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 GRR(Aa) = $P(D|Aa)/P(D|aa)$ and 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 GRR(AA) =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

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

(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

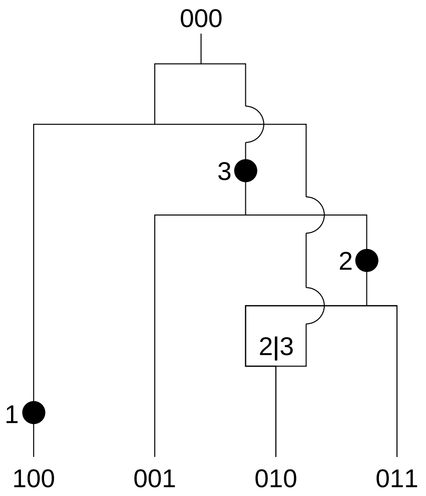

Figure 1.1: An Ancestral Recombination Graph

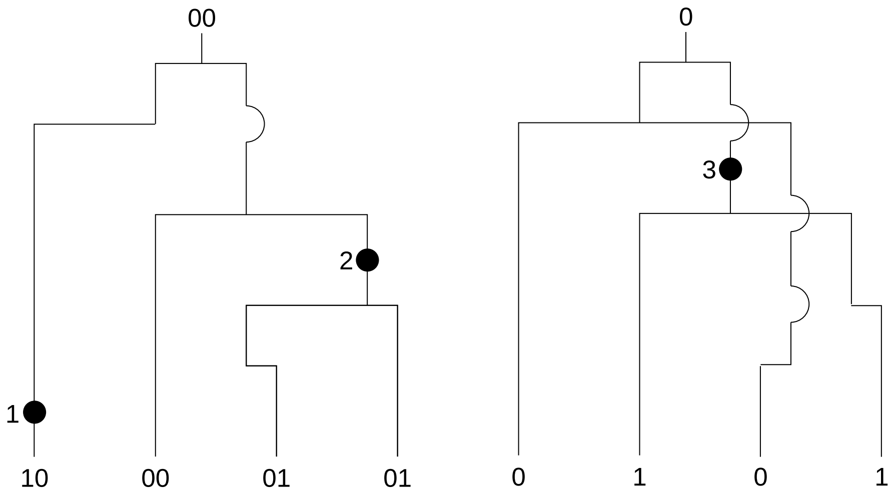

Figure 1.2: Marginal trees for the ARG in Figure 1.1.

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;

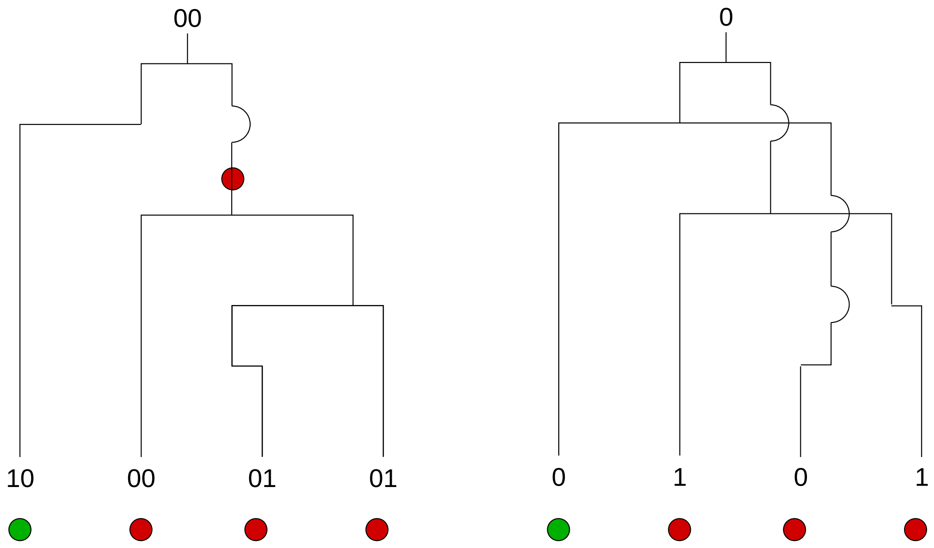

Figure 1.3: Disease mapping using the ARG. Suppose chromosomes denoted by a red dot are from disease individuals, then the branch with the red mutation shows the strongest clustering of cases beneath it.

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 $P\left(h_{k+1}|h_1, \ldots, h_k, \rho\right)$ 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.