JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 16, Number 8, 2009
© Mary Ann Liebert, Inc.
Pp. 1141–1150
DOI: 10.1089/cmb.2009.0105

Research Article

Inferring Ancestries Efficiently in Admixed Populations with Linkage Disequilibrium

SIVAN BERCOVICI and DAN GEIGER

ABSTRACT

Much effort has recently been invested in developing methods for determining the ancestral origin of chromosomal segments in admixed individuals. Motivations for this task are the study of population history such as bottleneck effects and migration, the assessment of population stratification for adequate adjustment of association studies, and the enhancement of mapping by admixture linkage disequilibrium (MALD). In this article, we present a novel framework for the inference of ancestry at each chromosomal location. The uniqueness of our method stems from the ability to incorporate complex probability models that account for linkage-disequilibrium in the ancestral populations. We provide an inference algorithm that is polynomial in the number of markers even though the underlying problem seems to be inherently exponential in nature. We demonstrate the validity of our model and conclude that, with sufficient ancestral haplotypes, this framework can provide higher accuracy in inferring ancestral origin.

Key words: algorithms, computational molecular biology, genetic mapping, genetic variations, machine learning, Markov chains.

1. INTRODUCTION

Much effort has recently been invested in developing methods for determining the ancestral origin of chromosomal segments in admixed individuals (Patterson et al., 2004; Hoggart et al., 2004; Tang et al., 2006). Motivations for this task are the study of population history such as bottleneck effects and migration, the assessment of population stratification for adequate adjustment of association studies, and the enhancement of mapping by admixture linkage disequilibrium (MALD).

A growing number of complex disease studies are currently being conducted using population-based association (The Wellcome Trust Case Control Consortium, 2007). The premise of this method is that affected individuals carry a common variant of disease susceptible gene which is in linkage disequilibrium with sampled markers, hence can be detected using a sufficiently large case/control pool. Even though methods of association studies are considered state-of-the-art for gene mapping, the existence of sub-populations in the examined population introduces a stratification problem, also called the Simpson paradox effect (Simpson, 1951), namely, high rates of false associations resulting from the population's sub-structure rather than due to a genuine disease gene. Inferring the sub-structure of the examined

Computer Science Department, Technion, Haifa, Israel.

population more accurately will enable better adjustments for this effect, and improve the statistical power of association studies.

Mapping by Admixture Linkage Disequilibrium (MALD) is a powerful gene mapping approach (Reich and Patterson, 2005; Smith and O'Brien, 2005). The method is used for the identification of genomic regions harboring disease susceptibility genes in recently admixed populations. This method is useful when the prevalence of a disease is considerably different between the ancestral populations from which the admixed population was formed. When such a disease is studied, admixed individuals carrying the hereditary disease are expected to show elevated frequencies of the ancestral population with the higher prevalence around the disease gene loci. The power of this method relies on the ability to infer ancestry along the chromosomes of admixed individuals. Recent successes of this method include the discovery of multiple risk alleles for prostate cancer (Haiman et al., 2007) and the discovery of alleles correlated with the behavior of white blood cell count (WBC)—an important clinical marker of which high levels are associated with heart disease and cancer (Analls et al., 2008).

Propelled by these important applications, several previous studies developed methods for ancestry inference. The work of Patterson et al. (2004) employed a hidden Markov model (HMM) for the estimation of ancestry along the genome. In this work, the HMM was integrated into a Markov chain Monte Carlo (MCMC) method to account for uncertainties in the model parameters. A drawback in this model is the assumption that markers are mutually independent given their ancestry; an assumption that does not hold in high-density marker maps where linkage disequilibrium (LD) exists. Tang et al. (2006) approached this concern by modeling LD using a Markov Hidden Markov model (MHMM), namely, each marker in this model depends on the previous marker and on the two corresponding ancestries.

In reality, however, marker data such as SNP data is not Markovian. Figure 1 illustrates that linkage disequilibrium is a complex phenomenon, and in particular, it implies that marker data exhibits rather general dependencies across chromosomal segments.

In this article, we present a novel framework for the inference of ancestry at each chromosomal location. The uniqueness of our method stems from the ability to incorporate complex probability models that account for linkage-disequilibrium in the ancestral populations.

We provide an inference algorithm that is quadratic in the number of markers and linear in the input conditional probability tables even though the underlying problem seems to be exponential in nature. When LD is modeled as a first-order Markov chain, the algorithm becomes equivalent to that of Tang et al. (2006). We show that the error in the prediction of ancestral origin in the simulated data decreased when LD is modeled as a first order Markov model, and further decreases by up to 45% when LD is modeled as a second order Markov model. We conclude that with sufficient ancestral haplotypes, this framework can provide higher accuracy in inferring ancestral origin.

2. BACKGROUND

The genome of a recently admixed individual is a mosaic of large chromosomal segments, where each segment originated from a single ancestral population. This fact was studied and employed in a variety of works such as Pfaff et al. (2001), Hoggart et al. (2004), Rosenberg et al. (2003), Patterson et al. (2004), and Bercovici et al. (2008). We use the following definitions to describe these segments in admixed individuals. This description is based on Bercovici et al. (2008).

Definition 1. *An **admixed chromosome** is a chromosome that originated from more than one ancestral population.*

Definition 2. *A **Post Admixture Recombination** (PAR) point is a recombination point in which either two chromosomes from different populations crossed, or two chromosomes crossed where at least one of the chromosomes is an admixed chromosome.*

Definition 3. *A **(PAR) block** is a chromosomal segment limited by two consecutive PAR points, or by a chromosome edge and its closest PAR point.*

An immediate implication of these definitions is that every PAR block originated from a single ancestral population, designated as the ancestry of the block, for otherwise the block would have been further divided.

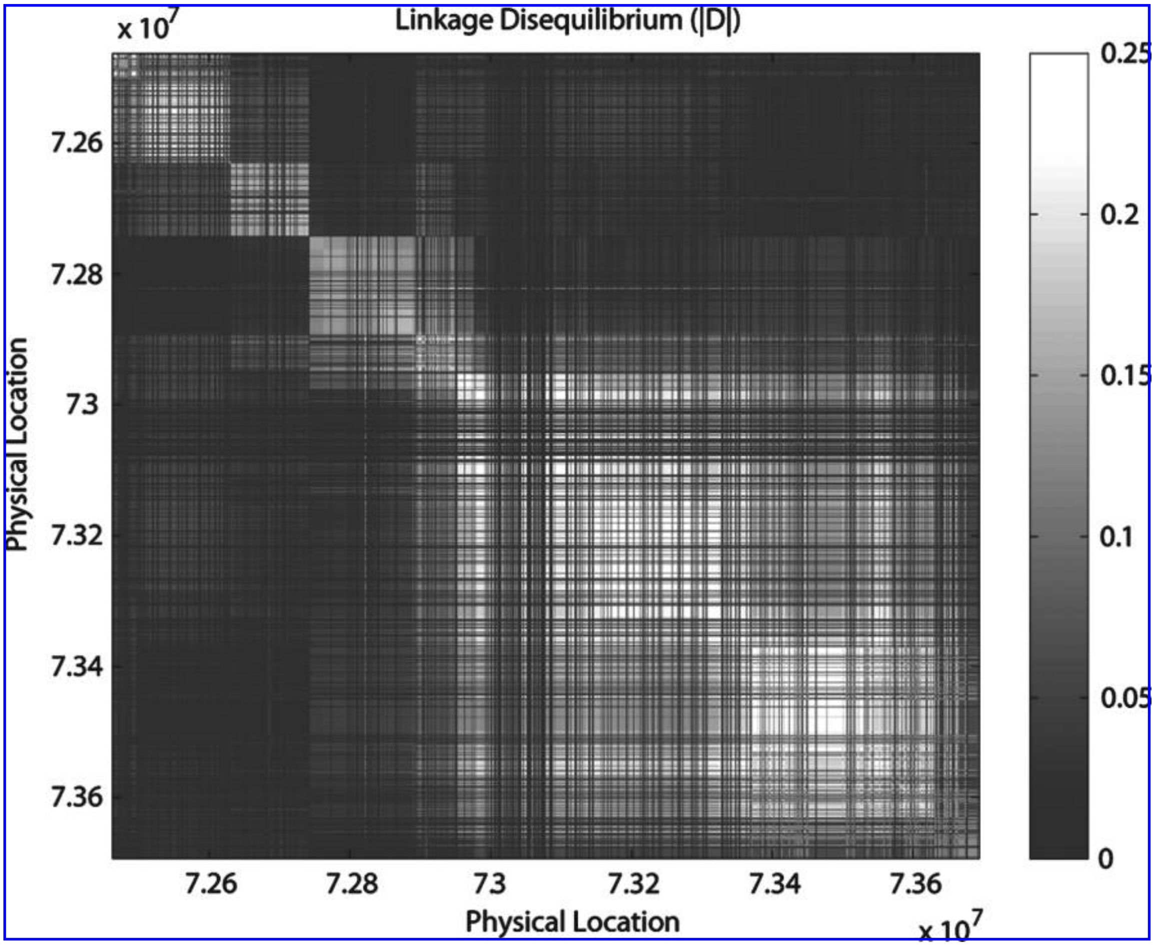

FIG. 1. Linkage disequilibrium (D) measured within 1-Mb segment on chromosome 1 in the European population. Data is taken from the HapMap project.

We denote the set of all observed markers by J, and the vector of an individual's PAR-blocks ancestries as Q. The set of an individual's PAR points defines a partition, denoted π, of the chromosomes into blocks (Fig. 2). We use the random variable Q_π to denote the vector of ancestries $(Q_1, Q_2, \dots)$ corresponding to the PAR-blocks determined by π, $Q_{\pi,i}$ to denote the ancestry (out of K possible ancestral populations) of the i^{th} PAR block in the given partition π, and the random vector $J_{\pi,i} = \{J_{\pi,i,1}, J_{\pi,i,2}, \dots, J_{\pi,i,m_i}\}$ to denote the

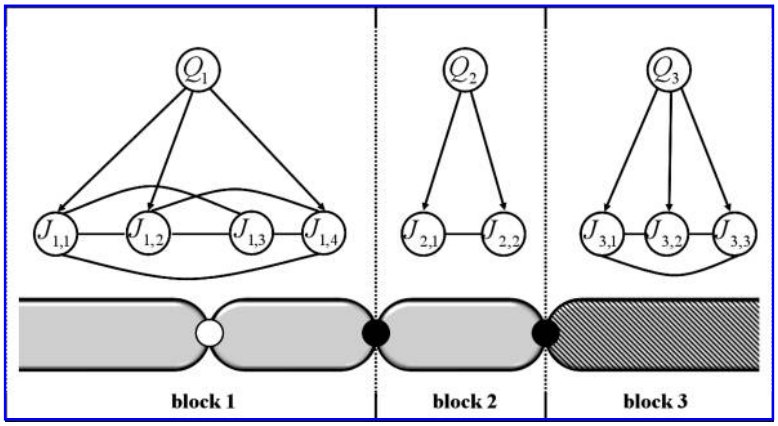

FIG. 2. Graphical model for $P(Q, J|\pi)$ assuming markers J_i within a PAR block are independent of markers in other blocks, and ancestries of PAR blocks are mutually independent.

haplotype assignment of the set of $m_{\pi,i}$ observed markers within this block. References to the subscript π will be omitted whenever π is clear from the context.

Markers within a PAR block are assigned according to the probability function of the corresponding ancestral population. We further assume that the ancestries of all PAR blocks of a given partition π are mutually independent. A graphical model showing these assumptions is given in Figure 2. In this figure, PAR points are represented by black dots and recombinations that are not PAR are represented by white dots.

The joint probability distribution described via the graphical model is given by

$$P(Q,J)=\sum_{\pi}P(\pi)\cdot\prod_{i=1}^{|Q_\pi|}P(Q_{\pi,i})\cdot P(J_{\pi,i}|Q_{\pi,i}). \tag{1}$$

In particular, when considering a specific point x on the genome, the joint probability for J and ancestry Q_x at that point is given by

$$P(Q_x,J)=\sum_{\pi}P(\pi)\cdot P(Q_x)\cdot P(J_{\pi,x}|Q_x)\cdot P_\pi(\bar{J}_{\pi,x}) \tag{2}$$

where $J_{\pi,x}$ are the markers within the same PAR-block as location x, and $P_\pi(\bar{J}_{\pi,x})$ is the probability of the set of all markers outside this block under the given partition π. We use this joint distribution to infer ancestry along the chromosome.

3. INFERENCE OF ANCESTRY

We develop an efficient computation for the conditional distribution $P(Q_x|J)$.

Based on the model presented in the previous section, the conditional distribution of the ancestry given the observed markers is:

$$P(Q_x|J)=\sum_{\pi}P(Q_x|\pi,J)\cdot P(\pi|J). \tag{3}$$

Using some algebra, Equation 3 can be rewritten as follows:

$$P(Q_x|J)=\sum_{\pi}\frac{P(J_{\pi,x}|Q_x,\pi)\cdot P(Q_x)}{P(J)}\cdot P(\pi)\cdot P(\bar{J}_{\pi,x}|\pi_x). \tag{4}$$

Computing the sum in Equations 3 and 4 is intractable in this form because the number of partitions grows exponentially in the number of markers, hence an alternative form is derived below.

Observe that for any two partitions π_1 and π_2 such that the PAR block that contains location x also contains the same set of markers indexed l through r ($J_{l,r}\subseteq J$), the term $P(J_{\pi,x}|Q_x,\pi)$ in Equation 4 is identical. Namely,

$$P(J_{\pi_1,x}|Q_x,\pi_1)=P(J_{\pi_2,x}|Q_x,\pi_2). \tag{5}$$

This is a key observation that enables efficient computation of Equation 3 as we now demonstrate.

Let $P(J_{\cdot,l})$ denote the probability of observing markers $J_1,\ldots,J_l$, and $P(J_{r,\cdot})$ denote the probability of observing markers J_r up to the last marker. Let $P(\pi_{l,r})$ denote the probability of a partition to contain a PAR-block which spans over markers J_l to J_r.

Since summing over all possible partitions is infeasible, we use Equation 5 to cluster such cases together, producing a computationally efficient form:

$$P(Q_x|J)=\frac{1}{P(J)}\sum_{l\in L}\sum_{r\in R}P(J_{l,r}|Q_x,\pi_{l,r})\cdot P(\pi_{l,r})\cdot P(Q_x)\cdot P(J_{\cdot,l})\cdot P(J_{r,\cdot}) \tag{6}$$

where L and R are indices of the markers to the left and right of x, inclusively.

The rest of this section develops efficient computations of the terms $P(J_{\cdot,l})$ and $P(J_{r,\cdot})$ in Equation 6, concluding with the specific case of $P(J)$. Consequently, Equation 3 becomes polynomial rather than exponential.

As described by Equation 1, PAR points divide the set of markers J into mutually independent sets, namely

$$P(J|\pi) = \prod_{i=1}^{|Q_\pi|} P(J_{\pi,i}). \tag{7}$$

Since PAR points themselves also occur independently, it is possible to employ a dynamic programming algorithm for the computation of $P(J)$ as follows.

Let $P(\overline{\mathbf{R}}_{i,k})$ denote the probability that no PAR point exists between markers J_i and J_k, and $P(\mathbf{R}_{i,k})$ denote the complementary event where at least one PAR point occurred in the corresponding segment. The following dynamic programming algorithm computes $P(J_{\cdot,i})$ for every marker in $J_i \in J$.

The term $P(J_{i,k})$ in Algorithm 1 is given by

$$P(J_{i,k}) = \sum_Q P(J_{i,k}|Q) \cdot P(Q). \tag{8}$$

Algorithm 1 Compute $P(J_{\cdot,l})$

```
for all $J_i \in J$, $J_i.p \leftarrow 0$
for $J_i \in J$, in chromosomal order do
  for $J_k \in J | k \geq i$ do
    $J_k.p \leftarrow J_k.p + J_{i-1}.p \cdot P(\mathbf{R}_{i-1,i}) \cdot P(J_{i,k}) \cdot P(\overline{\mathbf{R}}_{i,k})$
  end for
end for
```

When the algorithm terminates, the probability $P(J_{\cdot,i})$ for each marker J_i is stored in $J_i.p$. In particular, since by definition $P(J) = P(J_{\cdot,|J|})$, its value is stored in $J_{|J|}.p$. Note that both $J_0.p$ and $P(\mathbf{R}_{0,1})$ equal 1. The computation of $P(J_{r,\cdot})$ is achieved by a symmetrical algorithm, considering a reverse chromosomal order of computation.

As described in previous work, the common realization of $P(\mathbf{R}_{i,k})$ is via the Poisson distribution. In particular, $P(\mathbf{R}_{i,k}) = e^{-\lambda \cdot (l_k - l_i)}$ where l_i and l_k are the genomic locations of markers J_i and J_k, and λ is the rate of PAR points in the admixed individuals being examined (Bercovici et al., 2008). Similarly, $P(\pi_{l,r})$ from Equation 6 is written as the product of the three independent events that define it:

$$P(\pi_{l,r}) = P(\mathbf{R}_{l-1,l}) \cdot P(\overline{\mathbf{R}}_{l,r}) \cdot P(\mathbf{R}_{r,r+1}). \tag{9}$$

Once $P(J_{\cdot,i})$ and $P(J_{r,\cdot})$ are computed using our algorithms, Equation 6 can be efficiently computed for the examined ancestry Q_x. Using a Bayesian decision rule, the most probable assignment for Q_x is inferred to be the ancestry at location x.

A naive approach for inferring the ancestry along the entire chromosome would be to simply repeat the computation of Q_x on a grid of genomic locations. As the process of ancestry inference at different locations shares much of the computation, the inference of ancestry along the entire chromosome can (and has been) computed simultaneously. In particular, the term

$$P(J_{l,r}|Q_x, \pi_{l,r}) \cdot P(\pi_{l,r}) \cdot P(Q_x) \cdot P(J_{\cdot,l}) \cdot P(J_{r,\cdot}) \tag{10}$$

in Equation 6, being an identical additive term in $P(Q_x|J)$ of every Q_x where $l \leq x \leq r$, is computed once and added to the corresponding posterior probabilities. Hence, the computation of the posterior probability $P(Q_x|J)$ along a grid of points co-located with the markers is efficient, and can actually be achieved in the same complexity as the inference at a single location.

4. MOST PROBABLE ASSIGNMENT

The previous section developed ancestry inference based on the location-specific posterior probability $P(Q_x|J)$. Another appealing method is to base the ancestry inference on the most probable joint assignment of ancestries given the observations (MAP) via

$$\hat{Q} = \underset{Q,\pi}{\operatorname{argmax}}\ P(Q, \pi|J). \tag{11}$$

The properties of our admixture model, namely that the markers of PAR-blocks are mutually independent (Equation 7), enables the use of a dynamic programming algorithm for the computation of the most probable joint ancestry assignment. Algorithm 2 computes the most probable joint assignment of ancestry along a grid of locations. The algorithm maximizes the joint probability of the observations $J_{\cdot,i}$ for each marker $J_i \in J$ and ancestry. Rather than incrementally summing over the cases, as was done in Algorithm 1, we select the most probable assignment of ancestry for each PAR-block, resulting in a most probable joint assignment along the entire chromosome.

Algorithm 2 Infer Most Probable Ancestry

```
for all $J_i \in J$,
    $J_i.pred \leftarrow nil,\ J_i.p \leftarrow 0, J_i.Q \leftarrow nil$
for $J_i \in J$, in chromosomal order do
  for $J_k \in J | k \geq i$ do
    for $Q \in \{1..K\}$ do
      $prob \leftarrow J_{i-1}.p \cdot P(\mathbf{R}_{i-1,i}) \cdot P(\overline{\mathbf{R}}_{i,k})$
      if $J_k.p < prob \cdot P(J_{i,k}|Q) \cdot P(Q)$ then
        $J_k.p \leftarrow prob \cdot P(J_{i,k}|Q) \cdot P(Q)$
        $J_k.pred \leftarrow i$
        $J_k.Q \leftarrow Q$
      end if
    end for
  end for
end for
```

Once Algorithm 2 terminates, the most probable partition of the markers into PAR-blocks is defined by the predecessor list stored in the $J_i.pred$ variables. The most probable ancestry of each such PAR-block is stored in the $J_k.Q$ variable of its last marker.

At the functional level, Algorithm 1 and Algorithm 2 provide inference capabilities for our model equivalent to those provided by the *Forward-Backward* and *Viterbi* algorithms, respectively, for HMM models. The key difference is that our algorithms support arbitrary LD models.

When ancestry is inferred at points co-located with the sampled markers, the time complexity of both algorithms is quadratic in the number of markers $|J|$. Both algorithms iterate over all possible ranges (J_i, J_k) of markers, and for each of these $|J|^2$ ranges, a constant number of computations are performed. Hence, given the conditional probability tables, the time complexity is $O(|J|^2)$. The algorithms also fits the case where ancestry is inferred at specified locations, not necessarily collocated with the markers, in which case the complexity becomes quadratic in the size of the union of marker and specified locations.

5. EVALUATION

The framework developed in Section 3 and Section 4 decouples the admixture model from the ancestral population model. This separation allows the incorporation of general ancestry specific models of linkage disequilibrium at different resolutions.

The simplest model we explored assumes markers are independent, hence are assumed to be in linkage equilibrium. The assumption that the markers are independent given their ancestry underlies the model in the work of Patterson et al. (2004). We then considered a first-order Markov-chain as the marker linkage model, described by

$$P(J_{\pi,i}|Q_{\pi,i}) = P(J_{\pi,i,1}|Q_{\pi,i}) \cdot \prod_{j=2}^{m_i} P(J_{\pi,i,j}|J_{\pi,i,j-1}, Q_{\pi,i}) \tag{12}$$

Under this model, marker probabilities capture the LD present between every two adjacent markers in any particular ancestral population, similar to the model of Tang et al. (2006). We further evaluated a second-order Markov-chain model.

Table 1. Percentage of Error in Inferred Ancestry over 1000 Chromosomal Locations, Averaged over 50 Admixed Individuals

Method	*0*	*1*	*2*
Post	4.4655%	0.6%	0.24%
MAP	4.11%	0.29%	0.16%

In our experiments, we examined the accuracy of inferring ancestry in admixed African-American individuals, namely individuals who originated from the European and West-African ancestral populations. For our experiments, we used the haplotypes of 60 unrelated European parents and 60 unrelated West-African parents phased in the HapMap project (The International HapMap Project, 2005).

To synthesize our population, we used the Hybrid-Isolated (HI) admixture model (Long, 1991) which assumes a single admixture event at the first generation followed by random mating in all subsequent generations. For our experiment, we simulated 10 generations of admixture. The first-generation admixture uses a 0.2 European contribution, following the admixture dynamics of African-Americans as described in previous studies (Hoggart et al., 2004; Smith and O'Brien, 2005; Smith et al., 2004). We chose to focus our experiments on chromosome 1, in the area between 50 and 80 Mb. Within this segment, 1000 ancestry informative SNPs were greedily selected based on mutual information with ancestry (Rosenberg et al., 2003), satisfying the constraint of a 10-Kb minimal distance between markers. The density of the selected markers was chosen so as to be comparable to the Affymetrix 100K SNP chip density. The use of lower resolution panels nullifies the performance advantage of the complex models as the LD between observed markers decreases.

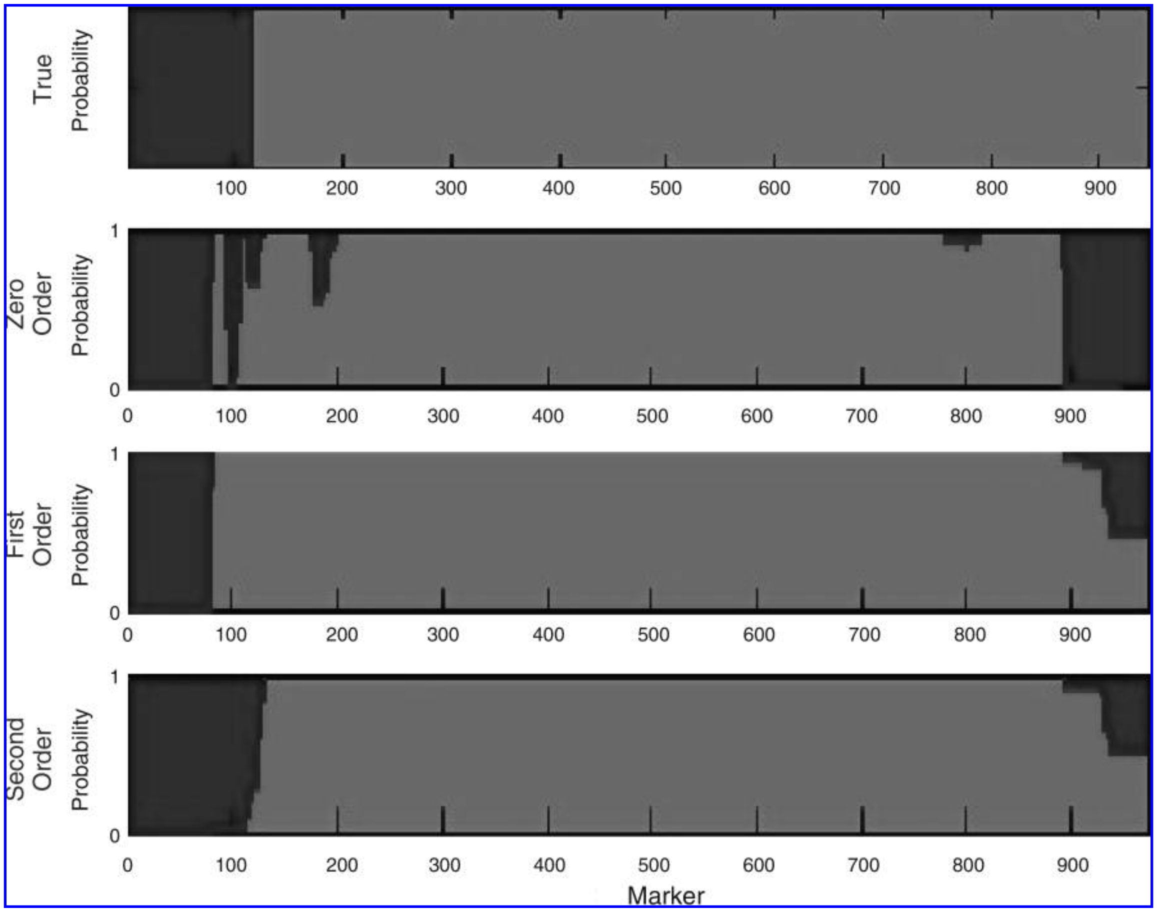

FIG. 3. Inferred ancestry based on the posterior probability of a single haplotype.

In our simulations we assumed the admixture coefficient $P(Q)$ and the number of generations that passed since admixture are given. Table 1 compares the accuracy in which ancestry is inferred across 50 simulated admixed individuals. We inferred ancestry using our two methods, namely posterior-based (denoted *Post*) and maximum a posteriori based (denoted *MAP*), under the three different LD models, namely the independent markers model, first-order Markovian and second-order Markovian (denoted 0,1 and 2, respectively). For example, this table shows that on average only 1.6 out of 1000 selected locations are not assigned their correct ancestry when a second order Markovian model is assumed and the *MAP* method is applied.

Figure 3 demonstrates inference of ancestry based on the posterior probability using the three LD models. The top row represents the true ancestry of a simulated admixed individual, followed by three rows representing the posterior probability based on the independent marker model, the first-order Markov-model, and the second-order Markov model. As can be observed in the figure, the first-order Markov model reduces the noise with respect to the independent model, and the second-order Markov model provides further accuracy as evident in the correct localization of the left-end ancestry switch. Moreover, we note that in this example, only the second-order Markov Model inferred the ancestry at the right-end correctly, as the posterior probability for the correct ancestry at the corresponding locations is above 0.5.

Figure 4 illustrates inference of ancestry based on maximum a posteriori estimates across 10 admixed individuals using the three different LD models. The top row represents the true ancestries across the 20 haplotypes, and the three subsequent rows correspond to our three LD models. Again, the results show that the inference accuracy improves as the LD model becomes more complex.

The ancestry in the reported experiments was inferred at 1000 points co-located with the markers. Under the examined marker density and in the case of the admixture characteristics of African-Americans, the ancestry of a region bounded by two consecutive points of ancestry evaluation has a very low probability ($<10^{-5}$) to disagree with the evaluation at the two endpoints. Such an event can only happen if at least two PAR points occur within the bounded region, and since this event occurs with low probability, the effect on accuracy is negligible.

6. DISCUSSION

We presented a novel probabilistic framework for the inference of ancestry at each chromosomal location. The main contribution is the ability to incorporate realistic probability models that account for linkage disequilibrium in the ancestral populations studied. The algorithms presented, one based on posterior probability of ancestry at each location and one on the MAP of ancestry on the entire chromosome, only grow quadratically as a function of the number of markers.

The power of our proposed algorithms becomes more prominent as more realistic models for the marker data of each ancestral population are used. We demonstrated that the use of second order Markov models reduces the error in inferring ancestry by approximately 45% compared to first order Markov models in the simulated data that we examined. Nonetheless, Figure 1 clearly shows that LD is a complex phenomenon, with effects spanning across chromosomal stretches. It is therefore reasonable to assume that better models of LD exist than the second order HMM that we have naively used in this paper. We hypothesize that complex models for LD may be learned once genomic data of high resolution will be made public beyond what is currently provided by HapMap. When such data becomes available, as expected shortly, our polynomial algorithms for inferring ancestry will be an efficient approach that is likely to dramatically improve the quality of ancestry inference and, as a result, the power of gene mapping techniques such as MALD.

The populations we simulated using the HapMap data, which follow the characteristics of the African-American admixed population, appear to represent a mild example of admixed individuals in the sense that even the simplest model that assumes no LD inferred ancestry rather well (only 4.5% error). For more complex admixtures, including cases in which the ancestral populations are not fully established, the error rate will increase, and the need to better model marker data will become more significant. In fact, this is precisely the situation in studies where unknown population stratification needs to be accounted for. In such cases our proposed algorithms are also appropriate and efficient.

Our algorithms can be extended to genotype data in two standard ways. One common method for the analysis of genotype data, when the resolution of the markers is sufficiently high, is to first apply a phasing step that produces haplotypes which can then be used as input to our proposed algorithms.

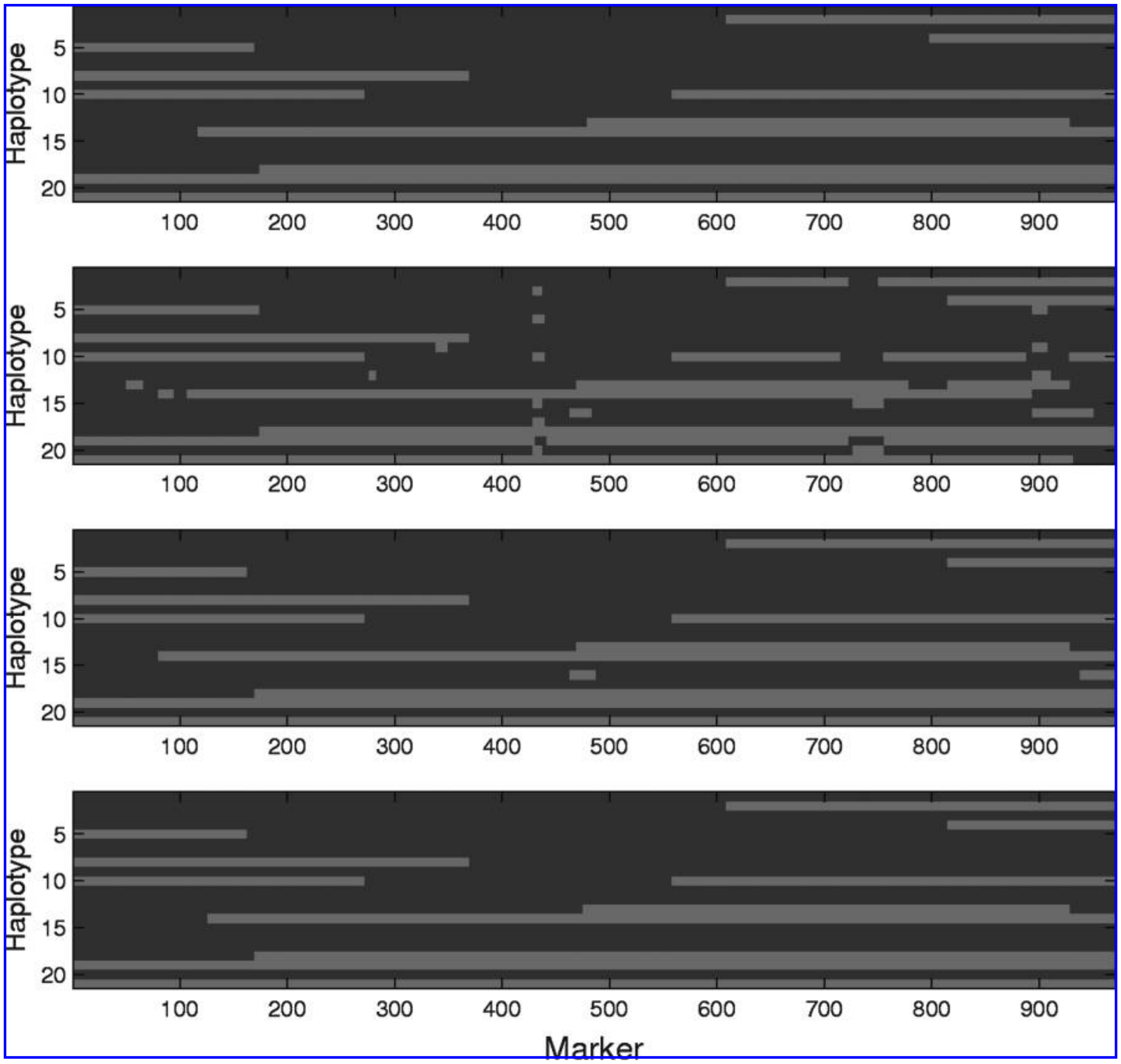

FIG. 4. Inferred ancestry based on maximal a posteriori estimation of several haplotypes.

There is also an immediate extension of our model to genotype data following the lines suggested in previous studies (Tang et al., 2006; Patterson et al., 2004). These alternatives need to be tested empirically.

Finally, we have demonstrated in Section 3 that the likelihood of the data given a model, namely the probability $P(J)$ of marker data can be efficiently computed for various parameters of the model. Consequently a maximum likelihood approach can be easily incorporated to learn any parameter of interest such as the admixture rate λ.

These and similar extensions are being evaluated and will be presented on our website along with the software developed (http://bioinfo.cs.technion.ac.il/MALD/).

ACKNOWLEDGMENTS

We thank Liran Shlush, Karl Skorecky, and Alan Templeton for fruitful discussions, and Tamar Aizikowitz for her helpful comments. We also thank the Wolfson Foundation for contributing a grid of 120 PCs that enabled us to conduct this research, and the Microsoft TCI grant.

DISCLOSURE STATEMENT

No competing financial interests exist.

REFERENCES

Analls, M., Wilson, J.G., Patterson, N.J., et al. 2008. Admixture mapping of white cell count: genetic locus responsible for lower white blood cell count in the Health ABC and Jackson heart studies. *Am. J. Hum. Genet.* 82, 81–87.

Bercovici, S., Geiger, D., Shlush, L., et al. 2008. Panel construction for mapping in admixed populations via expected mutual information. *Genome Res.* 18, 661–667.

Haiman, C.A., Patterson, N., Freedman, M.L., et al. 2007. Multiple regions within 8q24 independently affect risk for prostate cancer. *Nat. Genet.* 35, 638–644.

Hoggart, C.J., Shriver, M.D., Kittles, R.A., et al. 2004. Design and analysis of admixture mapping studies. *Am. J. Hum. Genet.* 965–978.

Long, J. 1991. The genetic structure of admixed population. *Genetics* 417–428.

Patterson, N., Hattangadi, N., Lane, B., et al. 2004. Methods for high-density admixture mapping of disease genes. *Am. J. Hum. Genet.* 979–1000.

Pfaff, C.L., Parra, E.J., Bonilla, C., et al. 2001. Population structure in admixed populations: effect of admixture dynamics on the pattern of linkage disequilibrium. *Am. J. Hum. Genet.* 68, 198–207.

Reich, D., and Patterson, N. 2005. Will admixture mapping work to find disease genes? *Philos. Trans. R. Soc. B* 360, 1605–1607.

Rosenberg, N.A., Li, L.M., Ward, R., et al. 2003. Informativeness of genetic markers for inference of ancestry. *Am. J. Hum. Genet.* 73, 1402–1422.

Simpson, E.H. 1951. The interpretation of interaction in contingency tables. *J. R. Statist. Soc.* 13, 238–241.

Smith, M.W., and O'Brien, S.J. 2005. Mapping by admixture linkage disequilibrium: advances, limitations and guidelines. *Nat. Rev. Genet.* 6, 623–632.

Smith, M.W., Patterson, N., Lautenberger, J.A., et al. 2004. A high-density admixture map for disease gene discovery in african americans. *Am. J. Hum. Genet.* 74, 1001–1013.

Tang, H., Coram, M., Wang, P., et al. 2006. Reconstructing genetic ancestry blocks in admixed individuals. *Am. J. Hum. Genet.* 79, 1–12.

The International HapMap Project. 2005. A haplotype map of the human genome. *Nature* 1299–1319.

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

Address correspondence to:
Sivan Bercovici
Computer Science Department
Technion
Haifa 32000, Israel

E-mail: sberco@cs.technion.ac.il

This article has been cited by:

1. Sivan Bercovici , Dan Geiger . 2011. Admixture Aberration Analysis: Application to Mapping in Admixed Population Using Pooled DNA. *Journal of Computational Biology* **18**:3, 237-249. [Abstract] [Full Text] [PDF] [PDF Plus]
2. S. Bercovici, C. Meek, Y. Wexler, D. Geiger. 2010. Estimating genome-wide IBD sharing from SNP data via an efficient hidden Markov model of LD with application to gene mapping. *Bioinformatics* **26**:12, i175-i182. [CrossRef]