Genotype Error Detection using Hidden Markov Models of Haplotype Diversity*

Justin Kennedy† Ion Măndoiu† Bogdan Paşaniuc†

July 5, 2008

Abstract

The presence of genotyping errors can invalidate statistical tests for linkage and disease association, particularly for methods based on haplotype analysis. Becker et al. have recently proposed a simple likelihood ratio approach for detecting errors in trio genotype data. Under this approach, a SNP genotype is flagged as a potential error if the likelihood associated with the original trio genotype data increases by a multiplicative factor exceeding a user selected threshold when the SNP genotype under test is deleted. In this paper we give improved error detection methods using the likelihood ratio test approach in conjunction with likelihood functions that can be efficiently computed based on a Hidden Markov Model of haplotype diversity in the population under study. Experimental results on both simulated and real datasets show that proposed methods have highly scalable running time and achieve significantly improved detection accuracy compared to previous methods.

1 Introduction

The sequencing of the human genome coupled with the initial mapping of human haplotypes by the HapMap project and rapid advances in SNP genotyping technologies have recently opened up the era of genome-wide

*A preliminary version of this paper will appear in *Proc. 7th Workshop on Algorithms in Bioinformatics (WABI)*, Springer Verlag Lecture Notes in Computer Science, 2007.

†Computer Science and Engineering Department, University of Connecticut, 371 Fairfield Way, Unit 2155, Storrs, CT 06269-2155. E-mail: {jlk02019,ion,bogdan}@engr.uconn.edu.

association studies, which promise to uncover the genetic basis of common complex diseases such as diabetes and cancer by analyzing the patterns of genetic variation within healthy and diseased individuals. However, the validity of associations uncovered in these studies critically depends on the accuracy of genotype data. Despite recent progress in genotype calling algorithms (Cutler et al., 2001; Di et al., 2005; Marchini et al., 2007; Nicolae et al., 2006; Rabbee & Speed, 2005; Xiao et al., 2007), significant error levels remain present in SNP genotype data due to factors ranging from human error and sample quality to sequence variation and assay failure, see (Pompanon et al., 2005) for a recent survey. A recent study of dbSNP genotype data (Zaitlen et al., 2005) found that as much as 1.1% of about 20 million SNP genotypes typed multiple times have inconsistent calls, and are thus incorrect in at least one dataset.

Recommended quality control procedures such as the use of external control samples from HapMap and duplication of internal samples (NCI-NHGRI Working Group on Replication in Association Studies, 2007) provide an estimate of error rates, but do not eliminate them. Although systematic errors such as assay failure can be detected by departure from Hardy-Weinberg equilibrium proportions (Hosking et al., 2004; Leal, 2005), and, when genotype data is available for related individuals, some errors become detectable as *Mendelian Inconsistencies* (MIs), a large fraction of errors remains undetected by these analyses, e.g., as much as 70% of errors in mother-father-child trio genotype data are undetected by Mendelian consistency analysis (Douglas et al., 2002; Gordon et al., 1999). Since even low error levels can lead to inflated false positive rates and substantial losses in the statistical power of linkage and association studies (Ahn et al., 2007; Abecasis et al., 2001; Cherny et al., 2001; Gordon et al., 2002; Mitchell et al., 2003; Zheng & Tian, 2005), detecting Mendelian consistent errors remains a critical task in genetic data analysis. This task becomes particularly important in the context of association studies based on haplotypes instead of single locus markers, where error rates as low as 0.1% may invalidate some statistical tests for disease association (Knapp & Becker, 2004).

A powerful approach of dealing with genotyping errors is to explicitly model them in downstream statistical analyses, see, e.g., (Cheng, 2007; Hao & Wang, 2004; Liu et al., 2007). While powerful, this approach often leads to complex statistical models and impractical runtime for large datasets such as those generated by genome-wide association studies. A more practical approach is to perform genotype error detection as

a separate analysis step following genotype calling. SNP genotypes flagged as putative errors can be either excluded from downstream analyses or retyped when high quality genotype data is required. Indeed, such a separate error detection step is currently implemented in all widely-used software packages for pedigree genotype data analysis including Mendel (Sobel et al., 2002), Merlin (Abecasis et al., 2002), Sibmed (Douglas et al., 2000), and SimWalk2 (Sobel & Lange, 1996; Sobel et al., 2002), all of which detect Mendelian consistent errors by independently analyzing each pedigree and identifying loci of excessive recombination. Unfortunately, these methods have very limited power to detect errors in genotype data from small pedigrees such as mother-father-child trios, and do not apply at all to genotype data from unrelated individuals. (Becker et al., 2006) have recently introduced the use of *population level* haplotype frequency information for genotype error detection in trio data via a simple likelihood ratio test. However, detection accuracy of their method is severely limited by the reliance on explicit enumeration of most frequent haplotypes within short blocks of consecutive SNP loci.

In this paper we propose novel methods for genotype error detection extending the likelihood ratio error detection approach of (Becker et al., 2006). While we focus on detecting errors in trio genotype data, our proposed methods apply with minor modifications to genotype data coming from unrelated individuals and small pedigrees other than trios. Unlike (Becker et al., 2006), we employ a hidden Markov model (HMM) to represent frequencies of all possible haplotypes over the set of typed loci. Similar HMMs have been successfully used in recent works (Kimmel & Shamir, 2005; Rastas et al., 2007; Scheet & Stephens, 2006; Schwartz, 2004) for genotype phasing and disease association. Two limitations of previous uses of HMMs in this context have been the relatively slow training based on genotype data and the inability to exploit available pedigree information. We overcome these limitations by training our HMM using haplotypes inferred by the pedigree-aware phasing algorithm of (Gusev et al., to appear), based on entropy minimization.

(Becker et al., 2006) use maximum phasing probability of a trio genotype as the likelihood function whose sensitivity to single SNP genotype deletions signals potential errors. The former is heuristically approximated by a computationally expensive search over quadruples of frequent haplotypes inferred for each window. We show that, when haplotype frequencies are implicitly represented using an HMM, computing the maximum trio phasing probability is, unfortunately, hard to approximate in polynomial time. Despite this

hardness result, we are able to significantly improve both detection accuracy and speed compared to (Becker et al., 2006) by using alternate likelihood functions such as Viterbi probability and the total trio genotype probability, both of which can be computed for commonly used unrelated and trio genotype data within a worst-case runtime that increases linearly in the number of SNP loci and that of genotyped individuals. Further improvements in detection accuracy for genotype trio data are obtained by combining likelihood ratios computed for different subsets of trio members. Empirical experiments show that this technique is very effective in reducing false positives within correctly typed SNP genotypes for which the same locus is mistyped in related individuals.

The rest of the paper is organized as follows. We introduce basic notations in Section 2 describe the structure of the HMM used to represent haplotype frequencies in Section 3, and present the likelihood ratio approach of (Becker et al., 2006) in Section 4. In Section 5 we show that the likelihood function in (Becker et al., 2006) cannot be approximated efficiently when an HMM is used to represent haplotype frequencies, and then, in Section 6, we give three alternative likelihood functions that can be computed efficiently based on an HMM. Finally, we give experimental results assessing the error detection accuracy of our methods on both simulated and real datasets in Section 7, and conclude with ongoing research directions in Section 8.

2 Preliminaries

We start by introducing basic terminology and notations used throughout the paper. We denote the major and minor alleles at a SNP locus by 0 and 1. A *SNP genotype* represents the pair of alleles present in an individual at a SNP locus. Possible SNP genotype values are $0/1/2/?$, where 0 and 1 denote homozygous genotypes for the major and minor alleles, 2 denotes the heterozygous genotype, and ? denotes missing data. SNP genotype g is said to be explained by an ordered pair of alleles $(\sigma, \sigma') \in \{0,1\}^2$ if $g =?$, or $g \in \{0,1\}$ and $\sigma = \sigma' = g$, or $g = 2$ and $\sigma \neq \sigma'$.

We denote by n the number of SNP loci typed in the population under study. A *multi-locus genotype* (or simply *genotype*) is a $0/1/2/?$ vector G of length n, while a *haplotype* is a $0/1$ vector H of length n. An ordered pair (H, H') of haplotypes explains multi-locus genotype G iff, for every $i = 1, \ldots, n$, the pair $(H(i), H'(i))$ explains $G(i)$. A *trio genotype* is a triple $T = (G_m, G_f, G_c)$ consisting of mother, father, and child multi-

locus genotypes. Assuming that no recombination takes place within the set of SNP loci of interest, we say that an ordered 4-tuple (H_1, H_2, H_3, H_4) of haplotypes explains trio genotype $T = (G_m, G_f, G_c)$ iff (H_1, H_2) explains G_m, (H_3, H_4) explains G_f, and (H_1, H_3) explains G_c. A *genotype duo* consisting of mother-child or father-child genotypes is defined similarly. An ordered 3-tuples of haplotypes (H_1, H_2, H_3) is said to explain such a duo iff (H_1, H_2) explains the parent genotype and (H_1, H_3) explains the child genotype.

3 Hidden Markov Model

Figure 1 belongs here.

The HMM used to represent haplotype frequencies has a similar structure to HMMs recently used in (Kimmel & Shamir, 2005; Rastas et al., 2007; Scheet & Stephens, 2006; Schwartz, 2004). This structure (see Figure 1) is fully determined by the number of SNP loci n and a user-specified *number of founders* K (typically a small constant, we used $K = 7$ in our experiments). Formally, the HMM is specified by a triple $M = (Q, \gamma, \varepsilon)$, where Q is the set of states, γ is the transition probability function, and ε is the emission probability function. The set of states Q consists of disjoint sets $Q_0 = \{q^0\}, Q_1, Q_2, \ldots, Q_n$, with $|Q_1| = |Q_2| = \cdots = |Q_n| = K$, where q^0 denotes the start state and Q_j, $1 \leq j \leq n$, denotes the set of states corresponding to SNP locus j. The transition probability between two states a and b, $\gamma(a, b)$, is non-zero only when a and b are in consecutive sets Q_i. The initial state q^0 is silent, while every other state q emits allele $\sigma \in \{0, 1\}$ with probability $\varepsilon(q, \sigma)$. The probability with which M emits a haplotype H along a path π starting from q^0 and ending at a state in Q_n is given by:

$$P(H, \pi|M) = \gamma(q^0, \pi(1))\varepsilon(\pi(1), H(1))\prod_{i=2}^{n}\gamma(\pi(i-1), \pi(i))\varepsilon(\pi(i), H(i)) \quad (1)$$

Intuitively, M represents founder haplotypes along high-probability paths of states, with recombination between pairs of founder haplotypes being captured via remaining transition probabilities.

In (Kimmel & Shamir, 2005; Rastas et al., 2007), similar HMMs were trained using genotype data via variants of the EM algorithm. Since EM-based training is generally slow and cannot be easily modified to take advantage of phase information that can be inferred from available family relationships, we adopted the following two-step approach for training our HMM. First, we use the highly scalable ENT algorithm of (Gusev et al., to appear) to infer haplotypes for all individuals in the sample based on entropy minimization.

ENT can handle genotypes related by arbitrary pedigrees, and has been shown to yield high phasing accuracy as measured by the so called *switching error*, which implies that inferred haplotypes are locally correct with very high probability. In the second step we use the classical Baum-Welch algorithm (Baum et al., 1970) to train the HMM based on the haplotypes inferred by ENT.

4 Likelihood ratio approach to genotype error detection

Our detection methods are based on the likelihood ratio approach of (Becker et al., 2006). We call *likelihood function* any function L assigning non-negative real-values to trio genotypes, with the further constraint that L is non-decreasing under data deletion. Let $T = (G_m, G_f, G_c)$ denote a trio genotype, $x \in \{m, f, c\}$ denote one of the individuals in the trio (mother, father, or child), and i denote one of the n SNP loci. The trio genotype $T_{(x,i)}$ is obtained from T by marking SNP genotype $G_x(i)$ as missing. The *likelihood ratio* of SNP genotype $G_x(i)$ is defined as $\frac{L(T_{(x,i)})}{L(T)}$. Notice that, by L's monotony under data deletion, the likelihood ratio is always greater or equal to 1. A SNP genotype $G_x(i)$ is flagged as a potential error whenever the corresponding likelihood ratio exceeds a user specified *detection threshold* t. A variant of this basic approach relies on simultaneously testing the mother/father/child SNP genotypes at a locus. In this variant, SNP locus i is flagged as a potential error whenever $\frac{L(T_i)}{L(T)} \geq t$, where T_i is the trio genotype obtained from T by deleting all three SNP genotypes $G_m(i)$, $G_f(i)$, and $G_c(i)$.

The likelihood function used by Becker et al. (Becker et al., 2006) is the maximum trio phasing probability,

$$L(T) = \max_{(H_1,H_2,H_3,H_4)} P(H_1)P(H_2)P(H_3)P(H_4) \tag{2}$$

where the above maximum is computed over all 4-tuples (H_1, H_2, H_3, H_4) of haplotypes that explain T. Clearly, the maximum phasing probability is monotonic under data deletion, since deleting SNP genotypes increases the number of compatible 4-tuples. The use of maximum trio phasing probability as likelihood function is intuitively appealing, since one does not expect a large increase in this probability when a single SNP genotype is deleted.

The computational complexity of computing the maximum trio phasing probability $L(T)$ depends on the encoding used to represent haplotype frequencies. When the $N = 2^n$ haplotype frequencies are given

explicitly, computing $L(T)$ can be trivially done in $O(N^4)$ time. Unfortunately, such an explicit representation can only be used for a small number n of SNP loci. To maintain practical running time, (Becker et al., 2006) adopted a heuristic that starts by creating a short list of haplotypes with frequency exceeding a certain threshold, followed by a pruned search over 4-tuples of haplotypes from this list. Due to the high computation cost of the search algorithm, the list of haplotypes must be kept very short – between 50 and 100 for the experiments reported in (Becker et al., 2006) – which makes the approach applicable only for windows of few consecutive SNP loci. This limits the amount of linkage information used in error detection, explaining at least in part the high number of false positives observed in (Becker et al., 2006) within correctly typed SNP genotypes located in the neighborhood of SNP genotypes that are mistyped in the same individual.

The HMM described in previous section provides a much more compact representation of haplotype frequencies, that can be used for large numbers of SNP loci. Although the probability of any given 4-tuple of haplotypes explaining a genotype trio can be computed efficiently based on this representation, approximating the maximum trio phasing probability is shown in next section to be computationally hard. To overcome this difficulty, in Section 6 we propose alternative likelihood functions that are efficiently computable based on an HMM representation of haplotype frequencies.

5 Inapproximability of maximum phasing probability

We first consider the problem of computing the maximum phasing probability for a single multi-locus genotype:

Maximum genotype phasing probability: Given an HMM model of haplotype diversity with n SNP loci and K founders and a genotype $G \in \{0, 1, 2, ?\}^n$, find

$$P(G|M) = \max_{(H_1,H_2)} P(H_1|M)P(H_2|M) \tag{3}$$

where the maximum is computed over all pairs (H_1, H_2) of haplotypes that explain G.

Theorem 1 *Maximum genotype phasing probability cannot be approximated within a factor of* $O(n^{\frac{1}{2}-\varepsilon})$ *for any* $\varepsilon > 0$*, unless ZPP=NP.*

Proof. We give a reduction from the problem of computing the size of the maximum clique in an undirected graph, refining the construction used in (Lyngsø & Pedersen, 2002) to show hardness of approximation for the consensus string problem.

Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be a graph with n vertices $\mathcal{V} = \{1, \ldots, n\}$. We will build an HMM $M_{\mathcal{G}}$ with $n+1$ SNP loci and a total of $K = 4n$ founders. In addition to the silent start state q^0, $M_{\mathcal{G}}$ contains for each vertex v of $\mathcal{G}$ and each SNP locus $i \in \{0, 1, \ldots, n\}$ four states denoted $q^i_{v,j}$, $j = 1, 2, 3, 4$ such that $q^i_{v,1}$ and $q^i_{v,3}$ emit 0 with probability 1, while $q^i_{v,2}$ and $q^i_{v,4}$ emit 1 with probability 1. For every $v \in \mathcal{V}$ there are two transitions from the start state q^0 to $q^0_{v,2}$ and $q^0_{v,3}$, each with probability $\frac{2^{\deg(v)}}{\gamma}$, where $\deg(v)$ denotes the degree of v in $\mathcal{G}$ and $\gamma = \sum_{v \in \mathcal{V}} 2^{\deg(v)}$ is a normalizing constant.

Remaining non-zero probability transitions take place only from a state $q^{i-1}_{v,j}$ to a state $q^i_{v,j'}$ with either $j, j' \in \{1, 2\}$ or $j, j' \in \{3, 4\}$. Non-zero probability transitions within the first two "rows" of states corresponding to vertex $v \in \mathcal{V}$ (i.e., states $q^i_{v,j}$ with $j = 1$ and $j = 2$) are as follows:

- For every SNP locus $i \in \{1, \ldots, n\} \setminus \{v\}$ such that i is not adjacent to v in $\mathcal{G}$, $M_{\mathcal{G}}$ has transitions with probability 1 from $q^{i-1}_{v,j}$, $j = 1, 2$, to $q^i_{v,1}$
- For every SNP locus $i \in \{1, \ldots, n\} \setminus \{v\}$ such that i is adjacent to v in $\mathcal{G}$, $M_{\mathcal{G}}$ has transitions with probability 1/2 from $q^{i-1}_{v,j}$, $j = 1, 2$, to both $q^i_{v,1}$ and $q^i_{v,2}$
- Finally, $M_{\mathcal{G}}$ has transitions with probability 1 from $q^{v-1}_{v,j}$, $j = 1, 2$, to $q^v_{v,2}$.

By construction, each haplotype emitted along a path within the first two rows of states corresponding to vertex v consists of a 1 followed by the characteristic vector of one of the $2^{deg(v)}$ subsets of $\mathcal{V}$ that contain v and zero or more of its neighbors.

Non-zero probability transitions within last two rows of states corresponding to vertex $v \in \mathcal{V}$ (i.e., states $q^i_{v,j}$ with $j = 3$ and $j = 4$) follow a symmetric pattern:

- For every SNP locus $i \in \{1, \ldots, n\} \setminus \{v\}$ such that i is not adjacent to v, $M_{\mathcal{G}}$ has transitions with probability 1 from $q^{i-1}_{v,j}$, $j = 3, 4$, to $q^i_{v,4}$
- For every SNP locus $i \in \{1, \ldots, n\} \setminus \{v\}$ such that i is adjacent to v, $M_{\mathcal{G}}$ has transitions with probability 1/2 from $q^{i-1}_{v,j}$, $j = 3, 4$, to both $q^i_{v,3}$ and $q^i_{v,4}$

- $M_{\mathcal{G}}$ has transitions with probability 1 from $q_{v,j}^{v-1}$, $j = 3, 4$, to $q_{v,3}^{v}$.

By construction, haplotypes emitted with non-zero probability along paths within v's last two rows consist of a 0 followed by the characteristic vector of the *complement* of a subset of $\mathcal{V}$ that contains v and zero or more of its neighbors. To illustrate the construction, Figure 2(b) gives the structure of $M_{\mathcal{G}}$ for the simple graph $\mathcal{G}$ in Figure 2(a). Figure 2 belongs here.

Note that, within the group of states corresponding to vertex v, a haplotype is emitted by $M_{\mathcal{G}}$ along a unique path whose probability $\frac{2^{\deg(v)}}{\gamma} \cdot \frac{1}{2^{\deg(v)}} = \frac{1}{\gamma}$ is independent of v and the haplotype itself. Thus, a haplotype H consisting of a 1 followed by the characteristic vector of a clique of size k of $\mathcal{G}$ is emitted by $M_{\mathcal{G}}$ with probability of k/γ, since there is exactly one path emitting H within each group of states corresponding to clique vertices. Conversely, any haplotype H starting with 1 that is emitted by $M_{\mathcal{G}}$ with probability of k/γ or more defines a clique of size k or more in $\mathcal{G}$ (consisting of vertices v whose groups of states emit H).

Let now G be the multi-locus genotype of length n that is heterozygous at every SNP locus. Clearly, G can only be explained by pairs (H_1, H_2) of haplotypes for which $H_2 = \overline{H_1}$, where $\overline{H}$ denotes the haplotype obtained by swapping 0's and 1's in H. Since the construction of $M_{\mathcal{G}}$ implies that $P(\overline{H}|M_{\mathcal{G}}) = P(H|M_{\mathcal{G}})$ for every haplotype H, it follows that

$$P(G|M_{\mathcal{G}}) = \Big(\max_{H} P(H|M_{\mathcal{G}})\Big)^2$$

and therefore $\mathcal{G}$ has a clique of size k or more iff $P(G|M_{\mathcal{G}}) \geq (k/\gamma)^2$. Since clique is hard to approximate within a factor of $O(|\mathcal{V}|^{1-\varepsilon})$ for any $\varepsilon > 0$ unless ZPP=NP (Håstad, 1999), the theorem follows. □

Remark. Computing maximum phasing probability is closely related to the optimal genotype phasing problem, which, given an HMM M and a multi-locus genotype G, asks for a pair (H_1, H_2) of haplotypes maximizing $P(H_1|M)P(H_2|M)$. Optimal genotype phasing was conjectured to be NP-hard by (Rastas et al., 2007). Since computing phasing probability $P(H_1|M)P(H_2|M)$ can be done in polynomial time for a given pair (H_1, H_2) of haplotypes by two runs of the forward algorithm, Theorem 1 trivially extends to the optimal genotype phasing problem.

The problem of computing the trio-based likelihood function of (Becker et al., 2006) based on an HMM representation of haplotype frequencies is formalized as follows:

Maximum trio phasing probability: Given an HMM model M of haplotype diversity with n SNP loci and K founders and a trio genotype $T = (G_m, G_f, G_c)$, find

$$L(T|M) = \max_{(H_1,H_2,H_3,H_4)} P(H_1|M)P(H_2|M)P(H_3|M)P(H_4|M) \tag{4}$$

where the maximum is computed over all 4-tuples (H_1, H_2, H_3, H_4) of haplotypes that explain T.

Theorem 2 *For every $\varepsilon > 0$, maximum trio phasing probability cannot be approximated within a factor of $O(n^{\frac{1}{4}-\varepsilon})$ for any $\varepsilon > 0$, unless ZPP=NP.*

Proof. We use a reduction similar to that in the proof of Theorem 1. When T consists of three genotypes that are heterozygous at every locus, the only 4-tuples of haplotypes that explain T are of the form $(H, \overline{H}, \overline{H}, H)$ for some haplotype H. Using the fact that $P(\overline{H}|M_{\mathcal{G}}) = P(H|M_{\mathcal{G}})$ it follows that $\mathcal{G}$ has a clique of size k or more iff $P(T|M_{\mathcal{G}}) \geq (k/\gamma)^4$, and the theorem follows again from the hardness of approximation established for the clique problem in (Håstad, 1999). □

6 Efficiently computable likelihood functions

In this section we consider three alternatives to the likelihood function used in (Becker et al., 2006), and describe efficient algorithms for computing them given an HMM model of haplotype diversity. As shown in Section 7, all three alternatives yield similar error detection accuracy, significantly higher than that obtained in (Becker et al., 2006).

6.1 Viterbi probability

The probability with which the HMM M emits four haplotypes (H_1, H_2, H_3, H_4) along a set of 4 paths $(\pi_1, \pi_2, \pi_3, \pi_4)$ is obtained by a straightforward extension of (1). The first proposed likelihood function is the *Viterbi probability*, defined, for a given trio genotype T, as the maximum probability of emitting haplotypes that explain T along four HMM paths. Viterbi probability can be computed using a "4-path" extension of the classical Viterbi algorithm (Viterbi, 1967) as follows.

For every 4-tuple $q = (q_1, q_2, q_3, q_4) \in Q_j^4$, let $V_f(j; q)$ denote the maximum probability of emitting alleles that explain the first j SNP genotypes of trio T along a set of 4 paths ending at states (q_1, q_2, q_3, q_4) (we will refer to these values as the *forward Viterbi values*). Also, let $\Gamma(q', q) = \gamma(q'_1, q_1)\gamma(q'_2, q_2)\gamma(q'_3, q_3)\gamma(q'_4, q_4)$ be the probability of transition in M from the 4-tuple $q' \in Q_{j-1}^4$ to the 4-tuple $q \in Q_j^4$. Then, $V_f(0; (q^0, q^0, q^0, q^0)) = 1$ and

$$V_f(j; q) = E(j; q) \max_{q' \in Q_{j-1}^4} \{V_f(j-1; q')\Gamma(q', q)\} \tag{5}$$

Here, $E(j; q) = \max_{(\sigma_1, \sigma_2, \sigma_3, \sigma_4)} \prod_{i=1}^4 \varepsilon(q_i, \sigma_i)$, where the maximum is computed over all 4-tuples $(\sigma_1, \sigma_2, \sigma_3, \sigma_4)$ that explain T's SNP genotypes at locus j. For a given trio genotype T, the Viterbi probability of T is given by $V(T) = \max_{q \in Q_n^4} \{V_f(n; q)\}$.

The time needed to compute forward Viterbi values with the above recurrences is $O(nK^8)$, where n denotes the number of SNP loci and K denotes the number of founders. Indeed, for each one of the $O(K^4)$ 4-tuples $q \in Q_j^4$, computing the maximum in (5) takes $O(K^4)$ time. A K^3 speed-up is obtained by identifying and re-using common terms between the maximums (5) corresponding to different 4-tuples q. Thus, instead of applying (5) directly we compute, for every j, the following:

- $m_1(j; q_1, q'_2, q'_3, q'_4) = \max_{q'_1 \in Q_j} \{V_f(j-1; (q'_1, q'_2, q'_3, q'_4))\gamma(q'_1, q_1)\}$ for each $(q_1, q'_2, q'_3, q'_4) \in Q_j \times Q_{j-1}^3$
- $m_2(j; q_1, q_2, q'_3, q'_4) = \max_{q'_2 \in Q_j} \{m_1(j; (q_1, q'_2, q'_3, q'_4))\gamma(q'_2, q_2)\}$ for each $(q_1, q_2, q'_3, q'_4) \in Q_j^2 \times Q_{j-1}^2$
- $m_3(j; q_1, q_2, q_3, q'_4) = \max_{q'_3 \in Q_j} \{m_2(j; (q_1, q_2, q'_3, q'_4))\gamma(q'_3, q_3)\}$for each $(q_1, q_2, q_3, q'_4) \in Q_j^3 \times Q_{j-1}$
- $V_f(j; q) = E(j; q) \max_{q'_4 \in Q_j} \{m_3(j; (q_1, q_2, q_3, q'_4))\gamma(q'_4, q_4)\}$ for each $q = (q_1, q_2, q_3, q_4) \in Q_j^4$

A similar speed-up idea was proposed in the context of single genotype phasing by (Rastas et al., 2007).

To apply the likelihood ratio test, we also need to compute Viterbi probabilities for trios with one of the SNP genotypes deleted. A naïve approach is to compute each of these probabilities from scratch using the above $O(nK^5)$ algorithm. However, this would result in a runtime that grows quadratically with the number of SNPs. A more efficient algorithm is obtained by also computing *backward Viterbi values* $V_b(j; q)$, defined as the maximum probability of emitting alleles that explain genotypes at SNP loci $j+1, \ldots, n$ of trio T along a set of 4 paths starting at the states of $q \in Q_j^4$. Once forward and backward Viterbi values are available, the Viterbi probability of a modified trio can be computed in $O(K^5)$ time by using again the

above speed-up idea, for an overall runtime of $O(nK^5)$ per trio. For unrelated individuals similar speed-up ideas lead to a runtime of $O(nK^3)$ per individual.

6.2 Probability of Viterbi haplotypes

The Viterbi algorithm described in previous section yields, together with the 4 Viterbi paths, a 4-tuple of haplotypes which we refer to as the *Viterbi haplotypes*. Viterbi haplotypes for the original trio can be computed by traceback. Similarly, Viterbi haplotypes corresponding to modified trios can be computed without increasing the asymptotic runtime via a bi-directional traceback. The second likelihood function that we considered is the probability of Viterbi haplotypes, which is obtained by multiplying individual probabilities of Viterbi haplotypes. The probability of each Viterbi haplotype can be computed using the standard forward algorithm in $O(nK)$ time. Unfortunately, Viterbi paths for modified trios can be completely different from each other, and the probability of each of them must be computed from scratch by using the forward algorithm. This results in an overall runtime of $O(nK^5 + n^2K)$ per trio, respectively $O(nK^3 + n^2K)$ per individual for genotype data from unrelated individuals.

6.3 Total trio genotype probability

The third considered likelihood function is the *total trio genotype probability*, i.e., the total probability $P(T)$ with which M emits any four haplotypes that explain T along any 4-tuple of paths. Using again the forward algorithm, $P(T)$ can be computed as $\sum_{q \in Q_n^4} p(n; q)$, where $p(0; (q^0, q^0, q^0, q^0)) = 1$ and

$$p(j; q) = E(j; q) \sum_{q' \in Q_{j-1}^4} p(j-1; q') \Gamma(q', q) \tag{6}$$

The time needed to compute $P(T)$ with the standard recurrence is $O(nK^8)$, but a K^3 speed-up can again be achieved by re-using common terms and computing, in order:

- $s_1(j; q_1, q_2', q_3', q_4') = \sum_{q_1' \in Q_{j-1}} p(j-1; (q_1', q_2', q_3', q_4')) \gamma(q_1', q_1)$ for each $(q_1, q_2', q_3', q_4') \in Q_j \times Q_{j-1}^3$
- $s_2(j; q_1, q_2, q_3', q_4') = \sum_{q_2' \in Q_{j-1}} s_1(j; (q_1, q_2', q_3', q_4')) \gamma(q_2', q_2)$ for each $(q_1, q_2, q_3', q_4') \in Q_j^2 \times Q_{j-1}^2$
- $s_3(j; q_1, q_2, q_3, q_4') = \sum_{q_3' \in Q_{j-1}} s_2(j; (q_1, q_2, q_3', q_4')) \gamma(q_3', q_3)$ for each $(q_1, q_2, q_3, q_4') \in Q_j^3 \times Q_{j-1}$

- $p(j;q) = E(j;q) \sum_{q'_4 \in Q_{j-1}} s_3(j;(q_1,q_2,q_3,q'_4))\gamma(q'_4,q_4)$ for each $q = (q_1,q_2,q_3,q_4) \in Q_j^4$

This allows computing $P(T|M)$ in $O(nK^5)$ time. By using a forward-backward algorithm, we can obtain within the same time bound all likelihood ratios for the SNP genotypes in the trio T. For unrelated individuals the runtime reduces to $O(nK^3)$ per individual.

7 Experimental results

7.1 Experimental setup

HMM-based genotype error detection algorithms using the three likelihood functions described in Section 6 were implemented in C++. We tested the performance of our methods on both synthetic datasets and a real dataset obtained from (Becker et al., 2006). Synthetic datasets were generated as follows. We started from the real dataset in (Becker et al., 2006), which consists of 551 trios genotyped at 35 SNP loci spanning a region of 91,391 base pairs from chromosome 16. The FAMHAP software (Becker & Knapp, 2004) was used to estimate the frequencies of the haplotypes present in the population. The 705 haplotypes that had positive FAMHAP estimated frequencies were used to derive synthetic datasets with 30-551 trios as follows. For each trio, four haplotypes were randomly picked by random sampling from the estimated haplotype frequency distribution. Two of these haplotypes were paired to form the mother genotype, and the other two were paired to form the father genotype. We created child genotypes by randomly picking from each parent a transmitted haplotype (assuming that no recombination is taking place). To make the datasets more realistic, missing data was inserted into the resulting genotypes by replicating the missing data patterns observed in the real dataset.

Finally, errors were inserted to the genotype data using four models simulating error types generated by commonly used genotyping technologies (Douglas et al., 2000):

- *Random allele model.* Under this model, we selected each (trio, SNP locus) pair with a probability of δ (δ was set to 1% in our experiments). For each selected pair, we picked uniformly at random one of the non-missing alleles and flipped its value.

- *Random genotype model.* Again, we selected each (trio, SNP locus) pair with probability δ. For each

selected pair, we picked uniformly at random one of the non-missing SNP genotypes and replaced it at random with one of the two other possible SNP genotypes, according to the expected Hardy-Weinberg equilibrium genotype frequencies (p^2, q^2, respectively $2pq$ for 0, 1, and 2 genotypes, where p is the estimated probability of allele 0 and $q = 1 - p$).

- *Heterozygous-to-homozygous model.* Each heterozygous SNP genotype was selected with probability δ, and selected genotypes were replaced with equal probability by one of the two homozygous SNP genotypes.

- *Homozygous-to-heterozygous model.* Each homozygous SNP genotype was replaced by the heterozygous SNP genotype with probability δ.

7.2 Results on synthetic datasets

Following the standard practice, we first removed the trivially detected MI errors by marking child SNP genotypes involved in MIs as missing (similar results were obtained by marking all three SNP genotypes as missing). To assess error detection accuracy of different methods in a threshold-independent manner we use receiver operating characteristic (ROC) curves, i.e., plots of achievable sensitivity vs. false positive rates, where

- the *sensitivity* is defined as the ratio between the number of Mendelian consistent errors flagged by the algorithm and the total number of Mendelian consistent errors inserted; and

- the *false positive rate* is defined as the ratio between the number of non-errors flagged by the algorithm and the total number of non-errors.

Figure 3 gives ROC curves for detection algorithms based on the three likelihood functions described in Section 6. These results are based on averages over 10 synthetic instances of 551 trios typed at 35 SNP loci, with errors inserted using the random allele model with $\delta = 1\%$. Since the detection accuracy achieved by the three likelihood functions is very similar in both parents and children, for the remaining experiments we use only the total trio genotype probability.

Figure 3 belongs here.

It is well known that there is an asymmetry in the amount of information gained from trio genotype data about children and parent haplotypes: while each of the two child haplotypes are constrained to be compatible with two genotypes, only one of the parent haplotypes has the same degree of constraint. This asymmetry is known to make errors in children more likely to result in MIs (Douglas et al., 2002; Gordon et al., 1999). As shown by the ROC curves in Figure 3, the asymmetry also leads to significantly higher detection sensitivity in children versus parents.

Figure 4 belongs here.

Figure 4 shows a different view of the asymmetry between children and parents. The top two histograms show the distributions of log-likelihood ratios (computed using the total trio genotype probability as likelihood function) for error and non-error SNP genotypes in both parents and children. Clearly, the separation between errors and non-errors is much sharper in children than in parents. Surprisingly, the histogram of log-likelihood ratios for non-error SNP genotypes in children also shows a significant peak between 3 and 4. Upon inspection, we found that these SNP genotypes are at loci for which parents have inserted errors. A similar bias towards higher false positive rates in correctly typed SNP genotypes for which the same locus is mistyped in related individuals has been noted for other pedigree-based error detection algorithms (Mukhopadhyaya et al., 2004). Since such a peak is not present in the distributions of log-likelihood ratios computed based on child-parent duos (see Figure 4), this suggests that reducing the above bias can be done by combining likelihood ratios computed for different subsets of trio members. We devised such a combined approach, referred to as TotalProb-Combined, whereby for each SNP genotype under test we compute three likelihood ratios using the total probability of (a) the trio genotype, (b) the duo genotypes formed by parent-child pairs, and (c) the individual's multi-locus genotype by itself. Likelihood ratios (b) and (c) can be computed without increasing the asymptotic running time via simple modifications of the algorithm in Section 6.3. A SNP genotype is then flagged as a potential error only if *all* above likelihood ratios exceed the detection threshold.

Figure 5 belongs here.

Figure 5 shows the ROC curves for TotalProb-Combined and flagging algorithms that use single log-likelihood ratios computed from the total probability of uno/duo/trio genotypes. We also included ROC curves for two versions of the algorithm of (Becker et al., 2006), which test one SNP genotype at a time (FAMHAP-1) or simultaneously test the mother/father/child SNP genotypes at a locus (FAMHAP-3). The

results show that simultaneous testing yields low detection accuracy, particularly in parents, and it is therefore not advisable. The combined algorithm yields the best accuracy of all compared methods. The improvement over the trio-based version is most significant in parents, where, surprisingly, uno and duo log-likelihood ratios appear to be more informative than the trio log-likelihood ratio. Figure 6 belongs here.

In next simulation experiments we attempted to quantify the robustness of TotalProb-Combined to changes in error type, sample size, and SNP density. Figure 6(a) gives ROC curves obtained by TotalProb-Combined on datasets generated using the four error models described in Section 7.1. The results show that TotalProb-Combined has high detection accuracy regardless of the error model. Indeed, detection accuracy seems to depends very little on the error model, with the largest difference arising between heterozygous-to-homozygous and random allele errors inserted in parents.

The error detection accuracy of TotalProb-Combined directly depends on the accurate representation of haplotype frequencies by the HMM. The quality of both the ENT phasing and HMM parameter estimation are expected to degrade with decreased sample size. To assess the effect of the number of trios on error detection accuracy we simulated testcases with 30, 129, and 551 trios in which errors were inserted using the random allele model with $\delta = 1\%$. Simulation results for the TotalProb-Combined method are shown in Figure 6(b). While detection accuracy does decrease with sample size, the method does retain high accuracy even for datasets with as few as 30 trios.

Finally, we ran experiments to assess the effect of SNP density on error detection accuracy. All previous results are based on simulated data derived from the real dataset of (Becker et al., 2006), which consists of a very dense (and hence tightly linked) set of 35 SNP loci spanning a region of 91,391 base pairs. We used the GENOME coalescent-based whole genome simulator (Liang et al., 2007) to generate 10 sets of 551 *unrelated* genotypes with 35 SNP loci for each of four different region lengths (10 kilobases, 100 kilobases, 1 megabase, and 10 megabases). All datasets were generated assuming recombination and mutation rates of 10^{-8} per generation per base pair. The ROC curves in Figure 6(c) show that, as expected, error detection accuracy decreases as the density of SNP loci is reduced. Even at comparable SNP density, error detection in unrelated individuals is significantly less accurate compared to parents from trio data. Part of this accuracy loss is explained by the reduced sensitivity of uno-based likelihood ratio tests (already apparent in Figure

5) compared to combined likelihood ratio tests. Remaining accuracy loss is due to the higher ambiguity in haplotype phase of unrelated genotypes compared to trio data, which leads to a less accurate HMM representation of haplotype frequencies. Table 1 belongs here.

7.3 Results on real data from (Becker et al., 2006)

For simplicity, in previous section we used the same detection threshold in both children and parents. However, histograms in Figure 4 suggest that better tradeoffs between sensitivity and false positive rate can be achieved by using differential detection thresholds. For the results on the real dataset from Becker et al. (Becker et al., 2006) (Table 1) we independently picked parent and children thresholds by finding the minimum detection threshold that achieves false positive rates of 0.1-1% under log-likelihood ratio distributions of simulated data.

Unfortunately, for this dataset we do not know all existing genotyping errors. Becker et al. resequenced all trio members at a number of 41 SNP loci flagged by their FAMHAP-3 method with a detection threshold of 10^4. Of the 41×3 resequenced SNP genotypes, 26 (12 in children and 14 in parents) were identified as being true errors, 90 were confirmed as originally correct. The error status of remaining 7 resequenced SNP genotypes is ambiguous due to missing calls in either the original or resequencing data. The "True Positive" columns in Table 1 give the number of TotalProb-Combined flags among the 26 known errors, the "False Positive" columns give the number of flags among the 90 known non-errors, and the "Unknown Signals" columns give the number flags among the 57,739 SNP genotypes for which the error status is not known (since resequencing was not performed or due to missing calls). With a predicted false positive rate of 0.1%, TotalProb-Combined detects 11 out of the 12 known errors in children, and 8 out of the 14 known errors in parents, with only 2 false positives (both in children). TotalProb-Combined also flags 72 SNP genotypes with unknown error status, 61 of which are in parents. We conjecture that most of these are true typing errors missed by FAMHAP-3, which, as suggested by the simulation results in Figure 5, has very poor sensitivity to errors in parent genotypes. We also note that the number of Mendelian consistent errors in parents is expected to be more than twice higher than the number of Mendelian consistent errors in children, due on one hand to the fact that there are twice more parents than children and on the other hand to the

higher probability that errors in parents remain undetected as Mendelian inconsistencies (Douglas et al., 2002; Gordon et al., 1999).

8 Conclusions

In this paper we have proposed high-accuracy methods for detection of errors in trio and unrelated genotype data based on Hidden Markov Models of haplotype diversity. The need for such methods is expected to increase in the future as genotype analysis methods shift towards the use of haplotypes. The runtime of our methods scales linearly with the number of trios and SNP loci, making them appropriate for handling the datasets generated by current large-scale association studies. Our simulation results further indicate the significant increase in detection accuracy when using genotype data for families of related genotypes such as trios. Parent-child relationships are well-known to help disambiguating a significant amount of phase uncertainty by application of simple Mendelian transmission rules. However, our results suggest that the value of incorporating family relationships in analysis methods can go well beyond these "first order" effects. A case in point is the sharp increase observed in children genotype error detection sensitivity due to the use of a trio-based likelihood function. A similar "virtuous cycle" effect was pointed out in ENT phasing accuracy: not only the number of ambiguous positions decreases significantly when phasing related versus unrelated genotypes, but the *relative* phasing accuracy of the algorithm increases significantly as well (Gusev et al., to appear).

In ongoing work we are extending the TotalProb-Combined method to arbitrary pedigrees. We are also exploring the use of locus dependent detection thresholds, methods for assigning p-values to error predictions, and iterative methods which use maximum likelihood to correct MIs and SNP genotypes flagged with a high detection threshold, then recompute log-likelihoods to flag additional genotypes. Finally, we are exploring integration of population-level haplotype frequency information with typing confidence scores for further improvements in error detection accuracy, particularly in the case of unrelated genotype data.

Acknowledgments

We would like to thank the authors of (Becker et al., 2006) for kindly providing us the real dataset used in their paper. This work was supported in part by NSF CAREER award IIS-0546457 and NSF award DBI-0543365.

References

Abecasis, G., Cherny, S., & Cardon, L. (2001). The impact of genotyping error on family-based analysis of quantitative traits. *Eur. J. Hum. Genet.*, *9*, 130–134.

Abecasis, G., Cherny, S., Cookson, W., & Cardon, L. (2002). Merlin-rapid analysis of dense genetic maps using sparse gene flow trees. *Nat Genet*, *30*, 97–101.

Ahn, K., Haynes, C., Kim, W., Fleur, R., Gordon, D., & Finch, S. (2007). The effects of SNP genotyping errors on the power of the Cochran-Armitage linear trend test for case/control association studies. *Ann. Hum. Genet.*, *71*, 249–261.

Baum, L., Petrie, T., Soules, G., & Weiss, N. (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *Ann. Math. Statist.*, *41*, 164–171.

Becker, T., & Knapp, M. (2004). Maximum-likelihood estimation of haplotype frequencies in nuclear families. *Genet. Epidemiol.*, *27*, 21–32.

Becker, T., Valentonyte, R., Croucher, P., Strauch, K., Schreiber, S., Hampe, J., & Knapp, M. (2006). Identification of probable genotyping errors by consideration of haplotypes. *European Journal of Human Genetics*, *14*, 450–458.

Cheng, K. (2007). Analysis of case-only studies accounting for genotyping error. *Ann. Hum. Genet.*, *71*, 238–248.

Cherny, S., Abecasis, G., Cookson, W., Sham, P., & Cardon, L. (2001). The effect of genotype and pedigree error on linkage analysis: Analysis of three asthma genome scans. *Genet. Epidemiol.*, *21*, S117–S122.

Cutler, D., Zwick, M., Carrasquillo, M., C.T.Yohn, Tobin, K., Kashuk, C., Matthews, D., Shah, N., Eichler, E., J.A.Warrington, & Chakravarti, A. (2001). High-throughput variation detection and genotyping using microarrays. *Genome Research*, *11*, 1913–1925.

Di, X., Matsuzaki, H., Webster, T., Hubbell, E., Liu, G., Dong, S., Bartell, D., Huang, J., Chiles, R., Yang, G., Shen, M.-M., Kulp, D., Kennedy, G., Mei, R., Jones, K., & Cawley, S. (2005). Dynamic model based algorithms for screening and genotyping over 100k SNPs on oligonucleotide microarrays. *Bioinformatics*, *21*, 1958–1963.

Douglas, J., Boehnke, M., & Lange, K. (2000). A multipoint method for detecting genotyping errors and mutations in sibling-pair linkage data. *AJHG*, *66*, 1287–1297.

Douglas, J., Skol, A., & Boehnke, M. (2002). Probability of detection of genotyping errors and mutations as inheritance inconsistencies in nuclear-family data. *AJHG*, *70*, 487–495.

Gordon, D., Finch, S., Nothnagel, M., & Ott, J. (2002). Power and sample size calculations for case-control genetic association tests when errors are present: Application to single nucleotide polymorphisms. *Human Heredity*, *54*, 22–33.

Gordon, D., Heath, S., & Ott, J. (1999). True pedigree errors more frequent than apparent errors for single nucleotide poloymorphisms. *Hum. Hered.*, *49*, 65–70.

Gusev, A., Paşaniuc, B., & Măndoiu, I. (to appear). Highly scalable genotype phasing by entropy minimization. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*.

Hao, K., & Wang, X. (2004). Incorporating individual error rate into association test of unmatched case-control design. *Human Heredity*, *58*, 154–163.

Håstad, J. (1999). Clique is hard to approximate within $n^{1-\varepsilon}$. *Acta Mathematica*, *182*, 105–142.

Hosking, L., Lumsden, S., Lewis, K., Yeo, A., McCarthy, L., Bansal, A., Riley, J., Purvis, I., & Xu, C. (2004). Detection of genotyping errors by hardy-weinberg equilibrium testing. *Eur J. Hum Genet*, *12*, 395–399.

Kimmel, G., & Shamir, R. (2005). A block-free hidden Markov model for genotypes and its application to disease association. *Journal of Computational Biology*, *12*, 1243–1260.

Knapp, M., & Becker, T. (2004). Impact of genotyping errors on type I error rate of the haplotype-sharing transmission/disequilibrium test (HS-TDT). *Am. J. Hum. Genet.*, *74*, 589–591.

Leal, S. (2005). Detection of genotyping errors and pseudo-SNPs via deviations from Hardy-Weinberg equilibrium. *Genet Epidemiol*, *29*, 204–214.

Liang, L., Zöllner, S., & Abecasis, G. (2007). GENOME: a rapid coalescent-based whole genome simulator. *Bioinformatics*, *23*, 1565–1567.

Liu, W., Yang, T., Zhao, W., & Chase, G. (2007). Accounting for genotyping errors in tagging SNP selection. *Am. J. Hum. Genet.*, *71*, 467–479.

Lyngsø, R., & Pedersen, C. (2002). The consensus string problem and the complexity of comparing hidden markov models. *J. Comput. Syst. Sci.*, *65*, 545–569.

Marchini, J., Spencer, C., Teo, Y., & Donnelly, P. (2007). A bayesian hierarchical mixture model for genotype calling in a multi-cohort study. in preparation.

Mitchell, A., Cutler, D., & Chakravarti, A. (2003). Undetected genotyping errors cause apparent overtransmission of common alleles in the transmission/disequilibrium test. *Am, J. Hum Genet*, *72*, 598–610.

Mukhopadhyaya, N., Buxbauma, S., & Weeks, D. (2004). Comparative study of multipoint methods for genotype error detection. *Hum. Hered.*, *58*, 175–189.

NCI-NHGRI Working Group on Replication in Association Studies (2007). Replicating genotype-phenotype associations. *Nature*, *447*, 655–660.

Nicolae, D., Wu., X., Miyake, K., & Cox, N. (2006). Gel: a novel genotype calling algorithm using emperical likelihood. *Bioinformatics*, *22*, 1942–1947.

Pompanon, F., Bonin, A., Bellemain, E., & Taberlet, P. (2005). Genotyping errors: causes, consequences and solutions. *Nat. Rev. Genet.*, *6*, 847–859.

Rabbee, N., & Speed, T. (2005). A genotype calling algorithm for affymetrix SNP arrays. *Bioinformatics*, *22*, 7–12.

Rastas, P., Koivisto, M., Mannila, H., & Ukkonen, E. (2007). Phasing genotypes using a hidden Markov model. In I.I. Măndoiu and A. Zelikovsky (Eds.), *Bioinformatics Algorithms: Techniques and Applications*. Wiley (to appear). Preliminary version in *Proc. WABI 2005*.

Scheet, P., & Stephens, M. (2006). A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *American Journal of Human Genetics*, *78*, 629–644.

Schwartz, R. (2004). Algorithms for association study design using a generalized model of haplotype conservation. *Proc. CSB* (pp. 90–97).

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

Sobel, E., Papp, J., & Lange, K. (2002). Detection and integration of genotyping errors in statistical genetics. *Am. J. Hum. Genet.*, *70*, 496–508.

Viterbi, A. (1967). Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory*, *13*, 260–269.

Xiao, Y., Segal, M., Yang, J., & Y., Y. R.-F. (2007). A multi-array multi-SNP genotyping algorithm for affymetrix SNP microarrays. *Bioinformatics*, *23*, 1459–1467.

Zaitlen, N., Kang, H., Feolo, M., Sherry, S. T., Halperin, E., & Eskin, E. (2005). Inference and analysis of haplotypes from combined genotyping studies deposited in dbSNP. *Genome Research*, *15*, 1595–1600.

Zheng, G., & Tian, X. (2005). The impact of diagnostic error on testing genetic association in case-control studies. *Statistics in Medicine*, *24*, 869–882.

	Total Signals			True Positives			False Positives			Unknown Signals		
FP rate	1%	0.5%	0.1%	1%	0.5%	0.1%	1%	0.5%	0.1%	1%	0.5%	0.1%
Parents	218	127	69	9	9	8	1	0	0	208	118	61
Children	104	74	24	11	11	11	3	3	2	90	60	11
Total	322	201	93	20	20	19	4	3	2	298	178	72

Table 1: Results of TotalProb-Combined on Becker et al. dataset.

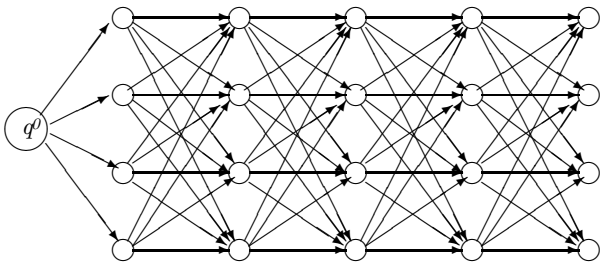
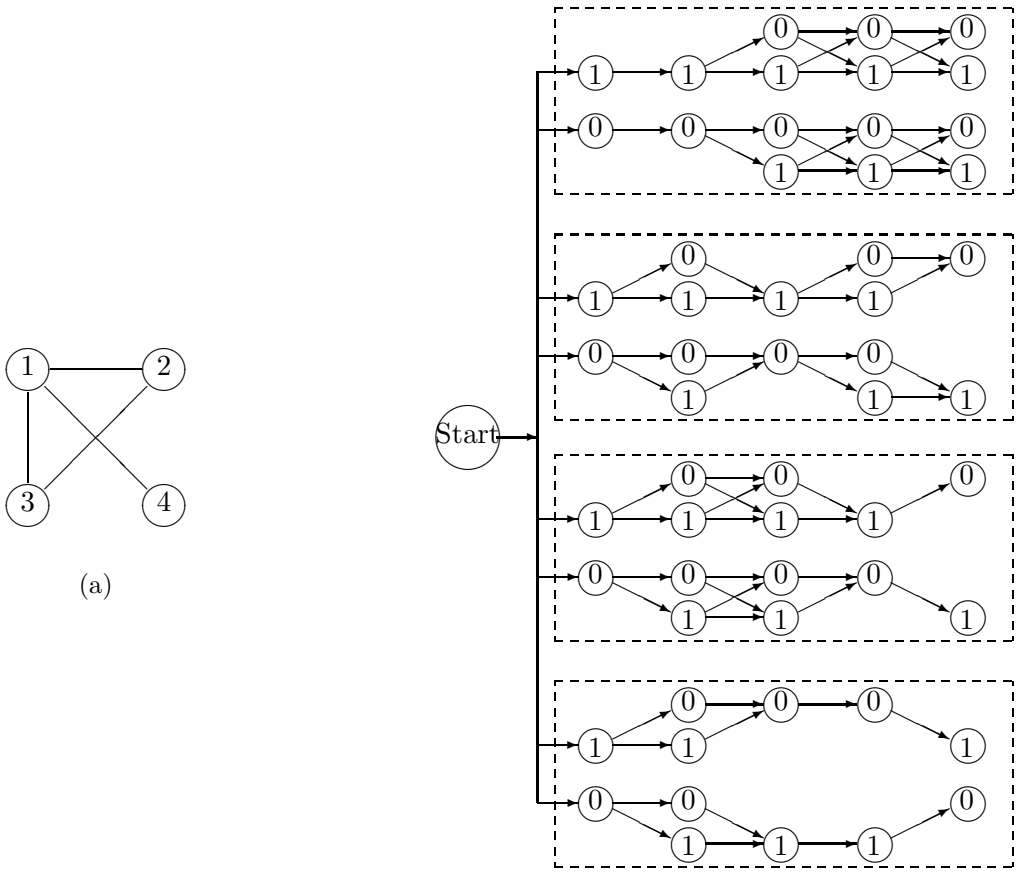


Figure 1: The structure of the Hidden Markov Model for n=5 SNP loci and K=4 founders.



(a)

(b)

Figure 2: A sample graph (a) and the corresponding HMM constructed as in the proof of Theorem 1 (b). The groups of states associated with each vertex are enclosed within dashed boxes. Only states reachable from the start state are shown, with each non-start state labeled by the allele emitted with probability 1.

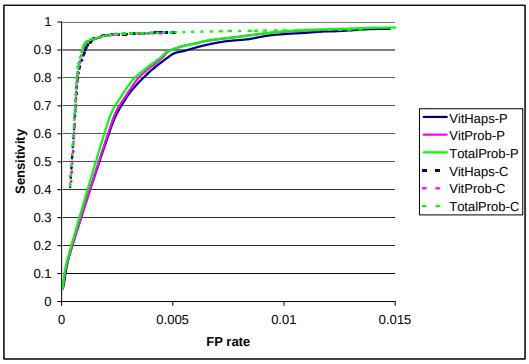


Figure 3: Detection ROC curves for parents (P) and children (C) using the three likelihood functions in Section 6.

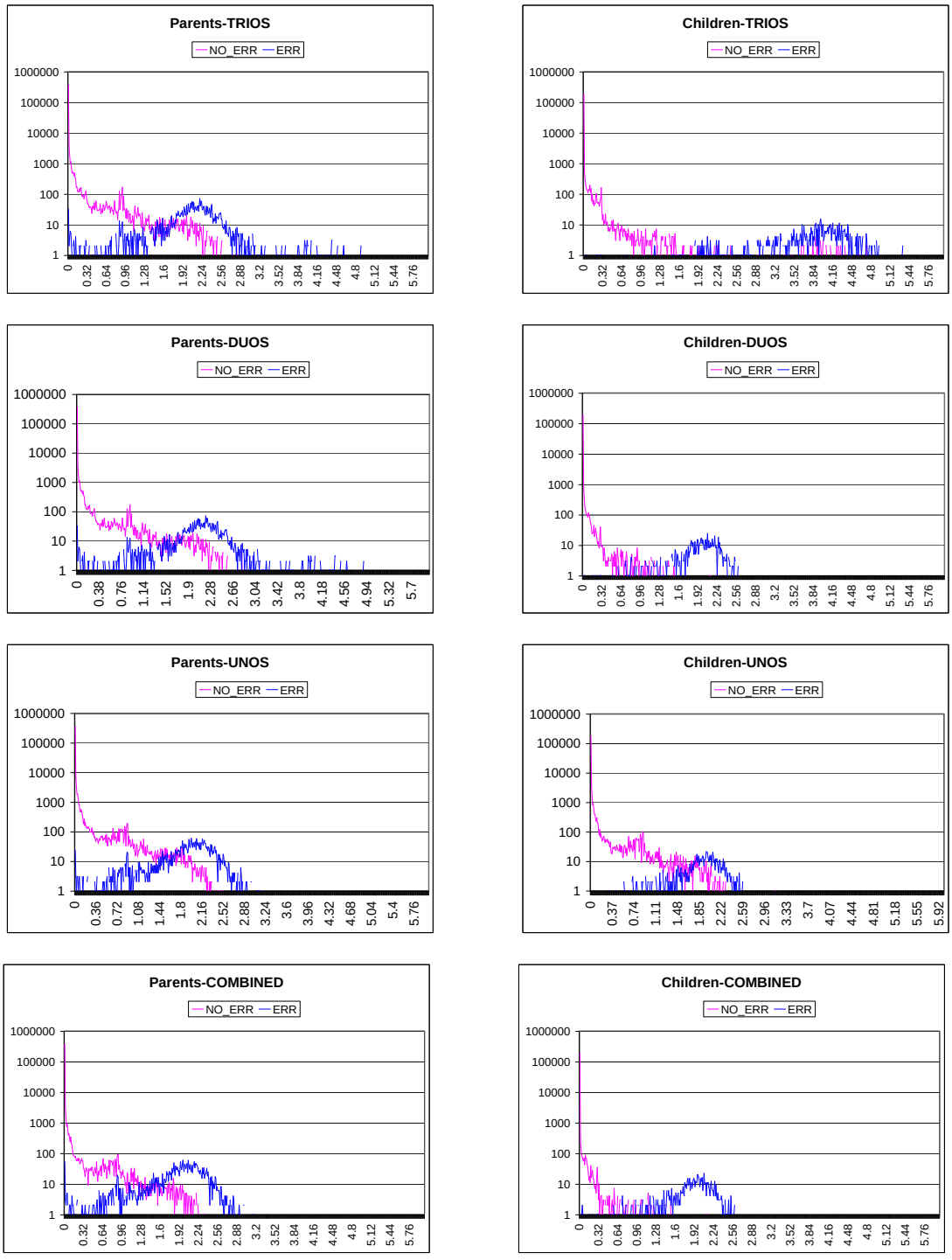


Figure 4: Histograms of log-likelihood ratios for parents (left) and children (right) SNP genotypes, computed based on trios, unos, duos, or the minimum of uno, duo, and trio log-likelihood ratios.

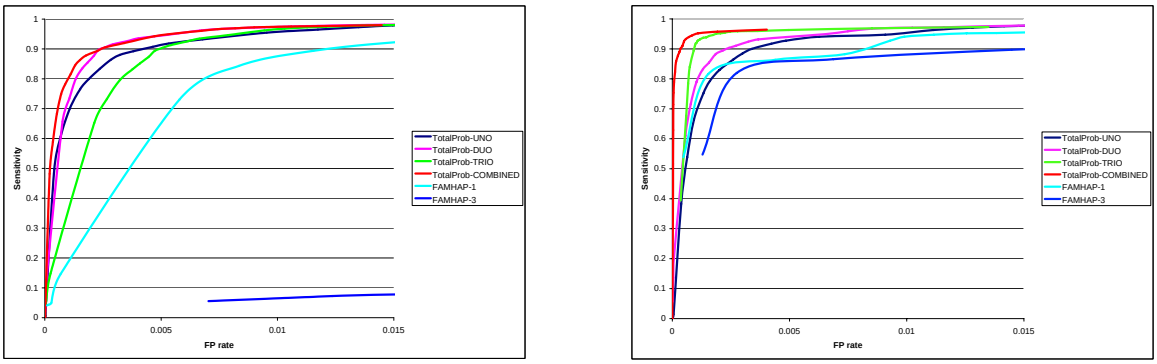
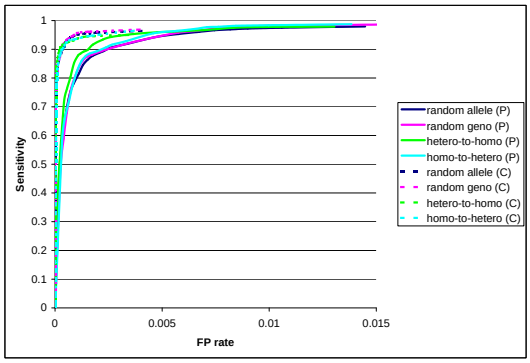
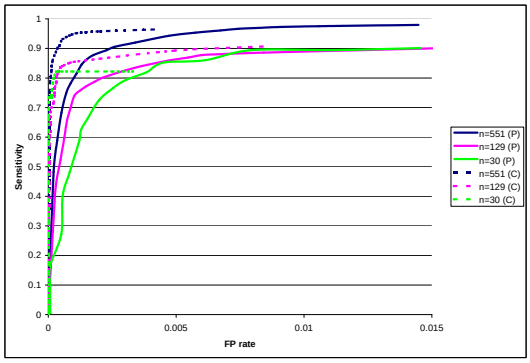


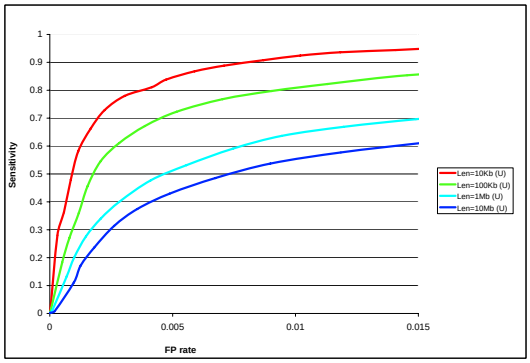
Figure 5: Comparison with FAMHAP accuracy for parents (left) and children (right).



(a)



(b)



(c)

Figure 6: Effect of the error model (a), sample size (b), and SNP density (c) on detection accuracy of TotalProb-Combined.