doi: 10.1111/j.1469-1809.2007.00379.x

Fine Mapping of Disease Genes Using Tagging SNPs

Arvid Sjölander[1*], Ola Hössjer[2], Linda Werner Hartman[3] and Keith Humphreys[1]

[1]*Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm*
[2]*Department of Mathematics, Stockholm University, Stockholm*
[3]*Department of Mathematical Statistics, Lund University, Lund*

Summary

We describe a haplotype clustering approach for localising a disease mutation within a fixed genomic region, which supplements *tagging* SNP (tSNP) information with (external) information on linkage disequilibrium. By applying our method to simulated data based on the coalescent, and on real haplotype data, we demonstrate that there are situations where significant gains can be made by incorporating *tagged* SNPs into the analysis. The issues we explore are important not only to these types of studies, but also to studies that select tSNPs based on (external) HapMap phase II data, and those that use genome-wide markers.

Keywords: Association, haplotype clustering, Bayesian inference

Introduction

For fine scale mapping of disease susceptibility loci it is widely recognised that association studies, which exploit linkage disequilibrium (LD; Risch & Merikangas, 1996) are more appropriate than family-based linkage studies. Recent years have seen many hypothesis-driven candidate gene association studies. Early candidate gene studies genotyped a handful of single nucleotide polymorphisms (SNPs) or microsatellites. As the cost of genotyping reduced, many studies looked at a larger number of SNPs and, say, concentrated on a couple of dozen genes. More recently several candidate gene association studies have concentrated on subsets of SNPs, referred to as tagging SNPs (tSNPs), that best predict any other SNP within (and around) studied genes. Subsets of tagging SNPs can be identified either by typing a large number of SNPs in a small, preliminary, sample of individuals (Haiman et al. 2003), or on the basis of haplotype information from public databases (Gibbs et al. 2003; The International HapMap Consortium, 2005). Presently, many genome-wide association, as opposed to candidate gene, studies are being carried out. The polymorphisms included in these studies can also be viewed as tagging SNPs, although the degree to which the typed polymorphisms "tag" other polymorphisms varies across the genome and across different genotyping platforms ("Barrett & Cardon, 2006).

This article is concerned with estimating the position of a disease mutation within a fixed genomic region. Until recently methods for fine mapping had been developed only for the analysis of relatively large genomic regions, with relatively sparse markers; see Clayton (2000) for a review, and Morris et al. (2002) for a fairly recent approach. Molitor et al. (2003) have proposed a method which is potentially useful in smaller genomic regions. This uses Bayesian inference to estimate the position of disease mutations, and works on the rationale that haplotypes which are similar around a disease mutation will contribute similarly to disease risk. The method essentially compares similarities of haplotypes within cases to similarities within controls. We extend the method of Molitor et al. to incorporate (external) information on LD, additional to the LD information captured by the tSNPs which are typed on cases and controls. We concentrate on a design where tSNPs are selected using a small, preliminary, sample of controls, in a case-control study, and only these tSNPs are subsequently typed in the full case-control sample. In this case the information on LD is collected in the first stage of a two-stage genotyping procedure, where both stages are carried out using the same population sample. This design is likely for researchers working with populations not well represented by any of the four Hapmap samples. Our method can also be used by researchers selecting tSNPs using Hapmap.

* Corresponding author: Dr. Arvid Sjölander, Department of Epidemiology and Biostatistics, Karolinska Institutet, Nobels Väg 12, 171 77 Stockholm, Sweden. Phone: +46 852483859; Fax: +46 8314975; E-mail: Arvid.Sjolander@ki.se

Assessing performance of our method in this context is less straightforward than it is for the design we concentrate on (see the Discussion).

Whilst the literature on tSNP selection (Chapman et al. 2003; Stram et al. 2003a) has not implied that tagged SNPs should be dis-regarded in association analyses, little attention has been given to how/when this information can/should be used. We show here that under certain circumstances pertaining to LD structure, estimating the position of a disease mutation can be considerably enhanced if information on tagged SNPs is incorporated. Whilst we focus on estimating the position of a disease mutation, we also describe how our method can be adapted for testing for the presence of a disease mutation in a genomic region. Thomas et al. (2004) have described an approach for constructing tests using genotypes of tagging and tagged SNPs. Our test is instead based on a haplotype analysis. We note that if interest is restricted to testing, there are other simpler haplotype methods than the Bayesian mapping one, such as that described by Schaid et al. (2002), which could be chosen to be developed to incorporate "tagged" SNPs.

We describe the basic clustering method and extend the method to handle additional information concerning haplotypes on SNPs outside of the main study. We define measures which we use to score the mapping performance of our method. We investigate the mapping performance of our method. For this purpose we use data simulated from a coalescent model, and from real LD structure. We investigate how performance is related to recombination rate.

Model and Definitions

The Clustering Method

We begin by considering a scenario where the same set of tSNPs is available for all individuals and no external information on LD is incorporated. The method we use to analyse this data is that of Molitor et al. (2003). We suppose that there are individuals $i = 1, \ldots, I$ with measured genotypes G_i (a vector where each entry corresponds to a typed SNP) and phenotypes γ_i($\gamma_i = 0$ if individual i is a control, $\gamma_i = 1$ if individual i is a case). In genetic association studies phase is typically not known and it is not possible to unambiguously resolve haplotype pairs, $H_i = (h_{i1}, h_{i2})$, $i \in I$, at heterozygous sites. In practice one can estimate the conditional probability of $\{H_i\}$ from $\{G_i\}$ and treat the most probable haplotype pair as "true". In this section we simply assume that $\{H_i\}$ is known. Logistic regression is used to model the association between H_i and γ_i:

$$\text{logit}(p(\gamma_i = 1)) = \gamma_{c_{h_{i1}}} + \gamma_{c_{h_{i2}}}, \quad (1)$$

h	0	0	0	0	0	0	0
$t1$	1	0	1	0	0	1	0
$t2$	0	0	1	0	0	1	1

Figure 1 Clustering of haplotypes.

where $c_{h_{ij}} \in \{1, \ldots, C\}$ is the cluster to which haplotype h_{ij} belongs and C is the number of clusters. All haplotypes within the same cluster c are assumed to contribute equally to the disease risk γ_c, and we define $\gamma = [\gamma_1, \ldots, \gamma_C]$. Note that the "true" number of clusters is unknown.

We cluster haplotypes based on allelic similarity in the hope that similar haplotypes have approximately the same probability of carrying a functional mutation at an unmeasured locus, and contribute approximately equally to disease risk. As a starting point we define x to be the putative location of the functional mutation. We express the similarity, $w_{h,h'}$, between two haplotypes h, h' as the number of alleles that the haplotypes have in common within a fixed length window of typed SNPs centred at x. Mostly, we have used a window size of 6 SNPs, i.e. three to the left and three to the right of x, but have also used other (fixed) window sizes. For computational convenience we consider each cluster c to have a "center" t_c (defined to be an observed haplotype), and we define $T = [t_1, \ldots, t_C]$. The haplotypes are assigned to the cluster with the closest centre, according to the defined similarity metric. As an illustration we consider the three haplotypes listed in Figure 1, where h is an ordinary haplotype, and t_1 and t_2 are haplotypes which are also chosen as cluster centres. We assume that the functional mutation is located somewhere between the third and the fourth SNP from the left. A six SNP window thus excludes the rightmost SNP. In this case $w_{ht_1} = 3$ and $w_{ht_2} = 4$ and h would be assigned to the cluster which has t_2 as its centre. If the functional mutation is instead somewhere between the fourth and the fifth SNP a six SNP window excludes the leftmost SNP; $w_{ht_1} = 4$, $w_{ht_2} = 3$ and h would be assigned to the cluster which has t_1 as its centre.

It follows from the definition of $w_{h,h'}$ that "moving" the functional mutation in the space between adjacent typed SNPs does not change the haplotype similarities, and hence not the cluster configuration. For example, if the functional mutation is between the fourth and the fifth SNP then $w_{ht_1} = 4$ and $w_{ht_2} = 3$, regardless of its exact location in this region.

The parameters for the model are thus the vector of risk parameters, γ, the mutation locus, x, and the vector of cluster centres, T. Since the true number of clusters is unknown, the model dimension, C, appears also as a parameter whose value must be estimated.

The Bayesian Perspective

We define $\gamma = [\gamma_1, \ldots, \gamma_I]$. The prospective likelihood, $p(\gamma|\gamma, x, T, C)$, is

$$p(\gamma|\gamma, x, T, C) = \prod_i \frac{e^{\gamma_i(\gamma_{\epsilon_{h_{i1}}} + \gamma_{\epsilon_{h_{i2}}})}}{1 + e^{\gamma_{\epsilon_{h_{i1}}} + \gamma_{\epsilon_{h_{i2}}}}}. \quad (2)$$

The parameters x, T and C do not explicitly appear in (2). Implicitly however they are important since together they determine the cluster configurations, i.e. the number of clusters and the cluster to which each haplotype belongs. If x, T and C were known we could use standard software for logistic regression to draw inference about odds ratios, ignoring the retrospective sampling scheme (Prentice & Pyke, 1979). When x, T and C are unknown exact inference is more complicated. Strictly speaking we can no longer ignore the retrospective sampling scheme. Despite this fact we use the prospective likelihood. The effect of ignoring the retrospective sampling scheme for estimating haplotype specific effects has been discussed by Stram *et al.* (2003b). If tSNP genotypes predict haplotypes well then there is little effect. For our model, maximum likelihood estimation of x, T and C is not computationally feasible. We adopt the fully Bayesian approach, described by Molitor et al. assigning a probability model $p(\gamma, x, T, C)$ (a prior) to the unknown quantities (note that we use "$p(\cdot)$" for both "probabilities" and "densities" here). In the Bayesian framework inference is drawn from the posterior probability function $p(\gamma, x, T, C|\gamma)$, which in general is calculated as

$$p(\gamma, x, T, C\,|\,\gamma) = \frac{p(\gamma|\gamma, x, T, C)\,p(\gamma, x, T, C)}{p(\gamma)}. \quad (3)$$

If we for example want to find the *most probable* location of a functional mutation we marginalize $p(\gamma, x, T, C|\gamma)$ over γ, T and C, and search for the mode of the resulting posterior distribution for x. Due to the complexity of the problem $p(\gamma, x, T, C|\gamma)$ cannot be calculated explicitly. By using Markov Chain Monte Carlo (MCMC) techniques we can, however, simulate from the posterior distribution. The algorithm is described in detail in Appendix S1.

Selection of Priors

We use a prior distribution which reflects our prior (vague) assumptions about the model parameters. We consider γ, x and T to be independent given C, and x independent of C, i.e. $p(\gamma, x, T, C) = p(\gamma|C)p(T|C)p(x)p(C)$. The elements of γ are considered to be mutually independent and normally distributed (mean 0, standard deviation 1) given C. The a-priori expected value of γ_c given C is thus 0, i.e. we expect no haplotype to be either (relatively) protective or harmful. If the analyzed genetic region does not contain a functional locus we expect all clusters to contribute with the same risk, or rather, all haplotypes to belong to the same (single) cluster. Placing a large prior probability at $C = 1$ corresponds in some sense to having a large prior belief in the null hypothesis of no functional locus. We let $p(C = 1) = 0.5$ and $p(C = k) = \frac{0.5}{C_{max}-1}$, $k \in \{2, \ldots, C_{max}\}$, where C_{max} is set to 15. Given C, we consider all possible T-vectors to be equally probable, i.e.

$$p(T|C) = \binom{Q}{C}^{-1}, \quad (4)$$

where Q is the number of observed haplotypes. We are completely ignorant about the location of the functional mutation, hence we place a uniform prior on x over the region between the leftmost and the rightmost typed SNP.

Incorporating Additional Information on Haplotypes

We have, until now, assumed that the set of tSNPs is the same for all individuals and that no additional information concerning LD is available and to be incorporated into the analysis. We now extend the mapping method to handle the situation where the tSNPs are, in fact, a subset of a larger set of markers which have been typed on a small subset (a preliminary sample) of individuals. This is the design considered by Thomas et al. (2004) and used, for example, by Haiman et al. (2003). In order to make use of the information captured in the tagged SNPs we modify the basic model as follows.

We denote an arbitrary haplotype, formed by the alleles at the full set of markers, by h. Similarly, we denote an arbitrary haplotype formed by the alleles only at the tSNPs by h^*. Given that we observe only the alleles at the tSNPs in the second stage sample, we can calculate the probability of h as

$$p(h|h^*) = \frac{p(h^*|h)\,p(h)}{p(h^*)} = \frac{I(h^* \sim h)\,p(h)}{\sum_h I(h^* \sim h)\,p(h)}, \quad (5)$$

where $I(h^* \sim h) = 1$ if h^* and h are equal at tSNPs, 0 otherwise, and the sum in the denominator is taken over all existing haplotypes h. We estimate $p(h|h^*)$ by taking the sum over all haplotypes observed in the preliminary sample, and replacing their true population frequencies, $p(h)$, with observed proportions. We treat this estimate $\hat{p}(h|h^*)$ as given. $\hat{p}(h|h^*)$ is consistent if subjects in the preliminary sample are exchangeable with subjects in the full sample. When the preliminary sample consists of controls only, as in the setting we consider, $\hat{p}(h|h^*)$ may be inconsistent. We expect our method to perform optimally when the

true value of $p(h|h^*)$ is used, and we thus expect the results we present below, which rely on a preliminary sample of controls only, to be conservative. Instead of expressing the similarity between haplotypes h_A^* and h_B^*, say, as the number of shared alleles within a window of tSNPs, we use the *expected* similarity within a window of consecutive SNPs (typed in the preliminary sample),

$$E[w_{h_A h_B}|h_A^*, h_B^*] = \sum_{h_A \sim h_A^*, h_B \sim h_B^*} w_{h_A h_B} p(h_A|h_A^*) p(h_B|h_B^*), \quad (6)$$

assuming both cases and controls are in HWE. Below we compare the performance of analysis based on this model, i.e. using all SNPs (tagging and tagged), with analysis based on a model using only tSNPs.

Mapping Performance Measures

The clustering mechanism assigns haplotypes which are similar around a specific locus, x, to the same cluster. The locus x is continuously updated in the MCMC-algorithm, yielding a new cluster configuration for each update. If x is a "true" mutation locus we expect a much better model fit (i.e. much larger value of $p(\gamma|\gamma, x, T, C)$) than if x is not. The posterior distribution for x, $p(x|\gamma)$, can hence be used as a mapping tool. A strong peak at a locus in the posterior distribution provides evidence of a disease mutation in the region, whilst a uniform posterior distribution does not. Because our main focus is on location estimation, we use as our first measure of performance a score which does not reflect (association) signal strength. This score, which we call the *Average coverage probability* (ACP), allows us to compare how methods would perform if follow-up genotyping/sequencing were used for different (fixed) region sizes. For a researcher limited to exploring a region of a fixed size it is sensible to choose the region as that where $p(x|\gamma)$ is largest. We calculate the ACP as the probability that the true location of the causative site is detected under this strategy (i.e. contained in the follow-up region). The ACP is similar to the cumulative distribution of distances between an estimated and true disease mutation which Zöllner & Pritchard (2005) used to assess their LD method. We denote the typed SNPs by $s_1, s_2, \ldots, s_J$, the region between s_j and s_{j+1} by a_j, and the length of a_j by $|a_j|$ (we will assume a normalized scale, i.e. $\sum_{j=1}^{J} |a_j| = 1$). Due to the clustering method, $p(x|\gamma)$ is flat between adjacent typed SNPs. We denote the value of $p(x|\gamma)$ in region a_j by p_j. Under the assumption that we are limited to sequence a region A of length $|A|$ less or equal to L, we define Ω as the set of indexes for which $\sum_{j\in\{\Omega\}} p_j$ is maximized under the constraint $\sum_{j\in\{\Omega\}} |a_j| \leq L$. We define $A = \{a_j\}, j \in \Omega$. A is thus a function of both γ and L, and to be explicit about this we write $A(\gamma, L)$. The functional mutation is detected if $x \in A$ and we thus define the ACP, as a function of L, as

$$p(x \in A(\gamma, L)) = \int_{x,\gamma} I(x \in A(\gamma, L)) p(x, \gamma) dx d\gamma, \quad (7)$$

where $I(x \in A(\gamma, L)) = 1$ if $x \in A(\gamma, L)$, and 0 otherwise. If L is a proportion then the ACP takes the value 0 for $L = 0$ (if we do not sequence a region we will not find x) and value 1 for $L = 1$ (if we sequence the whole region we will find x). It is instructive to compare any given ACP curve with the curve that would be obtained if $A(\gamma, L)$ is picked completely at random ("blind guessing"). In this case we have that ACP = L, that is, the probability of finding the true locus is directly proportional to the size of the followed-up region. Any method yielding an ACP curve which is consistently above the straight line from 0 to 1 is thus better than blind guessing. The ACP is estimated for simulated data below.

A researcher with extensive resources might, rather than fixing the region size in advance, continue to search in subregions until the true locus is found. With this strategy in mind, an interesting quantity is the average region size one needs to cover before finding x. We denote this quantity by Q. From the definition of the ACP,

$$Q = \int_L L \frac{dACP}{dL} dL = \int_L L \, dACP \quad (8)$$

If $A(\gamma, L)$ is picked completely at random, then Q is trivially equal to 0.5.

The ACP focuses on the estimation of the position of a disease mutation. In practice, one might be reluctant to search for a disease mutation in regions for which $p(x|\gamma)$ is low, no matter whether the resources to do so are available or not. Consider the following example. Suppose that two different methods give posterior distributions $N(x, \sigma^2)$ and $N(x, \sigma^2/k)$, $k > 1$. That is, the posterior distribution $p(x|\gamma)$ is not flat between adjacent typed SNPs. In this case our previous definition of $A(\gamma, L)$, together with the assumption that we now sequence a length of exactly L, yields $A(\gamma, L) = [x - L/2, x + L/2]$, for both methods. However, if k is large, meaning that the second method yields a very peaked posterior, an investigator would probably not waste resources on follow-up genotyping/sequencing in regions far away from the peak. We supplement the ACP and Q by reporting the spread of the posterior distributions, which we measure using the mean variance. The mean variance alone is not a sufficient measure of mapping performance, since a posterior distribution could have a very tight peak (i.e. low variance) at a "wrong" location. We therefore report the mean variance together with the mean distance between the posterior mode and x, that is, the mean bias in the mode as a point estimate of x. Because the posterior

is flat between markers we use midpoints. It can be shown that, if $A(\gamma, L)$ is picked completely at random, then the mean bias is 1/3 in expectation.

A Test Statistic

Although the main focus of this paper is on fine mapping, it is also possible to use the method for association testing. We have argued that we expect all haplotypes to fall in the same cluster if there is no functional mutation in the region. A reasonable choice of test statistic, D, is thus the posterior probability that $C > 1$, i.e.

$$D = \frac{\int_{C>1} p(\gamma, x, T, C|\gamma) d\gamma\, dx\, dT}{\int_{C=1} p(\gamma, x, T, C|\gamma) d\gamma\, dx\, dT}. \quad (9)$$

A test of association can be defined as

reject H_0 at significance level α if $D > K_\alpha$, where $\alpha \equiv p(D > K_\alpha | H_0)$. (10)

One approach for choosing a value of K_α is described below. We note that an alternative test could be based on a uniform posterior for x (under the null hypothesis of no disease mutation in the region).

Results

Simulations Based on a Coalescent Model

In order to examine the performance of the clustering method we simulated five different haploid populations from a *coalescent model* (Balding et al. 2001). From each population we repeatedly drew case-control samples which were analyzed with the clustering algorithm.

Simulation of the population, selection of markers and tSNPs, and simulation of case-control samples

The coalescent is a widely used modelling tool for population genetics. We used the publicly available software *ms* (Hudson, 2002) to generate several populations under the coalescent. The coalescent model is driven by two parameters, $\theta = 4N_0\mu$ and $\rho = 4N_0 r$, where $N_0$is the (effective) haploid population size, μ is the mutation probability per haplotype and generation, and r is the recombination probability per haplotype and generation. Five haploid populations were simulated (each using a different value of r; see below). The effective population size for each population was assumed to be 5000, and we considered regions of length 100kb. To begin with we used values $\mu = r = 2 \times 10^{-3}$, which can be considered to be realistic (Phillips et al. 2003). Note that because we use a region length of 100kb μ and r are equal to 2×10^{-8}, measured in base pairs. Figure 2 displays the LD pattern for this population. The pattern is similar to several 100kb regions observed in the ENCODE regions studied by the International HapMap Consortium (2005).

We created four additional populations by keeping μ fixed at 2×10^{-3} and varying r over four different values. For each population we removed all SNPs with a minor allele frequency (MAF) of less than 5%. We did this since it is known to be difficult to detect rare variants using standard association tests. From the remaining set of SNPs we selected one out of every 20 markers, as evenly spaced as possible. This procedure yields an approximate marker distance of 5kb, which is similar to the marker distance used in many case-control studies where tSNPs have been internally identified (e.g. Haiman et al. 2003; Einarsdottir et al. 2005).

We have used the abbreviation tSNP to denote (generally) a tagging SNP. Tagging SNPs can be selected to predict SNPs individually or to predict haplotypes. In this article we used the former approach. We used the squared correlation between true and predicted allele dosages, R_s^2, as described by Stram (2003a). From the set of markers created with an approximate marker distance of 5kb, we selected a subset of tSNPs as the minimal set which provided a minimum R_s^2 value of 0.8. We used a MAF cutoff at 5%, meaning that tSNPs were selected only to ensure good prediction of markers with a MAF above 5% (that is "all markers in the sample", since markers with a MAF less than 5% were already removed). Diploid case-control samples were randomly drawn from each population. Each sample consisted of 1500 cases and 1500 controls. We assumed a rare disease, a single functional mutation with a relative risk ψ (multiplicative penetrance), and a population in HWE. This implies that SNPs and haplotypes are in HWE both among controls and among cases, the controls having approximately the same haplotype frequencies as in the population, and the cases having modified haplotype frequencies

$$f_i^* = \frac{f_i \psi_i}{\sum_j f_j \psi_j}, \quad (11)$$

where f_i is the frequency for haplotype i in the population, and ψ_i are haplotype relative risks; $\psi_i = \psi$ if haplotype i carries the mutation, and $\psi_i = 1$ otherwise (Clayton, 2001). For our simulation we used a value of 1.4 for ψ which, given our choice of sample size, was (empirically) found to yield on average a power of 0.8 at 5% significance level for detection of a functional mutation. For each population we moved the location of the functional mutation over all SNPs (not only the subset of SNPs selected as markers in the preliminary sample) in the region between the leftmost and the rightmost tSNP. The reason for not letting

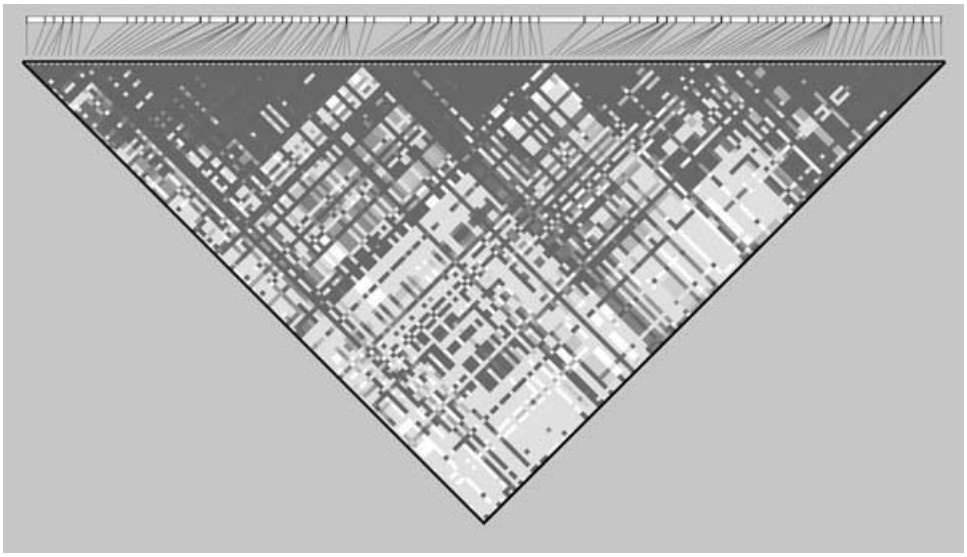

Figure 2 LD pattern (D') for the population with $\mu = r = 2 \times 10^{-3}$.

the functional mutation be located outside the region spanned by the tSNPs is explained below. This procedure yielded approximately 100 samples from each population.

Estimation of statistical power

Above we discussed how our approach, in principle, could be adapted for use in testing for association. We empirically evaluated the power of a test using simulated data based on the coalescent model. We used the population which we consider to be the most realistic in terms of mutation and recombination rates ($\mu = r = 2 \times 10^{-3}$). To ensure precise estimates of power, even for low significance levels, we tenfolded the number of samples for this specific population. In addition we randomly drew equally many (approximately 1000) case control samples from this population *assuming a relative risk of 1*. The latter samples can be viewed as being drawn under H_0 (with a relative risk of 1 a mutation is, by definition, non-functional). The former samples can be viewed as being drawn under the alternative hypothesis, H_A. The clustering method was applied to each case-control sample using (i) only tSNPs, and (ii) all of the markers observed in the preliminary sample of 100 controls and the tSNPs observed in the remaining 1400 controls and 1500 cases (i.e. tSNPs and tagged SNPs).

We empirically estimated the power of the test described above as a function of α. First we estimated K_α, for $\alpha \in \hat{K}_\alpha[0.01, 0.05]$, as the minimal value for which

$$\frac{n_{H_0}^{\hat{K}_\alpha}}{n_{H_0}} > \alpha, \tag{12}$$

where $n_{H_0}^{\hat{K}_\alpha}$ is the number of samples, simulated under H_0, for which $D > \hat{K}_\alpha$, and n_{H_0} is the total number of samples simulated under H_0. For fixed values of α we consistently estimated the power using (i) and (ii), as

$$\frac{n_{H_A}^{\hat{K}_\alpha}}{n_{H_A}}, \tag{13}$$

where $n_{H_A}^{\hat{K}_\alpha}$ is the number of samples, simulated under H_A, for which $D > \hat{K}_\alpha$, and n_{H_A} is the total number of samples simulated under H_A. As a comparison to our proposed test we also included the (estimated) power function of a standard likelihood ratio (LR) test based on the following logistic regression model:

$$\text{logit}\, p(y_i = 1 | A_i) = \beta_0 + \beta_A^T A_i \tag{14}$$

where $A_i = (A_{i1}, \ldots, A_{iK})$, K is the number of tSNPs, β_A is a $1 \times K$ vector, and

$$A_{ik} = \begin{cases} 0 \text{ if individual } i \text{ is homozygous carrier of reference allele at tSNP } k \\ 1 \text{ if individual } i \text{ is carrier of one reference allele at tSNP } k \\ 2 \text{ if individual } i \text{ is non-carrier of reference allele at tSNP } k \end{cases} \tag{15}$$

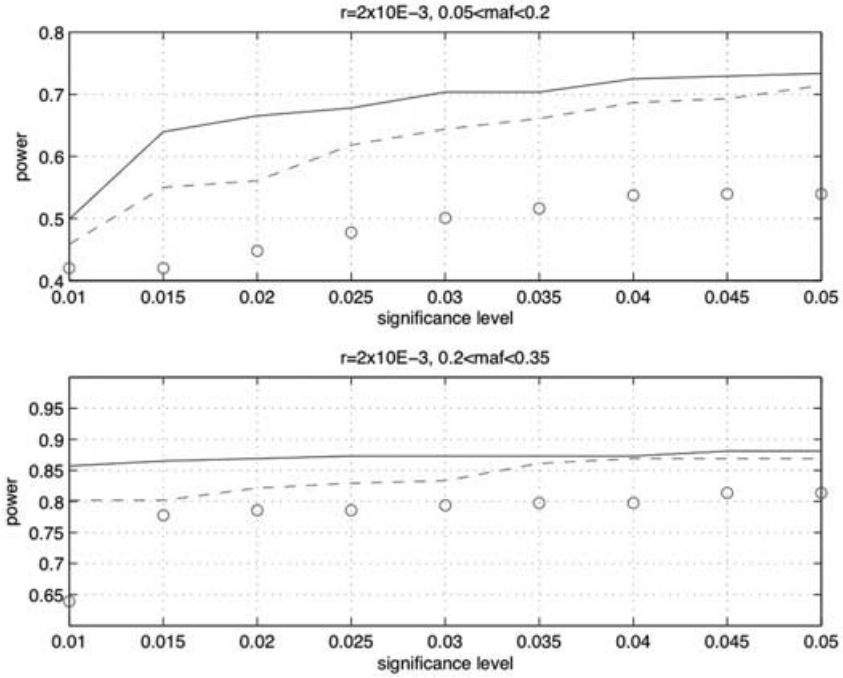

Figure 3 Empirical estimates of power by MAF. The dashed line is for tSNPs only, the solid line for tSNPs + tagged SNPs. The line of circles represents the LR test based on logistic regression.

We have plotted empirical estimates of power in Figure 3. We divided results by MAF at the disease locus (3 groups: MAF≤ 0.2, $0.2<$MAF≤ 0.35, $0.35<$MAF≤ 0.5). We first note that the power for all three methods increases with MAF. We excluded the plot for MAF>0.35, since in this region the power was observed to be very close to one for all three methods and all significance levels. The test based on haplotypes was consistently more powerful when tSNPs and tagged SNPs were used, in comparison to using tSNPs only. The power of our haplotype-based test using tSNPs only was observed as being consistently larger than the power for the test based on logistic regression using tSNP genotypes. The test which we have described assumes (complete) knowledge of haplotypes/LD, and hence requires development before it can be used for real data analysis. The simulation result does however suggest that, for this particular choice of recombination rate and mutation rate, there is potentially an advantage to dealing with both haplotype phase and tagged SNPs in testing for association.

Assessing mapping performance

We note that the prior distribution for x is only defined within the region between the leftmost typed SNP and the rightmost typed SNP. With the mapping method we propose, when using tSNPs only, we do not explicitly allow for a disease mutation to be located outside the region spanned by the tSNPs. To enable a fair comparison between using tSNPs only and using both tSNPs and tagged SNPs, in generating data we allow only for x to be between the leftmost and the rightmost tSNP.

In a practical setting one would probably test for association first, and then if the test falls out positive try to locate the functional mutation. For each simulated sample we tested for association using the LR test based on the additive logistic regression model in (14). The samples for which the test rejected the null hypothesis at 5% significance level (we refer to those samples as "verified") were used to estimate ACP, Q, mean bias and mean variance.

For each of the five populations we consistently estimated the ACP for using tSNPs only, and for using tSNPs and tagged SNPs separately, $L \in [0, 1]$, as

$$\frac{n_L}{n}, \tag{16}$$

where n_L is the number of samples in which x is covered by $|A| \leq L$, and n is the total number of samples. Figure 4 summarizes the results.

For a "very low" recombination rate ($r = 10^{-4}$) we have that ACP $= L$ for both using tSNPs only and using tSNPs and tagged SNPs. This means that the clustering algorithm performs no better than "blindly guessing". This is to be expected; in a region with a very low recombination

rate the magnitude of LD tends towards being independent of distance between loci. As the recombination rate increases we would expect LD to decrease as a function of distance, and as a result it will become possible to use markers for locating functional mutations more accurately. In our simulations we observed this pattern – the line for tSNPs and tagged SNPs moved away from the ACP $= L$ line as r increased from 10^{-4} to 10^{-3}. We saw that the line for tSNPs only joined the line for tSNPs and tagged SNPs from the right-hand side of each graph as r increased. If one searched for disease mutations in relatively small regions (e.g. $L = 0.2$), as might in practice be the case, it is clear that, for low to moderate recombination rates ($r = 0.5 \times 10^{-3}$ to $r = 2 \times 10^{-3}$) there is a significant gain in using tagged SNPs in addition to tSNPs. For large recombination rates the gain is minimal, because the proportions of SNPs in the preliminary sample, selected as tSNPs, approaches 1.

Table 1 displays the estimates of Q, the mean bias, and the mean variance as a function of recombination rate. As indicated by the ACP curve, for the lowest recombination rate, estimates of both Q and the mean bias were close to what we would expect from "blind guessing" . When comparing tSNPs only with tSNPs + tagged SNPs, we saw that estimates of Q, the mean bias, and the mean variance, are lower for tSNPs + tagged SNPs for almost all recombination rates. Exceptions arose for the bias, for the highest and lowest recombination rates. In those cases the estimates of the bias were very similar for the two methods, and the difference is likely to be due to sampling variability.

Table 1. Estimates of mean bias, mean variance and Q in coalescent simulations, for (i) tSNPs only and (ii) tSNPs + tagged SNPs

r	Q (i)	Q (ii)	bias (i)	bias (ii)	variance $\times 10^{-3}$ (i)	variance $\times 10^{-3}$ (ii)
10^-4	0.61	0.53	0.36	0.37	2.2	0.2
0.5×10^{-3}	0.48	0.32	0.26	0.2	2.6	0.22
10^{-3}	0.36	0.21	0.15	0.1	1.6	0.21
2×10^{-3}	0.34	0.27	0.13	0.09	1.1	0.21
4×10^{-3}	0.31	0.29	0.08	0.09	0.5	0.23

The importance of taking proper care of the two stage design can be emphasised even more if we consider situations where tSNPs are selected in a way that does not ensure good prediction of all stage 1 SNPs. We illustrated this by repeating the analysis for $r = 10^{-3}$ (low rate), $r = 2 \times 10^{-3}$ (moderate rate) and $r = 4 \times 10^{-3}$ (high rate), but with the tSNPs selected as the minimal set which provided a minimum R^2_h of 0.8 (Stram et al. 2003a) instead of R^2_s. A rare haplotype cutoff at 5% was used, meaning that tSNPs were selected only to ensure good prediction of haplotypes with a population frequency above 5%. In these three populations the common haplotypes (frequency > 5%)

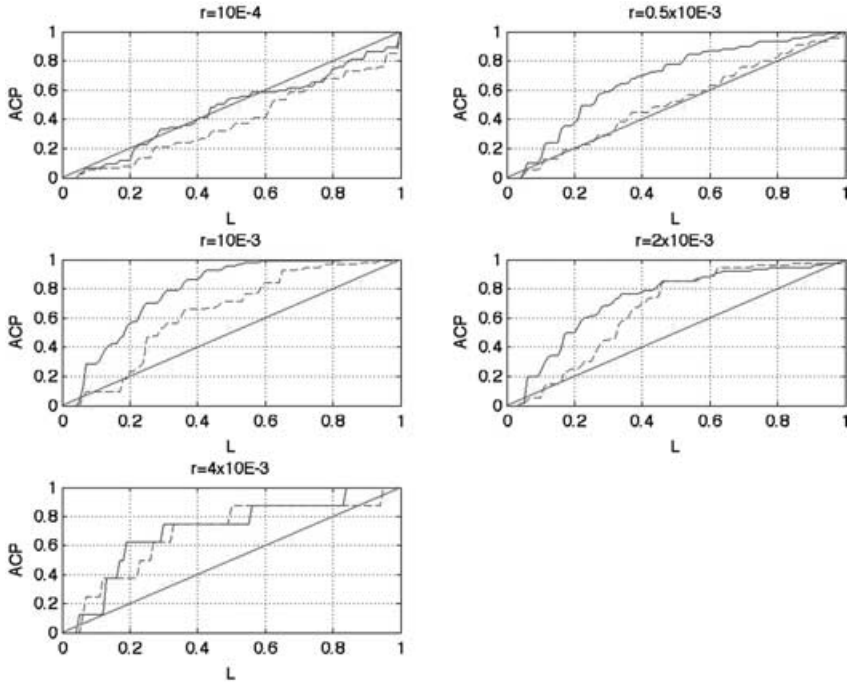

Figure 4 Estimating ACP. Dashed lines are for tSNPs only, solid lines for tSNPs + tagged SNPs. In each sub figure the line corresponding to ACP $= L$ is drawn.

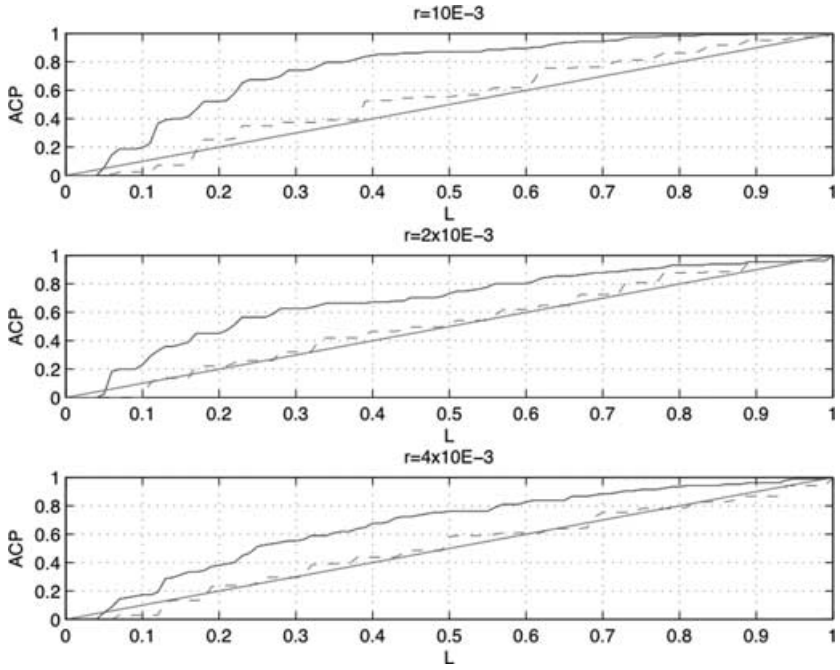

Figure 5 Estimating ACP when tSNPs are selected using R_h^2. Dashed lines are for tSNPs only, solid lines for tSNPs + tagged SNPs. In each graph the line corresponding to ACP = L is drawn.

accounted for 24%, 46% and 58% of the total haplotype pool. In practice tSNPs would not be selected according to such a criterion. One reason we chose to use this criterion was that we found that, unlike in the initial simulations (Figure 4), the proportion of stage 1 SNPs selected as tSNPs did not increase as the recombination rate increased. Figure 5 displays the results from the second set of simulations.

Figure 5 shows that, although the tSNPs alone perform almost as badly as blind guessing, there is a huge gain to be made by incorporating the tagged SNPs from stage 1 in the algorithm.

Our algorithm is based on a fixed window length, which is clearly not optimal (see Discussion). Although we have not introduced a window length parameter in our Bayesian algorithm, we did try out different choices of fixed window length. We repeated the analysis for $r = 10^{-3}$ and tSNPs selected according to the R_s^2 criterion, with 4 and 8 flanking SNPs, respectively. The result is displayed in Figure 6. Although the result seemed to be somewhat sensitive to window size, we note that for all three choices of window size the performance of the method was greatly improved when tagged SNPs were included, and that the 'best' ACP for tSNPs was only inferior to all ACPs for tSNPs + tagged SNPs.

Simulations Based on Real Haplotype Data

The simulations described above are based on a coalescent model under the assumption of constant recombination rates. We also studied the performance of our method using simulations based on real haplotype data collected in an ongoing candidate gene study of postmenopausal breast cancer risk (Einarsdottir et al. 2005). In each gene SNPs were selected from databases, aiming for a marker density of at least one SNP per 5kb, and typed on a preliminary sample of 92 controls. The tSNPs were typed on 1500 cases and 1500 controls. We used haplotype data from the preliminary sample for one particular gene, ESR1. More specifically we constructed a haploid population with haplotype frequencies equal to those which were observed (predicted) in the preliminary sample. We considered a sub-region of 131.5kb at the start of this gene (chr 6; 152300000..152168500). Fifty SNPs (MAF>0.05) were typed for the controls in the preliminary sample (i.e. the marker density was roughly 1 SNP per 2.6kb). Seventeen tSNPs were selected in this region, predicting common SNP genotypes with a minimum $R^2 \geq 0.8$. We noted that marker density in this preliminary sample was not much less than that of HapMap phase II in this region. HapMap phase II typed 78 SNPs (MAF>0.05) in this region and

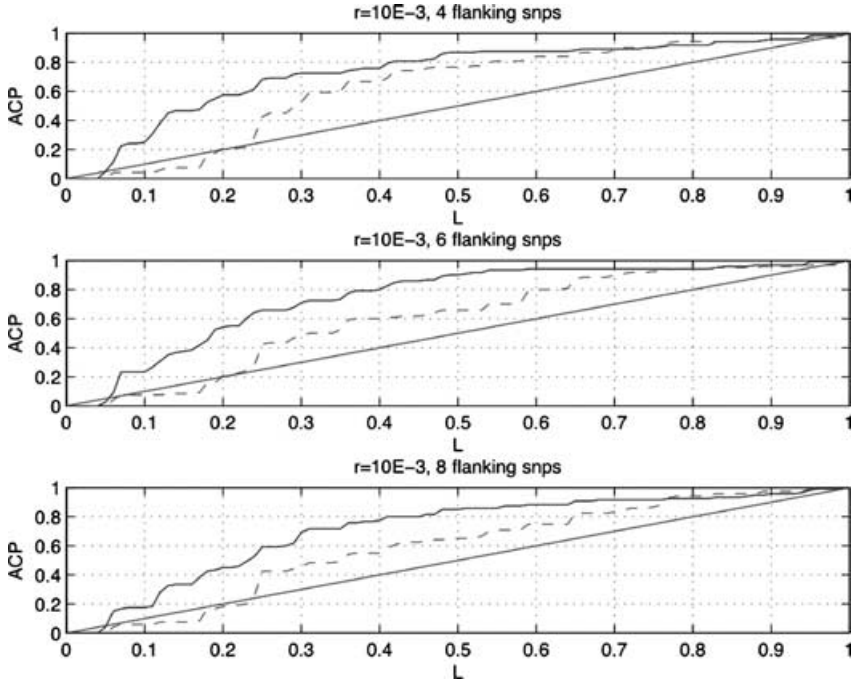

Figure 6 Estimating ACP with different fixed window sizes, $r = 10^{-3}$. Dashed lines are for tSNPs only, solid lines for tSNPs + tagged SNPs. In each graph the line corresponding to ACP $= L$ is drawn.

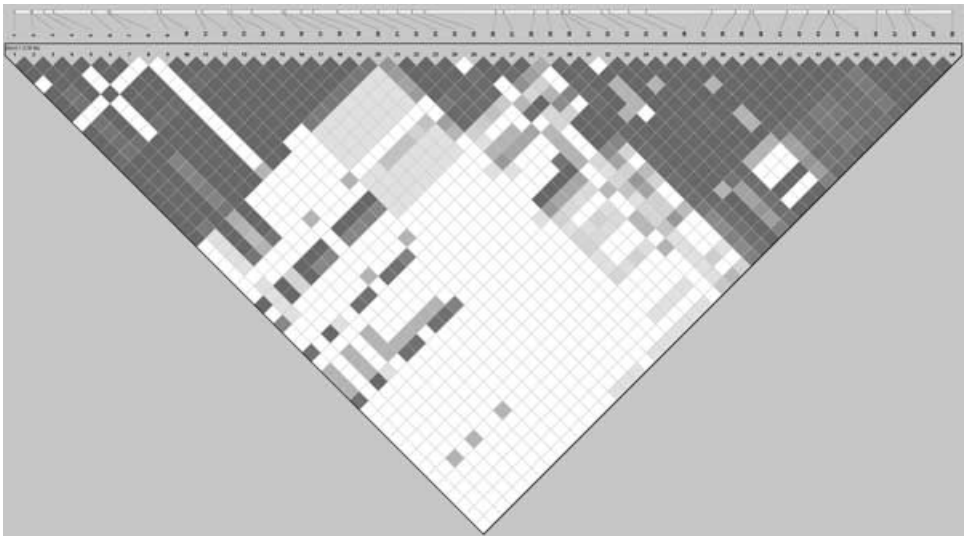

Figure 7 LD (D') in a subregion of ESR1

selected 24 SNPs based on its multimarker tSNP selection procedures. An LD plot for this region, based on the breast cancer study data, is shown Figure 7. Based on this haplotype data we performed a simulation study, drawing 100 samples of 1500 controls and 1500 cases, using a relative risk of $\psi = 1.4$, and moving the location of the functional mutation over all SNPs. The procedure used for drawing these samples was similar to that described for our simulation study based on the coalescent. We note that a drawback of this simulation, based on real haplotype data, is that haplotype diversity is underestimated since haplotype population frequencies are estimated using only 184 chromosomes. We plotted the ACP curve in Figure 8. For one particular sample we plotted the posterior distributions

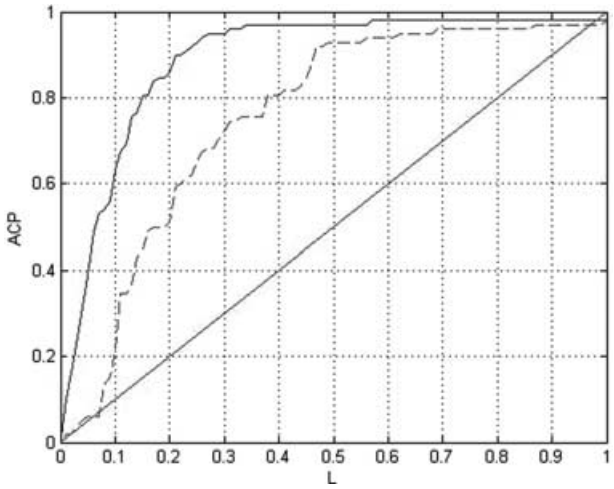

Figure 8 Estimating ACP within a subregion of ESR1. Dashed line is for tSNPs only, solid line is for tSNPs + tagged SNPs

for disease mutation position, x, and number of haplotype clusters (Figure 9). This is included only to provide a concrete example of how tagged SNPs can aid mapping. Of course incorporating tagged SNPs does not always improve localisation. It is the ACP curve that provides us with the aggregated measure of performance. It is clear that fine mapping in this region can benefit from incorporating haplotype information on tagged SNPs. Across the 100 datasets, the estimates of mean bias were 0.14 when using tSNPs only, and 0.08 when using tSNPs and tagged SNPs. The estimates of the mean variance were 1.2×10^{-3} and 1.2×10^{-4}, respectively, and the estimates of Q were 0.23 and 0.10, respectively. Including tagged SNPs improved accuracy of point estimation and yielded a more peaked posterior distribution for x.

Discussion

We have described a statistical method for estimating the position of a disease mutation within a fixed genomic region, which supplements tSNP information with (external) information on linkage disequilibrium. We have demonstrated that there is significant gain to be made by incorporating haplotype information on non-tagging SNPs. The performance of our method, which incorporates tagged SNPs, relative to the method based only on tSNPs is heavily dependent on pattern of LD. The absolute performance of our proposed method is heavily dependent on a number of factors (LD pattern, allele frequencies, penetrance etc). For this reason it is difficult to quantify exactly how well our method performs in specific situations. For mapping our simulation studies show that, given the region size and averaging over different positions for

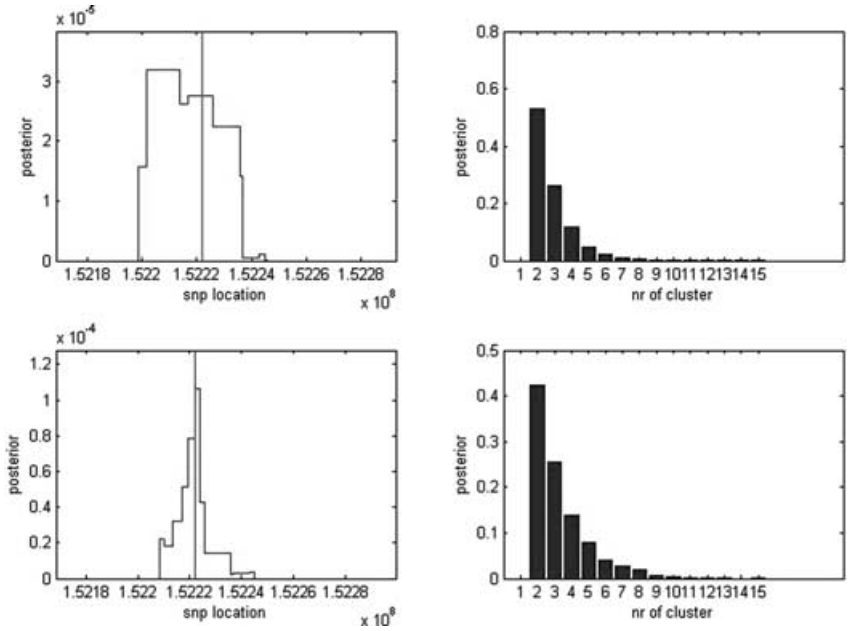

Figure 9 Example of posterior distributions for location of disease mutation (left) and for number of clusters (right), for a simulated dataset based on a subregion of ESR1. The vertical line shows position of the disease mutation. Upper plots are for tSNPs only, lower plots for tSNPs + tagged SNPs.

functional mutations, the method suggests regions which have good ACPs. But how to set the region size in a specific setting, in order to achieve a specific ACP, is an open question. Whilst our haplotype clustering method is usable in its current form, there are several ways in which the method could be further developed. In our simulation studies haplotype phase was known. One obvious and important extension to our method would be for dealing with uncertain haplotype phase. Although tSNPs often predict haplotypes (across tSNPs) well, there will be some loss in power and ability to localise disease variants (Molitor et al. 2005). Our algorithm would also benefit from allowing the number of flanking SNPs to vary locally, according to LD/haplotype diveristy; see Browning (2006) for a related LD mapping approach. This can be accomplished by incorporating a window size parameter in the algorithm. Throughout this article we have assumed a single disease predisposing allele within a fixed/studied region. Our method, of course, performs best in this situation. The extent of allelic heterogeneity at complex disease loci is still unknown, although some predictions have been made (Pritchard & Cox, 2002). Molitor et al. investigated the effect of allelic heterogeneity on mapping performance. They simulated samples assuming two functional mutations, and found that their algorithm successfully managed to track both loci. We don't expect our slightly modified algorithm to behave differently in this matter. We have demonstrated the performance of our method in a setting where identification of tSNPs and LD information on tagged SNPs are based on a small, preliminary sample. It is, of course, common to select tSNPs using Hapmap. In such a setting it is natural to obtain the external LD information from Hapmap as well. Hapmap, however, contains a mixture of several populations, and it is not obvious how well this mixture can help us to predict the LD structure in the study population at hand. The impact of using Hapmap is an aim of future work.

Software implemented in MATLAB can be obtained from the first author.

Acknowledgements

We would like to thank colleagues working on the candidate gene breast cancer study, at the Karolinska Institute and the Genome Institute of Singapore, for comments and data. A. Sjölander gratefully acknowledges financial support from the Swedish Research Council (621-2004-3940) and the Swedish Foundation for Strategic Research (A302:129).

References

Balding et al. (2001) *Handbook of Statistical Genetics*. John Wiley & Sons, Ltd.

Barrett, J.C. & Cardon, L.R. (2006) Evaluating coverage of genomewide association studies. *Nature Genetics* **38**, 659–662.

Browning, S.R. (2006) Multilocus association mapping using variable-length Markov chains. *Am J Hum Genet* **78**. 903–913.

Chapman, J.M., Cooper, J.D., Todd, J.A. & Clayton, D.G. (2003) Detecting disease associations due to linkage disequilibrium using haplotype tags: a class of tests and the determinants of statistical power. *Hum Hered* **56**, 18–31.

Clayton, D. (2000) Linkage disequilibrium mapping of disease susceptibility genes in human populations. *International Statistical Review* **68**, 23–43.

Clayton, D. (2001) Population Association. In: Balding, D.J., Bishop, M., Cannings, C., editors. Handbook of Statistical Genetics: John Wiley & Sons, Ltd.

Clayton, D., Chapman, J. & Cooper, J. (2004) The use of unphased multilocus genotype data in indirect association studies. *Genet Epidemiology* **27**, 415–428.

Einarsdottir, K., Humphreys, K., Bonnard, C., Palmgren, J., Iles, M.M., Sjölander, A., Li, Y., Chia, K.S., Liu, E.T., Hall, P., Liu, J. & Wedren, S. (2005) Comprehensive Assessment of the Association between CHEK2 Haplotypes and Breast Cancer Risk. *PLoS Medicine* **3**, e168.

Gelman, A., Carlin, J.B., Stern, H.S. & Rubin, D.B. (2004) *Bayesian Data Analysis*. Chapman & Hall/CRC.

Gibbs, R.A., Belmont, J.W., Hardenpol, P. et al. (2006). The International HapMap project. *Nature* **426**, 789–796.

Green, P. (1995) Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* **82**, 711–732.

Haiman, C.A., Stram, D.O., Pike, M.C., Kolonel, L.N., Buritt, N.P., Altshuler, D., Hirschhorn, J. & Henderson, B.E. (2003) A comprehensive haplotype analysis of CYP19 and breast cancer risk: the Multiethnic Cohort. *Hum Mol Genet* **12**, 2679–2692.

Hudson, R.R. (2002) Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics* **18**, 337–338.

Molitor, J., Marjoram, P. & Thomas, D. (2003) Fine scale mapping of disease genes with multiple mutations via spatial clustering techniques. *Am J Hum Genet* **73**, 1368–1384.

Molitor, J., Zhao, K. & Marjoram, P. (2005) Fine mapping – 19th Century style. *BMC Genetics* **6**, S63.

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

Phillips, M.S., Lawrence, R. & Sachidanandam, R. et al. (2003) Chromosome-wide distribution of haplotype blocks and the role of recombination hot spots. *Nature Genetics* **33**, 382–387.

Pritchard, J.K. & Cox, N.J. (2002) The allelic architecture of human disease genes: common disease-common variant or not? *Hum Mol Genet*, **11:** 2417–2423.

Prentice, R.L. & Pyke, R. (1979) Logistic disease incidence models and case-control studies. *Biometrika*. **66**, 403–411.

Risch, N., Merikangas, K. (1996) The future of genetic studies of complex diseases. *Science* **273**, 1516–1517.

Schaid, D., Rowland, C., Tines, D., Jacobson, R. & Poland, G. (2002) Score tests for association between traits and haplotypes when linkage phase is ambiguous. *Am J Hum Genet* **70**, 425–434.

Stram, D.O., Haiman, C.A., Hirschhorn, J.N., Altshuler, D., Kolonel, L.N. et al. (2003a) Choosing haplotype-tagging SNPS based on unphased genotype data using a preliminary sample of unrelated subjects with an example from the Multiethnic Cohort Study. *Hum Hered* **55**, 27–36.

Stram, D., Pearce, C.L., Bretsky, P., Freedman, M., Hirschhorn, J.N., Altshuler, D., Kolonel, L.N., Henderson, B.E. & Thomas, D.C. (2003b) Modeling and E-M Estimation of Haplotype-Specific Relative Risks from Genotype Data for a Case-Control Study of Unrelated Individuals. *Hum Hered* **55**, 179–190.

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

Thomas, D., Xie, R. & Gebregziabher, M., (2004) Two-Stage sampling designs for gene association studies. *Genetic Epidemiology* **27**, 401–414.

Zöllner, S. & Pritchard, J.K., (2005) Coalescent-based association mapping and fine mapping of complex trait loci. *Genetics* **169**, 1071-1092.

Supplementary Material

The following supplementary material is available for this article:

Appendix S1. The Mcmc Algorithm for Fine Mapping.

This material is available as part of the online article from: http://www.blackwell-synergy.com/doi/abs/10.1111/j.1469-1809.2007.00369.x
(This link will take you to the article abstract).

Please note: Blackwell Publishing are not responsible for the content or functionality of any supplementary materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.

Received: 13 January 2007
Accepted: 11May 2007