Multipoint Approximations of Identity-by-Descent Probabilities for Accurate Linkage Analysis of Distantly Related Individuals

Cornelis A. Albers,[1,*] Jim Stankovich,[2,3] Russell Thomson,[3] Melanie Bahlo,[2] and Hilbert J. Kappen[1]

We propose an analytical approximation method for the estimation of multipoint identity by descent (IBD) probabilities in pedigrees containing a moderate number of distantly related individuals. We show that in large pedigrees where cases are related through untyped ancestors only, it is possible to formulate the hidden Markov model of the Lander-Green algorithm in terms of the IBD configurations of the cases. We use a first-order Markov approximation to model the changes in this IBD-configuration variable along the chromosome. In simulated and real data sets, we demonstrate that estimates of parametric and nonparametric linkage statistics based on the first-order Markov approximation are accurate. The computation time is exponential in the number of cases instead of in the number of meioses separating the cases. We have implemented our approach in the computer program ALADIN (accurate linkage analysis of distantly related individuals). ALADIN can be applied to general pedigrees and marker types and has the ability to model marker-marker linkage disequilibrium with a clustered-markers approach. Using ALADIN is straightforward: It requires no parameters to be specified and accepts standard input files.

Introduction

Even in the new era of genome-wide association studies, the more traditional approach of linkage analysis with multiplex pedigrees remains a powerful, efficient method for mapping rare disease-susceptibility alleles.[1] It is powerful not only for the mapping of Mendelian disease alleles of complete penetrance but also for the mapping of complex disease alleles of incomplete penetrance. Generally, for alleles of incomplete penetrance, it is not possible to find clusters of closely related individuals presenting with disease. However, it might be possible to identify clusters of distantly related individuals, particularly in founder populations. Because haplotype sharing between more distantly related individuals is rarer, any observed sharing is more significant. An excellent example of the power of large pedigrees with distantly related individuals was provided by a recent study of pituitary adenoma predisposition in northern Finland.[2] Significant linkage was obtained with a single nine-generation pedigree containing just six affected individuals.

For nonparametric, affecteds-only linkage analysis of large families, the key computational challenge is the determination of identity by descent (IBD) sharing probabilities. These are the probabilities, given the available genotype data, that various clusters of affected individuals have inherited haplotype IBD from common ancestors at various loci. Exact and approximate linkage algorithms to calculate IBD-sharing probabilities formulate them in terms of configurations of the inheritance vector.[3] This vector has a binary component for each meiosis, recording whether a grandmaternal or grandpaternal allele is transmitted from parent to offspring. The number of possible states of the inheritance vector increases exponentially with pedigree size, so exact methods using the Lander-Green algorithm[4] (GENEHUNTER,[5] ALLEGRO,[6] MERLIN[7]) to enumerate the probabilities of all possible states become intractable for large pedigrees, even with the latest algorithmic improvements taking advantage of some symmetries (ALLEGRO2). For larger pedigrees, approximate Markov chain Monte Carlo (MCMC) sampling is generally used.[8–13] With run lengths sufficiently long, well-mixing MCMC algorithms converge to the exact solutions. However, the time required for the obtainment of an accurate solution can be very long for some MCMC samplers,[14] and some user experience is required for the determination of when an MCMC run has converged.

Other factors contribute to the computational complexity, as well. The new dense sets of SNP markers make it possible to determine patterns of IBD sharing with greater precision, particularly for multigenerational pedigrees with many untyped individuals, but increase computational burden. Although for the Lander-Green algorithm the increase is only linear, for large numbers of markers, the amount of computer memory and disk space required may be substantial. In addition, with dense marker sets, it is critical to make allowance for linkage disequilibrium (LD) between nearby markers. Only one Lander-Green program (MERLIN[7]) can handle LD between markers. The MCMC algorithm MCLINK[12] has been extended to allow for LD[15,16] but is still experimental. The two most commonly used MCMC programs, MORGAN[11] and SIMWALK2,[8] cannot yet handle LD.

For computation of linkage statistics on large pedigrees with many generations of untyped individuals, we propose

[1]Department of Biophysics, Institute for Computing and Information Sciences, Radboud University, 6525 EZ Nijmegen, The Netherlands; [2]Bioinformatics Division, Walter and Eliza Hall Institute of Medical Research, 1G Royal Parade, 3050 Parkville, VIC, Melbourne, Australia; [3]Menzies Research Institute, University of Tasmania, Private Bag 23, Hobart, TAS 7001, Australia
*Correspondence: k.albers@science.ru.nl
DOI 10.1016/j.ajhg.2007.12.016.

that it is not necessary to model explicitly all components of the inheritance vector for the untyped individuals be explicitly modeled. Instead, it suffices to model a variable Π that records the IBD-sharing configuration of the top generation of genotyped individuals. Although the correlations in Π are not first-order Markov for any but the simplest of pedigrees,[17–19] it has been successfully approximated as first-order Markov for homozygosity mapping[20–22] and the determination of genealogical relationships.[19] Our contribution is to present an approach for estimation of parametric and nonparametric linkage statistics for general numbers of cases with arbitrary degrees of relatedness, by using first-order Markov approximations. This requires computation of the probabilities of the various possible configurations of Π and computations of transition probabilities between configurations as a function of the recombination fractions. We use a dynamic programming algorithm based on variable elimination[23,24] to compute these probabilities exactly.

We have implemented the approximation in the computer program ALADIN (accurate linkage analysis of distantly related individuals). We evaluate accuracy in simulated data sets and a real data set and compare ALADIN with the MCMC program MORGAN in terms of accuracy and computational efficiency.

Material and Methods

We divide the group of individuals $\mathbf{P}$ that together form the pedigree into three disjoint groups of individuals $\mathbf{A}$, $\mathbf{T}$, and $\mathbf{D}$ such that $\mathbf{P} = \mathbf{A} \cup \mathbf{T} \cup \mathbf{D}$. The group of individuals $\mathbf{T}$ is defined by the requirement that individuals in $\mathbf{T}$ are related only by ancestors without genotype and phenotype information. This group of untyped ancestors is denoted by $\mathbf{A}$. Every pair of individuals in $\mathbf{T}$ is related by at least one common ancestor from $\mathbf{A}$. For every individual in $\mathbf{T}$, genotype or phenotype information is available, or both. The third group, $\mathbf{D}$, contains the remaining individuals.

The individuals of any pedigree can be grouped like this. ALADIN was designed with the situation in mind where $\mathbf{T}$ consists of a small number of cases (less than ten) diagnosed with the disease and $\mathbf{D}$ consists of a limited number of close relatives (spouses or children) of the cases that have been recruited to increase the information content. $\mathbf{A}$ is the group of ancestors through whom the cases are related. There is no limitation on the number of ancestors $\mathbf{A}$ or the number of generations they span, although the efficiency of ALADIN does depend on the complexity of the subpedigree formed by these ancestors (see below for details).

Exact computation of the IBD probabilities with the Lander-Green algorithm is exponential in the number of meioses separating the cases and can be prohibitively complex when $\mathbf{A}$ is large. We propose to approximate the likelihood of the original HMM formulation for multipoint linkage analysis[4,5] with a likelihood with lower computational complexity. This is accomplished through a change of variables based on the partitioning of the pedigree $\mathbf{P}$ into the groups $\mathbf{A}$, $\mathbf{T}$, and $\mathbf{D}$ and additional approximations to the prior probabilities of multilocus IBD configurations. The main idea is that exact inference with the Lander-Green algorithm in this approximate HMM can be performed efficiently.

In this section, we describe how to estimate the posterior probabilities of the IBD configurations given marker data in the approximate HMM. The nonparametric linkage statistics can be readily calculated from these posterior probabilities. In the Appendix (Estimation of Multipoint Parametric LOD Scores), we describe how parametric logarithm of odds (LOD) scores can be estimated.

A Change of Variables

We first consider a single marker locus. The exact probabilities of the possible IBD configurations of the affecteds (the cases) can be calculated from the posterior marginal distribution of the segregation indicators $\mathbf{s}_{\mathrm{nf}}$:

$$P(\mathbf{s}_{\mathrm{nf}} \mid \mathbf{M},\mathbf{f}) = P\left(\mathbf{s}_{\mathbf{D}_{\mathrm{nf}}},\mathbf{s}_{\mathbf{T}},\mathbf{s}_{\mathbf{A}_{\mathrm{nf}}} \mid \mathbf{M},\mathbf{f}\right) = \sum_{\mathbf{G}_{\mathbf{D}},\mathbf{G}_{\mathbf{T}},\mathbf{G}_{\mathbf{A}}} P\left(\mathbf{G}_{\mathbf{D}},\mathbf{G}_{\mathbf{T}},\mathbf{G}_{\mathbf{A}},\mathbf{s}_{\mathbf{D}_{\mathrm{nf}}},\mathbf{s}_{\mathbf{T}},\mathbf{s}_{\mathbf{A}_{\mathrm{nf}}} \mid \mathbf{M},\mathbf{f}\right), \quad (1)$$

where $\mathbf{M}$ denotes all observed marker genotypes, $\mathbf{f}$ the vector of allele frequencies, and the subscript nf a nonfounder in the pedigree. (Only nonfounders have segregation indicators. By definition, individuals in $\mathbf{T}$ are always nonfounders; hence, we discard the subscript for these segregation indicators.) $\mathbf{G}_{\mathbf{D}}$, $\mathbf{G}_{\mathbf{T}}$, and $\mathbf{G}_{\mathbf{A}}$ represent the ordered genotypes of the individuals in the three groups. The component of the inheritance vector $(\mathbf{s}_{\mathbf{T}},\mathbf{s}_{\mathbf{A}_{\mathrm{nf}}})$ uniquely determines the IBD configuration of the alleles of the individuals $\mathbf{T}$. However, different inheritance vectors may correspond to the same IBD configuration, and for this reason, we now introduce the variable Π, which summarizes the IBD configuration of the alleles of the individuals $\mathbf{T}$. Any configuration $(\mathbf{s}_{\mathbf{T}},\mathbf{s}_{\mathbf{A}_{\mathrm{nf}}})$ can be mapped to a single value of Π; all configurations mapped to the same value of Π have the same IBD configuration with respect to the alleles contained in $\mathbf{G}_{\mathbf{T}}$. When Π is substituted for $\mathbf{s}_{\mathbf{T}},\mathbf{s}_{\mathbf{A}_{\mathrm{nf}}}$ in Equation 1, it is clear that the nonparametric statistics can be computed alternatively from the posterior marginal distribution

$$P\left(\mathbf{s}_{\mathbf{D}_{\mathrm{nf}}},\Pi \mid \mathbf{M},\mathbf{f}\right) = \sum_{\left(\mathbf{s}_{\mathbf{T}},\mathbf{s}_{\mathbf{A}_{\mathrm{nf}}}\right)\in\Pi} P\left(\mathbf{s}_{\mathbf{D}_{\mathrm{nf}}},\mathbf{s}_{\mathbf{T}},\mathbf{s}_{\mathbf{A}_{\mathrm{nf}}} \mid \mathbf{M},\mathbf{f}\right), \quad (2)$$

where $(\mathbf{s}_{\mathbf{T}},\mathbf{s}_{\mathbf{A}_{\mathrm{nf}}})\in\Pi$ denotes the set of configurations with the same IBD configuration Π. The prior probability of an IBD configuration Π of $\mathbf{G}_{\mathbf{T}}$ is given by

$$P(\Pi) = \sum_{\left(\mathbf{s}_{\mathbf{T}},\mathbf{s}_{\mathbf{A}_{\mathrm{nf}}}\right)\in\Pi} P\left(\mathbf{s}_{\mathbf{T}},\mathbf{s}_{\mathbf{A}_{\mathrm{nf}}}\right). \quad (3)$$

With the assumption that a priori all meioses are independent and that paternal and maternal inheritance are equally probable, the evaluation of this sum for a given Π amounts to counting the number of configurations $(\mathbf{s}_{\mathbf{T}},\mathbf{s}_{\mathbf{A}_{\mathrm{nf}}})$ with the corresponding IBD configuration.

Figure 1 shows one of the possible descent graphs[8] for a pedigree of nine individuals $\mathbf{P}$ {1,…,9}. Suppose that individuals 1, 2, and 3 are affected and have genotype information, so that $\mathbf{T}$ = {1, 2, 3}, that their ancestors $\mathbf{A}$ = {4,…,9} are untyped, and that $\mathbf{D} = \emptyset$. Only the paternal alleles of individuals 1, 2, and 3 can be IBD through one of the ancestors 6 and 7; in the descent graph shown in the figure, the paternal alleles are all inherited IBD from individual 6. The maternal alleles can never be IBD because they are inherited from different founders. There are five possible IBD configurations of the paternal alleles: $(G_1^p\,G_2^p\,G_3^p)$, $(G_1^p)(G_2^p\,G_3^p)$, $(G_1^p\,G_2^p)(G_3^p)$, $(G_1^p\,G_3^p)(G_2^p)$, and $(G_1^p)(G_2^p)(G_3^p)$. Alleles within parentheses are defined to be IBD and are said to be in the same partition; a concatenation of partitions

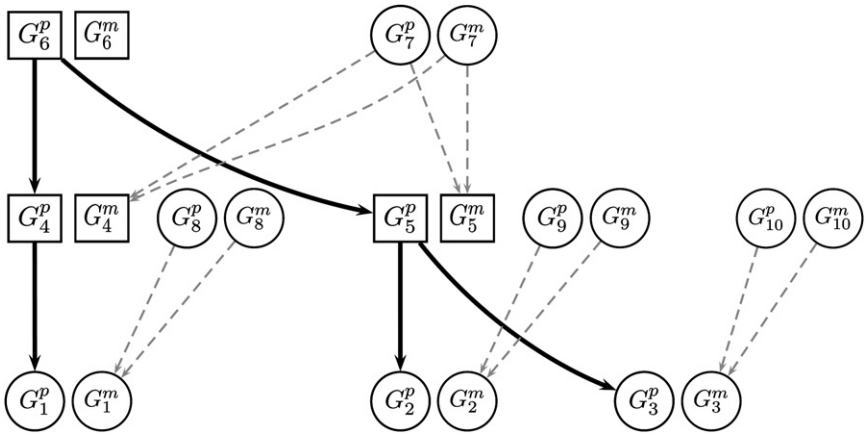

Figure 1. Descent Graph of the Top Configuration in Table 1

Squares represents alleles of males, and circles represent alleles of females. The solid black arrows indicate the transmission of founder allele G_6^p to individuals 1 and 2. Thus, the paternal allele G_1^p of individual 1 and the paternal allele G_2^p of individual 2 are IBD. The dashed gray arrows correspond to the p/m entries in Table 1.

defines an IBD configuration. Π takes as values the possible IBD configurations. Table 1 illustrates how different inheritance vectors are mapped to a state of the IBD variable Π, which in this example summarizes the IBD configuration of the paternal alleles 1, 2, and 3. The IBD configuration is uniquely determined by the seven segregation indicators listed in the table. The two possible values for each segregation indicator are paternal (p) and maternal (m). Segregation indicators for which both values yield the same IBD configuration are indicated with p/m. In total, there are 16 configurations of segregation indicators that imply IBD of G_1^p, G_2^p, and G_3^p simultaneously; the prior probability of this configuration thus is $P(\Pi = (G_1^p\, G_2^p\, G_3^p)) = 16/2^7 = 0.125$. In this example, the complexity has been reduced from 2^7 to 5 configurations.

Exact Single-Point Computation of Posterior IBD Probabilities

We first describe how the change of variables from $\mathbf{s}_{\mathrm{T_{nf}}}, \mathbf{s}_{\mathrm{A_{nf}}}$ to Π is implemented in the single-point case.

Table 1. Mapping of Configurations of Segregation Indicators s to IBD Configuration Π

Π	#	s_1^p	s_2^p	s_3^p	s_4^p	s_4^m	s_5^p	s_5^m
(1 2 3)	4	p	p	p	p	p/m	p	p/m
(1 2 3)	4	p	p	p	m	p/m	m	p/m
(1 2 3)	4	m	m	m	p/m	p	p/m	p
(1 2 3)	4	m	m	m	p/m	m	p/m	m
(1)(2 3)	4	p	p	p	p	p/m	m	p/m
(1)(2 3)	4	p	p	p	m	p/m	p	p/m
(1)(2 3)	16	m	p	p	p/m	p/m	p/m	p/m
(1)(2 3)	4	m	m	m	p/m	p	p/m	m
(1)(2 3)	4	m	m	m	p/m	m	p/m	p
(1)(2 3)	16	p	m	m	p/m	p/m	p/m	p/m
(1 2)(3)	4	p	p	m	p	p/m	p	p/m
(1 2)(3)	4	p	p	m	m	p/m	m	p/m
(1 2)(3)	4	m	m	p	p/m	p	p/m	p
(1 2)(3)	4	m	m	p	p/m	m	p/m	m
(1 3)(2)	4	p	m	p	p	p/m	p	p/m
(1 3)(2)	4	p	m	p	m	p/m	m	p/m
(1 3)(2)	4	m	p	m	p/m	p	p/m	p
(1 3)(2)	4	m	p	m	p/m	m	p/m	m
(1)(2)(3)	16	p	m	p	p/m	p/m	p/m	p/m
(1)(2)(3)	16	m	p	m	p/m	p/m	p/m	p/m

Π indicates the partitioning variable; paternal alleles of individuals within parentheses are IBD. # indicates the number of configurations of the segregation indicators.

By definition of the groups of individuals **D**, **T**, and **A**, the pedigree likelihood factors as follows (see Figure 2A)

$$P\left(\mathbf{M},\mathbf{G_D},\mathbf{G_T},\mathbf{G_A},\mathbf{s}_{\mathbf{D}_{\mathrm{nf}}},\mathbf{s_T},\mathbf{s}_{\mathbf{A}_{\mathrm{nf}}}|\mathbf{f}\right)=P\left(\mathbf{M},\mathbf{G_D},\mathbf{s}_{\mathbf{D}_{\mathrm{nf}}}|\mathbf{G_T},\mathbf{f}\right) \times P\left(\mathbf{G_T},\mathbf{G_A},\mathbf{s_T},\mathbf{s}_{\mathbf{A}_{\mathrm{nf}}}|\mathbf{f}\right). \quad (4)$$

Consequently, the variables $\mathbf{M},\mathbf{G_D},\mathbf{s}_{\mathbf{D}_{\mathrm{nf}}}$ are independent of $\mathbf{G_A},\mathbf{s_T},\mathbf{s}_{\mathbf{A}_{\mathrm{nf}}}$ conditioned on the ordered genotypes $\mathbf{G_T}$ (and the fixed allele frequencies $\mathbf{f}$). This implies that we only need to consider the right-most likelihood term $P(\mathbf{G_T},\mathbf{G_A},\mathbf{s_T},\mathbf{s}_{\mathbf{A}_{\mathrm{nf}}}|\mathbf{f})$ for the change of variables from $(\mathbf{s_T},\mathbf{s}_{\mathbf{A}_{\mathrm{nf}}})$ to Π.

Next, we derive the marginal likelihood of the variables $\mathbf{G_T}$ and Π from this likelihood term

$$P(\mathbf{G_T},\Pi\,|\,\mathbf{f}) = \sum_{\mathbf{G_A}} \sum_{(\mathbf{s_T},\mathbf{s}_{\mathbf{A}_{\mathrm{nf}}})\in\Pi} P\left(\mathbf{G_T},\mathbf{G_A},\mathbf{s_T},\mathbf{s}_{\mathbf{A}_{\mathrm{nf}}}|\mathbf{f}\right) = P(\mathbf{G_T}\,|\,\Pi,\mathbf{f})P(\Pi), \quad (5)$$

where $P(\Pi)$ is given by Equation 3. Equations 16–18 in the Appendix (Details of Single-Point Computations) provide a detailed derivation of Equation 5. Below, we give $P(\mathbf{G_T}|\Pi,\mathbf{f})$ for the example of Table 1. The conditional distribution of $P(\mathbf{G_T}|\Pi,\mathbf{f})$ does not depend on the particular configuration $(\mathbf{s_T},\mathbf{s}_{\mathbf{A}_{\mathrm{nf}}})\in\Pi$ but only on the IBD configuration determined by $(\mathbf{s_T},\mathbf{s}_{\mathbf{A}_{\mathrm{nf}}})$. This nontrivial result follows from the assumptions that the prior allele frequency distributions are the same for all founders, that no genotypes or phenotypes are observed for the individuals in **A**, and that paternal and maternal inheritance are equally likely a priori. It allows configurations $(\mathbf{s_T},\mathbf{s}_{\mathbf{A}_{\mathrm{nf}}})$ to be clustered with respect to their IBD configuration for the alleles in $\mathbf{G_T}$ in the computation of the posterior marginal distribution of Equation 2.

Combining Equations 1, 2, and 4 and the result Equation 5, we find that the single-point posterior is given by

$$P\left(\mathbf{s}_{\mathbf{D}_{\mathrm{nf}}},\Pi\,|\,\mathbf{M},\mathbf{f}\right) \propto \sum_{\mathbf{G_D},\mathbf{G_T}} P\left(\mathbf{M},\mathbf{G_D},\mathbf{s}_{\mathbf{D}_{\mathrm{nf}}}|\,\mathbf{G_T},\mathbf{f}\right)P(\mathbf{G_T}|\,\Pi,\mathbf{f})P(\Pi), \quad (6)$$

where the proportionality factor is given by $1/P(\mathbf{M}|\mathbf{f})$.

With the example of Table 1, we illustrate how the conditional probability $P(\mathbf{G_T}|\Pi,\mathbf{f})$ is calculated. Suppose we would like to know this probability for the IBD configuration $\Pi = (G_1^p)\,(G_2^p\, G_3^p)$, where allele G_2^p and G_3^p are IBD. G_2^p and G_3^p are a copy of the same founder allele, and G_1^p is a copy of a different founder allele. There are two contributions of the prior allele frequency distribution, one for each of the two transmitted founder alleles. This gives

$$P(G_1^p,G_2^p,G_3^p\,|\,\Pi=(G_1^p)(G_2^pG_3^p),\mathbf{f}) = P(G_1^p\,|\,\mathbf{f})P(G_2^p\,|\,\mathbf{f})\delta(G_2^p,G_3^p), \quad (7)$$

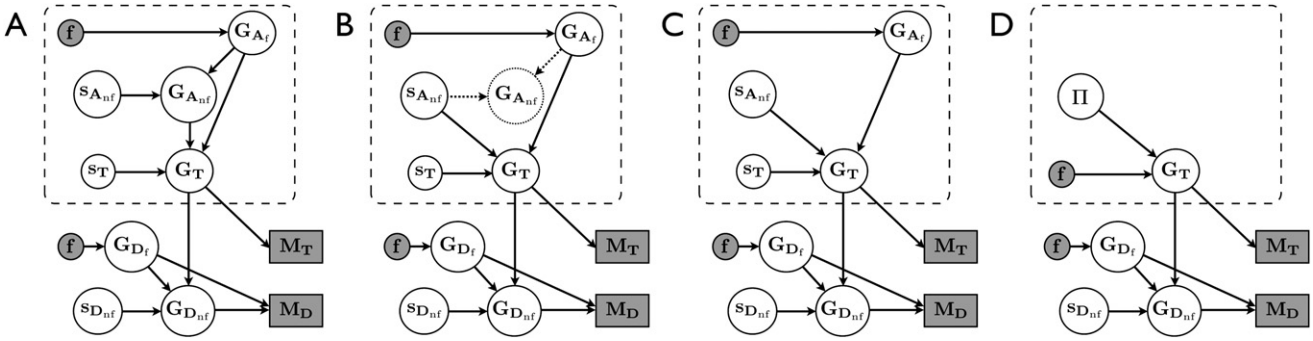

Figure 2. Graphical Model Representation

The graphical model reflects the conditional independencies of the single-point likelihood. White circles represent unobserved variables, gray circles represent model parameters assumed to be fixed and known, and gray rectangles represent observed variables, i.e., the marker genotypes. Panel (A) shows the graphical model corresponding to the single-point likelihood defined in terms of genotype variables and segregation indicators. Conditioned on the ordered genotypes $\mathbf{G_T}$, the variables shown in the dashed rectangle are independent of the variables outside the dashed rectangle. The graphical model can be alternatively constructed as in panel (B). As the group of ancestors (**A**) is defined to have no genotype or phenotype information, the corresponding ordered genotype variables (indicated by dotted lines) can be removed from the model, yielding the graphical model shown in panel (C). The HMM used by ALADIN is based on the model shown in panel (D). Here, the unobserved segregation indicators $\mathbf{s_T}$ and $\mathbf{s_{D_{nf}}}$ have been replaced by the IBD variable Π, which defines the IBD configuration of the alleles contained in $\mathbf{G_T}$. The ordered genotypes $\mathbf{G_A}$ of the ancestors (**A**) are not explicitly modeled.

where $\delta(x, y) = 1$ if $x = y$ and $\delta(x, y) = 0$ if $x \neq y$, with x and y discrete variables. This ensures that the conditional probability of $\mathbf{G_T}$ given Π is zero unless all alleles that are IBD have the same value. For each of the maternal alleles, there is a contribution of the allele frequency prior because these are never IBD on account of being inherited from founders:

$$P(G_1^m, G_2^m, G_3^m \mid \mathbf{f}) = P(G_1^m \mid \mathbf{f})P(G_2^m \mid \mathbf{f})P(G_3^m \mid \mathbf{f}). \quad (8)$$

The conditional probability distribution $P(\mathbf{G_T}|\Pi, \mathbf{f})$ is given by the product of the right hand sides of Equations 7 and 8.

Exact Multipoint Computation of Posterior IBD Probabilities

Given the results of the single-point case, the generalization to the multipoint case is straightforward. We define the multilocus IBD variable as $\Pi = \{\Pi^1, \Pi^2, \ldots, \Pi^L\}$, where L is the number of markers. (From here on, we will assume that unless a locus superscript l is specified, variables $\mathbf{G_T}$ represent multilocus genotypes.) The configurations of segregation indicators associated with a given multilocus IBD configuration Π are given by the cartesian product

$$\left\{\left(\mathbf{s}_\mathbf{T}^1, \mathbf{s}_{\mathbf{A}_{nf}}^1\right) \in \Pi^1\right\} \otimes \left\{\left(\mathbf{s}_\mathbf{T}^2, \mathbf{s}_{\mathbf{A}_{nf}}^2\right) \in \Pi^2\right\} \otimes \ldots \otimes \left\{\left(\mathbf{s}_\mathbf{T}^L, \mathbf{s}_{\mathbf{A}_{nf}}^L\right) \in \Pi^L\right\}. \quad (9)$$

Because we consider multilocus genotypes and segregation indicator configurations, the conditional independence of $(\mathbf{s}_{\mathbf{D}_{nf}}, \mathbf{G_D})$ and $(\mathbf{s_T}, \mathbf{s}_{\mathbf{A}_{nf}}, \mathbf{G}_{\mathbf{A}_f})$ given $\mathbf{G_T}$ still holds (see Figure 2B). As a result, the dependence of the posterior distribution on Π can again be determined from the likelihood term $P(\mathbf{G_T}, \mathbf{s_T}, \mathbf{s}_{\mathbf{A}_{nf}}, \mathbf{G}_{\mathbf{A}_f}|\mathbf{f}, \theta)$, where θ is the vector of recombination fractions. From Equation 5, the ordered genotypes $\mathbf{G}_\mathbf{T}^1$ for marker l are conditionally dependent only on Π^l and $\mathbf{f}^l$, so that the conditional probability distribution of $\mathbf{G_T}$ factorizes as a product of markers. Furthermore, in the exact HMM, there are conditional dependencies only between the segregation indicators (assuming linkage equilibrium). This means that the multipoint equivalent of Equation 5 is given by

$$P(\mathbf{G_T}, \Pi \mid \mathbf{f}, \theta) = P(\Pi \mid \theta) \prod_{l=1}^{L} P\left(\mathbf{G}_\mathbf{T}^l | \Pi^l, \mathbf{f}^l\right). \quad (10)$$

The prior distribution $P(\Pi|\theta)$ is explicitly given by

$$P(\Pi|\theta) = \sum_{(\mathbf{s}_{\mathbf{A}_{nf}}^1, \mathbf{s}_\mathbf{T}^1) \in \Pi^1} \cdots \sum_{(\mathbf{s}_{\mathbf{A}_{nf}}^L, \mathbf{s}_\mathbf{T}^L) \in \Pi^L} \prod_{i \in \mathbf{A}_{nf} \cup \mathbf{T}} \times \prod_{y=\{p,m\}} P\left(s_i^{1,y}\right) \prod_{l=2}^{L} P\left(s_i^{l,y} \mid s_i^{l-1,y}, \theta^{(l,l-1)}\right). \quad (11)$$

Here, the conditional distributions $P(s_i^{l,y}|s_i^{l-1,y}, \theta^{(l,l-1)})$ model the recombination between the markers l and $l-1$ for individual i, and y can be paternal (p) or maternal (m). In contrast with the prior distribution over the segregation indicators $\mathbf{s}$, the prior distribution $P(\Pi|\theta)$ is not first-order Markov in the IBD variables Π^l because of the summation and consequently does not factorize as a product over markers.

We can now write the multipoint generalization of Equation 6 in terms of $\mathbf{\Pi}$ by using Equation 10:

$$P\left(\mathbf{s}_{\mathbf{D}_{nf}}, \Pi \mid \mathbf{M}, \mathbf{f}, \theta\right) \propto P\left(\boldsymbol{M}, \mathbf{s}_{\mathbf{D}_{nf}}, \Pi \mid \mathbf{f}, \theta\right) = \sum_{\mathbf{G_D}, \mathbf{G_T}} P\left(\mathbf{M}, \mathbf{G_D}, \mathbf{s}_{\mathbf{D}_{nf}} \mid \mathbf{G_T}, \mathbf{f}\right) P(\Pi \mid \theta) \prod_l P\left(\mathbf{G}_\mathbf{T}^l \mid \Pi^l, \mathbf{f}^l\right). \quad (12)$$

The conditional probability distributions $P(\mathbf{G}_\mathbf{T}^l|\Pi^l, \mathbf{f}^l)$ are calculated for each locus as illustrated in the single-locus case above. The multipoint likelihood contains the product of a distribution conditional on Π and a prior distribution over Π, similar in form to the single-point likelihood (Equation 5). However, it is not particularly practical to work with because the summation over the subspace of Equation 9 required for the computation of $P(\Pi|\theta)$ is generally not feasible for large number of markers and prevents application of an efficient forward-backward algorithm.

ALADIN: Exact Inference in an Approximate HMM

We propose to approximate $P(\Pi|\theta)$ with a distribution that is first-order Markov in Π^l. We further assume that the conditional

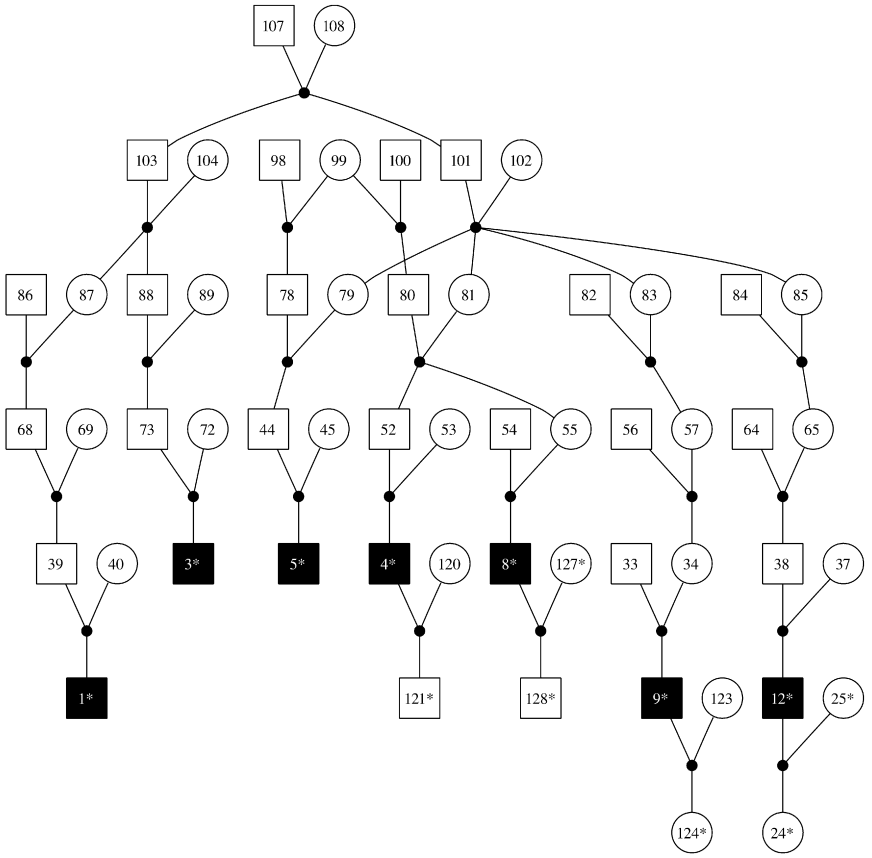

Figure 3. Prostate Cancer Pedigree I
Affected individuals are represented by a black symbol, and genotyped individuals are indicated with an asterisk.

probability distribution of Π^l given Π^{l-1} depends only on the recombination fraction $\theta^{(l,l-1)}$ between markers l and $l-1$. This first-order Markov approximation is given by

$$P(\Pi|\theta) \approx Q^{(1)}(\Pi \mid \theta) \equiv P(\Pi^1) \prod_{l=2}^{L} P'\left(\Pi^l \mid \Pi^{l-1}, \theta^{(l,l-1)}\right), \qquad (13)$$

where the superscripted (1) indicates the first-order approximation. $P(\Pi^1)$ is given by Equation 3. The conditional distribution is computed as follows:

$$P'\left(\Pi^l|\Pi^{l-1}, \theta^{(l,l-1)}\right) = \frac{P'\left(\Pi^l, \Pi^{l-1} \mid \theta^{(l,l-1)}\right)}{P\left(\Pi^{l-1}\right)}, \qquad (14)$$

where the prime indicates that we compute the marginal distribution in the numerator on the right-hand side exactly in a two-locus model.

The computation of these probabilities is a crucial step in ALADIN. Because the pedigree structure is the same for each marker, the prior IBD-partitioning probabilities are also the same for each marker,

$$P\left(\Pi^l\right) = P\left(\Pi^1\right) \; \forall \; l,$$

so that this computation has to be performed only once. The computation of the conditional probabilities must be performed for every recombination fraction, i.e., for every pair of adjacent markers. In the special case that all recombination fractions between the markers are the same, this computation also would have to be performed only once because the conditional probabilities depend only on the recombination fraction for a given pedigree. We use a dynamic programming algorithm based on variable elimination[24,25] to compute these probabilities exactly. This procedure is outlined in the Appendix (Computation of Prior IBD Probabilities with Variable Elimination). For practical application of the approximation, we require that exact computations of $P'(\Pi^l|\Pi^{l-1}, \theta^{(l,l-1)})$ in the two-locus model with the junction tree algorithm are feasible.

The full likelihood of the approximate model based on Equation 13 is obtained by substituting this equation into Equation 12:

$$\begin{aligned} P\left(\mathbf{s}_{\mathbf{D}_{\mathrm{nf}}}, \Pi \mid \mathbf{M}, \mathbf{f}, \theta\right) &\propto P\left(\mathbf{M}, \mathbf{s}_{\mathbf{D}_{\mathrm{nf}}}, \Pi \mid \mathbf{f}, \theta\right) \\ &\approx \sum_{\mathbf{G_D}, \mathbf{G_T}} P\left(\mathbf{M}, \mathbf{G_D}, \mathbf{s}_{\mathbf{D}_{\mathrm{nf}}} | \mathbf{G_T}, \mathbf{f}\right) \\ &\quad \times Q^{(1)}(\Pi \mid \theta) \prod_l P\left(\mathbf{G}_\mathbf{T}^l | \Pi^l, \mathbf{f}^l\right). \qquad (15) \end{aligned}$$

The main idea of ALADIN is to perform exact inference of the marginal distributions $P(\mathbf{s}^l_{\mathbf{D}_{\mathrm{nf}}}, \Pi^l|\mathbf{M}, \mathbf{f}, \theta)$ in the approximate HMM defined by Equation 15 with the Lander-Green algorithm. Exact inference in the approximate model will result in approximate IBD probabilities.

In summary of the procedures, the result of the change of variables from $(\mathbf{s_T}, \mathbf{s}_{\mathbf{A}_{\mathrm{nf}}})$ to Π is that the size of the hidden state space depends no longer on $|\mathbf{A}|$, the number of untyped ancestors, but only on $|\Pi|$, the number of possible IBD configurations of the alleles of the individuals in $\mathbf{T}$. The complexity of computing marginal distributions $P(\mathbf{s}^l_{\mathbf{D}_{\mathrm{nf}}}, \Pi^l|\mathbf{M}, \mathbf{f}, \theta)$ in the approximate HMM may be substantially lower than in the exact HMM. Therefore, application of the forward-backward algorithm to the approximate HMM can be feasible when application to the exact HMM is not.

Implementation

We have implemented the Lander-Green algorithm and the algorithm to compute the prior IBD probabilities in the computer program ALADIN. It calculates normalized $\mathrm{NPL}_{\mathrm{pairs}}$ and $\mathrm{NPL}_{\mathrm{all}}$ linkage statistics, as well as parametric LOD scores (as described in Appendix [Estimation of Multipoint Parametric LOD Scores]) for general pedigrees in the approximate HMM. It can analyze diallelic and multiallelic marker data sets. When dealing with dense SNP arrays, the use of which has rapidly become standard practice, it is important that linkage disequilibrium between the markers be accounted for. Therefore, we have implemented the same clustered-markers approach as MERLIN.[7] This approach clusters markers into haplotype blocks for which haplotype frequencies must be specified. Markers within a haplotype block may be in complete LD; markers in different haplotype blocks are assumed to be in linkage equilibrium. Absence of recombination is assumed for markers in the same haplotype block. ALADIN accepts standard LINKAGE-formatted locus and pedigree files and uses the

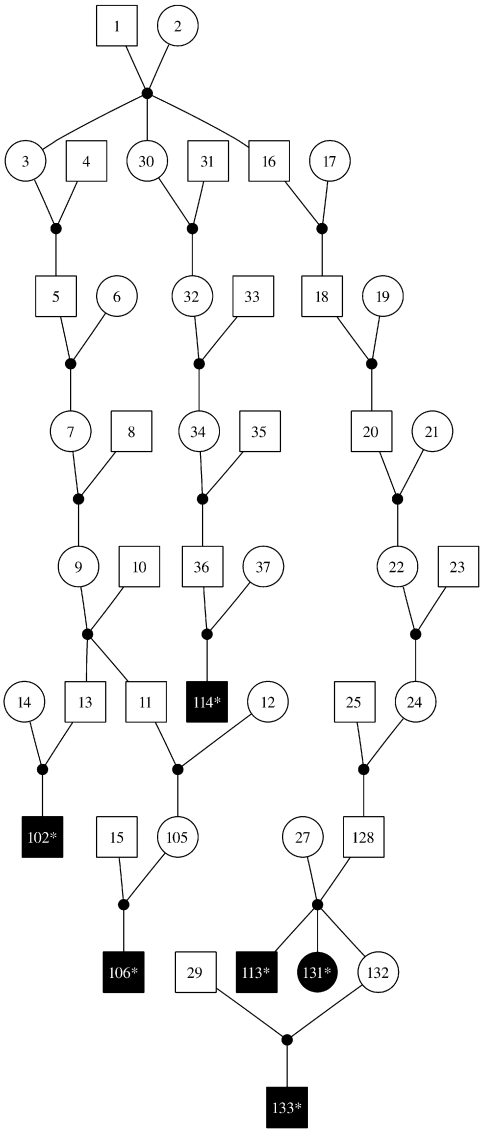

Figure 4. Pedigree II
The pedigree was taken from a pituitary adenoma study of Vierimaa et al.[2] Affected individuals are represented by a black symbol, and genotyped individuals are indicated with an asterisk.

MERLIN format to specify the haplotype blocks and haplotype frequencies.

Results

Setup

We evaluated the performance of ALADIN in simulated and real data sets. With data sets simulated in small pedigrees where exact multipoint computation with MERLIN was feasible, we evaluated the quality of the ALADIN approximation of NPL_{pairs} and NPL_{all}. We compared these results with those of the state-of-the-art MCMC sampler MORGAN, where we note that MORGAN only provides an approximation of the NPL_{pairs} statistic. In the comparison with MORGAN, we also assessed the accuracy of approximations of parametric LOD scores. With data sets simulated in large pedigrees where exact multipoint computation with MERLIN was not feasible, we estimated the type I error rate of ALADIN. We also compared the ALADIN and MORGAN approximation of NPL_{pairs} with the value of NPL_{pairs} of the true inheritance vector at each location for chromosomes where linkage was simulated. In a real data set where exact computation with MERLIN was feasible, we compared the accuracy of ALADIN and MORGAN. Finally, we compared computation time of ALADIN, MORGAN, and MERLIN.

We used two pedigrees taken from practical linkage studies in our evaluation. Pedigree I, shown in Figure 3, was taken from a prostate cancer (PC [MIM 176807]) study. Pedigree II, shown in Figure 4, was taken from a pituitary adenoma predisposition study.[2] Because these pedigrees are too large for exact computation with MERLIN (number of bits > 25, see Table 2), we considered three subpedigrees for our comparisons between ALADIN and MERLIN: (1) pedigree Ia, the subpedigree of I consisting of the cases 4, 8, 9, 12, their ancestors, spouses, and children, (2) pedigree Ib, the subpedigree of I consisting of only the cases 4, 8, 9, 12, and their ancestors, and (3) pedigree IIa, the subpedigree of II consisting of the cases 102, 114, 131, and the ancestors of these cases.

We used the lm_ibdtests and lm_markers programs in the MORGAN 2.8.1 package to obtain MCMC approximations of NPL_{pairs} and parametric LOD scores, respectively. MORGAN requires the user to specify a number of parameters. We used a default value of 100,000 scans for the sampler. The number of burn-in scans and sequential imputation steps for initialization of the Markov chain were set to the values recommended by Wijsman et al.[14] in a comprehensive evaluation of the MCMC programs MORGAN and SIMWALK2. With these settings, the authors showed that MORGAN can be expected to produce accurate approximations and that it is generally more efficient than SIMWALK2.

All analyses were performed on a small cluster of five AMD 64 bit machines with two dual-core 2.2 GHz processors. Each machine had 16 GB of physical memory available and 30 GB of swap space.

Simulation Study

Comparison with MERLIN

We assessed the accuracy of ALADIN and MORGAN in data sets simulated for pedigrees Ia, Ib, and IIa. We emulated a genome scan by generating marker data for the autosomal chromosomes. To see whether the linkage signal could be accurately detected, we simulated linked chromosomes where the inheritance vector at the middle marker was fixed such that all cases shared one allele IBD, which is the maximum for these pedigrees because there is no inbreeding. To determine a possible bias, we also simulated unlinked chromosomes where the inheritance vector at the middle marker was randomly generated.

Marker data was simulated under two conditions: linkage equilibrium (LE) and linkage disequilibrium (LD) between the markers. In the condition of LE, the founder alleles were sampled with the estimated allele frequencies

Table 2. Computation Time

Pedigree					Computation Time (min)[a]				
					LE			LD[b]	
Label	Cases	*T*	*P*	HMM Complexity[c]	MERLIN	ALADIN	MORGAN[d]	MERLIN	ALADIN
Ia	4	4	32	18	422.5	35.0	35460	11000	13.5
Ib	4	4	24	14	34.0	69.5	32855	148.5	24.5
IIa	3	3	35	18	288.5	10.5	38115	2191	4.0
I	7	7	53	36	∞[e]	49535	47510	∞	16282
II	6	4	44	27	∞	171.5	41760	∞	78

[a] Computation time is reported as estimated time required to analyze the 50K Affymetrix XbaI array in minutes.
[b] MORGAN cannot handle LD.
[c] Complexity is measured as the $\log_2$ of the number of hidden states of the exact HMM for a marker (the number of bits).
[d] MORGAN was run with 100,000 MCMC scans.
[e] ∞ indicates that exact computation with MERLIN was not feasible.

of the Affymetrix 50K XbaI array in 42 European ancestry samples provided by Affymetrix. In this dataset, there are 57,093 polymorphic autosomal markers, with mean minor allele frequency 0.22. In the condition of LD, founder haplotypes were sampled with haplotype block definitions and haplotype frequency estimates obtained with HAPLOVIEW[26] (spine-of-LD rule with $D' = 0.8$). HAPLOVIEW was run on data from Phase I of the International HapMap Project,[27] consisting of the genotypes of 90 individuals from Utah of northwestern European ancestry at the markers of the 50K XbaI array. There were 51,634 polymorphic, autosomal markers on the array that were in the Phase I dataset and passed HAPLOVIEW's quality control tests. They were assigned to 18,343 haplotype blocks (including 6103 blocks containing only one SNP). The maximum number of SNPs in any one block was 28. The number of haplotypes per block ranged from 2 to 31 (mean 3.4), and haplotype frequencies ranged from 0.01 to 0.98 (mean 0.29). Simulations were performed with the assumption of linkage equilibrium between neighboring blocks. Table 3 gives an overview of the number of replicates simulated in each condition.

Because the clustered-markers option for modeling LD in MERLIN required significantly more computation time than the standard analysis assuming LE, only 22 chromosomes in the condition of LD were simulated and analyzed with ALADIN and MERLIN. Because of the long computation time of MORGAN, we analyzed only one replicate of autosomal chromosomes 1, 3, 20, and 22 with MORGAN. Furthermore, MORGAN cannot model LD, so no chromosomes were analyzed in the LD condition (indicated by "NA" in the table).

Table 3. Number of Replicates Simulated for the Comparison with MERLIN

	LE		LD	
Type	ALADIN	MORGAN	ALADIN	MORGAN
unlinked	$2 \times 22 = 44$	$1 \times 4 = 4$	$1 \times 22 = 22$	NA
linked	$2 \times 22 = 44$	$1 \times 4 = 4$	$1 \times 22 = 22$	NA

NA indicates that MORGAN cannot model LD.

We used three measures to evaluate the accuracy: (1) Δ_{mean}, the mean absolute difference over all markers on the chromosome between the exact and approximate scores, (2) Δ_{max}, the maximum absolute difference over all markers on the chromosome between the exact and approximate scores, and (3) Δ_{mid}, the absolute difference evaluated at the middle marker on the chromosome, where for the linked chromosomes the inheritance vector was fixed.

Table 4 shows the error in the ALADIN approximation of NPL_{pairs} and NPL_{all} for the various pedigrees and the conditions of linked and unlinked chromosomes and LE and LD. Note that for the comparison of the absolute errors across different pedigrees, it is important that the range, i.e., the difference between the maximum and minimum value of NPL_{pairs} and NPL_{all}, be taken into account. The maximum score was attained in almost all linked chromosomes at the location of the middle marker. Overall, the approximation of NPL_{pairs} and NPL_{all} was accurate with values of Δ_{mean} less than 0.4 in all pedigrees and conditions. The absolute error at the middle marker Δ_{mid} was smaller than 0.08 for pedigrees Ia and Ib and smaller than 1.4 for pedigree IIa. The errors Δ_{mean} and Δ_{mid} were smaller for the unlinked chromosomes than for the linked chromosomes. The relative errors as a fraction of the exact value at the middle marker (shown between parentheses for the linked chromosomes) were smaller than 3%, except for the condition of LE for pedigree IIa. The large relative error for this pedigree was caused by two replicates with a large discrepancy at the location where linkage was simulated, which we examine in more detail below. On the remaining replicates, the mean error Δ_{mid} was 0.5386 (1.42%), comparable with the results for the linked chromosomes in the condition of LD for this pedigree. Thus, in general, ALADIN accurately detected the linkage signal.

The mean maximum errors (Δ_{max}) were larger than Δ_{mean} and Δ_{mid}, varying from 0.0912 to 0.7832 for the unlinked chromosomes and from 0.1973 to 6.807 for the linked chromosomes. We observed two typical situations

Table 4. Error of ALADIN in NPL_{pairs} and NPL_{all}

			NPL_{pairs}				NPL_{all}			
Pedigree	Type	*N*	Range	Δ_{mean}	Δ_{max}	Δ_{mid} (Relative)[a]	Range	Δ_{mean}	Δ_{max}	Δ_{mid} (Relative)
Linkage Equilibrium										
Ia	unlinked	44	9.90	0.0022	0.0912	0.0010	16.7	0.0020	0.0867	0.0008
Ia	linked	44	9.90	0.0047	0.1973	0.0011 (0.0155%)	16.7	0.0052	0.2243	0.0026 (0.0209%)
Ib	unlinked	44	9.90	0.0070	0.1647	0.0024	16.7	0.0067	0.1695	0.0023
Ib	linked	44	9.90	0.0154	0.3483	0.0107 (0.1476%)	16.7	0.0208	0.5199	0.0227 (0.1996%)
IIa	unlinked	44	50.7	0.0179	0.3017	0.0226	66.3	0.0177	0.2977	0.0223
IIa	linked	44	50.7	0.3868	6.807	1.357 (15.55%)	66.3	0.4585	9.295	1.874 (17.73%)
Linkage Disequilibrium										
Ia	unlinked	20[b]	9.90	0.0102	0.7533	0.0011	16.7	0.0092	0.6504	0.0011
Ia	linked	20[b]	9.90	0.0200	1.104	0.0070 (0.0854%)	16.7	0.0243	1.713	0.0160 (0.1159%)
Ib	unlinked	22	9.90	0.0149	0.7832	0.0025	16.7	0.0133	0.7296	0.0022
Ib	linked	22	9.90	0.0362	1.521	0.0776 (1.786%)	16.7	0.0460	2.333	0.1761 (3.498%)
IIa	unlinked	22	50.7	0.0032	0.1069	0.0060	66.3	0.0032	0.1054	0.0059
IIa	linked	22	50.7	0.2310	6.370	0.5279 (2.423%)	66.3	0.2845	8.881	0.7722 (2.641%)

Note that values of Δ are shown as means across *N* simulated chromosomes.

[a] Relative error is only reported for the linked chromosomes, where linkage was simulated at the location of the middle marker.

[b] MERLIN terminated with error status when chromosomes 1 and 2 were analyzed.

where these maximum errors occurred. The first situation was the most general: Here, the maximum error was found in a small region where the value of the statistic changed sharply as a function of the location and did not affect the value of the statistic at the location where linkage was simulated. Figure 5A illustrates this situation for a replicate of pedigree IIa. The second situation was found in only two replicates, both for pedigree IIa: Here, the error extended over a larger region that included the location where linkage was simulated, as illustrated in Figure 5B. For the condition of LD, we found that in pedigrees Ia and Ib, the value of Δ_{max} was mostly attained at the beginning of the chromosome. We believe that this may be partly explained by slight differences in our implementation of the clustered-markers approach and that of MERLIN regarding how inconsistencies due to recombination events within a haplotype cluster are dealt with. The number of inconsistencies was on average 1.3 ± 1.6 per chromosome per pedigree.

In Table 5, we compare the error of ALADIN and MORGAN in NPL_{pairs} on the subset of chromosomes analyzed by MORGAN. ALADIN had smaller values of Δ_{mean} and Δ_{max} than did MORGAN in pedigrees Ia and IIa and larger values in pedigree Ib. The values of Δ_{mid} of ALADIN were smaller than those of MORGAN in all pedigrees. In addition to the replicates shown in the table, we analyzed the two replicates for pedigree IIa where the error of ALADIN extended over a larger region (described above) with MORGAN. Interestingly, for both of these replicates, MORGAN produced the same result as ALADIN, with the corresponding error. Figure 5B illustrates this for one of these replicates. Here, we found a relatively large difference between the ALADIN approximation and the exact value of NPL_{pairs}: Both ALADIN and MORGAN yielded a similar

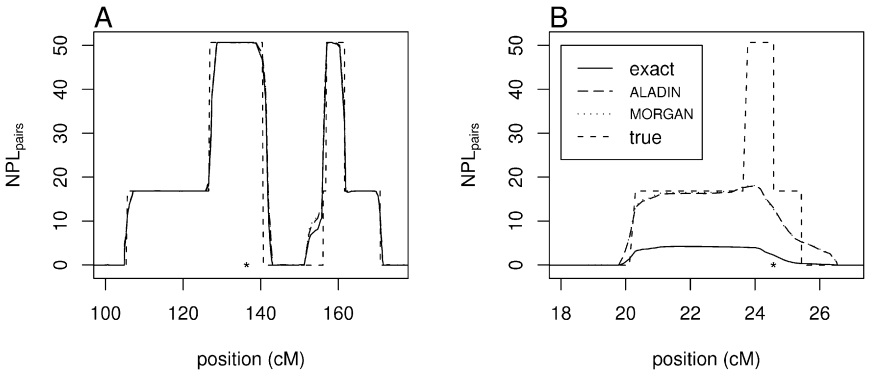

Figure 5. Comparison of ALADIN and MORGAN with Exact Method

The figure shows exact, approximate, and true simulated value of NPL_{pairs} for two typical replicates of pedigree IIa in the condition of linked chromosomes and linkage equilibrium. The location where linkage was simulated is indicated with the asterisk. Forty-two of the 44 simulated replicates were similar to (A). For this replicate, Δ_{mean} of ALADIN and MORGAN were 0.0913 and 0.1015, respectively. The maximum errors Δ_{max} were respectively 4.439 and 4.571, and were both attained at 142.4 cM, a region where NPL_{pairs} changed rapidly. In two replicates, the situation was as in (B): ALADIN and MORGAN produced similar scores that both overestimated NPL_{pairs} as compared to the value obtained with MERLIN but did not overestimate the value of NPL_{pairs} of the true inheritance vector. Here, the maximum errors of ALADIN and MORGAN were 14.05 and 14.15, respectively, and were attained at 23.96 cM by both methods.

Table 5. Comparison of Error in NPL_{pairs} of ALADIN and MORGAN

				Δ_{mean}		Δ_{max}		Δ_{mid} (Relative)	
Pedigree	Type	*N*	Range	ALADIN	MORGAN	ALADIN	MORGAN	ALADIN	MORGAN
Ia	unlinked	4	9.90	0.0051	0.0071	0.0923	0.3870	0.0059	0.0104
Ia	linked	4	9.90	0.0043	0.0103	0.1671	0.4826	0.0090 (0.1306%)	0.0431 (0.5548%)
Ib	unlinked	4	9.90	0.0085	0.0077	0.3206	0.1078	0.0011	0.0067
Ib	linked	4	9.90	0.0568	0.0274	0.6345	0.1528	0.0135 (0.1565%)	0.0623 (0.7058%)
IIa	unlinked	4	50.7	0.0044	0.0047	0.0489	0.0489	0.0001	0.0007
IIa	linked	4	50.7	0.3287	0.5099	4.396	4.814	1.000 (2.265%)	1.312 (3.000%)

Note that values of Δ are shown as means across *N* simulated chromosomes.

score that underestimated the true value. This score was closer to the value of the true inheritance vector than the exact score. We believe that this phenomenon is most likely due to multimodality of the posterior distribution that the approximate methods did not fully take into account. We found no replicates where ALADIN overestimated the value of NPL_{pairs}, and MERLIN (yielding exact results) did not overestimate it as well.

In Table 6, we compare the error of ALADIN and MORGAN in parametric LOD scores. We assumed a disease allele frequency of 0.001 and penetrance values (0.001, 0.20, 0.20), reflecting a dominant disease with low penetrance. We analyzed the same replicates as those used for Table 6. ALADIN and MORGAN were both accurate, with $\Delta_{mean} < 0.10$ and $\Delta_{mid} < 0.11$ for all pedigrees. The maximum errors Δ_{max} of ALADIN and MORGAN were similar and varied from 0.06 to 0.70. Because the LOD scores ranged from −2 to 3.5 for the linked chromosomes, the maximum errors in the parametric scores were similar to the maximum errors found for the nonparametric scores. We conclude that ALADIN was accurate and achieved a similar performance as MORGAN.

Large Pedigrees

We evaluated the performance of ALADIN in pedigrees I and II. Exact multipoint computation of NPL_{pairs} with MERLIN in these pedigrees is not feasible. However, it is possible to calculate the exact null distribution as single-locus computations are feasible in these pedigrees. The type I error rate (false-positive rate) of an exact method applied to fully informative marker data is given by the probability that the NPL_{pairs} statistic is larger than the significance threshold under the exact null distribution. We used an importance-sampling approach to estimate the type I error rate of ALADIN for high values of the significance threshold on NPL_{pairs} (i.e., small p values). For every possible value of NPL_{pairs}, we simulated 75 inheritance vectors for the middle marker location; genotypes for 200 markers were simulated conditional on the inheritance vector according to the specifications of the XbaI array. For pedigree I, there are 13 unique values of NPL_{pairs}, resulting in a total of 975 replicates. For pedigree II, there are 14 unique values of NPL_{pairs}, resulting in a total of 1050 replicates. The replicates were given the proper importance weights so that the fact that they were not drawn from the exact null distribution could be accounted for. ALADIN was used for the approximation of NPL_{pairs} at the middle marker location in each replicate. This procedure yields unbiased estimates of the type I error rate of ALADIN.

Table 6. Comparison of Error in Parametric LOD Scores of ALADIN and MORGAN

			Δ_{mean}		Δ_{max}		Δ_{mid}	
Pedigree	Type	*N*	ALADIN	MORGAN	ALADIN	MORGAN	ALADIN	MORGAN
Ia	unlinked	4	0.0128	0.0116	0.4135	0.5274	0.0253	0.1092
Ia	linked	4	0.0245	0.0081	0.6925	0.5284	0.0012	0.0015
Ib	unlinked	4	0.0152	0.0114	0.3298	0.3680	0.0018	0.0015
Ib	linked	4	0.0405	0.0121	0.3386	0.2233	0.0014	0.0023
IIa	unlinked	4	0.0052	0.0056	0.0733	0.0931	0.0003	0.0004
IIa	linked	4	0.0306	0.0719	0.4002	0.6154	0.0146	0.0150

Note that values of Δ are shown as means across *N* simulated chromosomes.

Figure 6 shows the empirical type I error rate for ALADIN and the type I error rate corresponding to the exact null distribution, with NPL_{pairs}. We find that ALADIN did not have an inflated type I error rate for the high values of the significance threshold that are relevant for genome-wide linkage analysis.

We also compared ALADIN and MORGAN in pedigree II for chromosomes where linkage was simulated at the middle marker. The inheritance vector at this marker was fixed such that IBD sharing between the cases was maximal. Marker data was simulated for eight autosomal chromosomes (1, 3, 5, 7, 11, 13, and 17) assuming linkage equilibrium between the markers. For each marker location, we computed the value of NPL_{pairs} as estimated from the marker data by using ALADIN and MORGAN. Because exact computation was not feasible, we compared these with the value of NPL_{pairs} of the true inheritance vector at each marker location.

For all eight chromosomes, the score estimated with ALADIN was very close to the score of the true inheritance vector at each marker (Figure 7). For two chromosomes, there were regions where MORGAN reported a score that was lower than the score of the true inheritance vector. In one case, this region included the marker location where linkage was simulated. Although the exact multipoint value of NPL_{pairs} given the marker data is not known, this figure suggests that the first-order Markov approximation of ALADIN has sufficient power to detect IBD sharing

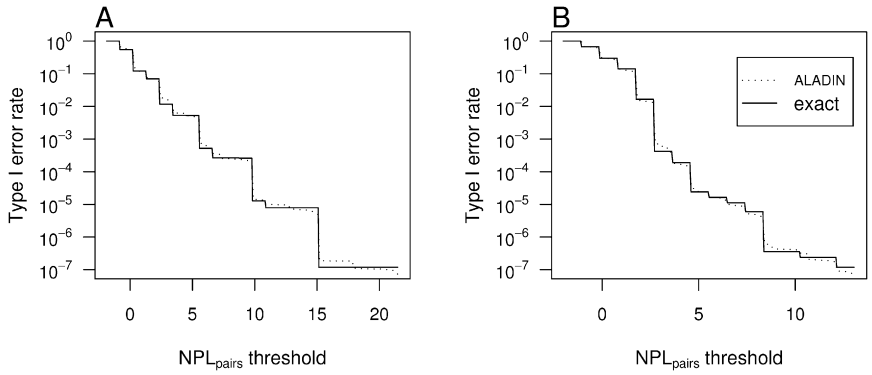

Figure 6. Empirical Type I Error Rate of ALADIN

(A) shows the empirical type I error rate for NPL_{pairs} in pedigree I. (B) shows empirical type I error rate for NPL_{pairs} in pedigree II. Note that the p value corresponding to a given significance threshold on NPL_{pairs} is given by the solid curve.

among the cases. Because we found that the type I error rate was not inflated, we infer that ALADIN might be expected to produce accurate results in data sets where exact computation is not feasible.

Application to Real Data

We compared the accuracy of ALADIN and MORGAN in the real data set of pedigree Ia. The marker data are from the SNPs of the Affymetrix 10K array. For each pair of SNPs in strong LD ($D' > 0.8$), one of the SNPs was removed from the data set for the prevention of spurious linkage results.

We first compared the accuracy of ALADIN and MORGAN for different numbers of MCMC scans for MORGAN, by using the NPL_{pairs} statistic. We focused the subset of chromosomes 1–10 in the real data set for pedigree Ia. The exact NPL_{pairs} scores were computed with MERLIN. Figure 8 shows that with 10,000 MCMC scans, ALADIN was more accurate than MORGAN; with 100,000 MCMC scans, ALADIN was less accurate than MORGAN. Thus, for a small number of MCMC scans, MORGAN was both slower and less accurate than ALADIN, whereas the accuracy of ALADIN was already high.

Second, we evaluated the accuracy of ALADIN by using all of the autosomal chromosomes. The mean error in NPL_{pairs} of ALADIN was 0.019, and the maximum error was 0.32. The mean and maximum errors NPL_{all} were 0.032 and 0.35, respectively. The maximum exact values of NPL_{pairs} and NPL_{all} were 7.65 and 12.8, respectively (at the same location); the relative errors of ALADIN at this peak were 0.0029% and 0.0037%, respectively. We conclude that the approximation of ALADIN in the real data set was accurate.

We also applied ALADIN to the full pedigree. Here, ALADIN replicated an analysis with a pedigree-splitting approximation[28] that found that there was just one suggestive linkage peak in this pedigree where four out of seven cases inherited a haplotype identical by descent. This haplotype sharing was confirmed by subsequent microsatellite genotyping and haplotype reconstruction with SIMWALK2 (L.M. FitzGerald, personal communication).

Computation Time

The computation time of MERLIN increases linearly with the number of markers and exponentially with the number of bits, which is the number of components of the inheritance vector required for exact computation in the HMM or, equivalently, the $\log_2$ of the number of hidden states of the exact HMM for a single marker. Computation time of ALADIN increases exponentially with T, the number of individuals in **T**, and linearly with the number of markers. MORGAN requires that single-locus exact computations are feasible. If the pedigree is not inbred, which was the case in all of our analyses, computation time increases linearly with P, the number of individuals in the pedigree.

Table 2 shows computation times for the various pedigrees we analyzed, reported as the estimated computation time required to analyze the 50K Affymetrix XbaI array. The computation time of ALADIN was mostly shorter than that of MERLIN for the pedigrees where exact computation with MERLIN was practical, especially when LD was modeled. Pedigree Ib without LD modeling forms an exception, which can be most likely attributed to the efficient implementation of MERLIN. Computation times of ALADIN were several orders of magnitude shorter than those of MORGAN. Analysis of pedigrees II and I with MERLIN was not practical. ALADIN was significantly faster than MORGAN in pedigree II. In pedigree I, the efficiency of ALADIN and MORGAN was comparable for the case of LE. Again, ALADIN was more efficient when LD was modeled.

We studied how computation time of ALADIN and MORGAN scaled with the number of markers. We found that computation time of MORGAN increased quadratically with the number of markers analyzed (Figure 9A), with a fixed number of 100,000 scans for the sampler. As expected, computation time of ALADIN increased linearly.

ALADIN was designed for the purpose of analyzing distantly related individuals. We therefore investigated computation time as a function of A, the number of untyped ancestors through which the cases are related, for a fixed value of $T = 4$. The structure of the pedigree used in this simulation was as follows: The four cases were related by two common ancestors 3–15 generations back, where each case was in a separate branch of the pedigree. The cases formed the group **T**, and the group **D** did not contain any individuals. Figure 9B shows computation time of ALADIN and MORGAN for replicates simulated with 100 markers. Computation time of ALADIN did not clearly

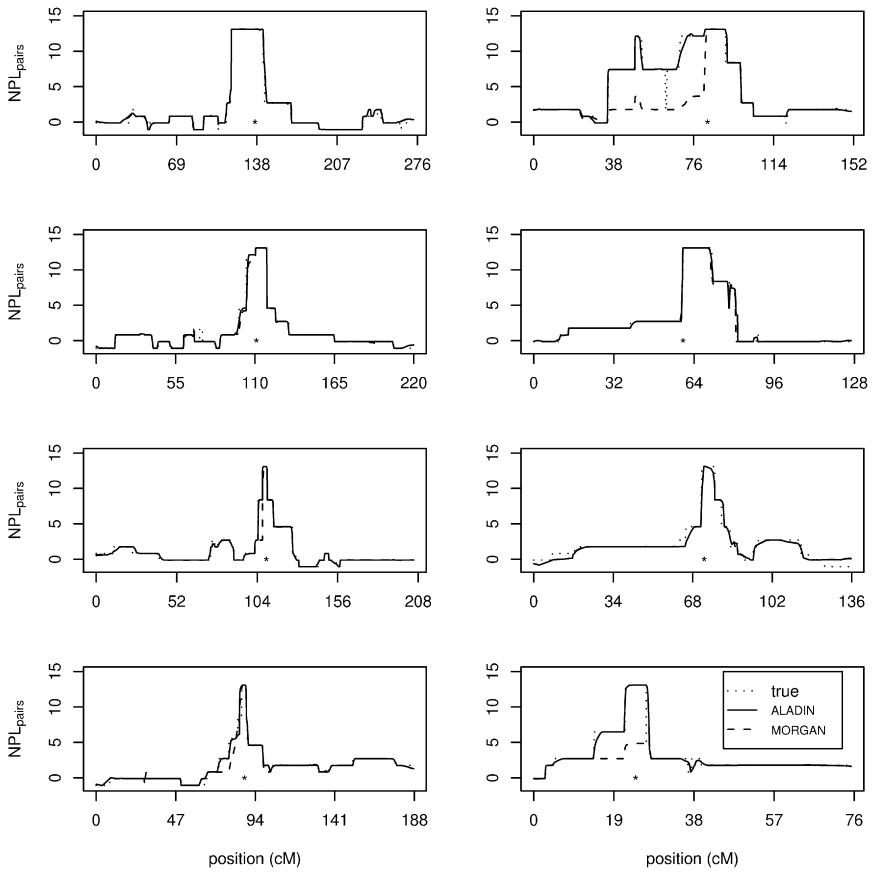

Figure 7. Evaluation of ALADIN in a Large Pedigree

The ALADIN and MORGAN estimates of NPL_{pairs} are compared to the NPL_{pairs} of the true inheritance vector for eight autosomal chromosomes (1, 3, 5, 7, 11, 13, and 17) for pedigree II. Exact multipoint computation of NPL_{pairs} was not feasible. The inheritance vector at the middle marker, indicated by the asterisk at the horizontal axis, was fixed such that all six cases shared one allele IBD at that location.

show an increase with A. Computation time of MORGAN increased linearly with A and was significantly higher than that of ALADIN. Thus, for small T and large A, ALADIN may be significantly more efficient than MORGAN.

Discussion

We presented ALADIN, a program for linkage analysis of distantly related individuals. ALADIN produced accurate estimates of nonparametric linkage and parametric linkage scores. ALADIN also produced accurate estimates when linkage disequilibrium was taken into account with the same clustered-markers approach as MERLIN. Accuracy was comparable with that of the state-of-the-art MCMC program MORGAN. We have shown that ALADIN is especially useful when a moderate number of cases are related through common ancestors many generations back and the pedigree is too large to analyze with exact methods. ALADIN may be several orders of magnitude more efficient than MORGAN, depending on the number of typed individuals and cases. However, it should be noted that we performed comparisons with version 2.8.1 of MORGAN, which uses a basic locus meiosis (LM) sampler that was not designed for the pedigrees with long descent chains considered in this paper. Nevertheless, we believe that it is the most powerful alternative to ALADIN for these pedigrees. The recently released version 2.8.2 of MORGAN uses a multimeiosis sampler[29] that will most likely perform better for linkage analysis of distantly related individuals. ALADIN may also be of use in pedigrees where analysis with MERLIN in principle is feasible but very time consuming, for instance, when the clustered-markers approach is used to account for linkage disequilibrium. In the current implementation, ALADIN does not require any parameters to be specified, which can be an advantage over MCMC-based programs, depending on the expertise of the user.

The calculation of the conditional probabilities between the states of the IBD variables of different loci may account for a large portion of the computation time of ALADIN. Because the pedigree structure is obviously the same for different pairs of adjacent markers, the only quantity varying is the recombination fraction. It is likely that many pairs of markers have very similar recombination fractions. One can make an additional approximation by using a discretized set of recombination fractions for which the conditional IBD probabilities will be computed. Then for any pair of markers encountered in the real data set, the IBD probabilities computed for the fraction in the discretized set that is closest to the true fraction can be used as an approximation. The computation of the conditional IBD probabilities can be easily performed in parallel. ALADIN currently has an option for creating and using a collection of recombination fractions so that the computation time can be reduced. As an example, for pedigree I, the computation time can be reduced to 4600 min from 16,282 min with a discretized set of 1000 marker distances. The relative error in the marker recombination frequencies due to discretization is small, at 0.36 % $\pm$ 0.26%. We expect that the effect of this approximation on the conclusions of the linkage study will be negligible.

An additional way to reduce computation time is the use of multiple IBD variables with fewer IBD configurations, thus optimizing the trade-off between time spent on the HMM calculations and the time spent on the computation of the prior IBD probabilities. In addition, when more

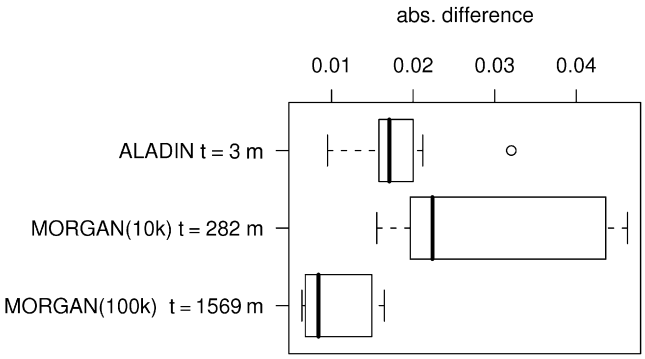

Figure 8. Evaluation of ALADIN with Real Data
The approximation error of ALADIN and MORGAN is compared on the real data set for subpedigree Ia for varying number of MCMC scans of MORGAN. The figure shows boxplots of the absolute difference between the approximate and exact NPL_{pairs} of all points on chromosomes 1–10. The number of MCMC scans is shown between parentheses, and the computation times are denoted by *t* on the horizontal axis.

segregation indicators are explicitly modeled and/or multiple IBD variables are used for subsections of the pedigree, more accurate approximations may be obtained. This is a direction for further research.

Although we found that the first-order Markov approximation was accurate for the pedigrees we considered, there are situations where the approximation is known to have problems. In particular for more distantly related individuals, there is an increasing tendency for segments of IBD sharing to cluster. If at three consecutive loci, A, B, and C, two individuals are non-IBD at the middle locus B, they are far more likely to be IBD at C if they are IBD at A than if they are non-IBD at A.[29] The reason for this is that a single recombination event between A and B is enough to break down the IBD sharing but that a recombination event in the same meiosis is sufficient to fully restore the sharing at locus C. The first-order Markov approximation underestimates the probability of IBD sharing in this situation.[29] Preliminary simulations suggest that the magnitude of this effect was small for the pedigrees studied in this paper. It may be beneficial to include an error model of some kind,[21] so that if a marker in the non-IBD segment absolutely precludes IBD, the approximation will not consider the two IBD segments as independent realizations of sharing. We plan on incorporating such an error model in a future release of ALADIN. As a diagnostic, one may sample from the posterior distribution of IBD configurations to identify these segments.

The problem of how to deal with linkage disequilibrium is an active area of research. Several LD models have been proposed to estimate haplotypes and haplotype frequencies for unrelated individuals.[30,31] In theory, it is straightforward to combine LD models with pedigree models: One can simply use one of the proposed LD models to model the prior distribution of haplotypes for the founder individuals in the pedigree.[16] The main issue then becomes how to deal with the significantly increased computational complexity of such a hybrid approach. Promising approaches using MCMC approximations have been recently proposed,[16] but these are still experimental.

We have chosen to use the clustered-markers approach of MERLIN to account for LD between the markers. This approach makes two simplifying assumptions to achieve high computational efficiency. First, it assumes absence of recombination between groups of markers that are in strong disequilibrium, clustering such groups into single "supermarkers" (the haplotype blocks). If the markers clustered together are very close, the impact of this limit on recombination is generally small. It may have a larger impact if markers spanning larger distances (~1 cM) are clustered.[32] In pedigrees with long lines of descent such as the ones considered here, there will be more recombinations within clusters than in the smaller pedigrees to which MERLIN is usually applied.

In our simulation of datasets, we allowed for recombination within clusters according to the marker map provided by Affymetrix. As a result, ALADIN and MERLIN found clusters where individuals' genotypes were inconsistent with the specified haplotypes in the cluster, that is, where recombination had occurred within clusters. We found that the number of inconsistencies in our simulations was on average 1.3 ± 1.6 per chromosome per pedigree. We expect that discarding of this small number of clusters will generally not result in much loss of information when high-density SNP arrays are used.

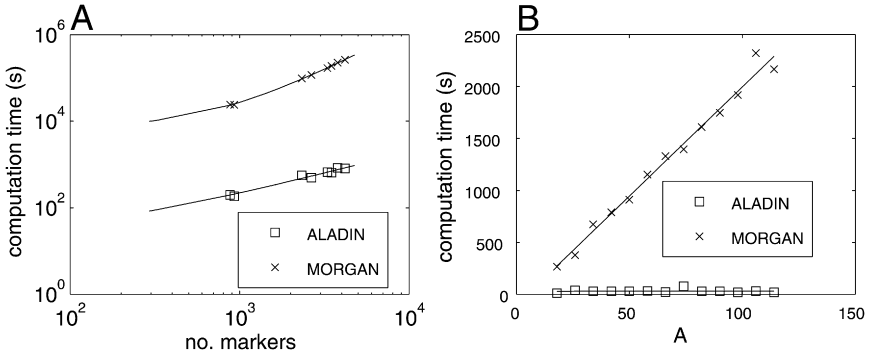

Figure 9. Scaling of Computation Time
(A) shows on a log-log scale the computation time as a function of the number of markers used in the multipoint analysis. Computation time of ALADIN scaled linearly with the number of markers, whereas that of MORGAN scaled quadratically with the number of markers. (B) shows computation time as a function of *A*, the number of untyped ancestors, for fixed number of individuals $T = 4$ and 100 markers.

Second, the clustered-markers approach of MERLIN assumes absence of LD between the haplotype blocks. Dealing with these lower levels of LD that remain after clustering without significantly reducing computational efficiency is very difficult and beyond the scope of this article. The limitations of the LD model should be kept in mind when practical data sets are analyzed.

Finally, maximum-likelihood haplotype reconstruction is straightforward in the framework we described. This option is not yet available in ALADIN but is planned for a future version.

Appendix A

Details of Single-Point Computations

The single-point likelihood of Equation 4 is given by

$$\begin{aligned}P\left(\mathbf{M},\mathbf{G_D},\mathbf{G_T},\mathbf{G_A},s_{\mathbf{D}_{\mathrm{nf}}},s_{\mathbf{T}},s_{\mathbf{A}_{\mathrm{nf}}}|\mathbf{f}\right)\\ =P(\mathbf{M_D}|\mathbf{G_D})P(\mathbf{G}_{\mathbf{D}_{\mathrm{nf}}}|s_{\mathbf{D}_{\mathrm{nf}}},\mathbf{G}_{\mathbf{D}_{\mathrm{f}}},\mathbf{G_T})P(\mathbf{G}_{\mathbf{D}_{\mathrm{f}}}|\mathbf{f})P(s_{\mathbf{D}_{\mathrm{nf}}})P(\mathbf{M_T}|\mathbf{G_T})\\ \times P(\mathbf{G_T}|s_{\mathbf{T}},\mathbf{G}_{\mathbf{A}_{\mathrm{nf}}},\mathbf{G}_{\mathbf{A}_{\mathrm{f}}})P(\mathbf{G}_{\mathbf{A}_{\mathrm{nf}}}|s_{\mathbf{A}_{\mathrm{nf}}},\mathbf{G}_{\mathbf{A}_{\mathrm{f}}})P(\mathbf{G}_{\mathbf{A}_{\mathrm{f}}}|\mathbf{f})P(s_{\mathbf{T}})P(s_{\mathbf{A}_{\mathrm{nf}}}),\end{aligned}$$

where the subscript f denotes a founder individual. The conditional independencies of this likelihood are represented by the graphical model[12,33] shown in Figure 2A. The likelihood term

$$\begin{aligned}P(\mathbf{G_T},\mathbf{G_A},s_{\mathbf{T}},s_{\mathbf{A}_{\mathrm{nf}}}|\mathbf{f})=P(\mathbf{G_T}|s_{\mathbf{T}},\mathbf{G}_{\mathbf{A}_{\mathrm{nf}}},\mathbf{G}_{\mathbf{A}_{\mathrm{f}}})\\ \times P(\mathbf{G}_{\mathbf{A}_{\mathrm{nf}}}|s_{\mathbf{A}_{\mathrm{nf}}},\mathbf{G}_{\mathbf{A}_{\mathrm{f}}})P(\mathbf{G}_{\mathbf{A}_{\mathrm{f}}}|\mathbf{f})P(s_{\mathbf{T}})P(s_{\mathbf{A}_{\mathrm{nf}}})\end{aligned}$$

is indicated by the rectangle in Figure 2. This term can be simplified by noting that the ordered genotypes $\mathbf{G}_{\mathbf{A}_{\mathrm{nf}}}$ are uniquely determined by $\mathbf{G}_{\mathbf{A}_{\mathrm{f}}}$ and $\mathbf{s}_{\mathbf{A}_{\mathrm{nf}}}$; the dependence of $\mathbf{G_T}$ on $(\mathbf{s_T},\mathbf{G}_{\mathbf{A}_{\mathrm{nf}}},\mathbf{G}_{\mathbf{A}_{\mathrm{f}}})$ can be written as a dependence on $(\mathbf{s_T},\mathbf{G}_{\mathbf{A}_{\mathrm{f}}})$. Consequently, $P(\mathbf{G_T},\mathbf{G_A},\mathbf{s_T},\mathbf{s}_{\mathbf{A}_{\mathrm{nf}}}|\mathbf{f})$ also satisfies the independencies of the graphical model shown in Figure 2B. Furthermore, the individuals in $\mathbf{A}$ have no genotype information. Thus we can sum over $\mathbf{G}_{\mathbf{A}_{\mathrm{nf}}}$, which removes this variable from Figure 2B, and obtain (see Figure 2C)

$$\begin{aligned}P(\mathbf{G_T},\Pi\,|\,\mathbf{f}) &= \sum_{(\mathbf{s_T},\mathbf{s}_{\mathbf{A}_{\mathrm{nf}}})\in\Pi}\sum_{\mathbf{G_A}} P(\mathbf{G_T},\mathbf{G_A},\mathbf{s_T},\mathbf{s}_{\mathbf{A}_{\mathrm{nf}}}|\,\mathbf{f})\\ &= \sum_{(\mathbf{s_T},\mathbf{s}_{\mathbf{A}_{\mathrm{nf}}})\in\Pi}\sum_{\mathbf{G}_{\mathbf{A}_{\mathrm{f}}}} P(\mathbf{G_T}|\,\mathbf{s_T},\mathbf{s}_{\mathbf{A}_{\mathrm{nf}}},\mathbf{G}_{\mathbf{A_f}})\\ &\quad\times P(\mathbf{G}_{\mathbf{A}_{\mathrm{f}}}|\,\mathbf{f})P(\mathbf{s_T})P(\mathbf{s}_{\mathbf{A}_{\mathrm{nf}}})\\ &= \sum_{(\mathbf{s_T},\mathbf{s}_{\mathbf{A}_{\mathrm{nf}}})\in\Pi} P(\mathbf{G_T}|\,\mathbf{s_T},\mathbf{s}_{\mathbf{A}_{\mathrm{nf}}},\mathbf{f})P(\mathbf{s_T})P(\mathbf{s}_{\mathbf{A}_{\mathrm{nf}}}).\end{aligned}\tag{16}$$

We now observe that the conditional probability distribution of $\mathbf{G_T}$ is independent of the particular configuration $(\mathbf{s_T},\mathbf{s}_{\mathbf{A}_{\mathrm{nf}}})\in\Pi$, i.e.,

$$P(\mathbf{G_T}\,|\,\mathbf{s_T},\mathbf{s}_{\mathbf{A}_{\mathrm{nf}}},\mathbf{f}) = P(\mathbf{G_T}|\,\Pi,\mathbf{f}),\ \forall\ (\mathbf{s_T},\mathbf{s}_{\mathbf{A}_{\mathrm{nf}}})\in\Pi.\tag{17}$$

This can be understood as follows. By the definition of Π, all configurations $(\mathbf{s_T},\mathbf{s}_{\mathbf{A}_{\mathrm{nf}}})\in\Pi$ imply the same IBD configuration of the alleles contained in $\mathbf{G_T}$. If a subset of alleles of individuals in $\mathbf{T}$ is IBD given Π, they are for any configuration $(\mathbf{s_T},\mathbf{s}_{\mathbf{A}_{\mathrm{nf}}}\Pi^{l})\in\Pi$ a copy of the same founder allele, although which allele of which founder does depend on $(\mathbf{s_T},\mathbf{s}_{\mathbf{A}_{\mathrm{nf}}})$. Because the general assumption is that the prior allele frequency distribution is the same for every founder, the independence follows.

We obtain Equation 5 from Equation 16 by using the independence in Equation 17:

$$\begin{aligned}P(\mathbf{G_T},\Pi\,|\,\mathbf{f}) &= P(\mathbf{G_T}\,|\,\Pi,\mathbf{f})\sum_{(\mathbf{s_T},\mathbf{s}_{\mathbf{A}_{\mathrm{nf}}})\in\Pi} P(\mathbf{s}_{\mathbf{A}_{\mathrm{nf}}})P(\mathbf{s_T})\\ &= P(\mathbf{G_T}\,|\,\Pi,\mathbf{f})P(\Pi),\end{aligned}\tag{18}$$

where we used Equation 3 to obtain the right equality. The graphical model corresponding to this marginal likelihood is shown in Figure 2D.

Computation of Prior IBD Probabilities with Variable Elimination

Gene-Dropping and Exhaustive Enumeration. First, we describe a naive procedure of calculating $P(\Pi)$ that consists of dropping founder alleles and exhaustively enumerating all configurations of segregation indicators. For this, we consider a single locus and assign to each founder allele a unique identifier allele:

$$\begin{aligned}G^{p}_{\mathrm{f}_1}=1, G^{m}_{\mathrm{f}_1}=2, G^{p}_{f_2}=3, G^{m}_{f_2}=4,\ldots, G^{p}_{\mathrm{f}_{|\mathbf{A}_{\mathrm{f}}|}}\\ =2\,|\,\mathbf{A}_{\mathrm{f}}\,|-1, G^{m}_{f_{|\mathbf{A}_{\mathrm{f}}|}}=2\,|\,\mathbf{A}_{\mathrm{f}}\,|,\end{aligned}$$

where f_i represents the i^{th} founder in the group of untyped founder ancestors $\mathbf{A}_{\mathrm{f}}$, $G^{p}_{f_i}$ is the associated paternal allele, and $G^{m}_{f_i}$ is the maternal allele. Recall that $\mathbf{A}$ contains the individuals through which individuals in $\mathbf{T}$ are related and that $\mathbf{T}$ does not contain founders. This procedure defines a marker with $2|\mathbf{A}_{\mathrm{f}}|$ possible alleles, where each allele observed in a nonfounder individual can be traced back to one of the founder alleles, because identity by state implies identity by descent. As a result, given allele values, it is not necessary to know the individual values of the segregation indicators in order to determine IBD status.

Probabilities of IBD configurations of the alleles in $\mathbf{G_T}$ can then be determined by the exhaustive enumeration all possible configurations of segregation indicators and checking for each configuration which alleles in $\mathbf{G_T}$ have the same value and which have different values. Because a priori all meioses are independent (for a single locus) and maternal and paternal inheritance are equally probable, $P(\Pi)$ is given by the number of configurations of segregation indicators $(\mathbf{s_T},\mathbf{s}_{\mathbf{A}_{\mathrm{nf}}})$ with a given IBD configuration Π divided by the total number of possible configurations.

Formal Definition. We shall now formalize this approach of gene dropping. We denote the probability that allele G_i^x and G_j^y, with $i, j \in \mathbf{T}$ and $x, y \in \{p, m\}$, are identical by descent as $P(\mathrm{IBD}(G_i^x, G_j^y))$. In the methods section, it was shown that these probabilities can be obtained from the likelihood term $P(\mathbf{G_T},\mathbf{G}_{\mathbf{A}_{\mathrm{nf}}},\mathbf{s_T},\mathbf{s}_{\mathbf{A}_{\mathrm{nf}}}|\mathbf{G}_{\mathbf{A}_{\mathrm{f}}},\mathbf{f})$ indicated by the dashed rectangle in Figure 2:

$$
\begin{aligned}
P\left(\mathrm{IBD}\left(G_i^x,G_j^y\right)\right) &= \sum_{\mathbf{G_T}} \delta\left(G_i^x,G_j^y\right) P(\mathbf{G_T}|\mathbf{G}_{\mathbf{A}_\mathrm{f}}=\mathbf{I}) \\
&= \sum_{\mathbf{G_T},\mathbf{G}_{\mathbf{A}_\mathrm{nf}}} \sum_{\mathbf{s_T},\mathbf{s}_{\mathbf{A}_\mathrm{nf}}} \delta\left(G_i^x,G_j^y\right) \\
&\quad \times P(\mathbf{G_T},\mathbf{G}_{\mathbf{A}_\mathrm{nf}},\mathbf{s_T},\mathbf{s}_{\mathbf{A}_\mathrm{nf}}|\mathbf{G}_{\mathbf{A}_\mathrm{f}}=\mathbf{I}) \\
&\propto \sum_{\mathbf{G_T},\mathbf{G}_{\mathbf{A}_\mathrm{nf}}} \sum_{\mathbf{s_T},\mathbf{s}_{\mathbf{A}_\mathrm{nf}}} \delta\left(G_i^x,G_j^y\right) \\
&\quad \times P(\mathbf{G_T},\mathbf{G}_{\mathbf{A}_\mathrm{nf}}|\mathbf{s_T},\mathbf{s}_{\mathbf{A}_\mathrm{nf}},\mathbf{G}_{\mathrm{A_f}}) \\
&\quad \times P(\mathbf{s_T},\mathbf{s}_{\mathbf{A}_\mathrm{nf}}) \prod_{k\in\mathbf{A}_\mathrm{f}} \prod_{z=\{p,m\}} \delta\left(G_k^z,I_{kz}\right),
\end{aligned}
\tag{19}
$$

where I_{kz} is an indicator function that assigns to each founder allele G_k^z a unique value as illustrated in the example above. The kronecker δ function $\delta\ (G_i^x,G_j^y)$ is equal to one if $G_i^x = G_j^y$, i.e., if G_i^x and G_j^y are IBD, and zero if $(G_i^x \neq G_j^y)$. The proportionality constant is given by $1/P(\mathbf{G}_{\mathbf{A}_\mathrm{f}} = \mathbf{I}|\mathbf{f})$. Thus, given the unique assignment of the alleles of the founders in $\mathbf{A}$, the sum over all configurations of ordered genotypes and segregation indicators yields the desired IBD probability.

Probabilities of IBD configurations of more than two alleles can be obtained by the replacement of $\delta(G_i^x,G_j^y)$ in Equation 19 with the appropriate product of δ functions. For instance, the IBD configuration denoted by the partitioning $(G_1^p,G_2^p)(G_3^p)$ (see the example in the Material and Methods), follows from the expression

$$
\delta\left(G_1^p,G_2^p\right)\left(1-\delta\left(G_1^p,G_3^p\right)\right).
$$

Here, G_1^p and G_2^p must have the same value (thus IBD given $\mathbf{G}_{\mathbf{A}_\mathrm{f}} = \mathbf{I}$), and G_3^p is required to have a different value than G_1^p (and hence G_2^p), yielding the desired IBD indicator function.

We compute the conditional probabilities $P(\Pi^{l+1}|\Pi^l,\theta)$ from the joint distribution $P(\Pi^{l+1}|\Pi^l,\theta^{(l+1,l)})$ in a two-locus model by using Equation 14. These joint distributions can be obtained by the generalization of Equation 19 to a two-locus model and the use of products of δ functions to define the IBD configurations of the two loci corresponding to the variables Π^{l+1} and Π^l. Note that these probabilities will depend on the recombination fractions θ.

Variable Elimination. In Equation 19 we have formulated the problem of calculating (multilocus) IBD probabilities as likelihood computations. Explicit evaluation of the sum is exponential in the number of variables and becomes infeasible as $\mathbf{A}$ grows large. However, given this formulation, it is straightforward to perform the summation more efficiently with the variable-elimination algorithm.[23,24] This technique has been successfully applied to the problem of genetic linkage analysis: It is one of the core operations of the program SUPERLINK,[24,34] and the Blocking Gibbs sampler for linkage analysis[12,25] uses related techniques to perform exact likelihood computations in pedigrees. We refer to their papers for detailed explanation of the methodology.

Here, we mention the most important details of how we tailored the variable-elimination algorithm to the problem of the computation of a priori IBD probabilities. For the calculation of $P(\Pi)$, we need to compute joint probabilities of the possible assignments of variables $\mathbf{G_T}$, where each variable $G_i^x \in \mathbf{G_T}$ may take $2|\mathbf{A}_\mathrm{f}|$ possible values. This makes the computation more complex as compared to the calculation of the likelihood of the observations $P(\mathbf{M}|\mathbf{f})$, but it can still be solved with the same variable-elimination technique. The elimination order is determined as described by Fishelson et al.;[34] however, the variables of interest $\mathbf{G_T}$ are not eliminated, but retained throughout the elimination procedure. This is the standard procedure for the calculation of joint-probability distributions in Bayesian networks.[35,36] In addition, we apply value abstraction[24,34,37] to reduce the number of configurations that have to be stored in memory; specifically, we do not consider all genotype configurations $\mathbf{G}_c$, $\mathbf{s}_c$ for a subset of variables c individually but cluster them with respect to their IBD configuration. This does not affect the joint probabilities but significantly improves efficiency.

Estimation of Multipoint Parametric LOD Scores

The nonparametric linkage scores $\mathrm{NPL}_{\mathrm{pairs}}$ and $\mathrm{NPL}_{\mathrm{all}}$ can be readily obtained from the marginal posterior distributions $P(\mathbf{s}^l_{\mathbf{D}_\mathrm{nf}},\Pi^l|\mathbf{M},\mathbf{f},\theta)$.[5] Given these posterior distributions, it is also possible to compute LOD scores under the assumption of a specific disease model.

We denote the vector of affection statuses for the individuals by $\mathbf{Z}$, where Z_i = {affected, unaffected, unkown} for individual i. Recall that by definition the affection status of the individuals in $\mathbf{A}$ is unknown; the individuals in $\mathbf{T}$ and $\mathbf{D}$ can have an affection status that is unkown, affected or not affected. The LOD score is then given by the ratio of the likelihood of the hypothesis that the disease locus is linked to the markers at location λ and the a priori likelihood of observing $\mathbf{Z}$:

$$
\mathrm{LOD}(\lambda) = \log_{10}\frac{P(\mathbf{Z}\,|\,\mathbf{M},\mathbf{f},\theta,\mathbf{d},\mathbf{p},\lambda)}{P(\mathbf{Z}\,|\,\mathbf{d},\mathbf{p})}. \tag{20}
$$

Here, $\mathbf{p}$ is the vector of penetrance values and $\mathbf{d}$ is the vector of allele frequencies of the disease locus. The likelihood term in the denominator of Equation 20 does not depend on the marker data and requires a single-locus computation that can be performed exactly. In the setting we consider, the numerator, however, is intractable to compute. Without loss of generality, assume that we wish to compute this likelihood for the location of marker l. In the approximate HMM (Equation 15) the numerator is approximated by

$$
\begin{aligned}
&P(\mathbf{Z}\,|\,\mathbf{M},\mathbf{f},\theta,\mathbf{d},\mathbf{p},\lambda=l) \\
&\approx \sum_{\mathbf{s}^l_\mathbf{D},\Pi^l} P\left(\mathbf{Z}\,|\,\mathbf{s}^l_{\mathbf{D}_\mathrm{nf}},\Pi^l,\mathbf{d},\mathbf{p}\right) P\left(\mathbf{s}^l_{\mathbf{D}_\mathrm{nf}},\Pi^l\,|\,\mathbf{M},\mathbf{f},\theta\right).
\end{aligned}
$$

Thus, after the marginal distributions $P(\mathbf{s}^l_{\mathbf{D}_\mathrm{nf}},\Pi^l|\mathbf{M},\mathbf{f},\theta)$ have been obtained with the Lander-Green algorithm, it is straightforward to estimate parametric LOD scores from these with single-locus computations.

Acknowledgments

We thank two anonymous reviewers for their valuable comments that helped improve the manuscript. C.A. is supported by the Dutch Technology Foundation (STW). M.B. is supported by a National Health and Medical Research Council of Australia R.D. Wright Fellowship.

Received: August 15, 2007
Revised: October 22, 2007
Accepted: December 11, 2007
Published online: March 6, 2008

Web Resources

The URLs for data presented herein are as follows:

ALADIN, http://www.mbfys.ru.nl/~keesa/aladin/
Online Mendelian Inheritance in Man (OMIM), http://www.ncbi.nlm.nih.gov/Omim/

References

1. Risch, N. (2001). Implications of multilocus inheritance for gene-disease association studies. Theor. Popul. Biol. *60*, 215–220.
2. Vierimaa, O., Georgitsi, M., Lehtonen, R., Vahteristo, P., Kokko, A., Raitila, A., Tuppurainen, K., Ebeling, T.M.L., Salmela, P.I., Paschke, R., et al. (2006). Pituitary adenoma predisposition caused by germline mutations in the AIP gene. Science *312*, 1228–1230.
3. Sobel, E., and Lange, K. (1993). Metropolis sampling in pedigree analysis. Stat. Methods Med. Res. *2*, 263–282.
4. Lander, E., and Green, P. (1987). Construction of multilocus genetic maps in humans. Proc. Natl. Acad. Sci. USA *84*, 2363–2367.
5. Kruglyak, L., Daly, M.J., Reeve-Daly, M.P., and Lander, E.S. (1996). Parametric and non-parametric linkage analysis: A unified multipoints approach. Am. J. Hum. Genet. *58*, 1347–1363.
6. Gudbjartsson, D.F., Jonasson, K., Frigge, M.L., and Kong, A. (2000). Allegro, a new computer program for multipoint linkage analysis. Nat. Genet. *25*, 12–13.
7. Abecasis, G.R., and Wigginton, J.E. (2005). Handling marker-marker linkage disequilibrium: Pedigree analysis with clustered markers. Am. J. Hum. Genet. *77*, 754–767.
8. Lange, K., and Sobel, E. (1996). Descent graphs in pedigree analysis: Applications to haplotyping, location scores, and marker-sharing statistics. Am. J. Hum. Genet. *58*, 1323–1337.
9. Thompson, E.A. (1994). Monte Carlo likelihood in genetic mapping. Stat. Sci. *9*, 355–366.
10. Thompson, E.A., and Heath, S.C. (1999). Estimation of conditional multilocus gene identity among relatives. In Statistics in Molecular Biology and Genetics. IMS Lecture Notes-Monograph Series, *Volume 33*, F. Seillier-Moiseiwitsch, ed. (Hayward, California: Institute of Mathematical Studies), pp. 95–113.
11. George, A.W., and Thompson, E.A. (2003). Discovering disease genes: Multipoint linkage analyses via a new Markov Chain Monte Carlo approach. Stat. Sci. *18*, 515–535.
12. Thomas, A., Abkevich, V., and Bansai, A. (2000). Multilocus linkage analysis by blocked Gibbs sampling. Stat. Comput. *10*, 259–269.
13. Baird, P.N., Foote, S.J., Mackey, D.A., Craig, J., Speed, T.P., and Bureau, A. (2005). Evidence for a novel glaucoma locus at chromosome 3p21–22. Hum. Genet. *117*, 249–257.
14. Wijsman, E., Rothstein, J.H., and Thompson, E.A. (2006). Multipoint linkage analysis with many multiallelic or dense diallelic markers: Markov chain-Monte Carlo provides practical approaches for genome scans on general pedigrees. Am. J. Hum. Genet. *79*, 846–858.
15. Thomas, A., and Camp, N.J. (2004). Graphical modeling of the joint distribution of alleles at associated loci. Am. J. Hum. Genet. *74*, 1088–1101.
16. Thomas, A. (2007). Towards linkage analysis with markers in linkage disequilibrium by graphical modelling. Hum. Hered. *64*, 16–26.
17. Donnelly, K.P. (1983). The probability that related individuals share some section of genome identical by descent. Theor. Popul. Biol. *23*, 34–63.
18. Feingold, E. (1993). Markov processes for modeling and analyzing a new genetic mapping method. J. Appl. Probab. *30*, 766–779.
19. McPeek, M., and Sun, L. (2000). Statistical tests for detection of misspecified relationships by use of genome-screen data. Am. J. Hum. Genet. *66*, 1076–1094.
20. Abney, M., Ober, C., and McPeek, M.S. (2002). Quantitative-trait homozygosity and association mapping and empirical genome-wide significance in large, complex pedigrees: fasting serum-insulin level in the Hutterites. Am. J. Hum. Genet. *70*, 920–934.
21. Leutenegger, A.L., Prum, B., Genin, E., Verny, C., Lemainque, A., Clerget-Darpoux, F., and Thompson, E.A. (2003). Estimation of the inbreeding coefficient through use of genomic data. Am. J. Hum. Genet. *73*, 516–523.
22. Leutenegger, A.L., Labalme, A., Genin, E., Toutain, A., Steichen, E., Clerget-Darpoux, F., and Edery, P. (2006). Using genomic inbreeding coefficient estimates for homozygosity mapping of rare recessive traits: Application to Taybi-Linder syndrome. Am. J. Hum. Genet. *79*, 62–66.
23. Dechter, R. (1996) Bucket elimination: A unifying framework for probabilistic inference. In: Proc. 12th Conference on Uncertainty in Artificial Intelligence. Portland, OR, USA, pp. 211–219.
24. Fishelson, J., and Geiger, D. (2002). Exact genetic linkage computations for general pedigrees. Bioinformatics *18*, S189–S198.
25. Jensen, C.S., and Kong, A. (1999). Blocking-Gibbs sampling for linkage analysis in large pedigrees with many loops. Am. J. Hum. Genet. *65*, 885–902.
26. Barrett, J.C., Fry, B., Maller, J., and Daly, M.J. (2005). Haploview: Analysis and visualization of LD and haplotype maps. Bioinformatics *21*, 263–265.
27. The International HapMap Consortium (2005). A haplotype map of the human genome. Nature *437*, 1299–1320.
28. Thomson, R., Quinn, S., McKay, J., Silver, J., Bahlo, M., FitzGerald, L., Foote, S., Dickinson, J., and Stankovich, J. (2007). The advantages of dense marker sets for linkage analysis with very large families. Hum. Genet. *121*, 459–468.
29. Thompson, E.A. (2000). Statistical Inferences from Genetic Data on Pedigrees. NSF-CBMS Regional Conference Series in Probability and Statistics, Volume 6 (Beachwood, OH: IMS).
30. Scheet, P., and Stephens, M. (2006). A fast and flexible statistical model for large-scale population genotype data: Applications to inferring missing genotypes and haplotypic phase. Am. J. Hum. Genet. *78*, 629–644.

31. Browning, S.R., and Browning, B.L. (2007). Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. Am. J. Hum. Genet. *81*, 1084–1097.
32. Albers, C.A., and Kappen, H.J. (2007). Modeling linkage disequilibrium in exact linkage computations: A comparison of first-order Markov approaches and the clustered-markers approach. BMC Proceedings *1* (*Suppl 1*), S159.
33. Pearl, J. (1988). Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference (San Mateo, CA: Morgan Kaufmann Publishers).
34. Fishelson, J., Dovgolevsky, N., and Geiger, D. (2005). Maximum-likelihood haplotyping for general pedigrees. Hum. Hered. *59*, 41–60.
35. Jensen, F.V. (1996). An Introduction to Bayesian Networks (London: UCL Press).
36. Lauritzen, S.L., and Sheehan, N.A. (2003). Graphical models for genetic analyses. Stat. Sci. *4*, 489–514.
37. Friedman, N., Geiger, D., and Lotner, N. (2000) Likelihood computation with value abstraction. In: Proc. 16th Conference on Uncertainty in Artificial Intelligence. Stanford, CA, USA, pp. 192–200.