Imputation-based local ancestry inference in admixed populations

Bogdan Paşaniuc[1], Justin Kennedy[2], and Ion Măndoiu[2]

[1] International Computer Science Institute, Berkeley, CA
bogdan@icsi.berkeley.edu
[2] CSE Department, University of Connecticut, Storrs, CT
{jlk02019,ion}@engr.uconn.edu

Abstract. Accurate inference of local ancestry from whole-genome genetic variation data is critical for understanding the history of admixed human populations and detecting SNPs associated with disease via admixture mapping. Although several existing methods achieve high accuracy when inferring local ancestry for individuals resulting from the admixture of genetically distant ancestral populations (e.g., African-Americans), ancestry inference in the case when ancestral populations are closely related remains challenging. Surprisingly, methods based on the analysis of allele frequencies at unlinked SNP loci currently outperform methods based on haplotype analysis, despite the latter methods seemingly receiving more detailed information about the genetic makeup of ancestral populations.

In this paper we propose a novel method for imputation-based local ancestry inference that exploits ancestral haplotype information more effectively than previous haplotype-based methods. Our method uses the 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 estimated within a neighborhood of the locus. Experiments on simulated data show that imputation-based ancestry assignment is competitive with best existing methods in the case of distant ancestral populations, and yields a significant improvement for closely related ancestral populations. Further demonstrating the synergy between imputation and ancestry inference, we also give results showing that the accuracy of untyped SNP genotype imputation in admixed individuals improves significantly when using estimates of local ancestry. The open source C++ code of our method, released under the GNU General Public Licence, is available for download at http://dna.engr.uconn.edu/software/GEDI-ADMX/.

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 [12]. 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 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 [19], SWITCH [13], HAPAA [18] 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 [14] uses an assumption of no recent recombination events within each window to estimate the ancestries, while WINPOP [9] 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 paper 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 [22] 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.

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-alelic 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

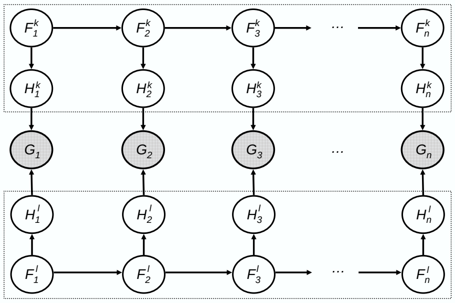

Fig. 1. Factorial HMM model for a multilocus SNP genotype $(G_1, \ldots, G_n)$ over an n-locus window within which one haplotype is inherited from ancestral population $\mathcal{P}_k$ and the other from ancestral population $\mathcal{P}_l$. For every locus i, F_i^k and H_i^k denote the founder haplotype, 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 $\mathcal{P}_l$.

at the corresponding locus, i.e., 0 and 2 encode homozygous major and minor genotypes, while 1 denotes a heterozygous genotype.

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 [4], genotype phasing [11,15], testing for disease association [6,16], and imputation of untyped SNP genotypes [5,7,8,15]. In this section we extend the imputation model in [5] 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) [3] referred to as $\mathcal{M}_{kl}$ and graphically represented in Figure 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 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 [1] 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 [6, 11].

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 $g \in \{0, 1, 2\}$

$$P_{\mathcal{M}_{kl}}(G_i = g|\mathbf{g}_{-i}) \propto P_{\mathcal{M}_{kl}}(\mathbf{g}[g_i \leftarrow g]) \tag{1}$$

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 2.2 temporarily marks as missing and imputes each SNP genotype, and thus requires computing probabilities (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:

$$\mathcal{F}^1_{f_1^k, f_1^l} = P(f_1)P(f_1') \tag{2}$$

$$\begin{aligned}\mathcal{F}^i_{f_i^k, f_i^l} &= \sum_{f_{i-1}^k=1}^{K_k} \sum_{f_{i-1}^l=1}^{K_l} \mathcal{F}^{i-1}_{f_{i-1}^k, f_{i-1}^l} \mathcal{E}^{i-1}_{f_{i-1}^k, f_{i-1}^l}(g_{i-1}) P(f_i^k|f_{i-1}^k) P(f_i^l|f_{i-1}^l) \\ &= \sum_{f_{i-1}^k=1}^{K_k} P(f_i^k|f_{i-1}^k) \sum_{f_{i-1}^l=1}^{K_l} \mathcal{F}^{i-1}_{f_{i-1}^k, f_{i-1}^l} \mathcal{E}^{i-1}_{f_{i-1}^k, f_{i-1}^l}(g_{i-1}) P(f_i^l|f_{i-1}^l) \quad (3)\end{aligned}$$

where

$$\mathcal{E}^i_{f_i^k, f_i^l}(g_i) = \sum_{\substack{h_i^k, h_i^l \in \{0,1\} \\ h_i^k + h_i^l = g_i}} P(h_i^k|f_i^k) P(h_i^l|f_i^l) \tag{4}$$

The innermost sum in (3) is independent of f_i^k, and so its repeated computation can be avoided by replacing (3) with:

$$\mathcal{C}^i_{f_{i-1}^k, f_i^l} = \sum_{f_{i-1}^l=1}^{K_l} \mathcal{F}^{i-1}_{f_{i-1}^k, f_{i-1}^l} \mathcal{E}^{i-1}_{f_{i-1}^k, f_{i-1}^l}(g_{i-1}) P(f_i^l|f_{i-1}^l) \tag{5}$$

$$\mathcal{F}^i_{f^k_i, f^l_i} = \sum_{f^k_{i-1}=1}^{K_k} P(f^k_i | f^k_{i-1}) \mathcal{C}^i_{f^k_{i-1}, f^l_i} \tag{6}$$

By using recurrences (2), (5), and (6), 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 $\mathcal{B}^i_{f^k_i, f^l_i} = P_{\mathcal{M}_{kl}}(f^k_i, f^l_i, g_{i+1}, \ldots, g_n)$ can be computed in $O(nK^3)$ time using similar recurrences:

$$\mathcal{B}^n_{f^k_n, f^l_n} = 1$$

$$\mathcal{D}^i_{f^k_{i+1}, f^l_i} = \sum_{f^l_{i+1}=1}^{K_l} \mathcal{B}^{i+1}_{f^k_{i+1}, f^l_{i+1}} \mathcal{E}^{i+1}_{f^k_{i+1}, f^l_{i+1}}(g_{i+1}) P(f^l_{i+1} | f^l_i)$$

$$\mathcal{B}^i_{f^k_i, f^l_i} = \sum_{f^k_{i+1}=1}^{K_k} P(f^k_{i+1} | f^k_i) \mathcal{D}^i_{f^k_{i+1}, f^l_i}$$

After computing forward and backward probabilities, posterior SNP genotype probabilities (1) can be evaluated in $O(K^2)$ time per SNP locus by observing that:

$$P_{\mathcal{M}_{kl}}(\mathbf{g}[g_i \leftarrow g]) = \sum_{f^k_i=1}^{K_k} \sum_{f^l_i=1}^{K_l} \mathcal{F}^i_{f^k_i, f^l_i} \mathcal{E}^i_{f^k_i, f^l_i}(g) \mathcal{B}^i_{f^k_i, f^l_i} \tag{7}$$

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

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 [9,13,14,18,19], 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 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

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 1) under each local ancestry model $\mathcal{M}_{kl}$
3. For each locus $i = 1, \ldots, n$,

$$\hat{a}_i \leftarrow \text{argmax}_{kl \in \mathcal{A}} \frac{1}{|W_i|} \sum_{j \in W_i} P_{\mathcal{M}_{kl}}(G_i = g_i | \mathbf{g}_{-i}) \tag{8}$$

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

Fig. 2. Single-window imputation-based ancestry inference algorithm.

and typically converges in a small number of iterations. As described in Section 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 [9, 14], 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 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 4 and Table 1) suggest that the multi-window strategy yields an average accuracy that is very close to (and, for some admixed populations, better

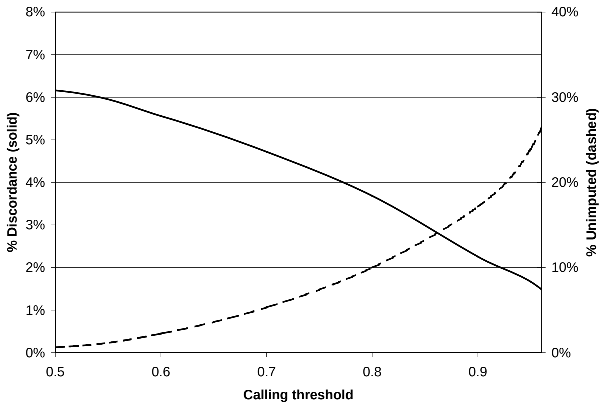

Fig. 3. Percentage of imputation errors (solid line) and unimputed genotypes (dashed line) at varying cutoff thresholds on posterior imputation probability for the WTCCC 1958 birth cohort dataset.

than) the maximum average accuracy achieved by running the single-window algorithm with any window size from the above set.

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 a second 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 [22]. Finally, we present results demonstrating the benefit of incorporating accurate local ancestry estimates when performing genotype imputation for admixed individuals.

3.1 SNP genotype imputation in homogeneous populations

To assess the accuracy achieved when imputing missing SNP genotypes based on the factorial HMM model described in Section 2.1, we used the 1,444 individuals of the 1958 birth cohort of the Wellcome Trust Case Control Consortium (WTCCC) [2]. For this homogeneous population imputation was performed using the GEDI package [5], based on a factorial model consisting of two identical left-to-right HMMs trained on CEU panel haplotypes from HapMap. SNP genotype imputation for admixed populations is further discussed in Section 3.3.

The individuals in the 1958 birth cohort were genotyped using the Affymetrix 500K GeneChip Assay. We masked as un-typed and then imputed 1% of the

SNPs on chromosome 22. We measured the error rate as the percentage of erroneously recovered genotypes from the total number of masked genotypes. Since the model provides the posterior probability for each imputed SNP genotype, one can get different tradeoffs between the error rate and the percentage of imputed genotypes by varying the cutoff threshold on posterior imputation probability. Figure 3 plots the achievable tradeoffs. For example, using a cutoff threshold of 0.95, HMM-based imputation has an error rate of 1.7%, with 24% of the genotypes left un-imputed.

3.2 Inference of local ancestry in admixed populations

The method described in Section 2.2 was implemented in an extension of the GEDI software package [5], 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 [19], SWITCH [13], and HAPAA [18]) are HMM based, while the other two (LAMP [14] and WIN-POP [9]) 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 information 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 [20], 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 [22] 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 [10, 17, 21]. 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.

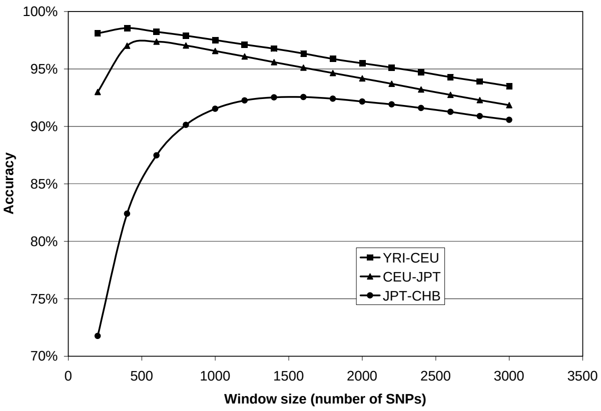

Fig. 4. Accuracy of local ancestry estimates obtained by GEDI-ADMX on the three HapMap admixtures using a single window of varying size.

Effect of window size on the local ancestry estimates. Figure 4 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 [9,14] 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 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 1). All remaining results were obtained using the multi-window approach.

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 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 popula-

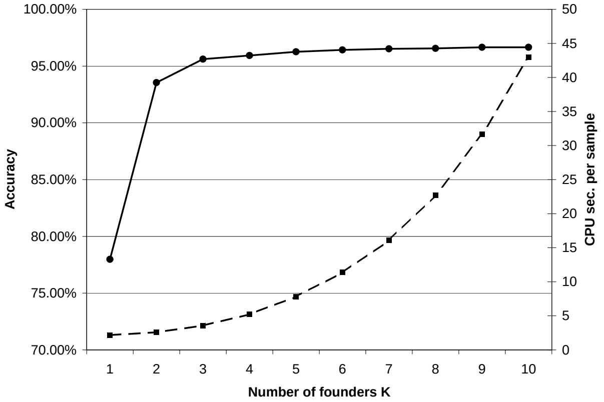

Fig. 5. GEDI-ADMX accuracy (solid line) and runtime (dashed line) for varying values of the number K of HMM founder haplotypes on the CEU-JPT dataset, consisting of $n = 38,864$ SNPs on Chromosome 1.

tions 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 5. 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 5 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 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 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 [9]. Notably, GEDI-ADMX accuracy on the JPT-CHB dataset exceeds the upper-bound for methods that do not model the LD. This underscores the

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

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

importance of exploiting ancestral haplotypes when performing local ancestry inference for admixtures of closely related populations.

3.3 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 3.2 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 2.2. We then imputed masked genotypes using the model in Section 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 [5], based on a factorial model similar to that in Section 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), respectively on (2) a haplotype list obtained by merging the panel haplotypes of the two ancestral populations (GEDI-2-Pop).

Table 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).

4 Discussion

In this paper 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 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.

Acknowledgments

BP was supported by NSF grant 0713254. IM and JK were supported in part by NSF CAREER award IIS-0546457 and NSF award DBI-0543365. This study makes use of data generated by the Wellcome Trust Case Control Consortium. A full list of the investigators who contributed to the generation of the data is available from www.wtccc.org.uk. Funding for the project was provided by the Wellcome Trust under award 076113.

References

1. L.E. Baum, T. Petrie, G. Soules, and N. Weiss. A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *Ann. Math. Statist.*, 41:164–171, 1970.
2. 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.
3. Z. Ghahramani and M.I. Jordan. Factorial hidden Markov models. *Mach. Learn.*, 29(2-3):245–273, 1997.
4. J. Kennedy, I.I. Măndoiu, and B. Paşaniuc. Genotype error detection using hidden markov models of haplotype diversity. *Journal of Computational Biology*, 15(9):1155–1171, 2008.

5. J. Kennedy, B. Paşaniuc, and I.I. Măndoiu. 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/` .

6. 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.

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

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

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

10. 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.

11. P. Rastas, M. Koivisto, H. Mannila, and E. Ukkonen. Phasing genotypes using a hidden Markov model. In I.I. Măndoiu and A. Zelikovsky, editors, *Bioinformatics Algorithms: Techniques and Applications*, pages 355–372. Wiley, 2008.

12. 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.

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

14. 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.

15. 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.

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

17. 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.

18. 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.

19. 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.

20. 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.

21. 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.

22. `http://www.hapmap.org/`.