JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 22, Number 5, 2015
© Mary Ann Liebert, Inc.
Pp. 451–462
DOI: 10.1089/cmb.2014.0151

A Spatial Haplotype Copying Model with Applications to Genotype Imputation

WEN-YUN YANG[1] FARHAD HORMOZDIARI[1]
ELEAZAR ESKIN[1,2] and BOGDAN PASANIUC[2,3]

ABSTRACT

Ever since its introduction, the haplotype copy model has proven to be one of the most successful approaches for modeling genetic variation in human populations, with applications ranging from ancestry inference to genotype phasing and imputation. Motivated by coalescent theory, this approach assumes that any chromosome (haplotype) can be modeled as a mosaic of segments copied from a set of chromosomes sampled from the same population. At the core of the model is the assumption that any chromosome from the sample is equally likely to contribute *a priori* to the copying process. Motivated by recent works that model genetic variation in a geographic continuum, we propose a new spatial-aware haplotype copy model that jointly models geography and the haplotype copying process. We extend hidden Markov models of haplotype diversity such that at any given location, haplotypes that are closest in the genetic-geographic continuum map are *a priori* more likely to contribute to the copying process than distant ones. Through simulations starting from the 1000 Genomes data, we show that our model achieves superior accuracy in genotype imputation over the standard spatial-unaware haplotype copy model. In addition, we show the utility of our model in selecting a small personalized reference panel for imputation that leads to both improved accuracy as well as to a lower computational runtime than the standard approach. Finally, we show our proposed model can be used to localize individuals on the genetic-geographical map on the basis of their genotype data.

Key words: 1000 Genomes, expectation maximization (EM) algorithm, genotype imputation, linkage disequilibrium, single nucleotide, polymorphism, spatial genetics, stochastic gradient descent.

1. INTRODUCTION

Complex population demography coupled with the presence of recombination hotspots have shaped genetic variation in the human genome into blocks of markers with similar recent ancestry (Gibbs et al., 2003; Consortium et al., 2010; Daly et al., 2001). This recent ancestry sharing induces dependencies among variants in the form of linkage disequilibrium (LD), that is, the nonrandom association of alleles at two or

[1]Department of Computer Science; [2]Department of Human Genetics; and [3]Department of Pathology and Laboratory Medicine, Geffen School of Medicine; University of California, Los Angeles, California.

more loci (Kruglyak, 1999). Therefore, the observed LD patterns across the genome are the result of a population's demographic history and are modeled in a wide range of problems, from population genetic inferences (Lohmueller et al., 2009; Pool et al., 2010) to medical population genetics (Marchini et al., 2007; Li et al., 2010). Most notably, LD has enabled the era of genome-wide association studies that use a small number of variants (as compared to all variation in the genome) to assay variation across the entire human genome (de Bakker et al., 2005). Thus, modeling population LD is a fundamental problem in computational genetics with applications ranging from genotype imputation and haplotype inference to locus-specific and genome-wide ancestry inference (Marchini et al., 2007; Howie et al., 2009, 2012a; Chung et al., 2013; Savage et al., 2013; Pasaniuc et al., 2009; Price et al., 2009).

Although many approaches for modeling LD have been proposed (Daly et al., 2001; Li and Stephens, 2003), one of the most successful framework has been introduced by Li and Stephens [widely referred to as the *haplotype copy model* (Li and Stephens, 2003)]. Drawing on coalescent theory, in this model, a haplotype sampled from a population is viewed as a mosaic of segments of previously sampled haplotypes. This mosaic structure can be efficiently modeled within a hidden Markov model to achieve very accurate solutions to many genetic problems such as genotype imputation (Marchini et al., 2007; Howie et al., 2009, 2012a), ancestry inference (Pasaniuc et al., 2009; Price et al., 2009), quality control in genome-wide association studies (Han et al., 2009), detection of identity by descent (IBD) segments (Browning, 2006; Browning and Browning, 2010), estimating recombination rates (Wegmann et al., 2011), haplotype phasing (Delaneau et al., 2012), migration rates (Roychoudhury and Stephens, 2007) and calling of genotypes at low coverage sequencing (Pasaniuc et al., 2012; Li et al., 2011).

At the core of the Li and Stephens (2003) model lies a hidden Markov model (HMM) that emits haplotypes through a series of segmental copies from the pool of previously observed haplotypes. The hidden states in the HMM indicate which haplotype from the reference panel to copy from while emission probabilities allow for potential mutation events observed since the most recent common ancestor of the target and the reference copy haplotype. Recombination events are modeled through the transition probabilities; the probability of copying from the same reference haplotype at successive loci is much higher than switching to another haplotype, based on the idea of the probability having a recombination between two neighboring loci is low. Motivated by coalescent theory in randomly mating populations, the *a priori* probability of switching the copy process to another haplotype is equally likely among all the previously observed haplotypes. However, since human populations show a tremendous amount of structure across geography (Novembre et al., 2008; Yang et al., 2012; Baran et al., 2013) (inline with isolation-by-distance models), it is likely that haplotypes physically closer in geography to the target haplotype contribute significantly more to the copy process. Furthermore, with the emergence of high-throughput sequencing that is generating massive amounts of data (Mardis, 2008; Schuster, 2008; Shendure et al., 2004), existing methods are increasingly computationally intensive due to the ever larger samples of haplotypes that can be used as reference. Although a commonly used approach for reducing computational burden is to downsample the reference panels (Howie et al., 2011; Pasaniuc et al., 2010; Liu et al., 2013) (often in an ad-hoc manner), a principled approach for selection of a reference panel for optimizing performance is currently lacking.

In this article, we propose a new approach to modeling genetic variation in structured populations that incorporates ideas from both the haplotype copying model (Li and Stephens, 2003) and the spatial structure framework that models genetic variation as function of geography (Yang et al., 2012; Baran et al., 2013). Thus, we propose a haplotype copy model that a priorly up weights the contribution of haplotypes closer in geographical distance to the copying process. We accomplish this by jointly modeling geography and the copying process. Each haplotype is associated with a geographical position; when copying into a new haplotype with known location, we instantiate an HMM that has switching transition probabilities upweighted for haplotypes closer in geographical space to the target haplotype.

We use real data from the 1000 Genomes project (Consortium et al., 2010) to show that our spatial-aware approach fits the data significantly better than the standard model. Through a masking procedure followed by a leave-one-out experiment we show that our spatial-aware method significantly increases imputation accuracy especially for lower frequency variation (e.g., an improvement of 6% [2%] for low-frequency [common] variation in Asian data). We also show that our approach can be used to select a small personalized reference panel for imputation that increases imputation accuracy while significantly reducing imputation runtime (up to 10-fold). Finally, we show how our model can be used in a supervised manner to infer locations on the genetic-geographic map for individuals based on their genetic data.

2. METHODS

2.1. The standard haplotype copying model

We start by briefly introducing the standard haplotype copying model (Li and Stephens, 2003) for modeling LD in a population. Let $H \in \{0, 1\}_{N \times L}$ be a matrix of haplotypes (which we will refer to as the *reference panel*), where $h_{ij} \in \{0, 1\}$ indicates if the i-th individual at the j-th position (SNP) contains the reference or the alternate allele. N denotes the number of haplotypes in the reference panel and L the number of SNPs in the data. Let $h \in \{0, 1\}_{1 \times L}$ be a multilocus haplotype that we will refer to as the *target haplotype*, where $h_i \in \{0, 1\}$ indicates the i-th SNP. The haplotype copy model views the target haplotype as being composed of a mosaic of segments from haplotypes of the reference panel.

Formally, we define a hidden Markov model (HMM) specified by a triple (S, τ, ω), where S is the set of states, τ is the transition probability, and ω is the emission probability function. The set S contains state variables $\{s_1, \ldots, s_L\}$ where $s_k = \{1, 2, \cdots N\}$ indicates from what reference haplotype is the k-th allele in the target haplotype copied from. The transition probability τ is nonzero only between pairs of states in consecutive sets of states S, which can be defined between SNP k and SNP $k + 1$ as follows

$$\tau_k(i, j) = \begin{cases} \theta_k + (1 - \theta_k)/N & \mathrm{i = j} \\ (1 - \theta_k)/N & \mathrm{i \neq j} \end{cases}, \quad \text{where} \quad \theta_k = \exp(-\rho \mathrm{d}_k).$$

Here d_k is the physical distance between SNP k and SNP $k + 1$ and $\rho = 4N_e c$ where N_e is the effective population size, c is the average rate of crossover per unit physical distance per meiosis (e.g., 10^{-8}). This can be easily extended to use recombination maps with varying recombination events at different loci in the genome. The emission probability mimics the mutation process and can be defined as follows

$$\omega(h_k, s_k; H) = \begin{cases} 1 - \epsilon & h_k = H_{s_k, k} \\ \epsilon & \text{otherwise} \end{cases}, \quad \text{where} \quad \epsilon = \frac{N}{N + \left(\sum_{m=1}^{N-1} 1/m\right)^{-1}}.$$

where N denotes the number of reference haplotypes. Intuitively the copying process is more accurate as the reference sample size grows, and it is more likely to find in the reference a haplotype closely matching the target one.

The likelihood of the target haplotype h is defined as:

$$P(h|S, H) = P(S) \prod_k P(h_k|s_k, H) = \prod_k \tau_k(s_{k-1}, s_k) \left(\prod_k \omega(h_k, s_k; H) \right) \tag{1}$$

and can be efficiently estimated using the forward/backward algorithm. Inference in this model is performed using standard HMM approaches such as Viterbi or posterior decoding. For example, if the target haplotype has any of the alleles missing, posterior decoding can be employed to estimate the most likely values conditional on the model and the rest of the target haplotype.

2.2. A spatial-aware haplotype copying model

A drawback of the standard haplotype copying model comes from the equal treatment of reference haplotypes; that is, *a priori* all haplotypes from the reference panel are equally likely to contribute to the target haplotype. This effect motivates us to propose the following approach to take spatial effect into account in the model. Let $X = \{x_1, \ldots, x_N\}$ indicate the locations for all N reference haplotypes and x indicate the location for target haplotype. In a scenario where the location of the individuals are not known, we estimate their locations from genotype data using methods such as PCA (Novembre et al., 2008), SPA (Yang et al., 2012) or LOCO-LD (Baran et al., 2013). Then, instead of using uniform switching probability across all reference haplotypes, we assign higher probability to haplotypes located closer to the target haplotype. Formally, we redefine the transition rate τ between SNP k and SNP $k + 1$ as:

$$\tau_k(i,j)=\begin{cases}\theta_k+(1-\theta_k)p_j & i=j\\ \\ (1-\theta_k)p_j & i\neq j\end{cases}\qquad \text{where}\quad p_j=\frac{\exp\left(-\lambda\psi(x,x_j)\right)}{Z}. \tag{2}$$

The function $\psi(x, x_j)$ denotes a distance function between x and x_j (e.g., Euclidean distance), and Z is a normalization factor to ensure the probability definition. The parameter λ specifies the effect of geographical distance. It is worth mentioning that this spatial-aware model can be reduced to standard haplotype copying model by setting $\lambda = 0$, such that $p_j = 1/N$; therefore, our approach can be viewed as a generalization of the standard Li and Stephens model. An illustration of our model is shown in Figure 1. Intuitively a large value for λ indicates a more pronounced spatial effect (less probability to copy from distant haplotypes), while $\lambda = 0$ reverts to assigning equal *a priori* probability.

The likelihood of the target haplotype is defined as before by summing on all paths in the model (Eq. 1). Inference in this model can be performed as in the standard haplotype copy model using a combination of Viterbi and posterior decoding as function of the particular application.

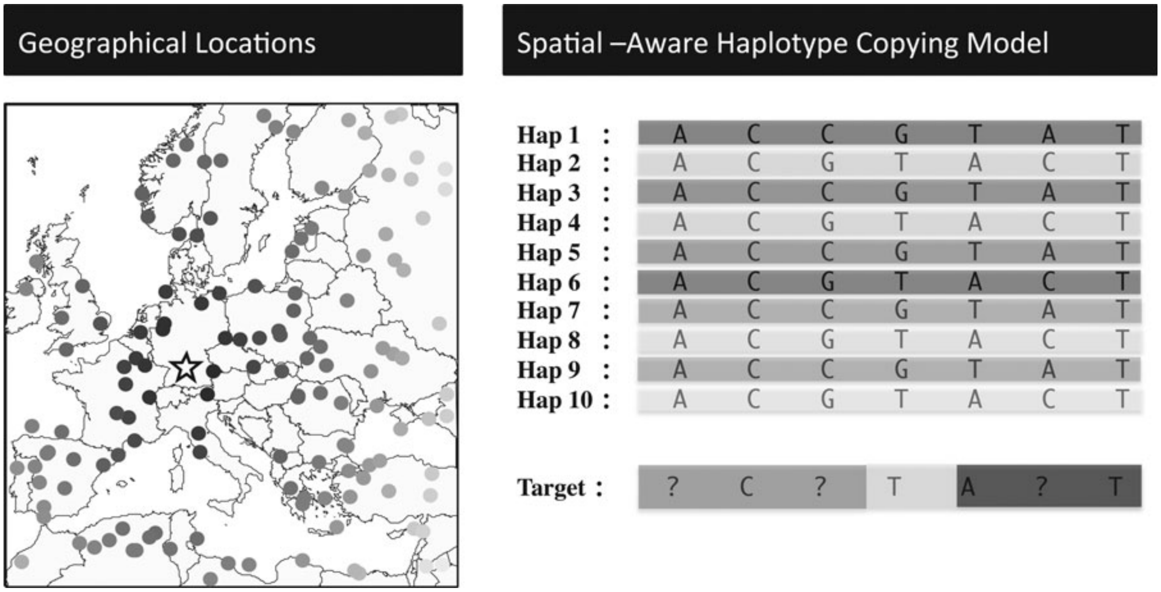
The standard forward-backward algorithm has a computational complexity $O(N^2L)$, since it has to enumerate all possible N^2 transitions between two nearby variables. However, we can take advantage of the transition probability structure to speed up the computation from $O(N^2L)$ to $O(NL)$. Note that the transition probability in Equation (2) consists of a term $(1 - \theta_k)p_j$, which is independent on the previous state variable s_i. Thus, we can precompute the following constant

$$\eta_k=\sum_i v_{ki}(1-\theta_k)p_j$$

where the variable v_{ki} denotes the inductive variables in forward-backward algorithm. Then we can simplify the induction rule for $v_{k+1,i}$ significantly as follows

$$v_{k+1,i}=v_{ki}\theta_k+\eta_k$$

This is only constant computation for $v_{k+1,i}$ given the precomputed constant ω_k. Therefore, the computation of the forward-backward algorithm induction can be easily reduced to $O(NL)$.



Hap 1 :	A	C	C	G	T	A	T
Hap 2 :	A	C	G	T	A	C	T
Hap 3 :	A	C	C	G	T	A	T
Hap 4 :	A	C	G	T	A	C	T
Hap 5 :	A	C	C	G	T	A	T
Hap 6 :	A	C	G	T	A	C	T
Hap 7 :	A	C	C	G	T	A	T
Hap 8 :	A	C	G	T	A	C	T
Hap 9 :	A	C	C	G	T	A	T
Hap 10 :	A	C	G	T	A	C	T
Target :	?	C	?	T	A	?	T

FIG. 1. An illustration of spatial haplotype copying model. In the left panel, the location for target haplotype is shown using the star. All haplotypes in the data are color coded using the distance to the target location (light are more distant, dark are closer). We enforce the transition rates (that encode the copy switching) to give higher weight to haplotypes closer to the target haplotype. A haplotype at the target location is more likely to contain mosaic segments from haplotypes that are closer to the target location.

2.3. *Estimation of spatial effect parameter λ*

All the other parameters θ and ϵ can be determined from the reference panel and SNP locations. But a prerequisite step in applying our model is the specification of λ. It is necessary to estimate the λ before using the model for various applications, as the value of λ could vary significantly across individuals or populations. We estimate λ through maximum likelihood estimation (MLE). Starting from the likelihood of the target haplotype h (Eq. 1), we marginalize over all possible values of hidden variables S to obtain likelihood as function of λ:

$$L(h;\lambda)=\sum_{S}P(h|S,H;\lambda) \tag{3}$$

However, this overall likelihood function is infeasible to optimize directly, as the number of all possible values of S is very large L^N. Although the likelihood computation can be reduced by forward-backward algorithm to $O(NL)$, the gradient is still very expensive to compute, as the calculation would involve a forward-backward in $O(NL)$ and a summation of $O(N^2L)$ terms. When the number of reference haplotypes is large, this gradient would be infeasible to compute. Fortunately, the gradient for the Q function in EM algorithm is much simpler to compute than the gradient of likelihood function in (3). It is also guaranteed that the gradient of the Q function will be an increasing direction for the original likelihood function, which is a theoretical property of the EM algorithm. Thus, we resort to compute the gradient of the Q function instead of the gradient of original likelihood function.

First, the Q function in EM algorithm can be written as follows

$$\begin{aligned} Q(\lambda,\lambda^{(t)}) &= \sum_{S}P(S)\ln P(h,S;\lambda) \\ &\propto \sum_{kij}P(s_{k-1}=i,s_k=j;\lambda^{(t)})\ln\tau_k(i,j;\lambda) \end{aligned} \tag{4}$$

The gradient for this Q function can be calculated as follows

$$\frac{\partial Q}{\partial\lambda}=-\sum_{kij}P(s_{k-1}=i,s_k=j;\lambda^{(t)})\left(\frac{\psi(x,x_j)-\sum_l\psi(x,x_l)p_l}{1+I(i=j)\left(\frac{\theta_k}{(1-\theta_k)p_j}\right)}\right) \tag{5}$$

where the identity function $I(i=j)$ is equal to 1 when $i=j$ and 0 otherwise. However, simple calculation of this gradient will also be inefficient with the complexity $O(N^2L)$, which is still expensive for thousands of reference haplotypes and millions of SNPs. We resort to computing a stochastic gradient for the Q function, and apply it to the original likelihood function as a searching direction. We estimate the gradient by sampling over N haplotypes, instead of enumerating all of them. In practice, between each pair of SNP k and SNP $k+1$, we randomly sample 1000 pairs of $s_{k\text{-}1}=i$ and $s_k=j$, instead of all N^2 pairs. The overall algorithm for efficient optimization of the spatial effect parameter λ is described in Algorithm 1.

Algorithm 1: Learning Algorithm for Parameter λ Estimation

```
Setting optimization parameters $R$ and $C$ (e.g., $R = 1 \times 10^3$ and $C = 20$)
Precomputing $\psi(x, x_j)$ for all reference haplotype $j$, and $\theta_k$ for all $k$.
Randomly initialize $\lambda^{(0)} > 0$
for $t$ from 0 to $T$ do
    Perform forward-backward algorithm to get the forward/backward probability
    Compute stochastic gradient $g(\lambda^{(t)})$ by sampling $R$ pairs of $i$ and $j$ in (5)
    Setting $\lambda^{(t+1)} = \lambda^{(t)} + \frac{1}{t+C} \cdot g(\lambda^{(t)})$
end for
Output $\lambda^{(T+1)}$
```

2.4. *Localization of individuals based on their genetic data*

Another appealing application for spatial-aware haplotype copying model is to localize individuals on the map. That is, given locations X for all reference panel haplotypes, we seek to find the best location x for the target haplotype to maximize the likelihood of the data. The algorithm follows similar procedure as above section 2.3. The difference mainly comes from a different Q function as follows

$$\begin{aligned} Q(x, x^{(t)}) &= \sum_{S} P(S) \ln P(h, S; x) \\ &\propto \sum_{kij} P(s_{k-1} = i, s_k = j; x^{(t)}) \ln \tau_k(i, j; x) \end{aligned} \tag{6}$$

which is parameterized by x instead of λ as in Equation (4). However, this change leads to nonconcavity of the function in general. But since there is only one parameter to estimate, and the function is well behaved in practice, we can still compute the gradient for the Q function and apply it to the stochastic gradient descent method. The gradient for the Q function in Equation (6) can be calculated as follows

$$\frac{\partial Q}{\partial x} = -\sum_{kij} P(s_{k-1} = i, s_k = j; x^{(t)}) \lambda \left(\frac{\frac{\partial \psi(x, X_j)}{\partial x} - \sum_l p_l \cdot \frac{\partial \psi(x, X_j)}{\partial x}}{1 + I(i = j)\left(\frac{\theta_k}{(1-\theta_k)p_j}\right)} \right) \tag{7}$$

we can use Euclidean distance $\psi(x, X_j) = \|x - X_j\|_2$ as a sufficient estimation of spatial distance. Thus, the gradient of the distance metric becomes

$$\frac{\partial \psi(x, X_j)}{\partial x} = \frac{x - X_j}{||x - X_j||_2}$$

The overall algorithm is similar to Algorithm 1 for optimizing λ, except for replacement of λ by x and the gradients correspondingly.

3. EXPERIMENTAL RESULTS

3.1. Estimation of spatial copying effect in the 1000 Genomes data

We applied our methods to data-generated part of the 1000 Genomes project (Consortium et al., 2010). A total of 1092 individuals were collected from 14 populations across the European, Asian, African, and American continents. For all of our simulations we used 157, 827 SNPs on chromosome 22, where 79.5% of SNPs are rare SNPs (allele frequency < 0.05), and the rest 20.5% are common SNPs; although the original data contained 473,481 SNPs, for computational efficiency we down-sampled to every third SNP. Among the considered SNPs, we assumed that only 2, 931 SNPs present on the Affymetrix 6.0 SNP array are collected and the remaining SNPs will be imputed using our model. This amounts to using 1.86% SNPs to impute the remaining 98.14% SNPs. We apply PCA (Novembre et al., 2008) to assign a geographical location to each individual in the dataset. Although we note that the imputation performance can be further improved if denser SNPs are assumed to be typed, we expect the general trends reported below to maintain.

The parameter θ for transition probability τ and the parameter ϵ for emission probability ω can be determined from the SNP locations, population size, and number of reference haplotypes, as given in section 2.1. Starting from the 2, 931 SNPs, we estimated the spatial effect parameter λ for each of the 2,184 haplotypes in the dataset. The average λ values are 1.54, 1.76, 1.30, and 1.32 for European, Asian, African, and American populations, respectively (Fig. 2). Generally, the higher value of λ corresponds to stronger spatial copying effect, which leads to more segments copied from nearby haplotypes. To test the significance of spatial effect, we compared the likelihoods of the data (the 2,184 haplotypes) within the model assuming no spatial effect ($\lambda = 0$) versus the model with spatial effect (λ* estimated from the data). The log likelihood ratio between spatial haplotype copying model and standard haplotype model is given in Figure 2. The likelihood is computed for each haplotype being emitted from the rest of the haplotypes. Across all populations we observe that the model with a spatial effect fits the data much better than the model with no spatial assignment. This is expected since we use haplotypes across all continents (except the target) in the reference panel, and it is expected that haplotypes share more continental-specific segments.

3.2. Spatial-aware model improves imputation accuracy

Having established that the model with spatial effect fits the data much better than the standard model with no spatial effect, we focused next on haplotype imputation [a standard approach in genome-wide

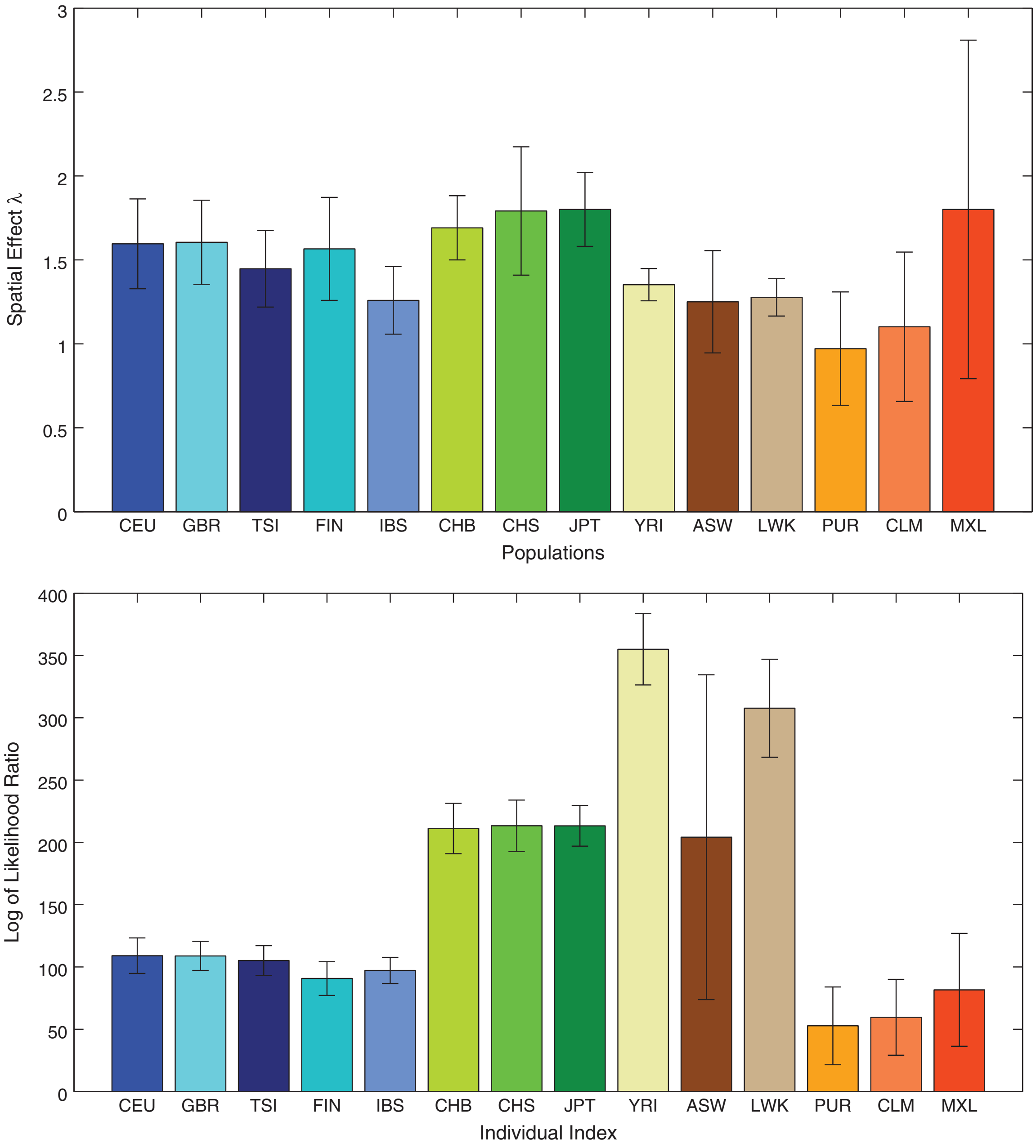


FIG. 2. Estimated spatial copying effects λ^* across different populations in 1000 Genomes data. The top panel shows the average λ^* across all individuals in a given population while the bottom displays the log likelihood ratio of the model with λ^* as compared to $\lambda = 0$. The error bars indicate the standard deviations for each population.

association studies through prephasing (Howie et al., 2012b)]. We carry out a leave-one-out procedure to perform the evaluation. In each round, we select one haplotype as a target and use the rest as the reference panel. To remove potential bias, instead of using all haplotypes, we randomly select one haplotype from each individual to use a total of 1,092 haplotypes (i.e., each round imputes one haplotype from the remaining 1,091). The imputation results are evaluated using the average per-SNP r^2 correlation coefficients averaged across all leave-one-out rounds for either all haplotypes or for data within each population.

We first demonstrate the effect of the lambda parameter on imputation accuracy by applying our model using a wide range of lambda parameter values. Compared with the baseline method ($\lambda = 0$), we observe that a clear improvement is obtained for a value of λ around 2, especially for European and Asian populations (see Fig. 3). This is consistent with the spatial model fitting those populations (see Fig. 2). We

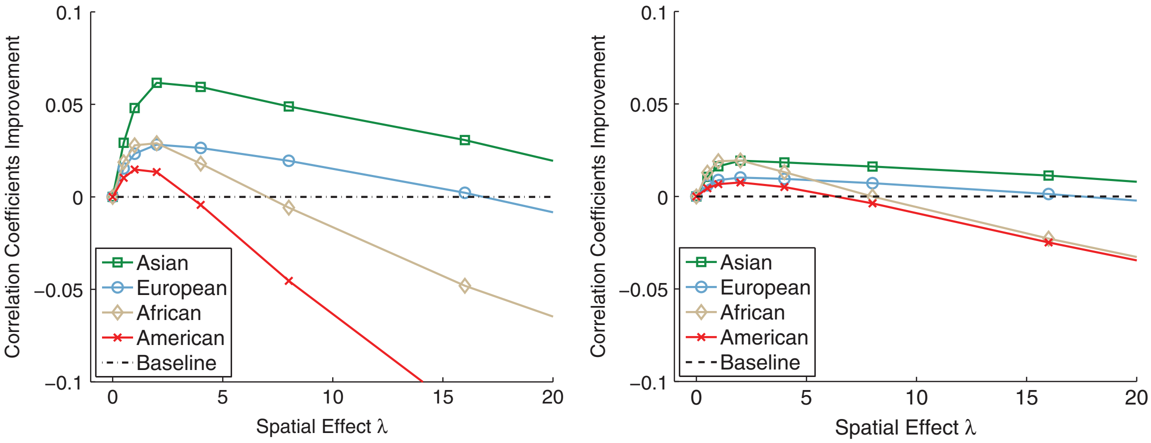


FIG. 3. Effect of spatial copying parameter λ on imputation accuracy. Left shows results for low frequency (1–5%) while right displays results for common variants ($>5\%$). The maximum accuracy is attained at a $\lambda \approx 2$, close to the maximum likelihood estimate for λ (1.3 to 1.7, see section 3.1).

also observe that the spatial model improves the imputation of rare variants more significantly than common variants, which is expected as the rare variants are more clustered geographically (Nelson et al., 2012). Moreover, the improvement for Asian and European populations is larger than for African and American populations.

Although we have shown that spatial model improves accuracy, in practice the value of λ is unknown and needs to be estimated from the data. We reassessed the accuracy of our approach by not setting λ to prespecified values but by estimating it from the data. The performance of the model using the maximum likelihood λ^* over baseline method is given in Table 1. As before, we observe larger improvements for rare variants than common variants. A plausible explanation for this effect is that rare variants are more clustered in geography (Nelson et al., 2012) than common variants. Overall for all populations, the improvement is highly correlated with allele frequency. The trend is shown in Figure 4, where we can see that the improvement is higher for SNPs with lower allele frequency.

3.3. Selection of a personalized reference panel for imputation to increase performance

Inspired by the significant spatial haplotype copying effect in experiments, we hypothesized that imputation efficiency can be improved by only using a personalized reference panel composed only from geographically close haplotypes (Pasaniuc et al., 2010; Howie et al., 2011). First, we expect that most of the reference haplotypes are not contributing haplotype segments to target haplotype. In Figure 5, we observe that the number of copied haplotypes decreases with higher λ (e.g., an average of 100 haplotypes are used in the copy process of a new target among 1091 reference haplotypes). On the other hand, in Figure 5, we plot the distance of those useful reference haplotypes from the target haplotype, weighted by the posterior. We observe there is a significant decrease of haplotype copying distance for higher λ value. It strongly suggests that the haplotype copying model can be significantly sped up by only keeping a small number of nearby haplotypes as reference panel. To assess this scenario, we reimputed the target data using gradual decreasing sizes for the reference panel (1091, 800, 600, 400, 200, 100, and 50), where we only keep the

TABLE 1. PERFORMANCE OF SPATIAL MODEL COMPARED TO THE STANDARD MODEL

	Methods	*European*	*Asian*	*African*	*American*
Low frequency variants	Baseline ($\lambda = 0$)	0.5560	0.4115	0.4833	0.5549
	Spatial model with λ^*	0.5834	0.4364	0.4912	0.5654
	Relative improvement	4.92 %	6.05 %	1.63 %	1.89 %
Common variants	Baseline ($\lambda = 0$)	0.7790	0.7189	0.6498	0.7701
	Spatial model with λ^*	0.7939	0.7326	0.6605	0.7765
	Relative improvement	1.90 %	1.91 %	1.64 %	0.84 %

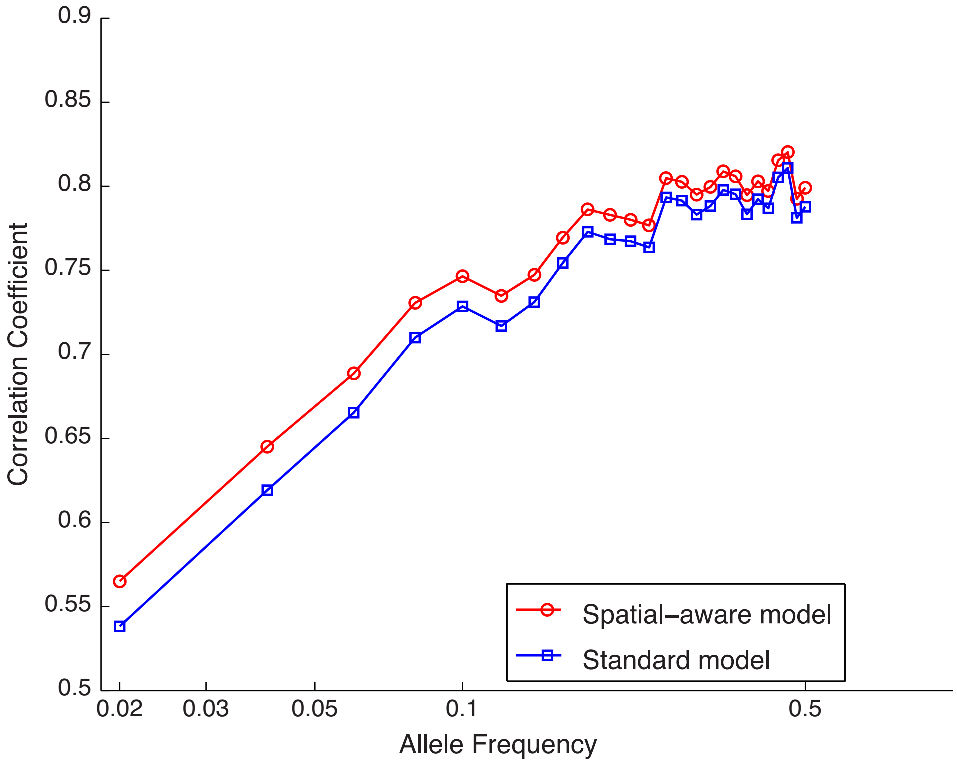


FIG. 4. Absolute imputation improvement across all spectrum of allele frequencies. Spatial-aware model uses λ^* inferred from the data.

most nearby haplotypes in geographical space. The relation between imputation correlation coefficients and computational CPU time is shown in Figure 6. We observe that the computational time can be improved linearly in the size of the reference panel but the imputation performance is also improved even using less number of reference haplotypes. For rare variants, the best imputation performance is obtained at 400 haplotypes and for common variants, the best imputation performance is obtained at 200 haplotypes.

3.4. Localization of individuals on a map

Finally, we explored whether we can use our approach to infer the location on the map of a new individual given data of individuals with known locations. We localized individual haplotypes using

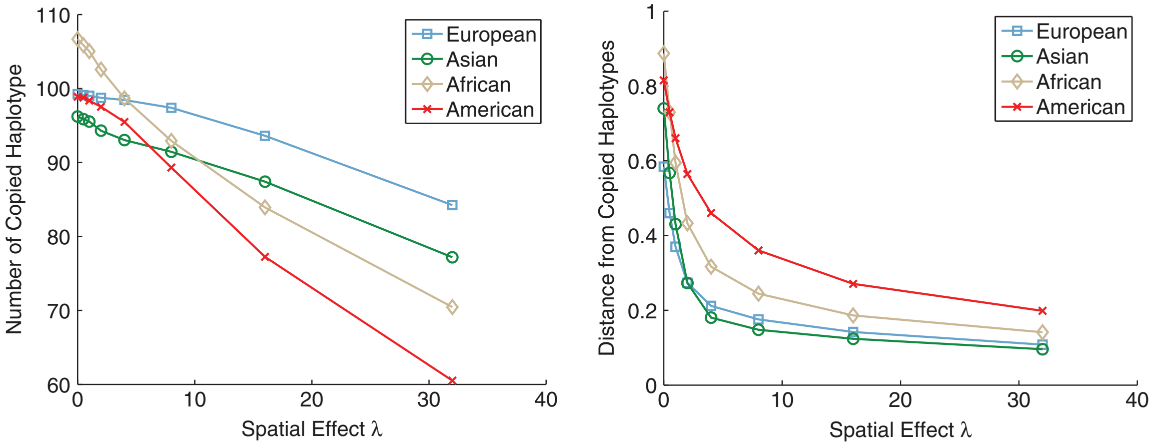


FIG. 5. Spatial effect on copied haplotypes from reference. Left shows that the number of copied haplotypes decreases while the spatial effect parameter is larger. Right shows that the averaged distance from copied haplotype decreases while the spatial effect parameter is larger.

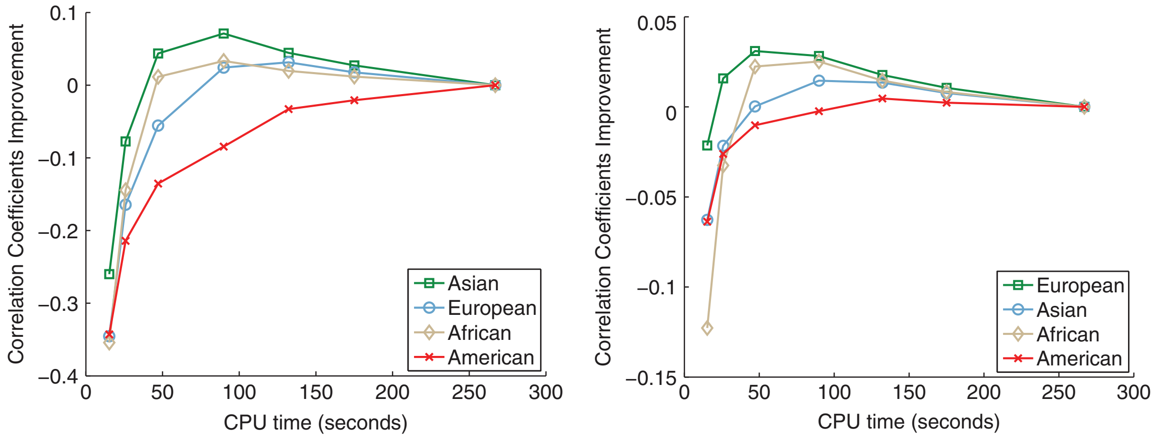


FIG. 6. Imputation accuracy versus computational time. Left shows low-frequency variants (1–5%) while right shows results over common variants ($>5\%$).

spatial-aware copying model with optimal λ value estimated before assuming known locations for the rest of the haplotype data. That is, in each round, we apply spatial-aware model to infer the optimal x^* for one individual using all other other individuals as reference panel (PCA was used to infer locations for the reference panel). We observe that spatial-aware model is able to well identify individual locations, in terms of the clear separating of different continents (see Fig. 7). We observe a high correlation coefficient between the PCA and our inferred geographical ($r = 0.87$), thus showing that our approach can potentially be used to localize individuals on a map given training data with known locations (see Fig. 7).

4. CONCLUSIONS

The haplotype copying model plays an important role in a wide variety of genetic applications. A major drawback is that the model assumes that all haplotypes in the reference panel equally contribute *a priori* to the observed haplotype. In this article, we have proposed a spatial-aware haplotype copying model that takes the spatial effects into account. We have also presented a highly efficient algorithm to estimate the spatial effect parameter before using the proposed model. We applied the proposed model to the 1000 Genomes data set for several applications. First, we estimate the likelihood ratio between the spatial-aware model and spatial-unaware model, and a significant improvement is observed. Second, we test the application of imputation

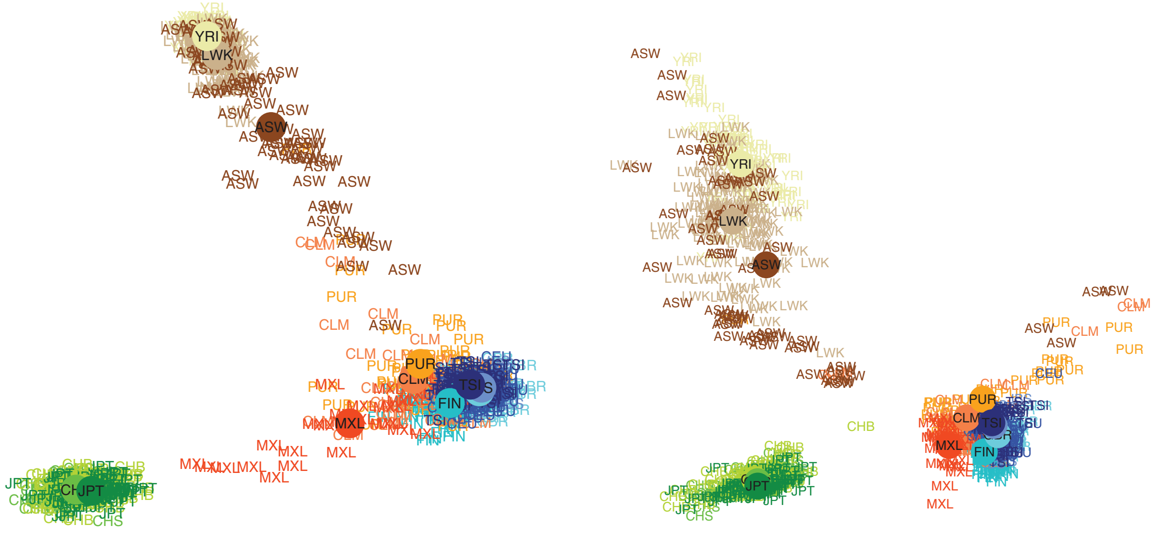


FIG. 7. The left shows results of PCA on chromosome 22 of the 1000 Genomes data while the right shows results of our leave-one-out procedure to localize 1000 Genomes individuals.

using spatial-aware model and obtain significant improvement over the standard model. Finally, we apply this model to localize individuals, and the results indicate high accuracy can be obtained.

ACKNOWLEDGMENTS

W.Y., F.H., and E.E. are supported by National Science Foundation grants 0513612, 0731455, 0729049, 0916676, 1065276, 1302448, and 1320589, and National Institutes of Health grants K25-HL080079, U01-DA024417, P01-HL30568, P01-HL28481, R01-GM083198, R01-MH101782, and R01-ES022282. W.Y. and B.P. are partially supported by National Institutes of Health awards R03-CA162200 and R01-GM053275. We acknowledge the support of the NINDS Informatics Center for Neurogenetics and Neurogenomics (P30 NS062691).

AUTHOR DISCLOSURE STATEMENT

The authors declare that no competing financial interests exist.

REFERENCES

Baran, Y., Quintela, I., Carracedo, A., et al. 2013. Enhanced localization of genetic samples through linkage-disequilibrium correction. *Am. J. Hum. Genet.* 92, 882–894.

Browning, S.R. 2006. Multilocus association mapping using variable-length Markov chains. *Am. J. Hum. Genet.* 78, 903–913.

Browning, S.R., and Browning, B.L. 2010. High-resolution detection of identity by descent in unrelated individuals. *Am. J. Hum. Genet.* 86, 526–539.

Chung, C.C., Kanetsky, P.A., Wang, Z., et al. 2013. Meta-analysis identifies four new loci associated with testicular germ cell tumor. *Nature Genetics* 45, 680–685.

Consortium, G.P., Abecasis, G.R., Altshuler, D., et al. 2010. A map of human genome variation from population-scale sequencing. *Nature* 467, 1061–1073.

Daly, M.J., Rioux, J.D., Schaffner, S.F., et al. 2001. High-resolution haplotype structure in the human genome. *Nature Genetics* 29, 229–232.

de Bakker, P.I.W., Yelensky, R., Pe'er, I., et al. 2005. Efficiency and power in genetic association studies. *Nat. Genet.* 37, 1217–1223.

Delaneau, O., Marchini, J., and Zagury, J.-F. 2012. A linear complexity phasing method for thousands of genomes. *Nature Methods* 9, 179–181.

Gibbs, R.A., Belmont, J.W., Hardenbol, P., et al., 2003. The international hapmap project. *Nature* 426, 789–796.

Han, B., Kang, H.M., and Eskin, E. 2009. Rapid and accurate multiple testing correction and power estimation for millions of correlated markers. *PLoS Genet* 5, e1000456+.

Howie, B., Fuchsberger, C., Stephens, M., et al. 2012a. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nature Genetics* 44, 955–959.

Howie, B., Marchini, J., and Stephens, M. 2011. Genotype imputation with thousands of genomes. *G3: Genes, Genomes, Genetics* 1, 457–470.

Howie, B.N., Donnelly, P., and Marchini, J., 2009. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genetics* 5, e1000529.

Kruglyak, L. 1999. Prospects for whole-genome linkage disequilibrium mapping of common disease genes. *Nature Genetics* 22, 139–144.

Li, N., and Stephens, M. 2003. Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics* 165, 2213–2233.

Li, Y., Sidore, C., Kang, H.M., et al. 2011. Low-coverage sequencing: implications for design of complex trait association studies. *Genome Res.* 21, 940–951.

Li, Y., Willer, C.J., Ding, J., et al. 2010. Mach: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genet. Epidemiol.* 34, 816–834.

Liu, E.Y., Li, M., Wang, W., and Li, Y. 2013. Mach-admix: Genotype imputation for admixed populations. *Genetic Epidemiology* 37, 25–37.

Lohmueller, K.E., Bustamante, C.D., and Clark, A.G. 2009. Methods for human demographic inference using haplotype patterns from genomewide single-nucleotide polymorphism data. *Genetics* 182, 217–231.

Marchini, J., Howie, B., Myers, S., et al. 2007. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nature Genetics* 39, 906–913.

Mardis, E.R. 2008. The impact of next-generation sequencing technology on genetics. *Trends Genet.* 24, 133–141.

Nelson, M.R., Wegmann, D., Ehm, M.G., et al. 2012. An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. *Science* 337, 100–104.

Novembre, J., Johnson, T., Bryc, K., et al. 2008. Genes mirror geography within Europe. *Nature* 456, 98–101.

Pasaniuc, B., Avinery, R., Gur, T., et al. 2010. A generic coalescent-based framework for the selection of a reference panel for imputation. *Genetic Epidemiology* 34, 773–782.

Pasaniuc, B., Rohland, N., McLaren, P.J., et al. Extremely low-coverage sequencing and imputation increases power for genome-wide association studies. *Nat. Genet.* 44, 631–635.

Pasaniuc, B., Sankararaman, S., Kimmel, G., and Halperin, E. 2009. Inference of locus-specific ancestry in closely related populations. *Bioinformatics* 25, i213–i221.

Pool, J.E., Hellmann, I., Jensen, J.D., and Nielsen, R. 2010. Population genetic inference from genomic sequence variation. *Genome Res.* 20, 291–300.

Price, A.L., Tandon, A., Patterson, N., et al. 2009. Sensitive detection of chromosomal segments of distinct ancestry in admixed populations. *PLoS Genetics* 5, e1000519.

Roychoudhury, A., and Stephens, M. 2007. Fast and accurate estimation of the population-scaled mutation rate, theta, from microsatellite genotype data. *Genetics* 176, 1363–1366.

Savage, S.A., Mirabello, L., Wang, Z., et al. 2013. Genome-wide association study identifies two susceptibility loci for osteosarcoma. *Nature Genetics* 45, 799–803.

Schuster, S.C. 2008. Next-generation sequencing transforms today's biology. *Nature Methods* 5, 16–18.

Shendure, J., Mitra, R.D., Varma, C., and Church, G.M. 2004. Advanced sequencing technologies: methods and goals. *Nat. Rev. Genet.* 5, 335–344.

Wegmann, D., Kessner, D.E., Veeramah, K.R., et al. 2011. Recombination rates in admixed individuals identified by ancestry-based inference. *Nature Genetics* 43, 847–853.

Yang, W.-Y., Novembre, J., Eskin, E. and Halperin, E. 2012. A model-based approach for analysis of spatial structure in genetic data. *Nature Genetics* 44, 725–731.

Address correspondence to:
Bogdan Pasaniuc
Department of Pathology and Laboratory Medicine UCLA
10833 Le Conte Ave.
CHS 33-365
Los Angeles, CA 90024

E-mail: bpasaniuc@mednet.ucla.edu