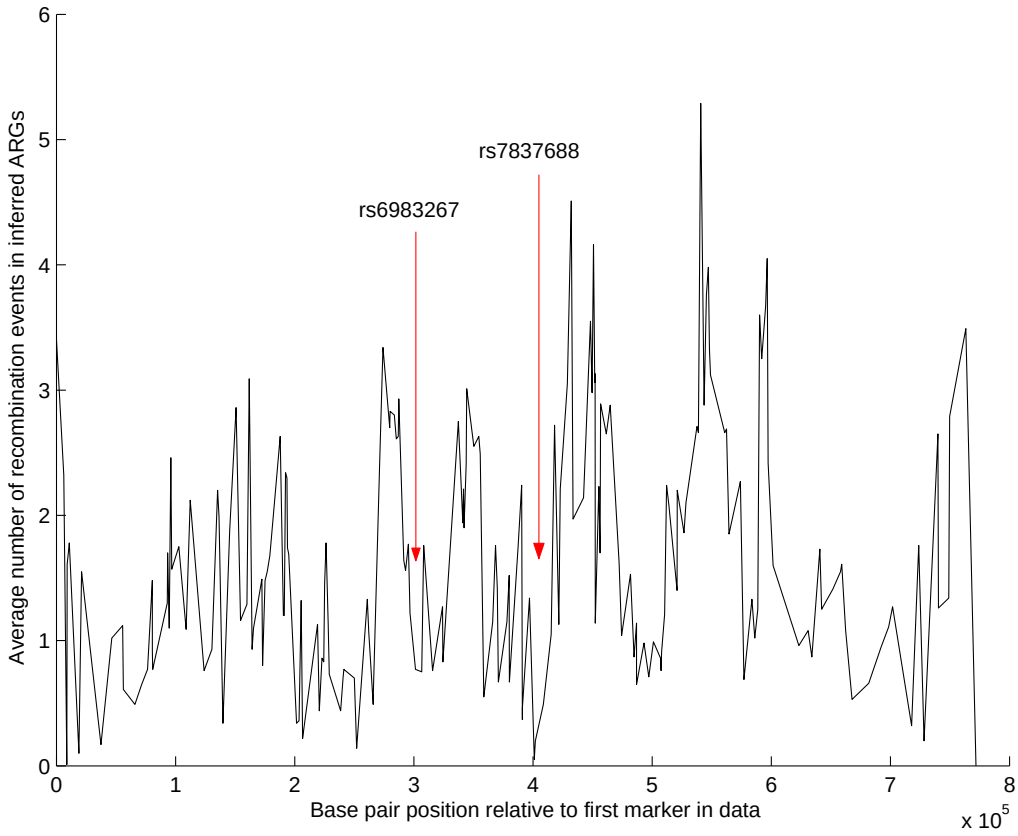


Figure 5.2: Number of inferred recombination events between markers in 8q24.

more complex recombination structure in the region, in particular one or more recombination hotspots.

The number of inferred recombination events between markers, averaged over 100 inferred ARGs, is shown in Figure 5.2. There is a pattern of peaks and valleys suggesting that there is local variation in recombination rate, including a peak between rs6983267 and rs7837688. However, it should be noted that MARGARITA is not a powerful method for estimating relative recombination rates because it only infers obligate recombination events, which are dependent on the frequencies of alleles.

The presence of a recombination hotspot separating those regions was confirmed by my collaborators at the NIH National Cancer Institute by applying the SequenceLDhot program of Fearnhead (2006). In the 130kb region covering the two peaks, SequenceLDhot identified a 5.5kb hotspot region which it estimated to contain 90% of the recombinations. This gives a population scaled recombination rate within the hotspot of 260, and 30 across the remainder

of the region. In Figure 5.1 the position of the recombination hotspot is shown with a yellow line.

Furthermore, the marginal trees from MARGARITA at the two peaks were found to be decorrelated, which, together with the evidence of a recombination hotspot, suggests that each region followed an independent history.

Since rs1447295 and rs7837688 both reside in the telomeric peak, and rs1447295 was identified as significant in previous studies, it was decided to focus on rs1447295 for the telomeric peak, rather than rs7837688, in the next analyses.

The ARGs were then used to estimate the frequencies of the causative alleles at rs1447295 and rs6983267. At rs6983267, the frequency of the inferred causative allele in the controls is 46% and at rs1447295 is 12%. These differences in inferred causative allele frequency further strengthen the hypothesis that the two signals are independent. These inferred causative allele frequencies also match the typed frequencies at rs6983267 and rs1447295, suggesting that the typed SNPs may be causative or essentially in complete association with the causative alleles. In the control populations, the predisposing allele of rs6983267 has frequency 48%, and at rs1447295 it has frequency 14%.

The ARGs were then used to guide a haplotype analysis, performed by my collaborators at the NIH National Cancer Institute. SNPs around each of rs6983267 and rs1447295 were phased using the program PHASE (Stephens and Donnelly, 2003): 20 SNPs around rs6983267, and 27 around rs1447295. The phase resolution with the greatest likelihood was taken, and 100 ARGs were constructed. The frequency with which each haplotype fell under the "best cut" for each region, that is, possessing an imputed causative polymorphism, was determined.

Figure 5.3 shows the results of this analysis. The haplotypes in green are those that tend to fall often under the best cut at the centromeric region (left) and the telomeric region (right), and those in red are those that tend to fall on the protective side of the best cut. The "Hap. freq" is the frequency of the haplotype in the data, and "Prediction" is the frequency with which the haplotype falls under the best cut.

For the centromeric region (around rs6983267), the protective haplotypes were found to be far less diverse than the susceptibility haplotypes, suggesting that the protective allele

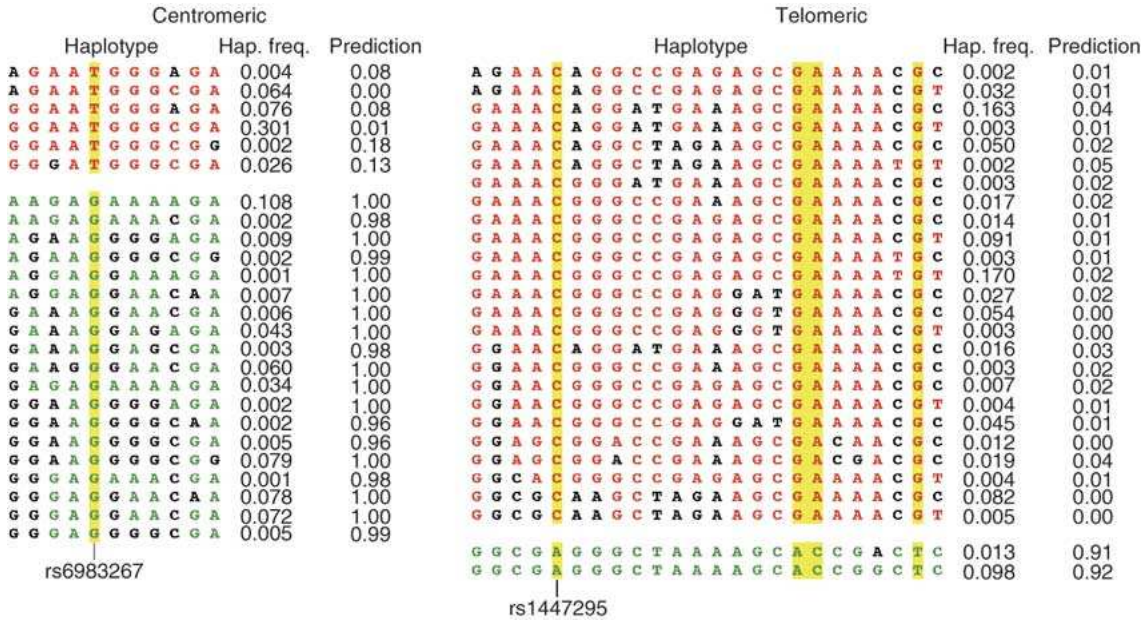


Figure 5.3: Haplotypes in the centromeric and telomeric peaks. Haplotypes coloured in red are those that tend to fall on the protective side of the marginal tree, and those coloured in green tend to fall under the imputed causative mutation. "Hap. freq." is the frequency of the haplotype in the data, and "Prediction" is the frequency with which the haplotype falls under the best cut.

is either more recent or positively selected. Further support for the hypothesis of positive selection comes from the observation that the protective allele at rs6983267 has frequency 52%, a high frequency given its relative haplotype diversity.

Conversely, the protective haplotypes in the telomeric region (around rs1447295) were found to be more diverse than the susceptibility haplotypes, suggesting that in this case the risk allele is either selected or more recent. Since the frequency of the risk allele of rs1447285 is 14%, there is better support here for the hypothesis that this risk allele is a recent mutation.

These conflicting genealogical accounts again support the hypothesis that the two markers correspond to distinct association signals, following independent histories. The deleterious mutation in the telomeric region is a more recent event than the protective mutation in the centromeric region.

5.4 Replication of Result

The signal at rs6983267 has been independently replicated in another case-control study (Haiman et al., 2007).

In Haiman et al. (2007) 2,973 SNPs were typed in up to 4,266 cases and 3,252 controls from five populations, across the 3.8Mb admixture peak of Freedman et al. (2006). They found three clusters of association, which they separated by comparing genetic and physical maps. Two of these clusters correspond to the "centromeric" and "telomeric" regions identified with the aid of MARGARITA in Yeager et al. (2007).

They then performed a stepwise logistic regression to determine the independence of the associated polymorphisms. This was done by incorporating each SNP into the model, in order of strength of association, and then repeating the analysis for the remaining SNPs, conditional on those already in the model. This analysis resulted in the identification of seven independent risk variants across 8q24, including rs6983267. However, the most strongly associated region of Haiman et al. (2007) was found sitting a few hundred kb centromeric from the signals found in the previously discussed case-control studies.

In another study, Gudmundsson et al. (2007) performed a genome-wide scan using 316,515 SNPs typed in 1,453 cases and 3,064 controls from Iceland. By testing each SNP separately, they replicated the previously identified signal at rs1447295, corresponding to the telomeric signal of Yeager et al. (2007). They then performed a haplotype block test, and identified another genome-wide significant signal in the same novel region as Haiman et al. (2007), and replicated these results in three other populations of European descent.

These studies together indicate that there are multiple independent risk alleles in 8q24, confirmed in multiple populations, but which do not lie in known genes. These variants could regulate nearby cancer causing genes, however, although there are genes in 8q24, such as the MYC oncogene, no differences in expression levels for genes in that region have been detected between carries and non-carriers of risk alleles (Gudmundsson et al., 2007). However, Haiman et al. (2007) note that 8q24 is the most frequently gained chromosomal region in prostate tumours, and speculate that the risk alleles make the region more prone to gain.

Chapter 6

Genotype Imputation

6.1 Motivation

There are two types of missing data in a case-control association study: some individuals have missing genotype values at loci which are otherwise successfully typed; and there are loci which are not typed at all. Most analysis methods use only the data that has been collected, but an alternative approach that has received a lot of attention in the statistical literature is to predict, or impute, missing values based on the observed data, and use the complete data for analyses (Rubin, 1987).

In what follows, *imputing missing genotypes* refers to imputing genotype values at loci which are typed in the sample of interest, but which are missing for a small fraction of the individuals in that sample. *Imputing untyped loci* refers to imputing genotypes at loci which are entirely untyped in the sample of interest, but which are typed in some other sample.

The ARG mapping approach described in Chapter 3 tests all branches on the marginal tree at a typed marker, that is, all possible SNPs (real or otherwise) that are compatible with the genealogy. A more direct approach is to test only those branches that correspond to known SNPs, typed or untyped. One way to do this would be to merge a sample of more densely genotyped individuals, such as from the HapMap Project, with the case-control sample. The ARG inference algorithm can then be used to impute genotypes at loci that are missing in the cases and controls but present in the denser sample. The imputed loci could

then be tested directly for association.

There are at least three additional reasons why we might want to estimate, or impute, genotypes that are not present in observed data.

First, any large-scale genotyping effort is likely to produce incomplete data. Depending on the subsequent analyses, and the processes generating the missing genotypes, incomplete data can lead to biased results. For example, in case-control studies, a systematic bias in the genotyping quality of an allele may result in false, or missed associations (Clayton et al., 2005), and incomplete data can hinder multi-SNP analyses of haplotypic or interaction effects. Furthermore some analysis methods are unable to operate on data that has any missing genotypes.

Second, imputation of untyped loci may help fine map disease causing variants. By imputing untyped loci at an association peak, additional informative association structure in that region may be found. If an imputed locus shows stronger association than the observed loci around it, it may be causative, or more strongly linked to the causative polymorphism(s).

Third, imputing untyped loci has the potential to give greater power to detect disease associations than relying on pairwise LD. As discussed in the Introduction, when the single marker chi-square test is applied to case-control association study data, the association signal at a SNP in LD with a causative polymorphism is dependent on the r^2 LD between those two loci (Pritchard and Przeworski, 2001). If no typed SNP is in sufficient r^2 LD with the untyped causative polymorphism, the association will be missed. Furthermore, it is conceivable that allelic heterogeneity and interaction can cause single marker tests at SNPs in strong r^2 LD with an untyped causative variant to show no significant signal (Terwilliger and Hiekkalinna, 2006). However, accurate imputation of the untyped causative polymorphism will recover the association signal. The scenario (albeit, a very unlikely one) constructed by Terwilliger and Hiekkalinna (2006) is as follows:

Consider a three SNP haplotype, with loci A, B and C, with alleles a, A; b, B; and c, C. Assuming that there is no recombination between these alleles, and no recurrent or back mutations, there can be at most four possible haplotype configurations. Suppose we observe the following haplotypes: A-B-C, A-B-c, a-B-C and A-b-c, each with equal frequencies, 0.25, in

the population. Now suppose that SNPs A and B are untyped and modify disease risk, while SNP C is typed but is not causative. If a and b affect the disease phenotype in a dominant fashion, with equivalent effect, i.e., $P(Disease|aa) = P(Disease|Aa) = P(Disease|bb) = P(Disease|Bb)$ and $P(Disease|AA, BB) = 0$, then the following contingency tables are possible:

Allele at A	Freq in Cases	Freq in Controls
A	0.667	0.757
a	0.333	0.243
		Odds Ratio 1.55

Allele at B	Freq in Cases	Freq in Controls
B	0.667	0.757
b	0.333	0.243
		Odds Ratio 1.55

Allele at C	Freq in Cases	Freq in Controls
C	0.5	0.5
c	0.5	0.5
		Odds Ratio 1.00

In this scenario, although loci A and C have $r^2 = 0.33$, the association at A will never be detected by genotyping C alone, regardless of sample size, because the odds ratio at C is 1.00. The same holds for detecting the association at B via C. In such a situation, A and B must be tested directly, or a multimarker method used. Accurate imputation of A and B and then "direct" testing would successfully identify the signal.

In this chapter, I evaluate the imputation performance of MARGARITA and compare it with fastPHASE (Scheet and Stephens, 2006), and apply the MARGARITA imputation approach to the 8q24/Prostate Cancer data of Yeager et al. (2007) to test loci not typed in the original

study.

6.2 Existing Methods for Imputing Genotype Data

There are a number of methods for imputing missing genotypes. The widely used PHASE (Stephens et al., 2001; Stephens and Donnelly, 2003; Stephens and Scheet, 2005) and fastPHASE (Scheet and Stephens, 2006) methods impute missing genotypes while inferring haplotype phase.

fastPHASE, the model for which is described in Chapter 2, can be used to impute missing genotypes in the following way: Once the model is fitted, the probability that a missing genotype takes a particular value is dependent on the probabilities of its haplotypes belonging to particular clusters, and the frequencies of the observed genotypes within those clusters. A point estimate of the missing genotype is taken by choosing the genotype with the greatest probability.

Within the context of association studies, there has been some discussion in the tagSNP literature (Goldstein et al., 2003) on explicitly estimating the genotypes of untyped SNPs from genotyped tagSNPs. Methods have been developed to do this (Evans et al., 2004; Souverein et al., 2006). It is also possible to select tagSNPs that optimise subsequent prediction accuracy of untyped SNPs (Nicolae, 2006; Paschou et al., 2007; Eyheramendy et al., 2007). However, only Souverein et al. (2006) and Paschou et al. (2007) apply their method to case-control association study data. I now briefly describe each of these approaches.

In Souverein et al. (2006), a linear or logistic regression model was fitted to a training data set containing genotype data for all SNPs, and then used to impute loci missing in the less densely typed data. The authors selected manually which predictor SNPs to use in the regression, and their method requires that the predictor SNPs are complete.

In Evans et al. (2004) windows of SNPs were taken, and population haplotype frequencies were determined from a training set. The probability that an individual has a particular value at a missing genotype was calculated by taking all the haplotypes matching the observed genotypes for the individual, and using these to fill in the missing value. The contribution of

each matching haplotype to the missing genotype was weighted by haplotype frequency. A similar approach was taken in Nicolae (2006).

Goldstein et al. (2003) discusses the idea of selecting tagSNPs on the basis of how well models involving those predict the untyped SNPs. This approach was adopted in Nicolae (2006); Eyheramendy et al. (2007) and Paschou et al. (2007).

Eyheramendy et al. (2007) describe a method for predicting non-tags using the Li and Stephens model (Li and Stephens, 2003). The model is fitted to training data, such as the HapMap, and then individuals sampled from the case-control study are introduced, and the missing genotypes imputed from the model. However, the method was only applied to the dense ENCODE HapMap data, not to an association study. They selected tagSNPs for the ENCODE data, fitted the model using a training set of haplotypes, and then measured how well the non-tags were reconstructed in the rest of the data, the same population.

Paschou et al. (2007) describe a method for choosing tagSNPs from a data set. During selection of tagSNPs, a singular value decomposition is performed on the genotype data matrix. The resultant eigenvectors give linear combinations of SNPs that capture the structure of the data, and can be used to select the SNPs that contribute the most information, which become the tagSNPs. When a population is typed with those tagSNPs, the information from the decomposition can be applied to reconstruct the missing genotypes. A drawback of this approach is that it requires exactly those tagSNPs, as defined by the singular value decomposition procedure, to be typed. Hence, when they applied it to an association study, they split the data into two: selecting tagSNPs from one half, then "assaying" those tagSNPs in the other half, and then imputing the untyped SNPs.

The advantage of methods such as fastPHASE and Margarita is that they can be applied to data with any pattern of missingness. Therefore, in this chapter, I compare the imputation performances of these most flexible methods.

Just prior to submission of this thesis, Servin and Stephens (2007) published a method that tackles the problem of imputing and testing untyped loci. They used fastPHASE to impute untyped loci in quantitative trait association studies, and then tested the imputed loci directly using a Bayesian regression approach. The advantage of using Bayesian regression

is that it naturally handles the uncertainty in imputed values. They found that compared to single SNP ANOVA tests and a linear regression test on tagSNPs only, the approach of imputing and testing with Bayesian regression appears to have greater power to detect rare variants.

The Wellcome Trust Case Control Consortium (2007) describes a genome wide association study for seven diseases, typed for 469,557 SNPs in 2,000 cases each, and using 3,000 shared controls. An multilocus method (Marchini et al., 2007) based on Li and Stephens (2003) was used to impute data at over 2 million HapMap SNPs not typed in the study, and these were tested for association. In one of the peaks associated with Type 2 Diabetes, an untyped SNP (imputed from the Phased II HapMap) was identified with stronger significance that the surrounding typed SNPs.

6.3 Imputing Missing Genotypes and Untyped Loci, and Testing for Association

As described in Chapter 2, missing data can be imputed using inferred ARGs. Missing alleles are imputed when two compatible sequences coalesce, where one sequence has an observed allele at the position which is missing in the other sequence. The missing position takes the allele of the other sequence, and this assignment is propagated down the ARG to the leaves.

I used two ARG strategies for imputing data:

- Margarita-full constructs ARGs for all the individuals in the data together. This approach can always be used to impute genotypes which are observed in some of the individuals.

- Margarita-one is only used when imputing loci that are untyped in a sample, by merging in a more densely typed sample. Rather than constructing ARGs for all the individuals together, ARGs are inferred for all the densely typed individuals and only one of the sparsely typed individuals at a time. The motivation for this is that additional individuals from the sparse data do not contribute to imputation at the untyped loci.

In order to obtain a single "best" estimate for missing data, I infer multiple ARGs and take the most frequently imputed genotype at each position as the consensus imputation. Where I compare the imputation accuracy of MARGARITA's consensus imputation with that from fastPHASE (Servin and Stephens, 2007), I use fastPHASE's default parameters.

However, for association testing, it is important to handle the uncertainty in genotype imputation, and I therefore analyse each of the ARG imputations rather than only the consensus imputation, and incorporate the uncertainty in imputation by using the methodology of Multiple Imputation (MI) (Rubin, 1987; Little and Rubin, 1987; Cordell, 2006; Souverein et al., 2006; Dai et al., 2006; Mensah et al., 2007).

MI is a simulation-based approach in which missing data are imputed multiple times, to give a number of complete data sets. Each complete data set is then analysed using some standard method. The statistic from the analysis is averaged over the imputations to give a single estimate, and the within- and between- imputation variances of the estimate are calculated. It is then possible to calculate significance for the estimated statistic (Rubin, 1987).

Specifically, I inferred $k = 30$ ARGs for each data set, giving k imputations. The choice of 30 was selected by considering the variance in the estimated odds ratios for imputed loci, which is given below. To test an imputed locus for association I calculated the log odds ratio $\hat{L}_i$ using the genotypes in imputation i. I then took the mean log odds ratio over the k imputations,

$$\bar{L} = \frac{1}{k} \sum_{i=1}^{k} \hat{L}_i,$$

to obtain an estimated log odds ratio for that locus. For imputing untyped loci in the Yeager et al. (2007) data, discussed later, $k = 30$ imputations gives a standard error of the mean, $\bar{L}$, of, on average, 9% of $\bar{L}$ (range 0.1% to 20%).

Note that the denser samples from which the missing loci are imputed are not included in calculating the log odds ratios.

In order to calculate confidence intervals and P-values for these estimated log odds ratios, the additional variance due to imputation uncertainty must be taken into account. I used

Rubin's rules to do this (Rubin, 1987), which combine the within-imputation variances (the variance for each $\hat{L}_i$), with the between-imputation variance (the variance of the sample mean $\bar{L}$). The estimated log odds ratio is distributed as a Student's t-distribution, with degrees of freedom and variance determined by the variance components just described. Testing for departure from the null hypothesis of no association is then achieved in the usual way, by comparison to this distribution. Rubin's rules are as follows:

The average within-imputation variance is

$$\bar{W} = \frac{1}{k}\sum_{i=1}^{k} W_i,$$

where the variance, W_i, for a log odds ratio, $\hat{L}_i$, is the sum of the reciprocals of the counts in each cell in the contingency table.

The between-imputation variance is

$$B = \frac{1}{k-1}\sum_{i=1}^{k}\left(\hat{L}_i - \bar{L}\right)^2,$$

and then the total variability associated with the estimate $\bar{L}$ is

$$T = \bar{W} + \frac{k+1}{k}B.$$

Then, for significance testing against $\bar{L} = 0$ and confidence interval estimation,

$$\frac{\bar{L}}{\sqrt{T}} \sim t_v$$

where the degrees of freedom for the t-distribution is

$$v = (k-1)\left(1 + \frac{1}{k+1}\frac{\bar{W}}{B}\right)^2.$$

MI approaches have already been used in two ways in association studies: First, to handle the uncertainty in haplotype phase when testing for haplotype specific effects (Cordell, 2006;

Mensah et al., 2007); and second, to handle the uncertainty in imputed genotypes (but not entirely missing loci) when testing for single SNP effects (Souverein et al., 2006; Dai et al., 2006). In Cordell (2006) and Mensah et al. (2007) haplotypes are estimated multiple times from the data, and disease model parameters are estimated by taking the mean of the estimates derived from each of the haplotype estimations. In Dai et al. (2006) three missing genotype imputation methods are compared for SNP data; missing genotypes were imputed multiple times and odds ratios calculated by MI.

In the experiments described below I compare my imputation approach to one where only the loci typed in the case-control sample are tested. The association signal at a typed locus is determined by calculating the log odds ratio from the observed data, and P-values calculated by comparison to a Normal distribution.

I performed experiments to test (1) the accuracy of missing genotype imputation; (2) the accuracy of untyped locus imputation; and (3) whether additional insights can be gained for fine-mapping.

6.4 Results for Imputing Missing Genotypes

To test the imputation accuracy for missing genotypes, I used two data sets:

- ASH. 163 Ashkenazi controls and 293 cases from a case-control study of association between a 10Mb region of chromosome 20 and Type 2 Diabetes (Barroso et al., 2007), typed with an average density of 1 SNP/2.5kb.
- NBS. 400 UK controls from the UK National Blood Service Control Cohort (The Wellcome Trust Case Control Consortium, 2007). I used chromosome 20, where there is an average density of 1 SNP/5kb.

The experiments were parameterised according to the population (ASH controls, ASH cases and controls, or NBS), and proportion of genotypes removed at random from the data (1%, 5%, 10% or 20%). For each parameterisation, 50 experiments were simulated, each one involving approximately 1Mb (400 SNPs for ASH, 200 SNPs for NBS) from a randomly chosen region, with the specified fraction of genotypes removed. Each data set had some

Population	% missing	MARGARITA-FULL	FASTPHASE
ASH controls	1	**0.019 (0.018,0.020)**	0.020 (0.019,0.021)
ASH controls	5	0.021 (0.020,0.022)	0.021 (0.020,0.022)
ASH controls	10	0.022 (0.021,0.023)	0.022 (0.021,0.023)
ASH controls	20	0.026 (0.025,0.027)	**0.025 (0.025,0.026)**
ASH cases and controls	1	**0.011 (0.011,0.012)**	0.017 (0.016,0.018)
ASH cases and controls	5	**0.013 (0.012,0.013)**	0.020 (0.019,0.021)
ASH cases and controls	10	**0.013 (0.012,0.013)**	0.019 (0.019,0.020)
ASH cases and controls	20	**0.017 (0.017,0.018)**	0.024 (0.023,0.025)
NBS	1	0.048 (0.044,0.051)	**0.034 (0.031,0.036)**
NBS	5	0.049 (0.047,0.052)	**0.036 (0.034,0.038)**
NBS	10	0.051 (0.048,0.054)	**0.036 (0.034,0.038)**
NBS	20	0.057 (0.054,0.060)	**0.038 (0.036,0.040)**

Table 6.1: Mean imputation error rates for missing genotypes, with standard error intervals in brackets.

missing data itself (0.8% for NBS chromosome 20 and 0.5% for ASH) which was imputed but not assessed.

MARGARITA-FULL and FASTPHASE were applied to these data sets, and their imputations compared to the held out observed genotypes. Table 6.1 reports the mean imputation error rate (the number of incorrect genotype imputations divided by the number of removed genotypes) for each parameterisation.

As can be seen from Table 6.1 and Figure 6.1 both methods perform better on the ASH data (with error rates of 1-3%) than on the NBS data (error rates 3-6%). There may be a number of reasons for this. First, the ASH data is typed more than twice as densely. Second, the markers in the ASH data were chosen to be non-redundant ($r^2 < 1$), whereas the NBS markers were not chosen with such a strong tagging requirement. Third, the Ashkenazi population is smaller and more homogeneous than the UK population; this can potentially give LD that extends over longer regions.

On the ASH data, the imputations are more accurate for the cases and controls combined than for the cases alone. This improvement is likely due to the increased sample size and increased homogeneity within the cases.

I also tested whether differences in the allele frequency spectrum between the ASH and NBS populations could affect imputation error rate. Figure 6.2 shows the minor allele fre-

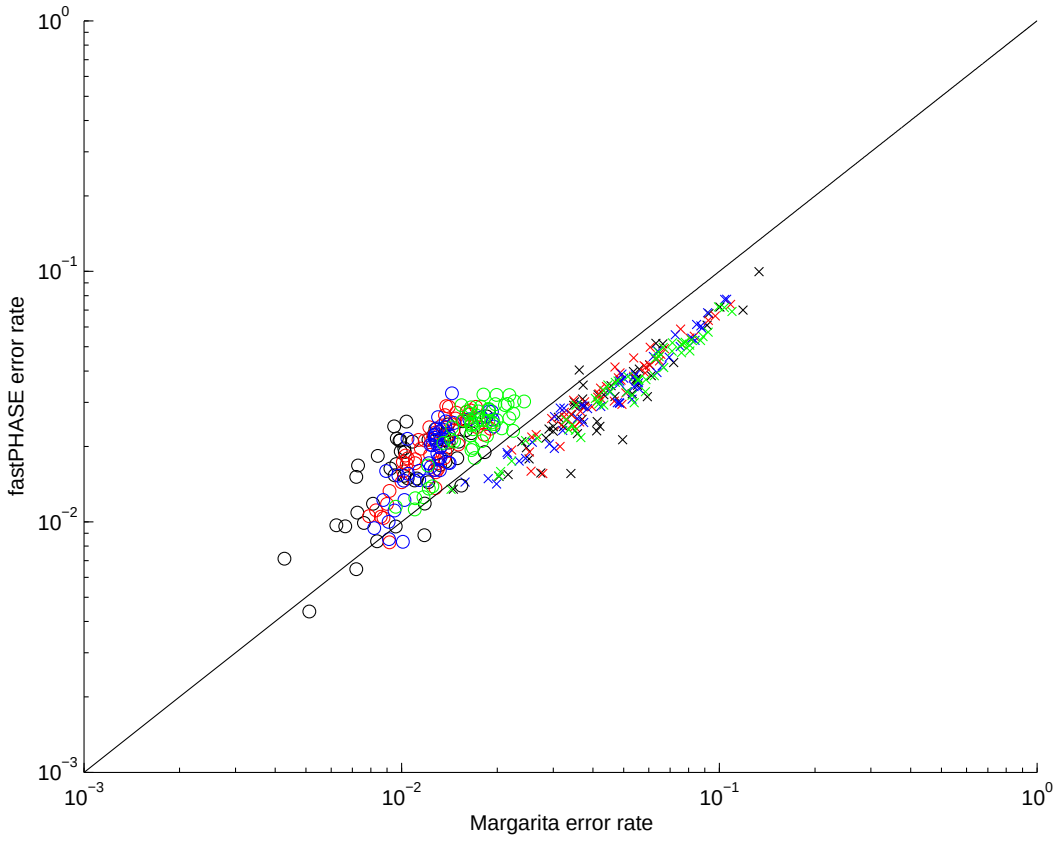


Figure 6.1: Error rates for imputation of genotypes removed at random from the data. Crosses are NBS, circles are ASH. Black is 1% missing data, red is 5%, blue is 10%, green is 20%.

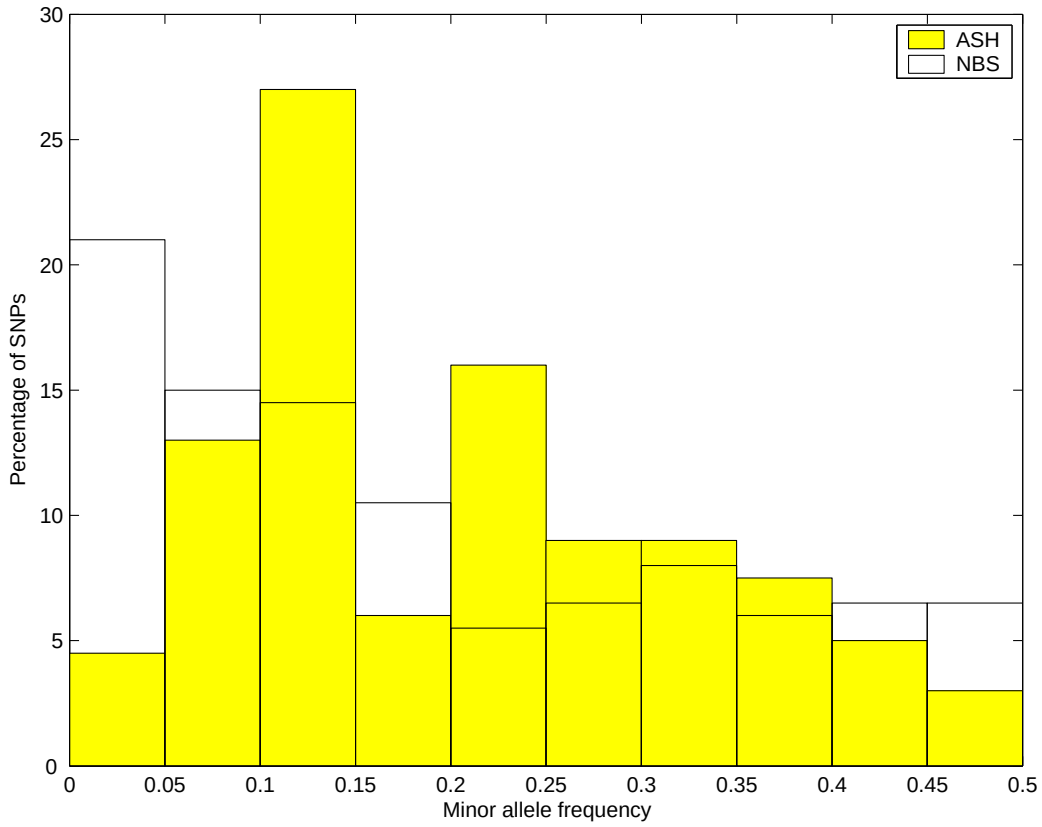


Figure 6.2: Minor allele frequency spectrum for a randomly chosen 1Mb of the ASH controls and NBS data.

quency spectrum for the two populations, and shows that the NBS has more rare alleles than the ASH controls. MARGARITA-FULL typically outperforms FASTPHASE on the ASH data, while FASTPHASE outperforms MARGARITA-FULL on the NBS data. One explanation may be that the NBS data has more rare alleles than the ASH data. We will observe in the next section that the ratio of MARGARITA-ONE error to FASTPHASE error has a slight tendency to be greater for less frequent alleles.

6.5 Results for Imputing Untyped Loci

In order to test how well loci which are untyped in one sample can be imputed from a more densely typed sample, I combined the NBS and HapMap data sets. For each simulation, a SNP typed in both samples was removed entirely from the NBS data, but kept in the HapMap data. I then took 200 NBS and/or HapMap SNPs either side of the removed locus,

Population	MARGARITA-FULL	MARGARITA-ONE	FASTPHASE
CEU+NBS	0.060 (0.049,0.071)	**0.058 (0.047,0.070)**	0.097 (0.081,0.114)
CEU+YRI+NBS	0.087 (0.072,0.102)	0.087 (0.071,0.103)	0.087 (0.068,0.106)
YRI+NBS	0.193 (0.159,0.227)	**0.173 (0.142,0.204)**	0.250 (0.211,0.289)

Table 6.2: Mean imputation error rates for untyped loci with standard error intervals in brackets.

corresponding to around 300-600kb, for 400 NBS individuals and the unrelated and unphased CEU and/or YRI HapMap. 50 of these experiments were performed using MARGARITA-FULL, MARGARITA-ONE and FASTPHASE. An additional 450 CEU+NBS experiments were performed using MARGARITA-ONE only, in order to gain further insight into how imputation performance varies with minor allele frequency at the untyped locus.

Table 6.2 gives the error rates for untyped locus imputation.

Using the CEU HapMap to impute missing loci gives a much lower error rate than using the YRI HapMap. This is not surprising; the NBS is a European population, and the CEU is a European-derived population, whereas the YRI is an African population. This result suggests that ensuring reasonable matching between the dense and case-control samples is important for accurate imputation.

When imputing using the CEU and YRI samples together, FASTPHASE achieves its best performance. The larger sample size may help. However, the performance of MARGARITA-ONE is intermediate between the NBS+CEU and NBS+YRI configurations.

Both MARGARITA-FULL and MARGARITA-ONE tend to perform better than FASTPHASE, with MARGARITA-ONE marginally better than MARGARITA-FULL.

Figure 6.3 shows how the relative error rate of MARGARITA-ONE and FASTPHASE varies with minor allele frequency at the missing locus. The relative error on the y-axis is the $\log_{10}$ of the MARGARITA-ONE error rate divided by the FASTPHASE error rate, for each of the 50 CEU+NBS experiments. There is a slight (non-significant) tendency for the ratio of MARGARITA-ONE error rate to FASTPHASE error rate to be lower for loci with greater minor allele frequency (for 50 CEU+NBS experiments, Spearman's rank correlation coefficient -0.07, $P = 0.64$).

Figure 6.4 gives the MARGARITA-ONE error rate versus minor allele frequency of missing

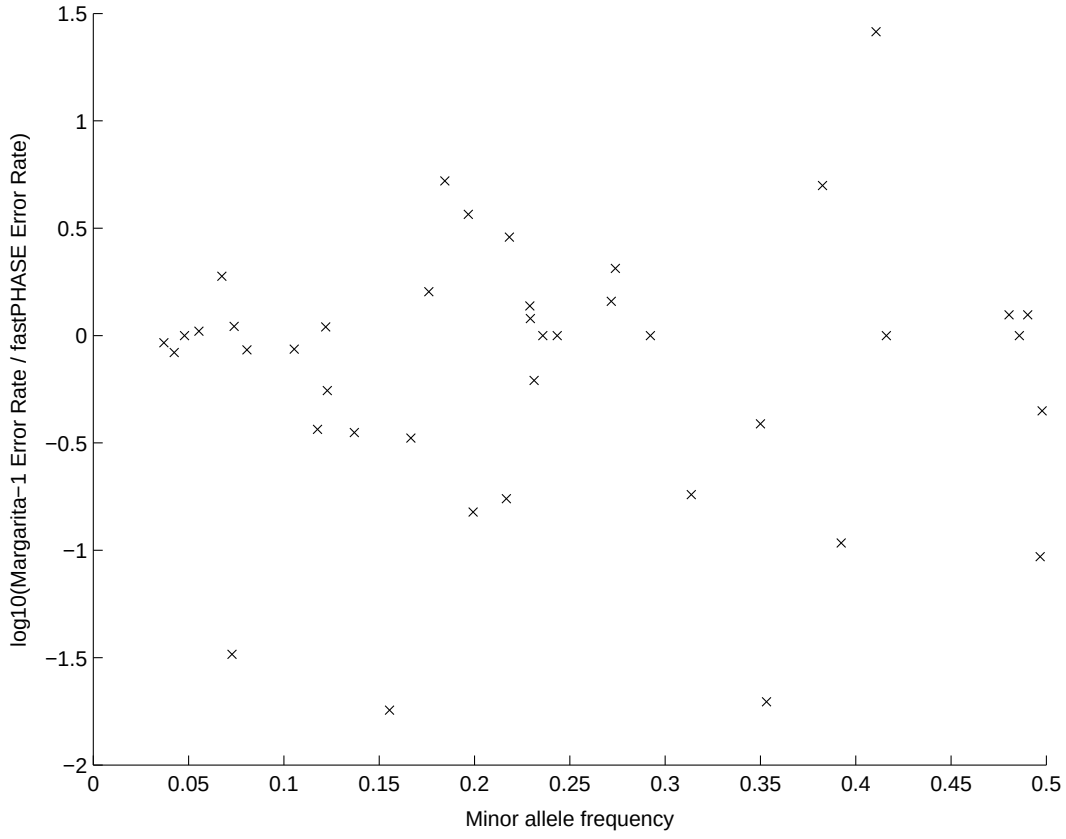


Figure 6.3: MARGARITA-ONE versus FASTPHASE error rates for imputation on 50 NBS+CEU data sets. Relative performance is plotted against minor allele frequency.

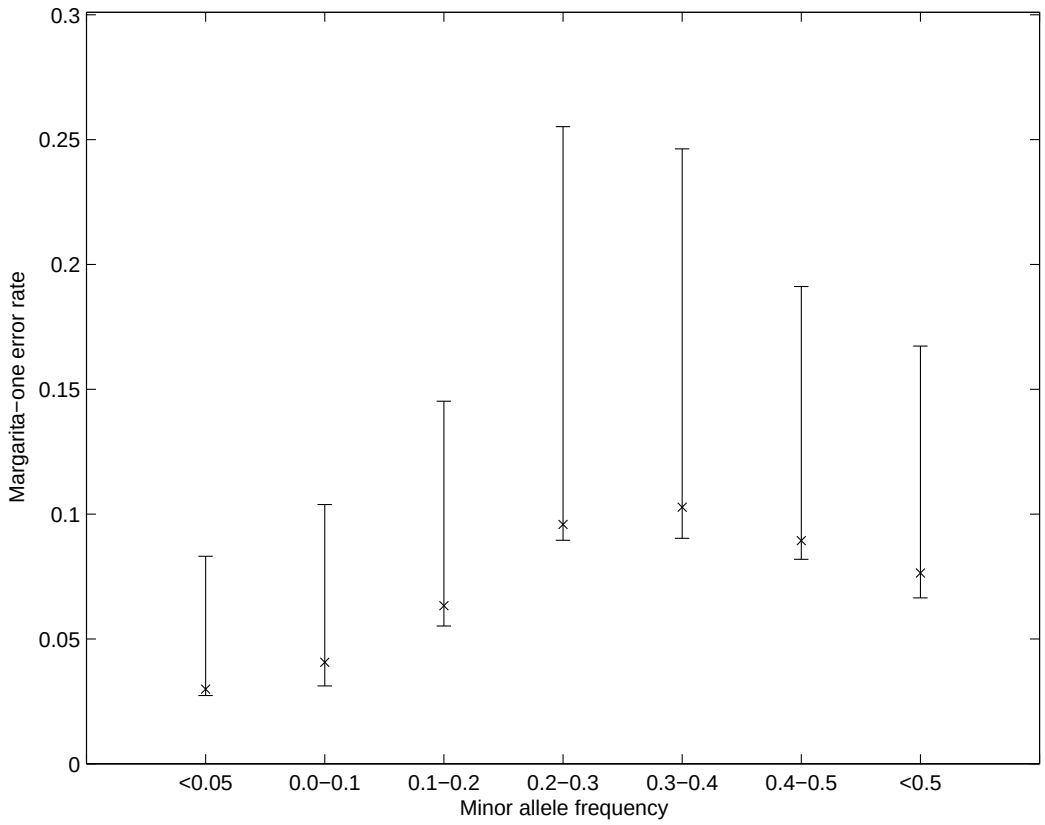


Figure 6.4: MARGARITA-ONE error rate (mean and inter-quartile range) versus minor allele frequency of imputed loci in 500 NBS+CEU experiments.

loci for 500 CEU+NBS experiments. The imputation error rate is lower for loci with lower minor allele frequencies (Spearman's rank correlation coefficient 0.11, P = 0.02). This may be expected, because there is inherently more uncertainty in the value of a genotype with greater minor allele frequency.

6.6 Results for Imputation of Untyped Loci in 8q24

A data set containing the European American samples analysed in Chapter 5 (Yeager et al., 2007) and the CEU HapMap was provided by G. Thomas of the NIH National Cancer Institute. This data set contains the 1169 prostate cancer cases and 1094 controls, and the 60 CEU parents, covering 1Mb of 8q24, with 250 SNPs typed in both the Yeager et al. (2007) sample and the HapMap, and 898 SNPs typed in the HapMap alone.

In order to test how well the association signal for an imputed locus matches the true

association signal, I constructed 250 experiments: one for each SNP typed in both the Yeager et al. (2007) sample and the HapMap. This was done by removing that SNP from the Yeager et al. (2007) data, but keeping it in the HapMap sample. 200 SNPs either side of the removed locus were then taken, corresponding to 300-600kb, and presented to MARGARITA-ONE.

The mean imputation error rate is 0.1213 and the median is 0.0898.

The P-values for the imputed log odds ratios and the observed log odds ratios are shown in Figure 6.5. These results show that loci which are significant when observed tend to be significant when imputed, and that imputation does not create false positives.

I also took the full 1Mb of data and imputed all loci typed in the HapMap, but untyped in the Yeager et al. (2007) sample. The results for all imputed loci are shown in Figure 6.6.

Table 6.3 gives the imputed odds ratios for SNPs untyped in the Yeager et al. (2007) sample but typed and found significantly associated across multiple populations in either Haiman et al. (2007) or Gudmundsson et al. (2007). Along with the imputed odds ratios, the table also gives the odds ratios reported in Haiman et al. (2007) and Gudmundsson et al. (2007). It should be noted that different samples are used between studies, which may have different allele frequencies and vary in size and population. Therefore, I only report the odds ratios for European-derived populations.

The SNPs rs13254738, rs6983561 and rs16901979 correspond to the most centromeric association signal found in Haiman et al. (2007) and Gudmundsson et al. (2007), for which there appears to be no significant evidence of association in our data set, imputed or observed.

However, rs13254738 and rs6983561 are not significant when only the European American population in Haiman et al. (2007) is considered, thus the most likely explanations for lack of imputed significance is that these SNPs are not causative, or have lower allele frequencies, or weaker effects in European-derived populations. Indeed, the odds ratios for these SNPs in the observed Haiman et al. (2007) European American data and the imputed odds ratios show rough agreement (Table 6.3).

Gudmundsson et al. (2007) report rs16901979 as being significant across multiple European populations; however, I do not find an imputed significant signal. This is most likely due to lack of power of the imputation approach, caused by there being only two copies of

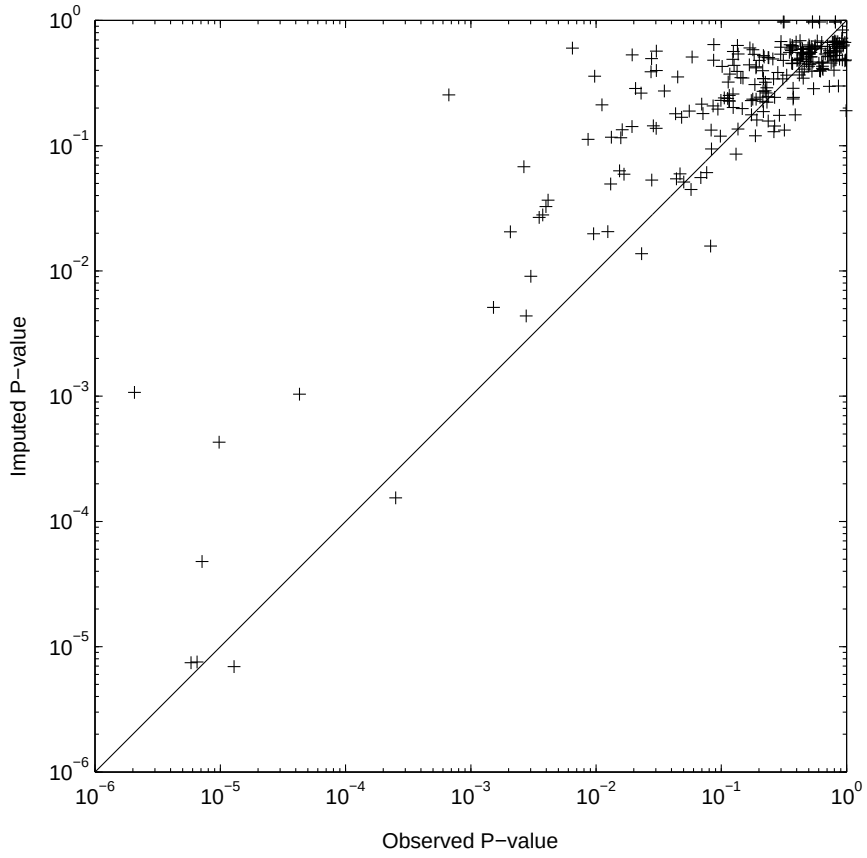


Figure 6.5: Observed association test P-values compared with those obtained by imputation, for SNPs typed in the Yeager et al. (2007) sample.

Study	SNP	Study OR	Imputed OR
Haiman et al.	rs13254738	1.11 (0.97-1.26)	1.06 (0.92-1.21)
Haiman et al.	rs6983561	1.16 (0.86-1.58)	1.22 (0.85-1.76)
Gudmundsson et al.	rs16901979	1.79 (1.53-2.11)	1.17 (0.85-1.63)
Haiman et al.	rs7000448	1.14 (0.98-1.40)	1.20 (1.06-1.36)
Haiman et al.	rs10090154	1.44 (1.17-1.76)	1.50 (1.24-1.81)

Table 6.3: Imputed odds ratios (ORs) and 95% confidence intervals for SNPs untyped in the Yeager et al. (2007) sample but which were found to be significant across multiple populations in other studies. Study ORs are from the European populations only (European Americans in Haiman et al. (2007) and all European populations combined in Gudmundsson et al. (2007)).

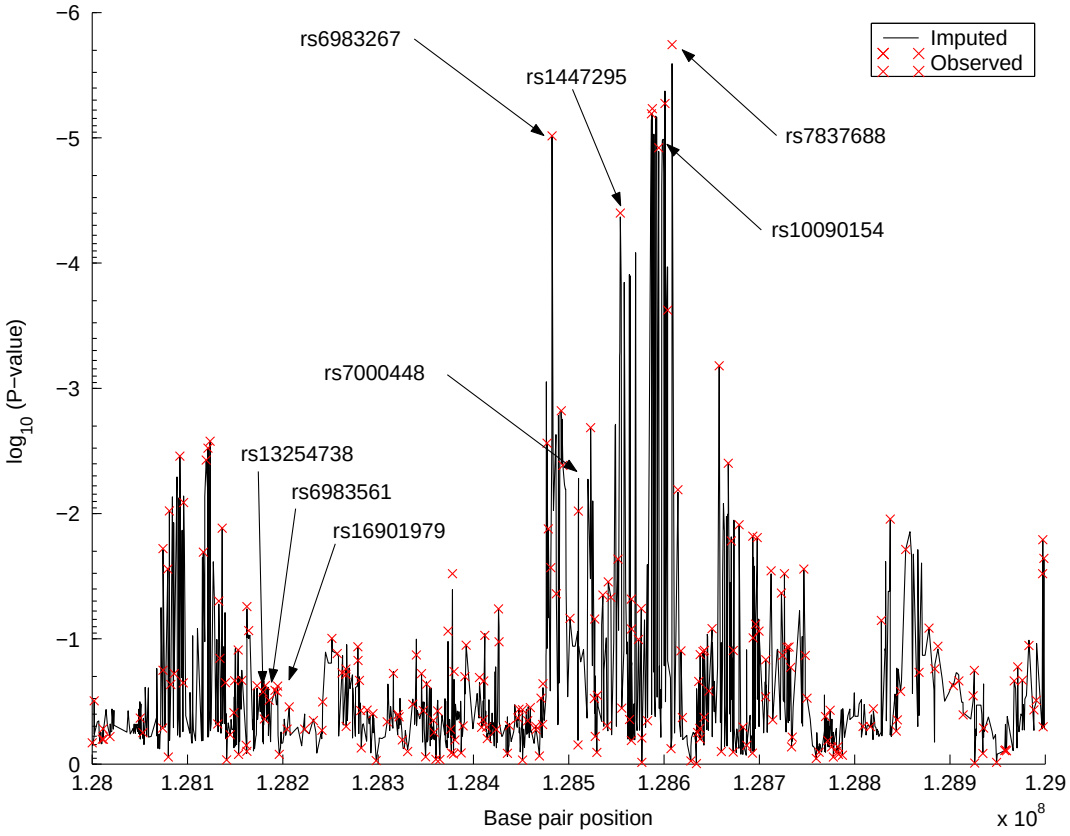


Figure 6.6: Imputed P-values for the Yeager et al. (2007) data. rs13254738, rs6983561, rs16901979, rs7000448 and rs10090154 are untyped in the Yeager et al. (2007) sample, but were found to be associated with prostate cancer in either Haiman et al. (2007) or Gudmundsson et al. (2007). rs6983267, rs1447295 and rs7837688 are typed in the Yeager et al. (2007) sample.

the rs16901979 risk allele in the CEU sample from which I impute.

The imputed association at rs7000448 has a greater odds ratio than that given in Haiman et al. (2007), where this association was originally detected. The significance at this imputed locus is SNP-wise P=0.005, which is suggestive of association. This SNP is in a region of high recombination and so could correspond to a causative polymorphism distinct from the association peaks either side of it (rs6983267 and rs1447295 in Figure 6.6).

The imputed SNP rs10090154 has strong significance (SNP-wise $P = 3.3 \times 10^{-5}$). This was previously detected in the Haiman et al. (2007) study, and is located within the telomeric peak of the observed Yeager et al. (2007) data. Many of the imputed and typed SNPs in the proximity of rs10090154 show strong significance, and are likely to correspond to the same causative variant. In the CEU HapMap, rs10090154 has $r^2 = 1$ with rs1447295 and rs7837688, which are typed in the Yeager et al. (2007) sample and are highly significant.

Although no new association peaks are found, the imputed data does show additional structure around rs1447295. The imputed odds ratios for untyped SNPs around rs1447295 are stronger than for rs1447295 itself, although after accounting for the increased variance due to imputation, the P-values for these SNPs are less significant. The fact that many of the imputed loci have greater odds ratios suggests that rs1447295 may be in LD with the causative variant(s) and not itself causal.

6.7 Discussion

Rather than testing each branch of inferred ARGs for disease association, a more direct approach is to test only those branches that correspond to segregating sites. I have done this by inferring ARGs for case-control data combined with more densely genotyped controls.

The performance of MARGARITA for imputing missing genotypes and untyped loci is in general comparable to that of FASTPHASE, with each appearing better in some conditions. In Scheet and Stephens (2006), FASTPHASE achieves an error rate of 3-4% on CEU HapMap data where genotypes are removed at random. In my experiments on more individuals less densely typed, similar error rates are achieved. Locus imputation, however, is harder, and

the error rates tend to be above 5%, which is likely to have a significant impact on the power of the imputation approach to detect true associations.

Because I am using HapMap samples to impute loci in cases and controls we might expect the association signals to be deflated, since the cases are being imputed from a relatively small sample which may not be ideally matched. I observed that the error rate for imputing loci in cases and controls from the Yeager et al. (2007) sample is greater than that for imputing loci in the NBS controls. Therefore, one way to increase the imputation accuracy at untyped loci may be to densely genotype (or in the future, resequence) case and control individuals from the case-control study, rather than using an external generic sample. This also suggests that a larger densely typed sample, e.g. a larger HapMap, would help imputation based approaches.

Another scenario in which this method may be valuable is where different case-control studies are combined, involving individuals from the same or related populations but genotyped on different platforms, and thus with mostly different typed SNPs. However, I do not consider this scenario here.

Analysis of the prostate cancer data in 8q24 of Yeager et al. (2007) suggests that untyped locus imputation may be useful for identifying additional association structure. For most SNPs that have been reported as associated in other publications, the imputation approach makes fairly accurate estimates of the odds ratios. One SNP which was found to be significant in Gudmundsson et al. (2007) was not imputed as significant in my experiment; however this SNP has minor allele frequency 1.7% in the CEU HapMap sample. This again suggests that a much larger set of densely typed individuals would be valuable.

Chapter 7

Additional Applications of the Algorithm

In this chapter I describe some additional applications of the method to questions in population genetics. These descriptions are not intended to be rigorous, but rather give illustrative examples towards how inferred ARGs could be used.

7.1 Detecting Selective Sweeps

A popular technique for identifying selective sweeps from population SNP data is to look for tracts of unexpectedly long haplotype sharing, called extended haplotype homozygosity (Sabeti et al., 2002; Nielsen et al., 2005; Hanchard et al., 2006; Voight et al., 2006). When a favoured allele increases rapidly in frequency, it will tend to reside on an unusually long haplotype of low diversity. This is because there has not been sufficient time for recombination to break down the LD. Meanwhile, chromosomes that do not carry the selected allele will tend to have levels of diversity and LD that are more typical of the genome as a whole. By comparing the rate of haplotype breakdown between one haplotype at a locus and the others at the same locus, it is possible to detect recent positive selective sweeps where the selected alleles have not yet reached fixation.

However, Reed and Tishkoff (2006) show that allele-specific recombination hotspots can

leave similar LD patterns to those just described, resulting in false positive signals. Specifically, an allele-specific recombination hotspot could result in reduced haplotype sharing for those chromosomes with that allele, which might then be interpreted as evidence for positive selection on the alternative allele.

If the true ARG for the data were known, it would be possible to distinguish between selective sweeps and allele-specific recombination hotspots. Three measures of an allele are encoded in the ARG: its age, its frequency and the number of recombination events occurring around it. A selective sweep would show up as a young allele with high frequency, while an allele-specific recombination hotspot would show up as an increased intensity of recombination events under the mutation.

Of course, the true ARG is unknown, and my current ARG inference algorithm uses haplotype homozygosity to order recombination and coalescence events, and hence is unlikely to powerfully distinguish between allele-specific recombination hotspots and selective sweeps.

Nevertheless, below I show that as it currently stands, the algorithm detects a signal indicative of positive selection at the Lactase gene, a known example of strong positive selection. This is done by comparing the frequencies of SNPs to their ages, where the relative ages of alleles are estimated using ARGs.

In most human populations, except those of European descent, the ability to digest lactose contained in milk disappears in childhood. Strong signals of selection have been identified in the Lactase gene (LCT) in European populations (Bersaglieri et al., 2004), agreeing with the hypothesis that selective advantage was gained by those with the ability to digest lactose as adults. Indeed, the selective signal at LCT is one of the strongest in the whole genome (The International HapMap Consortium, 2005; Voight et al., 2006).

In order to test the preliminary approach, I applied it to two approximately 1 Mb regions: 232 SNPs around the LCT gene on Chromosome 2 from the Phase 1 CEU HapMap, and a randomly selected region on Chromosome 2, again of 232 SNPs from the Phase 1 CEU HapMap. I then inferred 100 ARGs for the regions, and for each SNP calculated its frequency count (the number of the 120 CEU independent haplotypes possessing that allele) divided by its average age order (age 1 corresponds to the most recent SNP to be mutated in the ARG,

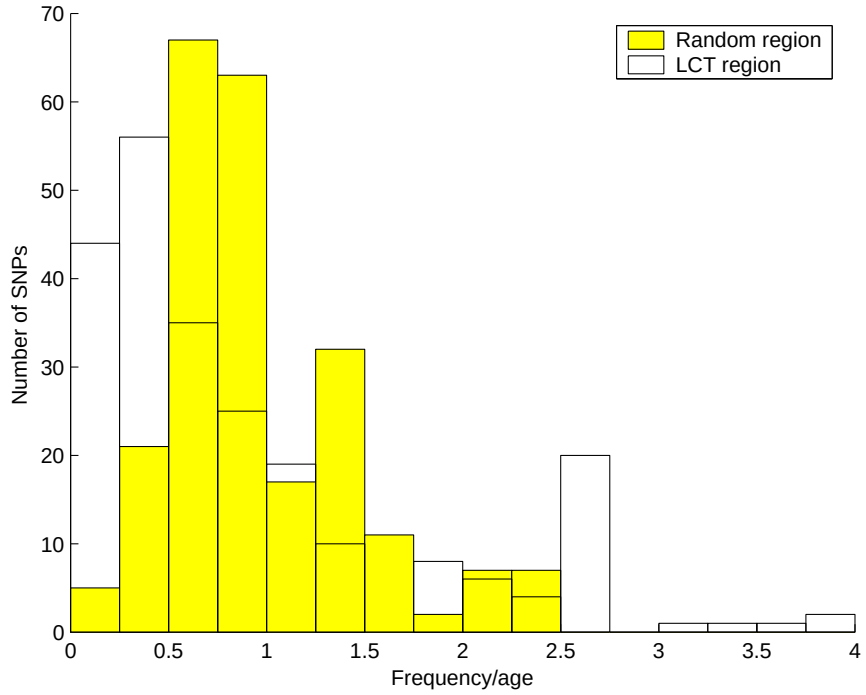


Figure 7.1: 232 SNPs in two 1 Mb regions (yellow = random region in Chromosome 2; clear = LCT region), their frequency count divided by their age ordering (averaged over 100 ARGs) in 60 unrelated CEU individuals.

and working backwards in time, age 232 corresponds to the SNP mutated highest up in the ARG).

Figure 7.1 shows the frequency/age distributions for these two regions. There is an excess of SNPs with high frequency/age in the LCT region, indicative of a positive selective sweep.

This analysis could be conducted for the whole genome, and then regions with SNPs in the tail of the frequency/age distribution could be considered as candidates for selection.

The comparison of frequency against age is what is required for identification of positive selection. In extended haplotype homozygosity tests, haplotype sharing is used as a proxy for age, as haplotype sharing decays with recombination over time. Hence, it may be expected that the direct comparison of frequency and age described above will have at least as much power to detect selection. However, it should be noted that the ARGs are inferred on the basis of shared segments, and so the differences between extended haplotype homozygosity and this approach may not be so marked.

It may be a good idea to not only consider the frequency/age distribution of typed SNPs,

but also of all branches of the inferred ARGs. This may have greater power to detect selection events when the alleles under selection are not typed, similar to Chapter 3 with fine mapping.

7.2 Sequence Imputation

It may soon be routine to resequence individuals for the purposes of association studies and other population genetic analyses (Balding, 2005; Romeo et al., 2007). However, full resequencing of many individuals is redundant because of LD structure, and while low coverage resequencing results in missing data and errors, it should be possible to impute missing genotypes and correct errors using an enhanced MARGARITA system. This could reduce the amount of genotyping effort required per individual, allowing more individuals from a greater number of populations to be sampled.

Imputing missing genotype data has been explored in Chapter 6. However, imputing sequence data is a slightly different problem for the following reasons:

- Resequencing data consists of nucleotides with quality scores attached to them, quantifying how accurate they are;
- For an individual, there will be contiguous tracts (reads) of observed nucleotides, and similar tracts which are entirely missing;
- The data will not be biallelic;
- There will be additional complexities such as copy number polymorphisms and rearrangements, which can lead to alignment errors.

To deal with these issues a sequence imputation system would need to incorporate sequence quality data, otherwise sequencing errors may mislead the ARG inference algorithm by giving evidence for recombination. One way to do this would be to adapt the shared segment calculation so that some mismatches are permitted when the quality score for a mismatch is low compared with surrounding evidence for a shared segment, suggesting that it is more likely to be an error than a genetic difference. The probability of a match or a mismatch is dependent on the quality scores of the two sequences at that position. So for each pair of

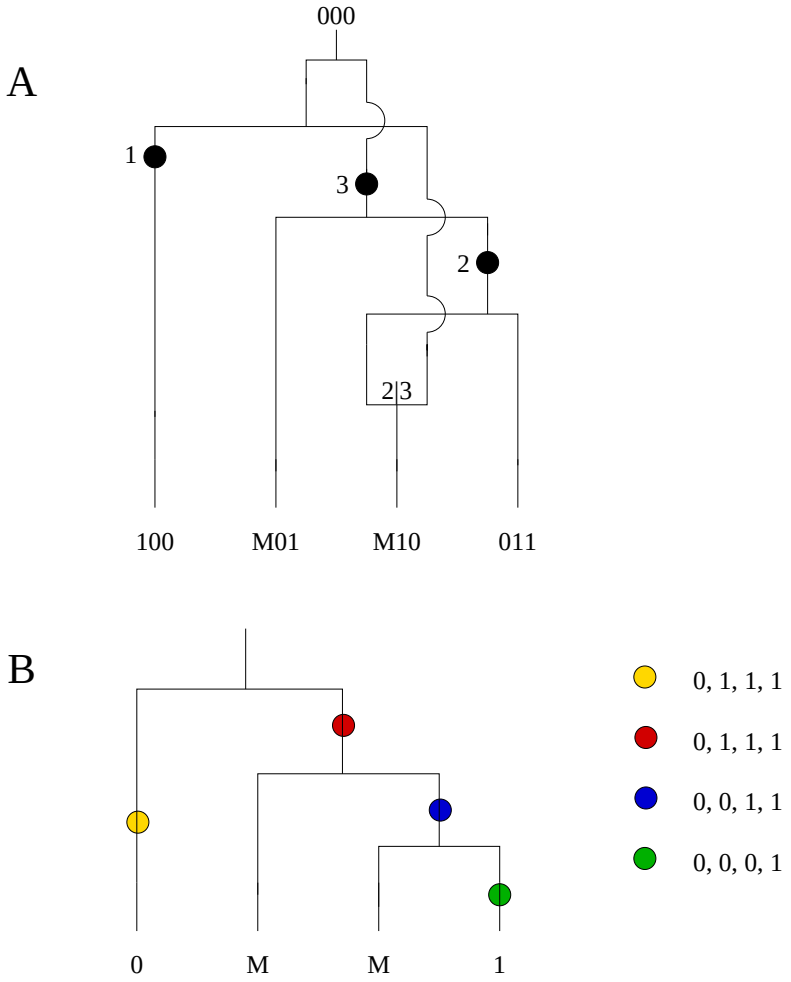


Figure 7.2: The positioning of the mutation affects the imputation.

sequences at a marker, a score could be calculated, such as the log odds of a match versus a mismatch. Maximal scoring shared segments could then be found (Ruzzo and Tompa, 1999) and used to guide the recombination and coalescence operations, as in the original algorithm.

When there are a large number of missing nucleotides, it may also be useful to enhance the way in which mutations are placed on the ARG. Figure 7.2 shows an example ARG where there are missing data. Given the ARG (Figure 7.2.A), both missing data points, denoted as Ms, would be imputed as 0s. However, considering the marginal tree (Figure 7.2.B) for the position with missing data, we see that there are four places where the mutation could be placed on the tree (denoted by the different coloured mutations), yielding four different imputations.

All four of these mutations fit legally into the ARG. The algorithm as described in Chapter 2 would give either the imputation represented by the yellow mutation, or the imputation represented by the red mutation. This is because a mutation is placed as early as possible, but after all missing data at that position is resolved. Nevertheless, the other two imputations are also valid, and it may be possible to increase imputation accuracy by considering them.

In collaboration with D. Carter at the Wellcome Trust Sanger Institute (Carter et al., 2007), the MARGARITA system has been extended. It now accommodates sequencing errors—early on in the ARG construction—by permitting mismatches in the shared segments. We also compute branch lengths for the ARG, and allow the implicit repositioning of mutations via use of the Felsenstein algorithm (Felsenstein, 1981) for calculating the probability of each nucleotide at the leaves.

In order to estimate branch lengths, we:

1. Throw out the explicit representation of mutation nodes in the ARG and then count the number of mutation events between coalescence and recombination events—or "nodes". This gives the number of mutation events on each edge of the ARG.

2. Calculate the active region for each edge of the ARG, that is the region of genetic material which is defined for that edge.

3. Estimate the ages of nodes in the ARG using a molecular clock assumption. This is done by maximising the likelihood of the number of mutations seen on each edge, subject to retaining the same global order of coalescence and recombination events. If an edge connects nodes with ages t_1 and t_2, and has an active region of length L, then the likelihood of it having k mutations is Poisson with mean $(t_2 - t_1)L$. An initial guess of the node ages is made from the coalescent-with-recombination, and ages are then updated by taking random horizontal "slices" through the ARG and allowing the ages of nodes above the slice to vary by a constant t, where t is chosen to maximise the joint likelihood of the mutation counts on all the edges cut by the slice. This update is performed several thousand times.

We then use the Felsenstein algorithm to assign a posterior probability to each nucleotide

on the leaf sequences, thereby correcting sequencing errors, as follows:

1. The marginal tree is extracted for each genomic interval between recombinations in the ARG.

2. We then use the Felsenstein algorithm (Felsenstein, 1981) to calculate the probability of a particular nucleotide at each locus and leaf, given the tree and branch lengths. Since this only uses the tree topology and lengths, it implicitly integrates over all permitted repositionings of mutations, as suggested in Figure 7.2. (In fact, it allows for multiple mutation events, relaxing the infinite sites assumption.)

3. The imputed nucleotide for a particular sequence and locus is the one with the highest posterior probability, averaged over multiple inferred ARGs; and the probability of error is the sum of the mean probabilities for the other possible nucleotide values.

We tested the quality of sequence imputation using *S. cerevisiae* resequencing data. Sanger shotgun sequencing was undertaken on 37 haploid strains at a coverage depth of 0.7 to 4.1, with mean depth of 1.64. We held out at random 10% of the read pairs and imputed their values using the above imputation procedure and the remaining data. Because the data is of varying quality, we only compared the imputed values to the assayed values with high quality scores. The system gives a mean error rate of 0.00126 on this data at polymorphic sites (sites which are monomorphic are trivial to impute, although some sites may appear monomorphic at low coverage which are in fact polymorphic).

We also evaluated this approach on simulated resequencing data, derived from the FREGENE population of Hoggart et al. (2005) used in Chapter 3, and matched to the sequencing characteristics of the *S. cerevisiae* data: read length, coverage and error probability. The results are shown in Table 7.1.

If high quality sequence is considered to have no more than 100 errors per million, then 7.1 gives us some idea of how this can be achieved with low coverage, imputation and error correction. The sequencing capacity which this frees can then be applied to resequencing more individuals, which has the important benefit of allowing more complete SNP discovery.

	Number of haplotypes					
Coverage	6	12	24	50	100	200
0.5	519	403	243	139	85	59
1.0	286	177	109	62	37	26
2.0	110	73	37	24	15	12
3.0	58	38	22	14	9	7

Table 7.1: Number of errors per million nucleotides for simulated resequencing data.

7.3 Detecting Population Substructure

Another question of interest in population genetics is that of identifying population substructure within data (Pritchard et al., 2000), either for the purpose of making demographic inferences, or for correcting for population effects in case-control studies (Clayton et al., 2005).

For the *S. cerevisiae* data, it may be of interest to identify regions of the genome where strains cluster together; for example, where those used for baking cluster, indicating that those parts of the genome may have been selected for. In Figure 7.3, I construct an ARG for 38 strains of *S. cerevisiae*, sequenced for Chromosome 1. In order to calculate the distance between strains, I take the average tree traversal distance between strains. For a particular marginal tree, the tree traversal distance between two strains is the number of coalescence events that must be traversed when travelling the path from the leaf corresponding to one of the strains to the other, divided by the maximum traversal distance, which is $n - 1$, where n is the number of sequences. This is averaged over all the polymorphic sites for Chromosome 1. The next step would be to analyse the clustering locally for patterns indicative of selection.

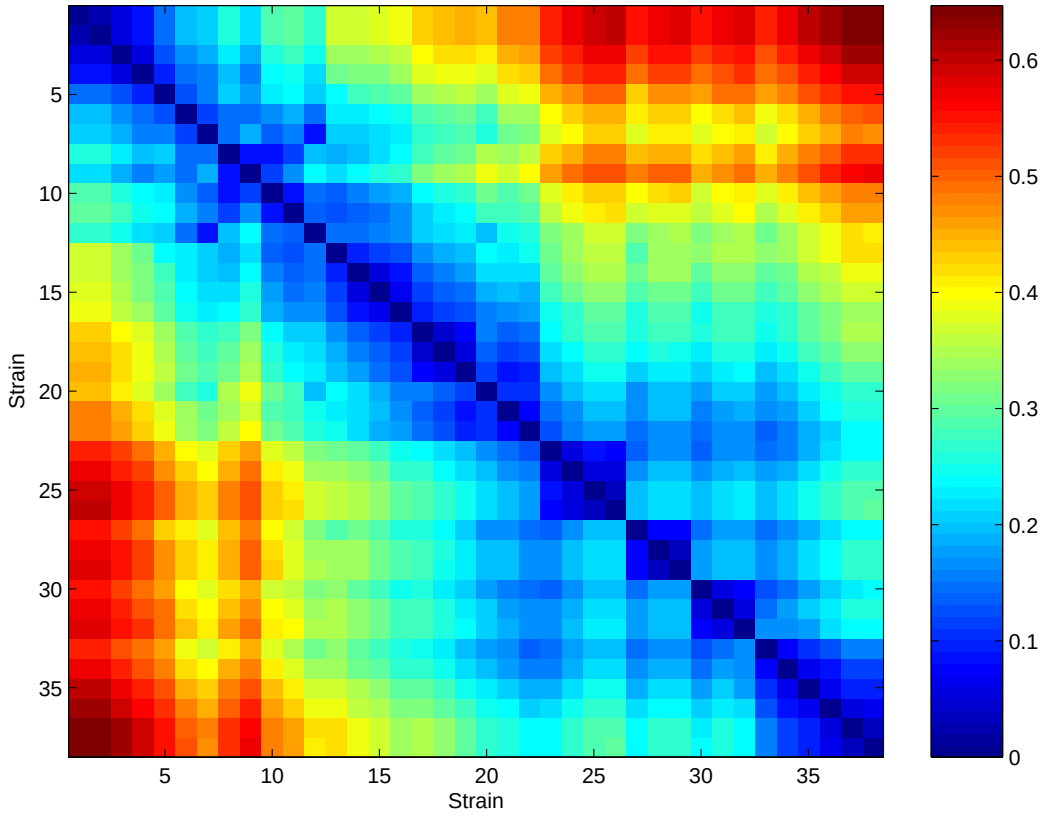


Figure 7.3: Pairwise tree traversal distances for 38 *S. cerevisiae* strains.

Chapter 8

Conclusions

In this thesis I have introduced a new method for constructing Ancestral Recombination Graphs from population genetic data. The ARG describes the history of a sample of sequences according to mutation, recombination and coalescence, and many analyses based on those features could be assisted by use of inferred ARGs. I have demonstrated this principally for disease mapping.

Analysing current case-control association data, involving thousands of individuals typed across the genome, requires an analysis method to be computationally efficient. This has been achieved by taking a heuristic approach to ARG inference—Margarita does not sample statistically from the coalescent-with-recombination, nor does it search for the ARG with the minimum number of obligate recombination events, but it does appear to capture some worthwhile features. This makes the method the first ARG mapping approach that can be applied to realistically sized data sets.

In order to detect disease associations, the local genealogies for loci, which are embedded in the ARG, are examined for a clustering of disease cases under a particular branch in the genealogy. Such a clustering suggests that a causative mutation may have occurred on that branch. In simulation studies, I found that this approach gives increased power to detect causative alleles after correcting for multiple testing by case-label permutation, and more accurate positioning of them, compared with the single marker chi-square test and a haplotype based method. Part of the increase in power after correcting for multiple testing is due to

the genealogies at nearby loci being more strongly correlated than r^2 linkage disequilibrium, meaning that nearby tests using my mapping method are more strongly correlated than tests with the chi-square. This results in a reduction in the multiple testing burden. This also gives more accurate positioning of causative loci; the decay in association with the ARG test is due to recombination events, whereas for the chi-square test, association is also affected by the relative timing of mutation events.

In addition to any increase in power and localisation, having an explicit estimate of the genealogy of a locus allows properties of untyped causative alleles to be inferred. On simulation studies, I showed that the inferred ARGs could be used successfully to infer the frequencies of untyped alleles.

In collaboration, I applied the ARG mapping method to two case-control studies. One of association between a 300kb region and Graves disease (Ueda et al., 2003), and the other of association between an 800kb region and prostate cancer (Yeager et al., 2007). In both cases, it was the additional interpretative power given by the ARG that proved to be the most useful aspect of the approach. For the Graves disease data, it was observed that there are multiple clusterings of disease individuals on the ARG, suggesting that those different clusters correspond to different causative alleles. This led to the identification of a weak signal of possible epistatic interaction. In the prostate cancer data the inferred ARGs helped show that the genealogies of two nearby association peaks are decorrelated due to a recombination hotspot, and thus correspond to independent signals. Analysis of the haplotypes extending around the association peaks showed a possible signal of selection.

Detecting disease association with the ARG approach is a missing data problem, where the untyped causative polymorphism is being imputed. In order to take a more direct approach, I considered using the ARG inference algorithm to impute missing values at known loci, specifically, loci which are untyped in the case-control data, but which are typed in a denser sample, such as the HapMap. The performance of the ARG approach was shown to be competitive with fastPHASE (Scheet and Stephens, 2006) as far as the quality of genotype imputation is concerned. However, it is not obvious from the experiments that this approach is useful in practice. There have been a number of signals detected in the 8q24 region showing

association with prostate cancer, and when I attempted to impute these using the Yeager et al. (2007) data combined with the CEU HapMap, the additional signals that had not previously been seen in the data set were not found.

However in the near future it may become routine to resequence individuals for disease studies (Balding, 2005; Romeo et al., 2007), in which case, any increases in power and localisation achieved by the ARG methods described here, over single marker tests, will be limited. When this situation arises, it is likely that the search for heterogeneity and epistatic interactions will begin in earnest, hence, the ARG approach will still be useful, as shown in my analysis of the CTLA4 data in Chapter 4. In fact, for that study (Ueda et al., 2003), the region was searched for new polymorphisms, and all of them were typed, hence it is arguable that this data is very similar to resequencing data; and still, the ARG approach was able to draw additional interesting inferences.

Indeed, as resequencing technology undergoes maturation, an important additional application of ARG inference could be to imputing missing nucleotides in resequencing data. Depending on the sequencing coverage there will be tracts of contiguous nucleotides which are observed, and other tracts that are completely unobserved. After appropriate modification, and coupled with some further enhancements, MARGARITA was applied to a population of *S. cerevisiae* sequences. The experiments show that resequencing can be performed at low coverage, and linkage disequilibrium relied upon in order to fill in the missing data. This is important because allowing resequencing to take place at lower coverage means that more individuals, from a wider range of populations, can be resequenced, allowing more complete SNP discovery.

Furthermore, many population genetic questions will still require sophisticated methods even when all nucleotides are assayed. As mentioned above, inferred ARGs may be useful in many analyses which rely on interpreting the recombination and mutation history. I gave some initial suggestions indicating how inferred ARGs could potentially be used to detect selective sweeps and population substructure.

Bibliography

J. M. Akey, G. Zhang, K. Zhang, L. Jin, and M. D. Shriver. Interrogating a high-density SNP map for signatures of natural selection. *Genome Res*, 12:1805–14, 2002.

L. T. Amundadottir, P. Sulem, J. Gudmundsson, A. Helgason, A. Baker, B. A. Agnarsson, A. Sigurdsson, K. R. Benediktsdottir, J. B. Cazier, J. Sainz, M. Jakobsdottir, J. Kostic, D. N. Magnusdottir, S. Ghosh, K. Agnarsson, B. Birgisdottir, L. L. Roux, A. Olafsdottir, T. Blondal, M. Andresdottir, O. S. Gretarsdottir, J. T. Bergthorsson, D. Gudbjartsson, A. Gylfason, G. Thorleifsson, A. Manolescu, K. Kristjansson, G. Geirsson, H. Isaksson, J. Douglas, J. E. Johansson, K. Balter, F. Wiklund, J. E. Montie, X. Yu, B. K. Suarez, C. Ober, K. A. Cooney, H. Gronberg, W. J. Catalona, G. V. Einarsson, R. B. Barkardottir, J. R. Gulcher, A. Kong, U. Thorsteinsdottir, and K. Stefansson. A common variant associated with prostate cancer in European and African populations. *Nat Genet*, 38:652–8, 2006.

S. E. Antonarakis and J. S. Beckmann. Mendelian disorders deserve more attention. *Nat Rev Genet*, 7:277–82, 2006.

V. Bafna and V. Bansal. Inference about recombination from haplotype data: lower bounds and recombination hotspots. *J Comput Biol*, 13:501–21, 2006.

D. Balding. The impact of low-cost and genome-wide resequencing on association studies. *Human Genomics*, 2:79–8, 2005.

D. J. Balding. A tutorial on statistical methods for population association studies. *Nat Rev Genet*, 7:781–91, 2006.

C. Bardel, V. Danjean, J. P. Hugot, P. Darlu, and E. Genin. On the use of haplotype phylogeny to detect disease susceptibility loci. *BMC Genet*, 6:24, 2005.

I. Barroso, J. Luan, E. Wheeler, P. Whittaker, J. Wasson, E. Zeggini, M. N. Weedon, S. Hunt, M. Delgado, R. Venkatesh, T. M. Frayling, R. J. Neuman, R. Sherva, B. Glaser, M. Walker, G. Hitman, M. I. McCarthy, A. T. Hattersley, M. A. Permutt, N. J. Wareham, and P. Deloukas. Population-specific risk of type 2 diabetes (T2D) conferred by HNF4A P2 promoter variants: a cautionary tale for genome-wide association studies. *Diabetes*, Submitted, 2007.

M. A. Beaumont. Detecting population expansion and decline using microsatellites. *Genetics*, 153:2013–29, 1999.

T. Bersaglieri, P. C. Sabeti, N. Patterson, T. Vanderploeg, S. F. Schaffner, J. A. Drake, M. Rhodes, D. E. Reich, and J. N. Hirschhorn. Genetic signatures of strong recent positive selection at the lactase gene. *Am J Hum Genet*, 74:1111–20, 2004.

L. R. Cardon and J. I. Bell. Association study designs for complex diseases. *Nat Rev Genet*, 2:91–9, 2001.

D. Carter, M. J. Minichiello, and R. Durbin. Imputing missing DNA values and discovering SNPs from low-coverage sequencing of multiple individuals. *In preparation*, 2007.

M. Chiano and D. Clayton. Fine genetic mapping using haplotype analysis and the missing data problem. *Ann Hum Genet*, 62:55–60, 1998.

G. A. Churchill and R. W. Doerge. Empirical threshold values for quantitative trait mapping. *Genetics*, 138:963–71, 1994.

A. G. Clark. Inference of haplotypes from PCR-amplified samples of diploid populations. *Mol Biol Evol.*, 7:111–22, 1990.

D. G. Clayton, N. M. Walker, D. J. Smyth, R. Pask, J. D. Cooper, L. M. Maier, L. J. Smink, A. C. Lam, N. R. Ovington, H. E. Stevens, S. Nutland, J. M. Howson, M. Faham, M. Moorhead, H. B. Jones, M. Falkowski, P. Hardenbol, T. D. Willis, and J. A. Todd.

Population structure and differential bias and genomic control in a large-scale and case-control association study. *Nat Genet*, 37:1243–6, 2005.

H. J. Cordell. Epistasis: what it means, what it doesn't mean and statistical methods to detect it in humans. *Hum Mol Genet*, 11:2463–8, 2002.

H. J. Cordell. Estimation and testing of genotype and haplotype effects in case/control studies: comparison of weighted regression and multiple imputation procedures. *Genet Epidemiol*, 30:259–75, 2006.

H. J. Cordell and D. G. Clayton. Genetic association studies. *Lancet*, 366:1121–31, 2005.

D. C. Crawford, T. Bhangale, N. Li, G. Hellenthal, M. J. Rieder, D. A. Nickerson, and M. Stephens. Evidence for substantial fine-scale variation in recombination rates across the human genome. *Nat Genet*, 36:700–6, 2004.

E. D. Crawford. Epidemiology of prostate cancer. *Urology*, 22:3–12, 2003.

J. Y. Dai, I. Ruczinski, M. LeBlanc, and C. Kooperberg. Imputation methods to improve inference in SNP association studies. *Genet Epidemiol*, 30:690–702, 2006.

M. J. Daly, J. D. Rioux, S. F. Schaffner, T. J. Hudson, and E. S. Lander. High-resolution haplotype structure in the human genome. *Nat Genet*, 29:229–32, 2001.

Z. Dawy, B. Goebel, J. Hagenauer, C. Andreoli, T. Meitinger, and J. C. Mueller. Gene mapping and marker clustering using Shannon's mutual information. *IEEE/ACM Trans Comput Biol Bioinform*, 3:47–56, 2006.

P. I. de Bakker, R. Yelensky, I. Pe'er, S. B. Gabriel, M. J. Daly, and D. Altshuler. Efficiency and power in genetic association studies. *Nat Genet*, 37:1217–23, 2005.

B. Devlin and N. Risch. A comparison of linkage disequilibrium measures for fine-scale mapping. *Genomics*, 29:311–22, 1995.

F. Dudbridge and B. P. Koeleman. Efficient computation of significance levels for multiple associations in large studies of correlated data and including genomewide association studies. *Am J Hum Genet*, 75:424–35, 2004.

C. Durrant, K. T. Zondervan, L. R. Cardon, S. Hunt, P. Deloukas, and A. P. Morris. Linkage disequilibrium mapping via cladistic analysis of single-nucleotide polymorphism haplotypes. *Am J Hum Genet*, 75:35–43, 2004.

A. O. Edwards, R. Ritter, K. J. Abel, A. Manning, C. Panhuysen, and L. A. Farrer. Complement factor H polymorphism and age-related macular degeneration. *Science*, 308:421–4, 2005.

D. M. Evans, L. R. Cardon, and A. P. Morris. Genotype prediction using a dense map of SNPs. *Genet Epidemiol*, 27:375–84, 2004.

L. Excoffier and M. Slakin. Maximum likelihood estimation of molecular haplotype frequencies in a diploid population. *Mol Biol Evol*, 12:921–7, 1995.

S. Eyheramendy, J. Marchini, G. McVean, S. Myers, and P. Donnelly. A model-based approach to capture genetic variation for future association studies. *Genome Res*, 17:88–95, 2007.

P. Fearnhead. SequenceLDhot: detecting recombination hotspots. *Bioinformatics*, 22:3061–6, 2006.

P. Fearnhead and P. Donnelly. Estimating recombination rates from population genetic data. *Genetics*, 159:1299–318, 2001.

P. Fearnhead, R. M. Harding, J. A. Schneider, S. Myers, and P. Donnelly. Application of coalescent methods to reveal fine-scale rate variation and recombination hotspots. *Genetics*, 167:2067–81, 2004.

J. Felsenstein. Evolutionary trees from DNA sequences: a maximum likelihood approach. *J Mol Evol*, 17:368–76, 1981.

J. Felsenstein. Confidence limits on phylogenies: an approach using the bootstrap. *Evolution*, 39:783–91, 1985.

M. L. Freedman, C. A. Haiman, N. Patterson, G. J. McDonald, A. Tandon, A. Waliszewska, K. Penney, R. G. Steen, K. Ardlie, E. M. John, I. Oakley-Girvan, A. S. Whittemore, K. A. Cooney, S. A. Ingles, D. Altshuler, B. E. Henderson, and D. Reich. Admixture mapping identifies 8q24 as a prostate cancer risk locus in African-American men. *Proc Natl Acad Sci U S A*, 103:14068–73, 2006.

S. B. Gabriel, S. F. Schaffner, H. Nguyen, J. M. Moore, J. Roy, B. Blumenstiel, J. Higgins, M. DeFelice, A. Lochner, M. Faggart, S. N. Liu-Cordero, C. Rotimi, A. Adeyemo, R. Cooper, R. Ward, E. S. Lander, M. J. Daly, and D. Altshuler. The structure of haplotype blocks in the human genome. *Science*, 296:2225–9, 2002.

D. B. Goldstein, K. R. Ahmadi, M. E. Weale, and N. W. Wood. Genome scans and candidate gene approaches in the study of common diseases and variable drug responses. *Trends Genet*, 19:615–22, 2003.

J. Graham and E. A. Thompson. Disequilibrium likelihoods for fine-scale mapping of a rare allele. *Am J Hum Genet*, 63:1517–30, 1998.

R. C. Griffiths and P. Marjoram. Ancestral inference from samples of DNA sequences with recombination. *J Comput Biol*, 3:479–502, 1996.

R. C. Griffiths and P. Marjoram. An ancestral recombination graph. *In: Donnelly P and Tavaré S (eds) and Progress in Population Genetics and Human Evolution and Springer Verlag*, pages 257–70, 1997.

R. C. Griffiths and S. Tavaré. Ancestral inference in population genetics. *Statistical Science*, 9:307–19, 1994a.

R. C. Griffiths and S. Tavaré. Sampling theory for neutral alleles in a varying environment. *Phil Trans R Soc Lond B*, 344:403–10, 1994b.

R. C. Griffiths and S. Tavaré. Simulating probability distributions in the coalescent. *Theor Popn Biol*, 46:131–59, 1994c.

J. Gudmundsson, P. Sulem, A. Manolescu, L. T. Amundadottir, D. Gudbjartsson, A. Helgason, T. Rafnar, J. T. Bergthorsson, B. A. Agnarsson, A. Baker, A. Sigurdsson, K. R. Benediktsdottir, M. Jakobsdottir, J. Xu, T. Blondal, J. Kostic, J. Sun, S. Ghosh, S. N. Stacey, M. Mouy, J. Saemundsdottir, V. M. Backman, K. Kristjansson, A. Tres, A. W. Partin, M. T. Albers-Akkers, J. G.-I. Marcos, P. C. Walsh, D. W. Swinkels, S. Navarrete, S. D. Isaacs, K. K. Aben, T. Graif, J. Cashy, M. Ruiz-Echarri, K. E. Wiley, B. K. Suarez, J. A. Witjes, M. Frigge, C. Ober, E. Jonsson, G. V. Einarsson, J. I. Mayordomo, L. A. Kiemeney, W. B. Isaacs, W. J. Catalona, R. B. Barkardottir, J. R. Gulcher, U. Thorsteinsdottir, A. Kong, and K. Stefansson. Genome-wide association study identifies a second prostate cancer susceptibility variant at 8q24. *Nat Genet*, 39:631–7, 2007.

K. L. Gunderson, F. J. Steemers, G. Lee, L. G. Mendoza, and M. S. Chee. A genome-wide scalable SNP genotyping assay using microarray technology. *Nat Genet*, 37:549–54, 2005.

D. Gusfield. Haplotyping as perfect phylogeny: conceptual framework and efficient solutions (extended abstract). *Proceedings of the 6th International Conference on Computational Molecular Biology*, page 166, 2002.

D. Gusfield, D. Hickerson, and S. Eddhu. An efficiently-computed lower bound on the number of recombinations in phylogenetic networks: Theory and empirical study. *Discrete Applied Mathematics*, 155:806–30, 2007.

C. A. Haiman, N. Patterson, M. L. Freedman, S. R. Myers, M. C. Pike, A. Waliszewska, J. Neubauer, A. Tandon, C. Schirmer, G. J. McDonald, S. C. Greenway, D. O. Stram, L. L. Marchand, L. N. Kolonel, M. Frasco, D. Wong, L. C. Pooler, K. Ardlie, I. Oakley-Girvan, A. S. Whittemore, K. A. Cooney, E. M. John, S. A. Ingles, D. Altshuler, B. E. Henderson, and D. Reich. Multiple regions within 8q24 independently affect risk for prostate cancer. *Nat Genet*, 39:638–44, 2007.

J. L. Haines, M. A. Hauser, S. Schmidt, W. K. Scott, L. M. Olson, P. Gallins, K. L. Spencer, S. Y. Kwan, M. Noureddine, J. R. Gilbert, N. Schnetz-Boutaud, A. Agarwal, E. A. Postel,

and M. A. Pericak-Vance. Complement factor H variant increases the risk of age-related macular degeneration. *Science*, 308:419–21, 2005.

E. Halperin and E. Eskin. Haplotype reconstruction from genotype data using imperfect phylogeny. *Bioinformatics*, 20:1842–9, 2004.

M. T. Hamblin, E. E. Thompson, and A. D. Rienzo. Complex signatures of natural selection at the Duffy blood group locus. *Am J Hum Genet*, 70:369–83, 2002.

N. A. Hanchard, K. A. Rockett, C. Spencer, G. Coop, M. Pinder, M. Jallow, M. Kimber, G. McVean, R. Mott, and D. P. Kwiatkowski. Screening for recently selected alleles by analysis of human haplotype similarity. *Am J Hum Genet*, 78:153–9, 2006.

D. A. Hinds, L. L. Stuve, G. B. Nilsen, E. Halperin, E. Eskin, D. G. Ballinger, K. A. Frazer, and D. R. Cox. Whole-genome patterns of common DNA variation in three human populations. *Science*, 18:1072–9, 2005.

J. N. Hirschhorn and M. J. Daly. Genome-wide association studies for common diseases and complex traits. *Nat Rev Genet*, 6:95–108, 2005.

J. N. Hirschhorn, K. Lohmueller, E. Byrne, and K. Hirschhorn. A comprehensive review of genetic association studies. *Genet Med*, 4:45–61, 2002.

C. J. Hoggart, T. Clark, R. Lampariello, M. D. Iorio, J. Whittaker, and D. Balding. FREGENE: Software for simulating large genomic regions. *Technical Report and Department of Epidemiology and Public Health and Imperial College and University of London*, 2005.

R. R. Hudson. Properties of a neutral allele model with intragenic recombination. *Theor Popul Biol*, 23:183–201, 1983.

R. R. Hudson. Two-locus sampling distributions and their application. *Genetics*, 159:1805–17, 2001.

R. R. Hudson. Generating samples under a wright-fisher neutral model of genetic variation. *Bioinformatics*, 18:337–8, 2002.

R. R. Hudson and N. L. Kaplan. Statistical properties of the number of recombination events in the history of a sample of DNA sequences. *Genetics*, 111:147–64, 1985.

J. P. Ioannidis, T. A. Trikalinos, E. E. Ntzani, and D. G. Contopoulos-Ioannidis. Genetic associations in large versus small studies: an empirical assessment. *Lancet*, 361:567–71, 2003.

A. J. Jeffreys, L. Kauppi, and R. Neumann. Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. *Nat Genet*, 29:217–22, 2001.

G. C. Johnson, L. Esposito, B. J. Barratt, A. N. Smith, J. Heward, G. D. Genova, H. Ueda, H. J. Cordell, I. A. Eaves, F. Dudbridge, R. C. Twells, F. Payne, W. Hughes, S. Nutland, H. Stevens, P. Carr, E. Tuomilehto-Wolf, J. Tuomilehto, S. C. Gough, D. G. Clayton, and J. A. Todd. Haplotype tagging for the identification of common disease genes. *Nat Genet*, 29:233–7, 2001.

J. F. C. Kingman. On the genealogy of large populations. *J Appl Prob*, 19A:27–43, 1982.

R. J. Klein, C. Zeiss, E. Y. Chew, J. Y. Tsai, R. S. Sackler, C. Haynes, A. K. Henning, J. P. SanGiovanni, S. M. Mane, S. T. Mayne, M. B. Bracken, F. L. Ferris, J. Ott, C. Barnstable, and J. Hoh. Complement factor H polymorphism in age-related macular degeneration. *Science*, 308:385–9, 2005.

L. Kruglyak. Prospects for whole-genome linkage disequilibrium mapping of common disease genes. *Nat Genet*, 22:139–44, 1999.

L. Kruglyak and D. A. Nickerson. Variation is the spice of life. *Nat Genet*, 27:234–6, 2001.

M. K. Kuhner, J. Yamato, and J. Felsenstein. Maximum likelihood estimation of recombination rates from population data. *Genetics*, 156:1393–401, 2000.

F. Larribe, S. Lessard, and N. J. Schork. Gene mapping via the ancestral recombination graph. *Theor Popul Biol*, 62:215–29, 2002.

S. Lesage and C. C. Goodnow. Organ-specific autoimmune disease: a deficiency of tolerogenic stimulation. *J Exp Med*, 194:31–36, 2001.

N. Li and M. Stephens. Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics*, 165:2213–33, 2003.

R. J. A. Little and D. B. Rubin. Statistical analysis with missing data. *John Wiley and Sons, Inc*, 1987.

K. E. Lohmueller, C. L. Pearce, M. Pike, E. S. Lander, and J. N. Hirschhorn. Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease. *Nat Genet*, 33:177–82, 2003.

J. Long, R. Williams, and M. Urbanek. An e-m algorithm and testing strategy for multiple-locus haplotypes. *Am J Hum Genet*, 56:799–810, 1995.

R. Lyngsø, Y. S. Song, and J. Hein. Minimum recombination histories by branch and bound. *Proceedings of Workshop on Algorithms in Bioinformatics, Lecture Notes in Computer Science*, 3692:239–50, 2005.

S. Macgregor and I. A. Khan. GAIA: An easy-to-use web-based application for interaction analysis of case-control data. *BMC Med Genet*, 7:34, 2006.

J. Marchini, D. Cutler, N. Patterson, M. Stephens, E. Eskin, E. Halperin, S. Lin, Z. S. Qin, H. M. Munro, G. R. Abecasis, and P. Donnelly. A comparison of phasing algorithms for trios and unrelated individuals. *Am J Hum Genet*, 78:437–50, 2006.

J. Marchini, B. Howie, S. Myers, G. McVean, and P. D. P. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat Genet*, 39:906–13, 2007.

H. Matsuzaki, S. Dong, H. Loi, X. Di, G. Liu, E. Hubbell, J. Law, T. Berntsen, M. Chadha, H. Hui, G. Yang, G. C. Kennedy, T. A. Webster, S. Cawley, P. S. Walsh, K. W. Jones, S. P. Fodor, and R. Mei. Genotyping over 100,000 SNPs on a pair of oligonucleotide arrays. *Nat Methods*, 1:109–11, 2004.

G. A. McVean. A genealogical interpretation of linkage disequilibrium. *Genetics*, 162:987–91, 2002.

G. A. McVean and N. J. Cardin. Approximating the coalescent with recombination. *Philos Trans R Soc Lond B Biol Sci*, 360:1387–93, 2005.

G. A. T. McVean, P. Awadalla, and P. Fearnhead. A coalescent-based method for detecting and estimating recombination from gene sequences. *Genetics*, 160:1231–41, 2002.

G. A. T. McVean, S. Myers, S. Hunt, P. Deloukas, D. R. Bentley, and P. Donnelly. The fine-scale structure of recombination rate variation in the human genome. *Science*, 304: 581–584, 2004.

F. K. Mensah, M. S. Gilthorpe, C. F. Davies, L. J. Keen, P. J. Adamson, E. Roman, G. J. Morgan, J. L. Bidwell, and G. R. Law. Haplotype uncertainty in association studies. *Genet Epidemiol*, 31:348–57, 2007.

M. J. Minichiello and R. Durbin. Mapping trait loci by use of inferred ancestral recombination graphs. *Am J Hum Genet*, 79:910–22, 2006.

J. Molitor, P. Marjoram, and D. Thomas. Fine-scale mapping of disease genes with multiple mutations via spatial clustering techniques. *Am J Hum Genet*, 73:1368–84, 2003.

A. P. Morris, J. C. Whittaker, and D. J. Balding. Fine-scale mapping of disease loci via shattered coalescent modeling of genealogies. *Am J Hum Genet*, 70:686–707, 2002.

A. P. Morris, J. C. Whittaker, and D. J. Balding. Little loss of information due to unknown phase for fine-scale linkage-disequilibrium mapping with single-nucleotide-polymorphism genotype data. *Am J Hum Genet*, 74:945–53, 2004.

C. Myers, L. Bottolo, C. Freeman, G. McVean, and P. Donnelly. A fine-scale map of recombination rates and hotspots across the human genome. *Science*, 310:321–4, 2005.

S. R. Myers and R. C. Griffiths. Bounds on the minimum number of recombination events in a sample history. *Genetics*, 163:375–94, 2003.

S. A. Narod, A. Dupont, L. Cusan, P. Diamond, J.-L. Gomez, R. Suburu, and F. Labrie. The impact of family history on early detection of prostate cancer. *Nat Med*, 1:99–101, 1995.

Nature Genetics Editorial. Freely associating. *Nat Genet*, 22:1–2, 1999.

D. L. Nicolae. Testing untyped alleles (TUNA)-applications to genome-wide association studies. *Genet Epidemiol*, 30:718–27, 2006.

R. Nielsen. Estimation of population parameters and recombination rates from single nucleotide polymorphisms. *Genetics*, 154:931–42, 2000.

R. Nielsen, S. Williamson, Y. Kim, M. J. Hubisz, A. G. Clark, and C. Bustamante. Genomic scans for selective sweeps using SNP data. *Genome Res*, 15:1566–75, 2005.

M. Nordborg and S. Tavaré. Linkage disequilibrium: what history has to tell us. *Trends Genet*, 18:83–90, 2002.

L. J. Palmer and L. R. Cardon. Shaking the tree: mapping complex disease genes with linkage disequilibrium. *Lancet*, 366:1223–34, 2005.

P. Paschou, M. W. Mahoney, A. Javed, J. R. Kidd, A. J. Pakstis, S. Gu, K. K. Kidd, and P. Drineas. Intra- and interpopulation genotype reconstruction from tagging SNPs. *Genome Res*, 17:96–107, 2007.

N. Patil, A. J. Berno, D. A. Hinds, W. A. Barrett, J. M. Doshi, C. R. Hacker, C. R. Kautzer, D. H. Lee, C. Marjoribanks, D. P. McDonough, B. T. Nguyen, M. C. Norris, J. B. Sheehan, N. Shen, D. Stern, R. P. Stokowski, D. J. Thomas, M. O. Trulson, K. R. Vyas, K. A. Frazer, S. P. Fodor, and D. R. Cox. Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. *Science*, 23:1719–23, 2001.

N. Patterson, N. Hattangadi, B. Lane, K. E. Lohmueller, D. A. Hafler, J. R. Oksenberg, S. L. Hauser, M. W. Smith, S. J. O'Brien, D. Altshuler, M. J. Daly, and D. Reich. Methods for high-density admixture mapping of disease genes. *Am J Hum Genet*, 74:979–1000, 2004.

A. L. Price, N. J. Patterson, R. M. Plenge, M. E. Weinblatt, N. A. Shadick, and D. Reich. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet*, 38:904–9, 2006.

J. K. Pritchard and M. Przeworski. Linkage disequilibrium in humans. *Am J Hum Genet*, 69:1–14, 2001.

J. K. Pritchard, M. Stephens, and P. Donnelly. Inference of population structure using multilocus genotype data. *Genetics*, 155:945–59, 2000.

S. Purcell, S. S. Cherny, and P. C. Sham. Genetic Power Calculator: design of linkage and association genetic mapping studies of complex traits. *Bioinformatics*, 19:149–50, 2003.

B. Rannala and J. P. Reeve. High-resolution multipoint linkage-disequilibrium mapping in the context of a human genome sequence. *Am J Hum Genet*, 69:159–78, 2001.

F. A. Reed and S. A. Tishkoff. Positive selection can create false hotspots of recombination. *Genetics*, 172:2011–4, 2006.

J. D. Rioux, M. J. Daly, M. S. Silverberg, K. Lindblad, H. Steinhart, Z. Cohen, T. Delmonte, K. Kocher, K. Miller, S. Guschwan, E. J. Kulbokas, S. O'Leary, E. Winchester, K. Dewar, T. Green, V. Stone, C. Chow, A. Cohen, D. Langelier, G. Lapointe, D. Gaudet, J. Faith, N. Branco, S. B. Bull, R. S. McLeod, A. M. Griffiths, A. Bitton, G. R. Greenberg, E. S. Lander, K. A. Siminovitch, and T. J. Hudson. Genetic variation in the 5q31 cytokine gene cluster confers susceptibility to Crohn disease. *Nat Genet*, 29:223–8, 2001.

N. Risch and K. Merikangas. The future of genetic studies of complex human diseases. *Science*, 13:1516–7, 1996.

N. J. Risch. Searching for genetic determinants in the new millennium. *Nature*, 405:847–56, 2000.

D. L. Rohde, S. Olson, and J. T. Chang. Modelling the recent common ancestry of all living humans. *Nature*, 431:562–6, 2004.

S. Romeo, L. A. Pennacchio, Y. Fu, E. Boerwinkle, A. Tybjaerg-Hansen, H. H. Hobbs, and J. C. Cohen. Population-based resequencing of ANGPTL4 uncovers variations that reduce triglycerides and increase HDL. *Nat Genet*, 39:513–16, 2007.

D. B. Rubin. Multiple imputation for nonresponse in surveys. *John Wiley and Sons*, 1987.

W. L. Ruzzo and M. Tompa. A linear time algorithm for finding all maximal scoring subsequences. *Proc Int Conf Intell Syst Mol Biol*, pages 234–41, 1999.

P. C. Sabeti, D. E. Reich, J. M. Higgins, H. Z. Levine, D. J. Richter, S. F. Schaffner, S. B. Gabriel, J. V. Platko, N. J. Patterson, G. J. McDonald, H. C. Ackerman, S. J. Campbell, D. Altshuler, R. Cooper, D. Kwiatkowski, R. Ward, and E. S. Lander. Detecting recent positive selection in the human genome from haplotype structure. *Nature*, 419:832–7, 2002.

R. Sachidanandam, D. Weissman, S. C. Schmidt, J. M. Kakol, L. D. Stein, G. Marth, S. Sherry, J. C. Mullikin, B. J. Mortimore, D. L. Willey, S. E. Hunt, C. G. Cole, P. C. Coggill, C. M. Rice, Z. Ning, J. Rogers, D. R. Bentley, P. Y. Kwok, E. R. Mardis, R. T. Yeh, B. Schultz, L. Cook, R. Davenport, M. Dante, L. Fulton, L. Hillier, R. H. Waterston, J. D. McPherson, B. Gilman, S. Schaffner, W. J. V. Etten, D. Reich, J. Higgins, M. J. Daly, B. Blumenstiel, J. Baldwin, N. Stange-Thomann, M. C. Zody, L. Linton, E. S. Lander, D. Altshuler, and I. S. M. W. Group. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature*, 409:928–33, 2001.

P. D. Sasieni. From genotypes to genes: Doubling the sample size. *Biometrics*, 53:1253–61, 1997.

D. J. Schaid and S. J. Jacobsen. Biased tests of association: comparisons of allele frequencies when departing from hardy-weinberg proportions. *Am J Epidemiol*, 149:706–11, 1999.

P. Scheet and M. Stephens. A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *Am J Hum Genet*, 78:629–44, 2006.

B. Servin and M. Stephens. Imputation-based analysis of association studies: candidate regions and quantitative traits. *PLoS Genetics*, preprint, 2007.

M. W. Smith, N. Patterson, J. A. Lautenberger, A. L. Truelove, G. J. McDonald, A. Waliszewska, B. D. Kessing, M. J. Malasky, C. Scafe, E. Le, P. L. D. Jager, A. A. Mignault,

Z. Yi, G. D. The, M. Essex, J. L. Sankale, J. H. Moore, K. Poku, J. P. Phair, J. J. Goedert, D. Vlahov, S. M. Williams, S. A. Tishkoff, C. A. Winkler, F. D. L. Vega, T. Woodage, J. J. Sninsky, D. A. Hafler, D. Altshuler, D. A. Gilbert, S. J. O'Brien, and D. Reich. A high-density admixture map for disease gene discovery in african americans. *Am J Hum Genet*, 74:1001–13, 2004.

Y. S. Song and J. Hein. On the minimum number of recombination events in the evolutionary history of DNA sequences. *J Math Biol*, 48:160–86, 2004.

Y. S. Song and J. Hein. Constructing minimal ancestral recombination graphs. *J Comput Biol*, 12:147–69, 2005.

Y. S. Song, R. Lyngsø, and J. Hein. Counting all possible ancestral configurations of sample sequences in population genetics. *IEEE/ACM Trans Comput Biol Bioinform*, 3:239–251, 2006.

O. W. Souverein, A. H. Zwinderman, and M. W. Tanck. Multiple imputation of missing genotype data for unrelated individuals. *Ann Hum Genet*, 70:372–81, 2006.

M. Stephens and P. Donnelly. A comparison of bayesian methods for haplotype reconstruction from population genotype data. *Am J Hum Genet*, 73:1162–69, 2003.

M. Stephens and P. Scheet. Accounting for decay of linkage disequilibrium in haplotype inference and missing-data imputation. *Am J Hum Genet*, 76:449–62, 2005.

M. Stephens, N. J. Smith, and P. Donnelly. A new statistical method for haplotype reconstruction from population data. *Am J Hum Genet*, 68:978–89, 2001.

S. Tavaré, D. J. Balding, R. C. Griffiths, and P. Donnelly. Inferring coalescence times from DNA sequence data. *Genetics*, 145:505–18, 1997.

A. R. Templeton. A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping or DNA sequencing. V. Analysis of case/control sampling designs: Alzheimer's disease and the apoprotein E locus. *Genetics*, 140:403–9, 1995.

A. R. Templeton, E. Boerwinkle, and C. F. Sing. A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping. I. Basic theory and an analysis of alcohol dehydrogenase activity in drosophila. *Genetics*, 117:343–51, 1987.

A. R. Templeton, K. A. Crandall, and C. F. Sing. A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping and DNA sequence data. III. Cladogram estimation. *Genetics*, 132:619–33, 1992.

A. R. Templeton, T. Maxwell, D. Posada, J. H. Stengard, E. Boerwinkle, and C. F. Sing. Tree scanning: a method for using haplotype trees in phenotype/genotype association studies. *Genetics*, 169:441–53, 2005.

A. R. Templeton and C. F. Sing. A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping. IV. Nested analyses with cladogram uncertainty and recombination. *Genetics*, 134:659–69, 1993.

A. R. Templeton, C. F. Sing, A. Kessling, and S. Humphries. A cladistic analysis of phenotype associations with haplotypes inferred from restriction endonuclease mapping. II. The analysis of natural populations. *Genetics*, 120:1145–54, 1988.

J. D. Terwilliger and T. Hiekkalinna. An utter refutation of the fundamental theorem of the HapMap. *Eur J Hum Genet*, 14:426–37, 2006.

The International HapMap Consortium. A haplotype map of the human genome. *Nature*, 437:1299–320, 2005.

The Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447:661–78, 2007.

H. Ueda, J. M. Howson, L. Esposito, J. Heward, H. Snook, G. Chamberlain, D. B. R. K. M. Hunter, A. N. Smith, G. D. Genova, M. H. Herr, I. Dahlman, F. Payne, D. Smyth, C. Lowe, R. C. Twells, S. Howlett, B. Healy, S. Nutland, H. E. Rance, V. Everett, L. J. Smink, A. C. Lam, H. J. Cordell, N. M. Walker, C. Bordin, J. Hulme, C. Motzo, F. Cucca, J. F. Hess,

M. L. Metzker, J. Rogers, S. Gregory, A. Allahabadia, R. Nithiyananthan, E. Tuomilehto-Wolf, J. Tuomilehto, P. Bingley, K. M. Gillespie, D. E. Undlien, K. S. Ronningen, C. Guja, C. Ionescu-Tirgoviste, D. A. Savage, A. P. Maxwell, D. J. Carson, C. C. Patterson, J. A. Franklyn, D. G. Clayton, L. B. Peterson, L. S. Wicker, J. A. Todd, and S. C. Gough. Association of the T-cell regulatory gene CTLA4 with susceptibility to autoimmune disease. *Nature*, 423:506–11, 2003.

B. Vaidya and S. Pearce. The emerging role of the CTLA-4 gene in autoimmune endocrinopathies. *Eur J Endocrinol*, 150:619–26, 2004.

B. F. Voight, S. Kudaravalli, X. Wen, and J. K. Pritchard. A map of recent positive selection in the human genome. *PLoS Biol*, 7:e72, 2006.

E. R. Waldron, J. C. Whittaker, and D. J. Balding. Fine mapping of disease genes via haplotype clustering. *Genet Epidemiol*, 30:170–9, 2006.

L. Wang, K. Zhang, and L. Zhang. Perfect phylogenetic networks with recombination. *J Comput Biol*, 8:69–78, 2001.

W. Y. Wang, B. J. Barratt, D. G. Clayton, and J. A. Todd. Genome-wide association studies: theoretical and practical concerns. *Nat Rev Genet*, 6:109–18, 2005.

I. J. Wilson and D. J. Balding. Genealogical inference from microsatellite data. *Genetics*, 150:499–510, 1998.

C. Wiuf. Inference on recombination and block structure using unphased data. *Genetics*, 166: 537–45, 2004.

S. Wright. Evolution in Mendelian populations. *Genetics*, 16:97–159, 1931.

M. Yeager, N. Orr, R. B. Hayes, K. B. Jacobs, P. Kraft, S. Wacholder, M. J. Minichiello, P. Fearnhead, K. Yu, N. Chatterjee, Z. Wang, R. Welch, B. J. Staats, E. E. Calle, H. S. Feigelson, M. J. Thun, C. Rodriguez, D. Albanes, J. Virtamo, S. Weinstein, F. R. Schumacher, E. Giovannucci, W. C. Willett, G. Cancel-Tassin, O. Cussenot, A. Valeri, G. L. Andriole, E. P. Gelmann, M. Tucker, D. S. Gerhard, J. R. Fraumeni, R. Hoover, D. J.

Hunter, S. J. Chanock, and G. Thomas. Genome-wide association study of prostate cancer identifies a second risk locus at 8q24. *Nat Genet*, 39:645–649, 2007.

K. Zhang, M. Deng, T. Chen, M. S. Waterman, and F. Sun. A dynamic programming algorithm for haplotype block partitioning. *Proc Natl Acad Sci U S A*, 28:7335–9, 2002.

S. Zollner and J. K. Pritchard. Coalescent-based association mapping and fine mapping of complex trait loci. *Genetics*, 169:1071–92, 2005.

G. Zsurka, Y. Kraytsberg, T. Kudina, C. Kornblum, C. E. Elger, K. Khrapko, and W. S. Kunz. Recombination of mitochondrial DNA in skeletal muscle of individuals with multiple mitochondrial DNA heteroplasmy. *Nat Genet*, 37:873–7, 2005.