Statistical Applications in Genetics and Molecular Biology

Volume 8, *Issue* 1 2009 *Article* 35

Ancestral Recombination Graphs under Non-Random Ascertainment, with Applications to Gene Mapping

Ola Hössjer, *Stockholm University*
Linda Hartman, *AstraZeneca*
Keith Humphreys, *Karolinska Institutet*

Recommended Citation:
Hössjer, Ola; Hartman, Linda; and Humphreys, Keith (2009) "Ancestral Recombination Graphs under Non-Random Ascertainment, with Applications to Gene Mapping," *Statistical Applications in Genetics and Molecular Biology*: Vol. 8: Iss. 1, Article 35.
DOI: 10.2202/1544-6115.1380

Ancestral Recombination Graphs under Non-Random Ascertainment, with Applications to Gene Mapping

Ola Hössjer, Linda Hartman, and Keith Humphreys

Abstract

Consider a sample of apparently unrelated individuals, for which marker genotype and phenotype data is available. When individuals are sampled on phenotypes, we propose an ascertained ancestral recombination graph (ARG) that models shared ancestry of the sample chromosomes given phenotype data along a region that possibly harbors a disease susceptibility gene. The ascertained ARG is used to define a gene mapping algorithm by means of a lod score and associated *p*-values based on permutation testing. Under certain modeling simplifications, the lod score and *p*-values can be computed exactly, without any Monte Carlo approximations, even for unphased chromosome genotype data. Our method handles incomplete penetrance, varying marker allele frequencies and neutral mutations, and is based on a Hidden Markov algorithm for subsets of disease mutated chromosomes. The performance of the method is investigated in a simulation study and for a real data set from a case-control study of breast cancer.

KEYWORDS: ancestral recombination graph, association analysis, case-control study, identical-by-descent, Hidden Markov Model, LOD score, multipoint, unknown haplotype phase

Author Notes: The work of Linda Hartman was supported by the Swedish foundation for Strategic Research (SSF), contract number A3 02:125. The work of Ola Hössjer and Keith Humphreys was supported by the Swedish Research Council, contract numbers 621-2005-2810 (for O.H.) and 523-2006-972 (for K.H.). We would like to thank colleagues working on the genetic association study of breast cancer, described in Section 7.3, for data, as well as valuable comments from the referees.

1 Introduction

1.1 Background

Population-based association studies are popular for gene mapping on a fine scale. Due to the higher number of meioses with possible recombination between the most recent common ancestor (MRCA) and today's apparently unrelated individuals, these studies yield higher resolution than is possible in family-based linkage studies.

Early association studies tested for association between disease and each marker separately. More recently efforts have been made to define more efficient multipoint methods that pinpoint loci x around which case (or disease mutated) chromosomes tend to cluster because of common inheritance from a quite recently mutated founder chromosome.

The simplest possiblity is to base clustering on some measure of identity-by-descent (IBS). For instance, Mailund et al. (2006) define a non-random phylogenic tree based on markers in close vicinity of the putative disease locus x. Then a statistic is chosen that quantifies the degree of clustering of case chromosomes in the phylogenic tree. Another possibility is to use haplotype clustering and cladograms, see e.g. Molitor et al. (2003), Durrant et al. (2004) and Waldron et al. (2006). The basis of these methods is an IBS-based haplotype similarity measure between pairs of haplotypes, such as the largest shared region around x, possibly normalized for varying allele frequencies.

A more elaborate approach, which focuses on identical by descent (IBD) sharing rather than IBS, is set forth by likelihood and Bayesian methods that define a stochastic model for linkage disequilibrium (LD) over a whole chromosome region as a result of common ancestry.

The population genetic tool for such gene mapping methods is the Ancestral Recombination Graph (ARG) of Hudson (1983), Griffiths and Marjoram (1997) and Wiuf and Hein (1999). It models how a population of chromosomes are related to each other along a given region through coalescence and recombination events. Typically, mutations are added along the branches of the ARG after its construction. The marginal coalescence tree, which describes shared ancestry at one position, can then be extracted from the ARG at each locus. Since the ARG as well as ancestral haplotypes are hidden variables, they must be summed over in the likelihood. Without further model simplifications, this cannot be done exactly, but has to be carried out by some Monte Carlo method, e.g. importance sampling or Markov chain Monte Carlo (MCMC). For instance, Larribe et al. (2002) incorporate the full ARG into gene mapping, using importance sampling. However, this method is hampered by its computational demand.

The main challenge is to develop computationally tractable yet effective gene mapping methods that incorporate some but not all of shared ancestry of the sampled chromosomes. An important first step in this direction was Terwilliger (1995), where star topology among case chromosomes was assumed and information from many markers combined, although dependence across markers was not taken into account. A generalization due to Service et al (1999), incorporates dependence between markers.

A closely related but slightly more systematic approach is to model shared ancestry among case chromosomes by means of a local ARG around the putative disease locus x, only incorporating recombination events closest to x, whereas control (or wildtype) haplotypes are considered unrelated with haplotype frequencies estimated from the population, typically approximated by means of an l:th order Markov chain, where l is 0 or 1. This reflects the fact that in association studies, there is an underlying hypothesis that cases are descendants from one (or a few) founders carrying the mutation. Then the genealogy of cases should be qualitatively different from that of the controls. The advantage of the local ARG is that only the coalescence tree at x needs to be modeled explicitly. For instance, McPeek and Strahs (1999) assume a star topology among cases, with a quasi likelihood correction factor that to some extent corrects for more general coalescence trees. The ancestral haplotype is interpreted as a nuisance parameter, haplotype phase is assumed to be known and the likelihood is evaluated using a Hidden Markov Model (HMM) algorithm. Sporadic cases are allowed for, as are, in principle, multiple disease mutations at x. Morris et al. (2000) use a similar approach, but within a Bayesian MCMC framework, treating ancestral haplotypes as a hidden variable that is summed over. Rannala and Reeve (2001) and Morris et al. (2002) also employ a Bayesian model and MCMC, but use more general coalescence trees for case chromosomes. In particular, the shattered coalescence tree of Morris et al. (2002) allows for both sporadic cases and multiple disease mutations. This is also true for Liu et al. (2001), although they assume a star topology for the subtrees of each disease mutation. Zöllner and Pritchard (2005) define the local ARG for *all* sampled chromosomes (not only cases) conditional on genotype data. Phenotypes are not incorporated when trees are generated, but added afterwards into the likelihood. In this way, multiple disease mutations and various penetrance models can be handled at little extra cost.

Other simplifications of the ARG have recently been suggested for gene mapping, as an alternative to the local ARG. Larribe and Lessard (2008) combine information of full ARGs from several small windows surrounding x, using a composite likelihood. Minchiello and Durbin (2006) develop a heuristic algorithm to infer plausible ARGs, whereas Wu (2007) employ ARGs that minimize the number of recombination events consistent with data. Gasbarra et al. (2007, 2009) employ the

DOI: 10.2202/1544-6115.1380

the discrete time genealogy of Gasbarra et al. (2005) for gene mapping. The resulting ARG with discrete time generations allows for tuning of offspring distributions and degree of monogacy, and is summed over by means of MCMC.

1.2 Outline of Paper

A major question is what given information in data the ARG should be conditioned on when applied to gene mapping in a retrospective study, for instance the case-control study. In such a study individuals are sampled (ascertained) non-randomly based on phenotypes. Ignoring the ascertainment may introduce bias, see e.g. Thomas et al. (2003). The conditional likelihood of genotype data given phenotypes is a natural choice that removes bias caused by ascertainment. This naturally leads to an *ascertained ARG*, by which we mean the ARG conditioned on phenotypes. It is true that when Monte Carlo approximation of the likelihood is employed, the conditional distribution of the ARG given genotype *and* phenotype data is ideal for computation. However, this ARG is very difficult to sample from and some approximation has to be used, see Sections 8.3-8.4. Instead, the ascertained ARG arises naturally from an expansion of the likelihood and is natural to use, at least when the likelihood can be computed exactly.

The two major contributions of the present paper are:

I) To present an ascertained (non-local) ARG along a given region. When there is one disease causing mutation, chromosomes are separated into two subpopulations of mutated and wildtype chromosomes respectively, for which the evolutionary process is differently modelled. Our approach is conditional on the age G of the disease mutation as well as on subpopulations sizes. This enables a novel asymptotic construction of the ascertained ARG, where time is counted in units of G (rather than in units of present populations size).

II) To use the ascertained ARG for defining a lod score for population based multipoint association studies. This lod score takes the retrospective sampling scheme as well as unknown haplotype phase into account and does not require Monte Carlo calculation for regions of short or medium length. A certain approximation, the pseudo lod score (PLOD), handles longer sequences as well.

The ascertained ARG in I) is general, but for computational reasons we suggest model simplifications and present a special case in II) for which the LOD score is tractable. These special cases include a) star topology coalescence tree among mutated chromosomes, b) single disease causing mutation, c) unrelated wildtype chromosomes, yielding background LE in the whole population and d) rare disease

allele frequency. In Section 8 we discuss how to relax these assumptions, whilst still retaining reasonable computation time.

Assumption d) is important since it naturally leads to a local ascertained ARG, similar to the genealogy employed by McPeek and Strahs (1999) and Morris et al. (2000). Our method is based on a Hidden Markov Model (HMM) with state space subsets of mutated chromosomes at a (varying) locus. It extends the HMM algorithm of McPeek and Strahs (1999) and handles a reasonably large number of markers, arbitrary phenotypes and genetic models, allows for neutral mutations, adapts to marker allele frequencies and, most importantly, unknown haplotype phase is handled at almost no extra computational cost. This is an important computational advantage of our method, since for most data sets multilocus genotypes are observed and halpotype phase cannot be resolved unambiguously. In contrast, for previous methods haplotypes are either assumed to be known, or, if not, the posterior distribution of haplotypes is either estimated in advance, using some haplotype reconstruction algorithm (e.g. Stephens et al., 2001), or estimated simultaneously with gene mapping using MCMC (Morris et al., 2004).

We further show how exact *p*-values can be estimated by means of a computationally feasible permutation procedure. Because of exchangeability of families under the null hypothesis of no disease gene, the *p*-values are non-biased under the null, even when the model simplifications a)-d) are incorrect.

Both Zöllner and von Haesler (2000) and Wang and Rannala (2004, 2005) introduce an ascertained ARG with subpopulations. These articles focus on using simulation to examine the performance of single locus association tests for chromosomes simulated under different scenarios. No multilocus gene-mapping algorithm is developed.

2 A LOD Score for Association Studies

In this section we define the basic gene mapping tools used in the rest of the paper, in particular the retrospective likelihood, the accompanying lod score and the regionwide *p*-value based on permutation testing.

Let τ denote a disease locus of an inheritable disease, and assume that two alleles exist at τ: a normal allele b and a disease causing allele B. The purpose of gene mapping is to estimate τ and/or test if a certain (small) chromosome region harbors τ. The region of interest is normalized as a unit interval $[0,1]$ in terms of genetic or physical map distance. The hypothesis testing problem of interest is

$$\begin{aligned} H_0 &: \quad \tau \notin [0,1], \\ H_1 &: \quad \tau \in [0,1]. \end{aligned}$$

DOI: 10.2202/1544-6115.1380

We assume a subset of m individuals with phenotypes $Y = (Y_1, \ldots, Y_m)$ is sampled. For each individual DNA is registered at a number of markers with positions $0 \le x_1 < \ldots < x_K \le 1$. Let $h_{2v-1} = (h_{2v-1,k})_{k=1}^K$ and $h_{2v} = (h_{2v,k})_{k=1}^K$ be the two homologous haplotypes of individual v and $h = (h_i)_{i=1}^n$ be the collection of all $n = 2m$ haplotypes. In general, because of phase uncertainty, (h_{2v-1}, h_{2v}) is not known for v but rather the unphased multilocus genotype g_v. Write $g = (g_1, \ldots, g_m)$ for the collection of all unphased multilocus genotypes. Based on marker data g and phenotypes Y we compute a test statistic $Z(x)$ for the pointwise test H_0 versus $H_1^x : \tau = x$ and reject H_0 when $Z(x)$ is large. Then

$$Z_{\max} = \max_{0 \le x \le 1} Z(x)$$

is a global test statistic for testing H_0 versus H_1, with large values of $Z_{\max}$ leading to rejection of H_0. Alternatively, we may estimate the disease locus as $\hat{\tau} = \arg\max_{0 \le x \le 1} Z(x)$ and compute an associated confidence region.

The test statistic $Z(x)$ should be large when affected individuals, or individuals with quantitative phenotypes indicating disease, tend to share DNA around x more often than expected by chance. This is so since under H_1^x, the mutated chromosome is segregated in close vicinity of x down to all mutated chromosomes of the sample.

To this end, we define the retrospective likelihood

$$L(x; \xi) = P_x(g|Y) \tag{1}$$

of genotype data given phenotypes, where the probability P_x is calculated under H_1^x. By conditioning on Y we do not need to know the sampling mechanism, as long as it is a function of Y only. This is an advantage, since the sampling scheme is often unknown in practice, see e.g. Kraft and Thomas (2000). All nuisance parameters that involve recombination, mutation, population growth and penetrance of the disease are contained in ξ. We assume that ξ is known or can be assigned an a-priori reasonable value, and use as test statistic the LOD score

$$Z(x) = \log_{10} \mathrm{LR}(x) = \log_{10} \frac{L(x)}{L(\infty)}, \quad 0 \le x \le 1. \tag{2}$$

Here $L(\infty)$ denotes the retrospective likelihood under H_0, since then τ is regarded as being unlinked to $[0,1]$, expressed formally as $\tau = \infty$. Hence $Z(x)$ is the base ten logarithm of the likelihood ratio $\mathrm{LR}(x)$ obtained when testing H_0 against H_1^x.

To assess the statistical significance of an observed maximal LOD score $Z_{\max} = z_{\max}$, we use permutation testing. Given any permutation γ of $\{1, \ldots, m\}$, let $Z_{\max,\gamma}$ be the maximal LOD score based on the retrospective likelihood $P_x(g|Y_\gamma)$,

where $Y_\gamma = (Y_{\gamma(1)}, \ldots, Y_{\gamma(m)})$ is the phenotype vector permuted according to γ. The p-value based on Q randomly chosen permutations $\gamma_1, \ldots, \gamma_Q$ is then then $\alpha(z_{\max})$, where

$$\alpha(z) = \frac{1}{Q} \sum_{i=1}^{Q} 1_{\{Z_{\max,\gamma_i} \geq z\}} \tag{3}$$

and 1_D is the indicator function of the event D. (With a Bayesian approach, permutation can be avoided though, cf. Section 8.4.

3 Expanding the Likelihood

In this section, we expand the retrospective likelihood (1) by summing over all genealogies across $[0,1]$ consistent with genotype and phenotype data. As we will see, this involves summing over all ARGs $\mathscr{A}$ corrected for disease status (and thus for ascertainment). The resulting likelihood handles dependence between and along the chromosomes.

The unconditional ARG $\mathscr{A}$ models the genealogy from today's generation back until the founder generation. This gives us the kinship relations of today's chromosomes, and thus a model for the dependencies. The expanded likelihood is

$$L(x) = \sum_{\mathscr{A}} P(g|\mathscr{A}) P_x(\mathscr{A}|Y) \tag{4}$$

for $x \in [0,1]$. Since we use a retrospective likelihood $P_x(g|Y)$, we see from (4) that an ascertained ARG $\mathscr{A}|Y$ is natural. For a prospective likelihood $P_x(Y|g)$, an ARG $\mathscr{A}|g$ would be of main interest (Fearnhead and Donnelly, 2002).

In order to define $\mathscr{A}$, we assume non-overlapping generations and follow the genealogy of the n chromosomes in the m sampled individuals along $[0,1]$ backwards in time until the disease causing mutation occurred, G generations ago (the founder generation). The ARG $\mathscr{A}$ involves both recombination and coalescence events (Hudson, 1983, Griffiths and Marjoram, 1997), as illustrated in Figure 1a for an ARG with $G = 3$ generations and $m = 2$ individuals. The point of recombination, X, is written above each recombination vertex. It means that two ancestral chromosomes c_1 and c_2 recombine at X, such that c_1 (the left hand chromosome) passes on genetic material $[0, X)$ and c_2 (the right hand edge), passes on genetic material $[X, 1]$, to the child chromosome c.

Let N_u denote the population size, i.e. the number of chromosomes of $N_u/2$ diploid individuals u generations back in time, $u = 0, 1, \ldots, G$. Let $\mathscr{S}_u = \{1, \ldots, n_u\}$ denote the set of chromosomes of Generation u that are ancestral to at least one of the n sampled chromosomes *somewhere* along $[0,1]$, assuming that the ancestral chromosomes of each generation have been numbered in some (arbitrary) way.

DOI: 10.2202/1544-6115.1380

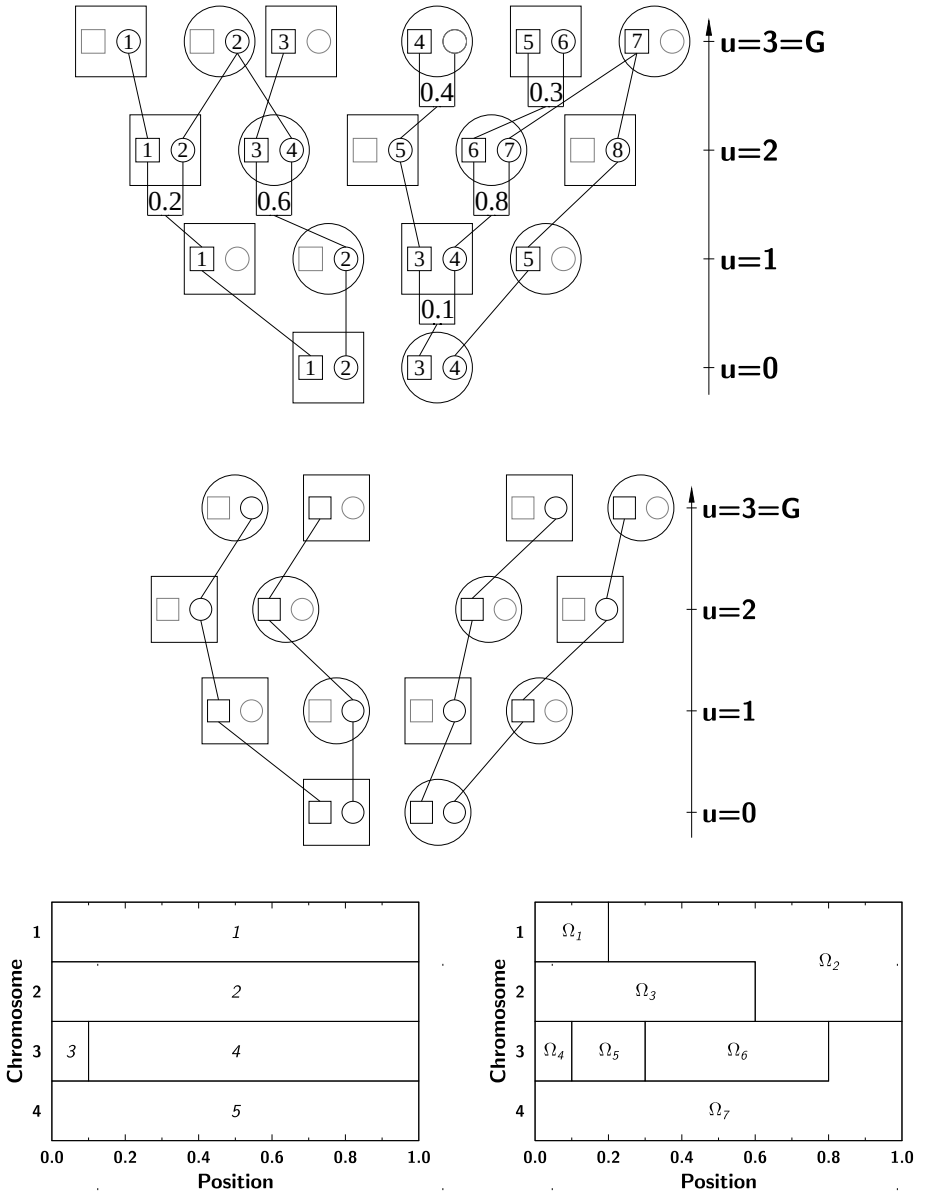

Figure 1: Ancestry of $m = 2$ individuals during $G = 3$ generations. (1a): ARG of the sample. A numbering $(1,\ldots,n_u)$ of the ancestral chromosomes of each Generation u is displayed. Points of recombination are written above recombination verteces. (1b): Coalescence tree at $x = 0.4$. (1c): Generation 1 ancestry $\mathscr{A}_1(x,i)$ of the sampled chromosomes (1a). (1d): Decomposition into IBD-regions Ω_j.

Then $1 \leq n_u \leq N_u$ and $n_0 = n$. We may write $\mathscr{A} = \{\mathscr{A}_u\}_{u=0}^{G}$, where[1] $\mathscr{A}_u : \{1,\ldots,n\} \times [0,1] \to \mathscr{S}_u$ is the ancestry of the sample in Generation u, and $\mathscr{A}_u(i,x) = j$ means that j is ancestral to i at x, see Figure 1c.

In particular, we say that DNA position x_1 of Chromosome i_1 is *identical by descent* (IBD) to DNA position x_2 of Chromosome i_2 if $\mathscr{A}_G(i_1,x_1) = \mathscr{A}_G(i_2,x_2)$. This IBD definition gives rise to a decomposition of $\{1,\ldots,n\} \times [0,1]$ into n_G disjoint regions $\Omega_j = \{(i,x); \mathscr{A}_G(i,x) = j\}$, $j = 1,\ldots,n_G$ as shown in Figure 1d.

An alternative ARG representation is $\mathscr{A} = \{\mathscr{T}(x); 0 \leq x \leq 1\}$, where $\mathscr{T}(x)$ is the marginal coalescence tree of the sample at x. $\mathscr{T}(x)$ only involves coalescence events and no recombination events. We obtain $\mathscr{T}(x)$ from $\mathscr{A}$ by following the n lineages from time 0, and, whenever a recombination event at X is passed, take the left parental edge if $x < X$ and the right parental line if $x \geq X$, see Figure 1b.

A chromosome belonging to any generation u is either mutated or wildtype. In the former case, it has received genetic material around τ from the mutated founder chromosome (allele B), and in the latter case not (allele b). Hence $N_u = N_{Mu} + N_{Wu}$, where N_{Mu} and N_{Wu} denote the number of mutated and wildtype chromosomes of Generation u. Further let $\mathscr{M}_u \subset \mathscr{S}_u$ and $\mathscr{W}_u \subset \mathscr{S}_u$ denote the subsets of mutated and wildtype chromosomes ancestral to the given sample of chromosomes. Then $n_u = n_{Mu} + n_{Wu}$, where $n_{Mu} = |\mathscr{M}_u|$ and $n_{Wu} = |\mathscr{W}_u|$.

We make a number of assumptions that will simplify our genealogy as well as the likelihood computations:

(i) The mutated chromosomes descend from a single chromosome of the founder generation, i.e. $n_{MG} = 1$.

(ii) G, $\{N_{Wu}\}_{u=0}^{G}$ and $\{N_{Mu}\}_{u=0}^{G}$ are known.

(iii) None of the marker loci are causal, i.e. $\tau \notin \{x_1,\ldots,x_K\}$.

(iv) τ is the only disease locus which, under H_1, is linked to $[0,1]$.

(v) Haplotypes (formed by alleles from loci $x_1,\ldots,x_K$) of the founder generation are independent with haplotype frequencies f.

(vi) All marker loci x_k are selectively neutral. Mutations occur at x_k with probability u_k per meiosis.

(vii) The disease mutation occurs at a randomly chosen chromosome $J \in \{1,\ldots,n_G\}$ of the founder generation, independently of founder haplotypes.

[1] A more precise definition is that $\mathscr{A}_u$ is invariant with respect to numbering of ancestral chromosomes of Generation $u-1$.

DOI: 10.2202/1544-6115.1380

(viii) Recombinations occur with probability r per chromosome and generation, independently of the mating in each generation. Given that a recombination occurs, it has density $\pi(\cdot)$ along $[0,1]$.

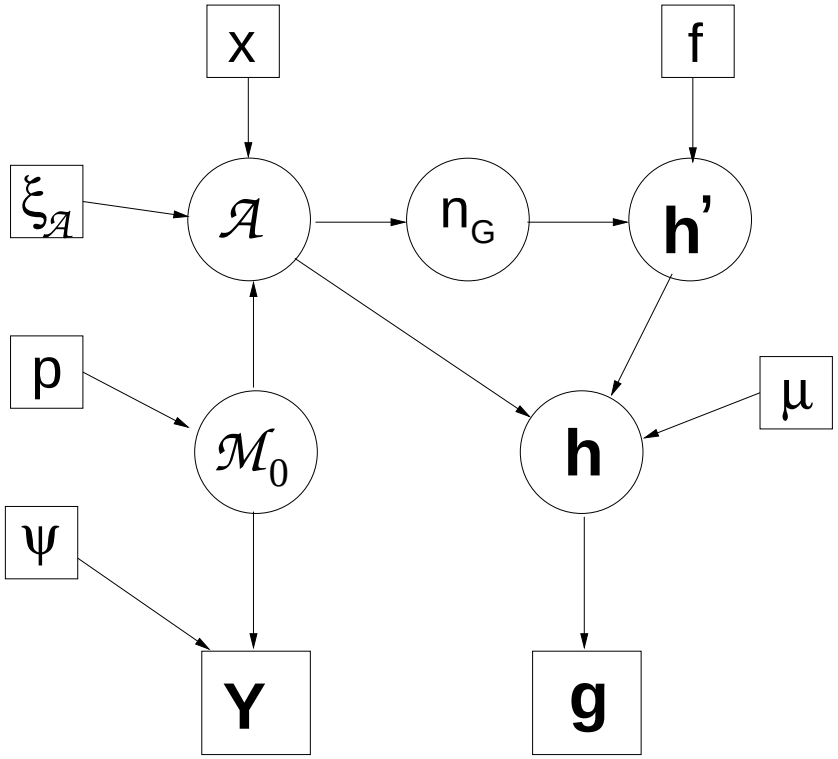

Figure 2: Directed Acyclic graph (DAG) of all parameters and variables relevant for the likelihood. Each arrow corresponds to a deterministic (e.g. $\mathscr{A} \to n_G$) or probabilistic causal relationsship, squares are constants (parameters or observed random variables) and circles unobserved random variables. The nuisance parameters $\xi = (\psi, p, \xi_{\mathscr{A}}, \mu, f)$ are defined in Table 1.

Figure 2 shows a directed acyclic graph (DAG) of all parameters and variables relevant for the likelihood and Table 1 summarizes the notation. Notice that $\mathscr{A}$ only describes the genealogy of the sample. It carries no information on mutation or marker data, but rather how to segregate founder haplotypes h', ancestral to the given sample, to obtain h, and how to segregate the mutated chromosome J to obtain $\mathscr{M}_0$. Given the value of a node in the DAG, its ancestors and descendants are conditionally independent. Hence Figure 2 implies that the likelihood (1) and (4)

can be expanded as

$$L(x) = \sum_{h,h',\mathscr{A},\mathscr{M}_0} P(g|h)P(h|\mathscr{A},h')P(h'|\mathscr{A})P_x(\mathscr{A}|\mathscr{M}_0)P(\mathscr{M}_0|Y). \tag{5}$$

The conditional dependencies of Figure 2 are consequences of (i)-(viii). For instance, (iv) implies that Y and $(\mathscr{A},h',h)$ are conditionally independent given $\mathscr{M}_0$. This implies that knowing the mutation status of all sampled chromosomes, we gain no further information on phenotypes from marker data.

In order to compute the likelihood we need to specify each of the five terms occurring on the right hand side of (5). Assuming individuals are sampled independently based on their phenotypes, the fifth term $P(\mathscr{M}_0|Y)$ in (5) can be written as

$$P(\mathscr{M}_0|Y) = \prod_{v=1}^{m} P(\mathscr{M}_0 \cap \{2v-1,2v\}|Y_v). \tag{6}$$

Each term of the right-hand side of (6) only depends on genetic model parameters of the disease. Define penetrances

$$\psi_{vj} = P(Y_v|v \text{ has } j \text{ disease alleles } B),$$

for $v = 1,\ldots,m$ and $j = 0,1,2$. According to (ix) in Appendix A, the disease genotype frequencies of Generation 0 are $(1-p)^2$, $2p(1-p)$ and p^2 for genotypes (bb), (Bb) and (BB), where $p = p_0$ is the disease allele frequency. Bayes' Theorem gives

$$\begin{array}{rcl} P(2v-1 \in \mathscr{M}_0, 2v \in \mathscr{M}_0|Y_v) &=& \psi_{v2}p^2/S \\ P(2v-1 \notin \mathscr{M}_0, 2v \in \mathscr{M}_0|Y_v) &=& \psi_{v1}p(1-p)/S \\ P(2v-1 \in \mathscr{M}_0, 2v \notin \mathscr{M}_0|Y_v) &=& \psi_{v1}p(1-p)/S \\ P(2v-1 \notin \mathscr{M}_0, 2v \notin \mathscr{M}_0|Y_v) &=& \psi_{v0}(1-p)^2/S, \end{array} \tag{7}$$

where $S = \psi_{v0}(1-p)^2 + 2\psi_{v1}p(1-p) + \psi_{v2}p^2$.

The fourth term $P_x(\mathscr{A}|\mathscr{M}_0)$ will be further discussed in the next section as a Markov chain backwards in time.

For the third term of (5), it follows from (v)-(vii), after conditioning on J, that

$$P(h'|\mathscr{A}) = P(h'|n_G) = \prod_{j=1}^{n_G} f(h'_j) = f(h'_J)\prod_{j\neq J} f(h'_j) \tag{8}$$

where $h' = \{h'_j\}_{j=1}^{n_G}$ is the collection of founder haplotype vectors $h'_j = (h'_{j1},\ldots,h'_{jK})$.

The second term of (5) depends on neutral mutations at the marker loci. Let $h_{jk} = \{h_{ik};\, (i,x_k) \in \Omega_j\}$. We obtain h_{jk} by spreading the founder allele h'_{jk} according to the coalescence tree $\mathscr{T}(x_k)$ to all chromosomes i that are IBD with j, possibly

Table 1: Variables and parameters.

τ	True disease locus
x	Hypothesized disease locus
n	Number of haplotypes
m	Number of individuals
Y	Sampled phenotypes
g	Sampled genotypes
h	Current generation haplotypes
h'	Founder haplotypes, at least *some* part of each haplotype ancestral to sample
G	Time (generation) of disease mutation
$\mathscr{M}_u$	Mutated chromosomes, Generation u
$\mathscr{W}_u$	Wildtype chromosomes, Generation u
N_u	Population size, Generation u
N_{Mu}	Population size, mutated subpopulation, Generation u
N_{Wu}	Population size, wildtype subpopulation, Generation u
n_u	Number of haplotypes ancestral to sample, Generation u
$\mathscr{A}$	ARG of sampled chromosomes along [0,1]
$p(t)$	Proportion of mutated chromosomes at time t
p	Disease allele frequency of current generation ($= p(0)$)
ψ	Penetrance parameters ($= \{\psi_{vj}\}$)
f	Founder haplotype frequencies (marginal marker allele frequencies f_k)
$\tilde{f}$	Current haplotype frequencies (marginal marker allele frequencies $\tilde{f}_k$)
μ	Mutation rates ($= \{\mu_k\}_{k=1}^K$)
$\xi_{\mathscr{A}}$	Continuous time ARG parameters ($= (\rho, \pi(\cdot), \lambda_M(\cdot), \lambda_W(\cdot))$)
ρ	Recombination rate
$\pi(x)$	Recombination density at locus x
$\lambda_M(t)$	Coalescent rate, mutated subpopulation at time t
$\lambda_W(t)$	Coalescent rate, wildtype subpopulation at time t

interrupted by neutral mutations at x_k. Assuming that mutations at different marker loci and founder chromosomes are independent we find that

$$P(h|\mathscr{A}, h') = \prod_{j=1}^{n_G} \prod_{k=1}^{K} P(h_{jk}|\mathscr{T}(x_k), h'_{jk}), \tag{9}$$

where $P(h_{jk}|\mathscr{T}(x_k), h'_{jk}) = 1$ whenever $h_{jk} = \emptyset$.

Finally, the first term of (5), P($g|h$), can take values 1 or 0 depending on if all genotypes g_v are consistent with haplotypes h_{2v-1} and h_{2v}, i.e.

$$P(g|h) = \prod_{v=1}^{m} P(g_v|h_{2v-1}, h_{2v}) = \prod_{v=1}^{m} 1_{\{g_v \sim (h_{2v-1}, h_{2v})\}},$$

where $g_v \sim (h_{2v-1}, h_{2v})$ means that the genotypes of v at all K marker loci are consistent with the corresponding alleles obtained from h_{2v-1} and h_{2v}.

4 The Distribution of an Ascertained ARG

In this section, we provide the distribution of $\mathscr{A}|\mathscr{M}_0$, i.e. the ascertained ARG under complete penetrance. This is the only term of the likelihood expansion (5) of the previous section left unspecified so far.

Assumption (viii) together with (ix), stated in Appendix A, imply that for fixed G, $\{\mathscr{A}_u\}_{u=0}^{G}|\mathscr{M}_0$ is a Markov chain in discrete time, see Appendix A for details. We will need the corresponding result for continuous time t. To this end, we assume $t \in [0,1]$, counting time in units of G generations, so that t corresponds to Generation $u = [tG]$, defined as the smallest integer less than or equal to tG. We also introduce the notation $n(t) = n_{[Gt]}$, $n_M(t) = n_{M[Gt]}$, $n_W(t) = n_{W[Gt]}$ and $\mathscr{A}(t) = \mathscr{A}_{[Gt]}$. Figure 3a illustrates $\mathscr{A} = \{\mathscr{A}(t); 0 \le t \le 1\}$ in continuous time for $n = 12$ sampled chromosomes (not showing the diploid structure). Each vertex corresponds to a recombination or coalescence event and each edge e is a line of descent between two such events. As t increases, a coalescence event decreases $n(\cdot)$ by (at least) one, and a recombination event increases $n(\cdot)$ by one. The corresponding marginal coalescent tree at $x = 0.3$ is displayed in Figure 3b, and the IBD regions for the sample are displayed in Figure 3c.

The effect of labeling one founder lineage J of $\mathscr{A}$ as mutated is to spread this mutation to a number of other edges. In fact, by definition of an edge e, all chromosomes (from any generation) of e are either mutated or wildtype. We write $e \in \mathscr{M}$ and $e \in \mathscr{W}$ for these two cases, where $\mathscr{M}$ and $\mathscr{W}$ are the sets of mutated and wildtype chromosomes ancestral to the given sample.

We show in Appendix B that, under certain regularity conditions, going from discrete generations to continuous time, the Markov chain $\{\mathscr{A}_u\}_{u=0}^{G}|\mathscr{M}_0$ converges to to a Markov process $\{\mathscr{A}(t); 0 \le t \le 1\}|\mathscr{M}_0$ in continuous time as $G \to \infty$. The proof assumes the existence of three functions $\lambda_M, \lambda_W, p : [0,1] \to \mathbb{R}$ and a constant $\rho > 0$ such that

$$\begin{aligned} \textstyle\sum_{u=1}^{[Gt]} \log(1 - N_{Mu}^{-1})^{-1} &\longrightarrow \textstyle\int_0^t \lambda_M(s)ds, \\ \textstyle\sum_{u=1}^{[Gt]} \log(1 - N_{Wu}^{-1})^{-1} &\longrightarrow \textstyle\int_0^t \lambda_W(s)ds, \\ N_{M[Gt]}/(N_{M[Gt]} + N_{W[Gt]}) &\longrightarrow p(t), \\ Gr &\longrightarrow \rho \end{aligned} \tag{10}$$

DOI: 10.2202/1544-6115.1380

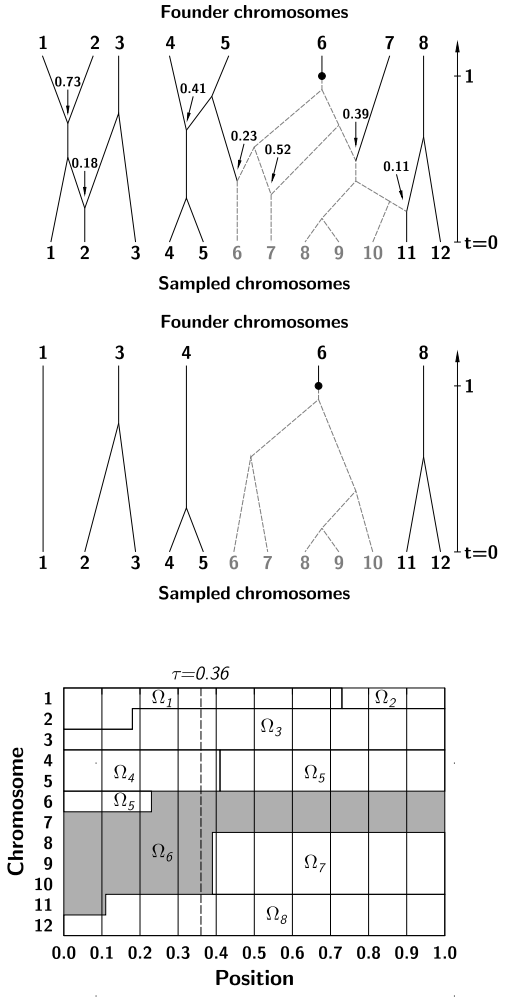

Figure 3: Representations of ancestry in continuous time for a sample of $n = 12$ chromosomes (not showing the diploid structure). The disease mutation is located at $\tau = 0.36$, on founder chromosome $J = 6$. Upper (3a): Ancestral Recombination Graph of the sample. The point of recombination is written above each recombination vertex. The sample contains five mutated chromosomes $\mathscr{M}_0 = \{6, \ldots, 10\}$, whereas the other 7 sampled chromosomes do not carry the mutation. The time-scale on the right measures time backward from today's sample until the founder generation. Middle (3b): Marginal coalescence tree at $x = 0.3$. Bottom (3c): IBD regions for the ARG of Figure 3a. The mutated region $\boldsymbol{\Omega}_J$ is displayed in grey. The location of the disease mutation $\tau = 0.36$, as well as 11 marker positions, $x_1 = 0$, $x_1 = 0.1, \ldots,$ $x_{11} = 1$ are displayed with vertical lines.

for all $0 \leq t < 1$ as $G \to \infty$. We refer to ρ as the recombination rate for a single lineage, and $\lambda_M(t)$ and $\lambda_W(t)$ are the coalescence rates of pairs of mutated and wildtype lineages. $p(t)$ is the disease allele frequency at time t and $p(0) = p$. Given $n_M(t) = k$ and $n_W(t) = l$, the coalescence rate for *some* pair of mutated lineages is $\binom{k}{2}\lambda_M(t)$ and the coalescence rate for *some* pair of wildtype lineages $\binom{l}{2}\lambda_W(t)$. If the former event occurs at time t, $n_M(\cdot)$ decreases by one at t, and in the latter case $n_W(\cdot)$ does so. The rate of recombination for *some* of the mutated lineages is $k\rho$ and similarly $l\rho$ for *some* of the wildtype lineages. Suppose a recombination occurs at time t for lineage e at X, with parental lineages e_1 and e_2 transmitting genetic matrial along $[0,X)$ and $[X,1]$ respectively. Then

$$\begin{array}{rcl} p(e_1 \in \mathscr{M} | e \in \mathscr{M}) & = & 1_{\{X>\tau\}} + p(t)1_{\{X\leq\tau\}}, \\ p(e_2 \in \mathscr{M} | e \in \mathscr{M}) & = & p(t)1_{\{X>\tau\}} + 1_{\{X\leq\tau\}}, \\ p(e_1 \in \mathscr{M} | e \in \mathscr{W}) & = & p(t)1_{\{X\leq\tau\}}, \\ p(e_2 \in \mathscr{M} | e \in \mathscr{W}) & = & p(t)1_{\{X>\tau\}}. \end{array} \tag{11}$$

Example 1 (Linear growth rate.) Suppose $N_{Mu} = 1 + a(G-u)$ and $N_{Wu} = cG + b(G-u)$ for some positive constants a, b, c. Here c is the haploid size of the wildtype founder population and a (b) the linear growth rate of the mutated (wildtype) population in units of G. It is easily seen that

$$\begin{array}{rcl} \lambda_M(t) & = & 1/(a(1-t)), \\ \lambda_W(t) & = & 1/(c+b(1-t)), \\ p(t) & = & a(1-t)/(c+(a+b)(1-t)). \end{array} \tag{12}$$

□

Example 2 (Exponential growth rate.) Suppose $N_{Mu} = \exp(a(G-u))$ and $N_{Wu} = (N_G - 1)\exp(b(G-u))$, where a (b) is the relative increase of population size per generation for the mutated (wildtype) subpopulation. In this case $\lambda_M = \lambda_W \equiv 0$ and

$$p(t) = \begin{cases} 1, & a > b, \\ 1/N_G, & a = b, \\ 0, & a < b. \end{cases}$$

□

After the ARG has been constructed, neutral mutations (vi) are added independently at the marginal coalescence trees at all loci. To get a non-trivial limit process in continuous time, we assume that

$$Gu_k \to \mu_k \text{ as } G \to \infty, \quad k = 1, \ldots, K. \tag{13}$$

DOI: 10.2202/1544-6115.1380

Then, in the limit, neutral mutations at x_k occur along the edges of $\mathscr{T}(x_k)$ according to a Poisson process with intensity μ_k.

5 Model Simplifications

In this section, we consider model simplifications that make exact computation of the retrospective likelihood (5) feasible, at least for moderately small data sets. We assume:

(x) The limiting proportion of mutated chromosomes in (10) is $p(t) = 0$ for $0 \leq t \leq 1$.

(xi) The wildtype coalescence rate $\lambda_W(t) = 0$ for $0 \leq t \leq 1$.

(xii) The mutated coalescence rate $\lambda_M(t) = 0$ for $0 \leq t < 1$.

(xiii) Linkage equilibrium (LE) in the founder population, i.e. $f(h'_j) = \prod_{k=1}^{K} f_k(h'_{jk})$, where f_k is the founder allele frequency at x_k.

(xiv) All markers are biallelic SNPs.

Condition (x) means that the disease allele is rare in the recent and previous generations. For consistency, this requires that we use a small value of p when calculating each term of $P(\mathscr{M}_0|Y)$ in (7). Conditions (xi)-(xii) state that wildtype chromosomes never coalesce and that all mutated chromosomes coalesce simultaneously at time $t = 1$. This is a good approximation if the subpopulations of wildtype and mutated chromosomes are both rapidly increasing (cf. Example 2).

Figure 4a shows an ARG satisfying (x)-(xii) and Figure 4b the corresponding IBD decomposition $\{\Omega_j\}_{j=1}^{n(1)}$. According to (xii), the subtree of $T(\tau)$ with $n_M = n_{M0}$ mutated edges has a star topology, i.e. they all coalesce at time $t = 1$. By (xi), there are no coalescence events between wildtype edges and (x) implies that any mutated edge of $\mathscr{A}$ that is parental in a recombination event must transmit a region containing τ to the child edge. Therefore, the n_M rows of Ω_J are all connected intervals containing τ, showing the portion of each $i \in \mathscr{M}_0$ that has descended from J. All other Ω_j, $j \neq J$ occupy a single row. Introduce the set

$$D = \{(i,k); 1 \leq i \leq n, 1 \leq k \leq K, (i,x_k) \in \Omega_J\}$$

of marker alleles which are inherited from the mutated founder chromosome. A more explicit definition of D, in terms of so called Nearest Recombination Events (McPeek and Strahs, 1999, Morris et al, 2002) is given in Appendix C.

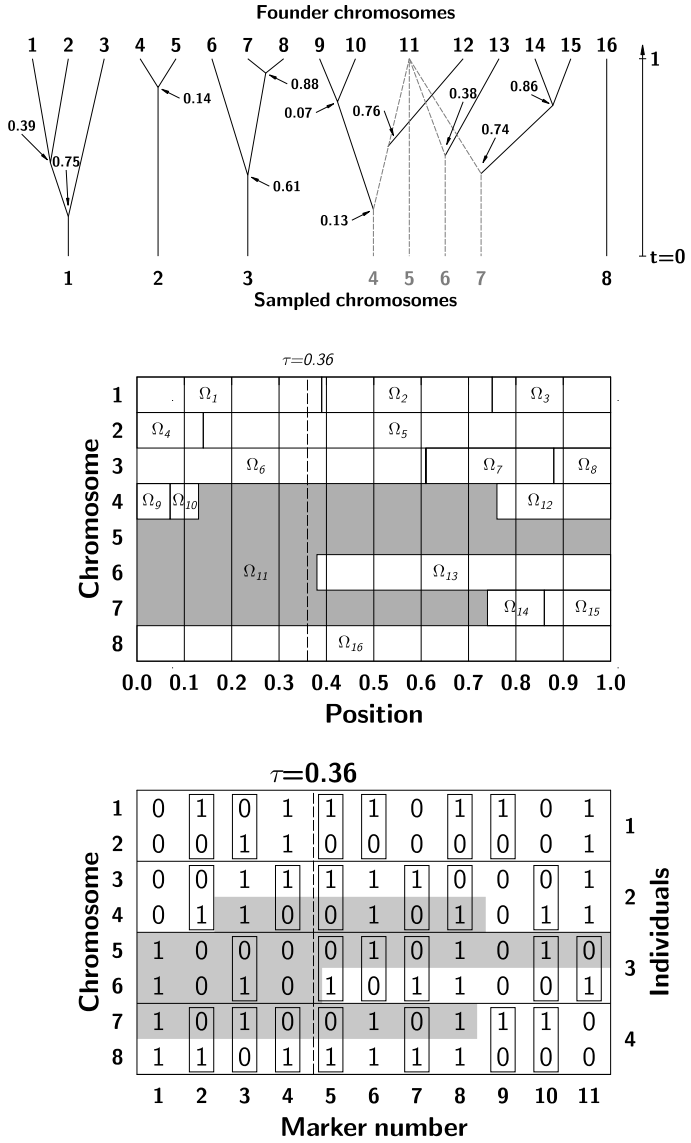

Figure 4: Ancestry fulfilling (x)–(xii) in continuous time for $n = 8$ chromosomes (not showing the diploid structure), with mutation at $\tau = 0.36$ on Founder chromosome $J = 11$. Upper: ARG. Middle: IBD regions for the ARG, with mutated region Ω_J displayed in grey. τ and 11 marker positions, $0, 0.1, \ldots, 1$ are vertically displayed. Lower: Example of marker haplotypes illustrating H, the set of heterozygous sites, as the union of the marked boxes. The mutated region D is displayed in grey (note that Chromosome 5 has mutated at marker 3). For each marker k, the set H_k consists of those chromosomes at heterozygous sites that belong to the k^{th} column of D, but where the homologous site does not. Thus $H_1 = \emptyset$, $H_2 = \{7\}$, $H_3 = \{7\}$, $H_4 = \{4,7\}$, $H_5 = \{4,5,7\}$, $H_6 = \{5\}$, $H_7 = \{4,5,7\}$, $H_8 = \{4\}$, $H_9 = \emptyset$, $H_{10} = \{5\}$, $H_{11} = \{5\}$. Further it follows that $n_{3H} = 1$, $n_{30} = 1$ and $n_{31} = 2$, etc.

DOI: 10.2202/1544-6115.1380

The likelihood computations simplify considerably using (x)-(xiv). It turns out that D contains all information about $\mathscr{A}$ needed for calculating (5), which can be written as

$$L(x) = \sum_{h,h',D} P(g|h)P(h|D,h')f(h')P_x(D|Y), \tag{14}$$

where $h' = h'_J = (h'_{J1},\ldots,h'_{JK})$ is the haplotype of the mutated founder and expressions for $P(h|D,h')$ and $P_x(D|Y)$ are given in Appendix D.

Another consequence of (x)-(xiv) is more effective simulation of LD structure under non-random ascertainment, as further discussed in Sections 7 and 8.3.

6 LOD Score Computation and Approximation

In this section, we will use the simplified likelihood (14) to obtain exact expressions for the likelihood ratio and lod score. We start by summing out h in (14) and dividing by $L(\infty)$ to obtain an expansion

$$\mathrm{LR}(x) = \sum_{h',D} \mathrm{LR}(h',D)f(h')P_x(D|Y) \tag{15}$$

of the likelihood ratio at x, where

$$\mathrm{LR}(h',D) = \frac{P(g|h',D)}{L(\infty)} \tag{16}$$

can be interpreted as the likelihood ratio when D and h' are known.

We will derive a very explicit expression for $\mathrm{LR}(h\prime,D)$. Let H be the set of heterozygous sites (i,k), i.e.

$$H = \cup_{k=1}^{K} \cup_{v=1}^{m} H_{vk},$$

where $H_{vk} = \{(2v-1,k),(2v,k)\}$ if g_{vk} is heterozygous ($h_{2v-1,k} \neq h_{2v,k}$) and H_{vk} is empty otherwise. Let H_k consist of those heterozygous sites that belong to the k^{th} column of D but the homologous site (i.e., the member of the same H_{vk}) does not. Then the k^{th} column of D has $n_{kH}+n_{k0}+n_{k1}$ elements, where $n_{kH} = |H_k|$, $n_{k0} = |\{i;\,(i,k) \in D \setminus H_k, h_{ik} = 0\}|$ and $n_{k1} = |\{i;\,(i,k) \in D \setminus H_k, h_{ik} = 1\}|$, see Figure 4c for en example. As shown in Appendix E, the likelihood ratio (16) can be written as

$$\begin{aligned} \mathrm{LR}(h',D) &= \prod_{k=1}^{K} \left(\left(P(0|h'_{Jk})/\tilde{f}_k(0)\right)^{n_{k0}} \left(P(1|h'_{Jk})/\tilde{f}_k(1)\right)^{n_{k1}} \right. \\ &\quad \left. \cdot \left(0.5P(0|h'_{Jk})/\tilde{f}_k(0) + 0.5P(1|h'_{Jk})/\tilde{f}_k(1)\right)^{n_{kH}}\right). \end{aligned} \tag{17}$$

where

$$\tilde{f}_k(a) = (1-q_k)f_k(a) + q_k f_k(1-a), \quad a = 0,1, \tag{18}$$

is the allele frequency of a at locus x_k and $q_k = (1-\exp(-2\mu_k))/2 \approx \mu_k$ is the

mutation probability at x_k, i.e. the probability that a founder allele at time $t = 1$ mutates an odd number of times down to $t = 0$, cf. (13) and (xiv). If μ_k is small, $\tilde{f}_k(a) \approx f_k(a)$.

Intuitively, the k^{th} term of (17) is large when there are no allelic mismatches between the chromosomes that have received genetic material at x_k from the mutated founder chromosome, i.e. when either n_{k0} or n_{k1} is zero. When this is not the case, the number of mismatches $\min(n_{k0}, n_{k1})$ at x_k penalizes the k^{th} term of (17) to an extent that depends on the mutation probability q_k.

6.1 Conditioning on Founder Haplotypes

In this approach we sum out D in (15) and write

$$\text{LR}(x) = \sum_{h'} \text{LR}(x; h') f(h'), \tag{19}$$

where $\text{LR}(x; h') = P_x(g|h', Y)/L(\infty)$ is the likelihood ratio when conditioning on missing data h'. Let $D_{\{v\}} = D \cap (\{2v-1, 2v\} \times \{1, \ldots, K\})$ denote the set of mutated sites (i,k) for Individual v. Then (15)-(17) imply

$$\text{LR}(x; h') = \prod_{v=1}^{m} \sum_{D_{\{v\}}} \text{LR}(h', D_{\{v\}}) P_x(D_{\{v\}}|Y_v) \tag{20}$$

where $\text{LR}(h'; D_{\{v\}})$ is the likelihood ratio obtained when conditioning on hidden data $(h', D_{\{v\}})$, i.e. replacing D by $D_{\{v\}}$ in (17). The crucial point is that conditionally on h' and Y, the rows of g are independent and hence LR can be written as a product of m terms. It is shown in Appendix F that each term of the outer product can be calculated with $O(K)$ operations, using a recursive Hidden Markov Model (HMM) algorithm. Hence the total complexity is $O(mK2^K)$ for evaluating $\text{LR}(x)$. This is a marked improvement compared to direct summation over h' and D, but still not feasible for large K. For large K, we may use a sliding window of $l < K$ marker loci. The window width l is chosen to make the computational complexity $O(ml2^l K)$ feasible.

To obtain p-values for the test, a permutation algorithm was proposed in (3). In general permutation tests are very computationally intensive, which constrict their practical applicability for tests that are already computationally demanding, such as ours. In the general setting, the test quantity must be calculated for each of the Q random permutations, which would give computational complexity $O(mK2^K Q)$. Since Q must be large, typically tens or hundreds of thousands, this is not feasible. However, in the case of binary phenotypes, we propose a procedure for the permuta-

DOI: 10.2202/1544-6115.1380

tion testing which reduces the computational demand. The algorithm exploits that $\sum_{D_{\{v\}}} \mathrm{LR}(h', D_{\{v\}}) P_x(D_{\{v\}}|Y_{\gamma(v)})$ is the same for all permutations where $Y_{\gamma(v)} = 1$, and similarly for all permutations where $Y_{\gamma(v)} = 0$. Thus, for each individual v the HMM must only be calculated twice, to obtain values of $\sum_{D_{\{v\}}} \mathrm{LR}(h', D_{\{v\}}) P_x(D_{\{v\}}|Y_v = 1)$ and $\sum_{D_{\{v\}}} \mathrm{LR}(h', D_{\{v\}}) P_x(D_{\{v\}}|Y_v = 0)$ respectively.

To obtain p-values the summation over h' and multiplication over v in (19) and (20) must be carried out for each of the Q permutations. The total complexity is thus $O(m2^K(2K+Q))$. Since typically $Q >> K$ the total complexity including permutation testing is $O(m2^K Q)$.

To estimate Z_{max}, LOD score is calculated at several positions $\tilde{x}_i,\ i = 1, \ldots, N_x$ within the interval $[0,\ 1]$, and $Z_{max} = \max_{i=1,\ldots,N_x} Z(\tilde{x}_i)$. As LOD score is calculated separately at each position, the total complexity is $O(m2^K Q N_x)$.

6.2 Conditioning on IBD Regions

In this approach we sum out h' in (15) and write

$$\mathrm{LR}(x) = \sum_D \mathrm{LR}(D) P_x(D|Y), \tag{21}$$

where $\mathrm{LR}(D) = P(g|D)/L(\infty)$ is the likelihood ratio obtained when conditioning on missing data D. This yields

$$\begin{aligned} \mathrm{LR}(D) \quad = \quad & \prod_{k=1}^K \big(a_k^{n_{kH}} ((1-q_k)/\tilde{f}_k(0))^{n_{k0}} (q_k/\tilde{f}_k(1))^{n_{k1}} f_k(0) \\ & + b_k^{n_{kH}} (q_k/\tilde{f}_k(0))^{n_{k0}} ((1-q_k)/\tilde{f}_k(1))^{n_{k1}} f_k(1) \big), \end{aligned} \tag{22}$$

$a_k = 0.5(1-q_k)/\tilde{f}_k(0) + 0.5 q_k/\tilde{f}_k(1)$ and $b_k = 0.5 q_k/\tilde{f}_k(0) + 0.5(1-q_k)/\tilde{f}_k(1)$.

In Appendix F, we describe a HMM algorithm for evaluating (21) with complexity $O(K2^{2m})$. This is not feasible for all but very small m, so we propose using a pseudo likelihood

$$\mathrm{PL}(x) = \prod_{V \in \mathscr{V}} L(x; V), \tag{23}$$

where $L(x;V)$ is the retrospective likelihood using only individuals from $V \subset \{1, \ldots, m\}$ and $\mathscr{V}$ a given collection of subsets V. The pseudo likelihood ratio and pseudo LOD score obtained from (23) are

$$\mathrm{PLR}(x) = \prod_{V \in \mathscr{V}} \mathrm{LR}(x; V)$$

and

$$\begin{aligned} \mathrm{PZ}(x) &= \tfrac{1}{|\mathscr{V}|}\log_{10}(\mathrm{PLR}(x)) \\ &= \tfrac{1}{|\mathscr{V}|}\textstyle\sum_{V\in\mathscr{V}} Z(x;V). \end{aligned}$$

For instance, if $\mathscr{V}$ contains all subsets of $\{1,\ldots,m\}$ of size $m_0 \geq 2$, we get computational complexity $O(Km^{m_0}2^{2m_0})$, which, for values of m of practical interest, is feasible only when m_0 equals 2.

When a permutation test is used to calculate the p-values the procedure would in general require $O(Km^{m_0}2^{2m_0}Q)$ operations for the PLOD score. For $m_0 = 2$ this is $O(Km^2 2^4 Q) = O(Km^2 Q)$. However, for binary phenotypes an effective algorithm can be developed, just as for the LOD score. The basis of this algorithm is that $\mathrm{LR}(x;V) = \sum_{D_V} \mathrm{LR}(D_V)P_x(D_V|Y_V)$ is constant for all permutations with the same Y_V. Here D_V is notation for which markers that are inherited IBD (from the mutated founder) for the individuals *within* subset V. As $\mathscr{V}$ consists of subsets of size 2, Y_V can take only four possible values, $Y_V = (0\ 0)$, $Y_V = (0\ 1)$, $Y_V = (1\ 0)$ or $Y_V = (1\ 1)$. Thus $\mathrm{LR}(x;V)$ must be calculated for each of these four cases, and for each permutation only the multiplication over all subsets in $\mathscr{V}$ remains. The complexity is thus $O(m^2(4K2^4 + Q))$. Since typically $Q \gg 2^6 K$ the complexity becomes $O(m^2 Q)$. To estimate $\mathrm{PZ}_{\max}$, PLOD score is calculated at several positions $\tilde{x}_i,\ i = 1,\ldots,N_x$ within the interval $[0,\ 1]$, and $\mathrm{PZ}_{\max} = \max_{i=1,\ldots,N_x} \mathrm{PZ}(\tilde{x}_i)$. As the PLOD score is calculated separately at each position, the total complexity is $O(m^2 Q N_x)$.

6.3 Software

The algorithms for simulation and calculation of the LOD and PLOD scores have been coded in Matlab. The algorithms, with inbuilt documentation, are available upon request from the authors.

7 Simulation Study and Real Data Analysis

To evaluate the performance of the proposed LOD and PLOD scores we carried out simulation studies and a real data analysis.

7.1 Simulating from the Ascertained ARG

As previously pointed out, the ascertained ARG is a powerful tool for simulation of case-control samples. For a prescribed number of cases and controls, the mutational status for each of a person's two alleles at the disease locus is simulated conditional on the person's disease status, according to (6). By simulation of re-

DOI: 10.2202/1544-6115.1380

combinations and coalescence of mutated and wildtype lineages, superimposed by neutral mutations, the marker alleles are obtained. Simulations in this subsection are obtained under the simplifications in Section 5. Then all coalescence events are deterministic and the only recombination events that need to be simulated are the nearest recombination events.

As an example of a commonly used genetic model, we account for simulations with multiplicative penetrance and binary phenotype. With genotype relative risk ratio λ, we have $\psi_1/\psi_0 = \psi_2/\psi_1 = \lambda$, where ψ_j is the probability that an individual with j disease alleles becomes affected. (Then $\psi_{vj} = \psi_j$ for all cases ($Y_v = 1$) and $\psi_{vj} = 1 - \psi_j$ for the controls ($Y_v = 0$).) The markers are equispaced in the interval $[0,1]$ (with $x_1 = 0, \ldots, x_K = 1$), with minor marker allele frequency $f_k = 0.5$ at all markers $k = 1, \ldots, K$. From the founder generation until today, the marker mutation rate is $q_k = q = 0.001$ and recombination rate in the interval is $\rho = 1.5$. In all simulations in this subsection the disease locus is positioned at $\tau = 0.36$, which is not a marker position. All parameter values can be chosen arbitrarily, although their values affect the power to detect association. Considering mutations that arose typically some hundred generations ago, the mutation rate 0.001 per marker is unrealistically high for SNPs, but still does not undermine the performance of our LOD score. On the other hand, the marker allele frequencies $f_k = 0.5$ are unrealistic to our favour. The accompanying decrease in sample size, that is made possible, is welcome for the computer demanding studies of power that we present here. However it does not change the fundamental behaviour of the LOD score, compared to arbitrary marker allele frequencies. We further test our algorithms for parameter values that do not fulfill all conditions of the approximation. In particular, the disease allele frequency is too high in the first simulation, and in that way more similar to what is assumed in real studies. (To pick up associations for diseases with weak penetrance would need unrealistically large samples if the disease allele frequency was very low.)

Each simulated data set is analyzed with the LOD score (19) and/or PLOD score (23), the latter with subsets of $m_0 = 2$ individuals. (For some data sets either of the methods is unfeasible due to the computational demand). To evaluate the performance, the p-value of the test statistic Z_{max} is found by permutation testing (3).

7.1.1 Comparisons with Single Marker Tests

For comparison we also calculated a global p-value based on carrying out Cochran-Armitage tests at each marker $k = 1, \ldots, K$. The global p-value is calculated in a similar way to the LOD and PLOD p-values (3), i.e. the observed maximum Cochran-Armitage test statistic, over the K markers, is compared to the maximum values obtained under the null (outcome-permuted) distribution.

7.1.2 LOD Score and PLOD Score

In three independent samples of 200 cases and 200 controls at $K = 5$ markers, the disease allele frequency was $p = 0.2$, relative risk ratio $\lambda = 3$, and baseline prevalence $\psi_0 = 0.0001$. Figure 5 displays the LOD and PLOD scores of the three data sets, each calculated at $N_x = 20$ equidistant locations interior of $[0,1]$. To be able to detect associations with a p-value as small as 10^{-4}, 100000 permutations were performed.

For the LOD score, the association is very clearly picked up by Z_{max}, and further it is clear that the largest $Z(x)$ is found close to the true maximum $\tau = 0.36$. Although the shape of the PLOD score curve is similar to that of LOD score, with its maximum close to $\tau = 0.36$, the p-values are considerably higher for PLOD. The p-value calculations further show that permutation testing is necessary, since the asymptotic χ^2-approximation (in which case LOD=3 corresponds to p-value 0.0002, which is commonly used to establish linkage) is not valid neither for LOD nor PLOD. Global Cochran-Armitage tests yielded p-values of 0.00028, 0.13032 and 0.00329 for the three data sets. For all three data sets p-values are smallest for the LOD score. For two of the three data sets the p-value from the PLOD score is smaller than the Cochran-Armitage test.

For the second simulated data set, presented above, we re-calculated LOD scores under mis-specification of disease allele frequency, p, and recombination rate, ρ. Neither the PLOD (correctly specified parameter values) nor the Cochran-Armitage test provided significant evidence of association for this data set. We wanted to see whether the LOD score could out-perform these two tests even under mis-specification of nuisance parameters. We recalculated LOD scores mis-specifying each parameter p and ρ, one-at-a-time, by multiplying the true parameter value by factors of both less than and greater than 1. Results are displayed in Figure 6. The p-values remain under 0.01 under all four parameter mis-specifications. Maximum LOD scores also remain close to the true disease location.

7.1.3 Power Calculations

To estimate the power of the tests we have performed tests for multiple simulated data sets. The results are plotted as a Receiver Operating Characteristic (ROC), i.e. power vs. significance level, see e.g. Bradley (1996). For each of N independent simulations from the genetic model, a p-value $\hat{\alpha}_i,\ i = 1,\ldots,N$, is estimated from the results of Q random permutations as in (3). The power β as a function of α could then be estimated by Monte Carlo as

$$\hat{\beta}(\alpha) = \frac{1}{N}\sum_{i=1}^{N} 1_{\{\hat{\alpha}_i < \alpha\}}$$

DOI: 10.2202/1544-6115.1380

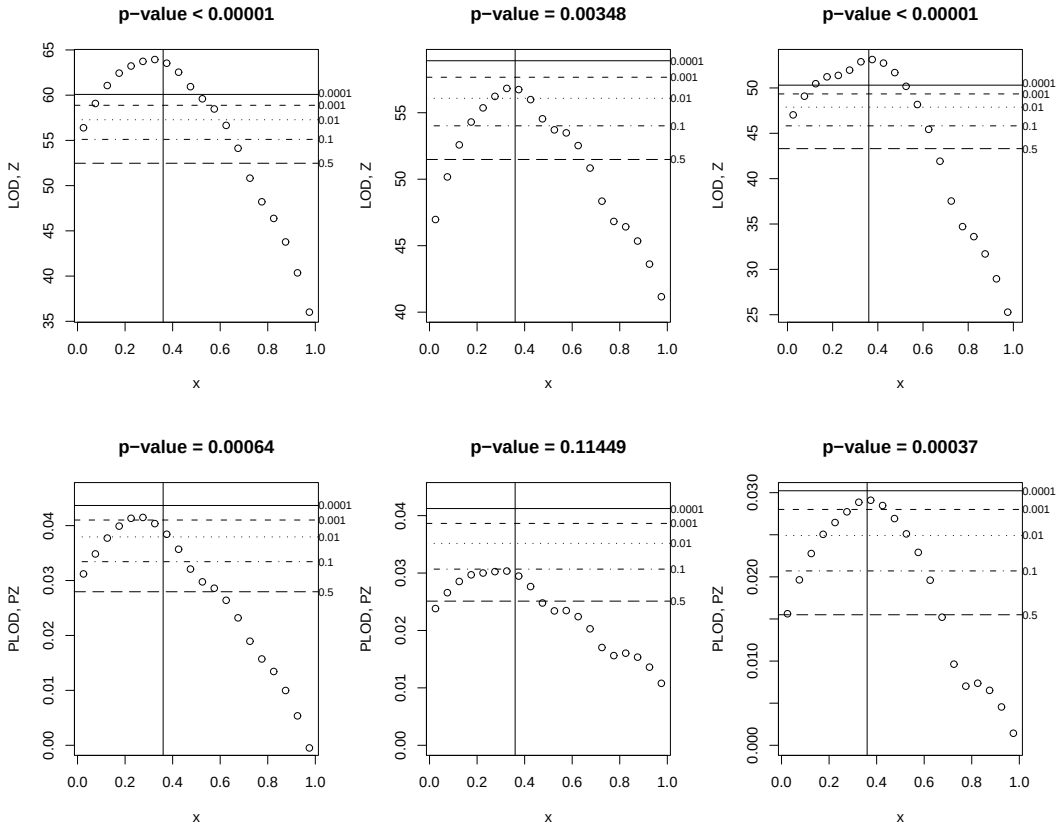

Figure 5: LOD and PLOD score calculated for three simulated data sets, at $N_x = 20$ positions along $[0,1]$. The genetic model is multiplicative penetrance with relative risk 3, disease allele frequency 0.2, marker allele frequency $f_k(1) = 0.5$ and mutation probability $q_k = 0.001$ $k = 1,\ldots,5$. 200 cases and 200 controls were simulated and analyzed. Within columns, LOD and PLOD are calculated for the same data set, and quantiles are estimated with the same random permutations. The horizontal lines show the critical limits for Z_{max} for different significance levels α (displayed on the right y-axes). Marker positions are indicated with dotted, vertical lines, and the true disease location $\tau = 0.36$ with a solid vertical line.

The ROC displayed in Figure 7 is an estimation based on 100 simulated data sets with $K = 10$ markers for 200 cases and 200 controls. The model parameters were $\lambda = 2$, $p = 0.1$ and $\psi_0 = 0.0001$. Each p-value was estimated from $Q = 10000$ permutations, and thus p-values larger than 10^{-3} could be estimated accurately. To cut down computation time for the ROC, while not altering the test performance, $Z(x)$ and $\mathrm{PZ}(x)$ were only calculated for $x = 0.2, 0.3, \ldots, 0.6$, i.e. $Z_{\max}$ and $PZ_{\max}$

were based on (P)LOD score at $N_x = 5$ non-marker positions around τ. We also constructed ROC curves for LOD scores under mis-specified values of p and ρ (indicated as LOD(p=0.3), LOD(rho=3) and LOD(rho=0.75) in Figure 7. We did this in order to confirm that the robustness of the LOD score to mis-specification of parameter values extends more generally, beyond what was observed for the example data set used in Figure 6. As the LOD score (with or without mis-specified p, λ) has a steeper ROC for small α it has better performance than both the PLOD score and the (global) Cochran-Armitage test. The power of the PLOD score is similar to the power of the Cochran-Armitage test in this example. Despite the relatively weak model, all tests turn out positively in a comparison with the baseline power $\beta(\alpha) = \alpha$, corresponding to a test that cannot discriminate between H_0 and H_1.

As a last example, Figure 8 displays the ROC for PLOD score from 100 runs with $Q = 10000$ permutations for the same genetic model as in Figure 7 ($p = 0.1$, $\lambda = 2$ and $\psi_0 = 0.0001$). The data set now consists of $K = 25$ markers for 200 cases and 200 controls. $PZ_{\max}$ is calculated from PZ at the same 5 positions in the vicinity of τ, as in Figure 7. The LOD score is not computationally tractable, but to use all 25 markers is possible with the PLOD score approximation. Comparison with Figure 7 shows that the performance of PLOD score has improved, but that it still gives worse results than LOD-score did with $K = 10$ markers. In other simulations with higher disease allele frequency $p = 0.2$ the quality of PLOD calculated from $K = 25$ markers almost matched that of LOD score with $K = 10$. For situations where calculation of LOD-score is not feasible, PLOD-score could be a potentially useful approximation, as long as the number of included individuals is not too large.

7.1.4 Time Consumption

Table 2 contains the mean computation times for the accounted LOD and PLOD scores. Computations were performed on one of the processors of a fast computer, a AMD Athlon(tm) 64 X2 Dual Core Processor 5000+ with 2.6GHz processor and total memory 2GB.

The computation times include permutation testing, and although highly dependent on the computer used, they demonstrate that the tests are feasible even for quite large data sets with many markers. There is potential for considerable decrease in computation time if implementation is done in a program language, instead of Matlab. We are currently working on an implementation in Fortran. A comparison between the empirical results and the theoretical complexity calculations of Sections 6.2 and 6.1 reveals quite large deviations. We believe this is mainly an artifact of memory constraints, which prevent us from proper vectorization of the code for the permutation test of the PLOD score. Also this would be avoided in a programming language that is effective for heavy computations.

DOI: 10.2202/1544-6115.1380

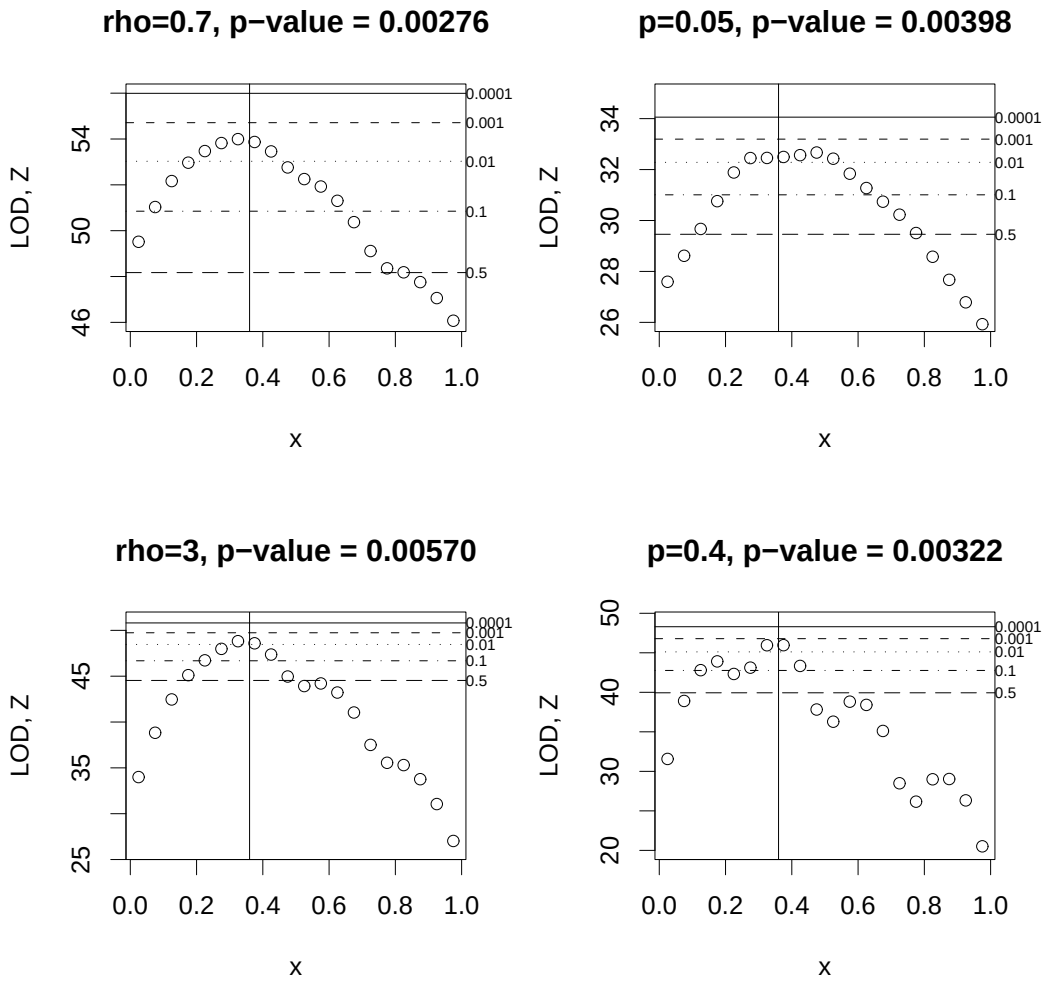

Figure 6: LOD scores calculated for the second simulated data set under the genetic model described in Figure 5 (column 2). That is, p=0.2 and ρ=1.5 for the true genetic model. Columns 1 and 2 display LOD scores calculated under misspecification of ρ (0.7,3) and p (0.05, 0.4), respectively.

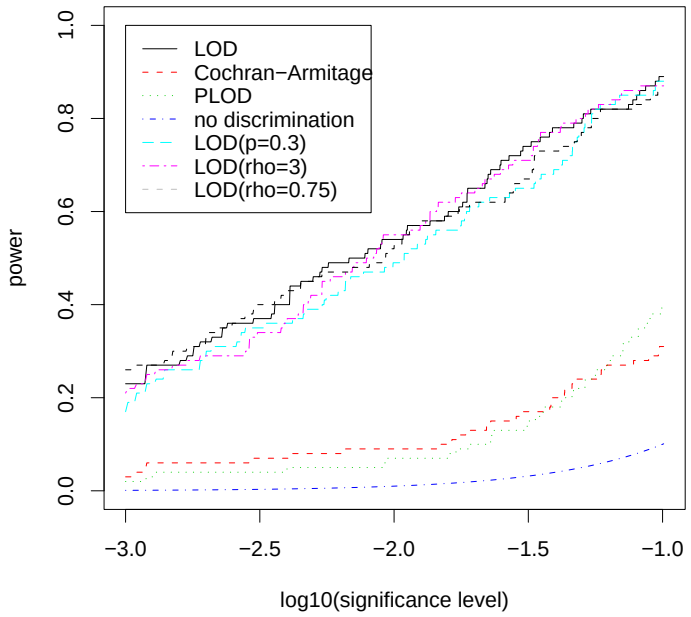

Figure 7: Estimated ROC curves calculated from $N = 100$ p-values, each calculated from $Q = 10000$ permutations. Genetic model is multiplicative with relative risk $\lambda = 2$, disease allele frequency $p = 0.1$, $K = 10$, $\rho = 1.5$, marker allele frequency $f_k(1) = 0.5$ and mutation probability $q_k = 0.001$ $k = 1,\ldots,10$. 200 cases and 200 controls were simulated and analyzed.

Table 2: Mean computation times for different sample sizes. The mean computation times for LOD and PLOD are measured in seconds.

parameters				time (s)		
m	K	N_x	Q	LOD	PLOD	Figure
400	5	20	100000	150	61000	Figure 5
400	10	5	10000	1000	1600	Figure 7
400	25	5	10000	—	1800	Figure 8

7.2 Simulating from an Alternative Genetic Model

To assess whether our ascertained ARG LOD score can perform well under a misspecified genetic model we performed a small simulation study. We used the software MS (Hudson, 2002) which simulates haplotypes under a Wright-Fisher neutral model of genetic variation. Since this model does not provide an automatic mechanism to simulate under non-random ascertainment, we deployed an extra level of

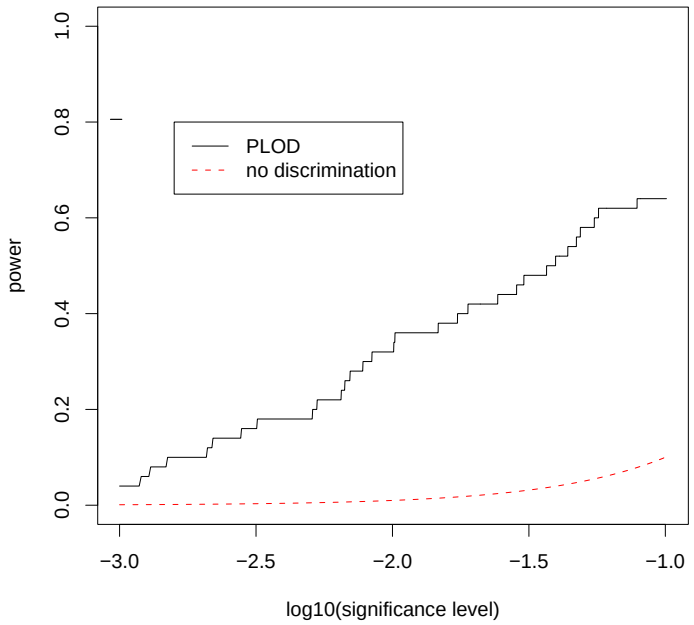

Figure 8: Estimated ROC curve for the PLOD score calculated from $N = 100$ p-values, each calculated from $Q = 10000$ permutations. Genetic model is multiplicative with relative risk $\lambda = 2$, disease allele frequency 0.1, $K = 25$, $\rho = 1.5$, marker allele frequency $f_k(1) = 0.5$ and mutation probability $q_k = 0.001$ $k = 1,\ldots,25$. 200 cases and 200 controls were simulated and analyzed.

simulation (of phenotypes) to mimic case-control sampling, the study design on which our ascertained ARG model is based. Specifically we simulated 6000 haplotypes under a model with mutation parameter value equal to 10, crossover rate parameter equal to 100, number of sites=100000 (using the parameters defined in MS; Hudson, 2002). After standardising the region length to a [0,1] interval we selected the segregating site within the region [0.35,0.65] which had the MAF closest to 0.2, as a disease locus. Haplotypes were grouped in pairs to represent 3000 individuals, whose case-control phenotype was simulated from the disease locus genotype with penetrances $\psi_0 = 0.25$, $\psi_1 = 0.38$ and $\psi_2 = 0.53$. Ten markers were then selected as the SNPs (MAF>0.03 and excluding the disease SNP) which were closest to positions 0.05, 0.15,..., 0.95. From the pool of 3000 individuals, 500 cases and 500 controls were then selected randomly. Ten data sets were simulated in this manner. LOD scores and Cochran-Armitage test statistics were estimated using the final (simulated) data sets of 1000 individuals. For both tests, p-values were calculated based on 100000 permutations. For calculating the LOD score we assumed (nuisance) parameter values of $p = 0.05$, $\rho = 1$, $\mu = 0.0005$, $\psi_0 = 0.05$

and $\lambda = 1.8$. LOD scores were calculated at 10 locations, spaced regularly across the [0,1] region. Results are displayed as ROCs in Figure 9. Although the number of simulated data sets is small, it is reasonable to conclude that the association test based on the ascertained ARG can compare favourably to the (global) Cochran-Armitage test even under a mis-specified genetic model.

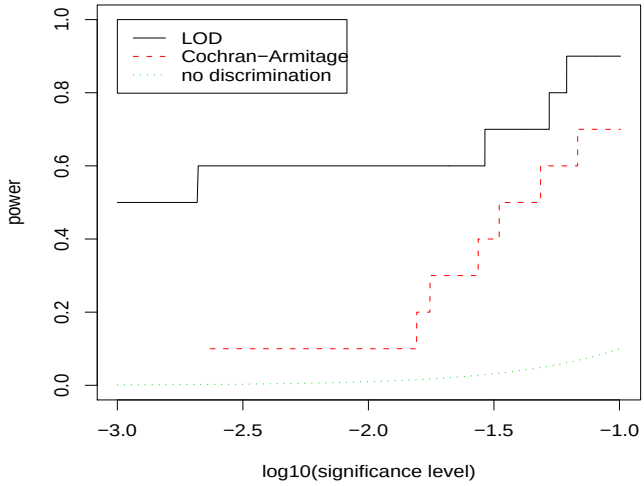

Figure 9: LOD score and Cochran-Armitage (C-A) test results, presented as estimated ROC curves, for ten data sets simulated under the neutral Wright-Fisher model.

7.3 Real Data Analysis

During 2007 two (independent) genome wide association scans (Hunter et al., 2007 and Easton et al., 2007) identified SNPs in the FGFR2 (Fibroblast growth factor receptor 2) gene to be associated with the risk of developing breast cancer (e.g. rs2981582 with 95% confidence interval for the per allele OR estimated as (1.23-1.30); Easton et al., 2007). We present here an analysis of SNPs in this gene, for 400 Swedish postmenopausal breast cases and 400 healthy Swedish controls. These cases and controls are part of an ongoing genome-wide association study and have been selected randomly from a larger sample of women which has been described in detail elsewhere (Einarsdóttir et al., 2006). We first selected a region of approximately 160,000 base pairs, within and downstream of FGFR2, containing 38 SNPs. The p-values from Cochran-Armitage trend tests are plotted in Figure 10a.

DOI: 10.2202/1544-6115.1380

The smallest (marginal) p-value was 0.0014, for rs1219648. The results from this subset of individuals closely resembles those displayed in Figure 2 of Hunter et al. (2007). We evaluated our ARG LOD score for a window of 10 adjacent SNPs, within which Cochran-Armitage test results show the strongest signal of association, both in our data and in Figure 2 of Hunter et al. (2007). We evaluated the LOD score at 19 equidistant locations within the selected region, assuming (nuisance) parameter values of $p = 0.2$, $\rho = 0.8$, $\mu = 0.0001$, $\psi_0 = 0.005$ and $\lambda = 1.3$. We obtained global p-values of 0.0102 and 0.0029 for tests based on the Cochran-Armitage and LOD scores, respectively. Results are displayed in Figure 10.

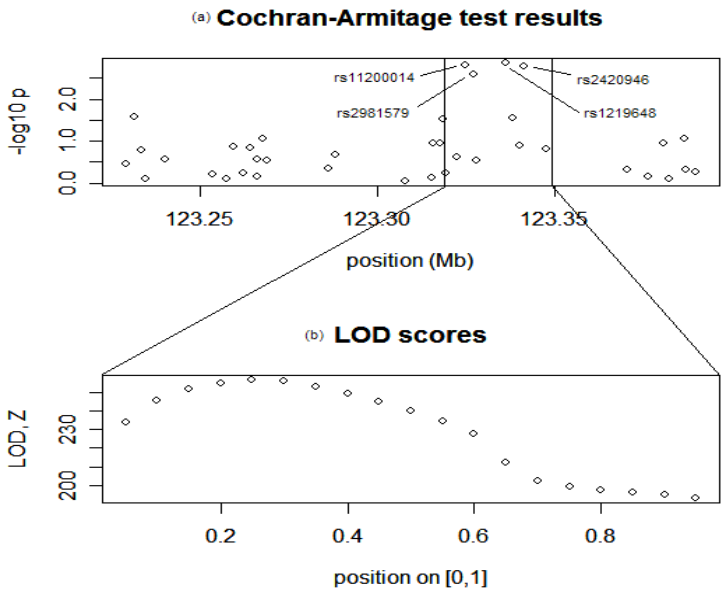

Figure 10: LOD score and Cochran-Armitage (C-A) test results for association between SNPs in FGFR2 and risk of postmenopausal breast cancer. LOD scores are evaluated within a region, standardised to [0,1] in (b), which is indicated in (a).

8 Discussion

8.1 Summary

In this article we have defined an ascertained version of the ARG, involving two subpopulations of mutated and wildtype chromosomes. We have developed a novel asymptotic scheme for such ARGs, conditioning on time of disease mutation G and subpopulations sizes, counting time in units of G generations. The ascertained ARG defines a general framework for chromosome evolution of a given sample, as

it handles arbitrary genetic models, adapts to marker allele frequencies and copes with neutral mutations. The model allows for varying but fixed population sizes of the mutated and wildtype populations.

As a major application of the ascertained ARG, we have shown how it can be used to calculate likelihood and LOD scores, i.e. we use the ascertained ARG directly for multipoint gene-mapping. The mapping procedure does not require known haplotype phase. Under certain modeling assumptions, the LOD score can be computed exactly, without Monte Carlo approximation, and *p*-values can be calculated based on permutation testing. The computational speedups are, to a large extent, based on a HMM algorithm for subsets of mutated chromosomes, assumed to evolve according to Markov chains across markers.

8.2 Extensions of the Proposed Likelihood

Some of our regularity conditions are more restrictive than in other papers, as discussed in Section 1. We mention a few possible extensions that should be investigated in more detail in future work:

A) One deficit with our model is that the nuisance parameters ξ, including the recombination rate ρ, haplotype frequencies f and mutation rates μ must be set by the modeler. In most real life situations these parameters are not known, and must in this case be estimated before the analysis is performed. Thus, the most important extension of our method would be to include estimation of the nuisance parameters that are treated as fixed but unknown and integrate over those that are treated as random (hidden variables).

Of the parameters given in Figure 2, we now argue that $\xi_{\mathscr{A}}$ and μ should be random. The reason why we have treated them as fixed is the conditioning on the age G of the disease causing mutation as well as subpopulation sizes $\{N_{Mu}\}$ and $\{N_{Wu}\}$. This simplifies the form of $\mathscr{A}|\mathscr{M}_0$ a lot because of its Markovian structure backwards in time. An alternative approach would be to condition only on total population sizes $\{N_u\}$, treating G, $\{N_{Mu}\}$ and $\{N_{Wu}\}$ as random. Wiuf and Donnelly (1999) derived the distribution of the coalescence tree $\mathscr{T}(\tau)|\mathscr{M}_0$ in a haploid framework. It would be interesting to extend their results to diploid populations and ARGs. After switching to continuous time, this essentially corresponds to using a likelihood

$$P_x(g|Y) = \int P_x(g|Y,\xi_{\mathscr{A}},\mu)dP(\xi_{\mathscr{A}},\mu), \tag{24}$$

since $\xi_{\mathscr{A}}$ and μ are the only nuisance parameters that are affected by the mutation age and population sizes. With assumptions (x)-(xii), (24) simplifies further to

$$P_x(g|Y) = \int P_x(g|Y,\rho,\mu)dP(\rho,\mu). \tag{25}$$

DOI: 10.2202/1544-6115.1380

Now $L(x) = L(x;\rho,\mu)$, whereas $L(\infty) = L(\infty;\mu)$. If we treat μ as a constant and only vary ρ in (25), we find that the null likelihood is constant, giving us a lod score which generalizes (2) as

$$Z(x) = \log_{10} \int \mathrm{LR}(x;\rho) dP(\rho), \tag{26}$$

where $\mathrm{LR}(x;\rho)$ computed as in Section 6. For instance, using a finite grid of ρ-values we only have to sum over a few terms in (26).

B) We also assumed (iii), that none of the marker loci are causal for the disease. Although computations will not collapse if LOD score is calculated at marker loci, there will in general be a discrepancy between disease allele frequency and marker allele frequency that will hamper the results. More precisely, even for calculations at a marker locus, the procedure does not require that exactly the chromosomes with disease mutation should have a certain allele. We are currently investigating extensions of our mapping procedure when $\tau = x_k$ is assumed for some k.

C) Multiple disease mutations at τ could be treated by generalizing (i) and considering M disjoint subpopulations of mutated chromosomes, descended from M founder mutations $b \to B_s$ (with possibly different penetrance functions) G_s generations back in time, $s = 1,\ldots,M$. The computational complexity in Section 6.1 increases from $mK2^K$ to $mK2^{MK}$, since we have to condition on M distinct founder haplotypes. The state space of the hidden Markov chain in Appendix F is enlarged slightly, since we need to keep track of the various subpopulatioins.

D) We have required background LE in (xiii) for two reasons; it allows fast computation of the likelihood for unphased data (Appendix E) and it gives a product structure when computing S in (A.10), which is needed for the HMM algorithm of Appendix F. An extension would be to model background LD in the current generation as a first order Markov chain (see e.g. Liu et al, 2001 and Morris et al., 2002). Assuming phased data and a likelihood $P_x(h|Y)$ this gives a likelihood ratio

$$L(h',D) = \prod_{i=1}^{n} \frac{\tilde{f}(h_{i,k_i^+ +1}|h_{i,k_i^-}) \prod_{k=k_i^-}^{k_i^+} P(h_{ik}|h'_k)}{\prod_{k=k_i^-}^{k_i^+ +1} \tilde{f}(h_{ik}|h_{i,k-1})},$$

where k_i^- and k_i^+ are the left and right hand end points of the i^{th} row of D. In Section 6.1, the resulting likelihood ratio $\mathrm{LR}(x)$ can be computed by means of an HMM algorithm with state space pairs of mutated subsets at neighbouring loci. The computational complexity is still proportional to $mK2^K$, but with a larger proportionality constant.

E) The perhaps most straightforward way of relaxing the star topology assumption of the ARG at the disease locus is to retain (x)-(xi) but relax (xii) and allow for $\mu_M(t) > 0$. This happens, for instance, in Example 1 when $a \ll b + c$ and $c \gg 1$. The resulting coalescence tree $T(\tau)$ is then similar to the unshattered version of the case chromosome topology in Morris et al. (2002). This implies that the nearest recombination events are no longer independent, see McPeek and Strahs (1999). As a result, the expansion (20) of the likelihood ratio is no longer valid. Hence some approximation of the lod score, such as the plod score, is necessary. Another complication is that the Markov property (A.12) of mutated subsets is lost, and the HMM algorithm of Appendix F cannot be applied without some model approximations. A possibility in this case would be to employ the Sequential Markov Coalescent (SMC) of McVean and Cardin (2005), which is an approximation of the ARG that is Markovian not only in time but also along sequences.

F) The rare disease allele condition (x) is imposed in order to assure that D, the set of markers inherited from the ancestral mutated haplotype, has connected rows, cf. (11). It is *not* needed for assuring that the number n_{M0} of mutated chromosomes is small. Relaxation of (x) would require a generalization of the Markov process along sequences in Appendix F to handle transitions back and forth between $\mathscr{M}_0$ and $\mathscr{W}_0$, perhaps using the SMC mentioned under E).

8.3 Importance Sampling

In Section 8.2, we discussed relaxing one regularity condition at a time. Without any of the the model approximations of Section 5, exact likelihood computation (5) is not feasible, because of the daunting summation over hidden data $H = (h, h', \mathscr{A}, \mathscr{M}_0)$. With importance sampling we use a *proposal distribution* $P_{\rm pr}$ for H and write

$$L(x) = \sum_H \frac{P_x(g, H|Y)}{P_{\rm pr}(H)} P_{\rm pr}(H),$$

which is approximated, without bias, by

$$\hat{L}(x) = \frac{1}{Q} \sum_{i=1}^{Q} \frac{P_x(g, H_i|Y)}{P_{\rm pr}(H_i)},$$

where $\{H_i\}$ is a random sample from $P_{\rm pr}$. The optimal choice $P_{\rm pr}(H) = P_x(H|g, Y)$ is not feasible to sample from, but it is crucial for $P_{\rm pr}$ to be close to this optimum. Larribe et al. (2002) employ, within a full likehihood framework $P_x(Y, g)$, a proposal distribution due to Griffiths and Tavaré (1994). More accurate choices of $P_{\rm pr}$

DOI: 10.2202/1544-6115.1380

have been suggested by Stephens and Donnelly (2000) and Fearnhead and Donnelly (2001,2002) in other contexts than disease locus testing/estimation. An interesting topic would be to extend their algorithms to our framework.

8.4 Bayesian Approach

Letting x (and possibly the whole or part of ξ) be random we can use the likelihood $L(x)$ for computing the posterior density

$$P(x|g,Y) = \sum_H P(x,H|g,Y)dH \propto L(x) \propto \mathrm{LR}(x), \tag{27}$$

when the prior of x is uniform, see e.g. Rannala and Reeve (2001), Morris et al. (2002) and Zöllner and Pritchard (2005). The Bayesian mapping procedure based on (27) is faster than (3), since no permutation procedure is needed for p-value computation. The Bayesian approach is also more suitable for disease locus estimation, using crediblity regions defined from the posterior. On the other hand, the pointwise p-value (3) is more robust towards model misspecification. We notice that, for each fixed x, the marginal distribution $P_x(H|g,Y)$ is identical to the optimal proposal distribution in Section 8.3.

8.5 IBD-Based Haplotype Similarity Measures

As mentioned in the introduction, a pairwise IBS-based similarity between haplotypes can be used for haplotype clustering and gene mapping. One application of ARGs is to define alternative IBD-based haplotype similarity measures . We have recently, in Hartman et al. (2008), used an alternative haplotype similarity metric based on either of the two IBD-sharing probabilities

$$\begin{aligned} P(i_1 \text{ and } i_2 \text{ IBD at } \tau|h_{i_1},h_{i_2}) &= P(i_1,i_2 \in \mathscr{M}_0|h_{i_1},h_{i_2}), \\ P(i_1 \text{ and } i_2 \text{ IBD at } \tau|h) &= P(i_1,i_2 \in \mathscr{M}_0|h) \end{aligned} \tag{28}$$

between Chromosomes $i_1 \neq i_2$. In both equalities of (28) we assume that no wild-type chromosomes coalesce, i.e. that $\mu_U(\cdot) = 0$. Then, the only way in which the two chromosomes could be IBD is that they both descend from the mutated founder chromosome.

8.6 LD Simulation

For the purpose of simulation from an ARG (with slight approximations) there are various software available, see e.g. Hudson (2002) and Marjoram and Wall (2006). The simulations include recombinations and neutral mutations, and can be modified to adapt to varying recombination or mutation rates. Thus, there are good methods

to generate random samples from a population. However, these algorithms provide no good way to obtain the kind of highly non-random samples that are used for case-control association studies, linkage-disequilibrium mapping studies and other gene-mapping purposes. We therefore believe one important application of our ascertained ARG is simulation of haplotype data h conditional on Y. This can be achieved by first simulating $\mathscr{M}_0$ conditional on Y, then $\mathscr{A}$ conditional on $\mathscr{M}_0$, h' conditional on $n_G = n_G(\mathscr{A})$ with specified founder haplotype frequencies, and finally h conditional on $\mathscr{A}$ and h'. In this way we mimic a non-random sample where chromosomes carrying a disease mutation tend to be more closely related than chromosomes not carrying the disease, and thus also more closely related than the population as a whole. Ascertainment corrected simulation has earlier been proposed by Zöllner and von Haessler (2000) and recently by Wang and Rannala (2004,2005), whose models show large similarities with the one we propose. Just as in this paper, the model of Wang and Rannala (2005) handles incomplete penetrance and genotype data. However, they use discrete generations, and account for varying or stochastic disease allele frequencies. This could also be incorporated into our ARG, by first generating the random disease allele frequencies $\{p(t); 0 \leq t \leq t\}$ and then generating $\mathscr{A}$ conditional on $p(\cdot)$, similarly as in (24). That the disease locus of Zöllner and von Haessler (2000) and Wang and Rannala (2004, 2005) is assumed to be on the same side of all markers, is probably easy to generalize to our approach, with markers on both sides of the disease locus. Further the SNP locations are simulated as part of the procedure, whereas ours appear at predetermined positions. This simpler approach allows researchers interested in a specified region to choose the marker locations among the SNP locations in the human genome, which are nowadays easily available, e.g. from the HapMap project (The International HapMap Consortium, 2003).

Appendix A: Markov Property of $\mathscr{A}|\mathscr{M}_0$ in Discrete Time

To show that we have a Markov property of $\mathscr{A}|\mathscr{M}_0$ backward in time we by introducing one more condition of random mating forwards in time:

(ix) For $u = G, G-1, \ldots, 1$, in formation of Generation $u-1$, there are $N_{u-1}/2$ random matings taking place between the $N_u/4$ males and $N_u/4$ females of Generation u, conditioned on the following: The number of produced males and females are equal, $N_{u-1}/4$, and the genotype frequencies at the disease locus within both the female and male subpopulations of Generation $u-1$ are q_{u-1}^2 for genotype (bb), $2p_{u-1}q_{u-1}$ for genotype (Bb) and p_{u-1}^2 for genotype (BB), where $q_u = 1 - p_u$ and $p_{u-1} = N_{M,u-1}/N_{u-1}$.

DOI: 10.2202/1544-6115.1380

Condition (ix) gives a diploid Wright-Fisher model for a population with two subpopulations (mutated and wildtype) with known varying population sizes. A consequence of (iii) is that half of the B-alleles of Generation u are in females and half in males for all $0 \leq u < G$.

It is well known that the ordinary diploid Wright-Fisher models gives rise to a Markov chain *backwards* in time for mutations and recombinations, see e.g. Nordborg (2001) and Nordborg and Tavaré (2002). We now modify these calculations to our framework of two diploid subpopulations.

Suppose there are $k \geq 2$ mutated and $l \geq 2$ wildtype chromosomes ancestral to the given sample in Generation $u-1$, where $2 \leq u \leq G-1$[2]. Let R be the event that any of the $k+l$ lines recombines, C_M the event that at least two of the mutated chromosomes coalesce and C_W the event that at least two of the wildtype chromosomes coalesce when going back one generation in time. Since recombinations are independent of the random mating for each generation, and Hardy-Weinberg equilibrium is kept within males and females of each generation, it follows that the probability of no coalescence or recombination event is

$$P(R^* \cap C_M^* \cap C_W^*) = P(R^*)P(C_M^* \cap C_W^*) = P(R^*)P(C_M^*)P(C_W^*)(1+O(N_u^{-1})),$$

where R^* is the complement to R, etc. When $k = 2$, we notice that

$$P(C_M) = \frac{N_{M,u-1}/2-1}{(N_{M,u-1}-1)} \times P(A|M) = 0.5P(A|M) + o(N_{u-1}^{-1}), \tag{A.1}$$

where $(N_{M,u-1}/2-1)/(N_{M,u-1}-1)$ is the probability that the two distinct mutated alleles have parents of the same sex, A is the event that two different alleles of Generation $u-1$ that originate from a parent of the same sex in Generation u also originate from the same parent and the same grandparental allele within that parent and M is the event that two randomly chosen alleles of Generation $u-1$ are both mutated. We further introduce the events H_i that the parent of Allele 1 has i B-alleles, $i = 0,1,2$. Then

$$P(A|M) = P(A|H_1,M)P(H_1|M) + P(A|H_2,M)P(H_2|M).$$

Without loss of generality, assume that the two alleles have male parents. Since there is Hardy Weinberg equilibrium within the male population of Generation u, it follows that $P(H_1) = 2p_uq_u$ and $P(H_2) = p_u^2$. Further $P(M|H_1) = 0.5$ and

[2]The computation of coalescence probabilities is slightly different when $u = 1$, due to the constaint that the n sampled chromosomes are from m pairs of individuals. Likewise, for $u = G$ the Hardy-Weinberg condition of the parental generation is not fulfilled. However, omitting a finite number of generations has no asymptotic effect when $G \to \infty$ below.

$P(M|H_2) = 1$, since both alleles of a parent are (unconditionally on mutation status) equally likely to be transmitted to the next generation. Hence, by Bayes' Rule, $P(H_1|M) = q_u$ and $P(H_2|M) = p_u$. Given that an allele of Generation $u-1$ is mutated, the probability that a given male of Generation u (out of all $N_u/4$ possible) is its father is $c/(N_u/4) = 2P(A|H_2,M)$ if the father has two B-alleles and $0.5c/(N_u/4) = P(A|H_1,M)$ if the father has one B-allele. To determine the proportionality constant c, we notice that the number of males with two B-alleles is $p_u^2 \cdot N_u/4$ and the number of males with one B-allele is $2p_uq_u \cdot N_u/4$. Hence

$$p_u^2 N_u/4 \cdot \frac{c}{N_u/4} + 2p_uq_uN_u/4 \cdot \frac{0.5c}{N_u/4} = 1 \Longleftrightarrow c = p_u^{-1}.$$

Putting things together, we find that

$$P(A|M) = \frac{2}{N_u p_u} q_u + \frac{2}{N_u p_u} p_u = \frac{2}{N_{Mu}}.$$

Inserting this into (A.1) we obtain a coalescence probability $P(C_M) = 1/N_{Mu} + o(1/N_u)$. For $k \geq 2$, let C_{ij} denote the event that chromosomes i and j coalesce. Then

$$\begin{aligned} P(C_M) &= P\left(\cup_{1\leq i<j\leq k} C_{ij}\right) = \textstyle\sum_{1\leq i<j\leq k} P(C_{ij}) + o(N_u^{-1}) \\ &= \binom{k}{2}/N_{Mu} + o(N_u^{-1}), \end{aligned} \tag{A.2}$$

since the probability that more than two lines coalesce at the same time has probability $o(N_u^{-1})$. In the same way, one shows

$$P(C_W) = \binom{l}{2}/N_{Wu} + o(N_u^{-1}).$$

Since the $l+k$ chromosomes recombine independently,

$$P(R) = 1 - (1-r)^{k+l}.$$

Given that a recombination occurs at X during formation of Chromosome e of Generation $u-1$, let e_1 and e_2 denote the two parental chromosomes of Generation u that transmitted material along $[0,X)$ and $[X,1]$ to e respectively. Then, the parent transmitting τ must have the same mutation status as e, whereas the other parent is selected randomly from one (male or female) half of the population with probability p_u of being mutated. Hence

DOI: 10.2202/1544-6115.1380

$$
\begin{array}{rcl}
p(e_1 \in \mathscr{M} | e \in \mathscr{M}) & = & 1_{\{X>\tau\}} + p_u 1_{\{X\le\tau\}}, \\
p(e_2 \in \mathscr{M} | e \in \mathscr{M}) & = & p_u 1_{\{X>\tau\}} + 1_{\{X\le\tau\}}, \\
p(e_1 \in \mathscr{M} | e \in \mathscr{U}) & = & p_u 1_{\{X\le\tau\}}, \\
p(e_2 \in \mathscr{M} | e \in \mathscr{U}) & = & p_u 1_{\{X>\tau\}}.
\end{array} \tag{A.3}
$$

□

Appendix B: Markov Process and Rates of $\mathscr{A}|\mathscr{M}_0$ in Continuous Time

Suppose there are k mutated and l wildtype lines at time $0 \le t_0 < 1$ (Generation $[t_0 G]$). Then, from the discrete time analysis of $\mathscr{A}|\mathscr{M}_0$, it follows that the probability that no coalescence or recombination event occurs up to time t_1 ($t_0 < t_1 < 1$) is

$$
\begin{array}{rl}
& \Pi_{u=[t_0G]+1}^{[t_1G]} \left(\left(1 - \binom{k}{2}/N_{Mu} \right) \left(1 - \binom{l}{2}/N_{Wu} \right) (1-r)^{k+l} \right) + o(1) \\
\longrightarrow & \exp\left(- \int_{t_0}^{t_1} \left(\binom{k}{2} \lambda_M(s) + \binom{l}{2} \lambda_W(s) + (k+l)\rho \right) ds \right)
\end{array}
$$

as $G \to \infty$, using (10) in the last step. Since t_0 and t_1 are arbitrary, recombination and coalescence events occur according to a Markov process in continuous time with rates as specified in Section 4. Moreover, if a recombination occurs at time t, the probability of the left and right hand parental lines being mutated or wildtype is given by (11), which corrresponds to (A.3) with $p(t)$ instead of p_u. □

Appendix C: Definition of D in Terms of Nearest Recombination Events

According to (11), Condition (x) implies that the set D of mutated sites (i,k) is a connected region, whose boundary to the left and right is defined using nearest recombination events as follows: Assume that H_1^x holds, and let $x_{k_0} \le x < x_{k_0+1}$ for some $k_0 = 0, 1, 2, \ldots, K$. (We put $x_0 = -\infty$ and $x_{K+1} = \infty$ to make k_0 well defined even when $x < x_1$ or $x \ge x_K$.) For each $i \in \mathscr{M}_0$, consider the i-lineage of $\mathscr{T}(x)$ from $t = 0$ back to $t = 1$. It is a union of several edges along the ARG. Each junction between two such edges corresponds to a recombination event. Let X_i^- and X_i^+ be the recombination points to the left and right of x that are nearest to x. (If there are no recombination points to the left of x we put $X_i^- = -\infty$ and similarly $X_i^+ = \infty$ if there are no recombination points to the right of x.) Then

$$
\begin{array}{rcl}
D & = & \{(i,k),\ i \in \mathscr{M}_0 \text{ and either } x_k \ge X_i^- \text{ for } k \le k_0 \\
& & \text{or } x_k < X_i^+ \text{ for } k \ge k_0 + 1\}.
\end{array} \tag{A.4}
$$

Appendix D: Derivation of (14)

The first step in simplifying (5) to (14) is to simplify the expression for $P(h|\mathscr{A},h')$ in (9). Conditions (x)-(xii) imply that $\mathscr{T}(x_k)$ is a union of disjoint straight lines for all lineages i such that $(i,k)\notin D$ and a remaining tree with star topology, connecting all edges i with $(i,k)\in D$ so that they all coalesce at time $t=1$, see Figure 4. Hence

$$P(h|\mathscr{A},h') = \prod_{(i,k)\in D} P(h_{ik}|h'_{Jk}) \cdot \prod_{(i,k)\notin D} P(h_{ik}|h'_{j_{ik}k}), \tag{A.5}$$

where $j_{ik}\neq J$ is the founder chromosome ancestral to (i,x_k). Using (xiv), we obtain

$$P(h_{ik}|h'_{jk}) = (1-q_k)^{\{h_{ik}=h'_{jk}\}} q_k^{\{h_{ik}\neq h'_{jk}\}}, \tag{A.6}$$

where q_k is the mutation probability at x_k.

We then sum over $h'_{(-J)}=\{h'_j;\, j\neq J\}$, use (A.5), the LE condition (xiii) and (8) to deduce

$$\begin{aligned} P(h|D,h') &= \textstyle\sum_{h'_{(-J)}} P(h|\mathscr{A},h')P(h'_{(-J)}|\mathscr{A}) \\ &= \textstyle\prod_{(i,k)\in D} P(h_{ik}|h'_{Jk}) \cdot \prod_{(i,k)\notin D} \tilde{f}_k(h_{ik}) \end{aligned} \tag{A.7}$$

where $\tilde{f}_k$ are the allele frequencies defined in (18).

We obtain (14) by summing over various variables in (5) in the following order: First we sum over $h'_{(-J)}$, as in (A.7), then over all $\mathscr{A}$ such that $D(\mathscr{A})=D$ and $\mathscr{M}_0$, noticing that

$$P_x(D|Y) = \sum_{\substack{\mathscr{A},\mathscr{M}_0 \\ D(\mathscr{A})=D}} P_x(\mathscr{A}|\mathscr{M}_0)P(\mathscr{M}_0|Y) = \sum_{\mathscr{M}_0} P_x(D|\mathscr{M}_0)P(\mathscr{M}_0|Y),$$

and, finally, we sum over h, h' and D.

Appendix E: Derivation of (17)

Let $\mathscr{H}=\{h;\, h\sim g\}$ consist of all haplotypes consistent with genotype data. We obtain $\mathscr{H}$ by switching all heterozygous sites (i,k) with the homologous one in $H_{[(i+1)/2]k}$. This is done independently for all pairs of heterozygous sites, so that $|\mathscr{H}|=2^{|H|/2}$. Hence, (16) implies

$$LR(h',D) = \sum_{h\in\mathscr{H}} \frac{P(h|h',D)}{L(\infty)}. \tag{A.8}$$

Since $P_\infty(D=\emptyset|Y)=1$, we obtain $L(\infty)$ by putting $D=\emptyset$ in (A.7), which yields

$$L(\infty)=\sum_{h\in\mathscr{H}}\prod_{(i,k)}\tilde{f}_k(h_{ik})=|\mathscr{H}|\prod_{(i,k)}\tilde{f}_k(h_{ik}). \tag{A.9}$$

Combining (A.7), (A.8) and (A.9) we get

$$LR(h',D)=\frac{1}{|\mathscr{H}|}\sum_{h\in\mathscr{H}}\prod_{(i,k)\in D}\frac{P(h_{ik}|h'_{Jk})}{\tilde{f}_k(h_{ik})}.$$

Carrying out the last summation independently for all nonempty pairs H_{vk} of heterozygous sites, we end up with (17).

Appendix F: HMM algorithm for Computing Likelihood Ratio (20) and (21)

Consider a set $V\subset\{1,\dots,m\}$ of individuals, let $I=\cup_{v\in V}\{2v-1,2v\}$ denote the corresponding set of haplotypes, $D_V=D\cap(I\times\{1,\dots,K\})$ the set of mutated markers (i,k) for chromosomes among individuals in V and $C_k=C_{x_k}$ the k^{th} column of D_V, $k=1,\dots,K$. Under the assumption that $\tau=x$, we will devise a HMM algorithm for computing

$$S=E_x\left(\prod_{k=1}^{K}U_k(C_k)|Y_V\right), \tag{A.10}$$

where $Y_V=\{Y_v,\,v\in V\}$ and $U_k(C_k)=U_k(C_k;g)$ a given function. We will apply this when

I) $V=\{v\}$, $S=\mathrm{LR}(h',D_{\{v\}})$ equals the v^{th} term of the outer product in (20) and U_k the k^{th} term on the right-hand side of (17), when $D_{\{v\}}$ replaces D on the left-hand side.

II) $V=\{1,\dots,m\}$, $Y_V=Y$, $D_V=D$, $S=\mathrm{LR}(x)$ is the likelihood ratio, expanded as in (21) and U_k the k^{th} term on the right-hand side of (22).

III) given V, $S=\mathrm{LR}(x;V)$ and U_k is the k^{th} term on the right-hand side of (21), when D_V replaces D on the left-hand side. This case is a generalization of II.

To begin with, we establish a Markov property of $\{C_k\}$. Because of (A.4), the columns of D_V to the left and right of the (assumed) disease locus, $\{C_k\}_{k=k_0}^{1}$ and $\{C_k\}_{k=k_0+1}^{K}$, evolve as two Markov chains with state space all subsets of I. The two chains are independent conditional on their starting values, which depend on

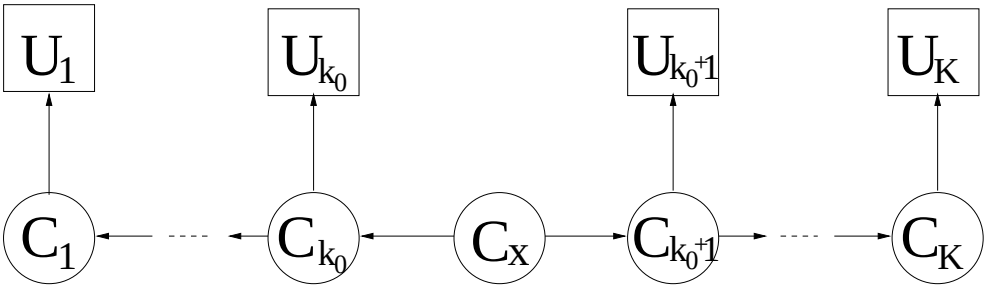

Figure 11: DAG showing the Markov chain $\{C_k\}$, incompletely observed by means of $U_k = U_k(C_k)$.

$C_x := \mathscr{M}_0 \cap I$. As the chains progress, C_k is non-increasing in both directions, with Lineage $i \in I$ lost at x_k when there are recombination events at X_i^- (X_i^+) just to the right (left) of x_k affecting i.

Since nearest recombination events occur as independent Poisson processes with rate $\rho\pi(\cdot)$ along different mutated i-lineages, it is easy to see that $\{X_i^-\}_{i\in\mathscr{M}_0}$ and $\{X_i^+\}_{i\in\mathscr{M}_0}$ are independent random variables with

$$\begin{aligned} P_x(X_i^- < x' | i \in \mathscr{M}_0) &= \exp\left(-\rho \textstyle\int_{x'}^{x} \pi(y)dy\right),\ 0 \le x' < x. \\ P_x(X_i^+ > x' | i \in \mathscr{M}_0) &= \exp\left(-\rho \textstyle\int_{x}^{x'} \pi(y)dy\right),\ x < x' \le 1. \end{aligned} \tag{A.11}$$

Let F^- and F^+ denote the distribution functions of X_i^- and X_i^+, and put

$$\begin{aligned} r_k &= (F^-(x_k) - F^-(x_{k-1}))/F^-(x_k), && 1 \le k \le k_0, \\ r_k &= (F^+(x_{k+1}) - F^+(x_k))/(1 - F^+(x_k)), && k_0 + 1 \le k \le K, \\ r_x^- &= F^-(x) - F^-(x_{k_0}), \\ r_x^+ &= F^+(x_{k_0+1}) - F^+(x). \end{aligned}$$

In words, when $k \le k_0$, r_k is the probability of a recombination between x_{k-1} and x_k given that no recombination has occurred between x_k and x. r_x^- is the probability of a recombination event between x_{k_0} and x for a mutated lineage $i \in \mathscr{M}_0$. The interpretation of r_k for $k \ge k_0 + 1$ and r_x^+ is similar. This gives rise to transition probabilities

$$\begin{aligned} P(C_{k-1} = C' | C_k = C) &= r_k^{|C|-|C'|}(1-r_k)^{|C'|},\ k = 1, \ldots, k_0, \\ P(C_{k+1} = C' | C_k = C) &= r_k^{|C|-|C'|}(1-r_k)^{|C'|},\ k = k_0+1, \ldots, K, \\ P_x(C_{k_0} = C' | C_x = C) &= (r_x^-)^{|C|-|C'|}(1-r_x^-)^{|C'|}, \\ P_x(C_{k_0+1} = C' | C_x = C) &= (r_x^+)^{|C|-|C'|}(1-r_x^+)^{|C'|}, \end{aligned} \tag{A.12}$$

DOI: 10.2202/1544-6115.1380

provided $C' \subseteq C$ (otherwise the transition probabilities are zero). Define

$$S_k(C) = \begin{cases} E_x(\prod_{l=1}^{k} U_l(C_l)|C_k = C), & 0 \le k \le k_0, \\ E_x(\prod_{l=k}^{K} U_l(C_l)|C_k = C), & k_0 + 1 \le k \le K+1. \end{cases}$$

and

$$S(C) = E_x(\prod_{l=1}^{K} U_l(C_l)|C_x = C).$$

Then, the recursive algorithm for computing S can be formulated in terms of $S_k(C)$, $S(C)$ and the transition probabilities as follows: Start with initial conditions

$$\begin{array}{rcll} S_0(C) & = & 1, & \\ S_{K+1}(C) & = & 1, & \end{array} \quad \forall C \subseteq I.$$

Then, define recursively for all $C \subseteq I$

$$\begin{array}{rcl} S_k(C) & = & \sum_{C' \subseteq C} S_{k-1}(C') U_k(C) P(C_{k-1} = C'|C_k = C), k = 1, \ldots, k_0, \\ S_k(C) & = & \sum_{C' \subseteq C} S_{k+1}(C') U_k(C) P(C_{k+1} = C'|C_k = C), k = K, \ldots, k_0 + 1. \end{array}$$

The final two steps are

$$\begin{array}{rcl} S(C) & = & \left(\sum_{C' \subseteq C} S_{k_0}(C') P(C_{k_0} = C'|C_x = C)\right) \\ & & \cdot \left(\sum_{C' \subseteq C} S_{k_0+1}(C') P(C_{k_0+1} = C'|C_x = C)\right), \ \forall C \subseteq I \\ S & = & \sum_{C \subseteq I} S(C) P(C_x = C|Y_V). \end{array}$$

In the last step we use (7) to evaluate $P(C_x = C|Y_V)$. The total complexity of the algorithm is $O(K2^{|I|})$, where $2^{|I|}$ is the state space size.

The HMM algorithm of McPeek and Strahs (1999) is close to a special case of ours, when $|I| = 1$. However, they introduce mutated alleles as a separate state (giving a total state space size $2^1 + 1 = 3$), whereas we incorporate mutation within the functions U_k. □

References

Bradley, A. (1996). ROC curves and the χ^2 test. *Pattern Recognition Letters*, 17(3):287–294.

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

Easton, D. F., Pooley, K. A., Dunning, A. M. et al. (2007). Genome-wide association study identifies novel breast cancer susceptibility loci. *Nature*, 447:1087–1094.

Einarsdóttir, K., Humphreys K., Bonnard C. et al. (2006) Linkage disequilibrium mapping of CHEK2: common variation and breast cancer risk. *PLoS Med*, 3:e168.

Fearnhead, P. and Donnelly, P. (2001). Estimating recombination rates from population genetic data. *Genetics*, 159(3):1299–1318.

Fearnhead, P. and Donnelly, P. (2002). Approximate likelihood methods for estimating local recombination rates. *J R Statist Soc Ser B*, 64:657–680.

Gasbarra, D., Sillanpää, M.J. and Eljas, A. (2005). Backward simulation of ancestors of sampled individuals, *Theor Popul Biol*, 67:75-83.

Gasbarra, D., Pirinen, M., Sillanpää, M.J. and Eljas, A. (2007). Estimating genealogies from linked marker data: a Bayesian approach, *BMC Bioinformatics* 8:411.

Gasbarra, D., Pirinen, M., Sillanpää, M.J. and Eljas, A. (2009). Bayesian quantitative trait locus mapping based on reconstruction of recent genetic histories. Manuscript.

Griffiths, R. and Marjoram, P. (1997). An ancestral recombination graph. In Donnelly, P. and Tavaré, S., editors, *Progress in Population Genetics and Human Evolution*, pages 257–270. Springer, New York.

Griffiths, R.C. and Tavaré, S. (1994). Ancestral inference in population genetics, *Statist. Sci.* **9**, 307-319.

Hartman, L., Humphreys, K. and Hössjer, O. (2008). Utilizing identity-by-descent probabilities for genetic fine-mapping in population based samples, via spatial smoothing of haplotype effects. *Computational Statistics and Data Analysis* **53**(5), 1802-1817.

Hudson, R. R. (1983). Properties of a neutral allele model with intragenic recombination. *Theor Popul Biol*, 23(2):183–201.

Hudson, R. R. (2002). Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics*, 18(2):337–338.

Hunter, D. J., Kraft, P., Jocobs, K. B. et al. (2007) A genome-wide association study identifies alleles in FGFR2 associated with risk of sporadic postmenopausal breast cancer. *Nature Genetics*, 39:870–874.

Kraft, P. and Thomas, D. (2000). Bias and efficiency in family-based gene-characterization studies: conditional, prospective, retrospective, and joint likelihoods. *Am J Hum Genet*, 66(3):1119–31.

Larribe, F. and Lessard, S. (2008). A composite-conditional likelihood approach for gene mapping based on linkeage disequilibrium in windows of marker loci. *Stat Appl Genet Mol Biol*, 7(1), Article 27.

Larribe, F., Lessard, S., and Schork, N. J. (2002). Gene mapping via the ancestral recombination graph. *Theor Popul Biol*, 62(2):215–229.

DOI: 10.2202/1544-6115.1380

Liu, J.S., Sabatti, C., Teng, J., Keats, B.J.B. and Risch, N. (2001). Bayesian analysis of haplotypes for linkage equilibrium mapping. *Genome Res* 11:1716-1724.

Mailund, T., Besenbacher, S. and Schierup, M.H. (2006). Whole genome association mapping by incompatibilities and local perfect phylogenies. *BMC Bioinformatics*, **7**, 454.

Marjoram, P. and Wall, J. D. (2006). Fast "coalescent" simulation. *BMC Genet*, 7:16.

Minichiello, M.J. and Durbin, R. (2006). Mapping trait loci using inferred ancestral recombination graphs. *Am J Hum Gen*, 79:910-922.

McPeek, M. and Strahs, A. (1999). Assessment of linkage disequilibrium by the decay of haplotype sharing, with application to fine-scale genetic mapping. *Am J Hum Genet*, 65(3):858–75.

McVean, G.A.T. and Cardin, N.J. (2005). Approximating the coalescent with recombination. *Phil. Trans. R. Soc. B*, 360, 1387-1393.

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

Morris, A.P., Whittaker, J.C. and Balding, D.J. (2000). Bayesian mapping of disease loci using hidden Markov models. *Am J Hum Genet*, 67:155-169.

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

Morris, A.P., Whittaker, J.C. and Balding, D.J. (2004). Little loss of information due to uknown phase for fine-scale disequilibrium mapping with single-nucleotide-polymorphism genotype data. *Am J Hum Genet*, 74:945-953.

Nordborg, M. (2001). Coalsescent theory. In *Handbook of Statistical Genetics*, eds. Balding, D.J., Bishop, M. and Cannings, C., 179-212, Wiley.

Nordborg, M. and Tavaré, S. (2002). Linkage disequilibrium: what history has to tell us. *Trends in Genetics* **18**, 83-90.

Rannala, B. and Reeve, J.P. (2001). High-resolution multipoint linkage disequilibrium mapping in the context of a human genome sequence. *Am J Hum Genet*, 69:159-178.

Service, S. K., Lang, D. W., Freimer, N. B., and Sandkuijl, L. A. (1999). Linkage-disequilibrium mapping of disease genes by reconstruction of ancestral haplotypes in founder populations. *Am J Hum Genet*, 64(6):1728–1738.

Stephens, M. and Donnely, P. (2000). Inference in molecular population genetics. *J R Stat Soc Ser B*, 62:605-655.

Stephens, M., Smith, N.J. and Donnelly, P. (2001). A new statistical method for haplotype reconstruction from population data. *Am J Hum Genet*, 68:978-989.

Terwilliger, J. D. (1995). A powerful likelihood method for the analysis of linkage disequilibrium between trait loci and one or more polymorphic marker loci. *Am J Hum Genet*, 56(3):777–787.

The International HapMap Consortium. (2003). The International HapMap Project. *Nature*, 426(6968):789–796.

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

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

Wang, Y. and Rannala, B. (2004). Simulating a coalescent process with recombination and ascertainment. In Istrail, S., Waterman, M., and Clark, A., editors, *Computational Methods for SNPs and Haplotype Inference*, volume 2983 of *Lecture Notes in Bioinformatics*, pages 84–95. Springer.

Wang, Y. and Rannala, B. (2005). In silico analysis of disease-association mapping strategies using the coalescent process and incorporating ascertainment and selection. *Am J Hum Genet*, 76(6):1066–1073.

Wiuf, C. and Donnelly, P. (1999). Conditional genealogies and the age of a neutral mutant. *Theor Popul Biol*, 56:183-201.

Wiuf, C. and Hein, J. (1999). The ancestry of a sample of sequences subject to recombination. *Genetics*, 151(3):1217–1228.

Wu, Y.D. (2007). Association mapping of complex diseases with ancestral recombination graphs. In *Research in Computational Molecular Biology*, eds. Speed, T. and Huang, H. Lecture Notes in Computer Science, vol. 4453, Speringer, Berlin, pp. 488-502.

Zöllner, S. and von Haeseler, A. (2000). A coalescent approach to study linkage disequilibrium between single-nucleotide polymorphisms. *Am J Hum Genet*, 66(2):615–628.

DOI: 10.2202/1544-6115.1380