Efficient Algorithms for SNP Genotype Data Analysis using Hidden Markov Models of Haplotype Diversity

Justin Kennedy

University of Connecticut, 2009

Advances in SNP genotyping technologies have played a key role in the proliferation of large scale genomic studies, leading to the discovery of hundreds of genes associated with complex human diseases. Currently, such studies involve genotyping thousands of cases and controls at up to millions of single nucleotide polymorphism (SNP) loci, generating very large datasets that require scalable analysis algorithms. For continued success, efficient algorithms that utilize accurate statistical models and are capable of processing massive amounts of data are needed.

This thesis presents several highly scalable algorithms which utilize Hidden Markov Models (HMMs) of haplotype diversity for SNP genotype data analysis problems. First, we propose novel likelihood functions utilizing these HMMs for the problems of genotype error detection, imputation of untyped SNPs, and missing data recovery. Empirical results show significant improvement when compared to other methods on real and simulated genotype datasets. Next, we contribute a novel method for imputation-based local ancestry inference that effectively exploits Linkage Disequilibrium (LD) information. Experiments on simulated admixed populations show that imputation-based ancestry inference has significantly better accuracy over the best current methods for closely related ancestral populations. Finally, we introduce a hierarchical-factorial HMM which integrates sequencing data with haplotype frequency information and is utilized by efficient decoding algorithms for genotype calling. We demonstrate that highly accurate SNP genotypes can be inferred from very low coverage shotgun using this HMM.

Efficient Algorithms for SNP Genotype Data Analysis using Hidden Markov Models of Haplotype Diversity

Justin Kennedy

M.Sc., Rensselaer, Hartford, CT, USA, 2002

A Dissertation

Submitted in Partial Fulfillment of the

Requirements for the Degree of

Doctor of Philosophy

at the

University of Connecticut

2009

Acknowledgments

My public thanks is first given to my advisor professor Ion Măndoiu for his invaluable advice and assistance throughout many years of dedicated mentoring. This work would not have been possible without his constant guidance. His ability to generate the absolute best in me has been inspiring and has affected me positively in profound ways, and I will always be grateful.

I also would like to thank the support of my advisory committee, which include professors Ion Măndoiu, Yufeng Wu, Sanguthevar Rajasekaran, and Dong-Guk Shin. Many thanks also go to my collaborators from which this thesis is based. Among these include Ion Măndoiu, Yufeng Wu, Bogdan Paşaniuc, Jorge Duitama, Sanjiv Dinakar, and Yozen Hernandez.

I would like to thank the authors of [8] for kindly providing us the real dataset used in their paper. I also would like to thank Genetic Association Information Network (GAIN) for the IMAGE dataset. This work was also supported in part by NSF awards IIS-0546457 and IIS-0916948.

Thanks also go to my friends who have been patient with me as I have sacrificed time with them, and yet regardless of this, have stuck by me. Finally, I would also like to thank all of my family. Specifically, I am most appreciative to all my parents and grandparents for their founding wisdom and diverse ways of support. Furthermore, my siblings have been very loyal and supportive, as we make this journey through the world in our generation. And gratitude is deserved to my children, as because of them, every effort in my life is made with the most to look forward to.

This work is dedicated to my wife Becky. This effort has required as much work and even more sacrifice from her as it has from me, and words cannot completely express the gratitude I have for all the love and support she has readily given. This belongs to us, and by everything we grow.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Recently, large scale Genome-Wide Association Studies (GWASs) have been made possible by the sequencing of the human genome [13,14] and the initial mapping of human haplotypes by the HapMap project [78]. Evidence supporting GWASs as being a powerful approach to identifying disease-gene associations is growing. One recent success, for example, was a joint GWAS using a large British population set identifying 24 statistically significant independent association signals across 6 diseases [16]. The genetic markers of choice in GWASs are Single Nucleotide Polymorphisms (SNPs), which account for most of the genomic variation in humans. A SNP is a single base pair mutation, or a variance in DNA nucleotides between organisms of the same species, at certain locations, called markers, along the chromosome. All possible nucleotides which exist for a large percentage of the population for a specific marker are called alleles. Most SNPs are biallelic, i.e., only two nucleotide variations are known to exist at that SNP marker. According to NCBI, the human genome consists of nearly thirteen million common SNPs, which have been cataloged in the most recent build (130) of the dbSNP database (`http://www.ncbi.nlm.nih.gov/projects/SNP/`).

In diploid organisms such as humans, cells contains two copies of each chromo-

some, one for each parent. The combination of alleles present at SNP markers on one parental chromosome is called a haplotype. The conflated allele information from the two haplotypes is called a genotype sequence. However, while the genotype sequence identifies the two alleles at each SNP marker, it does not specify which allele is assigned to a specific chromosome.

This thesis presents several highly scalable algorithms that utilize Hidden Markov Models (HMMs) of haplotype diversity which capture SNP information such as Linkage Disequilibrium (LD) observed in the population or populations under study. The algorithms are applied to address the problems of genotype error detection and correction, imputation-based local ancestry inference, and genotype calling with low coverage shotgun sequencing data.

The validity of associations uncovered in GWASs critically depends on the accuracy of genotype data. Despite recent progress in genotype calling algorithms [17,18,51,57,63,83], 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. Since even low error levels can lead to inflated false positive rates and substantial losses in the statistical power of linkage and association studies [4, 2, 12, 28, 53, 86], detecting and correcting genotype errors remains a critical task in genetic data analysis. Further, since causal SNPs are unlikely to be typed directly due to the limited coverage of current genotyping platforms, imputation of genotypes at untyped SNP loci has recently emerged as a powerful technique for increasing the power of association studies. Chapter 2 of this thesis proposes novel methods for genotype error detection which extends the likelihood ratio error detection approach of [8]. While we focus on detecting errors in parents-child trio genotype data, our proposed methods apply with minor modifications to genotype data coming from unrelated individuals and small pedigrees other than trios. Unlike previous approaches to genotype error detection [8], which use enumeration of

common haplotypes within a small window around each locus, we employ a hidden Markov model (HMM) to represent frequencies of all possible haplotypes over the set of typed loci. Empirical results shows significant improvement for both error accuracy and speed when compared to other methods on real and simulated datasets. The error detection approach can also be modified to address the problems of imputation of untyped SNP markers and missing genotype data recovery. We conclude the chapter by introducing GEDI, a software package that implements efficient algorithms for performing several common tasks in the analysis of population genotype data, including error detection and correction, imputation of both randomly missing and untyped genotypes, and genotype phasing.

One type of powerful tool used in disease-gene association studies that has emerged is admixture mapping. Admixture mapping relies on genotyping hundreds of thousands of single nucleotide polymorphisms (SNPs) across the genome in a population of recently admixed individuals and is based on the assumption that near a disease-associated locus there will be an enhanced ancestry content from the population with higher disease prevalence. Therefore, a critical step in admixture mapping is to obtain accurate estimates of local ancestry around each genomic locus. Chapter 3 contributes a novel method for imputation-based local ancestry inference that effectively exploits LD information using HMMs of haplotype diversity. While there are several methods that use a detailed HMM of LD (e.g. SABER [75], SWITCH [66],HAPAA [74]), surprisingly, existing methods that do not exploit LD outperform those that do. This second class of methods (e.g. LAMP [67] and WINPOP [60]) employs a window-based framework to achieve increased accuracy, however these methods differ from each other in the type of information ultimately used to make local ancestry inferences. Our novel method for imputation-based local ancestry inference more effectively exploits LD information by combining these two classes of methods. Our method employs multiple

HMMs trained on a set of ancestral haplotypes for each population in the admixture study to impute genotypes at all typed SNP loci (temporarily marking each SNP genotype as missing) under each possible local ancestry. We then assign to each locus the local ancestry that yields the highest imputation accuracy, as assessed using a weighted-voting scheme based on multiple SNP windows centered on the locus of interest. Preliminary experiments on simulated admixed populations show that imputation-based ancestry inference has accuracy competitive with best existing methods in the case of distant ancestral populations, and is significantly more accurate for closely related ancestral populations.

Recent massively parallel sequencing technologies deliver orders of magnitude higher throughput compared to classic Sanger sequencing. Sequencers like Roche/454 FLX Titanium, Illumina Genome Analizer II, ABI SOLiD 3 and Helicos HeliScope are able to provide millions of short reads in a single run which lasts just a few days in some cases and even less than one day in other cases. These advances promise to enable cost-effective shotgun sequencing of individual genomes. After recent publication of five complete individual genomes [9,22,52,62,80,23], ongoing efforts focus on increasing the quality and coverage of short reads and on improving algorithms for mapping, genotyping and variations discovery to sequence over a thousand more individual genomes [1]. While shotgun sequencing can discover new SNPs and other forms of sequence variation, its sensitivity of detecting heterozygous SNPs is limited by coverage depth. Chapter 4 demonstrates that highly accurate SNP genotypes can be inferred from very low coverage shotgun sequencing data by using a multilocus inference model that also exploits linkage disequilibrium (LD) information from HMMs of haplotype diversity. While shotgun sequencing can discover new SNPs and other forms of sequence variation, its sensitivity of detecting heterozygous SNPs is limited by coverage depth. It was estimated that a coverage depth of over 21× is required to achieve 99% sensitiv-

ity at detecting heterozygous SNPs based on the rule that each allele must be covered by two or more reads [23]. A coverage depth similar to that in [22, 23] (7.5×) detects only 75% of the heterozygous SNPs, and sensitivity drops rapidly at even lower coverage depths. The hierarchical-factorial HMM (HF-HMM) introduced in this chapter enables the integration of shotgun sequencing data with haplotype frequency information extracted from a reference panel. Efficient decoding algorithms for genotype calling that utilize this HF-HMM are introduced. Experimental results show that our algorithms achieve significant improvements in accuracy compared to previous methods. Based on publicly available reads from three different sequencing technologies, we show that we can achieve more than 95% accuracy on heterozygous SNP calls and more than 99% accuracy on homozygous SNP calls with just 5x coverage depth. Moreover, our proposed algorithms have a linear run time on the number of SNP loci and individuals to be analyzed.

Finally, a current status of this research, as well as a list of several improvements which build upon this research, is provided and will be explored as outlined in the concluding chapter.

Chapter 2

Genotype Error Detection Using Hidden Markov Models of Haplotype Diversity[1]

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 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 [17, 18, 51, 57, 63, 83], 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 [61] for a recent survey. A recent study of dbSNP genotype data [84] found that as much as 1.1% of about 20 million SNP genotypes typed multiple times have inconsistent calls, and are thus

[1]The results presented in this chapter are based on joint work with I. Mandoiu and B. Pasaniuc [38, 36, 37]

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 [55] 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 [32, 44], 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 [20, 29]. Since even low error levels can lead to inflated false positive rates and substantial losses in the statistical power of linkage and association studies [4, 2, 12, 28, 53, 86], 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 [42].

A powerful approach of dealing with genotyping errors is to explicitly model them in downstream statistical analyses, see, e.g., [11, 31, 49]. 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 [72], Merlin [3], Sibmed [19], and SimWalk2 [71, 72], 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. [8] 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.

2.1 Preliminaries

In this chapter we propose novel methods for genotype error detection extending the likelihood ratio error detection approach of [8]. 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 [8], 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 [41,64,68,69] 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 [30], based on entropy minimization.

The authors of [8] 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 [8] 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 chapter is organized as follows. We introduce basic notations in Section 2.1 describe the structure of the HMM used to represent haplotype frequencies in Section 2.2.1, and present the likelihood ratio approach of [8] in Section 2.2.2. In Section 2.2.3, we show that while the likelihood function in [8] cannot be approximated efficiently when an HMM is used to represent haplotype frequencies, 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 2.4.1, and conclude with ongoing research directions in Section 2.5.

We start by introducing basic terminology and notations used throughout the chapter. 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

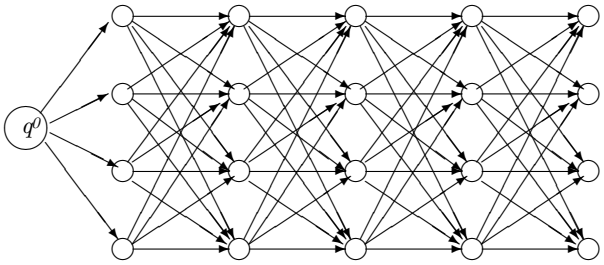


Figure 2.1: The structure of the Hidden Markov Model for n=5 SNP loci and K=4 founders.

$\sigma \neq \sigma'$.

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.

2.2 Methods

2.2.1 Hidden Markov Model

The HMM used to represent haplotype frequencies has a similar structure to HMMs recently used in [41,51,64,68,69]. This structure (see Figure 2.1) 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 $M = (Q, \gamma, \epsilon)$, 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 $\epsilon(q, \sigma)$. 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:

$$P(H, \pi|M) = \gamma(q^0, \pi(1))\epsilon(\pi(1), H(1))\prod_{i=2}^{n}\gamma(\pi(i-1), \pi(i))\epsilon(\pi(i), H(i)) \quad (2.1)$$

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

As noted above, the structure of our HMM is similar to that of other models proposed in the literature. However, there are also important differences. The model underlying the IMPUTE algorithm described in [51] defines HMM states at each SNP locus directly from reference haplotypes (thus, for N reference haplotypes there are N HMM states at each locus). Under the IMPUTE model the probability of switching from one reference haplotype to another is derived from the genetic distance between loci and the effective size for the population under study, and *does not depend* on the states (haplotypes) between which the transition occurs. Similar to our use of founder haplotypes, the model underlying the fastPHASE algorithm [68] reduces the number of HMM states by using at each SNP locus a state for each one of K *clusters* of reference haplotypes, where K is a user-

specified parameter. For each SNP locus, the fastPHASE model estimates K different different transition probabilities, with all transitions *into* a cluster being given an equal probability. As detailed above, our HMM model allows transition probabilities to depend on *both* the start and the end states (founder haplotypes), potentially providing more expressive power compared to the models in [68] and [51].

In HMMs nearly identical to our own [41, 64], training was accomplished 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 [30] 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 [6] to train the HMM based on the haplotypes inferred by ENT.

2.2.2 Likelihood Ratio Approach to Genotype Error Detection

Our detection methods are based on the likelihood ratio approach of [8]. 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. [8] is the maximum trio phasing probability,

$$L(T) = \max_{(H_1,H_2,H_3,H_4)} P(H_1)P(H_2)P(H_3)P(H_4) \tag{2.2}$$

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, [8] 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 [8] – 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 [8] 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 2.2.3 we propose alternative likelihood functions that are efficiently computable based on an HMM representation of haplotype frequencies.

2.2.3 Efficiently Computable Likelihood Functions

As noted in [36], Maximum genotype phasing probability cannot be approximated within a factor of $O(n^{\frac{1}{2}-\varepsilon})$. Also, for trios, maximum trio phasing probability cannot be approximated within a factor of $O(n^{\frac{1}{4}-\varepsilon})$ for any $\varepsilon > 0$, unless ZPP=NP. In this section we consider three alternatives to the likelihood function used in [8], and describe efficient algorithms for computing them given an HMM model of haplotype diversity. for any $\varepsilon > 0$, unless ZPP=NP. As shown in Section 2.4.1, all three alternatives yield similar error detection accuracy, significantly higher than that obtained in [8].

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

(2.1). 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 [79] 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

$$V_f(j; q) = E(j; q) \max_{q' \in Q_{j-1}^4} \{V_f(j-1; q')\Gamma(q', q)\} \tag{2.3}$$

Here, $E(j; q) = \max_{(\sigma_1, \sigma_2, \sigma_3, \sigma_4)} \prod_{i=1}^4 \epsilon(q_i, \sigma_i)$, 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 (2.3) takes $O(K^4)$ time. A K^3 speed-up is obtained by identifying and re-using common terms between the maximums (2.3) corresponding to different 4-tuples q. Thus, instead of applying (2.3) 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 [64].

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.

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.

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

$$p(j; q) = E(j; q) \sum_{q' \in Q_{j-1}^4} p(j-1; q') \Gamma(q', q) \tag{2.4}$$

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.

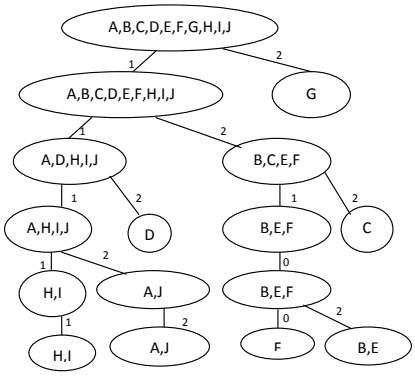
2.2.4 Trie Speed-up

For a dataset consisting of m unrelated samples (i.e., multi-locus genotypes of unrelated individuals), running the forward-backward algorithm independently on each sample results in a runtime of $O(mnK^3)$, where n is the number of SNP loci, and K is the number of founder haplotypes. However, due to the relatively limited genotype variation across individuals of the same population, independent processing of the samples leads to repeated computation of forward and backward probabilities corresponding to genotype prefixes (respectively suffixes) shared by multiple genotypes. To avoid this, we build PopTree, which is a prefix tree, or trie, from the given multilocus genotypes (see Fig. 2.2 for an example) and then computes probabilities by performing a preorder traversal of the trie. Specifically, the PopTree data structure for unrelated individuals in a population consists of up to n levels, where each node has up to 3 child edges- one for each possible genotype value $(0,1,2)$. Computation of backward probabilities is sped-up in a similar way using a trie of reversed genotypes.

The speed-up achieved by using the PopTree trie depends on the number and the similarity of the samples, as well as the number of SNP loci. See section 2.4.1 for an experiment which gives an approximate speed-up of 3× when using PopTree.

Ind	SNP1	SNP2	SNP3	SNP4	SNP5
A	1	1	1	2	2
B	1	2	1	0	2
C	1	2	2	0	2
D	1	1	2	2	2
E	1	2	1	0	2
F	1	2	1	0	0
G	2	1	2	1	1
H	1	1	1	1	1
I	1	1	1	1	1
J	1	1	1	2	2

(a)



(b)

Figure 2.2: Sample dataset over 5 SNPs (a) and corresponding trie (b).

2.3 Experimental results

2.3.1 Experimental Setup

HMM-based genotype error detection algorithms using the three likelihood functions described in Section 2.2.3 were implemented in GEDI using C++. We tested the performance of our methods on both synthetic datasets and a real dataset obtained from [8]. Synthetic datasets were generated as follows. We started from the real dataset in [8], which consists of 551 trios genotyped at 35 SNP loci spanning a region of 91,391 base pairs from chromosome 16. The FAMHAP software [7] 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 hap-

lotypes 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 [19]:

- *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.

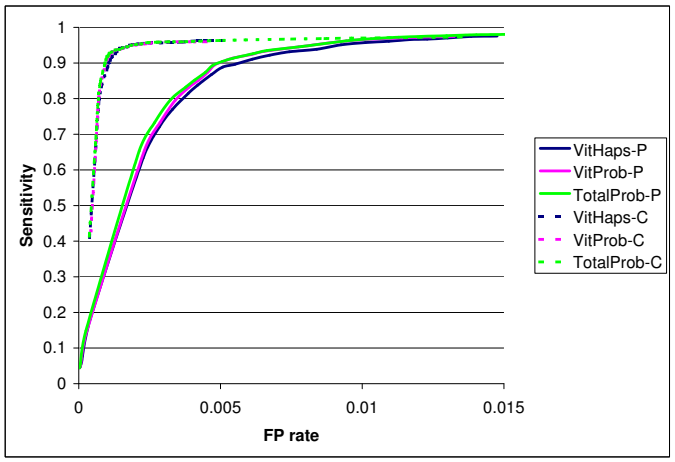


Figure 2.3: Detection ROC curves for parents (P) and children (C) using the three likelihood functions in Section 2.2.3.

2.3.2 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 2.3 gives ROC curves for detection algorithms based on the three likelihood functions described in Section 2.2.3. 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.

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 [20, 29]. As shown by the ROC curves in Figure 2.3, the asymmetry also leads to significantly higher detection sensitivity in children versus parents.

Figure 2.4 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 [54]. Since such a peak is not present in the distributions of log-likelihood ratios computed based on child-parent duos (see Figure 2.4), this suggests that reducing the above bias can be done by combining likelihood ratios computed for different subsets of trio members. We devised such

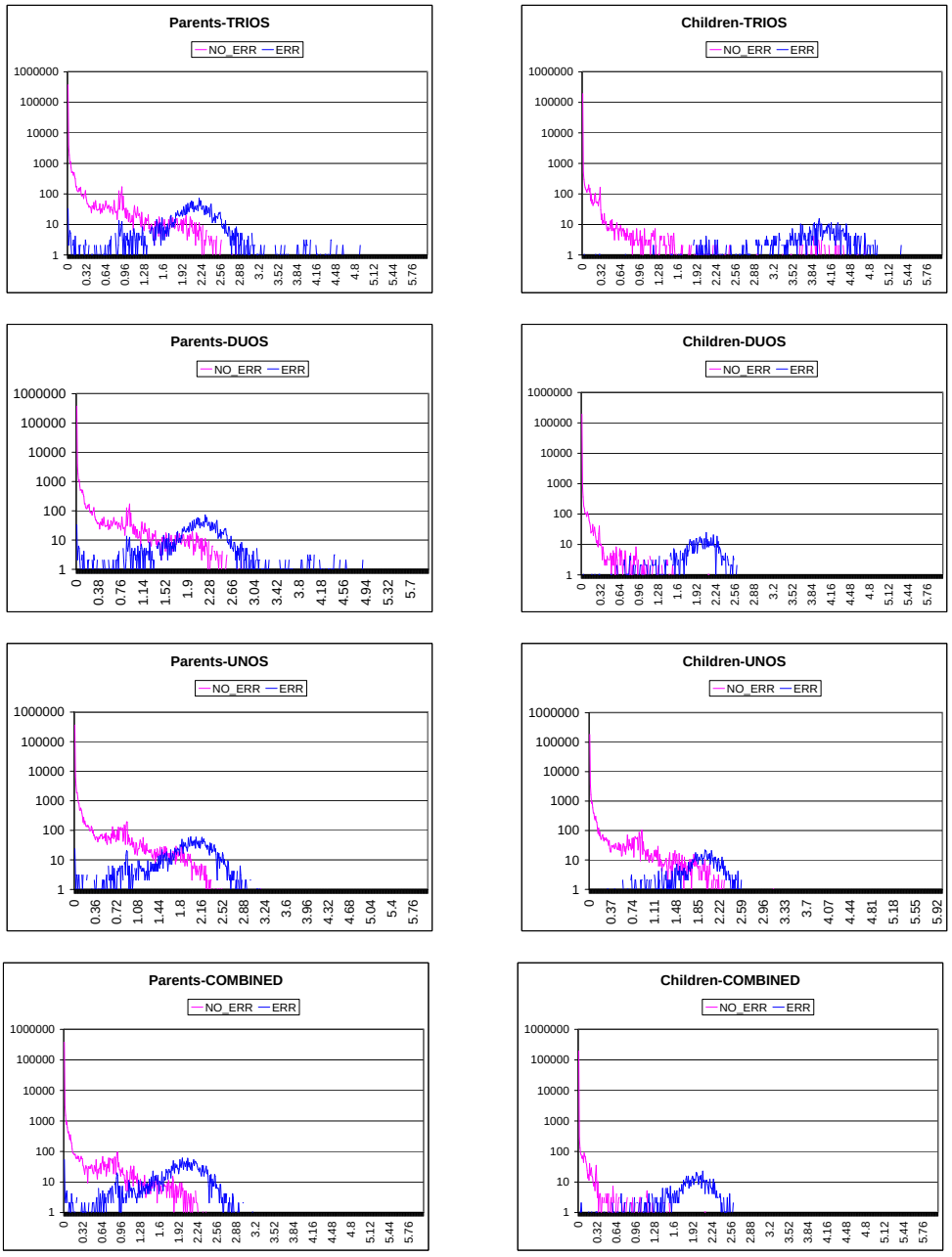


Figure 2.4: Histograms of log-likelihood ratios for parents (left) and children (right) SNP genotypes, computed based on trios, unos, duos, or the minimum of uno, duo, and trio log-likelihood ratios.

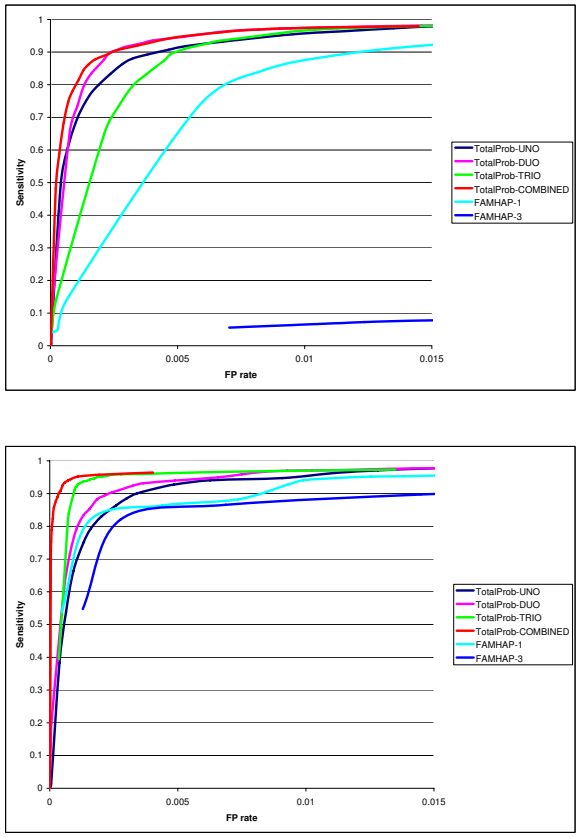


Figure 2.5: Comparison with FAMHAP accuracy for parents (top) and children (bottom).

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 2.2.3. A SNP genotype is then flagged as a potential error only if *all* above likelihood ratios exceed the detection threshold.

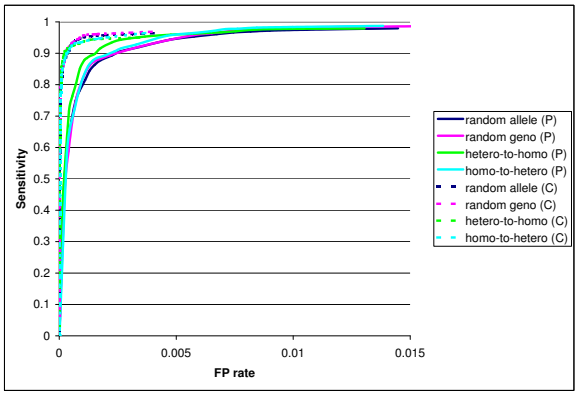
Figure 2.5 shows the ROC curves for TotalProb-Combined and flagging algorithms that use single log-likelihood ratios computed from the total probability of uno/duo/trio genotypes. We also included ROC curves for two versions of the

algorithm of [8], 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.

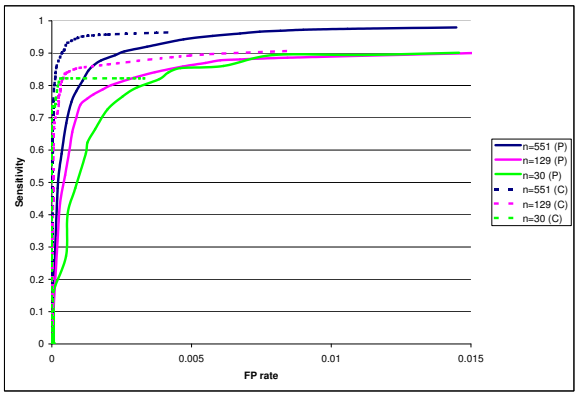
In next simulation experiments we attempted to quantify the robustness of TotalProb-Combined to changes in error type, sample size, and SNP density. Figure 2.6(a) gives ROC curves obtained by TotalProb-Combined on datasets generated using the four error models described in Section 2.3.1. The results 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 test cases 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 2.6(b). While detection accuracy does decrease with sample size, the method does retain high accuracy even for datasets with as few as 30 trios.

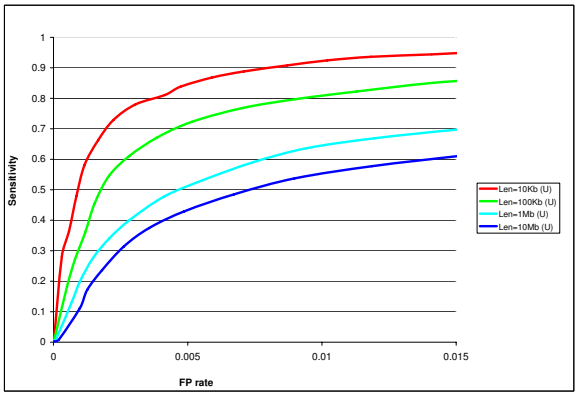
Finally, we ran experiments to assess the effect of SNP density on error detection accuracy. All previous results are based on simulated data derived from the real dataset of [8], which consists of a very dense (and hence tightly linked) set



(a)



(b)



(c)

Figure 2.6: Effect of the error model (a), sample size (b), and SNP density (c) on detection accuracy of TotalProb-Combined.

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

of 35 SNP loci spanning a region of 91,391 base pairs. We used the GENOME coalescent-based whole genome simulator [47] 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 2.6(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 2.5) 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.

2.3.3 Results on Real Data from [8]

For simplicity, in previous section we used the same detection threshold in both children and parents. However, histograms in Figure 2.4 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. [8] (Table 2.1) 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 re-sequencing data. The "True Positive" columns in Table 2.1 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 re-sequencing 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 2.5, 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 [20, 29].

2.4 GEDI Software

In this section we describe GEDI, a software package implementing efficient algorithms for several common tasks in the analysis of GWAS data:

- **Genotype error detection and correction.** GEDI implements the likelihood ratio method described in previous sections, using total genotype probability as likelihood function.

- **Missing data recovery.** High-throughput genotyping platforms leave uncalled large numbers of SNP genotypes. To complement quality control procedures that exclude SNP loci and samples with high proportions of missing genotypes, GEDI provides methods for maximum-likelihood inference of remaining missing genotypes. A missing SNP genotype at a typed SNP locus i is replaced by $\text{argmax}_x P_M(\mathbf{g}[g_i \leftarrow x])$.

- **Imputation of genotypes at untyped SNP loci.** Current genotyping platforms allow simultaneous typing of as many as a million SNP loci, but this is still just a fraction of the polymorphisms present in the human population. Imputation of genotypes at untyped SNP loci based on linkage disequilibrium information extracted from reference panels such as HapMap [34] is often performed to increase statistical power of GWAS studies, see, e.g., [51]. Furthermore, imputation is critical for performing meta-analysis of datasets generated using different platforms [85]. Similar to missing data recovery, imputation of genotypes at an untyped SNP locus is performed at an untyped locus i, and is replaced by $\text{argmax}_x P_M(\mathbf{g}[g_i \leftarrow x])$. Also, in the case of imputation, genotype probabilities are computed based on a "local" HMM model that spans the untyped locus and a user-specified number of typed SNP loci flanking it on each side.

- **Genotype phasing.** Haplotype based association tests can improve statistical power compared to single-SNP approaches, but have seen limited use in the analysis of GWAS data, in part due to the lack of haplotype inference methods that are both accurate and scalable. In an attempt to fill in this gap, GEDI includes an implementation of the highly-scalable phasing algorithm of [30], based on entropy minimization. This algorithm has been recently used by [5] in conjunction with a haplotyping sharing approach to implicate in Parkinson's disease a novel gene missed by traditional single-SNP analyses.

GEDI also handles genotype data of related individuals; in this case imputation probabilities are computed with a simple extension to small pedigrees, and log-likelihood ratios are computed jointly over nuclear families such as trios, as described in the previous chapter.

2.4.1 GEDI Results and Discussion

A comparison of imputation algorithms implemented by GEDI and several other publicly available software packages including [10, 21, 45, 48, 51, 68] is currently underway [26]. Here we present experimental results exploring the effect of GEDI's user-selected parameters on imputation accuracy.

Imputation experiments were performed on the Perlegen 600k genotype data (dbGaP accession number phs000016.v1.p1) generated by the International Multisite ADHD Genetics (IMAGE) project, comprising 958 parents-child trios from seven European countries and Israel. After excluding trios with one or more samples removed by data cleaning steps described in [56], we randomly selected 100 trios and phased them using the entropy minimization algorithm and pooled parental haplotypes with the 120 CEU haplotypes from HapMap release 22 to form a reference panel of 520 haplotypes. The test data consisted of the genotypes of re-

maining 2502 IMAGE individuals, treated as unrelated unless otherwise indicated. Specifically, we masked 9% of the typed SNP loci on chromosome 22 (530 out of 5835), and computed the imputation error rate as the percentage of discordant imputations out of the total of 1,326,060 masked SNP genotypes. In all imputation experiments we used 10 typed SNP loci on each side of masked loci, which, as shown in Fig. 2.7, yields an excellent tradeoff between accuracy and runtime.

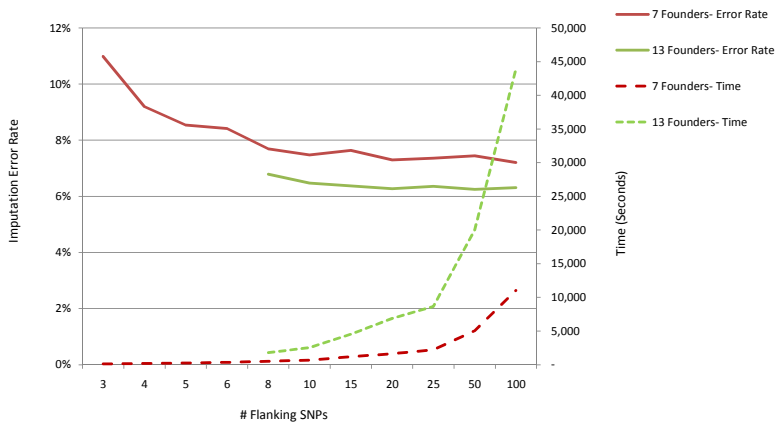


Figure 2.7: Imputation error rate and runtime for varying number of flanking typed SNP loci (IMAGE chr. 22 dataset, 520 training haplotypes).

Table 2.2: Imputation error rate on the IMAGE chr. 22 dataset for varying numbers of HMM founders and training haplotypes.

# Training Haplotypes	# Founders						
	3	5	7	9	11	13	15
30	17.02%	13.65%	13.11%	12.82%	12.27%	12.37%	12.47%
60	15.21%	11.10%	10.00%	9.75%	9.62%	9.59%	9.55%
90	14.82%	10.35%	9.58%	9.04%	8.63%	8.71%	8.57%
120	14.39%	10.11%	8.93%	8.52%	8.30%	8.23%	8.13%
220	13.73%	9.42%	8.28%	7.58%	7.26%	7.27%	7.16%
320	14.31%	9.53%	7.91%	7.37%	6.94%	6.81%	6.78%
420	14.10%	8.82%	7.70%	7.09%	6.75%	6.56%	6.51%
520	13.54%	9.38%	7.48%	6.86%	6.61%	6.47%	6.33%

Fig. 2.8 and Table 2.2 give GEDI imputation accuracy when the number of

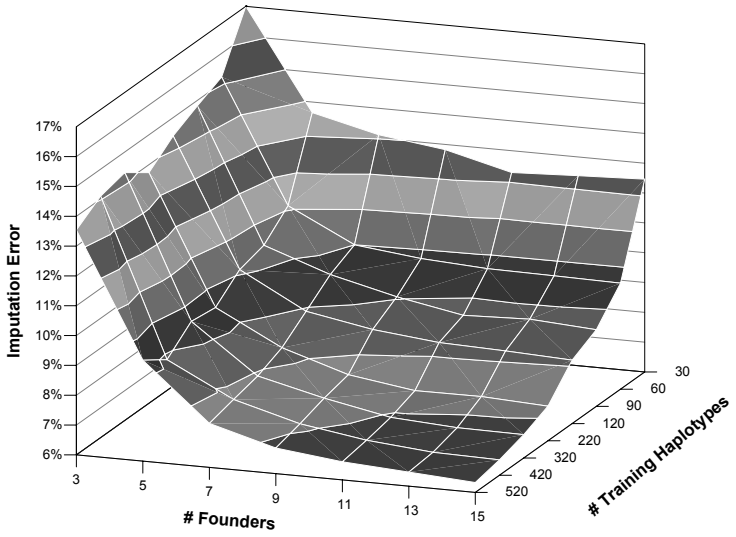


Figure 2.8: Imputation error rate on the IMAGE chr. 22 dataset for varying numbers of HMM founders and training haplotypes.

HMM founders is varied between 3 and 15 and the number of training haplotypes is varied between 30 and 520. Accuracy improves significantly when using reference panels larger than the commonly used HapMap panels, particularly in conjunction with increasing the number of HMM founders. For example, compared to the GEDI settings used in [26] (120 training haplotypes and 7 founders), increasing the number of training haplotypes to 520 and the number of founders to 15 yields an accuracy gain of over 2.5%.

Although the accuracy gained by using a larger number of HMM founders comes at the cost of increased imputation time, the latter remains practical for up to 15 founders, above which accuracy gains become very small. Indeed, as shown in Fig. 2.9, GEDI optimizations such as the PopTree trie speed-up described in Section 2.2.4 lead to sub-cubic runtime growth within the tested range of HMM founders, allowing users to better control the tradeoff between imputation speed and accuracy.

Indeed, the PopTree speed-up achieved by using tries depends on the number and the similarity of the samples, as well as the number of SNP loci. For exam-

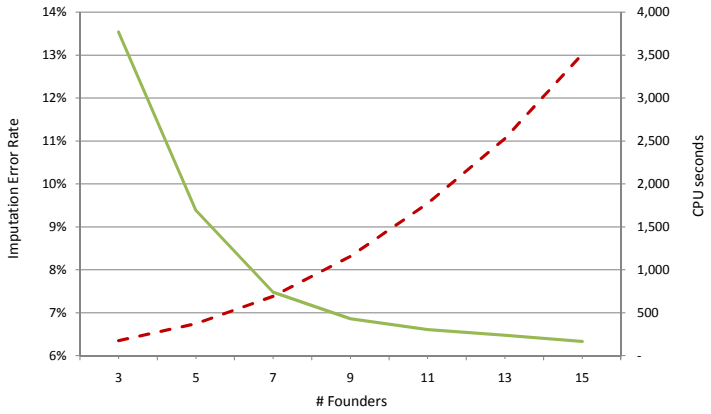


Figure 2.9: GEDI imputation error rate and runtime for varying number of founders (IMAGE chr. 22 dataset, 520 training haplotypes).

ple, when performing imputation using 10 flanking SNPs on the 2502 samples of the IMAGE dataset, using PopTree gives an approximate speed-up of 3×. See fig. (2.10) for the runtime comparison between using GEDI imputation with and without the speed-up (PopTree and Slow algorithms, respectively), when varying the number of flanking SNPs between 3 and 100 and the 7 or 13 HMM founders.

GEDI is also able to exploit pedigree information when available. For genotype data of related individuals, imputation probabilities (and log-likelihood ratios) are computed jointly over parents-child trios, using an extended version of the forward-backward algorithm as described in the previous chapter (see [39] for details). Fig. 2.11) compares the imputation error achieved by running GEDI with 13 HMM founders on the IMAGE dataset under two scenarios: (a) treating the 2502 test individuals as unrelated (as we have done in all previous experiments), and (b) analyzing them as 834 parents-child trios. Performing trio-based imputation reduces error rate by 0.22-0.44%, depending on the number of haplotypes used for training the model, pointing out to the value of using pedigree information.

Finally, we conducted experiments to assess the value of performing genotype error correction and missing data recovery prior to imputation. We generated a

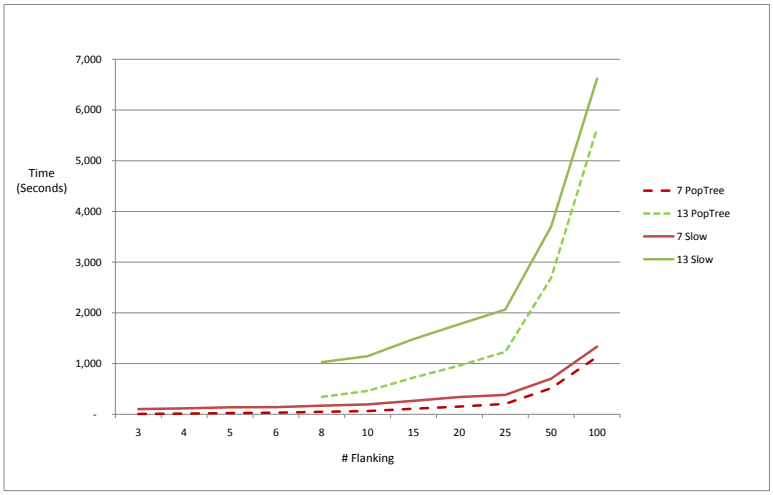


Figure 2.10: Runtime comparison between using GEDI imputation with and without PopTree

version of the IMAGE chr. 22 dataset generated by randomly inserting 1% errors and 1% missing data at typed SNP loci, and then ran two different analysis flows provided by GEDI:

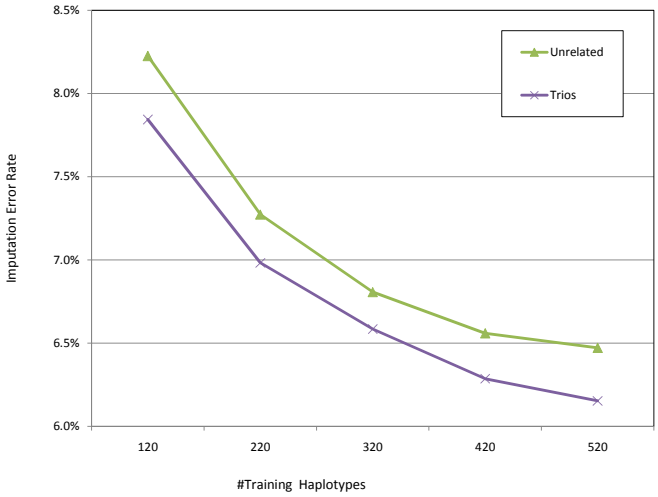


Figure 2.11: Effect of using pedigree information during imputation (IMAGE chr. 22 dataset, 13 HMM founders).

Table 2.3: Comparison of two GEDI imputation flows on a version of the IMAGE chr. 22 dataset generated by randomly inserting 1% errors and 1% missing data (520 training haplotypes).

GEDI flow	7 Founders		13 Founders	
	Error Rate	CPU sec.	Error Rate	CPU sec.
IMP	8.17%	410	7.20%	1,637
EDC+MDR+IMP	8.07%	4,153	6.91%	15,937

- In the first flow, referred to as IMP, genotypes at untyped SNP (the same as those used in previous experiments) were imputed based on the genotype data at typed SNPs and HMM models trained using 520 reference haplotypes.

- In the second flow, referred to as EDC+MDR+IMP, we first trained an HMM model over typed SNPs using the 520 reference haplotypes together with haplotypes inferred by phasing all test genotypes. This model was next used to run GEDI's error detection and correction and missing data recovery functions, replacing every SNP genotype g_i for which the likelihood ratio $\max_x P_M(\mathbf{g}[g_i \leftarrow x])/P_M(\mathbf{g})$ is greater than 10^3, respectively every missing SNP genotype g_i, by $\text{argmax}_x P_M(\mathbf{g}[g_i \leftarrow x])$. Finally, imputation was performed as in the IMP flow, but based on the modified genotype data for typed SNPs rather than the original genotypes.

Table 2.3 gives the error rate and runtime for running the two flows with 7, respectively 13 HMM founders. Performing the EDC+MDR+IMP flow improves accuracy over direct imputation in both cases, by almost 0.3% in the case of 13 founders. While EDC+MDR+IMP requires about 10× more time for the IMAGE dataset used in our experiment, the runtime increase should be much smaller for typical GWAS datasets, for which the number of typed loci is typically smaller than that of untyped loci. Indeed, for such datasets imputation time (which grows linearly with the number of untyped loci) is likely to dominate the time needed

for performing error detection and correction and missing data recovery (which is proportional to the number of typed loci).

While the accuracy gains obtained by using pedigree information or performing the EDC+MDR+IMP flow are small, they can translate in non-negligible cost savings. Indeed, as noted by [33], each 1% gain in imputation accuracy translates into a 5-10% reduction in the sample size needed to achieve a desired statistical power level.

2.5 Conclusion

In this chapter 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 individuals or trios and SNP loci, making them appropriate for handling the datasets generated by current large-scale association studies. Additionally, GEDI optimizations such as the PopTree trie speed-up lead to sub-cubic runtime growth within the tested range of HMM founders, allowing users to better control the tradeoff between runtime speed and accuracy. 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 [30]. Accuracy further benefits performing genotype error correction and missing data recovery prior to imputation using the GEDI software. As noted earlier, each 1the sample size needed to achieve a desired statistical power level.

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.

Chapter 3

Imputation-based Local Ancestry Inference in Admixed Populations[1]

3.1 Introduction

Rapid advances in SNP genotyping technologies have enabled the collection of large amounts of population genotype data, accelerating the discovery of genes associated with common human diseases. Admixture mapping has recently emerged as a powerful method for detecting risk factors for diseases that differ in prevalence across populations [65]. This type of mapping relies on genotyping hundreds of thousands of single nucleotide polymorphisms (SNPs) across the genome in a population of recently admixed individuals and is based on the assumption that near a disease-associated locus there will be an enhanced ancestry content from the population with higher disease prevalence. Therefore, a critical step in admixture mapping is to obtain accurate estimates of local ancestry around each genomic

[1]The results presented in this chapter are based on joint work with I. Mandoiu and B. Pasaniuc [59].

locus.

Several methods have been developed for addressing the local ancestry inference problem. Most of these methods use a detailed model of the data in the form of a hidden Markov model, e.g. SABER [75], SWITCH [66], HAPAA [74] but differ in the exact structure of the model and the procedures used for estimating model parameters. A second class of methods estimate the ancestry structure using a window-based framework and aggregate the results for each SNP using a majority vote: LAMP [67] uses an assumption of no recent recombination events within each window to estimate the ancestries, while WINPOP [60] employs a more refined model of recombination events coupled with an adaptive window size computation to achieve increased accuracy. Local ancestry inference methods also differ in the type of information used to make local ancestry inferences. Surprisingly, methods that do not model the linkage disequilibrium (LD) structure between SNPs currently outperform methods that model the LD information extracted from ancestral population haplotypes.

The main contribution of this chapter is a novel method for imputation-based local ancestry inference that more effectively exploits LD information. Our method uses a factorial HMMs trained on ancestral haplotypes to impute genotypes at all typed SNP loci (temporarily marking each SNP genotype as missing) under each possible local ancestry. We then assign to each locus the local ancestry that yields the highest imputation accuracy, as assessed using a weighted-voting scheme based on multiple SNP windows centered on the locus of interest. Preliminary experiments on simulated admixed populations generated starting from the four HapMap panels [78] show that imputation-based ancestry inference has accuracy competitive with best existing methods in the case of distant ancestral populations, and is significantly more accurate for closely related ancestral populations. We also give results showing that the accuracy of untyped SNP genotype imputation

in admixed individuals improves significantly when taking into account estimates of local ancestry.

3.2 Methods

In this work we consider the inference of locus-specific ancestry in recently admixed populations. We assume that for each admixed individual we are given the genotypes at a dense set of autosomal SNP loci, and seek to infer the two ancestral populations of origin at each genotyped locus. For simplicity we consider only bi-allelic SNPs. For every SNP locus, we denote the major and minor alleles by 0 and 1. A SNP genotype is encoded as the number of minor alleles at the corresponding locus, i.e., 0 and 2 encode homozygous major and minor genotypes, while 1 denotes a heterozygous genotype.

3.2.1 Genotype Imputation Within Windows with Known Local Ancestry

Various forms of left-to-right HMM models of haplotype diversity in a homogeneous population have been successfully used for numerous genetic data analysis problems including SNP genotype error detection [36], genotype phasing [64, 68], testing for disease association [41, 69], and imputation of untyped SNP genotypes [40, 45, 51, 68]. In this section we extend the imputation model described in the previous chapter to the case of individuals with known mixed local ancestry. Specifically, we assume that, over the set of SNPs considered, the individual has one haplotype inherited from ancestral population $\mathcal{P}_k$ and the other inherited from ancestral population $\mathcal{P}_l$, where $\mathcal{P}_k$ and $\mathcal{P}_l$ are known (not necessarily distinct) populations.

Multilocus SNP genotypes of individuals with such mixed ancestry are modeled statistically using a *factorial HMM* (F-HMM) [27] referred to as $\mathcal{M}_{kl}$ and

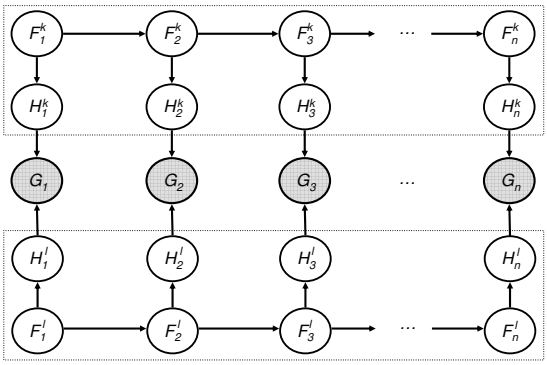


Figure 3.1: Factorial HMM model for a multilocus SNP genotype $(G_1, \ldots, G_n)$ over an n-locus window within which one haplotype is inherited from ancestral population $\mathcal{P}_k$ and the other from ancestral population $\mathcal{P}_l$. For every locus i, F_i^k and H_i^k denote the founder haplotype, respectively the allele observed on the haplotype originating from population $\mathcal{P}_k$; similarly, F_i^l and H_i^l denote the founder haplotype and observed allele for the haplotype originating from population $\mathcal{P}_l$.

graphically represented in Figure 3.1. At the core of the model are two left-to-right HMMs representing haplotype frequencies for the two ancestral populations (dotted boxes in Figure 3.1). Under these models, a haplotype from population $\mathcal{P}_j$, $j \in \{k, l\}$ is viewed as a mosaic formed as a result of historical recombination among a set of K_j founder haplotypes, where K_j is a population specific parameter (unless specified otherwise, we used $K_j = 7$ in our experiments).

Formally, for each SNP locus $i \in \{1, \ldots, n\}$, we let $G_i \in \{0, 1, 2\}$ be a random variable representing the genotype at locus i, $H_i^j \in \{0, 1\}$ be a random variable representing the allele inherited from population $\mathcal{P}_j$ at locus i, and $F_i^j \in \{1, \ldots, K_j\}$ be a random variable denoting the founder haplotype from which H_i^j originates. Values taken by these random variables are denoted by the corresponding lowercase letters (e.g., g_i, h_i^j, f_i^j). The model postulates that for each $j \in \{k, l\}$, F_i^j, $i = 1, \ldots, n$, form the states of a first order HMM with emissions H_i^j. We set $P(g_i | h_i^k, h_i^l)$ to be 1 if $g_i = h_i^k + h_i^l$ and 0 otherwise. Model training is completed by separately estimating probabilities $P(f_1^j)$, $P(f_{i+1}^j | f_i^j)$, and $P(h_i^j | f_i^j)$ using the

classical Baum-Welch algorithm [6] based on haplotypes inferred from a panel representing each ancestral population $\mathcal{P}_j$, $j \in \{k, l\}$. The parameters of the two left-to-right HMMs can alternatively be estimated directly from unphased genotype data using an EM algorithm similar to those in [41, 64].

Let $\mathbf{g} = (g_1, \ldots, g_n)$ be the multilocus genotype of a mixed ancestry individual and let $\mathbf{g}_{-i} = (g_1, \ldots, g_{i-1}, g_{i+1}, \ldots, g_n)$. If the individual's SNP genotype at locus i is unknown, it can be imputed based on the model $\mathcal{M}_{kl}$ by maximizing over $g \in \{0, 1, 2\}$

$$P_{\mathcal{M}_{kl}}(G_i = g|\mathbf{g}_{-i}) \propto P_{\mathcal{M}_{kl}}(\mathbf{g}[g_i \leftarrow g]) \tag{3.1}$$

where $\mathbf{g}[g_i \leftarrow g] = (g_1, \ldots, g_{i-1}, g, g_{i+1}, \ldots, g_n)$. The ancestry inference method described in Section 3.2.2 temporarily marks as missing and imputes each SNP genotype, and thus requires computing probabilities (3.1) for *all* n SNP loci. This computation can be done efficiently using a forward-backward algorithm, as described below.

For every $i \in \{1, \ldots, n\}$, $f_i^k \in \{1, \ldots, K_k\}$, and $f_i^l \in \{1, \ldots, K_l\}$, we let $\mathcal{F}^i_{f_i^k, f_i^l} = P_{\mathcal{M}_{kl}}(g_1, \ldots, g_{i-1}, f_i^k, f_i^l)$, which we refer to as the *forward probability* associated with the partial multilocus genotype $(g_1, \ldots, g_{i-1})$ and the pair of founder states (f_i^k, f_i^l) at locus i. The forward probabilities can be computed using the recurrence:

$$\mathcal{F}^1_{f_1^k, f_1^l} = P(f_1)P(f_1') \tag{3.2}$$

$$\begin{aligned}\mathcal{F}^i_{f_i^k, f_i^l} &= \sum_{f_{i-1}^k=1}^{K_k} \sum_{f_{i-1}^l=1}^{K_l} \mathcal{F}^{i-1}_{f_{i-1}^k, f_{i-1}^l} \mathcal{E}^{i-1}_{f_{i-1}^k, f_{i-1}^l}(g_{i-1}) P(f_i^k|f_{i-1}^k) P(f_i^l|f_{i-1}^l) \\ &= \sum_{f_{i-1}^k=1}^{K_k} P(f_i^k|f_{i-1}^k) \sum_{f_{i-1}^l=1}^{K_l} \mathcal{F}^{i-1}_{f_{i-1}^k, f_{i-1}^l} \mathcal{E}^{i-1}_{f_{i-1}^k, f_{i-1}^l}(g_{i-1}) P(f_i^l|f_{i-1}^l)\end{aligned} \tag{3.3}$$

where

$$\mathcal{E}^i_{f_i^k,f_i^l}(g_i) = \sum_{\substack{h_i^k,h_i^l\in\{0,1\}\\ h_i^k+h_i^l=g_i}} P(h_i^k|f_i^k)P(h_i^l|f_i^l) \tag{3.4}$$

The innermost sum in (4.18) is independent of f_i^k, and so its repeated computation can be avoided by replacing (4.18) with:

$$\mathcal{C}^i_{f_{i-1}^k,f_i^l} = \sum_{f_{i-1}^l=1}^{K_l} \mathcal{F}^{i-1}_{f_{i-1}^k,f_{i-1}^l}\mathcal{E}^{i-1}_{f_{i-1}^k,f_{i-1}^l}(g_{i-1})P(f_i^l|f_{i-1}^l) \tag{3.5}$$

$$\mathcal{F}^i_{f_i^k,f_i^l} = \sum_{f_{i-1}^k=1}^{K_k} P(f_i^k|f_{i-1}^k)\mathcal{C}^i_{f_{i-1}^k,f_i^l} \tag{3.6}$$

By using recurrences (4.17), (4.20), and (4.21), all forward probabilities can be computed in $O(nK^3)$ time, where n is the number of SNP loci and $K = \max\{K_k, K_l\}$.

Backward probabilities $\mathcal{B}^i_{f_i^k,f_i^l} = P_{\mathcal{M}_{kl}}(f_i^k, f_i^l, g_{i+1}, \ldots, g_n)$ can be computed in $O(nK^3)$ time using similar recurrences:

$$\begin{aligned}
\mathcal{B}^n_{f_n^k,f_n^l} &= 1\\
\mathcal{D}^i_{f_{i+1}^k,f_i^l} &= \sum_{f_{i+1}^l=1}^{K_l} \mathcal{B}^{i+1}_{f_{i+1}^k,f_{i+1}^l}\mathcal{E}^{i+1}_{f_{i+1}^k,f_{i+1}^l}(g_{i+1})P(f_{i+1}^l|f_i^l)\\
\mathcal{B}^i_{f_i^k,f_i^l} &= \sum_{f_{i+1}^k=1}^{K_k} P(f_{i+1}^k|f_i^k)\mathcal{D}^i_{f_{i+1}^k,f_i^l}
\end{aligned}$$

After computing forward and backward probabilities, posterior SNP genotype probabilities (3.1) can be evaluated in $O(K^2)$ time per SNP locus by observing that:

$$P_{\mathcal{M}_{kl}}(\mathbf{g}[g_i \leftarrow g]) = \sum_{f_i^k=1}^{K_k}\sum_{f_i^l=1}^{K_l} \mathcal{F}^i_{f_i^k,f_i^l}\mathcal{E}^i_{f_i^k,f_i^l}(g)\mathcal{B}^i_{f_i^k,f_i^l} \tag{3.7}$$

Thus, the total time for computing all posterior SNP genotype probabilities is $O(nK^3)$.

Input: multilocus genotype $\mathbf{g} = (g_1, \ldots, g_n)$, window half-size w, and reference haplotypes for ancestral populations $\mathcal{P}_1, \ldots, \mathcal{P}_N$
Output: inferred local ancestries $\hat{a}_i \in \mathcal{A}$ for each $i = 1, \ldots, n$

1. Train HMM models for each ancestral population and combine them to form factorial HMM models $\mathcal{M}_{kl}$ for every $kl \in \mathcal{A}$
2. For each locus i, compute posterior SNP genotype probabilities (Equation 3.1) under each local ancestry model $\mathcal{M}_{kl}$
3. For each locus $i = 1, \ldots, n$,

$$\hat{a}_i \leftarrow \text{argmax}_{kl \in \mathcal{A}} \sum_{j \in W_i} P_{\mathcal{M}_{kl}}(G_i = g_i | \mathbf{g}_{-i}) \tag{3.8}$$

where $W_i = \{\max\{1, i - w\}, \ldots, \min\{n, i + w\}\}$

Figure 3.2: Single-window imputation-based ancestry inference algorithm.

3.2.2 Local Ancestry Inference

Consider an individual coming from an admixture of (a subset of) of N ancestral populations $\mathcal{P}_1, \ldots, \mathcal{P}_N$. As in previous works [75, 67, 66, 74, 60], we view the local ancestry at a locus as an unordered pair of (not necessarily distinct) ancestral populations. The set of possible local ancestries is denoted by $\mathcal{A} = \{kl \mid 1 \leq k \leq l \leq N\}$.

Our local ancestry inference method is based on two observations: (1) for individuals from recently admixed populations the local ancestry of a SNP locus is typically shared with a large number of neighboring loci, and (2) the accuracy of SNP genotype imputation within such a neighborhood is typically higher when using the factorial HMM model $\mathcal{M}_{kl}$ corresponding to the correct local ancestry compared to a mis-specified model. These observations suggest using the algorithm in Figure 3.2 for inferring local ancestry based on imputation accuracy within windows centered at each SNP locus. More precisely, the algorithm assigns to each

SNP locus i the local ancestry that maximizes the average posterior probability for the true SNP genotypes over a window of up to $2w + 1$ SNPs centered at i (w SNPs downstream and w SNPs upstream of i).

Step 1 of the algorithm requires training N left-to-right HMMs based on haplotype data using the Baum-Welch algorithm, which takes $O(nK^2)$ per iteration and typically converges in a small number of iterations. As described in Section 3.2.1, Step 2 of the algorithm is implemented in $O(nK^3)$ time for each local ancestry model $\mathcal{M}_{kl}$. Once posterior SNP genotype probabilities are computed in Step 2, the window average probabilities required in Step 3 for each local ancestry model $\mathcal{M}_{kl}$ can be computed in $O(1)$ per window after precomputing in $O(n)$ time the sums of posterior probabilities for all prefix sets $\{1, \ldots, i\}$. Thus, since the number of possible ancestry models is $|\mathcal{A}| = O(N^2)$, the algorithm requires $O(nK^3N^2)$ time overall.

As previously observed for other window-based methods of local ancestry inference [67, 60], optimal window size selection plays a significant role in the overall estimation accuracy. Window-based methods must balance two conflicting requirements: on one hand, small window sizes may not provide enough information to accurately differentiate between the $|\mathcal{A}|$ possible local ancestries (particularly when ancestral populations are closely related) and on the other hand, large window sizes lead to more frequent violations of the assumption that local ancestry is uniform within each window. In the case of imputation-based ancestry inference we obtained good results by using a multi-window approach: for each SNP genotype g_i we run the algorithm of Figure 3.2 for all $w \in \{100, 200, \ldots, 1500\}$ and aggregate the results over all windows using a simple weighted voting scheme. Specifically, within each window we assign to each ancestry model $\mathcal{M}_{kl}$ a weight obtained by dividing the average posterior probability of the true genotypes, $\frac{1}{|W_i|} \sum_{j \in W_i} P_{\mathcal{M}_{kl}}(G_i = g_i | \mathbf{g}_{-i})$ by the sum of the averages achieved by all local ancestry models, and select

for each locus the model with maximum sum of weights over all windows. Preliminary experiments (see Figure 3.3 and Table 3.1) suggest that the multi-window strategy yields an average accuracy that is very close to (and, for some admixed populations, better than) the maximum average accuracy achieved by running the single-window algorithm with any window size from the above set.

3.3 Experimental results

In this section we present preliminary results comparing our approach to several state-of-the-art methods for local ancestry inference. We begin with results demonstrating the accuracy of imputation based on the factorial HMM model. In the first set of experiments, we compare our imputation-based algorithm to existing methods for local ancestry inference on admixture datasets simulated starting from the four populations represented in HapMap [78]. Finally, we present results demonstrating the benefit of incorporating accurate local ancestry estimates when performing genotype imputation for admixed individuals.

3.3.1 Inference of Local Ancestry in Admixed Populations

The method described in Section 3.2.2 was implemented in an extension of the GEDI software package [40], referred to as GEDI-ADMX. We compared GEDI-ADMX to several local ancestry inference methods capable of handling genome-wide data. Three of the competing methods (SABER [75], SWITCH [66], and HAPAA [74]) are HMM based, while the other two (LAMP [67] and WINPOP [60]) perform window-based estimation based on genotype data at a set of unlinked SNPs. When comparing various methods for ancestry inference one needs to take into account the fact that different methods use different types of information to make ancestry predictions. LAMP, WINPOP and SWITCH only require in-

formation about ancestral allele frequencies, while the other methods require the ancestral genotypes. In addition, HAPAA and GEDI-ADMX use additional information about ancestral haplotypes. Some of the methods also require the number of generations since the admixture process started. In general, we provided each method the maximum amount of information about the admixture process (e.g. number of generations g or the admixture ratio α) that it could take into account. Although these parameters can be estimated from genotype data when needed [76], we note that GEDI-ADMX does not require any additional parameters besides the ancestral haplotypes.

Experiments were performed on simulated admixtures using as ancestral populations the four HapMap [78] panels: Yoruba people from Ibadan Nigeria (YRI), Japanese from the Tokyo area (JPT), Han Chinese from Beijing (CHB) and Utah residents with northern European ancestry (CEU). We simulated admixtures for each of the YRI-CEU, CEU-JPT, and JPT-CHB pairs of populations as follows: we started the simulation by joining a random set of $\alpha \times n$ individuals from the first population and $(1-\alpha) \times n$ individuals from the second population. Within the merged panel we simulated g generations of random mating with a mutation and recombination rate of 10^{-8} per base pair per generation. We used only the 38,864 SNPs located on Chromosome 1 found on the Affymetrix 500K GeneChip Assay. For these simulations we used $n = 2000$, $g = 7$ and $\alpha = 0.2$ as it roughly corresponds to the admixture history of the African American population [77, 70, 58]. Our simulations result in an admixed population with known local ancestry. Each of the evaluated methods infers an ancestry estimate for every SNP genotype; we measure the accuracy as the fraction of SNP genotypes for which the correct ancestry is inferred.

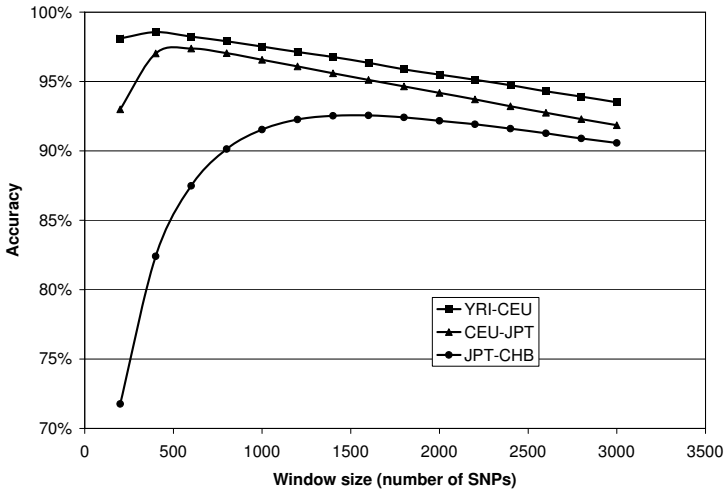


Figure 3.3: Accuracy of local ancestry estimates obtained by GEDI-ADMX on the three HapMap admixtures using a single window of varying size.

Effect of Window Size on the Local Ancestry Estimates.

Figure 3.3 plots the accuracy of the local ancestry prediction of GEDI-ADMX on the HapMap admixtures for different window sizes. As expected, the accuracy initially increases with window size for all three datasets, since more information is available to differentiate between ancestry models. However, very large window sizes lead to more violations in the assumption of uniform ancestry within each window, overshadowing these initial benefits. As previously reported in other window-based methods [67, 60] we also notice that the best window size employed by our method for the three datasets is correlated with the genetic distance between ancestral populations as closer ancestral populations require longer window size for accurate predictions. Finally, we notice that the combined multi-window approach described in Section 3.2.2 achieves accuracy close to the best window size for the YRI-CEU and CEU-JPT admixtures and better than any window size for the JPT-CHB admixture (see Table 3.1). All remaining results were obtained using the multi-window approach.

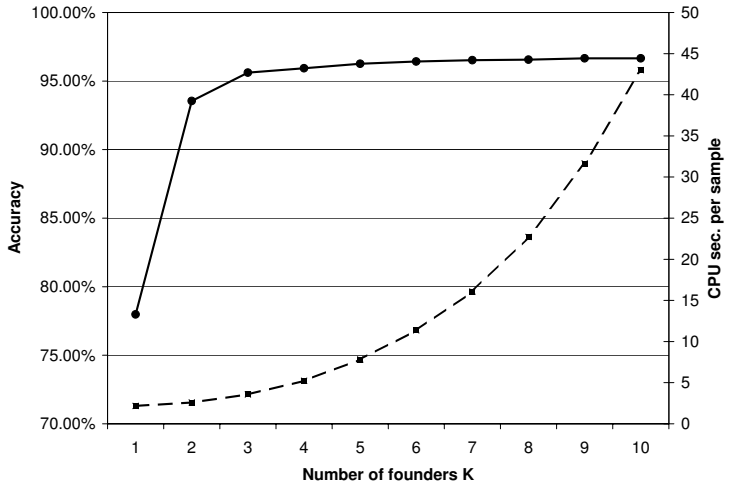


Figure 3.4: GEDI-ADMX accuracy (solid) and runtime (dashed) for varying values of the number K of HMM founder haplotypes on the CEU-JPT dataset, consisting of $n = 38,864$ SNPs on Chromosome 1.

Effect of Number of Founders on Local Ancestry Inference Accuracy and Runtime Scalability.

An important parameter of the HMM models used to represent the LD in ancestral populations is the number of founder haplotypes K. As discussed in Section 3.2.2, the runtime of the algorithms grows asymptotically with the cube of K, which renders the use of very large values of K impractical. Using very large values of K may also be problematic when the number of training haplotypes is limited, due to model overfitting. On the other hand, HMMs with very few founder haplotypes have a limited ability of capturing LD patterns in the ancestral populations, and is expected to lead to poor accuracy.

To assess these potentially complex tradeoffs between runtime and accuracy we run GEDI-ADMX on the CEU-JPT dataset using for both ancestral populations a number of founder haplotypes K varied between 1 and 10. The accuracy and runtime achieved by GEDI-ADMX for each value of K are plotted in Figure 3.4. Since for $K = 1$ our HMM model degenerates into a simple multinomial i.i.d. model

that captures allele frequency at each SNP but completely ignores LD, it is not surprising that ancestry inference accuracy is relatively poor (about 78%). For $K = 2$ accuracy improves significantly (to 93.5%), as the model is now able to represent pairwise LD between adjacent SNPs. As K is further increased, the model can capture more of the longer range LD, leading to further accuracy improvements. However, improvements in accuracy are quickly diminishing, with only 1% accuracy improvement achieved when increasing K from 3 to 10.

Although for small values of K lower order terms make the runtime growth in Figure 3.4 appear sub-cubic, the asymptotic cubic growth is already apparent for the largest tested values of K. For remaining experiments we used $K = 7$ since this setting achieves a good tradeoff between runtime and accuracy.

Comparison with Other Methods.

Table 3.1 presents accuracies achieved by the six compared methods on the three simulated HapMap admixtures. We note that GEDI-ADMX achieves similar accuracy to the best performing methods on the YRI-CEU and CEU-JPT admixture, while yielding a significant improvement in accuracy for the JPT-CHB dataset. Indeed, on the JPT-CHB admixture our method achieves an accuracy of 94.0%, which is an increase of more than 11% over the second best performing method WINPOP. Table 3.1 also reports an upper-bound on the maximum accuracy that can be obtained by methods that do not model the linkage disequilibrium (LD) between SNPs, computed as described in [60]. Notably, GEDI-ADMX accuracy on the JPT-CHB dataset exceeds the upper-bound for methods that do not model the LD. This underscores the importance of exploiting ancestral haplotypes when performing local ancestry inference for admixtures of closely related populations.

Method	YRI-CEU	CEU-JPT	JPT-CHB
SABER	89.4	85.2	68.2
HAPAA	93.7	88.2	72.0
SWITCH	97.8	94.8	74.8
LAMP	94.8	93.0	65.8
WINPOP	98.0	95.9	82.8
Upper Bound(no LD)	99.9	99.6	91.9
GEDI-ADMX	97.5	96.5	94.0

Table 3.1: Percentage of correctly recovered SNP ancestries on three HapMap admixtures with $\alpha = 0.2$.

3.3.2 SNP Genotype Imputation in Admixed Populations

In this section we present results that further demonstrate the synergy between SNP genotype imputation and local ancestry inference in admixed population. More specifically, we focus on assessing the utility of inferring locus-specific ancestries when performing imputation of genotypes for untyped SNPs.

For this experiment we generated three admixtures, corresponding to the YRI-CEU, CEU-JPT and JPT-CHB pairs of HapMap populations, using the same simulation procedure as described in Section 3.3.1 with parameters of $n = 2000$, $\alpha = 0.5$ and $g = 10$. We randomly chose 10% of the SNPs as untyped and we masked them from all the individuals in the admixture. We first ran GEDI-ADMX using unmasked SNP genotypes to infer local ancestries as described in Section 3.2.2. We then imputed masked genotypes using the model in Section 3.2.1 based on the ancestry inferred for the adjacent unmasked SNPs. We measured the error rate of the imputation procedure as the percentage of genotypes inferred erroneously (using no cutoff threshold on posterior imputation probability). To establish a baseline for the comparison, we also performed imputation using the GEDI package [40], based on a factorial model similar to that in Section 3.2.2 except that it consists of two identical left-to-right HMMs trained on either (1) panel haplotypes for only one of the ancestral populations (GEDI-1-Pop), re-

Method	YRI-CEU	CEU-JPT	JPT-CHB
GEDI-1-Pop Avg.	12.79	6.67	3.81
GEDI-2-Pop	7.31	3.90	3.02
GEDI-ADMX	4.34	2.81	2.74

Table 3.2: Imputation error rate, in percents, on three HapMap simulated admixtures with $\alpha = 0.5$.

spectively on (2) a haplotype list obtained by merging the panel haplotypes of the two ancestral populations (GEDI-2-Pop).

Table 3.2 shows the imputation accuracy achieved by the three compared methods. As expected, there is a large decrease in error rate when switching from using only one panel of ancestral haplotypes to using the combined panel consisting of haplotypes from both populations. Performing imputation based on the local ancestry inferred by GEDI-ADMX yields further improvements in accuracy. Accuracy gains are largest when admixed populations are distant (e.g. YRI-CEU).

3.4 GEDI-ADMX Software

The GEDI-ADMX package provides methods for the following in admixed populations based on whole-genome SNP genotype data and reference haplotype panels for ancestral populations:

- local ancestry inference
- SNP genotype error detection and correction
- imputation of missing genotypes at typed SNPs, and
- imputation of genotypes at untyped SNPs

Currently GEDI-ADMX handles genotype data from unrelated individuals. As described earlier in this chapter, local ancestry inference uses factorial HMMs

trained on ancestral haplotypes to impute genotypes at all typed SNP loci (temporarily marking each SNP genotype as missing) under each possible local ancestry. GEDI-ADMX assigns to each locus the local ancestry that yields the highest imputation accuracy, as assessed using a weighted-voting scheme based on multiple SNP windows centered on the locus of interest. Error detection and imputation of missing genotypes at typed SNPs is performed using multilocus genotype probabilities computed based on the factorial HMM corresponding to the inferred local ancestry. Imputation of genotypes at each untyped SNP is performed based on posterior genotype probabilities computed using a similar factorial HMM spanning k (default k=10) typed SNPs before and after the imputed locus.

3.4.1 Gene Admix Viewer

Gene Admix Viewer is a graphical user interface for GEDI-ADMX as well as a graphical viewer of local ancestry results. The graphical viewer is based on code developed by Christian Wanamaker (see figure 3.5). It allows the user to run GEDI-ADMX on Windows machines, as well as visually review inferred ancestry along a chromosome from GEDI-ADMX results, with gene information also available for analysis.

3.5 Discussion

In this chapter we propose a novel algorithm for imputation-based local ancestry inference. Experiments on simulated data show that our method exploits ancestral haplotype information more effectively than previous methods, yielding consistently accurate estimates of local ancestry for a variety of admixed populations. Indeed, our method is competitive with best existing methods in the case of admixtures of two distant ancestral populations, and is significantly more accurate

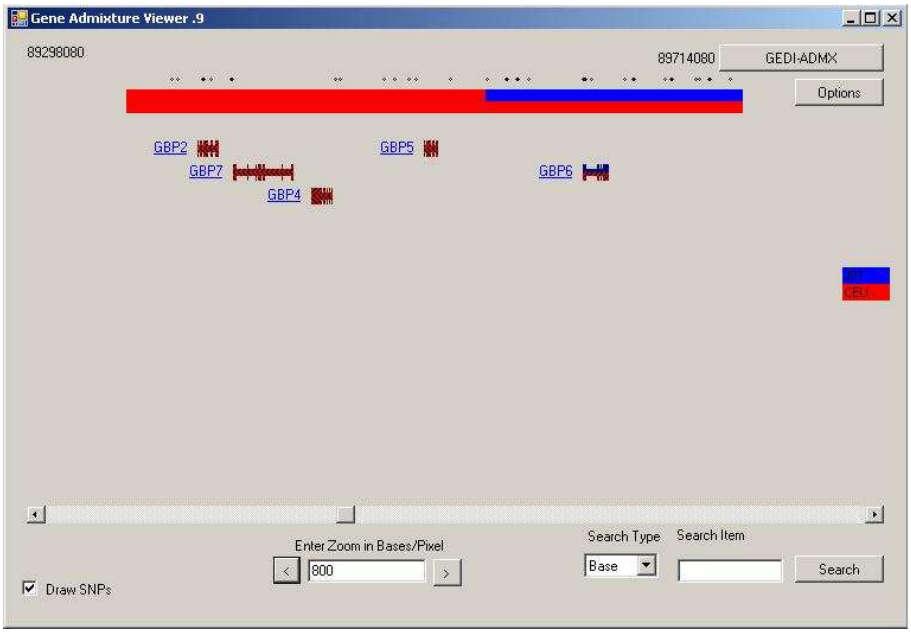


Figure 3.5: Gene Admix Viewer main screen

than previous methods for admixtures of closely related populations such as the JPT and CHB populations from HapMap. We also show that accurate local ancestry estimates lead to improved accuracy of untyped SNP genotype imputation for admixed individuals.

In ongoing work we are exploring methods that iteratively alternate between rounds of imputation-based ancestry inference and ancestry-based imputation for further improvements in accuracy. We are also conducting experiments to characterize the accuracy of our imputation-based local ancestry inference methods in the case of admixtures of more than two ancestral populations.

Chapter 4

Single Individual Genotyping from Low-Coverage Sequencing Data [1]

4.1 Introduction

Recent massively parallel sequencing technologies deliver orders of magnitude higher throughput compared to classic Sanger sequencing. Sequencers like Roche/454 FLX Titanium, Illumina Genome Analizer II, ABI SOLiD 3 and Helicos HeliScope are able to provide millions of short reads in a single run which lasts just a few days in some cases and even less than one day in other cases. These advances promise to enable cost-effective shotgun sequencing of individual genomes. After recent publication of five complete individual genomes [9, 22, 52, 62, 80, 23], ongoing efforts focus on increasing the quality and coverage of short reads and on improving algorithms for mapping, genotyping and variations discovery to sequence over a thousand more individual genomes [1].

[1] The results presented in this chapter are based on joint work with J. Duitama, S. Dinakar, Y. Hernandez, I. Mandoiu and Y. Wu [35].

These high-throughput sequence data allows one to reconstruct complete diploid genomes. One important aspect of reconstructing diploid genomes from sequence data is determining the genotypes of the sequenced individuals at single nucleotide polymorphisms (SNPs). This process is called genotyping. That is, given a set of sequence reads, we want to determine the alleles at given SNP sites of the sequenced individual. Accurate genotyping is critical in many applications like Genome Wide Association Studies (GWAS) where individual genotypes are the base information to establish correlation between a genome location and a disease. Sequence data based genotyping is in principle easy if we have a *large* number of sequence reads covering these SNP sites. Here, the number of reads spanning a SNP site is called sequencing depth or coverage depth. Despite some known biases during the sequencing process, it is assumed that reads are sampled with uniform distribution across the genome and hence raw coverage depth is calculated on average by dividing the total number of bases produced by the sequencer by the size of the genome.

The genotyping problem becomes more challenging when only sequencing data with low coverage depth is available (which is much cheaper to obtain). This is because: (a) when sequencing depth is low, many alleles at SNP sites may not be observed in any sequencing reads; (b) it is unknown from which of the two chromosome copies a sequencing read originates; (c) other sources of noise, such as sequencing errors or incorrect read mapping, can cause more uncertainty.

An obvious way to obtain high-quality SNP genotypes from sequence data is to simply increase the sequencing depth. It was estimated that a coverage depth of over $21\times$ on each SNP is required to achieve 99% sensitivity at detecting heterozygous SNPs [82], in absence of additional information. A coverage depth of ($7.5\times$) was found in [22, 23] to correctly identify only 75% of the heterozygous SNPs, and sensitivity drops rapidly as depth decreases. Clearly, obtaining sequence data with

high sequencing depth can be very expensive. Thus, an important technical problem is to develop new SNP genotype calling methods using *low* coverage sequence data.

Before we present our new genotyping method, we first briefly discuss how genotyping is performed now. Current techniques for genotype calling are based on the number of reads supporting the existence of each specific allele on each SNP, which is called allele coverage. In [22, 23], an allele is called if there are at least two reads supporting it. In [23], this rule is combined with a binomial test in which it is assumed as null hypothesis that the genotype is heterozygous and the allele counts follow a binomial distribution centered at 0.5. A p-value of less than 0.01 is required to reject this hypothesis. In [46] the binomial model is improved by taking into account base quality scores and dependency between close SNPs. While these techniques are adequate in absence of additional information, their reliability is not enough for many applications like GWAS.

An important source of information not used in the above approach is the correlation between alleles at nearby SNP sites (called linkage disequilibrium). We note that previous works [36,45,51,73,81] suggest that when LD information is available it is possible to impute with high accuracy the genotypes of untyped SNPs. Public databases like the one published by the international HapMap project [15] provide reference panels for many human populations, which are likely to improve in size and accuracy as new genotype information is gathered.

In this chapter we present a statistical model for multilocus genotype inference that fully exploits the linkage disequilibrium information contained in a reference panel of haplotypes to improve the accuracy of genotype calls based on short reads mapped to a reference sequence. We designed a hierarchical factorial hidden markov model (HF-HMM), which assumes that the individual haplotypes are the result of recombination events between a small set of founder haplotypes. We im-

plemented a posterior decoding algorithm which, after training with the reference haplotypes, combines the genotype probabilities given by the HMM with the genotype probabilities given the allele counts to calculate the most likely genotype. We show how this model allows to achieve more than 95% accuracy on heterozygous SNP calls and more than 99% accuracy on homozygous SNP calls with just 5× coverage depth. We also show how the reference panel information helps to make better guesses for loci with lack of enough reads support. Software implementing this model has been released under the GNU General Public License and is available at http://dna.engr.uconn.edu/software/GeneSeq/.

4.2 Methods

In this section we describe the statistical model that represents the assumed model of haplotype diversity, the problem formulation, and the algorithm implemented to solve this problem. After introducing basic notations we describe a simplified model that assumes independence between sites and that is useful to introduce some concepts and for comparison purposes. We then expand this model to include dependences between SNP alleles at different sites. We formalize the multilocus genotype problem in the context of the extended model and we show that computing the most likely multilocus genotype is computationally hard. Finally, we present several heuristics for inferring multi-locus genotypes

4.2.1 Notations

We use uppercase italic letters (e.g., X) to denote random variables and lowercase italic letters (e.g., x) to denote generic values taken by them. Vectors of random variables and generic variables are denoted by boldface uppercase (e.g., $\mathbf{X}$), respectively boldface lowercase letters (e.g., $\mathbf{x}$). When there is no ambiguity on

the underlying probabilistic event we use $P(x)$ to denote $P(X = x)$, with similar shorthands used for joint and conditional probabilities of multiple events.

For simplicity we consider only biallelic SNPs on autosomes. For every SNP locus, we denote the two possible alleles by 0 and 1, and the three genotypes by 0, 1, and 2, with 0 and 2 denoting the homozygous 0 and homozygous 1 genotypes, and 1 denoting the heterozygous genotype.

4.2.2 Single SNP Genotype Calling

In this section we describe a genotype inference model that assumes the SNPs to be unlinked as, e.g., in [23], but further incorporates allele uncertainty quantified by sequencing quality scores, read mapping uncertainty, and population genotype frequencies estimated from a reference panel.

Let r be a read mapped onto the genome. We denote by $m(r)$ the probability that r is aligned in the correct position. If r cover SNP locus i, we denote by $r(i)$ the allele observed in the read at this locus. Since our focus is on genotyping SNPs represented in a reference panel, we further assume that panel SNPs at which the individual under study has novel allele variants (observed at an estimated 0.02% of the markers in [23]) have been identified in a preliminary analysis, e.g., by using binomial probability tests as in [23]. Based on this assumption, all reads with alleles not represented in the panel population are discarded, and for remaining reads r, $r(i) \in \{0, 1\}$. The probability that allele $r(i)$ is affected by a sequencing error is denoted by $\varepsilon_{r(i)}$. In our model we set $\varepsilon_{r(i)} = 10^{-q_{r(i)}/10}$, where $q_{r(i)}$ denotes the phred quality score [24] of $r(i)$.

Let G_i be a random variable denoting the unknown SNP genotype at locus i, and let $\mathbf{r}_i = \{r_{i,1}, \ldots, r_{i,c_i}\}$ be the arbitrarily ordered set of shotgun reads covering locus i, where c_i is the coverage at this locus. Since for a homozygous genotype the allele of origin for a read is the same regardless of which chromosome is sampled,

we get:

$$P(\mathbf{r}_i|G_i = 0) = \prod_{\substack{r\in\mathbf{r}_i \\ r(i)=0}} (1-\varepsilon_{r(i)})^{m(r)} \prod_{\substack{r\in\mathbf{r}_i \\ r(i)=1}} (\varepsilon_{r(i)})^{m(r)} \tag{4.1}$$

and

$$P(\mathbf{r}_i|G_i = 2) = \prod_{\substack{r\in\mathbf{r}_i \\ r(i)=1}} (1-\varepsilon_{r(i)})^{m(r)} \prod_{\substack{r\in\mathbf{r}_i \\ r(i)=0}} (\varepsilon_{r(i)})^{m(r)} \tag{4.2}$$

We interpret a correct mapping probability $m(r) < 1$ as equivalent to observing a "fractional" read and view the contribution of such fractional reads as satisfying additivity. For example, we view two reads with 50% mapping confidence as equivalent to a single read with full confidence. This interpretation of mapping probabilities is enforced by raising the terms corresponding to read r in equations (4.1) and (4.2) to $m(r)$.

For a read r covering a heterozygous SNP locus i allele $r(i)$ can be observed either as the result of sampling r from the chromosome bearing allele $r(i)$ and correctly sequencing it, or as the result of sampling the other chromosome followed by a sequencing error. Hence:

$$\begin{aligned} P(\mathbf{r}_i|G_i = 1) &= \prod_{r\in\mathbf{r}_i} \left(\frac{1}{2}(1-\varepsilon_{r(i)}) + \frac{1}{2}\varepsilon_{r(i)}\right)^{m(r)} \\ &= \left(\frac{1}{2}\right)^{\sum_{r\in\mathbf{r}_i} m(r)} \end{aligned} \tag{4.3}$$

A natural approach to single-locus SNP genotyping is to call a genotype of $\widehat{g}_i = \text{argmax}_{g_i\in\{0,1,2\}} P(g_i|\mathbf{r}_i)$ for every SNP locus i for which the maximum posterior genotype probability exceeds a user-specified threshold. Posterior probabilities are obtained from (4.1)-(4.3) by applying Bayes' formula:

$$P(G_i = g_i|\mathbf{r}_i) = \frac{P(g_i)P(\mathbf{r}_i|G_i = g_i)}{\sum_g P(G_i = g)P(\mathbf{r}_i|G_i = g)} \tag{4.4}$$

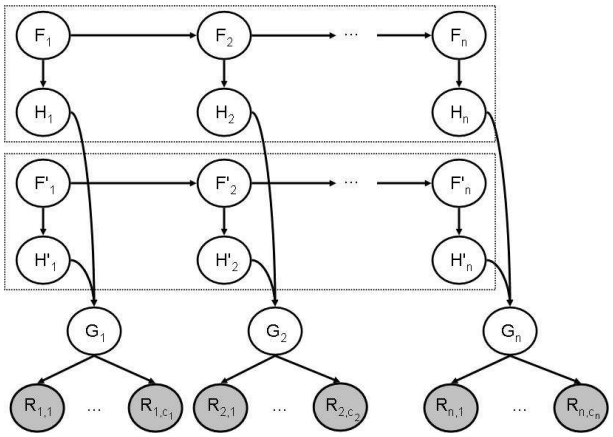


Figure 4.1: HF-HMM model for multilocus genotype inference.

where $P(G_i = g)$ denotes the population frequency of genotype g, estimated from the reference panel.

4.2.3 A Statistical Model for Multilocus Genotype Inference

In this section we introduce a statistical model that allows us to fully integrate shotgun sequencing data and LD information in the inference of SNP genotypes. Our model, represented graphically in Figure 4.1, can be thought of as a *hierarchical factorial HMM* (HF-HMM). Indeed, we use a distributed state (characteristic of factorial HMMs [27]) to exploit the independence between maternal and paternal chromosomes (implied by the assumption of random mating), while also employing a multilevel state representation as in hierarchical HMMs [25] to capture the structured nature of the data. The hierarchical factorial structure of the model leads to a reduced number of parameters and modular estimation procedures, and enables highly scalable inference algorithms, with runtime scaling linearly with the sum between the number of shotgun reads and that of SNP loci.

At the core of the model are two left-to-right HMMs M and M' representing

haplotype frequencies in the populations of origin of the sequenced individual's parents (dotted boxes in Figure 4.1). Under M and M' each haplotype is viewed as a mosaic formed as a result of historical recombination among a set of K founder haplotypes, where K is a population specific model parameter. Formally, for every SNP locus $i \in \{1, \ldots, n\}$, we let H_i (H'_i) be a random variable representing the allele observed at this locus on the maternal (paternal) chromosome of the individual under study, and F_i (F'_i) be a random variable denoting the founder haplotype from which H_i (respectively H'_i) originates. As in previous works [36,41,51,64,69], we assume that F_i form the states of a first order HMM with emissions H_i, and estimate probabilities $P(f_1)$, $P(f_{i+1}|f_i)$, and $P(h_i|f_i)$) using the classical Baum-Welch algorithm [6] based on haplotypes inferred from a panel representing the population of origin of the individual's mother. Probabilities $P(f'_1)$, $P(f'_{i+1}|f'_i)$, and $P(h'_i|f'_i)$ are estimated in the same way based on haplotypes inferred from a panel representing the population of origin of the individual's father.

We define $P(g_i|h_i, h'_i)$ to be 1 if $g_i = h_i + h'_i$ and 0 otherwise. Finally, we assume that each read covers a single SNP locus, and set

$$\begin{aligned} P(R_{i,j} = r|G_i = g_i) &= \frac{g_i}{2}(\varepsilon_{r(i)})^{1-r(i)}(1-\varepsilon_{r(i)})^{r(i)} \\ &\quad + \frac{2-g_i}{2}(\varepsilon_{r(i)})^{r(i)}(1-\varepsilon_{r(i)})^{1-r(i)} \end{aligned} \tag{4.5}$$

This implies that $P(\mathbf{r}_i|g_i)$ are given by Equations (4.1)-(4.3), and in the following we will assume that probabilities $P(\mathbf{r}_i|g_i)$ are precomputed in $O(m)$ time, where $m = \sum_{i=1}^{n} c_i$ is the total number of reads.

We can now formulate the following: **Multilocus Genotyping Problem (MGP)**

Given: *Two trained HMM models M, M' and a set of shotgun reads $\mathbf{r} = (\mathbf{r}_1, \ldots, \mathbf{r}_n)$*

Find: *Multilocus genotype* $\mathbf{g}^* \in \{0,1,2\}^n$ *with maximum posterior probability, i.e.,*

$$\mathbf{g}^* = \text{argmax}_{\mathbf{g}} P(\mathbf{g}|\mathbf{r}, M, M') \tag{4.6}$$

As shown below, MGP is NP-hard, and in fact, the maximum probability of a multilocus genotype is as hard to approximate as the maximum clique size of a graph. Formally, let us introduce the following optimization version of MGP:

Maximum Multilocus Genotype Probability Problem (MMGPP)

Given: *Two trained HMM models* M, M' *and a set of shotgun reads* $\mathbf{r} = (\mathbf{r}_1, \ldots, \mathbf{r}_n)$

Find: *The maximum multilocus genotype probability,*

$$\max_{\mathbf{g}} P(\mathbf{g}|\mathbf{r}, M, M') \tag{4.7}$$

Theorem 1 *For any* $\epsilon > 0$*, MMGPP cannot be approximated within* $O(n^{\frac{1}{2}-\epsilon})$ *unless P=NP, and it cannot be approximated within* $O(n^{1-\epsilon})$ *unless ZPP=NP. Furthermore, this holds even if* $M' = M$*.*

Proof. Lyngsø et al. [50] gave an approximation preserving reduction from the clique problem to the problem of computing the maximum probability of a string emitted by an HMM. It is not difficult to modify their construction to show that this reduction holds even for left-to-right HMMs that emit 0/1 strings of fixed length. Next, we show that computing the maximum probability of a string emitted by such an HMM M_0 can be reduced in approximation preserving manner to MMGPP with $M' = M$. The haplotype models M and M' are obtained from M_0 as follows (see the schematic state diagram in Figure 4.2):

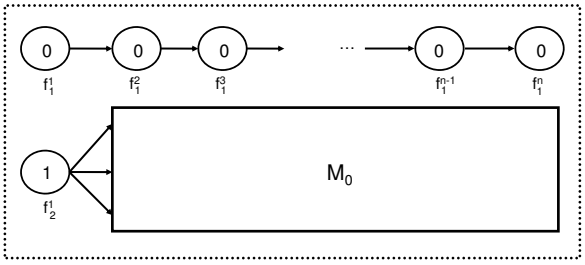


Figure 4.2: Schematic of reduction of the consensus string problem to MMGPP.

- The number of SNPs n is set to one plus the length of the strings emitted by M_0.
- At the first SNP, for two founder states f_1^1 and f_2^1 we have $P(f_i^1) = 1/2$; all other founder states have zero initial probability.
- For every SNP locus $i > 1$ we add a new founder f_1^i as well as a set of founders corresponding to the states at "column" $i - 1$ of M_0.
- All founder f_1^i, $i = 1, \ldots, n$, emit 0 with probability 1. Furthermore, $P(f_1^i | f_1^{i-1}) = 1$ for every $i = 2, \ldots, n$.
- Founder f_2^1 emits 1 with probability 1, and has transitions to founders f_j^2, $j > 1$, according to the initial probabilities of M_0.
- All other emission and transition probabilities are identical to those for the corresponding states of M_0.

Finally, we set $\mathbf{r} = \{r_0, r_1\}$ where r_0 is a read that supports allele 0 at first SNP and r_1 is a read that supports the allele 1 at first SNP. Error probabilities for both alleles are set to zero.

Note that $P(\mathbf{g}|\mathbf{r}, M, M') \neq 0$ only for multilocus genotypes with $g_1 = 1$ and

$g_i \in \{0,1\}$ for $i = 2, \ldots, n$. Furthermore, for such a genotype $\mathbf{g}$,

$$\begin{aligned} P(\mathbf{g}|\mathbf{r}, M, M') &= \frac{P(\mathbf{r}|\mathbf{g})P(\mathbf{g}|M, M')}{P(\mathbf{r})} \\ &= \frac{1}{4P(\mathbf{r})} P(\mathbf{g}|M, M') \\ &= \frac{1}{4P(\mathbf{r})} \frac{P(g_2, \ldots, g_n|M_0)}{2} \end{aligned} \tag{4.8}$$

The last equality comes from the fact that $\mathbf{g}$ can only be observed when the maternal haplotype is 0^n and the paternal haplotype is $\mathbf{g}$ or vice versa, and each of these configurations have a probability of $P(g_2, \ldots, g_n|M_0)/4$.

The inapproximability results follow from [50] since, by (4.8), $P(\mathbf{g}|\mathbf{r}, M, M')$ is constant fraction of $P(g_2, \ldots, g_n|M_0)$. □

It is easy to see that an algorithm similar to the forward algorithm for HMMs can be used to compute in polynomial time the marginal probability of a given genotype. Combined with Theorem 1, this observation implies the following:

Corollary 1 *MGP is NP-Hard.*

4.3 Efficient MGP Heuristics

4.3.1 Posterior Decoding

We next present several MGP heuristics, the first is similar to the posterior decoding algorithm for HMMs. Specifically, the algorithm selects for each SNP locus i the genotype $\widehat{g}_i$ with maximum posterior probability given the read data $\mathbf{r}$. Note that, unlike the single SNP genotype calling method describe in Section 4.2.2, where only the reads overlapping the SNP are taken into account, posterior decoding uses the *entire* set of reads.

Posterior Decoding Algorithm

1. *For each* $i = 1, \ldots, n$, $\widehat{g}_i \leftarrow \text{argmax}_{g_i} P(g_i|\mathbf{r})$
2. *Return* $\widehat{\mathbf{g}} = (\widehat{g}_1, \ldots, \widehat{g}_n)$

Below we explain how the posterior algorithm can be implemented in $O(m + nK^3)$ time. Since $P(g_i|\mathbf{r}) \propto P(g_i, \mathbf{r})$, the maximization in Step 1 of the posterior decoding algorithm can be equivalently restated as $\widehat{g}_i \leftarrow \text{argmax}_{g_i} P(g_i, \mathbf{r})$. Thus, we need to compute marginal probabilities $P(g_i, \mathbf{r})$ for every $i = 1, \ldots, n$ and $g_i \in \{0, 1, 2\}$.

For each SNP locus i and each pair of founders (f_i, f_i') we let the *forward probability* be $\mathfrak{F}^i_{f_i,f_i'} = P(\mathbf{r}_1, \ldots, \mathbf{r}_{i-1}, f_i, f_i')$, and the *backward probability* be $\mathfrak{B}^i_{f_i,f_i'} = P(\mathbf{r}_{i+1}, \ldots, \mathbf{r}_n|f_i, f_i')$, respectively. Using these forward and the backward probabilities, the marginal probability $P(g_i, \mathbf{r})$ can be written as

$$P(g_i, \mathbf{r}) = P(\mathbf{r}_i|g_i) \sum_{f_i=1}^{K} \sum_{f_i'=1}^{K} \mathfrak{F}^i_{f_i,f_i'} \mathfrak{B}^i_{f_i,f_i'} P(g_i|f_i, f_i')$$

where $P(g_i|f_i, f_i')$ is given by:

$$P(g_i|f_i, f_i') = \sum_{\substack{h_i, h_i' \in \{0,1\} \\ h_i + h_i' = g_i}} P(h_i|f_i) P(h_i'|f_i')$$

Thus all probabilities $P(g_i, \mathbf{r})$ can be computed in $O(nK^2)$ once the forward and backward probabilities $\mathfrak{F}^i_{f_i,f_i'}$ and $\mathfrak{B}^i_{f_i,f_i'}$ are available.

The forward probabilities can be computed using the recurrence:

$$\begin{aligned}
\mathfrak{F}^1_{f_1,f'_1} &= P(f_1)P(f'_1) && (4.9)\\
\mathfrak{F}^i_{f_i,f'_i} &= \sum_{f_{i-1}=1}^{K}\sum_{f'_{i-1}=1}^{K}\Big(\mathfrak{F}^{i-1}_{f_{i-1},f'_{i-1}}\mathfrak{E}^{i-1}_{f_{i-1},f'_{i-1}} \\
&\qquad P(f_i|f_{i-1})P(f'_i|f'_{i-1})\Big) \\
&= \sum_{f_{i-1}=1}^{K} P(f_i|f_{i-1}) \\
&\qquad \sum_{f'_{i-1}=1}^{K}\mathfrak{F}^{i-1}_{f_{i-1},f'_{i-1}}\mathfrak{E}^{i-1}_{f_{i-1},f'_{i-1}}P(f'_i|f'_{i-1}) && (4.10)
\end{aligned}$$

for every $f_i, f'_i \in \{1,\ldots,K\}$ and $i = 2,\ldots,n$, where

$$\mathfrak{E}^i_{f_i,f'_i} = \sum_{h_i,h'_i\in\{0,1\}} P(h_i|f_i)P(h'_i|f'_i)P(\mathbf{r}_i|G_i = h_i + h'_i) \qquad (4.11)$$

The inner sum in equation (4.10) is independent of f_i, and so its repeated computation can be avoided by replacing (4.10) with:

$$\begin{aligned}
\mathfrak{C}^i_{f_{i-1},f'_i} &= \sum_{f'_{i-1}=1}^{K}\mathfrak{F}^{i-1}_{f_{i-1},f'_{i-1}}\mathfrak{E}^{i-1}_{f_{i-1},f'_{i-1}}P(f'_i|f'_{i-1}) && (4.12)\\
\mathfrak{F}^i_{f_i,f'_i} &= \sum_{f_{i-1}=1}^{K} P(f_i|f_{i-1})\mathfrak{C}^i_{f_{i-1},f'_i} && (4.13)
\end{aligned}$$

A similar optimization can be applied when computing the backward probabilities, resulting in the following recurrence:

$$\mathfrak{B}^n_{f_n,f'_n} = 1 \qquad (4.14)$$

$$\mathfrak{D}^i_{f_{i+1},f'_i} = \sum_{f'_{i+1}=1}^{K} \mathfrak{B}^{i+1}_{f_{i+1},f'_{i+1}} \mathfrak{E}^{i+1}_{f_{i+1},f'_{i+1}} P(f'_{i+1}|f'_i) \quad (4.15)$$

$$\mathfrak{B}^i_{f_i,f'_i} = \sum_{f_{i+1}=1}^{K} P(f_{i+1}|f_i) \mathfrak{D}^i_{f_{i+1},f'_k} \quad (4.16)$$

Forward and backward probabilities can thus be computed in $O(nK^3)$ by using recurrences (4.9), (4.12), and (4.13), respectively (4.14), (4.15), and (4.16), resulting in an overall runtime of $O(m + nK^3)$.

4.3.2 Greedy Algorithm

Our second MGP algorithm picks SNP genotypes in left-to-right order using a greedy strategy. Notice that $P(\mathbf{g}|\mathbf{r}) \propto P(\mathbf{g})P(\mathbf{r}|\mathbf{g}) = P(g_1)P(\mathbf{r}_1|g_1) \prod_{i=2}^{n} [P(g_i|\mathbf{g}_{1:i-1})P(\mathbf{r}_i|g_i)]$. After having picked $\widehat{g}_1, \ldots, \widehat{g}_{i-1}$ in first $i-1$ iterations, in next iteration the algorithm makes its selection so that to maximize the i-th term of the above product:

Greedy Algorithm

1. $\widehat{g}_1 \leftarrow \text{argmax}_{g_1} P(g_1)P(\mathbf{r}_1|g_1)$
2. *For each* $i = 2, \ldots, n$,

   $\widehat{g}_i \leftarrow \text{argmax}_{g_i} P(g_i|\widehat{g}_1, \ldots, \widehat{g}_{i-1})P(\mathbf{r}_i|g_i)$
3. *Return* $\widehat{\mathbf{g}} = (\widehat{g}_1, \ldots, \widehat{g}_n)$

An $O(m + nK^3)$ time implementation of the algorithm is given as follows

Since for fixed $\widehat{g}_1, \ldots, \widehat{g}_{i-1}$ $P(g_i|\widehat{g}_1, \ldots, \widehat{g}_{i-1}) \propto P(\widehat{g}_1, \ldots, \widehat{g}_{i-1}, g_i)$, the maximization in Step 2 of the greedy algorithm can equivalently be restated as $\widehat{g}_i \leftarrow \text{argmax}_{g_i} P(\widehat{g}_1, \ldots, \widehat{g}_{i-1}, g_i)P(\mathbf{r}_i|g_i)$. Thus, we need to evaluate $P(\widehat{g}_1, \ldots, \widehat{g}_{i-1}, g_i)$

for for every $i = 2, \ldots, n$ and $g_i \in \{0, 1, 2\}$. These marginal probabilities can be computed efficiently using a forward algorithm as follows. For every $f_i, f_i' \in \{1, \ldots, K\}$, $i \in \{1, \ldots, n\}$, let $\mathcal{F}^i_{f_i,f_i'}(g_i) := P(\widehat{g}_1, \ldots, \widehat{g}_{i-1}, g_i, f_i, f_i')$, which we refer to as the *forward probability* associated with the partial multilocus genotype $(\widehat{g}_1, \ldots, \widehat{g}_{i-1}, g_i)$ and the pair of founder states f_i, f_i' at locus i. The forward probabilities can be computed using the recurrence:

$$\begin{aligned}
\mathcal{F}^1_{f_1,f_1'}(g_1) &= P(f_1)P(f_1')\mathcal{E}^1_{f_1,f_1'}(g_1) && (4.17)\\
\mathcal{F}^i_{f_i,f_i'}(g_i) &= \sum_{f_{i-1}=1}^{K}\sum_{f_{i-1}'=1}^{K} \mathcal{F}^{i-1}_{f_{i-1},f_{i-1}'}(\widehat{g}_{i-1})P(f_i|f_{i-1})P(f_i'|f_{i-1}')\mathcal{E}^i_{f_i,f_i'}(g_i) \\
&= \mathcal{E}^i_{f_i,f_i'}(g_i)\sum_{f_{i-1}=1}^{K} P(f_i|f_{i-1}) \sum_{f_{i-1}'=1}^{K} \mathcal{F}^{i-1}_{f_{i-1},f_{i-1}'}(\widehat{g}_{i-1})P(f_i'|f_{i-1}') && (4.18)
\end{aligned}$$

where $f_i, f_i' \in \{1, \ldots, K\}$, $i \in \{2, \ldots, n\}$, and

$$\mathcal{E}^i_{f_i,f_i'}(g_i) = \sum_{\substack{h_i,h_i' \in \{0,1\} \\ h_i + h_i' = g_i}} P(h_i|f_i)P(h_i'|f_i') \qquad (4.19)$$

The inner sum in (4.18) is independent of f_i and g_i, and so its repeated computation can be avoided by replacing (4.18) with:

$$\begin{aligned}
\mathcal{C}^i_{f_{i-1},f_i'} &= \sum_{f_{i-1}'=1}^{K} \mathcal{F}^{i-1}_{f_{i-1},f_{i-1}'}(\widehat{g}_{i-1})P(f_i'|f_{i-1}') && (4.20)\\
\mathcal{F}^i_{f_i,f_i'}(g_i) &= \mathcal{E}^i_{f_i,f_i'}(g_i)\sum_{f_{i-1}=1}^{K} P(f_i|f_{i-1})\mathcal{C}^i_{f_{i-1},f_i'} && (4.21)
\end{aligned}$$

By using recurrences (4.17), (4.20), and (4.21), all forward probabilities can be computed in $O(nK^3)$ time. Each multilocus genotype probability $P(\widehat{g}_1, \ldots, \widehat{g}_{i-1}, g_i)$ can then obtained as $\sum_{f_i=1}^{K}\sum_{f_i'=1}^{K} \mathcal{F}^i_{f_i,f_i'}(g_i)$, yielding an overall runtime of $O(m + nK^3)$ for the greedy algorithm (the $O(m)$ term comes from the preprocessing step

needed to compute conditional probabilities $P(\mathbf{r}_i|g_i)$).

4.3.3 Markov Approximation Algorithm

Our third algorithm uses dynamic programming to optimize an approximation of the posterior probability based on assuming Markovian dependence between SNP genotypes. As noted above, $P(\mathbf{g}|\mathbf{r}) \propto P(\mathbf{g})P(\mathbf{r}|\mathbf{g}) = P(g_1)P(\mathbf{r}_1|g_1)\prod_{i=2}^{n}[P(g_i|\mathbf{g}_{1:i-1})(\mathbf{r}_i|g_i)]$. If G_i is independent of $G_1, \ldots G_{i-2}$ conditional on G_{i-1}, then $P(g_i|\mathbf{g}_{1:i-1}) = P(g_i|g_{i-1})$ and we can view the genotypes as being generated by an inhomogeneous Markov chain with transition probabilities $P(g_i|g_{i-1})$. Although the above conditional independence assumption need not hold in our model, since long range dependencies are possible through the founder haplotypes, it appears to be a reasonable approximation. Thus, we seek a multilocus genotype $\widehat{\mathbf{g}}$ maximizing the posterior probability computed under the assumtion of Markovian dependence, i.e.,

$$\widehat{\mathbf{g}} = \text{argmax}_{\mathbf{g}} P(g_1)P(\mathbf{r}_1|g_1)\prod_{i=2}^{n}[P(g_i|g_{i-1})(\mathbf{r}_i|g_i)] \tag{4.22}$$

The optimum in (4.22) can be found efficiently by dynamic programming. Let $M^l(g_l)$ denote $\max_{g_1,\ldots,g_{l-1}} P(g_1)P(\mathbf{r}_1|g_1)\prod_{i=2}^{l}[P(g_i|g_{i-1})(\mathbf{r}_i|g_i)]$.

Markov Approximation Algorithm

1. *For each* $g_1 = 0, 1, 2$, $M^1(g_1) \leftarrow P(g_1)P(\mathbf{r}_1|g_1)$
2. *For each* $i = 2, \ldots, n$ *and* $g_i = 0, 1, 2$,

   $M^i(g_i) \leftarrow P(\mathbf{r}_i|g_i)\max_{g_{i-1}} M^{i-1}(g_{i-1})P(g_i|g_{i-1})$
3. $\widehat{g}_n \leftarrow \text{argmax}_{g_n} M^n(g_n)$
4. *For each* $i = n, \ldots, 2$,

$\widehat{g}_{i-1} \leftarrow \text{argmax}_{g_{i-1}} M^{i-1}(g_{i-1}) P(\widehat{g}_i | g_{i-1})$

5. *Return* $\widehat{\mathbf{g}} = (\widehat{g}_1, \ldots, \widehat{g}_n)$

The above dynamic programming algorithm requires $O(n)$ time assuming that probabilities $P(g_1)$ of first SNP locus genotypes and conditional probabilities $P(g_i|g_{i-1})$ are available. As described below, the latter can be computed in $O(nK^2)$ time, yielding an overall runtime of $O(m + nK^2)$.

The Markov approximation algorithm requires computing genotype probabilities $P(g_1)$ for the first SNP locus and all conditional probabilities of the form $P(g_i|g_{i-1})$. The former are given by $P(g_1) = \sum_{f_1=1}^{K} \sum_{f_1'=1}^{K} P(f_1) P(f_1') \mathcal{E}^1_{f_1, f_1'}(g_1)$ where $\mathcal{E}^1_{f_1, f_1'}(g_1)$ is given by (4.19). Since $P(g_i|g_{i-1}) = P(g_{i-1}, g_i)/P(g_{i-1})$, it suffices to compute all probabilities of the form $P(g_{i-1}, g_i)$ and $P(g_i)$. We next show how to compute these probabilities in $O(nK^2)$, resulting in an overall runtime of $O(m + nK^2)$ for the algorithm.

We begin by computing in $O(nK^2)$ probabilities $P(f_i)$ and $P(f_i')$, for every $f_i, f_i' \in \{1, \ldots, K\}$, $i = 2, \ldots, n$, using the recurrences

$$\begin{aligned} P(f_i) &= \sum_{f_{i-1}=1}^{K} P(f_i | f_{i-1}) \\ P(f_i') &= \sum_{f_{i-1}'=1}^{K} P(f_i' | f_{i-1}') \end{aligned}$$

Next, we compute in $O(nK)$ time all probabilities $P(h_i, h_i')$ for every $i = 1, \ldots, n$

and $h_i, h'_i \in \{0,1\}$ using the factorization

$$P(h_i, h'_i) = \left(\sum_{f_i=1}^{K} P(f_i)P(h_i|f_i)\right)\left(\sum_{f'_i=1}^{K} P(f'_i)P(h'_i|f'_i)\right)$$

This allows computing in $O(nK)$ all probabilities $P(g_i)$ using

$$P(g_i) = \sum_{\substack{h_i,h'_i \\ h_i+h'_i=g_i}} P(h_i, h'_i)$$

To compute $P(g_{i-1}, g_i)$ we use a similar method. We start by computing $P(h_{i-1}, h_i)$ for every $h_{i-1}, h_i \in \{0,1\}$ using the factorization

$$P(h_{i-1}, h_i) = \sum_{f_{i-1}=1}^{K} P(f_{i-1})P(h_{i-1}|f_{i-1})\alpha^i(f_{i-1}, h_i)$$

where

$$\alpha^i(f_{i-1}, h_i) = \sum_{f_i=1}^{K} P(f_i|f_{i-1})P(h_i|f_i)$$

This requires $O(nK^2)$ time. Probabilities $P(h'_{i-1}, h'_i)$ are computed within the same time bound using similar recurrences. Finally, we compute all probabilities $P(g_{i-1}, g_i)$ in $O(n)$ using

$$P(g_{i-1}, g_i) = \sum_{\substack{h_{i-1},h'_{i-1} \\ h_{i-1}+h'_{i-1}=g_{i-1}}} \sum_{\substack{h_i,h'_i \\ h_i+h'_i=g_i}} P(h_{i-1}, h_i)P(h'_{i-1}, h'_i)$$

4.4 Results

4.4.1 Datasets

We evaluated the HMM-based posterior decoding algorithm on shotgun sequencing datasets generated using three different sequencing technologies, as follows:

1. Watson 454: A set of 74.4 million reads downloaded from the NCBI SRA database (submission number SRA000065). The reads, with an average length of ∼265 bp, were generated using the Roche 454 FLX platform as part of James Watson's personal genome project. This is a subset of the 106.5 million 454 reads analyzed in [23]. Unless noted otherwise, the haplotype panel used to train identical HMM models for the maternal and paternal populations was obtained by phasing CEU trio genotypes from HapMap r23a [15] using the ENT algorithm of [30] and retaining parent haplotypes from each trio. As in [23], genotype calling accuracy was assessed using the SNP genotypes determined using duplicate hybridization experiments with Affymetrix 500k microarrays (only concordant genotypes were retained in the test set).

2. NA18507 Illumina: A set of 525 million paired-end reads downloaded from the NCBI SRA database (submission number: SRA000271). These 36bp reads, which were generated using the Illumina Genome Analyzer from a HapMap Yoruban individual identified as NA18507, are a subset of the dataset analyzed by [9]. For the analysis of this dataset the HMM models for maternal and paternal populations were trained using YRI haplotypes from HapMap r22, excluding the haplotypes of the YRI trio that contains NA18507. As gold standard we used the genotypes published as part of HapMap r22 for individual NA18507.

Table 4.1: Summary statistics for the three datasets used in evaluation

Dataset	Raw Reads	Raw Sequence	Mapped Reads	Test SNPs	Avg. Mapped SNP coverage
Watson 454	$74.2M$	$19.7Gb$	$49.8M$ (67%)	$443K$	5.85×
NA18507 Illumina	$525M$	$18.9Gb$	$397M$ (78%)	$2.85M$	6.10×
NA18507 SOLiD	$764M$	$21.15Gb$	$324M$ (42%)	$2.85M$	3.21×

3. NA18507 ABI SOLiD: A set of 764 million single ABI SOLiD reads with length between 20 and 44 bp, downloaded from the NCBI SRA database (submission number: SRA000272). The reads, also generated from the HapMap NA18507 individual, are a subset of those analyzed by [52]. HMM models and gold standard genotypes were determined in the same way as for the NA18507 Illumina dataset.

4.4.2 Read Mapping

We mapped 454 reads on build 36.3 of the reference human genome using the NUCMER tool of the MUMmer package [43] with default parameters. We discarded alignments matching less than 90% of the reference or with 10 or more errors (mismatches or indels). We then discarded surviving reads with multiple matching positions. We mapped Illumina and SOLiD reads using MAQ version 0.68 [46] with default parameters. We discarded alignments with mapping probability less than 0.9 or with sum of quality scores of mismatching bases higher than 60. Filtering was performed using the "submap" command of MAQ. Table 1 shows the number of initial and mapped reads for each dataset, the number of SNPs, and the average coverage per SNP obtained after mapping.

4.4.3 Genotype Accuracy

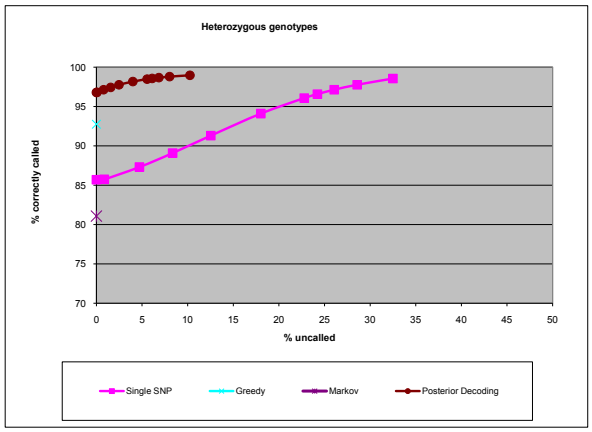
Fig. 4.3 gives accuracies for the three calling algorithms as described in Section 4.3. Since the calling accuracy achieved by the three likelihood functions vary, for the remaining experiments we use the best (posterior decoding) algorithm.

For each dataset of m mapped reads, we created four more different subsets of size $m/16$, $m/8$, $m/4$ and $m/2$ by picking reads at random, to evaluate the effects of read coverage on genotype calling. We called genotypes taking each subset as input and using separately the binomial test of [23] (with a threshold of 0.01), the single SNP posterior probability computed as described in Section 4.2.2, and by using the HMM posterior decoding algorithm presented in subsection 4.3.1. We measured the accuracy of a genotype calling method on a particular dataset by computing the percentage of SNP genotype calls that match the gold standard. As in previous papers [9, 52, 23], we separately report accuracy for homozygous and heterozygous SNPs.

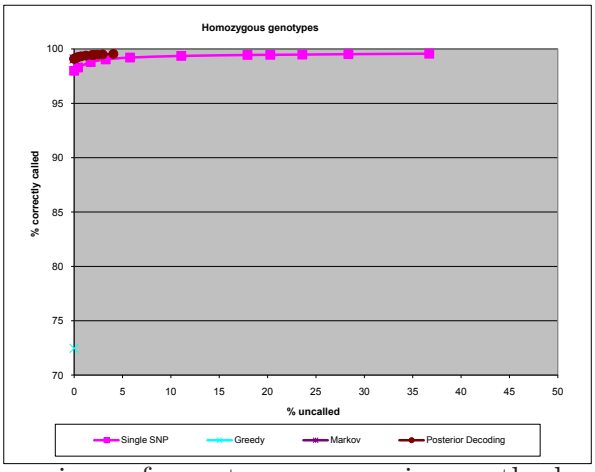
Figure 4.4(a) shows that, for both homozygous and heterozygous SNPs, the HMM-posterior decoding algorithm has higher accuracy than the binomial test at every considered coverage for the Watson 454 dataset. The improvement in accuracy is more pronounced for heterozygous SNPs, and as the average coverage goes down. This is not surprising since, as previously noted by [22, 23, 52], at low average coverage there is an increasingly high probability of not covering at least one of the alleles at a heterozygous SNP, and coverage of each allele is an implicit requirement of the binomial test method.[2] In contrast, the HMM-posterior algorithm does not have such a coverage requirement, and indeed, can accurately call genotypes even in the absence of any read coverage.

The above behavior is confirmed by the experiments on reads generated from

[2]The binomial test used by [22,23] actually requires for each allele to be covered at least twice; in this chapter we used the more relaxed requirement of covering each allele at least once.



(a)



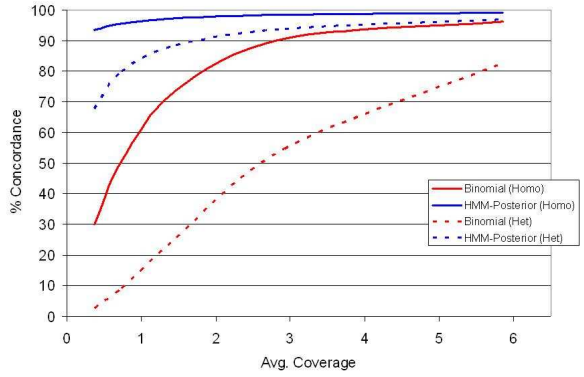
(b)

Figure 4.3: Comparison of genotype sequencing methods: Single SNP vs. Posterior Decoding vs. Greedy vs. Markov Approximation; Heterozygous (a), and Homozygous (b).

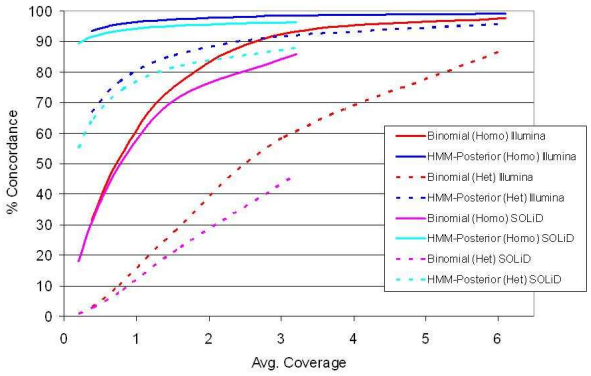
the NA18507 HapMap individual (Figure 4.4(b)). Genotype calling accuracy achieved on Illumina reads at various average coverages is similar to that observed for the Watson 454 reads. However, the accuracy achieved on SOLiD reads is slightly lower, even when normalized by average coverage. This is probably happens due at least part current mapping tools such as MAQ are better suited for reads in nucleotide space rather than in color space.

Since both methods based on posterior probabilities include as output the posterior probability of each genotype for each SNP, a minimum threshold can be applied to this probability leaving uncalled SNPs with low posterior probability. Figure 4.5(a) shows how the posterior decoding algorithm performed better than posterior based on allele frequencies for different percentages of uncalled SNPs obtained by varying the minimum threshold. This can be seen also in the results for heterozygous SNPs shown in Figure 4.5(b). The full set of Watson 454 reads was used in these experiments. Although no SNPs are left uncalled by the binomial test, we included the accuracy of this test on the plot for the Watson reads set as a single data point.

We performed some additional experiments to test how the posterior decoding algorithm behaves under different circumstances. Figure 4.6(a) shows the percentage of concordance for different local recombination rates. The percentage of SNPs in each category is included below in dashed lines. The plot shows how the accuracy increases as the recombination rate decreases, which is expected because in regions with low recombination rates the LD information is more helpful. Figure 4.6(b) shows the percentage of concordance for different SNP coverages. The accuracy increases with the SNP coverage until certain limit after which the results become unpredictable. This could be the effect of SNPs on repetitive regions or sequencing artifacts that distort the allele counts. Finally, Figure 4.6(c) shows the percentage of concordance for different panel sizes. For this experiment we

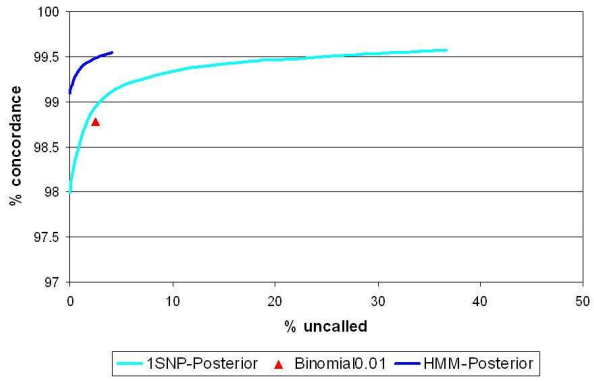


(a) Watson 454 reads

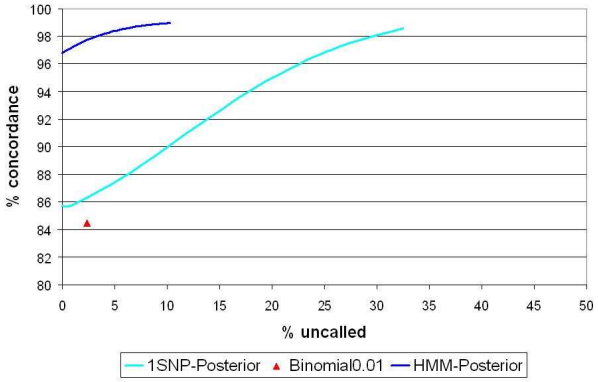


(b) NA18507 Illumina and SOLiD reads

Figure 4.4: Comparison between binomial and Multilocus genotype calling on percentage of concordance between predicted and gold standard genotype for different average coverages on the Watson dataset (a) and on the NA18507 datasets (b). Bold lines correspond to homozygous SNPs while dotted lines correspond to heterozygous SNPs

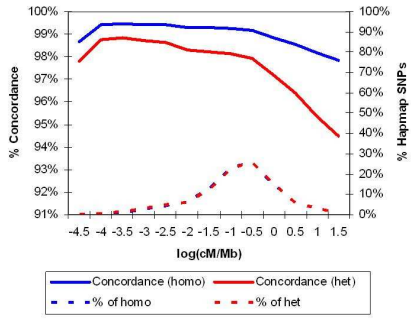


(a) Homozygous

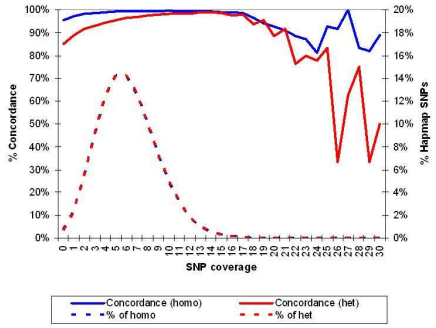


(b) Heterozygous

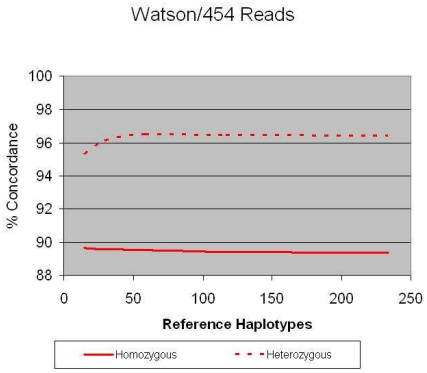
Figure 4.5: Comparison between single posterior and multilocus genotype calling on percentage of concordance between predicted and gold standard genotype for different probability thresholds expressed as uncalled genotype rates on the Watson dataset. Results of binomial genotype calling are shown as a single datapoint



(a) Recombination



(b) SNP coverage



(c) Panel Size

Figure 4.6: Effects of the recombination rate (a), SNP coverage (b) and Panel size (c) on concordance between predicted and gold standard genotype on the Watson dataset.

considered 242 different haplotypes available in Hapmap3 for the CEU population and, as in the experiments with variable coverage, we produced subsets of $n/16$, $n/8$, $n/4$ and $n/2$ haplotypes. The plot shows that not significant improvement is achieved by increasing the reference panel size and so just a few unrelated individuals of the population provide the LD information needed to improve the SNP calling accuracy.

4.5 Conclusions and Future Work

In this chapter we introduced a statistical model for multi-locus genotyping that integrates shotgun sequencing data with LD information extracted from a reference panel. Although finding the multi-locus genotype with maximum posterior probability under the integrated model is NP-Hard, experimental results suggest that a simple posterior decoding algorithm produces highly accurate genotype calls even from low-coverage sequencing data. Compared to current LD-oblivious genotype calling methods, our method allows researchers to achieve a desired accuracy target with reduced sequencing costs. For example, genotype calling accuracy achieved at 5-6× average coverage by a previously proposed binomial test is matched by the HMM-based posterior decoding algorithm using less than 1/4 of the reads. For a mapped coverage depth of 5-6× the HMM-based posterior decoding algorithm already achieves an accuracy comparable to that of microarray platforms, potentially over a much larger set of variants. Indeed, the number of interrogated variants is limited only by the available reference panels, and this limitation is expected to become less significant as ongoing large-scale sequencing efforts such as the 1000 Genomes project generate ever denser sets of genomic variants for an increasing number of human populations.

In ongoing work we are exploring efficient algorithms for LD-based haplotype reconstruction from shotgun sequencing reads. In particular, we seek to extend our multi-locus model in order to capture co-occurrence of SNP alleles in reads or mate-pairs; as read lengths continue to grow such co-occurrence information will become increasingly valuable.

Also, we are currently expanding our work to take population-level data into account as opposed to a single individual. We consider the problem of inferring genotypes from shotgun reads collected in a population. A potential benefit for

this genotyping approach is great cost reduction in accurately calling genotypes at very dense SNP sites.

In population-based genotype calling, we use the same HF-HMM structure as described in section 4.2.3. However, the training of this HMM via a reference panel of haplotypes is not done. Instead, we propose using the popular EM algorithm to train our HMM model based on the population-level provided. We choose an initial configuration and iteratively improve the likelihood value. To avoid getting stuck in a local maximum, we run multiple different initial configurations. An outline of our approach follows.

The input contains shotgun reads from a number of diploid individuals and we know which read comes from which individuals. We assume the reference genome is known, and reads mapping can be done with relatively small chance of error. We also assume that either through preprocessing or from existing knowledge, the SNP sites have been properly identified.

We are given a set of shotgun reads $R = (R_1, R_2, \ldots, R_n)$ that are from n diploid individuals from a population, such that reads R_i contains a number of reads for individual i. We assume the reads are short (e.g. from the Solexa sequencer), and thus one read only cover at most one SNP site. For simplicity, we do not consider reads with paired ends. That is, each read covers a contiguous (and short) genomic region. We further denote the set of reads in R_i that covers SNP site j as $R_{i,j}$. Note that there may be multiple reads in $R_{i,j}$. The goal is to infer the genotypes $G = (g_1, g_2, \ldots, g_n)$ from the given shotgun reads.

Our model and the method based on it is closely related to the previous work in section 4.2.3. Note that the difference from the previous work is that we do not have given genotypes, but just low coverage shotgun reads.

When the coverage is low, there is great uncertainty on the genotypes by the observed shotgun reads at a given site. We use the HMM as shown in figure

4.1, to represent the generation of a haplotype from a set of K founders. We currently still assume K is fixed. A founder F_i has m state, where each state $F_{i,j}$ with an emission probability $\theta_{i,j}$ to emit allele 1, and $1 - \theta_{i,j}$ to emit allele 0. To model recombination, a transition from $F_{i_1,j-1}$ to $F_{i_2,j}$ is allowed with probability $\tau_{i_1,i_2,j}$.Then a haplotype h in the population is modeled as the sequence of emitted symbols from M. To model genotypes, we can simply pair up two identical M, one for paternal haplotypes and one for maternal haplotypes.We use the popular EM algorithm to train our HMM. We choose an initial configuration and iteratively improve the likelihood value. To avoid getting stuck in a local maximum, we run EM multiple times from different initial configurations.

Initial Configuration

. We start EM by initializing the transition probability τ uniformly (i.e., $\tau_{i_1,i_2,j} = 1/K$ for all i_1, i_2, j). We then initialize emission probability of state $F_{i,j}$ based on the expected allele frequency at site s_j as follows. Suppose there are $n_{j,0}$ reads with allele 0 and $n_{j,1}$ reads with allele 1. We assign $\theta_{i,j}$ to $(n_{j,1} + 1)/(n_{j,0} + n_{j,1} + 2)$. Then we perturb the parameters by multiplying each parameter value by e^X. Here X is randomly drawn from $[-\eta, \eta]$. This way, each EM run starts with a different parameter configuration.

EM Algorithm

We want to find parameters (θ, τ) that maximize $P(R|\theta, \tau)$. The EM algorithm starts with initial values (θ^0, τ^0) and iteratively improves the parameters by setting

$$\theta^{t+1}, \tau^{t+1} = argmax_{\theta,\tau} \sum_{Z} P(Z|R, \theta^t, \tau^t) log P(R, Z|\theta, \tau)$$

Choosing Z is often important for the efficiency of the EM algorithm. Here

we choose $Z = (G, T, H)$ to make the complete likelihood $P(R, Z|\theta, \tau)$ easy to compute. Here, $G = (g_1, \ldots, g_n)$ is the set of multi-locus genotypes for the sampled individuals. T is a $n \times m \times 2$ matrix, where $T[i, j, k]$ indicates for diploid individual i, its paternal haplotype (if $k = 0$) or maternal haplotype (if $k = 1$) takes $F_{i,j}$ as its founder state. In other words, T specifies $2n$ paths across M, one for each of the $2n$ haplotypes. H is a $n \times m \times 2$ binary matrix, which indicates the emitted $2n$ haplotypes.

Chapter 5

Conclusions

The need for highly accurate and efficient methods for genotype data analysis is expected to increase in the future as genotype association studies grow in size. Ongoing efforts such as the 1000 Genomes project generate ever denser sets of genomic variants for an increasing number of human populations. Additionally, rapid advances in SNP genotyping technologies are expected to continue, and will be able to produce even larger amounts of population genotype data, accelerating the discovery of genes associated with common human diseases. For example, massively parallel sequencers like Roche/454 FLX Titanium, Illumina Genome Analyzer II, ABI SOLiD 3 and Helicos HeliScope are able to provide millions of short reads in a single run. In this thesis we attempted to address this need by developing scalable algorithms for several analysis problems.

Genotype Error Detection Using HMMs of Haplotype Diversity

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.

We also introduced GEDI, a software package that implements efficient algorithms for performing common tasks in the analysis of population genotype data, including error detection and correction, imputation of both randomly missing and untyped genotypes, and genotype phasing. By varying the user-selected parameters, we were able to display parameter settings for GEDI that yield an excellent tradeoff between imputation accuracy and runtime. We have also shown that accuracy improves significantly when using reference panels larger than the commonly used Hapmap panels, particularly in conjunction with the increase in the number of HMM founders. Accuracy further benefits from exploiting available pedigree information and performing genotype error correction and missing data recovery prior to imputation.

Imputation-based local ancestry inference in admixed populations

We proposed a novel algorithm for imputation-based local ancestry inference in Chapter 3. Experiments on simulated data show that our method exploits ancestral haplotype information more effectively than previous methods, yielding consistently accurate estimates of local ancestry for a variety of admixed populations. Indeed, our method is competitive with best existing methods in the case of admixtures of two distant ancestral populations, and is significantly more accurate than previous methods for admixtures of closely related populations such as the JPT and CHB populations from HapMap. We also show that accurate local ancestry estimates lead to improved accuracy of untyped SNP genotype imputation for admixed individuals. In ongoing work we are exploring methods that iteratively alternate between rounds of imputation-based ancestry inference and ancestry-based imputation for further improvements in accuracy. We are also conducting experiments to characterize the accuracy of our imputation-based local ancestry inference methods in the case of admixtures of more than two ancestral populations.

Single Individual Genotyping from Low-Coverage Sequencing Data

We introduced a statistical model for multi-locus genotyping that integrates shotgun sequencing data with LD information extracted from a reference panel. Although finding the multi-locus genotype with maximum posterior probability under the integrated model is NP-Hard, experimental results suggest that a simple posterior decoding algorithm produces highly accurate genotype calls even from low-coverage sequencing data. Compared to current LD-oblivious genotype calling methods, our method allows researchers to achieve a desired accuracy target with reduced sequencing costs. For example, genotype calling accuracy achieved at 5-6 average coverage by a previously proposed binomial test is matched by the HMM-based posterior decoding algorithm using less than 1/4 of the reads. For a mapped coverage depth of 5-6 the HMM-based posterior decoding algorithm already achieves an accuracy comparable to that of microarray platforms, potentially over a much larger set of variants. Indeed, the number of interrogated variants is limited only by the available reference panels, and this limitation is expected to become less significant as ongoing large-scale sequencing efforts such as the 1000 Genomes project generate ever denser sets of genomic variants for an increasing number of human populations. Also, we are currently expanding on the work which will take population-level data into account as opposed to a single individual. We consider the problem of inferring genotypes from shotgun reads collected in a population. The main benefit for our LD-based genotyping approach is the significant reduction in sequencing costs, which allows researchers to perform more and larger population sequencing studies within the same budget.

Bibliography

[1] The 1000 genomes project consortium.

[2] G.R. Abecasis, S.S. Cherny, and L.R. Cardon. The impact of genotyping error on family-based analysis of quantitative traits. *European Journal of Human Genetics*, 9:130–134, 2001.

[3] G.R. Abecasis, S.S. Cherny, W.O.C. Cookson, and L.R. Cardon. Merlin-rapid analysis of dense genetic maps using sparse gene flow trees. *Nature Genetics*, 30:97–101, 2002.

[4] K. Ahn, C. Haynes, W. Kim, R.S. Fleur, D. Gordon, and S.J. Finch. The effects of SNP genotyping errors on the power of the Cochran-Armitage linear trend test for case/control association studies. *Annals of Human Genetics*, 71:249–261, 2007.

[5] A.S. Allen and G.A. Satten. A novel haplotype-sharing approach for genome-wide case-control association studies implicates the calpastatin gene in Parkinson's disease. *Genet. Epidemiol.,* Epub ahead of print, 2009.

[6] L.E. Baum, T. Petrie, G. Soules, and N. Weiss. A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *Annals of Mathematical Statistics*, 41:164–171, 1970.

[7] T. Becker and M. Knapp. Maximum-likelihood estimation of haplotype frequencies in nuclear families. *Genetics Epidemiology*, 27:21–32, 2004.

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

[9] D.R. Bentley *et al.* Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, 456:53–59, 2008.

[10] B. L. Browning and S. R. Browning. Haplotypic analysis of Wellcome Trust case control consortium data. *Human Genetics*, 123(3):273–280, 2008.

[11] K.F. Cheng. Analysis of case-only studies accounting for genotyping error. *Annals of Human Genetics*, 71:238–248, 2007.

[12] S.S. Cherny, G.R. Abecasis, W.O.C. Cookson, P.C. Sham, and L.R. Cardon. The effect of genotype and pedigree error on linkage analysis: Analysis of three asthma genome scans. *Genetics Epidemiology*, 21:S117–S122, 2001.

[13] International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature*, 409:860–921, 2001.

[14] International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature*, 431:931–945, 2004.

[15] The International HapMap Consortium. A second generation human haplotype map of over 3.1 million snps. *Nature*, 449:851–861, 2007.

[16] The Wellcome Trust Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447:661–678, 2007.

[17] D.J. Cutler, M.E. Zwick, M.M Carrasquillo, C.T.Yohn, K.P. Tobin, C. Kashuk, D.J. Matthews, N.A. Shah, E.E. Eichler, J.A.Warrington, and A. Chakravarti. High-throughput variation detection and genotyping using microarrays. *Genome Research*, 10, 2006.

[18] X. Di, H. Matsuzaki, T.A. Webster, E. Hubbell, G. Liu, S. Dong, D. Bartell, J. Huang, R. Chiles, G. Yang, M.-M. Shen, D. Kulp, G.C. Kennedy, R. Mei, K.W. Jones, and S. Cawley. Dynamic model based algorithms for screening and genotyping over 100k SNPs on oligonucleotide microarrays. *Bioinformatics*, 21(9):1958–1963, 2005.

[19] J. Douglas, M. Boehnke, and K. Lange. A multipoint method for detecting genotyping errors and mutations in sibling-pair linkage data. *American Journal of Human Genetics*, 66:1287–1297, 2000.

[20] J. Douglas, A. Skol, and M. Boehnke. Probability of detection of genotyping errors and mutations as inheritance inconsistencies in nuclear-family data. *American Journal of Human Genetics*, 70:487–495, 2002.

[21] F. Dudbridge. Pedigree disequilibrium tests for multilocus haplotypes. *Genet. Epidemiol.*, 25:115–221, 2003.

[22] Levy et al. The diploid genome sequence of an individual human. *PLos Biology*, 5(10).

[23] Wheeler et al. The complete genome of an individual by massively parallel nda sequencing. *Nature*, 452:872–876, 2008.

[24] B. Ewing and P. Green. Base-calling of automated sequencer traces using phred. ii. error probabilities. *Genome Research*, 8(3):186–194, 1998.

[25] S. Fine, Y. Singer, and N. Tishby. The hierarchical hidden Markov model: Analysis and applications. *Mach. Learn.*, 32(1):41–62, 1998.

[26] GAIN Imputation Working Group. In preparation. 2009.

[27] Z. Ghahramani and M.I. Jordan. Factorial hidden Markov models. *Mach. Learn.*, 29(2-3):245–273, 1997.

[28] D. Gordon, S.J. Finch, M. Nothnagel, and J. Ott. Power and sample size calculations for case-control genetic association tests when errors are present: Application to single nucleotide polymorphisms. *Human Heredity*, 54(1):22–33, 2002.

[29] D. Gordon, S.C. Heath, and J. Ott. True pedigree errors more frequent than apparent errors for single nucleotide poloymorphisms. *Human Heredity*, 49:65–70, 1999.

[30] A. Gusev, I.I. Mandoiu, and B. Pasaniuc. Highly scalable genotype phasing by entropy minimization. *IEEE/ACM Trans. on Computational Biology and Bioinformatics*, 5(2):252–261, 2008.

[31] K. Hao and X. Wang. Incorporating individual error rate into association test of unmatched case-control design. *Human Heredity*, 58:154–163, 2004.

[32] L. Hosking, S. Lumsden, K. Lewis, A. Yeo, L. McCarthy, A. Bansal, J. Riley, I. Purvis, and CF. Xu. Detection of genotyping errors by hardy-weinberg equilibrium testing. *European Journal of Human Genetics*, 12:395–399, 2004.

[33] L. Huang, C. Wang, and N.A. Rosenberg. The relationship between imputation error and statistical power in genetic association studies in diverse populations. *Am. J. Hum. Genet., in press*, 2009.

[34] International HapMap Consortium. A second generation human haplotype map of over 3.1 million SNPs. *Nature*, 449(7164):851–861, 2007.

[35] Y. Hernandez I. Mandoiu J. Duitama, S. Dinakar and Y. Wu. Single individual genotyping from low-coverage sequencing data. *manuscript in preparation*.

[36] J. Kennedy, I.I. Mandoiu, and B. Pasaniuc. Genotype error detection using hidden markov models of haplotype diversity. *Journal of Computational Biology*, 15(9):1155–1171, 2008.

[37] J. Kennedy, I.I. Mandoiu, and B. Pasaniuc. Genotype error detection using hidden markov models of haplotype diversity. Technical report, arXiv.org, Cornell University, 2009.

[38] J. Kennedy, I.I. Măndoiu, and B. Paşaniuc. Genotype error detection using hidden Markov models of haplotype diversity. In *Proc. 7th Workshop on Algorithms in Bioinformatics*, Lecture Notes in Computer Science, pages 73–84, 2007.

[39] J. Kennedy, I.I. Măndoiu, and B. Paşaniuc. Genotype error detection using hidden markov models of haplotype diversity. *J. Comput. Biol.*, 15(9):1155–1171, 2008.

[40] J. Kennedy, B. Pasaniuc, and I.I. Mandoiu. GEDI: Genotype error detection and imputation using hidden markov models of haplotype diversity, manuscript in preparation. software available at at `http://dna.engr.uconn.edu/software/gedi/` .

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

[42] M. Knapp and T. Becker. Impact of genotyping errors on type I error rate of the haplotype-sharing transmission/disequilibrium test (HS-TDT). *American Journal of Human Genetics*, 74:589–591, 2004.

[43] S. Kurtz *et al.* Versatile and open software for comparing large genomes. *Genome Biology*, 5(2):R12, 2004.

[44] S.M. Leal. Detection of genotyping errors and pseudo-SNPs via deviations from Hardy-Weinberg equilibrium. *Genet Epidemiol*, 29(3):204–214, 2005.

[45] Y. Li and G. R. Abecasis. Mach 1.0: Rapid haplotype reconstruction and missing genotype inference. *American Journal of Human Genetics*, 79:2290, 2006.

[46] Durbin R Li H, Ruan J. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Research*, 18(1):1851–1858, 2008.

[47] L. Liang, S. Zöllner, and G.R. Abecasis. GENOME: a rapid coalescent-based whole genome simulator. *Bioinformatics*, 23(12):1565–1567, 2007.

[48] D.Y. Lin, Y. Hu, and B.E. Huang. Simple and efficient analysis of disease association with missing genotype data. *Am. J. Hum. Genet.*, 83(4):535–539, 2008.

[49] W. Liu, T. Yang, W. Zhao, and G.A. Chase. Accounting for genotyping errors in tagging SNP selection. *Am. J. Hum. Genet.*, 71(4):467–479, 2007.

[50] R.B. Lyngso and C.N.S. Pedersen. The consensus string problem and the complexity of comparing hidden markov models. *Journal of Computer Systems Science*, 65(3):545–569, 2002.

[51] J. Marchini, B. Howie, S. Myers, G. McVean, and P. Donnelly. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nature Genetics*, 39:906–913, 2007.

[52] K.J. McKernan *et al.* Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding. *Genome Research*, 19:1527–1541, 2009.

[53] A.A. Mitchell, D. Cutler, and A. Chakravarti. Undetected genotyping errors cause apparent overtransmission of common alleles in the transmission/disequilibrium test. *American Journal of Human Genetics*, 72:598–610, 2003.

[54] N. Mukhopadhyaya, S.G. Buxbauma, and D.E. Weeks. Comparative study of multipoint methods for genotype error detection. *Human Heredity*, 58:175–189, 2004.

[55] NCI-NHGRI Working Group on Replication in Association Studies. Replicating genotype-phenotype associations. *Nature*, 447:655–660, 2007.

[56] Neale *et al.* Genome-wide association scan of attention deficit hyperactivity disorder. *Am. J. Med. Genet. Part B*, 147b(8):1337–1344, 2008.

[57] D.L. Nicolae, X. Wu., K. Miyake, and N.J. Cox. Gel: a novel genotype calling algorithm using emperical likelihood. *Bioinformatics*, 22:1942–1947, 2006.

[58] E. J. Parra, A. Marcini, J. Akey, J. Martinson, M. A. Batzer, R. Cooper, T. Forrester, D. B. Allison, R. Deka, R. E. Ferrell, et al. Estimating african american admixture proportions by use of population-specific alleles. *Am J Hum Genet*, 63(6):1839–1851, December 1998.

[59] B. Pasaniuc, J. Kennedy, and I.I. Mandoiu. Imputation-based local ancestry inference in admixed populations. *Proc. 5th International Symposium on Bioinformatics Research and Applications/2nd Workshop on Computational Issues in Genetic Epidemiology*, pages 221–233, 2009.

[60] B. Pasaniuc, S. Sankararaman, G. Kimmel, and E. Halperin. Inference of locus-specific ancestry in closely related populations (under review).

[61] F. Pompanon, A. Bonin, E. Bellemain, and P. Taberlet. Genotyping errors: causes, consequences and solutions. *Nature Review Genetics*, 6:847–859, 2005.

[62] Quake S.R. Pushkarev D, Neff N.F. Single-molecule sequencing of an individual human genome. *Nature Biotechnology*, 27(9):847–850, 2009.

[63] N. Rabbee and T.P. Speed. A genotype calling algorithm for affymetrix SNP arrays. *Bioinformatics*, 22:7–12, 2005.

[64] P. Rastas, M. Koivisto, H. Mannila, and E. Ukkonen. Phasing genotypes using a hidden Markov model. In *Bioinformatics Algorithms: Techniques and Applications*, pages 355–373. Wiley, 2008, preliminary version in *Proc. WABI 2005*.

[65] D. Reich and Patterson N. Will admixture mapping work to find disease genes? *Philos Trans R Soc Lond B Biol Sci*, 360:1605–1607, 2005.

[66] S. Sankararaman, G. Kimmel, E. Halperin, and M.I. Jordan. On the inference of ancestries in admixed populations. *Genome Research*, (18):668–675, 2008.

[67] S. Sankararaman, S. Sridhar, G. Kimmel, and E. Halperin. Estimating local ancestry in admixed populations. *American Journal of Human Genetics*, 8(2):290–303, 2008.

[68] P. Scheet and M. Stephens. 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, 2006.

[69] R. Schwartz. Algorithms for association study design using a generalized model of haplotype conservation. In *Proc. CSB*, pages 90–97, 2004.

[70] M. W. Smith, N. Patterson, J. A. Lautenberger, A. L. Truelove, G. J. McDonald, A. Waliszewska, B. D. Kessing, M. J. Malasky, C. Scafe, E. Le, et al. A high-density admixture map for disease gene discovery in african americans. *Am J Hum Genet*, 74(5):1001–1013, May 2004.

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

[72] E. Sobel, J.C. Papp, and K. Lange. Detection and integration of genotyping errors in statistical genetics. *Americal Journal of Human Genetics*, 70:496–508, 2002.

[73] M. Stephens and P. Scheet. Accounting for decay of linkage disequilibrium in haplotype inference and missing-data imputation. *American Journal of Human Genetics*, 76:449–462, 2005.

[74] A. Sundquist, E. Fratkin, C.B. Do, and S. Batzoglou. Effect of genetic divergence in identifying ancestral origin using hapaa. *Genome Research*, 18(4):676–682, 2008.

[75] H. Tang, M. Coram, P. Wang, X. Zhu, and N. Risch. Reconstructing genetic ancestry blocks in admixed individuals. *Am J Hum Genet*, 79:1–12, 2006.

[76] H. Tang, Peng J., and Pei Wang P.and Risch N.J. Estimation of individual admixture: Analytical and study design considerations. *Genetic Epidemiology*, 28:289–301, 2005.

[77] C. Tian, D. A. Hinds, R. Shigeta, R. Kittles, D. G. Ballinger, and M. F. Seldin. A genomewide single-nucleotide-polymorphism panel with high ancestry information for African American admixture mapping. *Am J Hum Genet*, 79:640–649, 2006.

[78] `http://www.hapmap.org/`.

[79] A.J. Viterbi. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory*, 13:260–269, 1967.

[80] J Wang *et al.* The diploid genome sequence of an Asian individual. *Nature*, 456:60–65, 2008.

[81] X. Wen and D. L. Nicolae. Association studies for untyped markers with TUNA. *Bioinformatics*, 24:435–437, 2008.

[82] Wilson RK Wendl M. Aspects of coverage in medical dna sequencing. *BMC Bioinformatics*, 9:239, 2008.

[83] X. Yuanyuan, M.R. Segal, J. Yang, and Y. Ru-Fang Y. A multi-array multi-SNP genotyping algorithm for affymetrix SNP microarrays. *Bioinformatics*, pages 1–7, 2006.

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

[85] Zeggini *et al.* Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. *Nat. Genet.*, 40(5):638–645, 2008.

[86] G. Zheng and X. Tian. The impact of diagnostic error on testing genetic association in case-control studies. *Statistics in Medicine*, 24:869–882, 2005.