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,

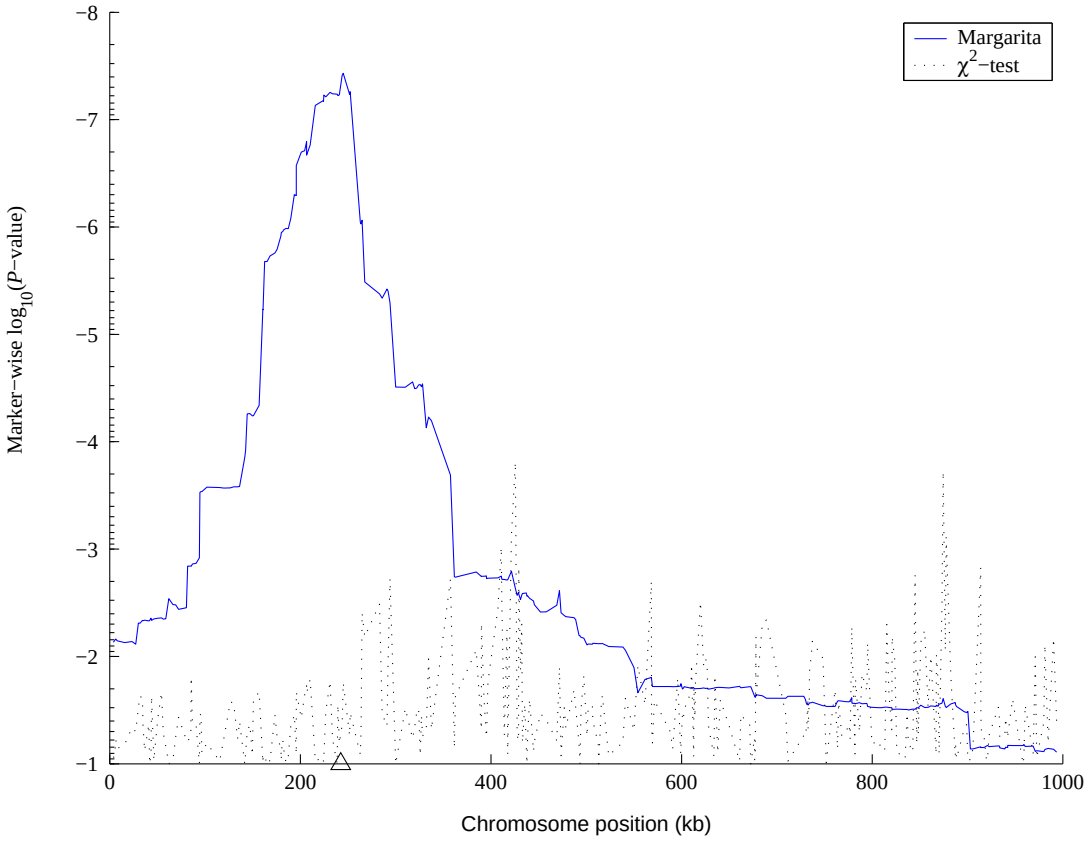

Figure 3.1: The association structure for a simulated case control study with disease parameters $GRR(Aa) = 2$, $GRR(AA) = 3$, $q = 0.04$ and $n_{cc} = 2000$, sampled from the "constant" population with the "full" ascertainment tag set. △ denotes the position of the (untyped) causative SNP.

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

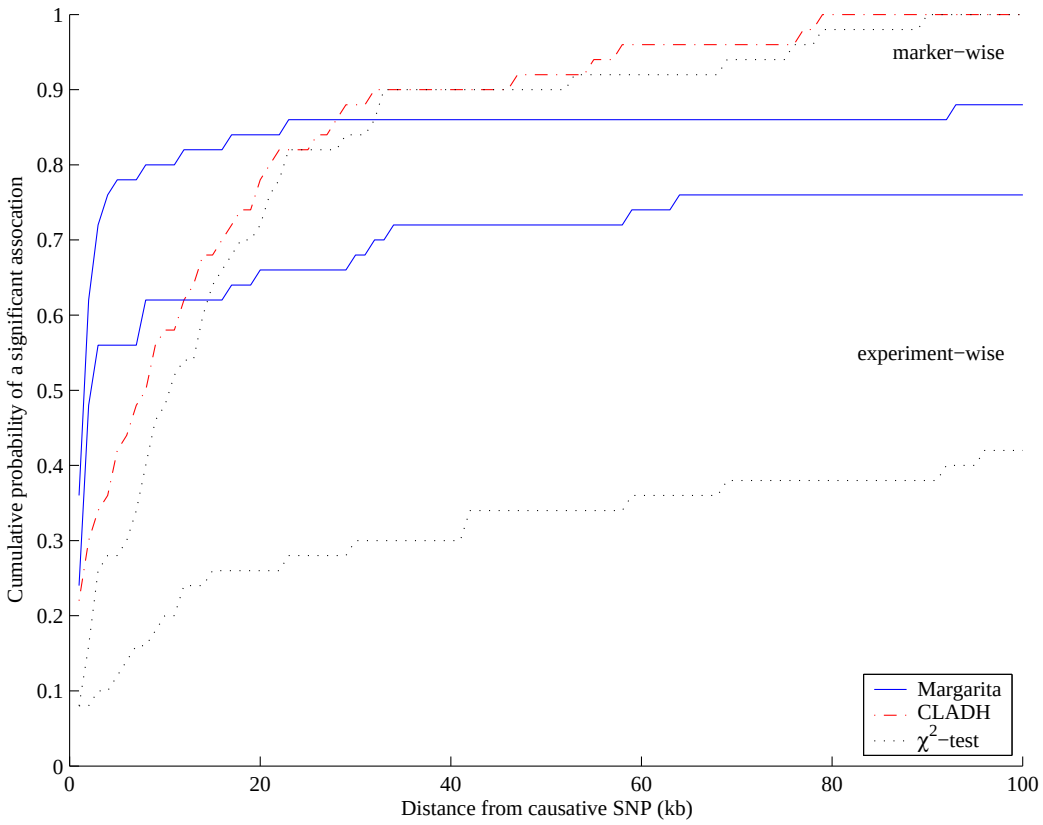

Figure 3.2: The probability of there being a significant association within an interval around the causative SNP. For a suite of case-control studies with disease parameters $GRR(Aa) = 2$, $GRR(AA) = 3$, $q = 0.04$ and $n_{cc} = 2000$, sampled from the "constant" population with the "full" ascertainment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

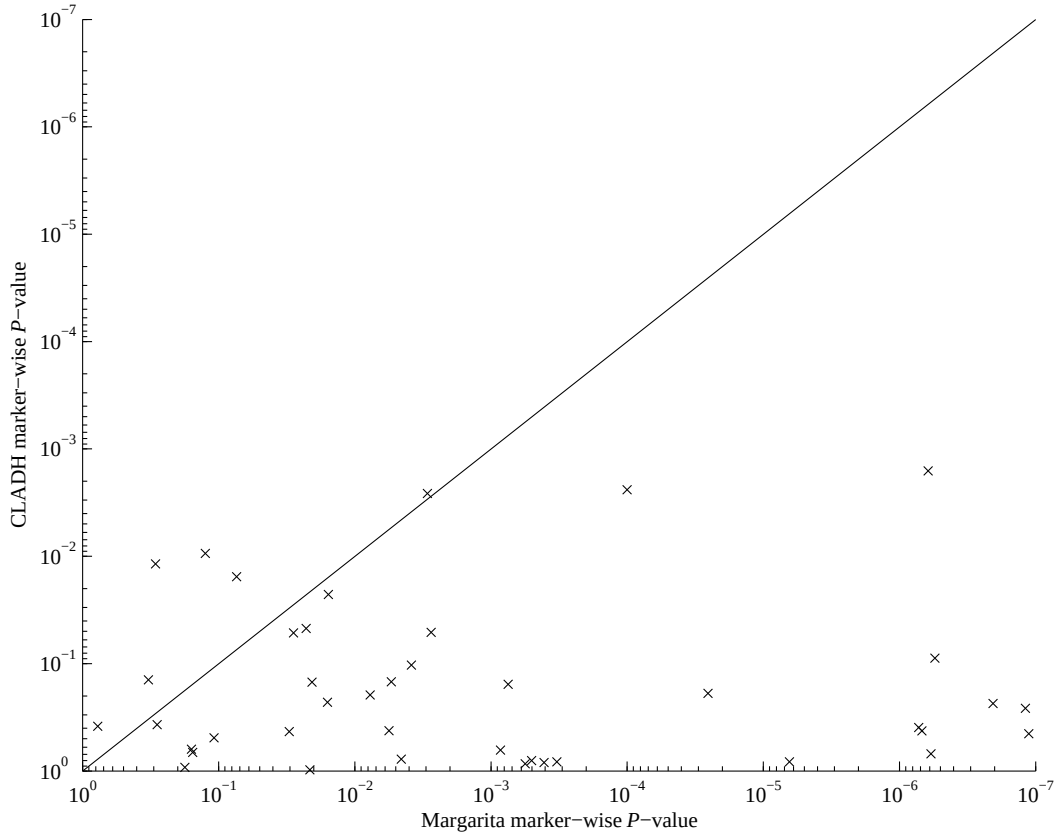

Figure 3.3: Marker-wise P-values at the marker closest to the causative SNP for 50 studies in a suite of case-control studies with disease parameters $GRR(Aa) = 2$, $GRR(AA) = 3$, $q = 0.04$ and $n_{cc} = 2000$, sampled from the "constant" population with the "full" ascertainment tag set.

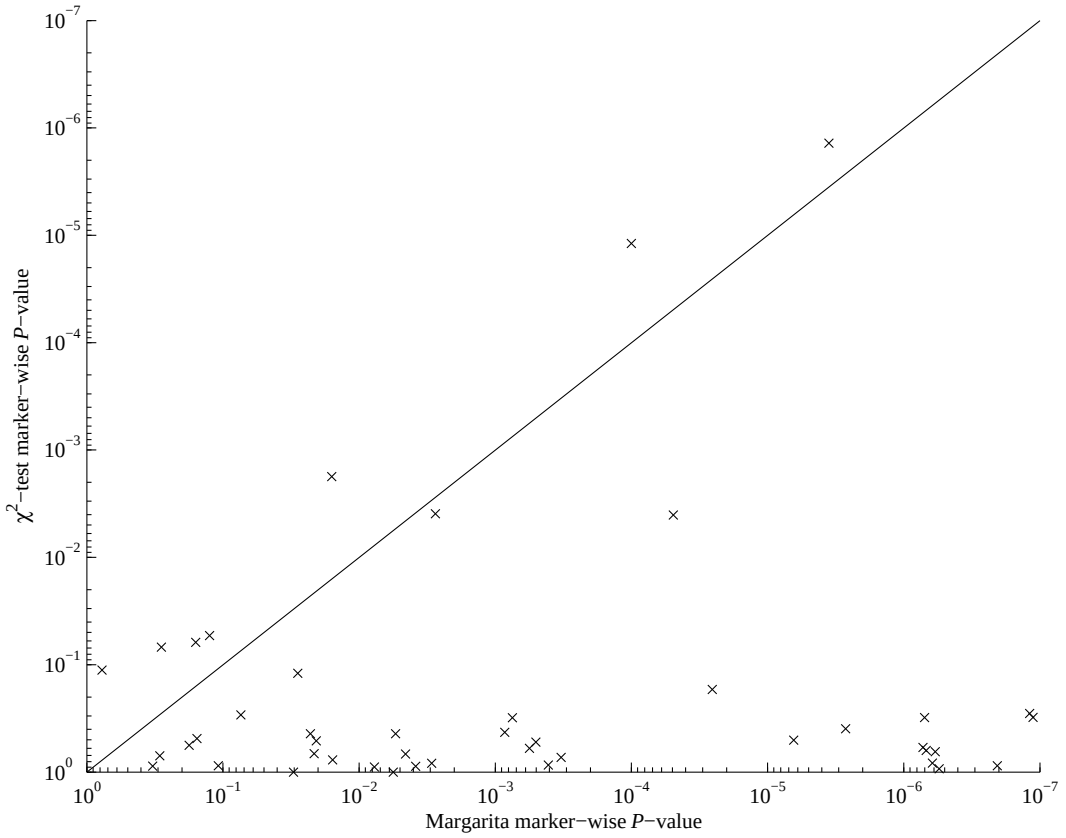

Figure 3.4: Marker-wise P-values at the marker closest to the causative SNP for 50 studies in a suite of case-control studies with disease parameters $GRR(Aa) = 2$, $GRR(AA) = 3$, $q = 0.04$ and $n_{cc} = 2000$, sampled from the "constant" population with the "full" ascertainment tag set.

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).

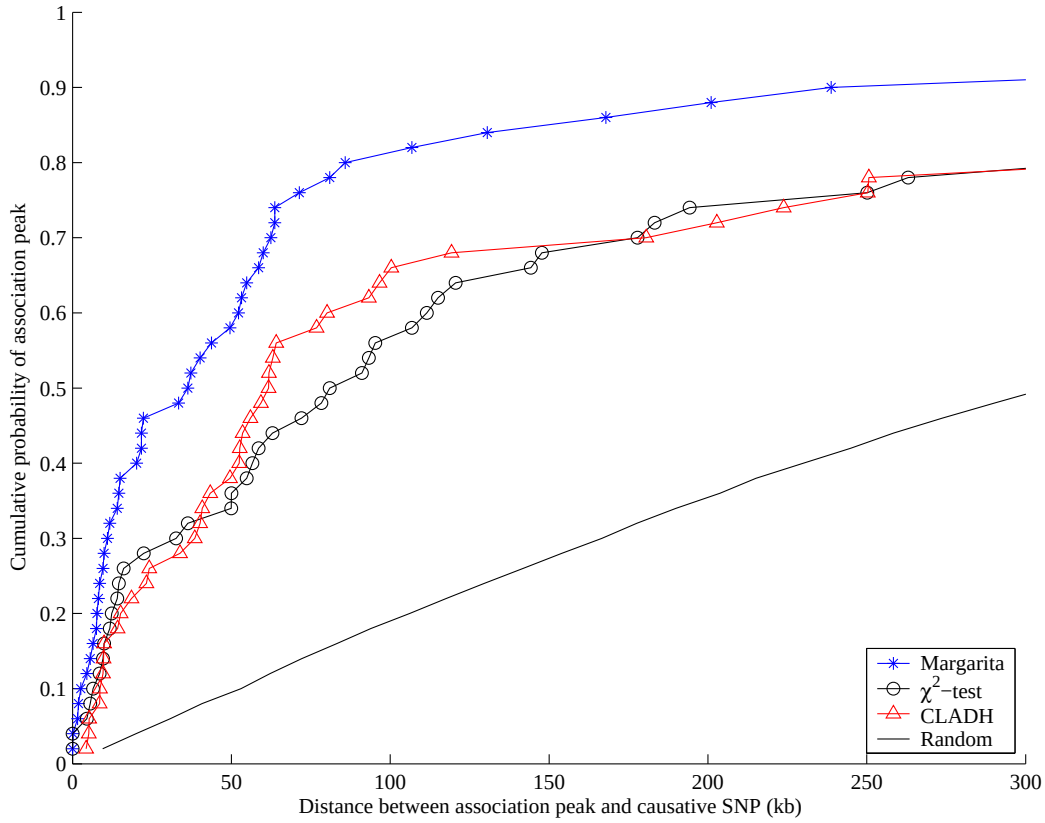

Figure 3.5: Cumulative distribution of distances between the association peak and the causative SNP. For a suite of case-control studies with disease parameters $GRR(Aa) = 2$, $GRR(AA) = 3$, $q = 0.04$ and $n_{cc} = 2000$, sampled from the "constant" population with the "full" ascertainment tag set.

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.

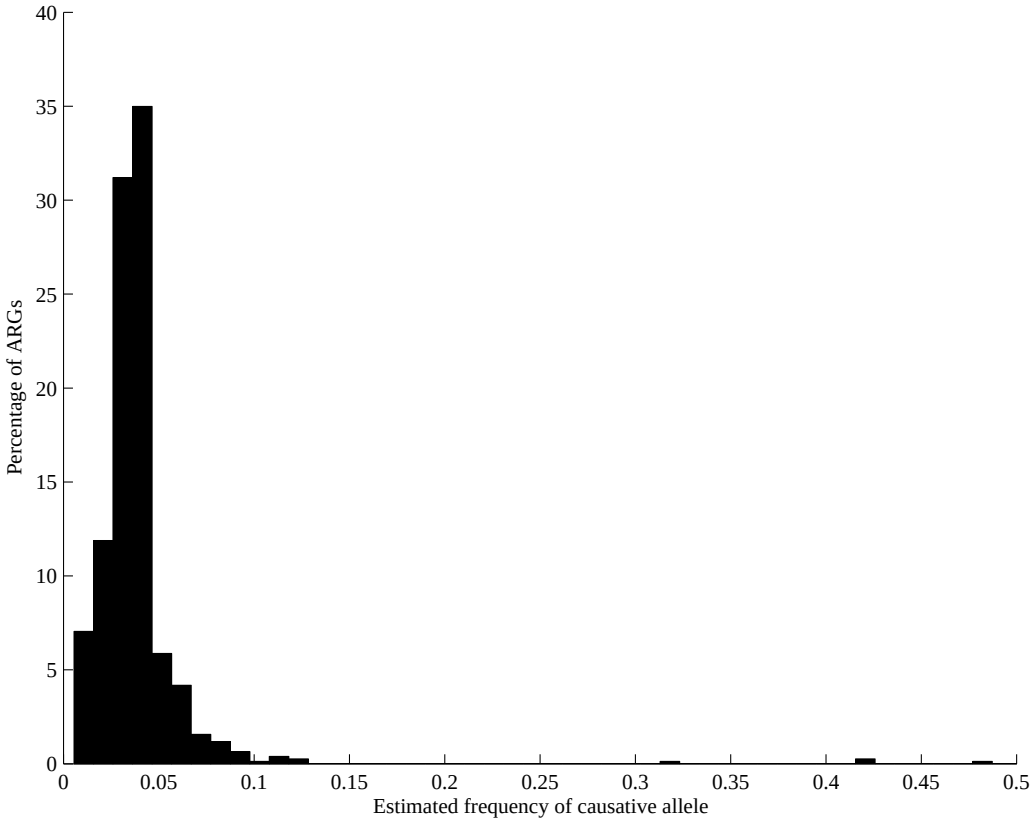

Figure 3.6: The distribution of estimated allele frequency in the general population. For a suite of case-control studies with disease parameters $GRR(Aa) = 2$, $GRR(AA) = 3$, $q = 0.04$ and $n_{cc} = 2000$, sampled from the "constant" population with the "full" ascertainment tag set.

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

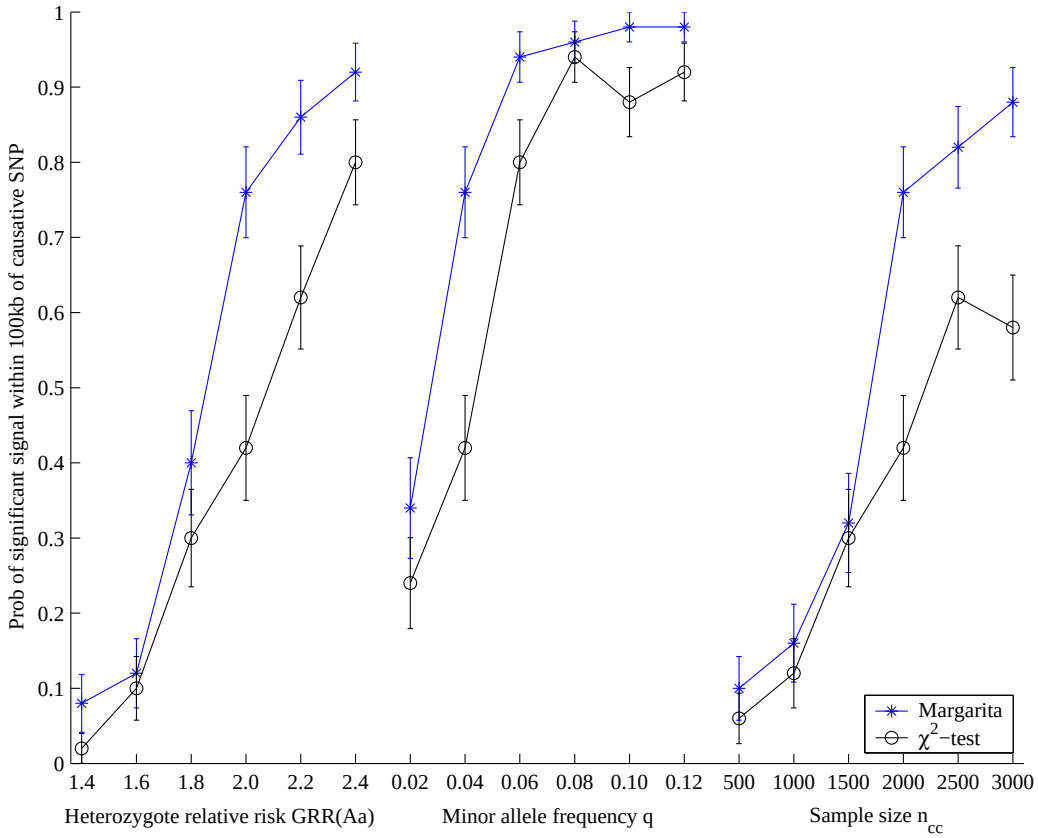

Figure 3.7: Probability of an experiment-wise significant signal within 100kb of the causative SNP (calculated as the proportion of studies in each suite that meet this criterion). Each point on the x-axis corresponds to a suite of 50 studies. Each of the disease parameters is varied between suites, 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.

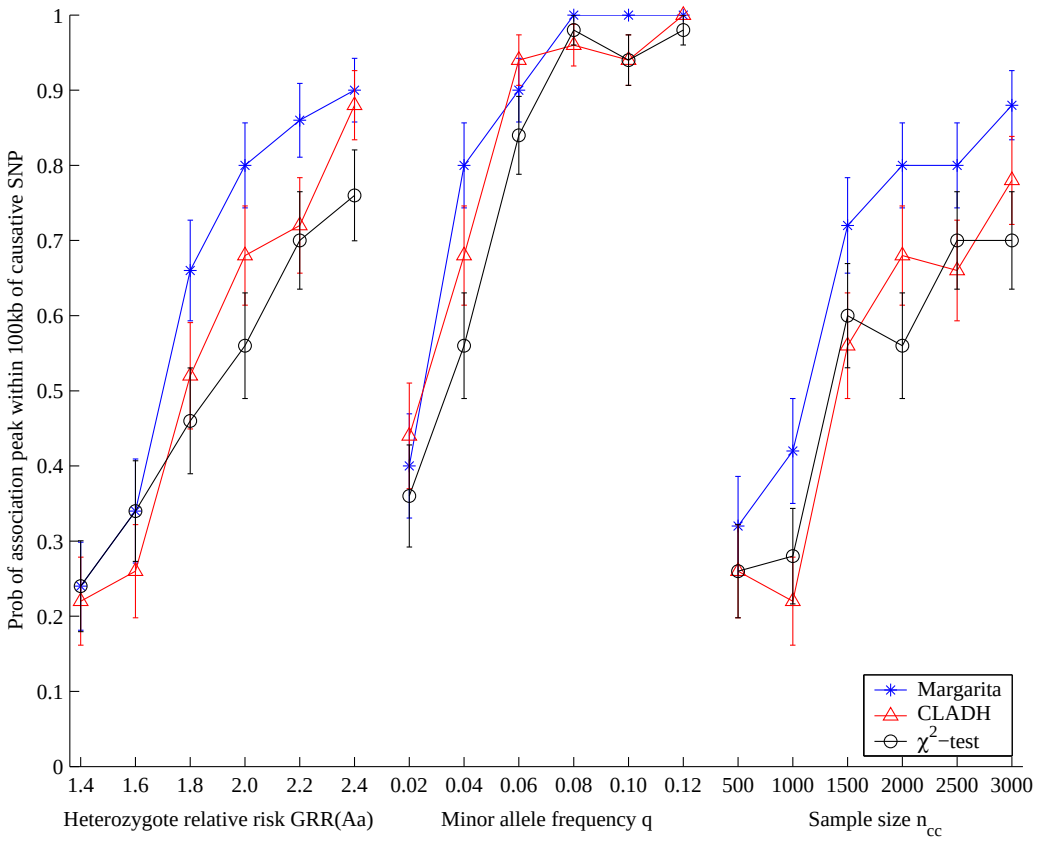

Figure 3.8: Probability that the association peak is within 100kb of the causative SNP. Each point on the x-axis corresponds to a suite of 50 studies. Each of the disease parameters is varied between suites, while the other parameters are held at "default" values of $GRR(Aa) = 2$, $GRR(AA) = 2 * GRR(Aa) - 1$, $q = 0.04$ 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

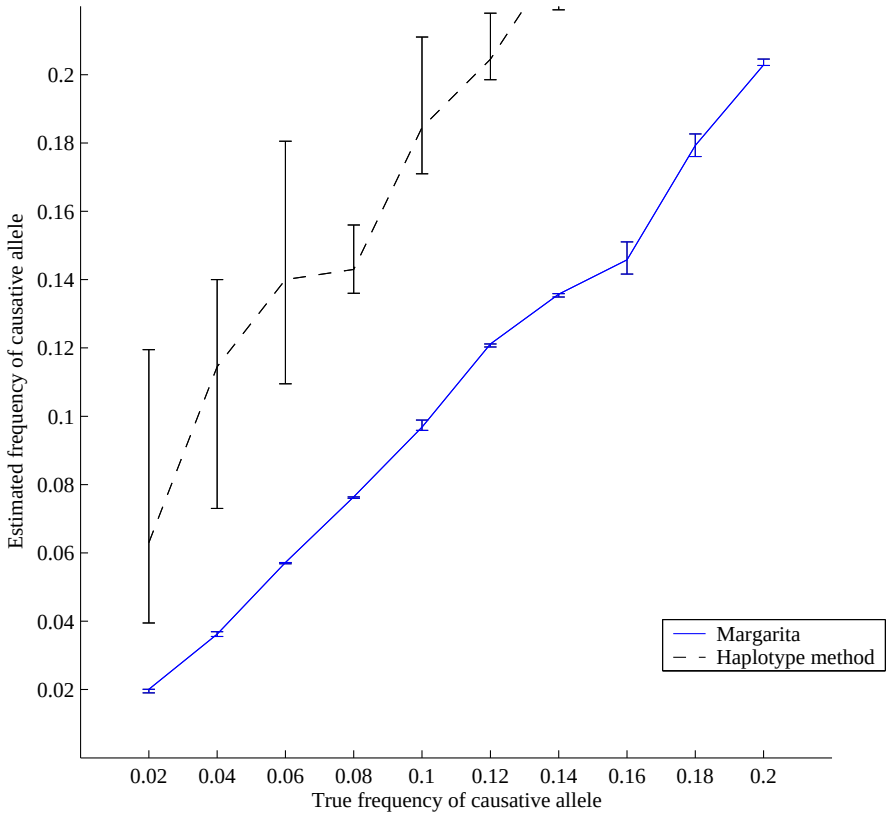

Figure 3.9: Estimated causative allele frequency versus true frequency q. Each point on the x-axis corresponds to a suite of 50 studies. q is varied while the other parameters are held at "default" values of $GRR(Aa) = 2$, $GRR(AA) = 2 * GRR(Aa) - 1$, $q = 0.04$ and $n_{cc} = 2000$. All studies are sampled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

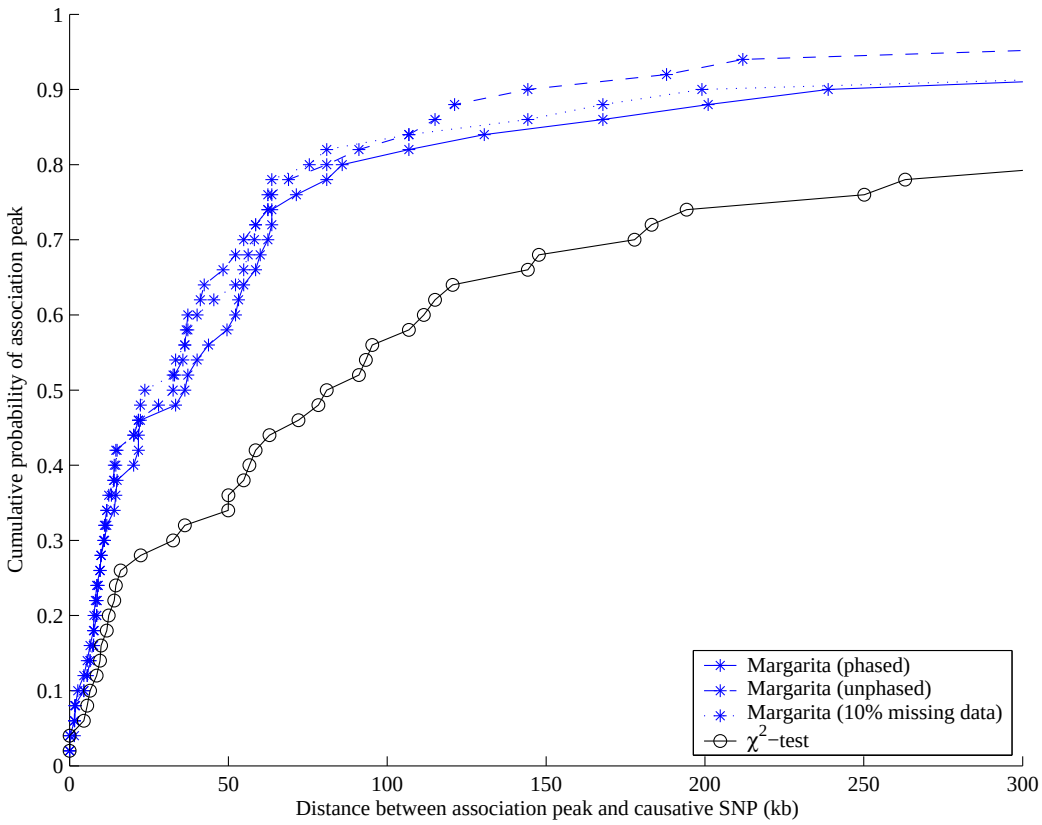

Figure 3.10: Performance on a suite of case-control studies with $GRR(Aa) = 2$, $GRR(AA) = 3$, $q = 0.04$, $n_{cc} = 2000$, sampled from the "constant" population and with the "full" ascertainment tag set. MARGARITA is applied to this suite under three scenarios: when the data is phased, when it is unphased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

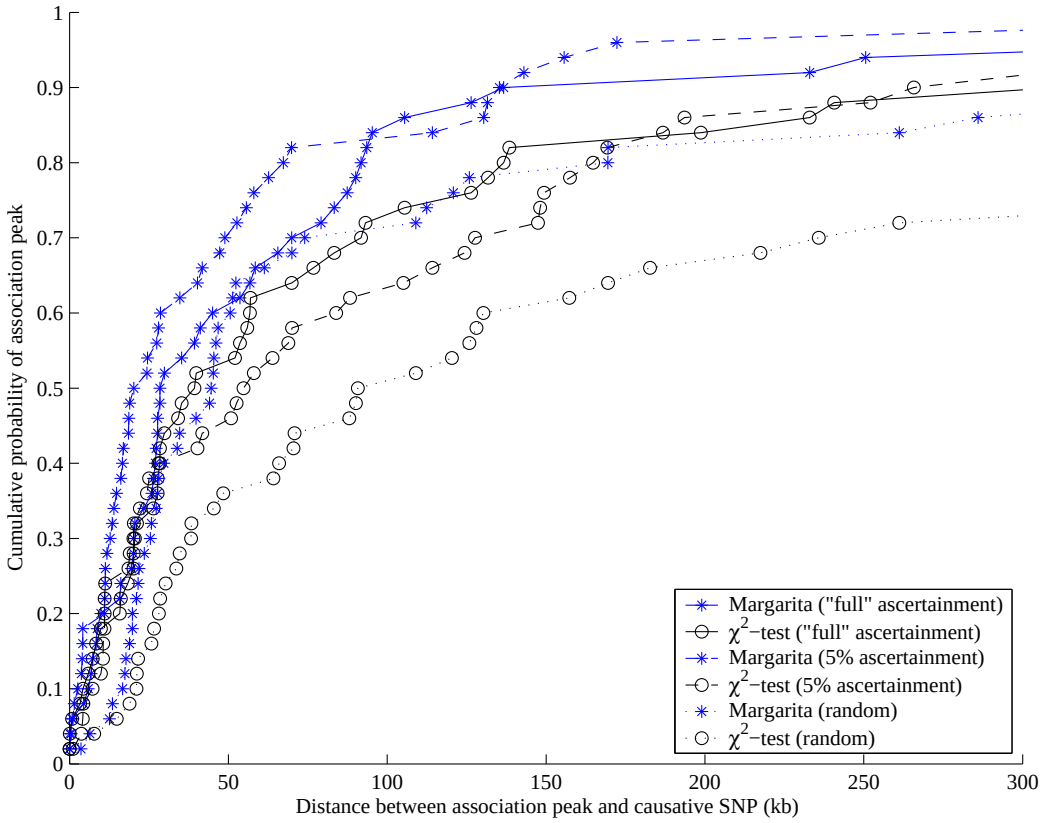

Figure 3.11: Localisation for different data, population and tag models. Performance on a suite of case-control studies sampled from the "hot" population (and with $GRR(Aa) = 2$, $GRR(AA) = 3$, $q = 0.04$, $n_{cc} = 2000$). Performance is compared using three different tagSNP ascertainment schemes.