Title

Spatial Localization of Recent Ancestors for Admixed Individuals

Journal

G3: Genes, Genomes, Genetics, 4(12)

ISSN

2160-1836

Authors

Yang, Wen-Yun
Platt, Alexander
Chiang, Charleston Wen-Kai
et al.

Publication Date

2014-12-01

DOI

10.1534/g3.114.014274

Spatial Localization of Recent Ancestors for Admixed Individuals

Wen-Yun Yang,* Alexander Platt,† Charleston Wen-Kai Chiang,† Eleazar Eskin,*,‡,§ John Novembre, and Bogdan Pasaniuc‡,§,††,1**

*Department of Computer Science, †Department of Ecology and Evolutionary Biology, ‡Interdepartmental Program in Bioinformatics, and §Department of Human Genetics, University of California, Los Angeles, California 90095, **Department of Human Genetics, University of Chicago, Chicago, Illinois 60637, and ††Department of Pathology and Laboratory Medicine, Geffen School of Medicine at University of California, Los Angeles, California 90095

ABSTRACT Ancestry analysis from genetic data plays a critical role in studies of human disease and evolution. Recent work has introduced explicit models for the geographic distribution of genetic variation and has shown that such explicit models yield superior accuracy in ancestry inference over nonmodel-based methods. Here we extend such work to introduce a method that models admixture between ancestors from multiple sources across a geographic continuum. We devise efficient algorithms based on hidden Markov models to localize on a map the recent ancestors (*e.g.*, grandparents) of admixed individuals, joint with assigning ancestry at each locus in the genome. We validate our methods by using empirical data from individuals with mixed European ancestry from the Population Reference Sample study and show that our approach is able to localize their recent ancestors within an average of 470 km of the reported locations of their grandparents. Furthermore, simulations from real Population Reference Sample genotype data show that our method attains high accuracy in localizing recent ancestors of admixed individuals in Europe (an average of 550 km from their true location for localization of two ancestries in Europe, four generations ago). We explore the limits of ancestry localization under our approach and find that performance decreases as the number of distinct ancestries and generations since admixture increases. Finally, we build a map of expected localization accuracy across admixed individuals according to the location of origin within Europe of their ancestors.

KEYWORDS
genetic variation
genetic continuum
admixture
localization
ancestry inference

Inference of ancestry from genetic data is a critical aspect of genetic studies, with applications ranging from controlling stratification in disease mapping to the inference of population history (Rosenberg *et al.* 2003; Novembre *et al.* 2008; Drineas *et al.* 2010; Price *et al.* 2010; Rosenberg *et al.* 2010; Seldin *et al.* 2011). Although many initial large-scale genetic association studies have focused primarily on homogeneous populations, increasingly studies are addressing samples in which individuals have more complex backgrounds, including admixed ancestry (*i.e.*, emerging from the mixing of genetically diverged ancestors; Bryc *et al.* 2010; Hinch *et al.* 2011; N'diaye *et al.* 2011; Wegmann *et al.* 2011; Jarvis *et al.* 2012; Perera *et al.* 2013). Such studies depend crucially on accurate and unbiased ancestry inference both at a genome-wide level as well as at each locus in the genome (Seldin *et al.* 2011; Pasaniuc *et al.* 2013).

Traditional ancestry inference from genetic data has been focused on modeling populations as discrete units. As a result, traditional genome-wide ancestry inference estimates the proportion of sites in the genome coming from a set of source populations (continental or subcontinental), and locus-specific inference aims to assign each allele in the genome to one of the considered populations [Pritchard *et al.* 2000; Falush *et al.* 2003; Alexander *et al.* 2009; Price *et al.* 2009; Shringarpure and Xing, 2009; Baran *et al.* 2012; Pasaniuc *et al.* 2013). More recently, alternative approaches model population structure in a geographic continuum, capitalizing on the correlation of

doi: 10.1534/g3.114.014274
Manuscript received July 7, 2014; accepted for publication October 27, 2014; published Early Online November 3, 2014.

Supporting information is available online at http://www.g3journal.org/lookup/suppl/doi:10.1534/g3.114.014274/-/DC1
[1]Corresponding author: Department of Pathology & Laboratory Medicine, Geffen School of Medicine at University of California, Los Angeles, 10833 Le Conte Ave, CHS 33-365, Los Angeles, CA 90095. E-mail: bpasaniuc@mednet.ucla.edu

genetics and geography expected in isolation by distance models (Price *et al.* 2006; Wasser *et al.* 2007; Yang *et al.* 2012; Baran *et al.*, 2013; Elhaik *et al.* 2014) and observed in many organisms (Guillot *et al.* 2009; Storfer *et al.* 2010). Spatial assignment offers three (related) advantages beyond simple population assignment. First, it appropriately acknowledges that nature rarely provides neat boundaries between distinct populations of exchangeable individuals. Second, it allows for model-based inference to exploit the geographic structure of allele frequencies for increased power, and third allows for the accurate assignment of ancestors in otherwise un-sampled or under-sampled regions.

Spatial analysis of genetic data often are performed through principal components analysis (PCA) (Price *et al.* 2006; Seldin *et al.* 2006; Paschou *et al.* 2007; Wasser *et al.* 2007; Novembre *et al.* 2008; Paschou *et al.* 2008; McVean 2009; Zakharia *et al.* 2009; Brisbin *et al.* 2012), a general procedure for reducing the dimensionality of the data, whereas alternative approaches focus on explicit modeling of the relationship between patterns of genetic variation and geography (Wasser *et al.* 2004; Yang *et al.* 2012; Baran *et al.* 2013). These approaches typically assume that an individual's genotype is drawn from the genetic variation present at a single geographic location, or (as in Brisbin *et al.* 2012) assume that ancestral locations are specified *a priori* and then assign individual loci accordingly. These assumptions are clearly violated when individuals have ancestors from multiple unknown geographic regions, as occurs with recently admixed populations in the Americas (such as African-Americans) and more generally, individuals who have ancestry from multiple regions within the same continent (*e.g.*, individuals with recent ancestors from multiple regions of Europe). Recent works have circumvented this issue by first inferring segments of different continental ancestry (*i.e.*, locus-specific) followed by independent application of PCA only on segments of specific continental ancestry (*e.g.*, European segments; Johnson *et al.* 2011; Moreno-Estrada *et al.* 2013). A critical component of such an approach is the performance of locus-specific ancestry inference, which has been shown to attain high accuracy for continental ancestries but to be less accurate in inferring subcontinental ancestry (*e.g.*, country of origin; Price *et al.* 2009; Baran *et al.* 2012; Maples *et al.* 2013). Other approaches to address admixed individuals have only considered the limited case of a first-generation admixed individual (*i.e.*, one parent from each of two locations), in which case local ancestry need not be inferred as individuals are heterozygous at all loci for both ancestries (Wasser *et al.* 2004; Yang *et al.* 2012).

In this work, we introduce models of admixture across varying number of generations and ancestries in a geographic continuum. We model admixed genomes as having recent ancestors from several locations on a genetic-geographical map. We perform ancestry inference by simultaneously localizing on the map the recent ancestors of an admixed individual and partitioning the admixed genome into segments inherited from the same ancestor (*i.e.*, locus-specific ancestry). Assigning a small number of ancestors helps maintain computationally tractability and is biologically meaningful for various important use cases. Where admixture in the recent past is suspected, there will be a particular genomic signature of repeated observations of the same ancestral locations across many unlinked portions of the genome. Modeled this way, we both uncover the nature of the admixture events as well as combine the signal from each of these loci for increased accuracy of the locations of each contributing lineage. We also take advantage of the simple observation that if one allele is inherited from a specific ancestor, then most likely the neighboring alleles are also inherited from the same ancestor. Specifically, we use a model-based framework for genetic variation in the geographical continuum (Yang *et al.* 2012) and use hidden Markov modeling (HMM) of the admixture process (Patterson *et al.* 2004; Gravel, 2012). We develop efficient optimization algorithms that allow us to accurately predict the geographic location of the recent ancestors of an admixed individual in conjunction with locus-specific ancestry inference. The results allow the localization on a geographical map of each allele in recently admixed individuals.

We use empirical genotype data from the Population Reference Sample (POPRES) project (Nelson *et al.* 2008) to validate our approach. The POPRES project has genotyped more than 3000 individuals with ancestry distributed throughout Europe and has recorded the self-reported ancestry (typically at the level of country) for both individuals and their parents/grandparents. We use 1385 POPRES individuals with homogeneous ancestry (*i.e.*, all reported grandparents having the same ancestry) to infer patterns of variation across geography in Europe (Yang *et al.* 2012) and use our method to localize the recent ancestors of individuals with self-reported admixed ancestry (*i.e.*, grandparents with multiple ancestries in Europe). Our method is able to localize the grandparents of the admixed individuals in POPRES data within an average of 470 km of their reported ancestry. The accuracy is dependent on the specific ancestries and ranges from 305 km for individuals with Swiss and French ancestry to 701 km for those with Spanish and Portuguese ancestry. We use simulations from the POPRES genotype data to show that the localization accuracy within Europe decreases with increased number of ancestors and with the number of generations since the admixture. We also show that inference accuracy (at the genome-wide and locus-specific level) increases as distance among ancestors increases. Finally, we provide an analysis of ancestry localization error across all pairs of countries in Europe as resource for the community interested in subcontinental ancestry in Europe.

With the growing availability of high-quality reference panels for nonhuman and nonmodel systems, we expect our proposed methods to be broadly applicable to other organisms as well. Particularly, it could be used to identify the contributing strains to novel hybridized plants, to identify origins of recombinant human pathogens, to identify colonization origins of invasive species, and to investigate environmental refugia/dispersal sources in general source-sink dispersal systems. A software package implementing our methods is freely available at http://bogdan.bioinformatics.ucla.edu/software/.

MATERIALS AND METHODS

Overview of spatial localization for admixed individuals

In this work, we consider models of ancestry for admixed individuals in a geographical continuum. We view the mixed ancestry genome as being generated from several geographical locations on a map, corresponding to the locations of their recent ancestors (see Figure 1). For example, consider the case of an individual with one maternal grandparent from Italy and the other one from Great Britain (see Figure 1A). The maternal copy of its genome will be composed of segments originating from the two locations in Europe (see Figure 1B). Each position in the genome has its own function that describes the population allele frequencies at that site as a function of geography. The approach we take follows spatial ancestry analysis (SPA) (Yang *et al.* 2012) and assumes these functions take on logistic gradient shapes.

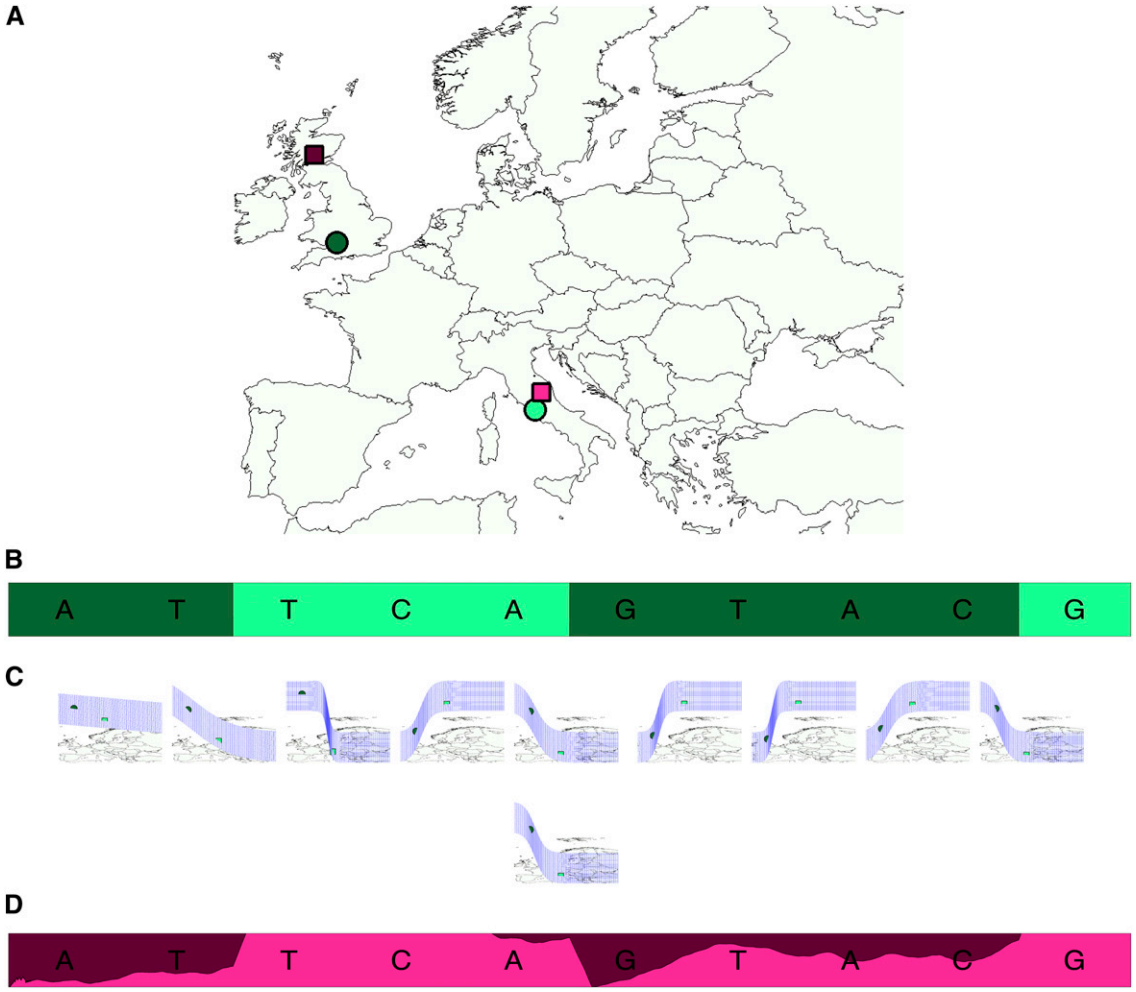

Figure 1 SPAMIX model for admixed individuals. (A) Example of haploid individual with two ancestry locations in Europe (circles denote the true ancestry locations). (B) The admixture process induces segments of different ancestry backgrounds. (C) SPAMIX uses logistic gradients to describe allele frequencies as a function of geographic map to instantiate an admixture hidden Markov modeling for each pair of locations on a map. Each location on the map is associated to a particular allele frequency at all sites in the genome. (D) SPAMIX finds the location of ancestors on a map (denoted by squares in A) and the locus-specific ancestry at each site in the genome by maximizing the likelihood of genotype data.

Some variants may have steep gradients (*i.e.*, frequencies that change drastically with location) whereas other variants may not vary at all with geography (see Figure 1C). Although these types of functions clearly do not explain all correlation between genetics and geography, it has been previously shown that when there are basic isolation-by-distance patterns, such simple functions carry sufficient information to be very informative of ancestry status across individuals (Yang *et al.* 2012; Baran *et al.* 2013) and lead to likelihood functions that are simple to optimize. Other types of functions such as linear functions (Baran *et al.* 2013) or more complex functions could also be employed in our framework.

Having estimated gradient functions at each site in the genome, we extend standard HMMs for admixture to incorporate variation at each position on the map by allowing the emission probabilities to vary according to these gradients (see *Spatial modeling of allele frequency*). We perform inference in this model to find the ancestor locations on the map that maximize the likelihood of the observed genome (Figure 1A). After finding the location of the recent ancestors, we assign each allele in the mixed genome to one of the ancestor locations. This provides a locus-specific ancestry call across the genome. Figure 1D shows an output of our method (SPAMIX) with locations in the admixed genome being labeled according to the inferred ancestral location on the map.

Although we presented a simple example of our framework with two ancestors on the map, the model flexibly handles diploid data with arbitrary number of generations and ancestors localized on the map (*e.g.*, diploid genome with four ancestors localized on the map two generations ago, diploid genome with eight ancestor locations three generations ago) (see *Diploid data with admixed ancestry*). As the number of generations since admixture increases, the total number of ancestors to localize increases dramatically, making the inference problem very challenging. To account for this effect we limit the number of different locations for the recent ancestors for the maternal (paternal) haplotypes to a fixed constant (*e.g.*, $M(N)$, see below) with varying amount of contributions to ancestry (see below). We devise efficient algorithms to jointly optimize the locations of the ancestors as well as the proportion they contribute to the genome of the admixed individual (see below). For example, in the case of three generations ago, one ancestry location may contribute 1/8th to the admixed genome if only

one ancestor comes from that location and may contribute 1/2 to the admixture process if half the ancestors come from that specific location. Finally, we note that the diploid model is symmetric making M and N interchangeable.

Spatial modeling of allele frequency

Although our base method for explicit modeling of genetic variation as function of geography has been described elsewhere (Yang *et al.* 2012; Baran et al. 2013), we briefly present here the generative model. We view each of the alleles of an individual as an independent Bernoulli draw from an allele frequency that changes across the map and we parametrize the allele frequency function through a logistic gradient as a function of position ($\vec{x} = (x_1, x_2)$) in the map. Formally, the probability of observing the reference allele in single-nucleotide polymorphism (SNP) j at position $\vec{x}$ on the map $f_j(\vec{x})$, is defined as:

$$f_j\left(\vec{x}\right) = \frac{1}{1 + \exp\left(-\vec{a}_j^T \vec{x} - b_j\right)} \tag{1}$$

where $\vec{a}_j$ and b_j are parameters specific to SNP j. We estimate $\vec{a}_j$ and b_j from data containing individuals with known homogeneous locations (Yang *et al.* 2012) and then use these coefficients in the inference of ancestries of mixed individuals.

Although easy to manipulate mathematically, the logistic functions we use here clearly do not capture all genetic variation (for example, variants that have multiple modes or peaks in the allele frequency surface as may be typical of rare variants). However, these functions have been shown to capture general trends in common-variant frequencies sufficiently well enough in isolation-by-distance samples to produce highly accurate spatial assignment in individuals with nonmixed ancestry (Yang *et al.* 2012). We hypothesize that such simple-to-manipulate functions are sufficient for accurate localization of recent ancestors in individuals with mixed subcontinental ancestries.

Haploid data with admixed ancestry

Spatial model for admixed haploid data: For simplicity, we start by introducing the model for haploid data and extend it to genotype data in the next section. Denote by $h = (h_1, \ldots, h_L)$ the multisite haplotype of an admixed haplotype, where L is the number of SNPs typed across the genome and $h_i \in \{0, 1\}$ encodes the number of reference alleles at SNP i. Due to the admixture process, the haplotype can be viewed as a mosaic of segments coming from ancestors from multiple locations on the map. We define variables $Z = (z_1, \ldots, z_L)$ as indicators for an allele coming from ancestry location j ($z_i = j$ if allele at locus i is from j-th ancestry location) and write the likelihood of the haplotype data as function of ancestry locations X. The likelihood for a given admixed haplotype data having M ancestry locations $X = (x_1, \ldots, x_M)$ where each ancestry contributes proportionally with $\Pi = (\pi_1, \ldots, \pi_M)$ is defined as:

$$L(h; X, \Pi) = \sum_Z P(Z; \Pi) \prod_{i=1}^{L} P(h_i | z_i; X) \tag{2}$$

The hidden variable Z encodes the mosaic structure of the admixed haplotype (*i.e.*, inheritance within the past generations for recent admixture, admixture-linkage disequilibrium [LD]) and can be modeled using a Markov chain as follows:

$$P(Z; \Pi) = P(z_1; \Pi) \prod_{i=1}^{L-1} P(z_{i+1} | z_i; \Pi)\, P(z_1 = j; \Pi) = \pi_j$$
$$P(z_{i+1} = j | z_i; \Pi) = \begin{cases} (1 - \tau_i) + \tau_i \pi_j & z_{i+1} = z_i \\ \tau_i \pi_j & z_{i+1} \neq z_i \end{cases}$$

where the parameters $\tau = \{\tau_1, \ldots, \tau_{L-1}\}$ stand for the recombination probability (within the past g generations) between each two neighbor loci. The alleles present at a site i on a haplotype is modeled as a Bernoulli variable with a success probability given by the allele frequency $f_i(x_{z_i})$ as follows:

$$P(h_i | z_i; X) = \left(\frac{1}{1 + \exp\left(-a_i^T x_{z_i} - b_i\right)}\right)^{h_i} \left(\frac{1}{1 + \exp\left(a_i^T x_{z_i} + b_i\right)}\right)^{(1-h_i)}$$

where the parameters a and b are estimated beforehand. Thus, they should be regarded as fixed parameters in this article. An illustration of the model is given in Figure 1. We note that our model makes the assumptions of independence of alleles conditional on local ancestry (no modeling of background LD or deviations from Hardy-Weinberg proportions).

Spatial ancestry inference for haploid data: Under the generative model shown previously, spatial ancestry inference is reduced to inferring the M ancestral locations given data for an admixed haplotype, followed by posterior decoding in the HMM to obtain locus-specific predictions. This can be achieved by maximizing the likelihood function (2) with respect to X. By treating X as parameters and Z, Π as hidden variables, this maximization falls within the procedure of the standard expectation-maximization (EM) algorithm (Dempster *et al.* 1977):

E step: The expectation step is similar to the forward-backward algorithm for HMM, which calculates the posterior probability of hidden variables Z given current estimation of ancestral locations $X^{(t)}$ and ancestral proportion $\Pi^{(t)}$:

$$P\left(z_i = j | h; X^{(t)}, \Pi^{(t)}\right) = \frac{\alpha_i(j)\beta_i(j)}{\sum_j \alpha_L(j)}$$

where α/β are standard forward/backward HMM functions and can be efficiently calculated (see Supplementary Note).

M step: The maximization step alternatively optimizes the Q functions in X and in Π (Dempster *et al.* 1977). In detail, it first optimizes the Q function in X by fixing Π. Second, it optimizes the Q function in Π by fixing X. These two steps are performed alternatively until the maximization converges. The Q function in X in the first step can be derived as

$$Q\left(X; X^{(t)}, \Pi^{(t)}\right) = \sum_Z P\left(Z | h; X^{(t)}, \Pi^{(t)}\right) \ln\left(P(Z; \Pi) \prod_i P(h_i | z_i; X)\right)$$
$$\propto \sum_{i,j} C_{ij} q_i(x_j)$$

where C_{ij} denotes the posterior $P(z_i = j | h, X^{(t)}, \Pi^{(t)})$ from the *E step*, and the shorthand $q_i(x_j)$ is defined as:

$$q_i(x_j) = \begin{cases} -\ln\left(1 + \exp\left(a_i^T x_j + b_i\right)\right) & h_i = 0 \\ -\ln\left(1 + \exp\left(-a_i^T x_j - b_i\right)\right) & h_i = 1 \end{cases}$$

The Q function in Π in the second step can be derived as follows

$$Q\left(\Pi; X^{(t)}, \Pi^{(t)}\right) = \sum_Z P\left(Z \middle| h; X^{(t)}, \Pi^{(t)}\right) \ln\left(P(Z;\Pi)\prod_i P(h_i|z_i;X)\right) \propto \sum_{i,j} D_{ij} \ln \pi_j + E_{ij} \ln\left(1-\tau_i\left(1-\pi_j\right)\right)$$

where D_{ij} and E_{ij} denote constants calculated from the posterior $P(z_i = j|h, X^{(t)}, \Pi^{(t)})$ in the *E step*.

We perform the maximization by taking advantage of the convex properties of the equation and using analytical forms for the Hessian of the function. The complete derivations are given in Supporting Information, File S1.

To give an overview of the whole EM algorithm, an illustration of the aforementioned EM algorithm is given in Figure 2.

Locus-specific spatial ancestral inference for haploid data: Having obtained the maximum likelihood geographical locations X^*, we compute the posterior probability for Z, which leads to a locus specific assignment of ancestry at each allele in the genome. The most probable local ancestral locations are found by maximizing

$$\arg\max_Z P(Z|h;X^*) = \arg\max_Z P(h|Z;X^*)P(Z)$$

which can be efficiently solved by the Viterbi algorithm (Viterbi 2006). To compute a posterior probability of each locus-specific ancestry, we use the forward-backward algorithm (see File S1).

Diploid data with admixed ancestry

Spatial model for admixed diploid data: We next extend the haploid model to genotypes by considering M paternal ancestry locations $X = (x_1, \ldots, x_M)$ with proportions $\Pi = (\pi_1, \ldots, \pi_M)$ and N maternal ancestry locations $Y = (y_1, \ldots, y_N)$ with proportions $\Omega = (\omega_1, \ldots, \omega_N)$. Denote by $g = (g_1, \ldots, g_L)$ the multisite genotype of an admixed genotype, where L is the number of SNPs typed across the genome and $g_i \in \{0, 1, 2\}$ encodes the number of reference alleles at SNP i. Then the likelihood becomes:

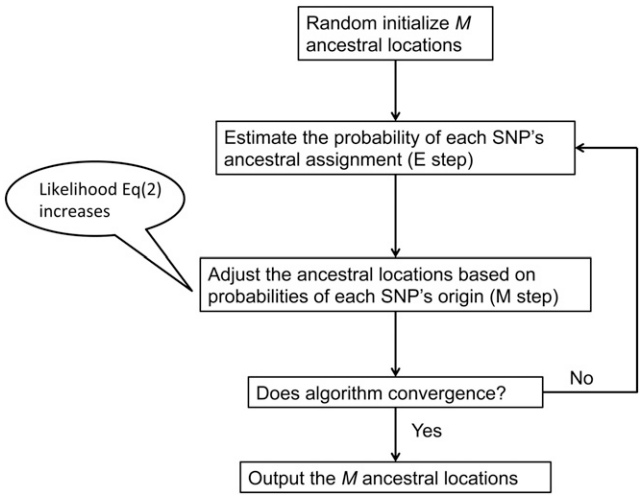

Figure 2 An illustration of the expectation-maximization (EM) algorithm for spatial ancestry inference for haploid data. The E-step and M-step are performed alternatively until the EM algorithm converges. The last M ancestral locations are used as the output of EM algorithm. SNPs, single-nucleotide polymorphisms.

$$L(g;X,Y,\Pi,\Omega) = \sum_Z P(Z,\Pi,\Omega)\prod_{i=1}^{L} P\left(g_i|z_i^p, z_i^m; X, Y\right) \quad (3)$$

The variables Z^p and Z^m now encode the ancestry status of the paternal (maternal) alleles ($z_i^p = j$ denotes that the paternal allele at locus i is from j-th paternal ancestry) and can be modeled through the same Markovian process as:

$$P(Z;\Pi,\Omega) = \left(P(Z_1^p;\Pi)\prod_{i=1}^{L-1} P(z_{i+1}^p|z_i^p;\Pi)\right)\left(P(z_1^m;\Omega)\prod_{i=1}^{L-1} P(z_{i+1}^m|z_i^m;\Omega)\right)$$
$$P(z_1^p = j;\Pi) = \pi_j$$
$$P(z_1^m = k;\Omega) = \omega_k$$
$$P(z_{i+1}^p = j|z_i^p;\Pi) = \begin{cases}(1-\tau_i)+\tau_i\pi_j & z_{i+1}^p = z_i^p \\ \tau_i\pi_j & z_{i+1}^p \neq z_i^p\end{cases}$$
$$P(z_{i+1}^m = k|z_i^m;\Omega) = \begin{cases}(1-\tau_i)+\tau_i\omega_k & z_{i+1}^m = z_i^m \\ \tau_i\omega_k & z_{i+1}^m \neq z_i^m\end{cases}$$

Given the origin of alleles, the likelihood of the admixed individual genotype is modeled as two Bernoulli draws:

$$P(g_i|z_i^p, z_i^m; X, Y) = \begin{cases}\left(1-f_i\left(x_{z_i^p}\right)\right)\left(1-f_i\left(y_{z_i^m}\right)\right) & g_i = 0 \\ \left(1-f_i\left(x_{z_i^p}\right)\right)f_i\left(y_{z_i^m}\right)+f_i\left(x_{z_i^p}\right)\left(1-f_i\left(y_{z_i^m}\right)\right) & g_i = 1 \\ f_i\left(x_{z_i^p}\right)f_i\left(y_{z_i^m}\right) & g_i = 2\end{cases}$$

The function f_i is the allele frequency function in logistic form (1). The probability $P(Z)$ models the recombination events in paternal and maternal ancestries, and the probability $P(g_i|z_i^p, z_i^m; X, Y)$ models the probability of generating the genotype from two ancestral geographical locations.

Spatial ancestry inference for diploid data: We would like to infer $M + N$ ancestral locations for a given mixed individual genotype. This can be achieved by maximizing the likelihood function (3) with respect to X and Y, which, analogous to the haploid case, can be performed using the EM algorithm (Dempster *et al.* 1977).

E step: In short, the expectation step is similar with forward-backward algorithm in HMM, which calculates the posterior probability of hidden variables Z given current estimation of ancestral locations $X^{(t)}$ and $Y^{(t)}$.

$$P\left(z_i^p = j, z_i^m = k|g;X^{(t)},Y^{(t)},\Pi^{(t)},\Omega^{(t)}\right) = \frac{\alpha_i(j,k)\beta_i(j,k)}{\sum_{j,k}\alpha_L(j,k)}$$

where α and β can be calculated recursively using a procedure similar to the forward-backward algorithm for HMMs.

M step: The maximization step alternatively optimizes the Q functions in X, Y Π and Ω. In detail, it first optimizes the Q function in X and Y by fixing Π and Ω. Second, it optimizes the Q function in Π by fixing X and Y and Ω. Third, it optimizes the Q function in Ω by fixing X, Y and Π. These two steps are performed alternatively until the maximization converges. The Q function in X and Y in the first step can be derived as follows

$$Q\left(X,Y;X^{(t)},Y^{(t)},\Pi^{(t)},\Omega^{(t)}\right) = \sum_{Z^p,Z^m} P\left(Z^p,Z^m|g;X^{(t)},Y^{(t)},\Pi^{(t)},\Omega^{(t)}\right)\ln \times\left(P(Z^p;\Pi)P(Z^m;\Omega)\prod_i P\left(g_i|z_i^p,z_i^m;X,Y\right)\right) \propto \sum_{i,j,k} C_{ijk}q_i\left(x_j,y_k\right)$$

where C_{ijk} denotes the posterior $P\left(z_i^p = j, z_i^m = k | g, X^{(t)}, Y^{(t)}\right)$ computed from *E step*, and the shorthand $q_i(x_j, y_k)$ is defined as:

$$q_i(x,y) = \begin{cases} -\ln\left(1+\exp\left(a_i^T x + b_i\right)\right) - \ln\left(1+\exp\left(a_i^T y + b_i\right)\right) & g_i = 0 \\ \ln\left(\begin{array}{l} \dfrac{1}{\left(1+\exp\left(a_i^T x + b_i\right)\right)\left(1+\exp\left(-a_i^T y - b_i\right)\right)} \\ + \dfrac{1}{\left(1+\exp\left(-a_i^T x - b_i\right)\right)\left(1+\exp\left(a_i^T y + b_i\right)\right)} \end{array}\right) & g_i = 1 \\ -\ln\left(1+\exp\left(-a_i^T x - b_i\right)\right) - \ln\left(1+\exp\left(-a_i^T y - b_i\right)\right) & g_i = 2 \end{cases}$$

The Q function in Π in the second step can be derived as

$$\begin{aligned} &Q\left(\Pi; X^{(t)}, Y^{(t)}, \Pi^{(t)}, \Omega^{(t)}\right) \\ &\quad = \sum_{Z^p, Z^m} P\left(Z^p, Z^m | g; X^{(t)}, Y^{(t)}, \Pi^{(t)}, \Omega^{(t)}\right) \\ &\qquad \times \ln\left(P(Z^p;\Pi)P(Z^m;\Omega)\prod_i P\left(g_i | z_i^p, z_i^m; X, Y\right)\right) \\ &\qquad \propto \sum_{i,j} D_{ij}\ln\pi_j + E_{ij}\ln\left(1-\tau_i\left(1-\pi_j\right)\right) \end{aligned}$$

where D_{ij} and E_{ij} denote constants calculated from the posterior $P(z_i^p = j, z_i^m = k | g, X^{(t)}, Y^{(t)})$ in the *E step*. We omit the Q function in Ω in the third step here as it is very similar with the aforementioned function.

As in the haploid case, we leverage the convexity of the function and analytical forms for the Hessian to efficiently optimize the Q function. The complete derivations and optimization details are given in File S1. As noted in File S1, the Q function $Q(X, Y; X^{(t)}, Y^{(t)}, \Pi^{(t)}, \Omega^{(t)})$ in the first step is not concave in general. However, we can still use convex optimization techniques to get a local optimal solution. In practice, we observe that the function is concave almost all the time.

Locus-specific spatial ancestral inference for diploid data: Having obtained the maximum likelihood geographical locations X^* and Y^* for each ancestry, we can compute the posterior probability for Z^p and Z^m, which leads to the spatial local ancestry inference. The most probable local ancestral states are obtained by maximizing

$$\arg\max_Z P(Z|g; X^*, Y^*) = \arg\max_Z P(g|Z; X^*, Y^*)P(Z)$$

which can be efficiently solved by the Viterbi algorithm (Viterbi, 2006). The posterior of local ancestries for each allele can be obtained using a forward-backward algorithm following the *E step* in the algorithm (see Supplementary Note).

Homogeneous paternal and maternal ancestries: In the aforementioned notations we derived the general solution that allows for paternal and maternal ancestries to be different from each other, which is suitable for applications of inference of parental locations or grandparent locations. A simplifying case is when maternal and paternal ancestries are homogeneous, *i.e.*, the paternal haplotype and maternal haplotype are from the same set of ancestral populations. We allow for this case by setting $M = N$ and enforce a constraint $x_j = y_j$ in the M step.

POPRES data set

We applied our methods to a data set collected from European populations, which was assembled and genotyped as part of the larger POPRES project (Nelson *et al.* 2008) and accessed via dbGAP accession number phs000145.v4.p2. A total of 3192 European individuals were genotyped at 500,568 loci using the Affymetrix 500K SNP chip. The same stringency criteria as in Novembre *et al.* (2008) were applied to create the training data. We removed SNPs with low confidence scores and low call-rate (Nelson *et al.* 2008; Novembre *et al.* 2008). We filtered individuals to avoid sampling individuals from outside of Europe, to create more even sample sizes across Europe, and to remove individuals whose self-reported data have grandparents with different origins. We note that this is the same sample set used in Novembre *et al.* (2008) and Baran *et al.* (2013). For the remaining individuals who have reported grandparental data, we use that origin for the individual. Otherwise, we use the individual-level self-reported country of birth. From these data, we infer logistic gradients starting from genotype data from 447,245 autosomal loci in 1385 individuals from 36 populations. A total of 77.4% of SNPs are common SNPs (allele frequency > 0.05), and the remaining 22.6% have low frequencies (< 0.05). As in Novembre *et al.* (2008), we use country geographical center as the geographical locations for all the individuals from that country population. For testing, we identified an additional 470 individuals from the POPRES data that have self-reported grandparental ancestry from two or more countries in Europe. A summary of homogeneous ancestry individuals used in estimating logistic gradients (1385) and with subcontinental European admixed ancestry (470) are given in Table 1.

Background LD

Although our approach models admixture LD, it assumes that markers are independent conditional on local ancestry (no background LD). Preliminary results (not shown) that used transition rates based on the assumed number of generations and recombination rate (*i.e.*, similar to simulations, $(g-1)\phi(d_{i+1} - d_i)$ where d_i's are the locations of each SNPs, g denotes the number of generations, and ϕ is the probability of one recombination per generation per base-pair (Paşaniuc *et al.* 2009) resulted in an increased number of inferred short ancestry windows which led to decreased accuracy. This effect is likely due to lack of modeling of background LD in the model. To remove short ancestry windows (likely spurious, induced by residual LD) we first performed LD pruning at a level of $r^2 < 0.2$ (72,418 SNPs retained) followed by adjustment of the transition rates in our model by a factor of 10^{-2}. This adjustment factor is used to regularize the recombination rate, such that the number of inferred short ancestry windows is appropriate for the number of generations and distinct ancestries and will vary with these parameters. Results at different LD pruning levels and the corresponding adjustment factors are reported in Table S1 and Table S2.

Simulation setup

We use BEAGLE to phase the POPRES data then simulate offspring admixed individuals by modeling recombinations within the last couple of generations. The recombination probability between each SNPs is approximated as $(g-1)\phi(d_{i+1} - d_i)$ where d_i's are the locations of each SNP in base pair, g denotes the number of generations, and

Table 1 Self-reported grandparental ancestry (location of origin) of the POPRES data individuals (1906 in total)

	Number of Different Ancestries					
		2				
	1	(2/2)	(3/1)	Total	3	4
Number of individuals	1385	261	153	414	54	2
Percentage of total	74.7%	14.1%	8.2%	22.3%	2.9%	0.1%

For individuals with grandparental ancestry from 2 different countries, we also report the number of individuals with two grandparents from one location and two from the other (2/2) *vs.* individuals with three grandparents from one country (3/1). POPRES, Population Reference Sample.

ϕ is the probability of one recombination per generation per base pair (Paşaniuc *et al.* 2009). For the recombination map, we assumed a flat recombination rate of $\phi = 10^{-8}$ per base pair. For given number of M paternal ancestries and N maternal ancestries, we randomly select from the POPRES data a set of $M + N$ individuals, each of which has four grandparents from the same location and randomly select one haplotype from each individual. We simulate the admixed haplotypes independently for the maternal and paternal haplotypes using the standard Poission process of admixture block distribution (Price *et al.* 2009). If specified as homogeneous paternal and maternal ancestries, we pick M instead of $M + N$ ancestries and use the same M ancestries for both paternal and maternal haplotype simulation.

For the SPAMIX haploid model, the simulated haplotypes are used as input directly. Also, we always use the correct number of ancestries M or N as input. For the SPAMIX diploid model, the combined genotype from two simulated paternal and maternal haplotypes are used as input. To avoid testing bias, we estimate the allele frequency logistic gradients each time using the POPRES individuals with the $M + N$ simulation ancestors excluded from the training set. We do not optimize over the ancestry proportions but provide them as input to SPAMIX. The ancestry proportions are fixed as uniformly distributed among all ancestors.

We use several metrics to assess performance of SPAMIX in simulations and real data. For the ancestral location prediction, we evaluate the results by computing the average geographical distance between predicted locations and true locations in simulations (*prediction error*). To account for the distance among ancestries, we also compute the *relative prediction error*, defined as the ancestral location prediction error divided by the distance between the true ancestry locations used in simulations. Note that we use as the "true" ancestral locations for the admixed individual the set of country centers from the $M + N$ ancestries.

For locus-specific inference, we propose two different metrics. The first one is the *local ancestry prediction error*, which is the average distance between predicted location and true location at each locus. The second metric we use is the *local ancestry prediction accuracy*, defined as the percentage of loci across the genome with correct assignment of ancestry. To account for the ambiguity in matching the true to inferred ancestries, we permute the inferred ancestries to find the closest match in terms of inferred location to true location.

RESULTS

Performance of continuous ancestry inference in simulations

We investigated the performance of our model through simulations from the POPRES data (Nelson *et al.* 2008). The POPRES data measures genome-wide genetic variation in a large number of individuals with ancestries across Europe (with a larger proportion of individuals with ancestry from the Central and Western Europe). For each individual, the self-reported ancestry (typically at the level of country) of parents and grandparents was tabulated. To produce a large number of admixed individuals on which to test the data, we first randomly selected individuals with homogenous ancestry (*i.e.*, all four grandparents from the same country of origin) from various areas in Europe to serve as "ancestor" genomes. We used them to simulate an admixed individual and attempted to recover the original ancestral locations using the simulated genome and a set of logistic gradients inferred from the remaining unused individuals (see the section *Materials and Methods*). SPAMIX attains an ancestry localization accuracy (*i.e.*, average distance between true and inferred locations of the recent ancestors) for individuals with two recent ancestors in Europe of 550 km (see Table 2). There is a large variance (334 km) across different sets of ancestral pairs, showing the high variability in performance across subjects. Contributing to this effect is the variable sampling density of the POPRES data (which is denser toward the center of Europe) and variability of goodness-of-fit of the logistic gradients across the map (*e.g.*, poorer fit at range boundaries).

To test how much is gained by explicit modeling of correlations among SNPs induced by segments of recent shared ancestry (admixture LD), we also inferred the recent ancestry location using a naïve model that assumes all SNPs to be independent (as in Yang *et al.* 2012). It can also be understood as SPAMIX with completely random transition probability, which is equivalent to independent SNP assumption. We observe a significant increase in the average distance between true and inferred locations in this naïve model (880 km) *vs.* the 550 km for SPAMIX, thus showing that modeling admixture LD significantly increases performance. We also quantified the effect of correlations among markers conditional on local ancestry (background LD) in our approach. Eliminating loci found in strong disequilibrium with each other (LD pruning) was observed to increase accuracy even though the model had less data to use (see Table S1 and Table S2); therefore, all results in the main text are obtained after LD pruning ($r^2 < 0.2$, see the section *Materials and Methods*).

It is increasingly often the case that access to pedigree data allows haplotypes to be determined with high accuracy. Therefore, we quantified the gain in ancestry localization accuracy arising from having access to phased haplotype data (*i.e.*, haploid data) compared with unphased diploid data. Table 2 shows that accurate phasing significantly increases localization accuracy. For example, having access to perfect phasing allows for the inference of the four ancestral locations (two ancestors for each haplotype) within 557 km of the simulated location where the diploid model for four ancestral locations attains an average of 639 km of its simulated locations.

An important parameter of our model is the number of generations since admixture; with more generations, more recombination events have the opportunity to shuffle ancestry across the genome thus reducing the average length of the ancestry segments. We observe a slight decrease in performance from two to eight generations (548 to 562 km), which we expect to continue as the

Table 2 Average distance between inferred and true ancestry locations in simulated admixed individuals from POPRES data

No. Ancestries	1	2	3	4
Naive model	443 ± 4 (265)	880 ± 5 (491)	898 ± 10 (530)	880 ± 9 (578)
SPAMIX haploid model	458 ± 4 (273)	557 ± 4 (334)	620 ± 7 (392)	665 ± 7 (449)
SPAMIX diploid model	443 ± 4 (265)	550 ± 4 (326)	591 ± 7 (367)	639 ± 7 (423)
SPAMIX (logistic)	75 ± 1 (41)	236 ± 5 (131)	363 ± 6 (215)	419 ± 6 (247)

The inferred locations are global for the admixed individuals, which marginalizes over all possible uncertainty on the local genome blocks assignment. Simulations assume four generations in the mixture process. Naive model denotes the extension of SPA that ignores admixture-LD. SPAMIX (logistic) represents simulation results starting from haplotypes generated at a location on a map using a Bernoulli sampling from the logistic gradients (see the section *Materials and Methods*). Parentheses denote the SD, whereas SEM is computed as SD divided by square of number of simulations in each category. POPRES, Population Reference Sample; SPA, spatial ancestry analysis; LD, linkage disequilibrium.

Table 3 Average distance between inferred and true ancestry locations in simulated admixed individuals from POPRES data as function of number of generations in the mixture process

No. Generation	2	4	6	8
Naive model	899 ± 17 (487)	880 ± 5 (491)	864 ± 10 (466)	927 ± 11 (491)
SPAMIX	548 ± 12 (329)	550 ± 4 (326)	541 ± 7 (295)	562 ± 8 (336)

The inferred locations are global for the admixed individuals, which marginalizes over all possible uncertainty on the local genome blocks assignment. Two ancestral locations were assumed for this simulation. Parenthesis denote the SD, whereas SEM is computed as SD divided by square of number of simulations in each category. POPRES, Population Reference Sample.

number of generations increases (in the limit of extremely large number of generations, our model is equivalent to the naïve model that does not model admixture-LD; see Table 3), since the recombination rate will be large enough to produce random switches.

Our framework models genetic variation as function of geography by assuming a logistic gradient for the spatial distribution of genetic variations (see the section *Materials and Methods*). That is, the frequency of a given variant is allowed to change in a given direction on a map only according to a parametrized logistic function. Although this approach has been shown to provide a good approximation of common variation that leads to accurate ancestry inference, we hypothesize that the error in fitting logistic gradients to real data limits the method's accuracy. To assess this scenario, instead of using real individual's haplotype data, we simulated admixed haplotypes directly from the logistic gradients we inferred from POPRES data (see the section *Materials and Methods*). We observe a large increase in accuracy in this idealized scenario as compared to simulations from real haplotype data (*e.g.*, 236 *vs.* 550 km for two ancestries four generations ago, Table 2), thus indicating that logistic gradients do not account for all the correlation between geography and genetic variation. This suggests that further work on functions linking geography to genetics within our framework may yield additional improvements (see the section *Discussion*).

We investigated the performance of our approach as we increase the number of ancestral locations (*M*/*N*, see the section *Materials and Methods*) to estimate for a given admixed individual. For a fixed number of generations (four), we varied the number of ancestry locations to estimate. The parental inference is different from two ancestry inference, as the parental inference assumes that one haplotype is from paternal ancestry and one from maternal ancestry. However, the two-ancestry inference assumes that both of the haplotypes are mosaic of two ancestries ($M = N = 2$). As expected, we observe decreases in performance as the number of ancestral locations increases. For example, the average prediction error increases from 550 for two ancestries to 639 km for four ancestral locations (Table 2).

Increased distance between ancestral locations improves performance

It is well known that accuracy of ancestry inference correlates with genetic distance between ancestral populations. Discrete local ancestry can be inferred with very high degree of accuracy in mixtures of highly diverged populations (*e.g.*, African Americans) compared with closely related ones (*e.g.*, subcontinental mixtures) (Basu *et al.* 2008; Paşaniuc *et al.* 2009; Price *et al.* 2009; Maples *et al.* 2013). Because geography correlates with genetic distance, we hypothesized that the accuracy of continuous ancestry inference in recently admixed individuals also correlates with distance among ancestries on the map. Indeed, we observe that the relative prediction error (*i.e.*, the difference between predicted and true locations normalized by the distance between the true ancestry locations, see the section *Materials and Methods*) decreases with the distance between ancestries in Europe (Figure 3A). For example, if the ancestries are 500 km apart, we observe a relative prediction error of 0.75 compared with 0.50 when the ancestries are located 2000 km apart. Interestingly, when not normalizing for the distance between ancestries (Figure 3B), we observe that prediction error increases with increased distance. This shows that although the task of separating the ancestry locations becomes simpler, the localization accuracy becomes poorer (*e.g.*, two ancestors located 500 km apart are localized within 450 km of their true locations, whereas two ancestors located 3000 km apart are localized within 1000 km of their true locations). This effect is presumably due to

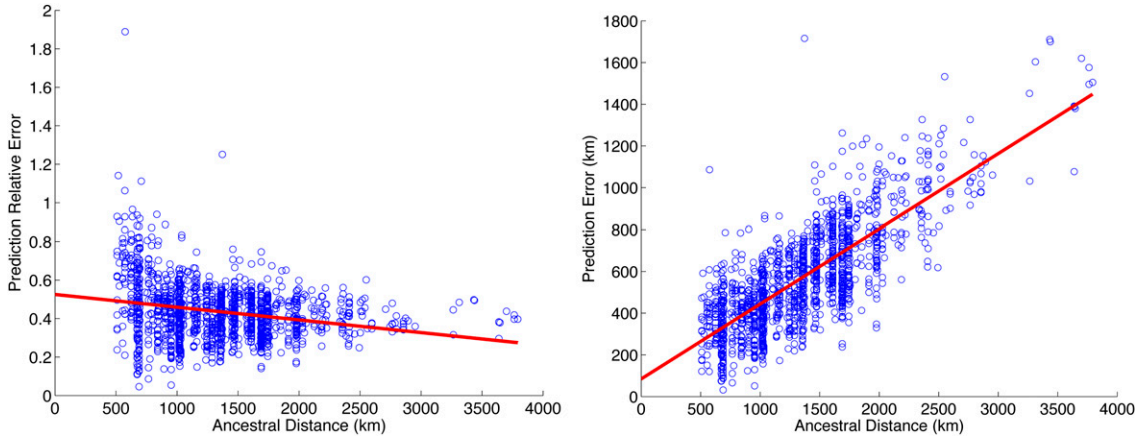

Figure 3 Ancestral location prediction error as a function of distance between ancestral locations in simulations over Population Reference Sample data. Left, the prediction error normalized by the distance between the ancestral locations used in simulations; right, plot of the prediction error. Simulations use the haploid model with two generations in the mixture.

assignment errors in the local ancestry that have a much bigger impact if the ancestral locations are further apart. Although fewer local ancestry errors are being made with increased distance (see *Locus-specific inference*), these errors have a stronger impact on the ancestral localization due to their higher distance to the true location.

Inference of number of ancestors

In the aforementioned simulations, we have assumed that the true number of different ancestry locations is known. We investigated whether our approach can also be used to predict the number of distinct ancestries on the map for a given genome. We used the standard Akaike information criterion (AIC; Bozdogan 1987) that balances the goodness of fit with the number of parameters in the model (more ancestries to infer increases the number of parameters in our method). Starting from POPRES individuals with homogeneous ancestry we simulated admixed individuals with up to 4 ancestry locations 4 generations ago under the constraint that the ancestries are at least 600 kilometers apart. For each simulated admixed individual, we ran our method SPAMIX using $N = 1, 2, 3, 4$ ancestry locations and used the AIC to infer the number of ancestors. Figure 4 shows that this procedure will on average estimate the number of ancestries correctly, but the error rate is expected to be high for any single case of inference.

Locus-specific inference

An advantage of our framework is that in addition to identifying the most likely locations of the recent ancestry of admixed individuals, it can also provide an assignment of each allele in the genome to each ancestry location. We observe that local ancestry prediction accuracy (*i.e.*, the proportion of alleles assigned to the correct ancestry, see the section *Materials and Methods*) increases with the distance between ancestral locations (Figure 5) from 55% of loci assigned accurately for very closely related ancestries (less than 500 km apart) to more than 70% for ancestries 2500 km apart (Figure S1). Similar to the ancestor localization, we observe that although the total number of assignment errors is reduced with increased distance, these errors have a bigger impact when averaging across all sites to compute the average allele

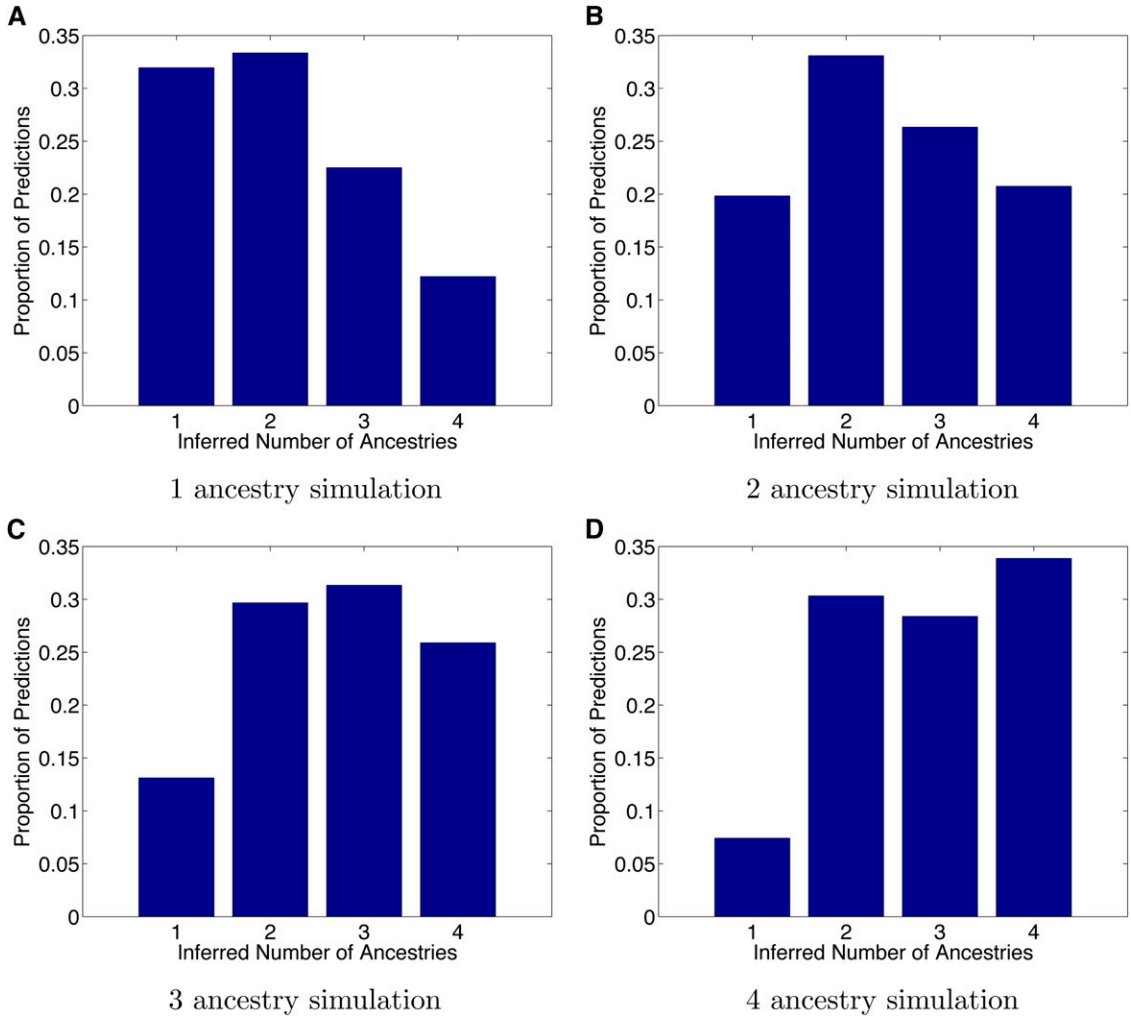

Figure 4 Inference of number of distinct ancestries using the Akaike information criterion (AIC). We simulated 1000 admixed individuals with up to four distinct ancestry sources in Europe and used the AIC within the SPAMIX model to infer the number of ancestries. (A–D) Proportion of inferred number of ancestries (y-axis) as function of number of simulated ancestries (x-axis). Although we observed a large variance in the number of predicted ancestries, we note that the histogram is centered on the correct simulated number of ancestries, thus suggesting that AIC could be used to infer the number of distinct ancestors.

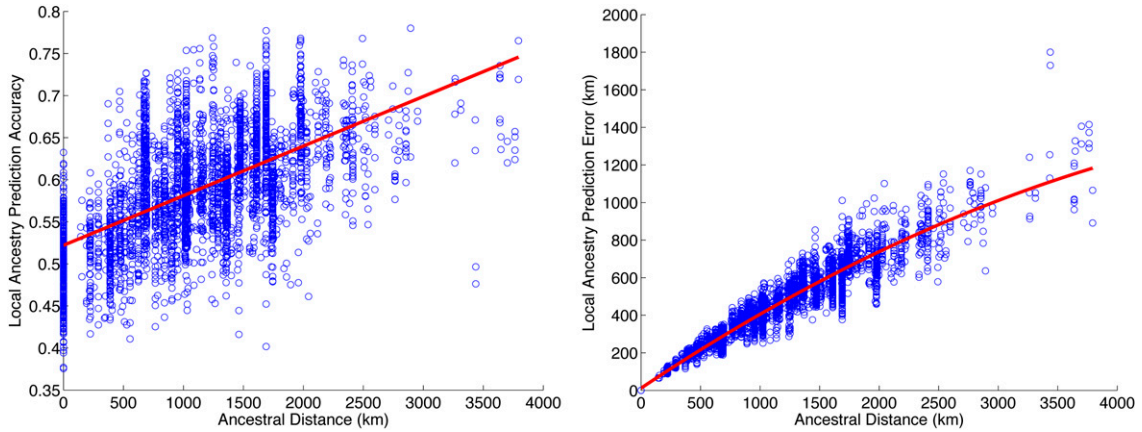

Figure 5 SPAMIX locus-specific ancestry prediction accuracy as function of distance between ancestral locations. Left, local ancestry prediction accuracy, defined as the percentage of all loci with correct assignment of ancestry. Right, average distance to true locations for each allele in the genome (local ancestry prediction error). Simulations use the haploid model with two generations in the mixture.

localization error. Therefore, we observe that the average local ancestry prediction error is increased as distance between ancestral locations is increased.

Map of accuracy across Europe

We also investigated the variance in performance according to the ancestor's labeled origin (*i.e.*, typically to level of country). Figure 6 shows the prediction error for admixed individuals with ancestry from pairs of origins in Europe. In general, we observe decreased performance for populations at the boundary of the European map (*e.g.*, Portugal, Spain, Italy), and increased performance for subcontinental admixtures from populations located geographically in the center of Europe (*e.g.*, France, Switzerland) (Figure S2 and Figure S3). This can be an effect of biased sampling in the POPRES data, that sampled more individuals from European center, but also can be an effect of having more information to localize individuals in SPAMIX. In general, we observe a prediction accuracy ranging from 411 km for admixtures from Spain and Italy to 641 km for individuals with recent ancestors from Spain and the United Kingdom.

Analysis of real admixed individuals from POPRES data

Finally, we investigated whether high accuracies observed in simulations also can be attained in real data. Using SPAMIX, we localized the recent ancestry of all admixed European individuals from POPRES (see the section *Materials and Methods*). A total of 470 admixed individuals were analyzed using SPAMIX (see Table 1 and Figure 7). As "ground truth" ancestral locations, we used the center of the self-reported grandparent country of origin. Therefore, we assume the mixed individuals from POPRES have two to four ancestry locations to infer. Across all 470 individuals, we observe an average prediction error distance of 470 km which decreases to 426 if outlier individuals defined as those with prediction errors greater than 1000 km are removed (all such outlier individuals are reported in Table S3). The error distance is lower than simulated experiments likely due to the large proportion of the admixed individuals of French and Swiss ancestries, which can be accurately localized (average of 305 km). As discussed previously, we note that SPAMIX ancestor localization performance varies greatly across Europe with ancestors from pairs of countries localized at the boundary of European map being harder to localize (*e.g.*, an average of 701 km for ancestor localization for Spanish/Italian mixed individuals).

DISCUSSION

We have introduced new models for predicting the geographical origins of multiple recent ancestors for individuals with recent mixed ancestry. Existing methods for local ancestry inference in admixed populations either focus on discrete ancestry assignment or use locus-specific ancestry inference followed by PCA on subsets of the data. We introduce models that leverage the spatial structure of genetic variation using HMMs for the admixture process to achieve high accuracy in localizing the recent ancestry of a given individual on a geographical map. Our proposed model can be viewed as a generalization of the parental localization model proposed in Yang *et al.* (2012) to account for admixture-LD while allowing for multiple generations and ancestries. Our algorithm is very efficient. For the scale of 100,000 SNPs and four ancestral locations, computation is typically 10 min and uses less than 100 Mb of memory.

Although in our framework we use standard logistic gradient functions that were previously used to link geography and genetic variation, it is worth mentioning that such functions do not capture the whole variability observed in empirical data. To that extent, introducing more flexibility in these functions within the framework for admixture we described here is more likely to provide considerable improvements in accuracy with a tradeoff of computational time. We view this as a promising direction for future study. This is especially important for handling sequencing data, as rare variants rarely are fit well by the gradient functions (results not shown). For example, we are working on using a logistic quadratic function to model the spatial genetic structure, which is a generalization of the original SPA model. It will further enhance the admixed ancestor inference.

Another area for further developments is extending the framework to model background LD (correlations among variants on the same ancestral backgrounds). We found it necessary to modify the transition rates used in our inference by a multiplicative factor based on the level of LD pruning applied to the SNP list (Table S2). Such LD adjustments have proved fruitful in improving localization accuracy for unadmixed individuals (Baran *et al.* 2013) and are likely to improve inference for admixed individuals as well. Although we leave this for future work, one potential approach would be to perform inference within short windows (to account for the local structure of LD) and merge the information within each window into the overall likelihood.

We also note that we used a simplified model of ancestry switching along chromosomes that approximates the pedigree structure. In

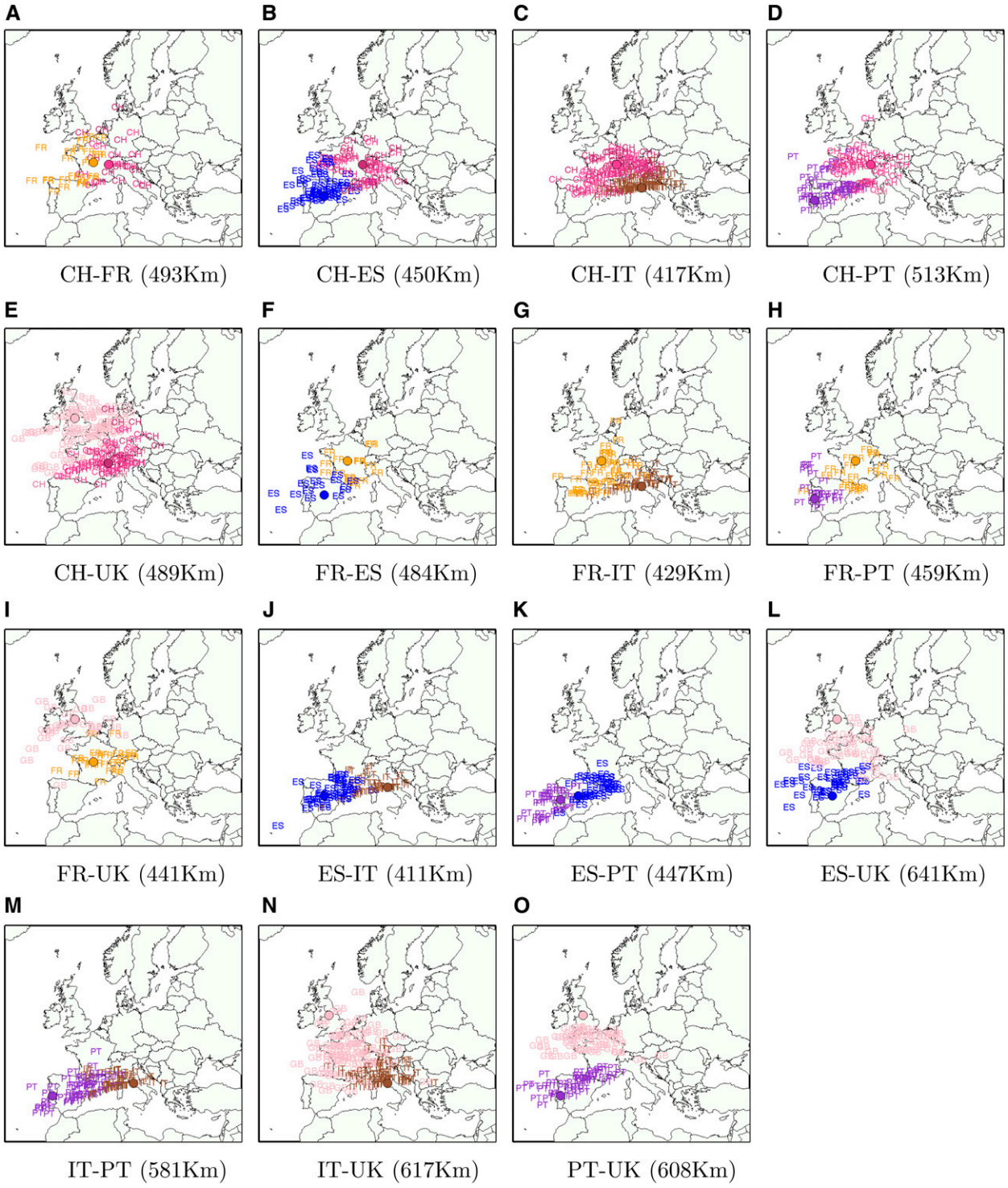

Figure 6 Ancestral location prediction error in simulations of European individuals with ancestry from two locations in Europe, stratified by the country of origin of each location (the country of origin is displayed in different colors). The assumed true locations are displayed by shaded circles. Results in parenthesis denote the average ancestral location prediction error across all simulations. In each simulation the reference data (used to estimate logistic gradients) is disjoint from data used to simulate admixed genomes (see the section *Materials and Methods*). The admixed genome is simulated as four generations ago, and SPAMIX diploid model is used for the inference. The number of simulated pairs can be found in Figure S3.

effect, our approach assumes a fixed effective time-scale of admixture, and ignores the structured transition matrices that are expected due to a fixed pedigree. Future work that explicitly considers the pedigree structure could allow one to address questions regarding the specific timing and configuration of admixed ancestries. For example, for a mixed individual with one Italian and three British grandparents, we

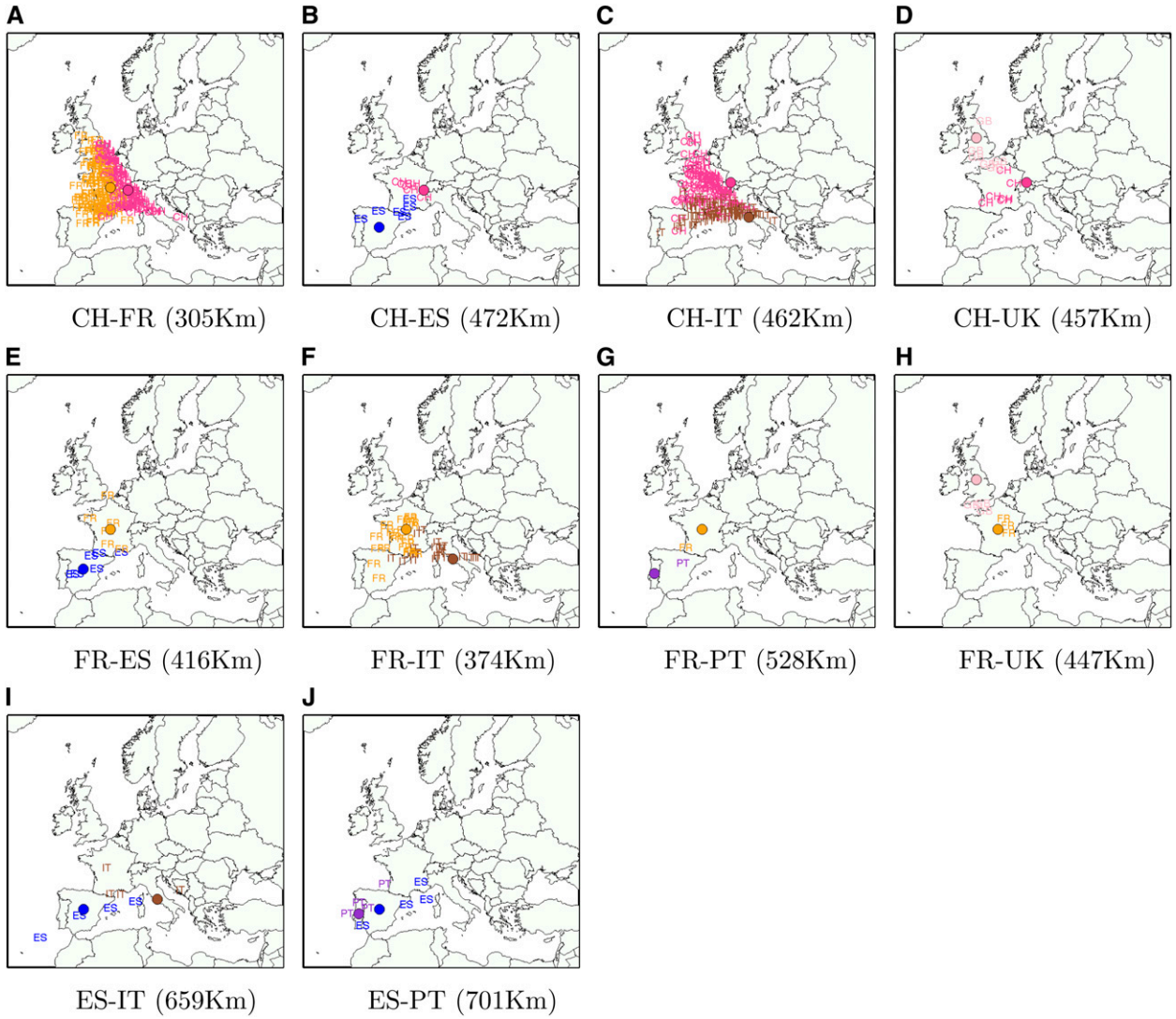

Figure 7 Ancestral location prediction error in real POPRES admixed individuals, stratified by the country of origin of each location. Letters are the inferred locations, and the shaded circles are the assumed true locations.

could incorporate the specific inheritance pattern in the HMM transition rates. The question of whether the ancestors themselves were admixed could be investigated by assigning local ancestry followed by analyzing the length distribution of the ancestry blocks, and we leave this as future work.

For most of the results presented, we assumed that the number of ancestral locations is known. In practice, such information will often not be available. To address this we developed a procedure for inferring the number of ancestries using AIC. In Figure 4 we showed that although on average the correct number of ancestries will be inferred, in a high proportion of cases the inferred number of ancestries will be mistaken. This type of model selection problem is akin to estimating K in the admixture model of STRUCTURE/admixture (Falush *et al.* 2003; Alexander *et al.* 2009) and is typically challenging. In future work, pedigree-based models should lead to constraints on the possible observations that make the number of ancestors more straightforward to infer.

In this work, we focus on the prediction of ancestral locations using an EM algorithm, which is a deterministic method to produce point estimates of the parameters of interest (geographic origins of ancestors) and missing data (the local ancestry of each allele copy). Alternative inference approaches can be taken, for example, the likelihoods we define in Equations (2) and (3) could be used in a Bayesian Markov Chain Monte Carlo approach method to sample from the posterior distribution of the spatial prediction (Wasser *et al.* 2007). In such an approach, an efficient starting point would be from the point estimate obtained via the EM algorithm, which could significantly expedite the convergence of the Markov Chain Monte Carlo approach.

Our work provides a framework of predicting ancestral locations for admixed individuals which can be further improved. For example, due to the continuous nature of the approach, ancestral locations can be predicted to be outside of standard geographic boundaries (*e.g.*, ocean). Future work could improve on our framework by providing constraints to the optimization procedure or by a postoptimization adjustment (selecting the closest location that fits geographical boundaries). In addition, the underlying SPA approach works best in the simple isolation-by-distance settings and is not expected to work well in for complex scenarios with barriers or irregular geometries. We leave a full investigation of such modifications of our framework as

ongoing work and recommend the use of cross-validation approaches with individuals of known ancestry to assess the suitability of the model prior to application.

A direct benefit of the proposed model is that it leads to efficient optimization procedures for tackling inference problems such as localization of ancestors in the genetic-geographic map. In this work we have presented such an algorithm based on the well-known EM procedure, which leverages the HMM of the admixture process joint with the gradient representation of genetic variation as function of geography.

ACKNOWLEDGMENTS

This work is supported in part by the National Institutes of Health (R03-CA162200, R01-GM053275). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

LITERATURE CITED

Alexander, D. H., J. Novembre, and K. Lange, 2009 Fast model-based estimation of ancestry in unrelated individuals. Genome Res. 19: 1655–1664.

Baran, Y., B. Pasaniuc, S. Sankararaman, D. Torgerson, C. Gignoux *et al.*, 2012 Fast and accurate inference of local ancestry in Latino populations. Bioinformatics 28: 1359–1367.

Baran, Y., I. Quintela, A. Carracedo, B. Pasaniuc, and E. Halperin, 2013 Enhanced localization of genetic samples through linkage-disequilibrium correction. Am. J. Hum. Genet. 92: 882–894.

Basu, A., H. Tang, X. Zhu, C. C. Gu, C. Hanis *et al.*, 2008 Genome-wide distribution of ancestry in Mexican Americans. Hum. Genet. 124: 207–214.

Bozdogan, H., 1987 Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. Psychometrika 52: 345–370.

Brisbin, A., K. Bryc, J. Byrnes, F. Zakharia, L. Omberg *et al.*, 2012 PCAdmix: principal components-based assignment of ancestry along each chromosome in individuals with admixed ancestry from two or more populations. Hum. Biol. 84: 343–364.

Bryc, K., C. Velez, T. Karafet, A. Moreno-Estrada, A. Reynolds *et al.*, 2010 Genome-wide patterns of population structure and admixture among Hispanic/Latino populations. Proc. Natl. Acad. Sci. USA 107: 8954–8961.

Dempster, A. P., N. M. Laird, and D. B. Rubin, 1977 Maximum likelihood from incomplete data via the EM algorithm. J. R. Stat. Soc., B 39: 1–38.

Drineas, P., J. Lewis, and P. Paschou, 2010 Inferring geographic coordinates of origin for Europeans using small panels of ancestry informative markers. PLoS One 5: e11892.

Elhaik, E., T. Tatarinova, D. Chebotarev, I. S. Piras, C. M. Calò *et al.*, 2014 Geographic population structure analysis of worldwide human populations infers their biogeographical origins. Nat. Commun. 5: 3513.

Falush, D., M. Stephens, and J. K. Pritchard, 2003 Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. Genetics 164: 1567–1587.

Gravel, S., 2012 Population genetics models of local ancestry. Genetics 191: 607–619.

Guillot, G., R. Leblois, A. Coulon, and A. C. Frantz, 2009 Statistical methods in spatial genetics. Mol. Ecol. 18: 4734–4756.

Hinch, A., A. Tandon, N. Patterson, Y. Song, N. Rohland *et al.*, 2011 The landscape of recombination in African Americans. Nature 476: 170–175.

Jarvis, J., L. Scheinfeldt, S. Soi, C. Lambert, L. Omberg *et al.*, 2012 Patterns of ancestry, signatures of natural selection, and genetic association with stature in Western African pygmies. PLoS Genet. 8: e1002641.

Johnson, N., M. Coram, M. Shriver, I. Romieu, G. Barsh *et al.*, 2011 Ancestral components of admixed genomes in a Mexican cohort. PLoS Genet. 7: e1002410.

Maples, B. K., S. Gravel, E. E. Kenny, and C. D. Bustamante, 2013 RFMix: a discriminative modeling approach for rapid and robust local-ancestry inference. Am. J. Hum. Genet. 93: 278–288.

McVean, G., 2009 A genealogical interpretation of principal components analysis. PLoS Genet. 5: e1000686.

Moreno-Estrada, A., S. Gravel, F. Zakharia, J. L. McCauley, J. K. Byrnes *et al.*, 2013 Reconstructing the population genetic history of the Caribbean. PLoS Genet. 9: e1003925.

N'Diaye, A., G. K. Chen, C. D. Palmer, B. Ge, B. Tayo *et al.*, 2011 Identification, replication, and fine-mapping of loci associated with adult height in individuals of African ancestry. PLoS Genet. 7: e1002298.

Nelson, M. R., K. Bryc, K. S. King, A. Indap, A. R. Boyko *et al.*, 2008 The population reference sample, POPRES: a resource for population, disease, and pharmacological genetics research. Am. J. Hum. Genet. 83: 347–358.

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

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

Pasaniuc, B., S. Sankararaman, D. G. Torgerson, C. Gignoux, N. Zaitlen *et al.*, 2013 Analysis of Latino populations from GALA and MEC studies reveals genomic loci with biased local ancestry estimation. Bioinformatics 29: 1407–1415.

Paschou, P., E. Ziv, E. G. Burchard, S. Choudhry, W. Rodriguez-Cintron *et al.*, 2007 PCA-correlated SNPs for structure identification in worldwide human populations. PLoS Genet. 3: e160.

Paschou, P., P. Drineas, J. Lewis, C. M. Nievergelt, D. A. Nickerson *et al.*, 2008 Tracing sub-structure in the European American population with PCA-informative markers. PLoS Genet. 4: e1000114.

Patterson, N., N. Hattangadi, B. Lane, K. E. Lohmueller, D. A. Hafler *et al.*, 2004 Methods for high-density admixture mapping of disease genes. Am. J. Hum. Genet. 74: 979–1000.

Perera, M. A., L. H. Cavallari, N. A. Limdi, E. R. Gamazon, A. Konkashbaev *et al.*, 2013 Genetic variants associated with warfarin dose in African-American individuals: a genome-wide association study. Lancet 382: 790–796.

Price, A., N. Patterson, R. Plenge, M. Weinblatt, N. Shadick *et al.*, 2006 Principal components analysis corrects for stratification in genome-wide association studies. Nat. Genet. 38: 904–909.

Price, A., A. Tandon, N. Patterson, K. Barnes, N. Rafaels *et al.*, 2009 Sensitive detection of chromosomal segments of distinct ancestry in admixed populations. PLoS Genet. 5: e1000519.

Price, A., N. Zaitlen, D. Reich, and N. Patterson, 2010 New approaches to population stratification in genome-wide association studies. Nat. Rev. Genet. 11: 459–463.

Pritchard, J. K., M. Stephens, and P. Donnelly, 2000 Inference of population structure using multilocus genotype data. Genetics 155: 945–959.

Rosenberg, N. A., L. M. Li, R. Ward, and J. K. Pritchard, 2003 Informativeness of genetic markers for inference of ancestry. Am. J. Hum. Genet. 73: 1402–1422.

Rosenberg, N. A., L. Huang, E. M. Jewett, Z. A. Szpiech, I. Jankovic *et al.*, 2010 Genome-wide association studies in diverse populations. Nat. Rev. Genet. 11: 356–366.

Seldin, M. F., R. Shigeta, P. Villoslada, C. Selmi, J. Tuomilehto *et al.*, 2006 European population substructure: clustering of northern and southern populations. PLoS Genet. 2: e143.

Seldin, M., B. Pasaniuc, and A. Price, 2011 New approaches to disease mapping in admixed populations. Nat. Rev. Genet. 12: 523–528.

Shringarpure, S., and E. P. Xing, 2009 mStruct: inference of population structure in light of both genetic admixing and allele mutations. Genetics 182: 575–593.

Storfer, A., M. A. Murphy, S. F. Spear, R. Holderegger, and L. P. Waits, 2010 Landscape genetics: where are we now? Mol. Ecol. 19: 3496–3514.

Viterbi, A., 2006 Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. IEEE Trans. Inf. Theory 13: 260–269.

Wasser, S. K., A. M. Shedlock, K. Comstock, E. A. Ostrander, B. Mutayoba *et al.*, 2004 Assigning African elephant DNA to geographic region of

origin: applications to the ivory trade. Proc. Natl. Acad. Sci. USA 101: 14847–14852.

Wasser, S. K., C. Mailand, R. Booth, B. Mutayoba, E. Kisamo *et al.*, 2007 Using DNA to track the origin of the largest ivory seizure since the 1989 trade ban. Proc. Natl. Acad. Sci. USA 104: 4228–4233.

Wegmann, D., D. Kessner, K. Veeramah, R. Mathias, D. Nicolae *et al.*, 2011 Recombination rates in admixed individuals identified by ancestry-based inference. Nat. Genet. 43: 847–853.

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

Zakharia, F., A. Basu, D. Absher, T. L. Assimes, A. S. Go *et al.*, 2009 Characterizing the admixed African ancestry of African Americans. Genome Biol. 10: R141.

Communicating editor: S. I. Wright