
Mathematical Statistics
Stockholm University

# Utilizing Identity-By-Descent Probabilities for Genetic Fine-mapping in Population Based Samples, via Spatial Smoothing of Haplotype Effects

Linda Hartman, Keith Humphreys and Ola Hössjer

Research Report 2007:15

ISSN 1650-0377

## Postal address:

Mathematical Statistics
Dept. of Mathematics
Stockholm University
SE-106 91 Stockholm
Sweden

## Internet:

http://www.math.su.se/matstat

Mathematical Statistics
Stockholm University
Research Report **2007:15**,
http://www.math.su.se/matstat

# Utilizing Identity-By-Descent Probabilities for Genetic Fine-mapping in Population Based Samples, via Spatial Smoothing of Haplotype Effects

Linda Hartman, Keith Humphreys and Ola Hössjer

20071122

**Abstract**

Genetic fine mapping can be performed by exploiting the notion that haplotypes that are structurally similar in the neighbourhood of a disease predisposing locus are more likely to harbour the same susceptibility allele. Within the framework of Generalized Linear Mixed Models this can be formalized using spatial smoothing models, i.e. inducing a covariance structure for the haplotype risk parameters, such that risks associated with structurally similar haplotypes are dependent. In a Bayesian procedure a local similarity measure is calculated for each update of the presumed disease locus. Thus, the disease locus is searched as the place where the similarity structure produces risk parameters that can best discriminate between cases and controls.

We describe an approach which takes a population genetics perspective to theoretically motivate the use of an identity-by-descent based similarity metric. We compare this approach to other more intuitively motivated models and similarity measures based on identity-by-state, suggested in the literature.

KEY WORDS: genetic association analysis, spatial smoothing, Generalized Linear Mixed Model, population genetics, IBD, IBS, SNP

# 1 Introduction

In genetic studies with dense markers, the dependence between markers due to high linkage disequilibrium has simultaneously been both the major tool for fine-mapping as well as an obstacle for the analysis. Analyses based on haplotypes, i.e. a collection of alleles at closely linked loci on the same chromosome, utilize the increased polymorphism obtained when simultaneously studying several loci. The biological importance of haplotypes is at least twofold. Firstly, proteins consist of a linear sequence of amino acids, which is read off from the DNA content on a chromatid, as captured in the haplotypes. Thus haplotypes may capture the interaction of several *cis*-acting susceptibility variants found within the gene, that can be hard to detect when markers are studied one at a time. Secondly, and what is important in this article, the genetic variation in the population is inherently structured in haplotypes, and thus the haplotype structure mirrors the population history of genetic drift, recombination, mutation, selection etc. A review of conditions for when haplotype based analyses are more powerful than analyses based

on single markers, or than multi-locus analyses without regard to haplotype phase is found in Schaid (2004).
We will in this article concentrate on case-control studies, where information on haplotypes is collected. The methods and algorithms we present are intended for diseases with moderate genetic effects, but with a caution that the algorithms can be prone to numerical difficulties if the number of sampled individuals is large.
To use haplotypes for gene mapping, we will utilize a *local* similarity measure, that is calculated for each pair of haplotypes at putative disease loci. The disease locus is searched for as the chromosomal position where the local similarity measure best discriminates between cases and controls. Just as in Molitor et al. (2003a) we use a likelihood based on spatial smoothing to determine where this discrimination is optimal. The term spatial is used here to refer to a multidimensional space, on which a distance metric is constructed, such that similar haplotypes are separated by a short distance. This in turn implies a large dependence, just as in two-dimensional spatial statistics. A spatial smoothing model based on a local similarity metric directly exploits the increased polymorphism, obtained when several linked loci are studied together, for the purpose of fine mapping, and simultaneously handles the problems that might arise if there are many rare haplotypes. Molitor et al. (2003a) used a spatial smoothing model with a conditional auto-regression (CAR) formulation, where the weights in the auto-regression are determined by the similarity measure based on alleles shared identical by state(IBS). Other spatial models suggested in the literature include Bayesian clustering (Molitor et al., 2003b; Waldron et al., 2006) or cladistic analysis (Durrant et al., 2004; Durrant and Morris, 2005). These models also use IBS measures for defining haplotype similarity.
Using spatial models for genetic mapping is thus not a new idea, but as pointed out in Schaid (2004), further research is required to construct a dependence structure allowing for covariances determined by shared ancestry. In this article we suggest a similarity metric based on identity-by-descent (IBD) instead of IBS relationships. In population based samples (such as the type we focus on), no pedigree data is available, but information on IBD sharing is captured solely from multi-marker data. Utilizing the ideas and likelihood ratio calculations of Hartman and Hössjer (2007), we show how haplotype IBD probabilities can be calculated either strictly pairwise, ignoring all other haplotypes in the sample, or pairwise but utilizing the full haplotype sample. In general the IBD probabilities must be calculated through simulation. Under a simplified model based on a star-topology for cases the IBD probabilities can be calculated analytically, still allowing for mutation, recombination and varying allele frequency. Alternative calculations

of strictly pairwise IBD probabilities used for QTL for unrelated individuals in animal genetics can be found in Meuwissen and Goddard (2001), with a recent extension to longer haplotypes in Meuwissen and Goddard (2007).
In this article we show how to use a CAR proposal for risk parameters, together with an IBD metric. We also suggest a spatial model that we derive from a population genetic perspective – this places an even stronger emphasis on dependences being due to shared ancestry than the CAR model with IBD metric. In this latter model it turns out that the covariance matrix for the haplotype risk parameters consist of IBD probabilities for each pair of haplotypes.

# 2 Spatial smoothing models for gene mapping

The idea of spatial smoothing models for gene mapping is to use a model where a covariance structure is imposed on the haplotype risk parameters, such that risk parameters corresponding to haplotypes with a high structural similarity get assigned high dependence. By defining a *local* similarity metric, that is calculated around a putative disease locus $x$, and letting this putative disease locus be updated in the estimation procedure, the methods are well suited for mapping of disease genes. Thus, the disease locus is searched as the place where the similarity structure produces risk parameters that can best discriminate between cases and controls. The general idea of spatial smoothing for gene mapping was introduced in Thomas et al. (2001) and has been elaborated on in Molitor et al. (2003a), whose methodology we follow to large extent. The main differences between our work and Molitor's are that we use a covariance matrix based on population genetic reasoning, see Section 2.3, and a less ad hoc similarity metric based on Identity-by-descent, see Section 3.1.

## 2.1 Notation and Bayesian framework

Assume data that consist of observed phenotypes $y_v$, and single nucleotide polymorphism (SNP) marker genotypes $\boldsymbol{g}_v$, $v = 1, \ldots, m$, of $m$ investigated individuals. The genotypes are measured at $K$ marker loci, i.e. $\boldsymbol{g}_v = (g_{v1}, \ldots, g_{vK})$. The region of interest is normalized as a unit interval $[0, 1]$ in terms of genetic or physical map distance, with marker positions $0 \leq x_1 < x_2 < \ldots < x_K \leq 1$. We assume that the genotype phase, i.e. which alleles belong to the same chromosome, is known. Thus, for each individual the genotype can be separated into two *haplotypes*, $\boldsymbol{g}_v = (\boldsymbol{h}_{2v-1}, \boldsymbol{h}_{2v})$. Each haplotype $\boldsymbol{h}_i$ thus comprises $K$ markers, $\boldsymbol{h}_i = (h_{i1}, \ldots, h_{iK})$, where the two

possible alleles for each $h_{ik}$ are coded as 0 and 1.
We consider models where each of a person's two haplotypes contribute additively to the total risk, i.e.

$$g(P(Y_v = 1)) = \mu + b_{2v-1} + b_{2v}, \tag{1}$$

where $g$ is a link function, typically a logit link, i.e $g(p) = \ln(p/(1-p))$.
Written in vector form we obtain

$$g(E(\boldsymbol{Y})) = \mu\mathbf{1} + \boldsymbol{b}\,\boldsymbol{Z}, \tag{2}$$

where $\boldsymbol{Y} = (Y_1, \ldots, Y_m)$ is a random vector of (binary) phenotypes, $\mathbf{1} = (1, \ldots, 1)$, $\boldsymbol{b} = (b_1, \ldots, b_{2m})$ and $\boldsymbol{Z} = (Z_{iv})$ is a $2m \times m$ design matrix with non-zero elements $Z_{2v-1,v} = Z_{2v,v} = 1$. More generally, environmental covariates can be incorporated into (1) – (2), by letting $\mu$ depend on $v$. For haplotype data $\boldsymbol{h} = (\boldsymbol{h}_1, \ldots, \boldsymbol{h}_{2m})$, and Gaussian risk parameters,

$$\boldsymbol{b}|\boldsymbol{h} \sim N(\mathbf{0},\, \sigma_b^2\boldsymbol{\Sigma}), \tag{3}$$

(2) becomes a *Generalized Linear Mixed Model (GLMM)*, (McCulloch and Searle, 2001; Breslow and Clayton, 1993).
The covariance matrix $\sigma_b^2\boldsymbol{\Sigma}$, is specifed to depend on the putative disease locus $x$, and is defined such that structurally similar haplotypes get positively dependent parameter values. Introducing a covariance structure based on structural similarity is a way to mimic the notion that haplotypes that are structurally similar in the neighbourhood of a disease predisposing locus are more likely to harbour the same susceptibility allele and hence should have similar risks. Further, this approach implicitly deals with rare haplotypes in an elegant manner.
We adopt a Bayesian approach, with unknown parameter vector $(\mu, \boldsymbol{b}, x, \sigma_b, \boldsymbol{\xi})$, where $\boldsymbol{\xi}$ contains the parameters used in the calculation of the local similarity metric. For $\boldsymbol{y} = (y_1, \ldots, y_m)$, i.e. the observed value of $\boldsymbol{Y}$, the joint posterior distribution is

$$\pi(\mu, \boldsymbol{b}, x, \boldsymbol{\xi}, \sigma_b|\boldsymbol{y}, \boldsymbol{h}) \propto \pi(\boldsymbol{y}|\mu, \boldsymbol{b})\pi(\mu)\pi(\boldsymbol{b}|\boldsymbol{\Sigma}(x, \boldsymbol{\xi}, \boldsymbol{h}), \sigma_b)\pi(x)\pi(\boldsymbol{\xi})\pi(\sigma_b), \tag{4}$$

where $\pi(\boldsymbol{y}|\mu, \boldsymbol{b})$ and $\pi(\boldsymbol{b}|\boldsymbol{\Sigma}(x, \boldsymbol{\xi}, \boldsymbol{h}), \sigma_b)$ are defined in (1)–(2) and (3) respectively. Figure 1 contains a graphical interpretation of the model as a directed acyclic graph (*DAG*). To estimate the model parameters, of which $x$ is of particular interest, the parameters are updated in an MCMC algorithm, which we describe in detail in Appendix A.

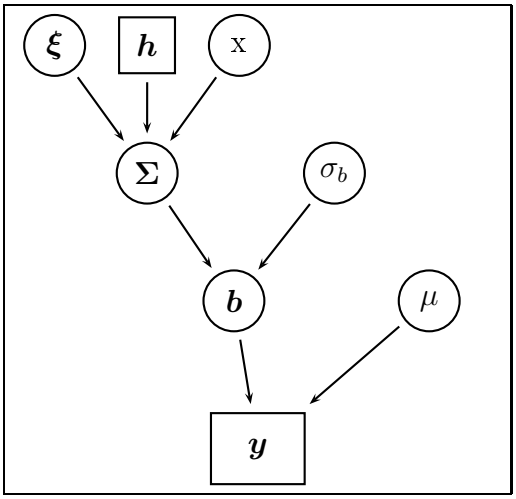


Figure 1: DAG for model with known phase. Circles represent parameters whereas squares represent observed quantities.

## 2.2 Conditional auto regression (CAR) model

A popular way to model dependence between parameters, is through a *Conditional AutoRegression (CAR)* model. A CAR model for spatial smoothing was first implemented for gene mapping by Molitor et al. (2003a).
The model is defined through its conditional distributions:

$$b_i|\boldsymbol{b}_{(-i)}, \sigma_b^2 \in N(\bar{b}_i,\, \sigma_b^2/\sum_{j\neq i} w_{ij}), \tag{5}$$

where $w_{ij}$ quantifies similarity between haplotypes $i$ and $j$, $\boldsymbol{b}_{(-i)}$ denotes for the vector of all $b_j$s except $b_i$, and $\bar{b}_i$ is the average of the other risk parameters, weighted by their similarity with $i$, i.e. $\bar{b}_i = \sum_{j\neq i} w_{ij} b_j / \sum_{j\neq i} w_{ij}$. From a similarity matrix $\boldsymbol{W} = (w_{ij})_{i,j=1}^{2m}$ (with $w_{ii}$ set to 0), let $\boldsymbol{M}^{-1} = \mathrm{diag}(\sum_{j\neq i} w_{ij})$ and $\boldsymbol{C} = \mathrm{diag}(1/\sum_{j\neq i} w_{ij})\boldsymbol{W}$. For a given $\sigma_b^2$, which here denotes the *conditional variance*, the joint distribution (3) then has precision (i.e. inverse covariance) matrix $\sigma_b^{-2}\boldsymbol{Q} = \sigma_b^{-2}\boldsymbol{\Sigma}^{-1}$ where

$$\boldsymbol{Q} = \boldsymbol{M}^{-1}(\boldsymbol{I}_{2m} - \boldsymbol{C}). \tag{6}$$

In many practical applications, it is not ensured that the CAR definition leads to proper distributions for $\boldsymbol{b}$, but since the covariance matrix is updated in the MCMC procedure, it is important that it does so here. By definition, (6) leads to an intrinsic distribution, but addition of (a small) $\varepsilon > 0$ on the diagonal of $\boldsymbol{Q}$, makes $\boldsymbol{Q}$ diagonal dominant, and thus positive definite, as

long as $w_{ij}$ are all non-negative. The CAR model thus allows great flexibility in defining the similarity metric. The resulting distribution has covariance matrix $\sigma_b^2 \boldsymbol{\Sigma} = \sigma_b^2 \left( \boldsymbol{M}^{-1}(\boldsymbol{I}_{2m} - \boldsymbol{C}) + \varepsilon \boldsymbol{I}_{2m} \right)^{-1}$.

## 2.3 Population genetic model

A CAR model produces spatially smoothed parameters, where each haplotype risk is a weighted mean of the other haplotype risks. The weights are determined by the similarities between the haplotypes. The CAR model is not defined from a population genetic perpective, but is an "all-purpose" algorithm to obtain correlated parameters.
As an alternative, assume that the studied sample of today's chromosomes have been inherited from $N'$ founder haplotypes $\boldsymbol{h}'_1, \ldots, \boldsymbol{h}'_{N'}$, and that a disease gene existed in one of the founder chromosomes, $\boldsymbol{h}'_{I_{\text{mut}}}$, at position $x$ in the studied interval. The material of $\boldsymbol{h}'_{I_{\text{mut}}}$ at and around $x$ has since then segregated down to some of today's haplotypes, through the process of recombination and mutation, causing an increased risk of disease in carriers of those haplotypes. The disease risk of each of today's haplotypes, is thus determined by which founder haplotype has segregated its material at the disease locus $x$. Let $b'_I$ be the random penetrance effect of founder chromosome $I$. One possibility is to let $b'_I$ be binary, with a value depending on whether $I$ is mutated or not, see Appendix B for further details.
However, following a traditional simplification in quantitative genetics, we assume for the Bayesian analysis (4) that

$$b'_I \sim N(0, \sigma_b^2) \text{ are i.i.d., } I = 1, \ldots, N' \tag{7}$$

Let $I_i$ denote the founder chromosome that is passed to Chromosome $i$ of today's generation, so that $b_i = b'_{I_i}$ for $i = 1, \ldots, 2m$. We assume that

$$\boldsymbol{b}' = (b'_I)_{I=1}^{N'} \text{ is independent of } (\boldsymbol{I}, \boldsymbol{h}), \tag{8}$$

where $\boldsymbol{I} = (I_1, \ldots, I_{2m})$. Formula (8) corresponds to no genetic drift of the disease allele frequency (independence of $\boldsymbol{b}'$ and $\boldsymbol{I}$), that $x$ is not a marker locus and that the founder generation is in linkage equilibrium (see Appendix B for a motivation when $b'_I$ is binary). It implies

$$\begin{aligned}
\text{Cov}(b_i, b_j|\boldsymbol{h}) &= \text{Cov}(b'_{I_i}, b'_{I_j}|\boldsymbol{h}) \\
&= \text{E}\left(\text{Cov}(b'_{I_i}, b'_{I_j}|\boldsymbol{I}, \boldsymbol{h})\right) + \text{Cov}\left(\text{E}(b'_{I_i}|\boldsymbol{I}, \boldsymbol{h}), \text{E}(b'_{I_j}|\boldsymbol{I}, \boldsymbol{h})\right) \\
&= \sigma_b^2 P(I_i = I_j|\boldsymbol{h}) + 0 \\
&= \sigma_b^2 P(i, j \text{ IBD at } x|\boldsymbol{h}),
\end{aligned} \tag{9}$$

where the inner expectation/covariance is with respect to $\boldsymbol{b}'$ and the outer one with respect to $\boldsymbol{I}$. Since all $b'_I$ are Gaussian, the marginal distributions of $\boldsymbol{b}|\boldsymbol{h}$ is $N(0, \sigma_b^2)$. We simplify further and assume that $\boldsymbol{b}|\boldsymbol{h}$ is multivariate Gaussian with covariance matrix as in (9). This implies that (3) holds with $\boldsymbol{\Sigma} = (P(i, j \text{ IBD at } x|\boldsymbol{h}))_{i,j=1}^{2m}$.
As an approximation of (9),

$$\text{Cov}(b_i, b_j|\boldsymbol{h}_i, \boldsymbol{h}_j) = P(i, j \text{ IBD at } x|\boldsymbol{h}_i, \boldsymbol{h}_j), \tag{10}$$

could be used. As noted in the discussion, it is then important to confirm that the approximate covariance matrix is still positive definite.
The idea of the spatial models was that structurally similar haplotypes should have dependent risk parameters, as they are likely to harbour the same susceptibility gene. The above population genetic argument shows that if similarities are measured as IBD probabilities, these enter directly as the elements of the covariance matrix. Hence it is intuitive to specify $\boldsymbol{\Sigma} = \boldsymbol{W} = (w_{ij})_{i,j=1}^{2m}$ with similarity metrics

$$w_{ij} = P(i, j \text{ IBD at } x|\boldsymbol{h}) \tag{11}$$

and

$$w_{ij} = P(i, j \text{ IBD at } x|\boldsymbol{h}_i, \boldsymbol{h}_j), \tag{12}$$

which we calculate in Section 3.1.2 and 3.1.1, respectively.
For QTL linkage analysis with an additive genetic variance component, (2), (3) and (9) could be combined, with a linear link function $g$. The major difference from what we present here is that $\boldsymbol{h}$ has to be redefined, to contain not only marker data but also the known pedigree structure of all $m$ family members see Amos (1994), Almasy and Blangero (1998) and Sham et al. (2002). See also Meuwissen et al. (2002), where the approximation (10) is used for association analysis in animal genetics.

# 3 Haplotype similarity measures

As described in the previous section, a local similarity metric $w_{ij}$ is calculated between all pairs of haplotypes $i$ and $j$ for each putative disease locus $x$. Whereas the population genetic model of Section 2.3 suggests a very specific similarity matrix $\boldsymbol{W} = (w_{ij})_{i,j=1}^{2m}$ based on (11)(or its approximation (12)), the CAR model of Section 2.2 allows a wide range of similarity matrices.

## 3.1 Similarity measures based on IBD

The IBD similarities (11)–(12) directly estimate the probability that two haplotypes are descendants from the same founder, and thus share the same susceptibility allele. IBD is at the basis of many gene mapping strategies, such as linkage analysis. IBD calculations typically rely on marker data from several generations from families with known pedigree structures. In population based studies, such that we consider here, only marker data of the present generation are available, so pedigree structure is unknown. Meuwissen and Goddard (2001) estimate pairwise IBD-probabilities from marker data in a region with strong LD, by modelling the coalescence of all markers, for one pair of chromosomes at a time. The algorithms presume known phase, that haplotypes were randomly sampled and that no mutations have occurred, and require known effective population size and time since the most recent mutation. Recently Meuwissen and Goddard (2007) published a new algorithm allowing for mutations, which estimates the effective population size as part of the algorithm. Due to computational constraints, haplotypes are only compared pairwise, i.e. the coalescence trees are not built simultaneously for all haplotypes in the sample.
In Hartman and Hössjer (2007) an *Ancestral Recombination Graph (ARG)*, for retrospectively sampled data was developed, and a likelihood ratio (LR) test for gene mapping suggested. Under general population genetic conditions, extensive simulations of the ARG must be performed in order to estimate the likelihood ratios. However, assuming linkage equilibrium (LE) in founder haplotypes, different founders for unmutated chromosomes, and a star topology for the ancestral tree of the mutated chromosomes, analytic expressions for the LR-test can be obtained.
In the following we will use the approach of Hartman and Hössjer (2007) to calculate IBD-probabilities, both pairwise, i.e. only accounting for the two haplotypes in the pair, and pairwise but conditional on the full sample. The genealogical model used for the IBD calculations assumes that all or many of the diseased individuals carry the same disease allele (which is presumed to be rare) together with a small chromosome segment from the founder.

### 3.1.1 Pairwise IBD calculations

Strictly pairwise IBD calculations imply utilizing only the two haplotypes $\boldsymbol{h}_i$ and $\boldsymbol{h}_j$ in the calculation of the IBD-probability, just as in the proposed IBS similarity metrics of Section 3.2 or the IBD similarity metric of Meuwissen and Goddard (2007). Let

$$\alpha = P(i, j \text{ IBD at } x) \tag{13}$$

be the prior probability that two arbitrary chromosomes in the sample are IBD. Under a star topology this is identical to the probability that both chromosomes carry the mutated disease chromosome. The value of $\alpha$ is affected by the ascertainment scheme, which is an issue we discuss further in Section 3.1.3.
The pairwise IBD probabilities can be calculated as

$$\begin{aligned} w_{ij} &= P(i,j \text{ IBD at } x|\boldsymbol{h}_i,\boldsymbol{h}_j) \\ &= \frac{P(\boldsymbol{h}_i,\boldsymbol{h}_j|i,j \text{ IBD at } x)\alpha}{P(\boldsymbol{h}_i,\boldsymbol{h}_j|i,j \text{ IBD at } x)\alpha + P(\boldsymbol{h}_i,\boldsymbol{h}_j|i,j \text{ not IBD at } x)(1-\alpha)} \\ &= \frac{\text{LR}_{ij}}{\text{LR}_{ij} + (1-\alpha)/\alpha}, \end{aligned} \tag{14}$$

where

$$\text{LR}_{ij} = \frac{P(\boldsymbol{h}_i,\boldsymbol{h}_j|i,j \text{ IBD at } x)}{P(\boldsymbol{h}_i,\boldsymbol{h}_j|i,j \text{ not IBD at } x)}.$$

Analytical expressions for the likelihood ratio $L_{ij}$ can be found in Appendix C.

### 3.1.2 IBD calculations conditional on the full sample

Using the ARG of Hartman and Hössjer (2007) under the star-topology we can however not only calculate IBD probabilities pairwise, but also conditional on the full haplotype sample $\boldsymbol{h}$.
To this end, define $\boldsymbol{C} = (C_1, ..., C_{2m})$, where $C_i = 1$ if chromosome $i$ is mutated at $x$ and $C_i = 0$ otherwise. Further let $\boldsymbol{C}_{ij} = \{\boldsymbol{C}; C_i = C_j = 1\}$ represent all $\boldsymbol{C}$ for which $i$ and $j$ are IBD at $x$ (according to the star topology). Then

$$\begin{aligned} w_{ij} &= P(i,j \text{ IBD at } x|\boldsymbol{h}) \\ &= P(C \in \boldsymbol{C}_{ij}|\boldsymbol{h}) \\ &= P(\boldsymbol{h}|C \in \boldsymbol{C}_{ij})P(C \in \boldsymbol{C}_{ij})/P(\boldsymbol{h}) \\ &= \frac{\alpha P(\boldsymbol{h}|C \in \boldsymbol{C}_{ij})/P_0(\boldsymbol{h})}{P(\boldsymbol{h})/P_0(\boldsymbol{h})}, \end{aligned} \tag{15}$$

$P(C \in \boldsymbol{C}_{ij}) = P(i,j \text{ IBD at } x) = \alpha$. $P_0(\boldsymbol{h})$ is the probability of $\boldsymbol{h}$ under the null hypothesis of no disease locus, thus $P_0(\boldsymbol{h}) = \prod_{q=1}^{2m}\prod_{k=1}^{K}\tilde{f}(h_{qk})$, where $\tilde{f}(\cdot)$ are the allele frequencies of today's generation.
In Appendix D , $\widetilde{\text{LR}} = P(\boldsymbol{h})/P_0(\boldsymbol{h})$ and $\widetilde{\text{LR}}_{ij} = P(\boldsymbol{h}|C \in \boldsymbol{C}_{ij})/P_0(\boldsymbol{h})$ are calculated by summation over all possible mutated founder haplotypes $\boldsymbol{h}'$.

### 3.1.3 Choice of $\alpha$

The prior probability that two haplotypes in the sample are IBD, $\alpha$, is central to the calculations of IBD probabilities. Under random sampling $\alpha = p^2$, where $p$ is the disease allele frequency. Under non-random sampling, such as a case-control study, $\alpha$ can be greatly increased.
We have used a simple approach to address the non-random sampling, which incorporates information on the number of cases and controls (together with penetrance and disease allele frequency).
For $m_0$ controls and $m_1 = m - m_0$ cases, we use the Empirical Bayes choice $\alpha = P(i, j \text{ IBD}|m_0, m_1)$ of prior probability in (13). The IBD similarity metrics in (14) and (15) should more correctly be written as $w_{ij} = P(i, j \text{ IBD at } x|\boldsymbol{h}_i, \boldsymbol{h}_j, m_0, m_1)$ and $w_{ij} = P(i, j \text{ IBD at } x|\boldsymbol{h}, m_0, m_1)$. To find a general expression for $\alpha$ we let $\psi_0, \psi_1$ and $\psi_2$ denote penetrance parameters, i.e. the probabilities of being affected when having 0, 1 and 2 disease alleles respectively. Assume $i \in (2v-1, 2v)$, is one of the two chromosomes of Individual $v$. Under Hardy-Weinberg equilibrium we define

$$p_{\text{ctrl}} = P(i \text{ mutated}|Y_v = 0) = \frac{p^2(1-\psi_2) + p(1-p)(1-\psi_1)}{1-S}$$

and

$$p_{\text{case}} = P(i \text{ mutated}|Y_v = 1) = \frac{p^2\psi_2 + p(1-p)\psi_1}{S},$$

where $S = \psi_0(1-p)^2 + \psi_1 2p(1-p) + \psi_2 p^2$ is the prevalence. Then, assuming $i$ and $j$ are drawn randomly from a pool of cases of controls of relative sizes $m_1/m$ and $m_0/m$, we get

$$\alpha = \left(\frac{m_0}{m}\right)^2 p_{\text{ctrl}}^2 + 2\frac{m_0}{m}\frac{m_1}{m}p_{\text{ctrl}}p_{\text{case}} + \left(\frac{m_1}{m}\right)^2 p_{\text{case}}^2. \tag{16}$$

Notice that for a recessive disease we obtain $\alpha = (m_1/m)^2$, in the limit when $p \to 0$.
Another approach for handling non-random sampling would be to include $\alpha$ in the MCMC algorithm, e.g. by assigning a prior on $[0, 1]$, and updating $\alpha$ in each step of the chain. For $w_{ij} = P(i, j \text{ IBD}|\boldsymbol{h}_i, \boldsymbol{h}_j)$, this is possible since $\alpha$ enters only in the quotient of likelihood ratios in (14). Thus the computationally demanding likelihood ratios could still be calculated only once, at the beginning of the algorithm. We tried out this approach in preliminary analyses. Updating $\alpha$ in the MCMC algorithm didn't improve our results, so we let $\alpha$ be fixed in our final MCMC algorithm. For $w_{ij} = P(i, j \text{ IBD}|\boldsymbol{h})$, $\alpha$ enters already in the prior probability $P(C_i)$ that a chromosome is mutated, c.f. (30) and (31) in the appendix. Thus each update of $\alpha$ would require

calculation of new likelihood ratios, which is not computationally feasible except for with very small data sets.

## 3.2 Similarity measures based on IBS

Similarity metrics for spatial smoothing, suggested in the literature on fine-mapping, have typically been based on the notion of *Identity-by-state (IBS)*. The IBS similarity metrics are approximately monotone functions of the probability that two haplotypes are descendants of the same founder at the disease locus, and thus share the properties determined at the disease locus. IBS similarity metrics have been used also for clustering models, where each haplotype is compared with each cluster-centre haplotype, and a clustering is sought that can best discriminate between cases and controls.
As a simple IBS measure Molitor et al. (2003b) suggest the length shared IBS around $x$ between the two haplotypes, in connection with the CAR model. We have implemented their approach to fine mapping for comparison with ours. Although $x$ could vary continuously in the measured region, for computational reasons we have only calculated the IBS (and later the IBD) similarity metrics at a discrete set of locations, chosen as the midpoint of each marker interval. In this case,

$$w_{ij} = R_{ij}(x) - L_{ij}(x) \tag{17}$$

where $R_{ij}(x) = \frac{x_{r'}+x_r}{2}$ where $r$ is the the first marker to the right of $x$ where a difference is encountered between haplotypes $\boldsymbol{h}_i$ and $\boldsymbol{h}_j$, and $r' = r-1$. (If no difference is encountered to the right let $r' = r = K$.) $L_{ij}(x)$ is defined similarly to the left of $x$. If a difference is encountered at the closest marker to the right or left of $x$ (but not both), this algorithm includes half of the interval $[x_{k_0}, x_{k_0+1}]$ in the shared length, where $k_0$ is the closest marker to the left, i.e. $x_{k_0} < x < x_{k_0+1}$. The above algorithm could also be extended to allow for mismatches due to mutations, by letting intervals after an encountered difference be included in the IBS-measure, but assess a penalty to the shared length after such a difference. Waldron et al. (2006) define an alternative IBS-measure that sums up similarity scores over all possible windows around the putative locus, and uses the maximum window score as the overall similarity score. In order to assign higher scores to matches of rare alleles, than matches of more common alleles, and allow for mutation, the SNP similarity score is defined from the odds against a match if the haplotypes are unrelated. Thus the score is

$$\begin{array}{cl} (1-p)/p & \text{if the alleles match} \\ 0 & \text{if any allele is missing} \\ -\gamma p(1-p) & \text{if the alleles do not match,} \end{array}$$

where $\gamma$ is a mismatch penalty parameter and $p$ is allele frequency. Durrant et al. (2004) used a similar score for cladistic analysis, with $1-p$ as matching score and no mismatches allowed for, thus in effect $\gamma = \infty$.

# 4 Simulations

As an example of the performance, we show results from analysing a simulated data set. The simulated data consisted of 11 markers for 100 cases and 100 controls (thus 400 haplotypes in total) and the disease allele frequency was 0.1. In order for this small sample size to be sufficient for detecting association we used a full recessive penetrance model, i.e. $\psi = [0, 0, 1]$. Thus, we tried our gene mapping algorithms for a data set generated from a different (stronger) penetrance model to that assumed in (1). In Appendix B a connection between binary penetrance effects and the logit-model for continuous risk parameters is described.
We simulated the case-control sample, using the retrospective ARG of Hartman and Hössjer (2007), with a star topology. Thus, for each individual the mutational status for each of the two alleles at the disease locus is simulated conditional on the disease status. Each unmutated chromosome has markers in LE, whereas each mutated chromosome carries the alleles of the (simulated) mutated founder, up to a simulated recombination point to the left and right of the disease locus, respectively. Outside of the recombination points marker alleles are in linkage equilibrium. We ran the simulations without neutral mutations; these are otherwise superimposed independently at all chromosomes and markers.
The 11 markers were equidistantly spread, with $x_1 = 0,\ x_2 = 0.1, \ldots, x_{11} = 1$, and we used marker allele frequencies $f_k(0) = f_k(1) = 0.5,\ k = 1, \ldots, 11$. The disease location is at 0.65, i.e. in between markers 7 and 8, and the expected number of recombinations since the founder mutation within the chromosomal region, $\rho = 4$.
We fitted five models to the simulated data

- CAR with IBS similarity
- CAR with IBD (pairwise)
- CAR with IBD (full)
- Population genetic model with IBD (full)
- Population genetic model with IBD (pairwise)

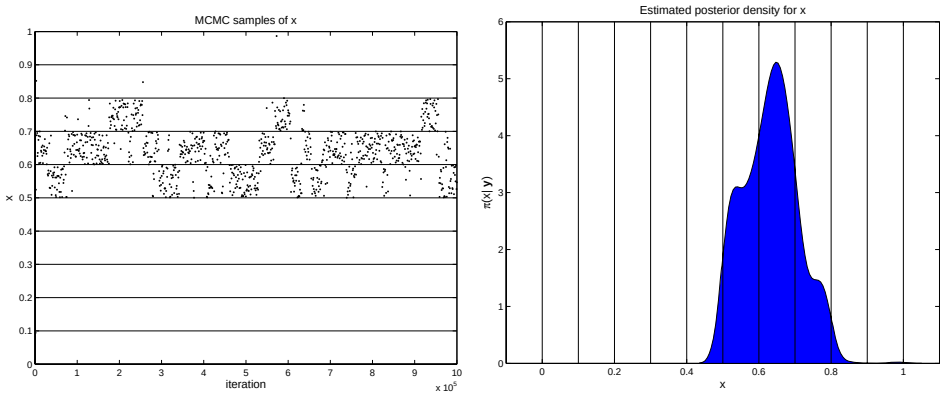


Figure 2: MCMC output and estimated posterior density for disease locus $x$ for a CAR model with IBS similarity metric (17). True disease locus is at 0.65.

Results from fitting these models are displayed in Figures 2–6, respectively. In the gene mapping analyses, marker allele frequencies were estimated from (the simulated) data. To calculate IBD probabilities the correct values of $\rho$, and $\psi = (\psi_0, \psi_1, \psi_2)$ were assumed known. Using (16), with $\psi_0 = \psi_1 = 0$ and $\psi_2 = 1$, we obtained $\alpha = 0.30$. This value of $\alpha$ was used in the IBD similarity metrics, except in Figure 6 where we used $\alpha = 0.05$ to obtain a positive definite covariance matrix, see discussion below. The density estimations to the right in the figures were produced by a normal density kernel smoother with automatic bandwidth (Matlab's function `ksdensity.m`). For all analyses with CAR models we obtained a positive definite covariance matrix by adding a fixed $\varepsilon = 0.0001$ on the diagonal of the precision matrix, see Section 5 for further discussion.

All models gave reasonable results, in terms of their ability to point out the correct marker interval. The IBD based similarity metrics (used in both CAR and the population genetic model) gave for this dataset slightly better results than the IBS based similarity metric (which can be used only with the CAR model).

The analyses were sometimes hampered by numerical difficulties. As the population genetic model of Section 2.3 uses the similarity matrix $\boldsymbol{W}$ directly as covariance matrix, it demands a positive definite covariance matrix. For this data set the approximation $w_{ij} = P(i, j \text{ IBD}|\boldsymbol{h}_i, \boldsymbol{h}_j)$ did not produce a positive definite matrix. In the IBD calculations the parameter $\alpha$ is central,

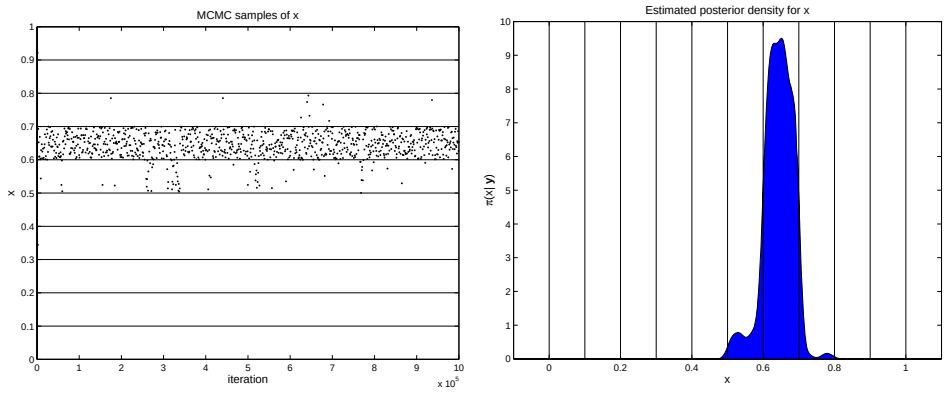


Figure 3: MCMC output and estimated posterior density for disease locus $x$ for a CAR model with IBD similarity metric $w_{ij} = P(i, j \text{ IBD}|\boldsymbol{h}_i, \boldsymbol{h}_j)$. True disease locus is at 0.65.

both when conditioning on the full sample $\boldsymbol{h}$ or just the pair of haplotypes $\boldsymbol{h}_i$, $\boldsymbol{h}_j$. A small $\alpha$ amounts to low prior probability that two haplotypes are IBD, and will thus make the posterior probabilities $w_{ij} = P(i, j \text{ IBD}|\boldsymbol{h})$, or the approximation $w_{ij} = P(i, j \text{ IBD}|\boldsymbol{h}_i, \boldsymbol{h}_j)$, smaller. As the corresponding similarity matrices $\boldsymbol{W}$ will always have unit diagonal (a haplotype is by definition IBD with itself), a smaller $\alpha$ will thus produce a similarity matrix that is further from the limit of singularity. Thus, for the approximate version $w_{ij} = P(i, j \text{ IBD}|\boldsymbol{h}_i, \boldsymbol{h}_j)$ the chance that $\boldsymbol{W}$ will be positive definite increases, if a small $\alpha$ is used. For the population genetic model we have therefore tried analyses with an $\alpha < 0.30$. In Figure 6 we used $\alpha = 0.05$, as this ( but not $\alpha = 0.1$) was small enough (for this particular dataset) to get a positive definite $\boldsymbol{W}$ even for the approximation $w_{ij} = P(i, j \text{ IBD}|\boldsymbol{h}_i, \boldsymbol{h}_j)$.
For some of the analyses where the covariance matrix was positive definite, such as all analyses with the CAR model, numerical difficulties also occured. The seemingly accurate result of Figure 4, is probably also an artefact of bad mixing. When we ran several separate MCMC-chains with this model (CAR model with $w_{ij} = P(i, j \text{ IBD}|\boldsymbol{h})$), the analyses most often gave similar results as that displayed. Sometimes however, and more often for starting values close to $x = 0$, the MCMC got stuck in another marker interval. The seemingly accurate results from this model are thus not reliable for this data set.
For this particular data set all proposed models and similarity metrics gave

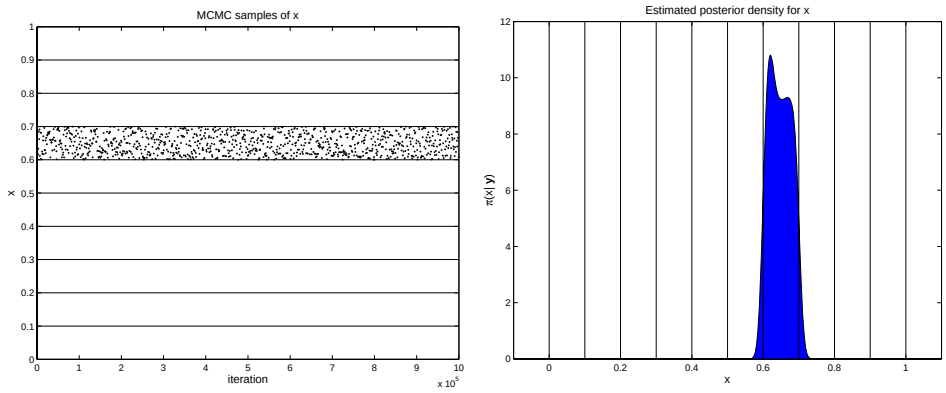


Figure 4: MCMC output and estimated posterior density for disease locus $x$ for a CAR model with IBD similarity metric $w_{ij} = P(i, j \text{ IBD}|\boldsymbol{h})$. True disease locus is at 0.65.

reasonable results. If the CAR model with $w_{ij} = P(i, j \text{ IBD}|\boldsymbol{h})$ (Figure 4) is excluded due to bad mixing, the best results were obtained with $w_{ij} = P(i, j \text{ IBD}|\boldsymbol{h}_i, \boldsymbol{h}_j)$, either in the CAR model (Figure 3) or in the population genetic motivated model (Figure 6). In the latter case $\alpha = 0.05$ was used in order to get $\boldsymbol{\Sigma}$ to be positive definite. The population genetic motivated model with $w_{ij} = P(i, j \text{ IBD}|\boldsymbol{h})$ (Figure 5) also gave better results than the CAR model with IBS based similarity metric (Figure 2) for this data set.
Additional simulations (not shown here) with data sets of different sizes, disease models, recombination rates etc, confirm that a CAR model with $w_{ij} = P(i, j \text{ IBD}|\boldsymbol{h})$ seems to be the model which is most prone to bad mixing. The strictly pairwise IBD metric $w_{ij} = P(i, j \text{ IBD}|\boldsymbol{h}_i, \boldsymbol{h}_j)$ seems to give reliable results in general, both when used directly as entries in the covariance matrix (after positive definiteness is assured) or in a CAR specification. Sometimes however, even these analyses seem prone to bad mixing, and require a smaller $\alpha$ than suggested in (16) to negate the problem. The IBS based similarity metric, that can only be used in the CAR model, was least prone to numerical difficulties.
We expected that the theoretically motivated model of Section 2.3 with $w_{ij} = P(i, j \text{ IBD}|\boldsymbol{h})$ would give the best results, since we evaluate data simulated from the same genealogies that are assumed in the analysis. We believe that the above mentioned numerical difficulties might at least partially explain why we did not observe this model to outperform the others. In the

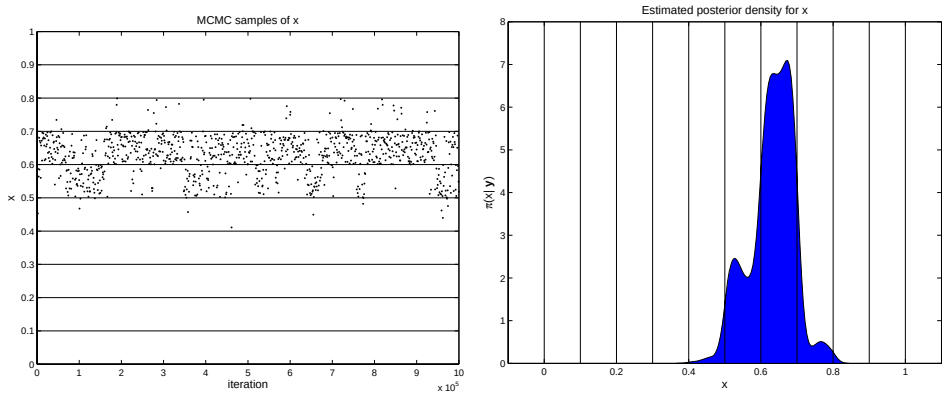


Figure 5: MCMC output and estimated posterior density for disease locus $x$, for the population genetic covariance matrix with $w_{ij} = P(i, j \text{ IBD}|\boldsymbol{h})$. True disease locus is at 0.65.

population genetic models the similarity matrix $\boldsymbol{W}$ enters directly as the covariance matrix $\boldsymbol{\Sigma} = \boldsymbol{W}$. For the IBD based similarity metrics, $\boldsymbol{W}$ gets close to singular for tightly linked markers. Numerical problems may then arise, e.g. when the density of $\boldsymbol{b}$ is calculated as part of the MCMC algorithm. Problems with bad mixing also increase with the size of the data set. Apart from problems with close to singular matrices, the large dependence between variables updated in different blocks in the MCMC algorithm might also be an issue for mixing, as discussed briefly in Section 5.

# 5 Discussion

We have in this paper studied two different spatial smoothing models for haplotype risk parameters, in an algorithm for genetic fine mapping using population based data. The CAR model has been suggested for fine mapping in earlier articles. It has the advantage that it can be used with a wide range of similarity metrics, but lacks population genetic interpretation. We have derived an alternative model from a population genetic perspective, which results in a covariance matrix consisting of pairwise IBD probabilities. We have studied how IBD probabilities can be calculated for population based data with tightly linked markers. Under a star shaped topology we retrieve analytical formulas for $P(i \text{ and } j \text{ IBD at x}|\boldsymbol{h})$, and $P(i \text{ and } j \text{ IBD at x}|\boldsymbol{h}_i, \boldsymbol{h}_j)$,

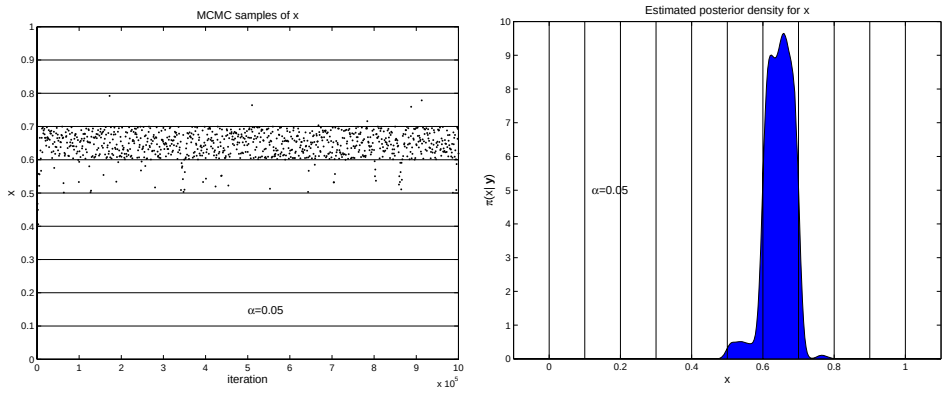


Figure 6: MCMC output for disease locus $x$, for the population genetic covariance matrix with $w_{ij} = P(i, j \text{ IBD}|\boldsymbol{h}_i, \boldsymbol{h}_j)$, with $\alpha = 0.05$. For this case-control study we obtained $\alpha = 0.30$ in (16), but the covariance matrix $\boldsymbol{W}$ was then not positive definite. True disease locus is at 0.65.

where the latter can be regarded as an approximation utilizing only the markers of the two respective haplotypes. Whereas IBS similarity metrics can be calculated without much background knowledge, the suggested IBD probabilities require (and adapt to) estimates of marker allele frequencies, marker mutation frequencies, and the (possibly varying) recombination rate. When used for gene mapping (on simulated data), the more advanced similarity metrics based on IBD probabilities gave slightly higher quality.

The CAR model, which explicitly smooths parameter estimates by writing the conditional mean of each parameter as a weighted average of all other parameters, has been extensively used e.g. in image analysis, geostatistical applications, spatial epidemiology and environmental statistics. Rue and Held (2005) and Banerjee et al. (2004) give a wealth of references and examples. The model has great numerical advantages in settings where the induced precision matrix $\boldsymbol{Q}$ is sparse, in which case $\boldsymbol{b}$ is a so called Markov Random field. From the definition in (6), the precision matrix and thus the covariance matrix is positive semi-definite (as long as the similarity matrix has only non-negative entries). In applications where the precision matrix is fixed, it is often not crucial that the precision matrix is positive definite. Since the disease locus $x$, and thus the precision matrix is here updated as part of the MCMC algorithm it is important that the parameters have a density, i.e. that the covariance matrices are all positive definite.

To get a proper distribution with the CAR formulation a popular approach, discussed by Banerjee et al. (2004) and used by Molitor et al. (2003a), is to introduce $\lambda$, $0 < \lambda < 1$ in $\boldsymbol{Q} = (\boldsymbol{M}^{-1}(\boldsymbol{I}_{2m} - \lambda \boldsymbol{C}))$. The "propriety parameter" $\lambda$ is then updated in the MCMC procedure along with the other parameters. To get non-negligible spatial dependence $\lambda$ must be very close to 1. We have instead added (a small) $\varepsilon > 0$ on the diagonal of the precision matrix. In earlier simulations we updated $\lambda$ or $\epsilon$ within the MCMC algorithm, but in the final algorithm we have instead added a fixed $\varepsilon$ of size $10^{-4}$, which gave a more effective MCMC algorithm, without impaired quality of the rest of the estimators.

Also for the model of Section 2.3, the issue of positive definiteness is of concern. As the covariance matrix is here taken as the (IBD) similarity matrix, it is required that the similarity matrix is in itself positive definite. This is ensured for $w_{ij} = P(i, j \text{ IBD at } x|\boldsymbol{h})$, whereas the approximation $w_{ij} = P(i, j \text{ IBD at } x|\boldsymbol{h}_i, \boldsymbol{h}_j)$ will often require an adaptation, especially for tightly linked markers. One possibility is addition of a positive number on the diagonal, but the larger addition that is needed, the further the model will depart from the derived model. The IBS-based similarity metrics of Section 3.2 in general produce a similarity matrix that is far from positive definite, and could thus not be used to directly define a covariance matrix. Although $w_{ij} = P(i, j \text{ IBD at } x|\boldsymbol{h})$ will always produce a positive definite similarity matrix, it may be so close to singularity, that numerical problems may occur. These problems are accentuated for large datasets with tightly linked markers. Addition of a small positive $\epsilon$ on the diagonal might thus be needed here too, to get good mixing.

Both CAR and the population genetic model require one risk parameter for each haplotype. If the number of individuals $m$ and haplotypes $2m$ is large, the model dimensionality gets very high. A computationally tractable alternative could then be to assign a risk parameter to each unique haplotype, $\boldsymbol{\beta} = (\beta_1, \ldots \beta_H)$ where $H$ is the number of unique haplotypes present in the sample. This is what Molitor et al. (2003a) did in their CAR model. The population genetic interpretation would be that equal haplotypes in today's sample implies the same founder haplotype at $x$. In an area of strong Linkage Disequilibrium (LD), $H$ is typically much smaller than $2m$. Another alternative that will probably scale better to data sets with very many haplotypes (as is common in studies of complex diseases) is to use a clustering approach just as in Molitor et al. (2003b), Waldron et al. (2006), etc. The IBD based similarity metrics could be used in these models as well, to split the haplotypes in clusters.

Apart from close to singular covariance or precision matrices, another reason for bad mixing in an MCMC analysis is high dependence in the posterior

distribution between parameters in different blocks. An improvement of our algorithm would thus be to update $\boldsymbol{b}$ in a block together with $\sigma_b^2$, and possibly also $\mu$. Rue and Held (2005, Ch. 4) describe how such block-wise updates could be achieved, which for a logit link, requires auxiliary variables with Kolmogorov-Smirnov distribution, for which the distribution function is only known as an infinite series.
The methods presented were developed for data with known genotype phase, which appears e.g. in data from case-parent triads. The methods will also work well when phase can be inferred with little uncertainty, which can be the case for very tightly linked loci. If phase is not known, it must be inferred, and the uncertainty should then be taken care of in the estimation procedure, as discussed e.g. by Thomas et al. (2003). This could be accomplished by treating haplotypes as unobserved data which is updated in the MCMC procedure. For genotypes where phase cannot be unambiguously resolved, the unobserved haplotypes can alternate between the possible states, or at least between these with an estimated frequency above a minimal threshold. If the possible haplotype pairs are found in a pre-analysis, together with an estimated probability for each of the states, these probabilities can be used when phase is updated in the MCMC algorithm, see e.g. Morris (2005), where this is done in a clustering algorithm.


# Acknowledgements

We are grateful to Juni Palmgren and Arvid Sjölander at Karolinska Institutet in Stockholm, Sweden for discussions and helpful comments during this project.


# A MCMC algorithm

We fit the model using MCMC, where $\boldsymbol{\xi} = (f_k, q_k, \rho, \ldots)$ and $\boldsymbol{h}$ are considered fixed and are therefore left out in the notation. The algorithm is close to that of Molitor et al. (2003a), although the parameters are here updated in larger blocks to obtain faster mixing. Due to the conditional independence relations that are illustrated in the DAG of Figure 1, the involved steps are comparably easily calculated. Notation $\boldsymbol{\theta} = (\mu, \boldsymbol{b}, \sigma_b, x)$ is for the collection of all model parameters and $\boldsymbol{\theta}_{-z}$ is used for all parameters except $z$, e.g. $\boldsymbol{\theta}_{-\mu} = (\boldsymbol{b}, \sigma_b, x)$. The algorithm works with both model types presented in Section 2, although $\sigma_b^2$ has different interpretation. For identifiability reasons, the risk parameters in the vector $\boldsymbol{b} = (b_1, \cdots, b_{2m})$ are constrained to sum

to 0, and thus (4) becomes

$$\pi(\boldsymbol{\theta}|\boldsymbol{y}) \propto L(\boldsymbol{y};\boldsymbol{\theta})\pi(\mu)\pi(\boldsymbol{b}|\mathbf{1}^T\boldsymbol{b}=0,x,\sigma_b^2)\pi(x)\pi(\sigma_b)$$

where $L(\boldsymbol{y};\boldsymbol{\theta}) = \pi(\boldsymbol{y}|\boldsymbol{\theta})$ is the likelihood. The distribution $\pi(\boldsymbol{b}|\mathbf{1}^T\boldsymbol{b}=0)$ (where other parameters are omitted in the notation) fulfils

$$\pi(\boldsymbol{b}|\mathbf{1}^T\boldsymbol{b}) = \frac{\pi(\boldsymbol{b})\pi(\mathbf{1}^T\boldsymbol{b}|\boldsymbol{b})}{\pi(\mathbf{1}^T\boldsymbol{b})}.$$

Thus we obtain

$$\pi(\boldsymbol{b}|\mathbf{1}^T\boldsymbol{b}=0) = \frac{\pi(\boldsymbol{b})|\mathbf{1}^T\mathbf{1}|^{-1/2}}{(2\pi)^{-1/2}|\mathbf{1}^T\sigma_b^{-2}\boldsymbol{Q}^{-1}\mathbf{1}|^{-1/2}} = \frac{\pi(\boldsymbol{b})(2m)^{-1/2}}{(2\pi\sigma_b^2)^{-1/2}|\mathbf{1}^T\boldsymbol{Q}^{-1}\mathbf{1}|^{-1/2}},$$

where $\pi(\boldsymbol{b})$ is as in Section 2. (For the simplified model described in Section 5, $2m$ is exchanged with $H$.) This constraint is accounted for in the proposal of $\boldsymbol{b}$, and in the acceptance probabilities of $\sigma_b$ and $x$, that influence the covariance matrix for $\boldsymbol{b}$. The term $\mathbf{1}^T\boldsymbol{Q}^{-1}\mathbf{1}$ is to be calculated and saved once and for all, for each possible $\boldsymbol{Q}$, i.e. for each marker interval.

## A.1 Likelihood

For the logit link model the likelihood is

$$L(\boldsymbol{y};\boldsymbol{\theta}) = \prod_{v=1}^{n} \frac{(\exp(\mu + b_{2v-1} + b_{2v}))^{y_v}}{1+\exp(\mu + b_{2v-1} + b_{2v})}.$$

## A.2 Update of $\mu$

We use a MH-step to update $\mu$. With prior density $\pi(\mu) \sim N(0,\sigma_{\mu_0}^2)$ and a normal random-walk (RW) proposal $\mu_{new} = \mu_{old} + z_\mu$ where $z_\mu \in N(0,\sigma_{mRW}^2)$ we obtain an acceptance probability

$$p_{acc} = \min\left(1, \frac{L(\boldsymbol{y};\mu_{new},\boldsymbol{\theta}_{-\mu})\pi(\mu_{new})}{L(\boldsymbol{y};\mu_{old},\boldsymbol{\theta}_{-\mu})\pi(\mu_{old})}\right).$$

## A.3 Update of $\boldsymbol{b}$

To gain computational speed and better mixing we update $\boldsymbol{b}$ in one block with a RW-proposal, and sample $\boldsymbol{b}_{new} = \boldsymbol{b}_{old} + \boldsymbol{z} - \bar{\boldsymbol{z}}$ where $\boldsymbol{z} \in N(0,\sigma_{bRW}^2\boldsymbol{I}_{2m})$ where $\boldsymbol{I}_{2m}$ is the identity matrix of dimension $2m$. Accept $\boldsymbol{b}_{new}$ with probability

$$p_{acc} = \min\left(1, \frac{L(\boldsymbol{y};\boldsymbol{b}_{new},\boldsymbol{\theta}_{-\boldsymbol{b}})\pi(\boldsymbol{b}_{new}|\boldsymbol{Q},\sigma_b^2)}{L(\boldsymbol{y};\boldsymbol{b}_{old},\boldsymbol{\theta}_{-\boldsymbol{b}})\pi(\boldsymbol{b}_{old}|\boldsymbol{Q},\sigma_b^2)}\right),$$

where $\pi(\boldsymbol{b}|\boldsymbol{Q},\sigma_b^2) \sim N(0,\sigma_b^2\boldsymbol{Q}^{-1})$, i.e. the unconstrained density for $\boldsymbol{b}$.

## A.4 Update of $\sigma_b$

The standard deviation, $\sigma_b$, of the risk parameters is updated in a Gibbs sampling step. Inverse gamma distribution is used as prior for $\sigma_b^2$, i.e. $\sigma_b^{-2} \in \Gamma(a_s, b_s)$ resulting in conditional distribution

$$\pi(\sigma_b^{-2}|\boldsymbol{\theta}_{-\sigma_b}, \boldsymbol{y}) \sim \Gamma(a_s + (2m-1)/2 - 0.5, \frac{2b_s}{2 + b_s \boldsymbol{b}^T \boldsymbol{Q} \boldsymbol{b}}).$$

The restriction $\sum b_i = 0$ enters as a subtraction of $-1/2$ in the shape parameter of the resulting gamma-distribution.

## A.5 Update of $x$

For disease locus $x$ we use a flat prior over the measured area. For better utilization of information in the more densely marked regions we use a two step procedure to sample proposal values $x_{new}$. First sample a marker location $k$ from the discrete uniform distribution on $\{1, \ldots, K\}$, then sample $x_{new}$ uniformly from the interval $\left(\frac{x_{k-1}+x_k}{2}, \frac{x_k+x_{k+1}}{2}\right)$. When $k$ is an end marker ($k = 1$ or $k = K$) $x_{new}$ is sampled uniformly from $\left(x_1, \frac{x_1+x_2}{2}\right)$ and $\left(\frac{x_{K-1}+x_K}{2}, x_K\right)$ respectively.
Accept $x_{new}$ with probability

$$p_{acc} = \min\left(1, \frac{\pi(\boldsymbol{b}|\boldsymbol{Q}(x_{new}), \sigma_b, \mathbf{1}^T\boldsymbol{b} = 0)(x_{k+1}^{new} - x_{k-1}^{new})}{\pi(\boldsymbol{b}|\boldsymbol{Q}(x_{old}), \sigma_b, \mathbf{1}^T\boldsymbol{b} = 0)(x_{k+1}^{old} - x_{k-1}^{old})}\right),$$

where

$$\pi(\boldsymbol{b}|\boldsymbol{Q}(x), \sigma_b, \mathbf{1}^T\boldsymbol{b} = 0) \propto \\ (\mathbf{1}^T\boldsymbol{Q}^{-1}(x)\mathbf{1})^{1/2}|\boldsymbol{Q}(x)|^{1/2} \exp(-0.5\sigma_b^{-2}\boldsymbol{b}^T\boldsymbol{Q}(x)\boldsymbol{b}).$$

# B Binary penetrance effects in the population genetic model

In Section 2.3 we derived (3) for a population genetic model with covariance matrix (9). It is more realistic though to assume that $b_i$ is binary, with the larger value attained when $I_i$ is mutated, and the lower value when it is not. Assume that only one founder allele $I_{\text{mut}}$ is mutated and has uniform distribution on $1, \ldots, N'$. It follows then from (8) that the disease allele frequency

$$p = P(I_i = I_{\text{mut}}) = \mathrm{E}(P(I_{\text{mut}} = I|I_i = I)) = 1/N'$$

In order to retain $\mathrm{E}(b_i|\boldsymbol{h}) = 0$ and $\mathrm{Var}(b_i|\boldsymbol{h}) = \sigma_b^2$ from (3), a simple calculation reveals that the two levels of $b_i$ have to be chosen according to

$$b_i = \begin{cases} \sqrt{q/p}\sigma_b, & I_i = I_{mut}, \\ -\sqrt{p/q}\sigma_b, & I_i \neq I_{mut}, \end{cases} \tag{18}$$

where $q = 1 - p$. It can be verified that (9) only holds in the limit $p \to 0$ for the binary model. The reason is that $\{b'_I\}$ are otherwise not i.i.d. but negatively correlated.
Let $\psi_l$ denote the probability that an individual with $l$ copies of the disease causing allele becomes affected. In view of (1) we thus have

$$\begin{cases} \psi_0 &= g^{-1}\left(\mu - 2\sqrt{p/q}\sigma_b\right) \\ \psi_1 &= g^{-1}\left(\mu + \left(\sqrt{q/p} - \sqrt{p/q}\right)\sigma_b\right) \\ \psi_2 &= g^{-1}\left(\mu + 2\sqrt{q/p}\sigma_b\right) \end{cases} \tag{19}$$

For the binary case we now motivate why (8) requires linkage equilibrium in the founder generation and that $x$ is not a marker locus.
For instance, suppose $x = x_k$, so that the disease locus is at $x_k$ and write $\boldsymbol{h}'_I = (h'_{I1}, ..., h'_{IK})$, so that $h'_{Ik}$ is the allele at $x_k$ of founder haplotype $\boldsymbol{h}'_I$. Then, by (18), we have

$$b'_I = -\sqrt{p/q}\sigma_b + (\sqrt{q/p} + \sqrt{p/q})\sigma_b h'_{Ik},$$

provided we encode $h'_{Ik}$ to have value 1 if $I$ is mutated, i.e. $I = I_{\mathrm{mut}}$. If there are no other mutations than the disease causing one, $h'_{Ik}$ equals the common value of $\{h_{ik};\, I_i = I\}$. If $x$ is close to $x_k$, $b'_I$ and $h'_{Ik}$ are still correlated if there is linkage disequilibrium in the founder generation.

# C Calculation of likelihood ratios conditional on $\boldsymbol{h}_i$ and $\boldsymbol{h}_j$ under the star topology

Following the notation of Hartman and Hössjer (2007), let $M_s$ denote the set of mutated chromosomes, put $\Omega = \{i, j\} \times \{1, \ldots, K\}$ and let $D \subseteq \Omega$ be the set of mutated sites. Further let $f_k$ denote the allele frequency at marker $k$ in the founder generation, and $q_k$ denote the probability of a mutation at marker $k$ between the founder generation and today's sample, so that today's allele frequency is $\tilde{f}_k(a) = (1 - q_k)f_k(a) + q_k(1 - f_k(a)),\ a = 0, 1$. Apply Hartman

and Hössjer (2007, Eq. 21), adapted to known phase, to subsample $\{i, j\}$ and sum over all possible $D$ to obtain

$$\mathrm{LR}_{ij} = \sum_D \mathrm{LR}(D) P_x(D|i, j \in M_s), \tag{20}$$

where $\mathrm{LR}(D) = P(\boldsymbol{g}|D)/\prod_{(i,k)\in\Omega} \tilde{f}_k(h_{ik})$. Let $n_{k0}(n_{k1})$ be the number of 0(1) alleles at marker $k$, that belong to the area $D$, see Figure 7. Since the phase is known, Hartman and Hössjer (2007, Eq. 22) implies

$$\mathrm{LR}(D) = \frac{P(\boldsymbol{h}_i, \boldsymbol{h}_j|D)}{\prod_{k=1}^K \tilde{f}_k(h_{ik})\tilde{f}_k(h_{jk})} = \prod_{k=1}^K \mathrm{LR}_k, \tag{21}$$

where

$$\mathrm{LR}_k = \frac{(1-q_k)^{n_{k0}} q_k^{n_{k1}} f_k(0) + q_k^{n_{k0}} (1-q_k)^{n_{k1}} f_k(1)}{\tilde{f}_k(0)^{n_{k0}} \tilde{f}_k(1)^{n_{k1}}}.$$

Let further $n_k = n_{k0} + n_{k1}$ be the number of mutated sites at locus $x_k$. We notice that $0 \le n_k \le 2$ and that $\mathrm{LR}_k = 1$ when $n_k < 2$. Hence we can rewrite (21) as

$$\mathrm{LR}(D) = \prod_{k;n_k=2} \mathrm{LR}_k. \tag{22}$$

Assume $x_{k_0} \le x < x_{k_0+1}$ and let $X^-$and $X^+$denote the crossovers closest to the left and right of $x$. Define $K^- \in \{1, \ldots, k_0+1\}$ and $K^+ \in \{k_0, \ldots, K\}$ by

$$\begin{array}{l} x_{K^--1} < X^- \le x_{K^-}, \\ x_{K^+} \le X^+ < x_{K^++1}. \end{array} \tag{23}$$

We can now rewrite (22) as

$$\mathrm{LR}(D) = \prod_{k=K^-}^{k_0} \mathrm{LR}_k \cdot \prod_{k=k_0+1}^{K^+} \mathrm{LR}_k, \tag{24}$$

where the products are interpreted as 1 when $K^- = k_0 + 1$ and $K^+ = k_0$ respectively. By construction of the retrospective ARG, $K^-$ and $K^+$ are independent random variables.
Hence, combining (20) and (24) we get

$$\mathrm{LR}_{ij} = \left( \sum_{k=1}^{k_0+1} P(K^- = k) \prod_{l=k}^{k_0} \mathrm{LR}_l \right) \cdot \left( \sum_{k=k_0}^{K} P(K^+ = k) \prod_{l=k_0+1}^{k} \mathrm{LR}_l \right). \tag{25}$$

Let $\rho$ be the expected number of recombinations within the chromosomal region since founder generation. Given that a recombination occurs, it has

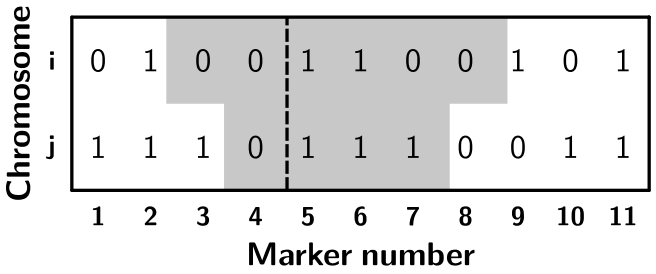


Figure 7: Illustration of the set $\Omega$ for two example haplotypes, $i$ and $j$, with $K = 11$ markers. The dashed vertical line displays the position of the disease mutation, and the shaded area shows $D$, the set of mutated sites. Thus, in this example $k_0 = 4$, $K^- = 4$ and $K^+ = 7$. Further $n_1 = 0 = n_{10} = n_{11},\ n_3 = 1$,
$n_{30} = 1,\ n_{31} = 0,\ n_4 = 2 = n_{40},\ n_{41} = 0,\ n_8 = 1 = n_{80},\ n_{81} = 0$, etc.

density $\pi$ on $[0,1]$. To calculate the distributions of $K^-$ and $K^+$ note that recombinations occur along different mutated $i$-lineages as independent Poisson processes with rate $\rho\pi(\cdot)$. Thus $\{X_i^-\}_{i\in M_s}$ and $\{X_i^+\}_{i\in M_s}$ are independent random variables with

$$\begin{array}{rcll} P_x(X_i^- < x') & = & \exp(-\rho\int_{x'}^{x}\pi(s)ds), & 0 \le x' \le x, \\ P_x(X_i^+ > x') & = & \exp(-\rho\int_{x}^{x'}\pi(s)ds), & x \le x' \le 1. \end{array} \tag{26}$$

If the recombination rate is uniform along the chromosome, i.e.
$\pi(s) \equiv 1,\ 0 \le s \le 1$ then

$$\begin{array}{rcll} P_x(X_i^- < x') & = & \exp(-\rho(x-x')), & 0 \le x' \le x, \\ P_x(X_i^+ > x') & = & \exp(-\rho(x'-x)), & x \le x' \le 1. \end{array} \tag{27}$$

Due to the star-topology of mutated chromosomes, the closest recombinations $X^- = \max(X_i^-, X_j^-)$ and $X^+ = \min(X_i^+, X_j^+)$ (to the left or right) between two different $i$-lineages are independent, and thus

$$\begin{array}{rcll} P_x(X^- < x') & = & \exp(-2\rho(x-x')), & 0 \le x' \le x, \\ P_x(X^+ > x') & = & \exp(-2\rho(x'-x)), & x \le x' \le 1. \end{array} \tag{28}$$

Thus

$$P(K^- = k) = \begin{cases} \exp(-2\rho x) & k = 1 \\ \exp(-2\rho(x - x_k))(1 - \exp(-2\rho(x_k - x_{k-1}))) & 2 \le k \le k_0 \\ 1 - \exp(-2\rho(x - x_{k_0})) & k = k_0 + 1, \end{cases} \tag{29}$$

and similarly for the recombinations to the right of $x$.
Thus LR, and hence also $w_{ij}$, can be computed in $O(K)$ time, and thus the full similarity matrix $\boldsymbol{W}$ is calculated in $O(Km^2)$ time. Notice that this is true also when neutral mutations are allowed for, i.e. $q_k > 0$, or when the recombinations do not appear uniformly.

# D Calculation of likelihood ratios conditional on $\boldsymbol{h}$ under the star topology

To calculate Equation (15), first note that $\widetilde{\mathrm{LR}} = P(\boldsymbol{h})/P_0(\boldsymbol{h})$ could be found by modifying Hartman and Hössjer (2007, Eqn. 19–20) to known phase, i.e.

$$\widetilde{\mathrm{LR}} = \sum_{\boldsymbol{h}'} f(\boldsymbol{h}') \prod_{i=1}^{2m} \sum_{C_i=0}^{1} \widetilde{\mathrm{LR}}(\boldsymbol{h}', i, C_i) P(C_i). \tag{30}$$

Here $P(C_i = 1) = P(\text{Chromosome } i \text{ mutated})$. Since we condition on neither marker data nor disease status, $P(C_i)$ is the same for all $i$. Thus we can utilize $P(C_i = 1) = \sqrt{P(C \in C_{ij})} = \sqrt{\alpha}$, where the star-topology gives the last identity. For non-mutated chromosomes $\widetilde{\mathrm{LR}}(\boldsymbol{h}', i, 0) = 1$, while for mutated chromosomes $\widetilde{\mathrm{LR}}(\boldsymbol{h}', i, 1) = \sum_{R_i} \left( \prod_{k \in R_i} P(h_{ik}|h'_k)/\tilde{f}_k(h_{ik}) \right) P_x(R_i|C_i = 1)$, where $R_i \subseteq \{1, \ldots, K\}$ denotes the set of all mutated marker loci for Chromosome $i$.
Similarly $\widetilde{\mathrm{LR}}_{ij} = P(\boldsymbol{h}|C \in \boldsymbol{C}_{ij})/P_0(\boldsymbol{h})$ is

$$\widetilde{\mathrm{LR}}_{ij} = \sum_{\boldsymbol{h}'} f(\boldsymbol{h}') \widetilde{\mathrm{LR}}(\boldsymbol{h}', i, 1) \widetilde{\mathrm{LR}}(\boldsymbol{h}', j, 1) \prod_{q \neq i,j} \sum_{C_q=0}^{1} \widetilde{\mathrm{LR}}(\boldsymbol{h}', q, C_q) P_x(C_q). \tag{31}$$

The calculation is $O(2^K m^2)$ in time, where the main effort is calculation of $2^K 2m$ likelihood ratios $\widetilde{\mathrm{LR}}(\boldsymbol{h}', i, 1)$. Calculating $P(i, j \text{ IBD at } x|\boldsymbol{h})$ is thus more computationally intense than calculating $P(i, j \text{ IBD at } x|\boldsymbol{h}_i, \boldsymbol{h}_j)$, but still feasible for reasonably small $K$ and $m$.

# References


Almasy, L., Blangero, J., May 1998. Multipoint quantitative-trait linkage analysis in general pedigrees. Am J Hum Genet 62 (5), 1198–1211.

Amos, C., 1994. Robust variance-components approach for assessing genetic linkage in pedigrees. Am J Hum Genet 54, 535–543.

Banerjee, S., Carlin, B. P., Gelfand, A. E., 2004. Hierarchical modeling and analysis for spatial data. Chapman & Hall/CRC.

Breslow, N. E., Clayton, D. G., 1993. Approximate inference in generalized linear mixed models. J Amer Statist Assoc 88 (421), 9–25.

Durrant, C., Morris, A., Dec 2005. Linkage disequilibrium mapping via cladistic analysis of phase-unknown genotypes and inferred haplotypes in the Genetic Analysis Workshop 14 simulated data. BMC Genet 6 Suppl 1, S100.

Durrant, C., Zondervan, K. T., Cardon, L. R., Hunt, S., Deloukas, P., Morris, A. P., Jul 2004. Linkage disequilibrium mapping via cladistic analysis of single-nucleotide polymorphism haplotypes. Am J Hum Genet 75 (1), 35–43.

Hartman, L., Hössjer, O., 2007. Retrospective ancestral recombination graphs with applications to gene mapping. Tech. Rep. 2007:7, Centre for Mathematical Sciences, Lund University.

McCulloch, C. E., Searle, S. R., 2001. Generalized, linear, and mixed models. Wiley series in probability and statistics. Wiley, New York.

Meuwissen, T. H., Goddard, M., Jun 2007. Multipoint IBD prediction using dense markers to map QTL and estimate effective population size. Genetics.

Meuwissen, T. H., Goddard, M. E., 2001. Prediction of identity by descent probabilities from marker-haplotypes. Genet Sel Evol 33 (6), 605–634.

Meuwissen, T. H. E., Karlsen, A., Lien, S., Olsaker, I., Goddard, M. E., May 2002. Fine mapping of a quantitative trait locus for twinning rate using combined linkage and linkage disequilibrium mapping. Genetics 161 (1), 373–379.

Molitor, J., Marjoram, P., Thomas, D., Sep 2003a. Application of Bayesian spatial statistical methods to analysis of haplotypes effects and gene mapping. Genet Epidemiol 25 (2), 95–105.

Molitor, J., Marjoram, P., Thomas, D., Dec 2003b. Fine-scale mapping of disease genes with multiple mutations via spatial clustering techniques. Am J Hum Genet 73 (6), 1368–84.

Morris, A. P., Sep 2005. Direct analysis of unphased SNP genotype data in population-based association studies via Bayesian partition modelling of haplotypes. Genet Epidemiol 29 (2), 91–107.

Rue, H., Held, L., 2005. Gaussian Markov Random Fields: Theory and Applications. Vol. 104 of Monographs on Statistics and Applied Probability. Chapman & Hall, London.

Schaid, D. J., Dec 2004. Evaluating associations of haplotypes with traits. Genet Epidemiol 27 (4), 348–364.

Sham, P. C., Purcell, S., Cherny, S. S., Abecasis, G. R., Aug 2002. Powerful regression-based quantitative-trait linkage analysis of general pedigrees. Am J Hum Genet 71 (2), 238–253.

Thomas, D. C., Morrison, J. L., Clayton, D. G., 2001. Bayes estimates of haplotypes effects. Genet Epidemiol 21(Suppl 1), S712–S717.

Thomas, D. C., Stram, D. O., Conti, D., Molitor, J., Marjoram, P., 2003. Bayesian spatial modeling of haplotype associations. Hum Hered 56 (1-3), 32–40.

Waldron, E. R. B., Whittaker, J. C., Balding, D. J., Feb 2006. Fine mapping of disease genes via haplotype clustering. Genet Epidemiol 30 (2), 170–179.