Genotype Error Detection using Hidden Markov Models of Haplotype Diversity

Justin Kennedy, Ion Măndoiu, and Bogdan Paşaniuc

CSE Department, University of Connecticut, Storrs, CT 06269
{jlk02019,ion,bogdan}@engr.uconn.edu

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 achieve significantly improved detection accuracy compared to previous methods with highly scalable running time.

1 Introduction

Despite recent advances in typing technologies and calling algorithms, significant error levels remain present in SNP genotype data (see [1] for a recent survey). A recent study of dbSNP genotype data [2] 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. When genotype data is available for related individuals, some errors become detectable as *Mendelian inconsistencies* (MIs). However, a large proportion of errors (as much as 70% in mother-father-child trio genotype data [3, 4]) remains undetected by Mendelian consistency analysis.

Since even low error levels can lead to substantial losses in the statistical power of linkage and association studies [5–7], error detection 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 [8].

An indirect approach to handling genotyping errors is to explicitly model them in downstream statistical analyses, see, e.g., [9, 10]. While powerful, this approach often leads to complex statistical models and impractical runtimes for large datasets such as those generated by current large-scale 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 then be excluded from downstream analyses or can be retyped when high quality genotype data is required. Error detection is currently implemented in all widely-used software packages for pedigree genotype data analysis such as SimWalk2 [11] and Merlin [12], which detect Mendelian consistent errors by independently analyzing each pedigree and identifying loci of excessive recombination. Unfortunately, these methods are not appropriate for error detection in genotype data from unrelated individuals or small pedigrees such as mother-father-child trios, which require using population level linkage information.

In this paper we propose novel methods for genotype error detection extending the likelihood ratio error detection approach recently proposed by Becker et al. [13]. While we focus on detecting errors in trio genotype data, our proposed methods can be applied with minor modifications to genotype data coming from unrelated individuals and small pedigrees other than trios Unlike Becker et al., who adopt a window-based approach and rely on creating a short list of frequent haplotypes within each window, we use a hidden Markov model (HMM) to represent frequencies of *all* haplotypes over the set of typed loci. Similar HMMs have been successfully used in recent works [14–17] for genotype phasing and disease association. Two limitations of previous uses of HMMs in this context have been the relatively slow (typically EM-based) training on genotype data, and the inability to exploit available pedigree information. We overcome these limitations by training our HMM on haplotypes inferred using the pedigree-aware ENT phasing algorithm of [18], based on entropy minimization.

Becker et al. [13] use the maximum phasing probability of a trio genotype as the likelihood function whose high 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. When all haplotype frequencies are implicitly represented using an HMM, we show that computing the maximum trio phasing probability is in fact hard to approximate in polynomial time. Despite this result, we are able to significantly improve both detection accuracy and speed compared to [13] by using alternate likelihood functions such as Viterbi probability and the total trio genotype probability. We show that these alternate likelihood functions can be efficiently computed for small pedigrees such as trios, with a worst-case runtime increasing linearly in the number of SNP loci and the number of trios. Further improvements in detection accuracy 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 and describe the structure of the HMM used to represent haplotype frequencies in Section 3. Then, in Section 4 we present the likelihood ratio framework for error detection, and in Section 5 we describe three likelihood functions that can be efficiently computed using the HMM. Finally, we give experimental results assessing the error detection accuracy of our methods on both simulated

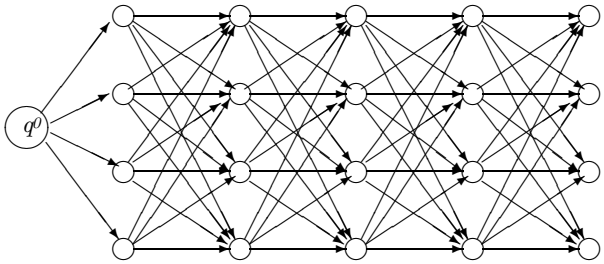

Fig. 1. The structure of the Hidden Markov Model for n=5 SNP loci and K=4 founders.

and real datasets in Section 6, and conclude with ongoing research directions in Section 7.

2 Preliminaries

We start by introducing the basic definitions 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 $\sigma = \sigma' = g$ when $g \in \{0,1\}$, or $\sigma \neq \sigma'$ when $g = 2$.

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* $T = (G_m, G_f, G_c)$ consists of multi-locus genotypes for the mother, father, and child of a nuclear family. An ordered 4-tuple (H_1, H_2, H_3, H_4) of haplotypes is said to explain a trio $T = (G_m, G_f, G_c)$ iff (H_1, H_2) explains G_m, (H_3, H_4) explains G_f, and (H_1, H_3) explains G_c.

3 Hidden Markov Model

The HMM used to represent haplotype frequencies has a similar structure to HMMs recently used in [14–17] (see Figure 1). This structure 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. The initial state q^0 is a silent state, 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:

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

In [14, 15], similar HMMs were trained from genotype data using 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 the HMM. First, we use a highly scalable algorithm based on entropy minimization [18] to infer haplotypes for all individuals in the sample. The phasing algorithm can handle genotypes related by arbitrary pedigrees, and has been shown to yield high phasing accuracy as measured by the so called *switching error*. In the second step we use the classical Baum-Welch algorithm to train the HMM based on the inferred haplotypes.

4 Likelihood ratio approach to error detection

Our detection methods are based on the likelihood ratio approach of Becker et al. [13]. 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 $G_x(i)$ will be flagged as a potential error whenever the corresponding likelihood ratio exceeds a user specified *detection threshold* t.

The likelihood function used by Becker et al. [13] is the maximum trio phasing probability,

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

where the above maximum is computed over all 4-tuples (H_1, H_2, H_3, H_4) of haplotypes that explain T. 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 all $N = 2^n$ haplotype frequencies are given explicitly, computing $L(T)$ can be trivially done in $O(N^4)$ time. Unfortunately this representation can only be used for a small number n of SNP loci. To maintain practical running time,

Becker et al. [13] adopted a heuristic approach that relies on creating a short list of haplotypes with frequency exceeding a certain threshold (computed using the FAMHAP software package [19]) followed by a pruned search over 4-tuples of haplotypes from the 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 [13]), which makes the approach applicable only for short windows of consecutive SNP loci. This limits the amount of linkage information that could be used in error detection, explaining at least in part the high number of false positives observed in [13] within correctly typed SNP genotypes located in the neighborhood SNP genotypes that are mistyped in the same individual.

The HMM described in previous section provides a compact implicit representation of all haplotype frequencies that can be used for large numbers of SNP loci. The problem of computing $L(T)$ based on the HMM is formalized as follows:

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

$$L(T|M) = \max_{(H_1,H_2,H_3,H_4)} P(H_1|M)P(H_2|M)P(H_3|M)P(H_4|M) \qquad (3)$$

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

Computing $P(H|M)$ for a given haplotype H can be easily done in $O(nK)$ time by using a standard forward algorithm, and thus the probability of any given 4-tuple (H_1, H_2, H_3, H_4) that explains T can also be computed within the same time bound. Unfortunately, as stated in the following theorem whose proof we omit due to space constraints, approximating the HMM-based maximum trio phasing probability is hard under some standard computational complexity assumption.[1]

Theorem 1. *For every $\varepsilon > 0$, $L(T|M)$ cannot be approximated within a factor of $O(n^{\frac{1}{4}-\varepsilon})$ for any $\varepsilon > 0$, unless ZPP=NP.*

In next section we propose alternative likelihood functions that are efficiently computable based on an HMM model of haplotype diversity, even for very large numbers of SNP loci.

[1] A proof similar to that of Theorem 1 shows that, when haplotype frequencies are represented using an HMM, computing the maximum phasing probability for a single multi-locus genotype is hard to approximate within a factor of $O(n^{\frac{1}{2}-\varepsilon})$ for any $\varepsilon > 0$, unless ZPP=NP, thus solving a problem left open in [15].

5 Efficiently computable likelihood functions

In this section we consider three alternatives to the likelihood function used in [13], and describe efficient algorithms for computing them given an HMM model of haplotype diversity.

5.1 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 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

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

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 (4) takes $O(K^4)$ time. A $O(K^3)$ speed-up is achieved by computing, in order:

$$\begin{aligned}
Pre_1(j; q_1, q'_2, q'_3, q'_4) &= \max_{q'_1 \in Q_j} \{V_f(j; (q'_1, q'_2, q'_3, q'_4))\gamma(q'_1, q_1)\} \\
Pre_2(j; q_1, q_2, q'_3, q'_4) &= \max_{q'_2 \in Q_j} \{Pre_1(j; (q_1, q'_2, q'_3, q'_4))\gamma(q'_2, q_2)\} \\
Pre_3(j; q_1, q_2, q_3, q'_4) &= \max_{q'_3 \in Q_j} \{Pre_2(j; (q_1, q_2, q'_3, q'_4))\gamma(q'_3, q_3)\} \\
V_f(j+1; q) &= E(j+1; q) \max_{q'_4 \in Q_j} \{Pre_3(j; (q_1, q_2, q_3, q'_4))\gamma(q'_4, q_4)\}
\end{aligned}$$

for each SNP locus $j = 1, \ldots, n$ and all 4-tuples $(q_1, q'_2, q'_3, q'_4) \in Q_{j+1} \times Q_j^3$, $(q_1, q_2, q'_3, q'_4) \in Q_{j+1}^2 \times Q_j^2$, $(q_1, q_2, q_3, q'_4) \in Q_{j+1}^3 \times Q_j$, respectively $q = (q_1, q_2, q_3, q_4) \in Q_{j+1}^4$. A similar speed-up idea was used in the context of single genotype phasing by Rastas et al. [15].

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.

5.2 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 a standard traceback algorithm. 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 a 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.

5.3 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

$$p(j;q) = \sum_{q' \in Q_j^4} p(j-1;q')\Gamma(q',q) \sum_{(\sigma_1,\sigma_2,\sigma_3,\sigma_4)} \prod_{i=1}^{4} \varepsilon(q_i,\sigma_i) \qquad (5)$$

The second sum in last equation is computed over all 4-tuples $(\sigma_1,\sigma_2,\sigma_3,\sigma_4)$ that explain T's SNP genotypes at locus j. Using the speed-up techniques from Section 5.1, we obtain an overall runtime of $O(nK^5)$ per trio.

6 Experimental results

6.1 Experimental setup

HMM-based genotype error detection algorithms using the three likelihood functions described in Section 5 were implemented in C++. Since the detection accuracy of the three likelihood functions is very similar, we report here accuracy results only for the total trio genotype probability.

We tested the performance of our methods on both synthetic datasets and a real dataset obtained from Becker et al. [13]. Synthetic datasets were generated following the methodology of [13]. We started from the real dataset in [13], which consists of 551 trios genotyped at 35 SNP loci spanning a region of 91,391 base pairs from chromosome 16. The FAMHAP software [19] 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 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.

Errors were inserted to the genotype data using the *random allele model* [20]. Under this model, we selected each (trio, SNP locus) pair with a probability of δ (δ was set to 1% in all our experiments). For each selected pair, we picked uniformly at random one of the non-missing alleles and flipped its value. Similar detection accuracy was obtained in experiments in which we simulated recombination rates of up to 0.01 between adjacent SNPs, and in experiments where errors were inserted using the random genotype, heterozygous-to-homozygous, and homozygous-to-heterozygous error models described in [20].

6.2 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).

Figure 2 shows 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. 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 $\delta = 1\%$.

It is 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 was shown to make errors in children more likely to result in MIs [3, 4]. As shown by the histograms in Figure 2, the asymmetry also results in a much sharper separation between errors and non-errors in children than in parents. Surprisingly, the histogram of log-likelihood ratios for non-error SNP genotypes in children has 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

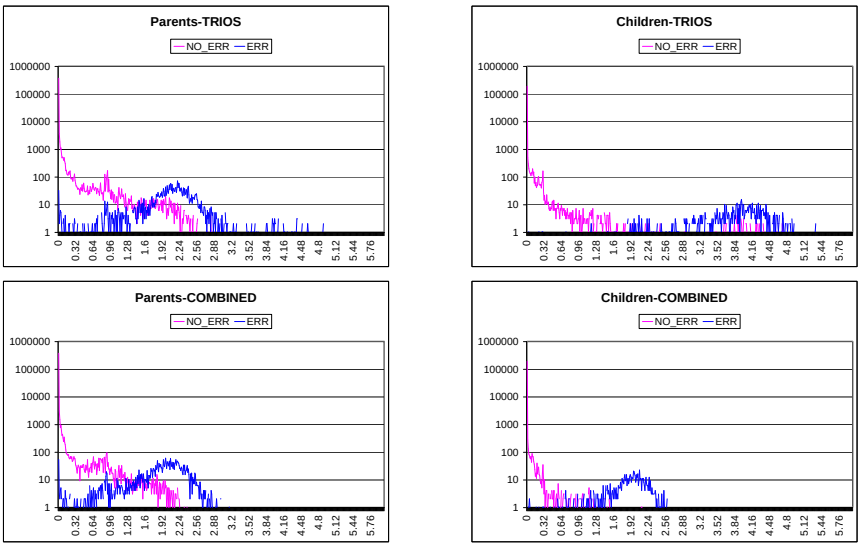

Fig. 2. Histograms of log-likelihood ratios for parents (left) and children (right) SNP genotypes, computed based on trios (top) or by using the minimum of uno, duo, and trio log-likelihood ratios.

other pedigree-based error detection algorithms [21]. To reduce this bias, we propose a simple technique of combining multiple likelihood ratios computed for different subsets of trio members. Under this combined approach, henceforth referred to as TotalProb-Combined, for each SNP genotype we compute 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 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 5.3. A SNP genotype is then flagged as a potential error only if *all* above likelihood ratios exceed the detection threshold.

To assess the accuracy of our error detection methods 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 shows the ROC curves for TotalProb-Combined and for flagging algorithms that use single log-likelihood ratios computed from the total proba-

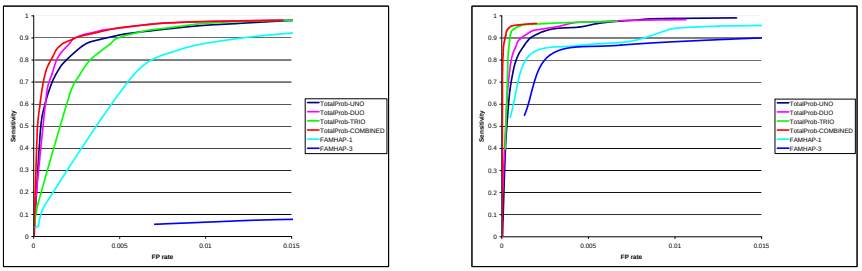

Fig. 3. Comparison with FAMHAP accuracy for parents (left) and children (right).

bility of uno/duo/trio genotypes. We also included ROC curves for two versions of the algorithm of [13], 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.

6.3 Results on real data from [13]

For simplicity, in previous section we used the same detection threshold in both children and parents. However, histograms in Figure 2 suggest that better trade-offs 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. [13] (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

	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.

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 3, 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 [3, 4].

7 Conclusions

In this paper we have proposed high-accuracy methods for detection of errors in trio genotype data based on Hidden Markov Models of haplotype diversity. 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. In ongoing work we are 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 [13] 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

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

2. Zaitlen, N., Kang, H., Feolo, M., Sherry, S.T., Halperin, E., Eskin, E.: Inference and analysis of haplotypes from combined genotyping studies deposited in dbSNP. Genome Research **15** (2005) 1595–1600
3. Douglas, J., Skol, A., Boehnke, M.: Probability of detection of genotyping errors and mutations as inheritance inconsistencies in nuclear-family data. AJHG **70** (2002) 487–495
4. Gordon, D., Heath, S., Ott, J.: True pedigree errors more frequent than apparent errors for single nucleotide poloymorphisms. Hum. Hered. **49** (1999) 65–70
5. Ahn, K., Haynes, C., Kim, W., Fleur, R., Gordon, D., Finch, S.: 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** (2007) 249–261
6. Abecasis, G., Cherny, S., Cardon, L.: The impact of genotyping error on family-based analysis of quantitative traits. Eur. J. Hum. Genet. **9** (2001) 130–134
7. Cherny, S., Abecasis, G., Cookson, W., Sham, P., Cardon, L.: The effect of genotype and pedigree error on linkage analysis: Analysis of three asthma genome scans. Genet. Epidemiol. **21** (2001) S117–S122
8. Knapp, M., Becker, T.: Impact of genotyping errors on type I error rate of the haplotype-sharing transmission/disequilibrium test (HS-TDT). Am. J. Hum. Genet. **74** (2004) 589–591
9. Cheng, K.: Analysis of case-only studies accounting for genotyping error. Ann. Hum. Genet. **71** (2007) 238–248
10. Liu, W., Yang, T., Zhao, W., Chase, G.: Accounting for genotyping errors in tagging SNP selection. Am. J. Hum. Genet. **71**(4) (2007) 467–479
11. Sobel, E., Papp, J., Lange, K.: Detection and integration of genotyping errors in statistical genetics. Am. J. Hum. Genet. **70** (2002) 496–508
12. Abecasis, G., Cherny, S., Cookson, W., Cardon, L.: Merlin-rapid analysis of dense genetic maps using sparse gene flow trees. Nat Genet **30** (2002) 97–101
13. Becker, T., Valentonyte, R., Croucher, P., Strauch, K., Schreiber, S., Hampe, J., Knapp, M.: Identification of probable genotyping errors by consideration of haplotypes. European Journal of Human Genetics **14** (2006) 450–458
14. Kimmel, G., Shamir, R.: A block-free hidden Markov model for genotypes and its application to disease association. Journal of Computational Biology **12** (2005) 1243–1260
15. Rastas, P., Koivisto, M., Mannila, H., Ukkonen, E.: Phasing genotypes using a hidden Markov model. In: Bioinformatics Algorithms: Techniques and Applications, Wiley (to appear, preliminary version in Proc. WABI 2005)
16. Scheet, P., Stephens, M.: 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 (to appear) (2006)
17. Schwartz, R.: Algorithms for association study design using a generalized model of haplotype conservation. In: Proc. CSB. (2004) 90–97
18. Gusev, A., Paşaniuc, B., Măndoiu, I.: Highly scalable genotype phasing by entropy minimization. IEEE Transactions on Computational Biology and Bioinformatics (to appear)
19. Becker, T., Knapp, M.: Maximum-likelihood estimation of haplotype frequencies in nuclear families. Genet. Epidemiol. **27** (2004) 21–32
20. Douglas, J., Boehnke, M., Lange, K.: A multipoint method for detecting genotyping errors and mutations in sibling-pair linkage data. AJHG **66** (2000) 1287–1297
21. Mukhopadhyaya, N., Buxbauma, S., Weeks, D.: Comparative study of multipoint methods for genotype error detection. Hum. Hered. **58** (2004) 175–189