BIOINFORMATICS

Vol. 26 ISMB 2010, pages i175–i182
doi:10.1093/bioinformatics/btq204

Estimating genome-wide IBD sharing from SNP data via an efficient hidden Markov model of LD with application to gene mapping

Sivan Bercovici[1,*], Christopher Meek[2], Ydo Wexler[2] and Dan Geiger[1,2]

[1]Technion-Israel Institute of Technology, Computer Science Department Haifa, 32000 Israel and [2]Microsoft Research, Redmond, WA 98052, USA

ABSTRACT

Motivation: Association analysis is the method of choice for studying complex multifactorial diseases. The premise of this method is that affected persons contain some common genomic regions with similar SNP alleles and such areas will be found in this analysis. An important disadvantage of GWA studies is that it does not distinguish between genomic areas that are inherited from a common ancestor [identical by descent (IBD)] and areas that are identical merely by state [identical by state (IBS)]. Clearly, areas that can be marked with higher probability as IBD and have the same correlation with the disease status of identical areas that are more probably only IBS, are better candidates to be causative, and yet this distinction is not encoded in standard association analysis.

Results: We develop a factorial hidden Markov model-based algorithm for computing genome-wide IBD sharing. The algorithm accepts as input SNP data of measured individuals and estimates the probability of IBD at each locus for every pair of individuals. For two g-degree relatives, when $g \geq 8$, the computation yields a precision of IBD tagging of over 50% higher than previous methods for 95% recall. Our algorithm uses a first-order Markovian model for the linkage disequilibrium process and employs a reduction of the state space of the inheritance vector from being exponential in g to quadratic. The higher accuracy along with the reduced time complexity marks our method as a feasible means for IBD mapping in practical scenarios.

Availability: A software implementation, called IBDMAP, is freely available at http://bioinfo.cs.technion.ac.il/IBDmap.

Contact: sberco@gmail.com

1 INTRODUCTION

Association analysis is the method of choice for studying complex multifactorial diseases with low penetrance. The basic idea is to compute a correlation between each measured SNP and the affection status, and then further study areas that contain SNPs with high scores. The premise of this method is that affected persons contain some common genomic regions with similar SNP alleles and such areas will be found in this analysis. The large number of tests ($\geq$500000) requires correction for multiple testing (Benjamini and Yekutieli, 2005; Han *et al.*, 2009; Peer *et al.*, 2008). To capture the fact that dense SNPS are not independent, one can use various blocks of linked SNPs and find correlation to the blocks rather than to individual SNPs (Cardon and Abecasis, 2003; Greenspan and Geiger, 2004; Halperin *et al.*, 2005). The main advantage of association analysis is that it can potentially identify small suspect areas provided that sufficiently many individuals are sampled.

An important disadvantage of this method is that it does not distinguish between genomic areas that are inherited from a common ancestor [identical by descent (IBD)] and areas that are identical merely by state [identical by state (IBS)]. Clearly, areas that can be marked with higher probability as IBD and have the same correlation with the disease status of identical areas that are more probably only IBS, are better candidates to be causative, and yet this distinction is not encoded in standard association analysis.

The popular PLINK program for association analysis, among its many functions, provides a method that challenges this common practice (Purcell *et al.*, 2007). PLINK accepts SNP data of affected and healthy individuals, then infers the IBD areas for each pair and consequently uses a score (Equation 6) to evaluate the extent to which a SNP is suspected to predispose or cause a disease. The input of this formula relies on inferring the IBD status of the SNP for each pair of input individuals. The inference is done via an hidden Markov model (HMM) with a small hidden state space that encodes the IBD status at the current locus using three states. The power of this mapping method highly depends on the precision of the IBD inference.

Our work herein replaces the part of inferring IBD with an improved method using a more accurate HMM model with a state space that grows quadratically with the number of generations g that distinguish two individuals. We develop a factorial HMM-based algorithm for computing genome-wide IBD sharing. The algorithm accepts as input SNP data of measured individuals and estimates the probability of IBD at each locus for every pair of individuals. To estimate performance, we measure precision, which is the number of correctly identified IBD positions divided by the total number of position inferred as IBD, and recall, which is the number of correctly identified IBD positions divided by the number of IBD positions. For two g-degree relatives, when $g \geq 8$, the computation yields a precision of IBD tagging of over 50% higher than previous methods for 95% recall.

Our algorithm uses a reduction of the state space of the inheritance vector used in MERLIN from being exponential in g to quadratic and combines it with a first-order Markovian model for the linkage disequilibrium (LD) process. In essence, we apply sophisticated techniques from linkage analysis to genetic studies without pedigree input. The application of our method to gene mapping shows a noticeable improvement.

The rest of this article is organized as follows. Section 2 defines HMMs and explains the inferences done with them. Section 3

*To whom correspondence should be addressed.

provides background information regarding genetic analysis. Section 4 develops a model for LD and an appropriate inference procedure for it. Section 5 explicates the results for IBD inference and for gene mapping. Finally, Section 6 discusses limitations of current IBD inference procedures and future directions.

2 HMM AND FACTORIAL HMM

Consider a HMM (Rabiner and Juang, 1986) with hidden variables S_i and observed variables X_i, $i=1,\ldots,L$, as described in Geiger *et al.*, 2008. The (*hidden*) *state space* is the set S of possible values for S_i. The state space is identical for every slot i. The likelihood of data $(x_1,\ldots,x_L)$ for L slots is specified via two main components. The single slot likelihood of data $P(x_i|s_i)$ at slot i given a state s_i for S_i and the transition probabilities $P(S_i=s_i|S_{i-1}=s_{i-1})$ from a state at slot $i-1$ to slot i.

$$\begin{aligned} P(data)=&\sum_{s_1}P(s_1)P(x_1|S_1=s_1)\\ &\sum_{s_2}P(S_2=s_2|S_1=s_1)P(x_2|S_2=s_2)\cdots\\ &\sum_{s_L}P(S_L=s_L|S_{L-1}=s_{L-1})P(x_L|S_L=s_L). \end{aligned} \tag{1}$$

The time complexity of computing this sum grows quadratically with the size of the state space $|S|$ and linearly in the number of slots L. The time complexity is $O(L|S|^2+cL|S|)$ where c is an upper bound for computing the single slot likelihood $P(x_i|s_i)$. We note that in many HMM applications, including IBD sharing and linkage analysis, the goal is to also compute the marginal probabilities $P(S_i|x_1,\ldots,x_L)$ for all $i=1,\ldots,L$ rather than to compute just the likelihood of data. This task can be completed using the junction-tree algorithm with only twice the computational cost (Lauritzen and Spiegelhalter, 1988).

Factorial HMMs (Ghahramani and Jordan, 1997) are HMMs in which the hidden variable is a vector $S_i=(S_i^1,\ldots,S_i^k)$ with values drawn from some Cartesian product $H_1\times\ldots\times H_k$ and with a transition probability defined component by component for $i=2,\ldots,L$ via

$$\begin{aligned} &P(S_i=(s_i^1,\ldots,s_i^k)|S_{i-1}=(s_{i-1}^1,\ldots,s_{i-1}^k))\\ &\quad=\prod_{j=1}^{k}P_j(s_i^j|s_{i-1}^j) \end{aligned} \tag{2}$$

and for the first slot, $P(S_1=(s_1^1,\ldots,s_1^k))=\prod_{j=1}^{k}P_j(s_1^j)$. In our application, each S_i^j is called a selector, and the vector S_i determines the inheritance pattern in the pedigree. We model the recombination probabilities as $P(S_i^j|S_{i-1}^j\neq S_i^j)=1-e^{-l}$, where l is the genetic distance, in Morgans, between location i and $i-1$.

Factorial HMMs offer computational benefits when computing the likelihood of data. Ghahramani and Jordan (1997) Section 3.2 show how specifying the probabilities $P(S_i|S_{i-1})$ via a product as in Equation 2 reduces the time complexity to $O(L|S|\log|S|+cL|S|)$. Their algorithm is a special case of bucket elimination (Dechter, 1998). We note that computing the L marginal probabilities $P(S_i|x_1,\ldots,x_L)$ in a factorial HMM can also be performed with only twice the amount of computations using the junction-tree algorithm (Lauritzen and Spiegelhalter, 1988).

In applications, where S is possibly very large such as for linkage analysis where it grows exponentially in, roughly, the number of persons in the pedigree, the dominating factor $|S|\log|S|$ can be further reduced substantially if the state space S can be partitioned into equivalence classes $[s]$ for which the likelihood of data is constant. This effectively changes the sum over the state space at each slot to a sum over equivalence classes. The dominating complexity will now depend on the number of equivalence classes rather than on the number of states in S (Browning and Browning, 2002; Geiger *et al.*, 2008; Markianos *et al.*, 2001).

The likelihood is computed for one representative of each equivalence class via

$$\begin{aligned} P(data)=&\sum_{[s_1]}P([s_1])P(x_1|S_1=[s_1])\\ &\sum\nolimits_{[s_2]}P(S_2=[s_2]|S_1=[s_1])P(x_2|S_2=[s_2])\cdots\\ &\sum\nolimits_{[s_L]}P(S_L=[s_L]|S_{L-1}=[s_{L-1}])P(x_L|S_L=[s_L]) \end{aligned} \tag{3}$$

where the prior for a class $[s]$ is the sum over the priors of its constituent states, namely, $P([s])=\sum_{s\in[s]}P(s)$. Note that $[s_i]$ is used to denote the class containing s_i, as in $s_i\in[s_i]$, and also a representative from the class containing state s_i, as in $S_i=[s_i]$.

The equivalence of Equations 1 and 3 stems from two general conditions:

CONDITION I. The single slot likelihood given a hidden state s is equal for all states in the equivalence class $[s]$, namely, $P(x_i|s)=P(x_i|s')$ for all s and s' in the same equivalence class. Hence, we can safely define the single slot likelihood given an equivalence class via $P(x_i|[s])=P(x|s)$.

CONDITION II. Denote by $P([s]|s')=\sum_{s\in[s]}P(s|s')$ the transition probability from state s' to an equivalence class $[s]$. The condition is that this transition probability does not distinguish between two states in the same equivalence class, namely, $P([s]|s')=P([s]|s'')$ for all s' and s'' in the same equivalence class. Hence, we can safely define the transition probabilities between equivalence classes via $P([s]|[s'])=P([s]|s')$.

These two natural conditions are sufficient to ensure that Equations 1 and 3 are equivalent (Geiger *et al.*, 2008).

3 GENETIC ANALYSIS

Genetic linkage and association analysis seek to locate genomic regions that are likely to contain genes that increase the probability of inheritable traits such as hereditary diseases. In the case of linkage analysis, the input are pedigrees of families that segregate a disease, genetic marker information such as SNP data and affection status of some or all family members. The main idea is that genetic markers that are found in the same vicinity on the chromosome are more likely to stay together during meiosis. Thus, based on the topology of the pedigree and the marker readings, it is possible to compute how likely it is for a predisposing gene to be located on the chromosome near each of the markers (Elston and Stewart, 1971; Lander and Green, 1987; Lange, 1997; Ott, 1999). Genetic association analysis shares the same goal of linkage analysis but uses a very different approach. The input to the basic design of association analysis is genetic marker data of affected individuals and of matched healthy controls. The output is a correlation, or more generally some score,

relating marker data at specific genomic locations with the trait under study (Carlson *et al.*, 2004; Halperin and Stephan, 2009; Peer *et al.*, 2006; Wang *et al.*, 2005).

There are pros and cons to these methods and both are used in countless number of mapping projects ranging from single gene rare Mendelian diseases to common complex diseases that are caused by an array of genetic, behavioral, environmental and other factors. Genetic linkage analysis is very successful in underpinning Mendelian diseases with high penetrance; there are hundreds of successfully verified discoveries using linkage. For such studies, a sparse array of genetic markers is needed to reach a valid conclusion–around 6000 SNP genome-wide at an average distance of 0.5 M base pairs. With such distance, the alleles of the founders of the pedigree can be considered to be independent, that is in *linkage equilibrium*, and this assumption is used in the standard model of linkage analysis. The strength of linkage analysis is the identification of stretches of SNPs that are more often passed from affected persons to affected offspring than to non-affected offspring. Such stretches have high likelihood of odds (LOD) to contain a predisposing or causative gene.

There are several scoring methods commonly used for linkage analysis. They differ in how the scoring function depends on the probability of the possible inheritance patterns in the pedigree. Examples of such functions are $S_{\text{all}}, S_{\text{pairs}}$ and LOD scores (Kruglyak *et al.*, 1996). As the number of such inheritance patterns grows exponentially in the number of markers and roughly in the number of persons in the pedigree, computationally sophisticated methods were proposed for this task. A common structure shared by most exact scoring methods is a HMM (Rabiner and Juang, 1986) backbone, which is in fact a Factorial HMM with a state space defined by a set of variables S_i called selectors that determine the inheritance pattern in the pedigree (Abecasis *et al.*, 2002; Gudbjartsson *et al.*, 2005, 2000; Ingolfsdottir and Gudbjartsson, 2005; Kruglyak and Lander, 1998; Kruglyak *et al.*, 1995, 1996; Lander and Green, 1987; Markianos *et al.*, 2001). Other techniques based, sometimes implicitly, on Bayesian networks (Lauritzen, 1996; Pearl, 1988) focus on larger pedigrees with fewer measurements (Cottingham *et al.*, 1993; Elston and Stewart, 1971; Fishelson and Geiger, 2002; O'Connell and Weeks, 1995; Silberstein *et al.*, 2006; Sobel and Lange, 1996; Thompson, 1994).

The main computer programs that analyze pedigrees with SNP data are GENEHUNTER, ALLEGRO and MERLIN (Abecasis *et al.*, 2002; Gudbjartsson *et al.*, 2000; Markianos *et al.*, 2001). Due to the increasing number of markers on standard SNP panels, these programs basically use the same HMM model employing the forward–backward algorithm as developed by Lander and Green (1987) in their seminal work in this area. The differences lie in the details of how transition matrices are represented and multiplied, and how the emission probabilities are computed, both of which can dramatically affect the time complexity of the forward–backward algorithm.

In several works (Kruglyak and Lander, 1998; Kruglyak *et al.*, 1995, 1996) with increasingly better methods for multiplying a vector by the specialized form of the transmission matrix $P(S_i|S_{i-1})$ used in this application, of size $N \times N$, complexity of multiplication dropped to $N \log N$ and the effective size of N dropped to N' by considering symmetries in the transition matrix. For certain pedigrees, this drop is exponential in the number of persons decreasing N from 2^{2n} to 2^{2n-f} and for some pedigrees to 2^{2n-f-c} where n is the number of non-founders in the pedigree, f is the number of founders and c is the number of first grandchildren whose grandparents are all founders in the pedigree (Markianos *et al.*, 2001). Further exponential reductions of the state space for special pedigrees such as pedigrees that contain long chains are also available (Browning and Browning, 2002; Geiger *et al.*, 2008).

Other works (Abecasis *et al.*, 2002; Gudbjartsson *et al.*, 2005, 2000) improved the representation and computation of the emission probabilities, namely, the probability of data at slot i given the state at slot i. There are two factors that make this computation demanding. First, to compute the probability $P(x_i|s_i)$ for a specific state s_i requires to sum over all possible assignments of alleles f_i to the founders in the pedigree, namely,

$$P(x_i|s_i) = \sum_{f_i} P(x_i|s_i, f_i) P(f_i). \tag{4}$$

A naive approach will be exponential in the number of founders in the pedigree that are not genotyped, but Sobel and Lange (1996) managed to use the conditional independence fact that once a state s_i is given, only the alleles of a few founders determines $P(x_i|s_i, f_i)$. Using what they termed descend trees, they compute $P(x_i|s_i)$ in polynomial time for every single state s_i. Still, to compute this likelihood for every s_i requires repeating this operation $2^{|S|}$ times, which often grows too large. The modern software packages MERLIN and ALLEGRO both decrease the magnitude of this problem by reusing computations from one state to another and the fact that after a partial vector of s_i is known, the remaining component of the state do not change the computation of $P(x_i|s_i)$. These improvements still leave the worst case time complexity unchanged, but in practice considerably reduce the run time. We refer to the resulting factorial HMM that we just described (Equations 1 and 2) as the *standard model*.

The standard model has some clear deviations from reality. First, it assumes that the pedigree is known with certainty, which often is not the case. Second, it assumes that the hidden states are first-order Markovian, which means in genetics language, that a recombination event at slot i (encoded as $s_i^j \neq s_{i+1}^j$) is independent of a recombination event at the previous slot (encoded as $s_{i-1}^j \neq s_i^j$), which does not hold for close markers; the violation of this assumption is termed *chiasma interference*. A third assumption is that the two alleles for each founder are independent of each other; this assumption is called the Hardy–Weinberg equilibrium. A fourth and final assumption of the standard HMM model for linkage analysis is that the founder's alleles at slot i do not depend on the founder's alleles at slot $i-1$, hence $P(x_i|s_i)$ can be written as $\sum_f P(x_i|s_i, f_i) P(f_i)$ at each slot where $P(f_i)$ is the prior probability of the founder alleles. This assumption, which does not hold for close markers, is called *linkage equilibrium*.

The dependency between alleles of nearby markers is called *LD* (Purcell *et al.*, 2007). Our main contribution in this article is to model LD within the HMM framework, improving accuracy of inference, and still perform the forward–backward algorithm sufficiently fast for some common pedigree structures. In the experimental section, we show the improvement of accuracy as a result of modeling LD.

4 MODELING LD AND INFERRING IBD

When markers become closer, and therefore dependent, a stretch of SNPs that was inferred to be inherited as a block by affected individuals and thus have high LOD score can actually be only IBS rather than IBD. In other words, the stretch of SNPs shared by some affected individuals has not been inherited from a single source by inheritance (IBD), but inherited from multiple sources which happen to be the same (IBS). For example, a stretch of n independent markers with allele A have a probability $p_{A,1} \cdot p_{A,2} \cdot \ldots \cdot p_{A,n}$ while due to LD it is possible that only a handful of the 2^n possible allele assignments is possible, say in the extreme case only $AA\ldots A$ and $BB\ldots B$ are possible. In such case a LOD score that does not take LD into account is inflated and increases type I error, weakening the power to differentiate true signals from false ones. Consequently, as denser maps of SNPs are used in genetic analysis, modeling the effects of LD becomes increasingly important.

The availability of denser and denser SNP panels weakens the validity of the assumption of linkage equilibrium; almost every two nearby markers are dependent. MERLIN is an efficient linkage program that allows the modeling of LD. Furthermore, MERLIN's developers provide a clear example that indicates the benefits of some type of models for LD for genetic analysis, showing how two loci are found related to a disease when ignoring LD, while only one locus remains suspect when LD is taken into account (Abecasis and Wigginton, 2005). MERLIN's method consists of the following steps: cluster nearby SNPs, yielding a single marker with more than two alleles, specify the prior of each of the cluster's alleles using EM or some data set such as HapMap (Frazer *et al.*, 2007), and finally, use the new markers as input to the standard (HMM) model. The assumptions underlying this solution are that there are no recombinations within the selected clusters and that there is no dependency between clusters. We now develop an alternative to this approach that does not make these assumptions.

For the task of adding LD to the model, consider an HMM with hidden variables S_i and F_i and observed variables X_i, $i=1,\ldots,L$. The state space is now the set $S \times F$ of possible values for $S_i \times F_i$, where F_i are founder alleles at slot i. The state space is identical for every slot i. The single slot likelihood of data $P(x_i|s_i,f_i)$ at slot i is given for every state (s_i,f_i) and the transition probabilities from a state at slot $i-1$ to slot i have the product form

$$P(s_i,f_i|s_{i-1},f_{i-1}) = P(s_i|s_{i-1}) \cdot P(f_i|f_{i-1})$$

Further factorization of the transition matrix is given by

$$P(f_i = (f_i^1,\ldots,f_i^k)|f_{i-1} = (f_{i-1}^1,\ldots,f_{i-1}^k)) = \prod_{j=1}^{k} P_j(f_i^j|f_{i-1}^j), \qquad (5)$$

and for the first slot, $P(f_1 = (f_1^1,\ldots,f_1^k)) = \prod_{j=1}^{k} P_j(f_1^j)$. A similar factorization is assumed also for $P(s_i|s_{i-1})$ as given by Equation 2. We refer to the factorial HMM represented by Equations 2 and 5 as the *full model*.

We note that if every f_i is independent of f_{i-1}, then the full model reduces to the standard model described in the previous section by setting $P(x_i|s_i) = \sum_f P(x_i|s_i,f_i)P(f_i)$. In our genetics application, f_i represents alleles of the founders at slot i, and f_i being independent of f_{i-1} means that there is linkage equilibrium between the markers at slot i and $i-1$. Hence, the full model generalizes the standard model by encoding LD via $P(f_i|f_{i-1})$. These conditional probability tables, which we call *LD terms*, are estimated directly from data sets such as HapMap that provides sets of haplotypes.

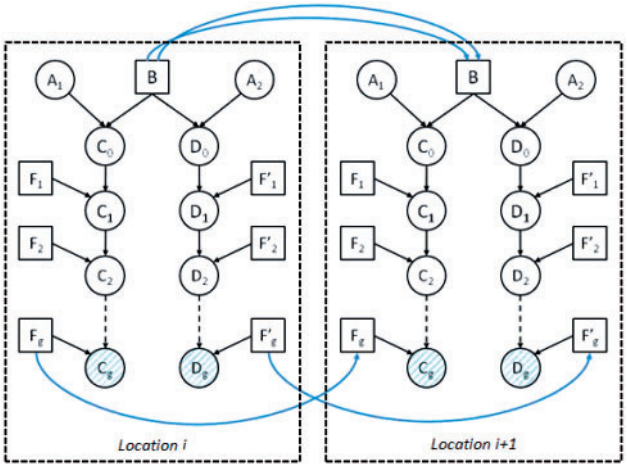

Fig. 1. Two slots of the g-degree relatives pedigree, where each node represents an individual at a specific location. The last generation is genotyped (measured). The full model has in addition two LD terms between the founder F_i in one slot and F_i in the next slot, for every founder in the pedigree. For clarity, these links are not shown. The model drawn has one LD term for each of the common founder B's alleles, and one additional LD term for each of the founders F_g and F_g'.

We have experimented with an implementation of the model and report the results in the experimental section. There is an increase of precision in all experiments with respect to current software for IBD sharing. However, the full model has a significantly higher computational complexity than the standard model that severely limits the applicability to specific pedigree topologies. The source of increased time complexity stems mainly from the increase of the hidden state space from S to $H = S \times F$. The magnitude of this increment depends on the relative size of F versus S which depends strongly on the pedigree topology. For example, a pedigree that consists of two parents and c children has a state space size of the order $S = O(2^c)$ while F is a constant of size 16. Hence for a nuclear family with many children, the increase of the state space can be acceptable. On the other hand, for g-degree relatives pedigree the state space without LD is $O(2^{2g})$, while the state space with LD grows to $O(2^{6g})$ because each founder F_i and F'_i adds two LD terms to the standard model, one for each founder allele. See Figure 1 for an illustration. Notably, the full model is still a factorial HMM and therefore, the forward–backward algorithm has a matrix multiplication complexity of $O(|H|\log|H|)$ rather than $O(|H|^2)$, which means that it can run for small pedigrees such as nuclear families or small three generation family pedigrees. Such families are often used in genetics studies. Consequently, since the full model exhibits some improvement in precision, it should be used whenever possible.

The time complexity of the full model increases significantly with larger pedigrees, because the factorial HMM is augmented with an LD term $P(f_i^j|f_{i-1}^j)$ for every allele for every founder in the pedigrees. However, since this addition burdens the computations, one can settle with modeling the LD only between a subset of founders. Major reduction of the running time can be achieved using this approximation. In this article we focus on the g-degree relatives

model with two LD terms for the single common founder and a single LD term for each of the two direct parents of the typed individuals. We term this model the *4-track* model, reflecting the fact that four LD chains are retained from the full model. We further call the model containing only two LD terms for the single common founder the *2-track* model.

The size of the hidden state space for the 4-track model is $O(2^{2g})$, which yields exponential time and space requirements. We now describe a reduction of the state space to $O(g^2)$ such that the likelihood computed in the 4-track model is identical to the likelihood computed when the state space is reduced. The ideas that lead to this major reduction of time and space complexity are based on Geiger *et al.* (2008), where we analyzed the g-degree relatives models without modeling LD. The state space reductions are formed by clustering the selectors S_i and partitioning the states of each cluster so that Conditions I and II, stated in Section 2, are satisfied.

We first define a class of clusters that satisfies Condition II and then make a specific choice within this class that also satisfies Condition I. A selector S can have two complement states: ON and OFF. For a cluster C with r such selectors, a state $[j]$ of the reduced state space of C is the equivalence class that contains all vectors of size r that have j entries being ON and $r-j$ being OFF. So, we have $c(j,r)=r!/j!(r-j)!$ vectors in state $[j]$ for $j=0,\ldots,r$. This set of $r+1$ equivalent classes is called the *counting partition*.

For the counting partition, the transition probability $P([i]|c_j)$ for switching from a state c_j of C with j positions ON to any one state with i positions ON is specified below. Let θ be the probability of switching from state ON to state OFF and of switching from state OFF to state ON. The other two transitions have probability $1-\theta$. The probability of switching from a state c_j, where j selectors are ON to the state $[i]$ in which some arbitrary i selectors are ON, is given by

$$P([i]|c_j)= \sum_t c(t,j)\cdot c(i-t,r-j)(1-\theta)^{r-(i+j-2t)}\cdot\theta^{i+j-2t}$$

where t is the number of selectors that are ON both in $[i]$ and in state c_j, ranging from $max(0,i+j-r)$ to $\min(i,j)$. With these transition probabilities, the next theorem states that the reduced state space satisfies Condition II.

THEOREM 1. *Let $S=(S^1,\ldots,S^k)$ be a vector of selectors and let $\mathcal{C}=\{C_1,\ldots,C_m\}$ be a set of disjoint clusters with $r_1,\ldots,r_m$ selectors, respectively, in each cluster, where $k=\sum_{j=1}^m r_j$. Then a factorial HMM in which the hidden variable has values drawn from the Cartesian product $[\mathcal{C}]=[C_1]\times\cdots\times[C_m]\times F$, where F is some fixed state space, $[C_l]$ is the set of equivalence classes of cluster C_l generated by using the counting partition, satisfies Condition II.*

This theorem has been stated and proven in Geiger *et al.* (2008) for the case $F=\emptyset$, namely, when no LD terms are included. Since the model is a factorial HMM and F is part of the transition probability that does not change in the reductions, the proof given earlier for $F=\emptyset$ still holds verbatim, and Condition II holds for counting partitions also when some or all the LD links are added to the standard model. The theorem holds for models of arbitrary pedigrees.

Satisfying also Condition I requires limiting the way we cluster selectors. For the g-degree relatives pedigree, the 4-track model includes a set of selectors along the chain of inheritance to one individual, and a chain of selectors of the same length to the other individual.

THEOREM 2. *Consider the 4-track model for g-degree relatives. Let $S_a=(S_a^0,\ldots,S_a^g)$ be a vector of selectors for the first chain and let $S_b=(S_b^0,\ldots,S_b^g)$ be a vector of selectors for the second chain in this model. Let S_a^C be a binary variable with a value* ON *if $S_a^i=$* ON *for $i=1,\ldots,g$, and a value* OFF *if $S_a^i=$* OFF *for at least some i, $1\le i\le g$. Let S_b^C defined similarly w.r.t. the second chain. Then the likelihood of SNP data at each slot is determined by S_a^C,S_a^0,S_b^C,S_b^0, the founder alleles f^1,f^2,f^3,f^4 and the prior allele distribution.*

PROOF. The proof analyzes all possible assignments to the selectors in the 4-track model and shows that all $O(2^{2g})$ assignments map into the 16 possibilities defined by the four binary variables S_a^C,S_a^0,S_b^C and S_b^0. The data consists of four alleles, two for each individual. Each individual receives one allele from the parent not on the chain to the common ancestor, namely f^3 and f^4. In addition, each individual receives the second allele either from the common founder or from another founder off the chain of inheritance.

A state (s_i,f_i) is *consistent* with the data x_i if the inheritance defined by state s_i connects each measured allele to a founder allele that equals to it, or to an allele of a founder who is off the chain of inheritance. Out of the four measured alleles, $N=0,1,2$ are inherited from the set $\{f^1,f^2\}$ — the common founder's alleles. Let $p_1,\ldots p_4$ be the marginal probabilities of the four measured allele. The probability of data is 0 if the state is not consistent with the data. Otherwise, the probability of the data is a function of these four marginals. Since consistency and the number of shared alleles N is determined by $S_a^C,S_a^0,S_b^C,S_b^0,f^1,\ldots,f^4$, so is the probability of data. ■

The next theorem summarizes the main claim.

THEOREM 3. *The likelihood computed in the 4-track model with and without the state space reduction are identical. The time complexity is $O(g^2\log g)$.*

PROOF. We have shown that the state space reductions for the 4-track model satisfy Conditions I and II. Consequently, the likelihood computed in the 4-track model is identical to the likelihood computed with the reduced state space. The time complexity of multiplication for factorial HMM is given by $O(|H|\log|H|)$ where H is the domain of the hidden state space. Here $H=2^4\times g\times g$, yielding the claimed complexity. ■

Similar results also hold for arbitrary pedigrees by adding any subset of LD links to the standard model. The statement and proof of the most general claim requires several definitions and a lengthy derivation that are beyond the scope of this article. However, the arguments follow closely the line of reasoning pursued in Geiger *et al.* (2008). Using that paper's terminology, the definition of a chain needs a slight adjustment. Each chain must now be split to two chains at each individual for which an LD term is added. That paper explains how the state space reductions are formed and, using the revised definition of a chain, the arguments for correctness change only slightly. We refer the reader to Geiger *et al.* (2008) for details.

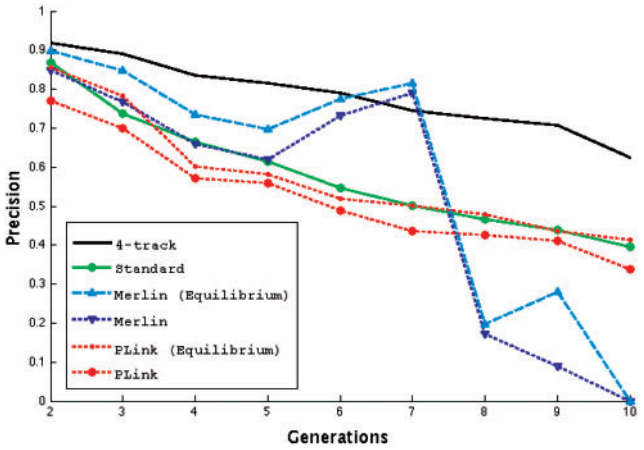

Fig. 2. The 4-track model achieves a significantly higher precision in comparison to the other methods under most tested scenarios, increasing the performance gap for $g \geq 8$ with over 50% improvement. Results correspond to over 95% recall for the 4-track model, and a lower recall for PLINK and MERLIN.

5 EXPERIMENTAL RESULTS

We divide the experiments into two main components. First, we demonstrate the improvements in the IBD inference precision using simulated data. We compare our results to the standard programs MERLIN and PLINK. Second, we apply the association mapping technique offered by PLINK, which uses inferred IBD status to score genome loci as suspect areas for predisposing genes. We show that the increased IBD inference accuracy gained using our method improves the performance in the task of gene mapping. The higher precision of our method when inferring IBD status is well illustrated in Figure 2. Each point is the average of 100 runs. The data was simulated as follows. We used a pedigree representing a pair of g-degree relatives and their ancestors (as depicted in Fig. 1). Haplotype data of the founders were randomly sampled from HapMap's CEPH (Utah residents with ancestry from northern and western Europe) set of phased individuals. We simulated the offspring along the chain of inheritance using the corresponding pedigree topology and recombination probabilities $P(S_i|S_{i-1})$. Both PLINK and MERLIN were applied using default parameters and the original marker set used in simulations. As both models assume that markers are independent, we further evaluated the performance of these two methods on a reduced marker set that accommodates this linkage equilibrium assumption. The markers were selected by applying PLINK's VIF (Variance Inflation Factor) pruning, with the recommended setting.

The precision values shown in Figure 2 are for recall of approximately 95%. The results indicate that the 4-track model is superior to previous methods in terms of precision, except for MERLIN when $g=7$, while maintaining a higher recall rate in all the examined cases. Namely, the recall of the 4-track model was above 95% in all reported results, whereas previous models had at most 95% recall and often lower. All methods deteriorate as a function of g, but the 4-track model deteriorates slower than others. As expected, pruning markers in linkage disequilibrium improves the performance of both PLINK and MERLIN (denoted by *Equilibrium* in the graph).

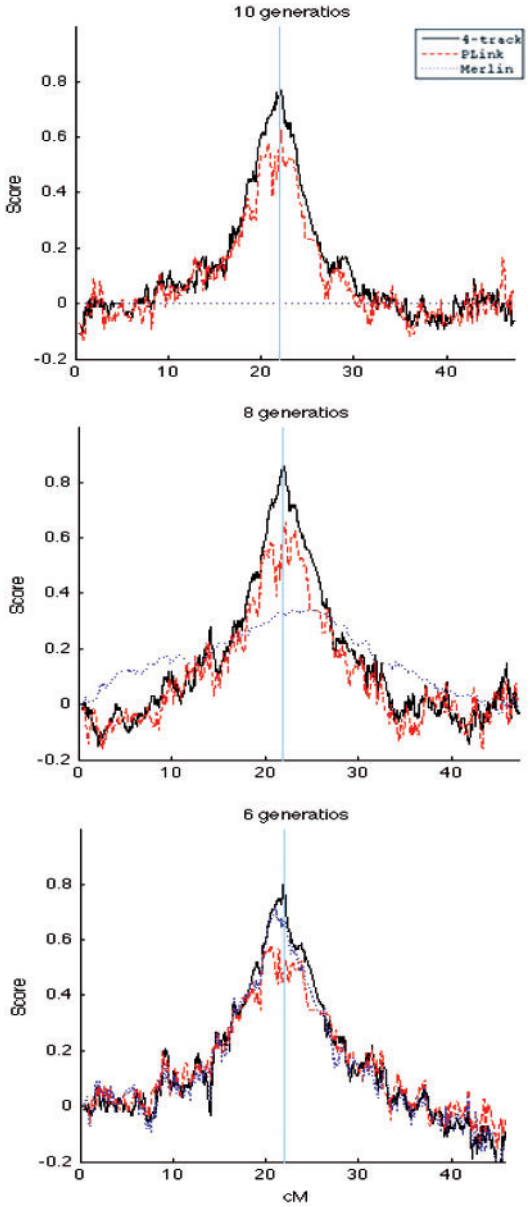

Fig. 3. Finding a disease seeded at 22 cm (location indicated by the horizontal line) using IBD predictions. Pairs of individuals were simulated using $g=6,8,10$, assuming a dominant disease with complete penetrance that originated from a common founder.

As an extreme example, we consider data simulated with $g=30$. The ability to apply our program successfully to such data means that one can hope to identify common genomic areas of individuals that have a common ancestor ~700 years ago, around the population bottleneck event caused by the black death that killed approximately half of Europe's population. When using the 2-track model, and assuming $g=10$ during the inference, IBDMAP achieves a precision of 35% for a recall of 85% and a precision of 47% for a recall of 75%. This configuration was used to expedite the results.

The second step of our experiments has been aimed at examining the extent of improvement in gene mapping techniques as a result of the increased precision in IBD inference. The program PLINK, among

many other functions, accepts SNP data of affected individuals, then infers the IBD areas for each pair and consequently uses the following formula to score to what extent an SNP is suspected to predispose or cause a disease:

$$S_i = \frac{M_i^{\text{aa}} - \bar{M}^{\text{aa}}}{N^{\text{aa}}} - \frac{M_i^{!\text{aa}} - \bar{M}^{!\text{aa}}}{N^{!\text{aa}}} \quad (6)$$

where N^{aa} and $N^{!aa}$ represent the number of pairs of affected and unaffected individuals, respectively, and M_i^{aa} and $M_i^{!aa}$ represent the number of pairs with IBD at marker i that are both affected or where at least one of the pair is unaffected, respectively (Purcell *et al.*, 2007). The terms $\bar{M}^{aa}$ and $\bar{M}^{!\text{aa}}$ are the averages of M_i^{aa} and $M_i^{!\text{aa}}$. The input of this formula relies on the inferred IBD status of an SNP for each pair of input individuals. Our test replaces the IBD inference mechanism and plots the resulting score. We simulated data from g-degree relatives using a dominant disease model with complete penetrance, designating a specific founder's haplotype location as the disease origin. Founder haplotypes, including those of the common ancestor, were randomly selected from the HapMap CEPH data, using a separate set from the one used for learning the LD. Figure 3 details the performance of the different methods in the task of disease gene mapping. For $g=6$, the results show that all three models correctly identify the disease locus, placed at 22 cm. However, for $g=8$, the peak around the genuine locus is made much more evident when using the 4-track model than when using either PLINK or MERLIN, maintaining this relative performance for $g=10$. Repeating these experiments with different randomly chosen SNPs to be the location of the disease yielded similar results. We note that these experiments were done over a 50 cm region on Chromosome 1, using SNP density comparable to the 500 k genome-wide panels.

6 DISCUSSION

Measurements of LD in recent years become more accurate as data of haplotypes accumulate. Yet, even with current public data, the knowledge of LD already upgrades the performance of various methods for gene mapping such as linkage analysis, mapping by admixture linkage disequilibrium (MALD), and association studies (Abecasis and Wigginton, 2005; Bercovici and Geiger, 2009; Eskin, 2008). In this article we incorporated a first-order Markov model of LD within models of family inheritance and showed how this improves the analysis of shared IBD areas. The proposed model becomes computationally efficient because we have devised a suitable partition of the state space, which satisfies Conditions I and II, and as a result reduces an exponential complexity to a quadratic one.

The fundamental difficulty in estimating IBD status is that recombination events erase the trace of inheritance after a few dozens of generations leaving small IBD areas that compete against random areas that are equal only by state. The errors produced by PLINK are mostly due to interpreting an IBS area as being IBD, and our method that uses a novel first-order HMM for modeling LD reduces these mistakes. An important challenge remains to incorporate better models of LD within better models of family inheritance to obtain higher accuracy of IBD inference. The result of more accurate models that are still computationally feasible will be the ability to identify smaller regions of IBD whose origin is more distant than what can be inferred with current models. In addition, our model can be extended so as to take into account genotyping errors via appropriate adjustment of Equation 4. Specifically, error in genotyping can be expressed through the conditional probability $P(x_i|s_i,f_i)$.

Our model assumes a pedigree structure of an equidistance single ancestor, based on the observation that the closest common ancestor contributes most to the IBD between two individuals. As such, the model efficiently approximates more complex pedigrees for which exact computation of IBD status is intractable. The performance of our method with respect to an arbitrary given pedigree can be evaluated via simulations. It further follows that a misestimation of the number of generations since the common ancestor affects our method's performance. For example, in the case of 5-degree relatives, a mis-specified $g=3,4,6,7$ during the inference will reduce the precision from 89% to 82%, for a recall of 92%. One possible approach for the estimation of the parameter g is by learning from the SNP data of the two individuals, using adjusted maximum likelihood (Schwarz, 1978).

The main contribution of this article is towards mapping techniques that do no not use pedigrees as input, such as association studies, IBD mapping and homozygosity mapping. We assume, as with PLINK, that pairs of affected persons have a higher likelihood for a common ancestor, and when such an ancestor is inferred, the IBD areas can be detected with high probability. We modeled the relationship between two input individuals using a g-degree relatives pedigree, namely, a single common ancestor with a separate path to each individual. The computational breakthrough that allowed this computation to take place for large enough g has been the clustering of selectors that represent meiosis events along the two inheritance paths, into two variables, each with $g+1$ states, and this reduced the complexity of the state-of-the-art linkage program MERLIN from exponential to quadratic. A software implementation, called IBDMAP, is freely available at *http://bioinfo.cs.technion.ac.il/IBDmap*. The higher accuracy along with the reduced time complexity marks our method as a feasible means for IBD mapping in practical scenarios.

Funding: Israel Science Foundation; Azrieli Fellowship, Azrieli Foundation (to S.B.).

Conflict of Interest: none declared.

REFERENCES

Abecasis,G. and Wigginton,J. (2005) Handling marker-marker linkage disequilibrium: pedigree analysis with clustered markers. *Am. J. Hum. Genet.*, **77**, 754–767.

Abecasis,G.R. *et al.* (2002) Merlin-rapid analysis of dense genetic maps using sparse gene flow trees. *Nat. Genet.*, **30**, 97–101.

Benjamini,Y. and Yekutieli,D. (2005) Quantitative traits loci analysis using the false discovery rate. *Genetics*, **171**, 783–789.

Bercovici,S. and Geiger,D. (2009) Inferring ancestries efficiently in admixed populations with linkage disequilibrium. *J. Comput. Biol.*, **16**, 1141–1150.

Browning,S. and Browning,B. (2002) On reducing the statespace of hidden markov models for the identity by descent process. *Theor. Popul. Biol.*, **62**, 1–8.

Cardon,L. and Abecasis,G. (2003). Using haplotype blocks to map human complex trait loci. *Trends Genet.*, **19**, 135–140.

Carlson,C. *et al.* (2004) Mapping complex disease loci in whole-genome association studies. *Nature*, **429**, 446–452.

Cottingham,R.W. *et al.* (1993) Faster sequential genetic linkage computations. *Am. J. Hum. Genet.*, **53**, 252–263.

Dechter,R. (1998) Bucket elimination: a unifying framework for probabilistic inference. *Proceedings of the NATO Advanced Study Institute on Learning in graphical models*, Kluwer Academic Press, Erice, Italy, pp. 75–104.

Elston,R. and Stewart,J. (1971) A general model for the analysis of pedigree data. *Hum. Hered.*, **21**, 523–542.

Eskin,E. (2008) Increasing power in association studies by using linkage disequilibrium structure and molecular function as prior information. *Genome Res.*, **18**, 653–660.

Fishelson,M. and Geiger,D. (2002) Exact genetic linkage computations for general pedigrees. *Bioinformatics*, **18**(Suppl. 1), S189–S198.

Frazer,K. *et al*. (2007) A second generation human haplotype map of over 3.1 million snps. *Nature*, **449**, 851–861.

Geiger,D. *et al*. (2008) Speeding up HMM algorithms for genetic linkage analysis via chain reductions of the state space. *Bioinformatics*, **25**, i196–i203.

Ghahramani,Z. and Jordan,M. (1997) Factorial hidden Markov models. *Mach. Learn.*, **29**, 245–273.

Greenspan,G. and Geiger,D. (2004) High density linkage disequilibrium mapping using models of haplotype block variation. *Bioinformatics*, **20**(Suppl. 1), i137–i144.

Gudbjartsson,D. *et al*. (2005) Allegro version 2. *Nat. Genet.*, **37**, 1015–1016.

Gudbjartsson,D. *et al*. (2000) Allegro, a new computer program for multipoint linkage analysis. *Nat. Genet.*, **25**, 12–13.

Halperin,E. and Stephan,D. (2009) Maximizing power in association studies. *Nat. Biotecmol.*, **27**, 255–256.

Halperin,E. *et al*. (2005) Tag SNP selection in genotype data for maximizing SNP prediction accuracy. *Bioinformatics*, **21**(Suppl. 1), i195–i203.

Han,B. *et al*. (2009) Rapid and accurate multiple testing correction and power estimation for millions of correlated markers. *PloS Genet.*, **5**, e1000456.

Ingolfsdottir,A. and Gudbjartsson,D. (2005) Genetic linkage analysis, algorithms and their implementation. *Trans. Comput. Syst. Biol.*, **3737**, 123–144.

Kruglyak,L. and Lander,E. (1998) Faster multipoint linkage analysis using Fourier transform. *J. Comput. Biol.*, **5**, 1–7.

Kruglyak,L. *et al*. (1995) Rapid multipoint linkage analysis of recessive traits in nuclear families including homozygosity mapping. *Am. J. Hum. Genet.*, **56**, 519–527.

Kruglyak,L. *et al*. (1996) Parametric and nonparametric linkage analysis: a unified multipoint approach. *Am. J. Hum. Genet.*, **58**, 1347–1363.

Lander,E. and Green,P. (1987) Construction of multilocus genetic maps in humans. *Proc. Natl Acad. Sci.*, **84**, 2363–2367.

Lange,K. (1997) *Mathematical and Statistical Methods for Genetic Analysis.* Springer-Verlag, New York.

Lauritzen,S.L. (1996) *Graphical Models.* Oxford University Press, Oxford, UK.

Lauritzen,S.L. and Spiegelhalter,D.J. (1988) Local computations with probabilities on graphical structures and their application to expert systems (with discussion). *J. Roy. Stat. Soc. Series B stat. Methodol.*, **50**, 157–224.

Markianos,K. *et al*. (2001) Efficient multipoint linkage analysis through reduction of inheritance space. *Am. J. Hum. Genet.*, **68**, 963–977.

O'Connell,J. and Weeks,D. (1995) The VITESSE algorithm for rapid exact multilocus linkage analysis via genotype set-recoding and fuzzy inheritance. *Nat. Genet.*, **11**, 402–408.

Ott,J. (1999) *Analysis of Human Genetic Linkage.* Johns Hopkins University Press, Baltimore, Maryland.

Pearl,J. (1988) *Probabilistic Reasoning in Intelligent Systems.* Morgan Kaufmann, San Francisco, CA.

Peer,I. *et al*. (2006) Evaluating and improving power in whole genome association studies using fixed marker sets. *Nat. Genet.*, **38**, 663–667.

Peer,I. *et al*. (2008) Estimation of the multiple testing burden for genomewide association studies of nearly all common variants. *Genet. Epidemiol.*, **32**, 381–385.

Purcell,S. *et al*. (2007) Plink: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.*, **81**, 559–575.

Rabiner,L.R. and Juang,B.H. (1986) An introduction to hidden Markov models. *IEEE Acoust. Speech sign. Process. Mag.*, pages 4–15.

Schwarz,G. (1978) Estimating the dimension of a model. *Ann. Stat.*, **6**, 461–464.

Silberstein,M. *et al*. (2006) Online system for faster multipoint linkage analysis via parallel execution on thousands of personal computers. *Am. J. Hum. Genet.*, **78**, 922–935.

Sobel,E. and Lange,K. (1996) Descent graphs in pedigree analysis: applications to haplotyping, location scores, and marker sharing statistics. *Am. J. Hum. Genet.*, **58**, 1323–1337.

Thompson,E.A. (1994) Monte Carlo likelihood in genetic mapping. *Stat. Sci.*, **9**, 355–366.

Wang,W. *et al*. (2005) Genome-wide association studies: theoretical and practical concerns. *Nat. Rev. Genet.*, **6**, 109–118.