Multiple-Ancestor Localization for Recently Admixed Individuals

Yaron Margalit[1,⋆], Yael Baran[1,⋆], and Eran Halperin[1,2,3]

[1] The Blavatnik School of Computer Science, Tel Aviv University, Israel
[2] Department of Molecular Microbiology and Biotechnology, Tel-Aviv University, Israel
[3] International Computer Science Institute, Berkeley, CA, 94704

Abstract. Inference of ancestry from genetic data is a fundamental problem in computational genetics, with wide applications in human genetics and population genetics. The treatment of ancestry as a continuum instead of a categorical trait has been recently advocated in the literature. Particularly, it was shown that a European individual's geographic coordinates of origin can be determined up to a few hundred kilometers of error using spatial ancestry inference methods. Current methods for the inference of spatial ancestry focus on individuals for whom all ancestors originated from the same geographic location.

In this work we develop a spatial ancestry inference method that aims at inferring the geographic coordinates of ancestral origins of recently admixed individuals, *i.e.* individuals whose recent ancestors originated from multiple locations. Our model is based on multivariate normal distributions integrated into a two-layered Hidden Markov Model, designed to capture both long-range correlations between SNPs due to the recent mixing and short-range correlations due to linkage disequilibrium. We evaluate the method on both simulated and real European data, and demonstrate that it achieves accurate results for up to three generations of admixture. Finally, we discuss the challenges of spatial inference for older admixtures and suggest directions for future work.

Keywords: admixture, ancestry inference, spatial model, hidden markov model, multivariate-normal distribution

1 Introduction

Determining the ancestry of individuals based on their DNA sequence is a common and useful task in the study of human genetics: In addition to its multiple applications in population genetics [5,10,9,24,7], it has a critical role in correcting for confounders in disease studies [19]: If different ancestral composition in cases and controls is not accounted for, ancestry-informative markers will appear to be disease-related. Admixed individuals, *i.e.* individuals whose ancestry is mixed,

⋆ Contributed equally to this work.

are of special importance in discovering disease-associated traits. The analysis of individuals from admixed populations, such as African-Americans and Latinos, have allowed for the detection of multiple disease association signals [28,8,12] using Mapping by Admixture Linkage Disequilibrium (MALD) [22] as well as other computational techniques [21]. An accurate characterization of the ancestral origins of study samples is therefore critical when population structure exists in the dataset; improved accuracy should lead to more accurate control of confounding, which in turn leads to increased statistical power to detect disease-related markers.

Until recently, the ancestry of admixed individuals was treated as a discrete trait, and individuals were classified into pre-defined classes (*e.g.* Tuscan, Sicilian, Catalan). For admixed individuals, different methods were developed to determine, under the assumption that their multiple ancestries are known, the fraction of genome originating from each one [20,1], and the ancestral origin of each genomic region [18,2,13,6]. In reality, however, ancestries are not discrete, because different populations go through constant mixing whose rate is determined to a large extent by the geographic distance between them. Ancestries are therefore better described as a continuum strongly correlated with geographic structure. One recent work attempts to deal with this challenge by modeling individuals as a mixture of a large panel of reference populations, in a procedure that does not require any prior knowledge about the number or identity of the mixture components; this approach moves closer to a continuous representation, but still suffers from the principal drawbacks of other discrete approaches. Other works tried to cope with this challenge by first classifying the parts of the genomes into continents of origins, and then obtaining within-continent continuous spatial separation using Principal Component Analysis (PCA) [15,11]. One serious problem with this approach is that when the mixing populations are from the same continent, *e.g.* when attempting to separate two different European origins [13], the preliminary classification stage is inaccurate. A more fundamental problem is that the use of PCA for localization is merely a heuristic, and although PCA maps sometimes fit well with the geographic map [17], this is not always the case [23].

Motivated by above difficulties, a few probabilistic spatio-genetic models have recently been developed [26,3,27]. These models describe the allele frequency of each SNP as a continuous function of the geographic coordinates. In addition to improved accuracy of localization compared with PCA, these models can be naturally extended to describe individuals of mixed origin. Indeed, one of these methods named SPA [26] included an immediate extension of its model for the inference of two different origins, paternal and maternal. As we show below, this trivial extension does not provide accurate localization. In addition, the inference becomes more challenging in the presence of multiple generations of mixture, *e.g.* when each of the four grandparents originates from a different ancestral location. Here too, a recent effort to address this problem has been made by the method SPAMIX [27], but as we show below this method does not

provide accurate localization either. Multiple-ancestor geographic localization is therefore an open challenge.

In this paper we provide LIZARD (LocalIZAtion of Recently aDmixed individuals), a method for estimating the multiple geographic coordinates of origin for individuals of recent admixed ancestry. Our approach is based on modeling both the long- and short-range correlations that exist in the genetic sequence of admixed individuals. The long-range correlations, also termed *Admixture Linkage Disequilibrium*, result from the recent mixing, due to which the chromosomes become mosaics of long haplotypic segments originating from the same location; because origin affects sequence, the alleles of SNPs residing on the same ancestral haplotype are correlated. Next, given a haplotype's location of origin, short-range correlations exist between nearby SNPs due to Linkage Disequilibrium (LD). Our model captures the short-range correlations by modeling haplotypes within short genomic windows using the multivariate normal distribution (MVN). The long-range correlations are captured by combining these windows into Hidden Markov Models (HMMs) whose structure and between-window transition rates are determined by the mixing pattern.

We evaluate LIZARD on both simulated and real data of European individuals. LIZARD attains high accuracy in localizing the two parental locations of individuals whose father and mother originate from different geographic regions in Europe, with a median error of 374 km, considerably better than other existing approaches. We use real individuals of admixed ancestry from Europe whose multiple origins are known to validate the method on real data, and observe that LIZARD's localization accuracy remains high and similar to the simulation results. Next, we test our approach of individuals of 2 and 3-generation admixture. As expected, the method's performance deteriorates as g increases, though the results remain useful for downstream applications (median error of 478 and 571 km for 2 and 3 generations, respectively). We discuss the limitations of our approach for large values of g and suggest directions for future work. A software package implementing our method will be freely available upon publication of the manuscript.

2 Materials and Methods

We define an individual to be *of homogeneous ancestry* (or simply *homogenous*) if all of their ancestors originated from the same geographic location, *i.e.* from identical geographic coordinates. In addition, we say that an admixed individual resulted from a *g-generations admixture* if each of their 2^g ancestors from g generations ago is of homogeneous ancestry, and g is the smallest number for which this holds. In reality, locations are never identical and individuals are never homogenous, as we are all mixed to some extent, but we can use them as approximations when the scale of geographic variation in the study sample is large enough.

We assume that a reference panel of $2n$ haplotypes whose locations of origin are known is available through phasing the genotypes of n individuals of

homogenous ancestry. We denote the haplotypes by $H = (h_1, ..., h_{2n})$ and the corresponding locations by $X = (x_1, ..., x_{2n})$. Each location is a vector which contains longitude and latitude coordinates, and since the individuals are homogenous $x_{2i-1} = x_{2i}\ \forall i \in \{1 \dots n\}$. Given an individual of recent admixed ancestry, our goal is to predict their geographical origins by utilizing the information in the reference panel.

We begin by introducing our spatial model and a procedure for estimating its parameters from a group of homogenous individuals. We then present a model for individuals of 1-generation admixture, *i.e.* whose parents originate from different locations but are themselves of homogeneous ancestry, and show how it can be used to estimate the two parental origins. Finally, we extend this model to localize individuals of g-generations admixture (for instance, a case of two generations with four different ancestries).

2.1 Estimation of spatial parameters from training data

We split the haplotypes into L non-overlapping contiguous *windows* of l SNPs per window, and denote by h_{ij} the part of haplotype h_i confined to window j. Similarly to [3], our model assumes that given h_i's location of origin, x_i, h_{ij} is sampled from a multivariate normal distribution (MVN), with window-specific and location-dependent expectation $\beta_j x_i$ and window-specific covariance Σ_j. Here β_j is an $l \times d$ matrix and Σ_j is an $l \times l$ matrix, where d is the dimension of the spatial representation - 2 in our case, for latitude and longitude. The probability of observing haplotype h_i in window j given location x_i equals the multivariate normal likelihood:

$$\begin{aligned} L(h_{ij} \mid \beta_j, \Sigma_j, x_i) &= MVN(h_{ij}; \beta_j x_i, \Sigma_j) \\ &= \frac{1}{(2\pi)^{\frac{l}{2}} \mid \Sigma_j \mid^{\frac{1}{2}}} e^{-\frac{1}{2}(\beta_j x_i - h_{ij})^T \Sigma_j^{-1} (\beta_j x_i - h_{ij})} \end{aligned} \tag{1}$$

Despite the fact that the multivariate normal distribution is continuous while genotypes are discrete, MVNs has been shown to model genotypic data well in multiple contexts [25,14,6], including localization [3]. The advantage of using MVNs is the ability to derive closed-form, efficiently-computable maximum likelihood solutions for the model parameters while accounting for pairwise correlations between SNPs.

The training stage in which we estimate the parameters of the multivariate normal distribution from the reference haplotypes has been previously derived [3], and we describe it here briefly. Denote by H_j the matrix whose ith column is haplotype h_{ij}, and by X the matrix whose ith column is the corresponding location vector x_i. Then the maximum likelihood estimator of β_j, Σ_j have the following closed form solution:

$$\hat{\beta}_j = H_j X^T (X X^T)^{-1} \tag{2}$$

$$\hat{\Sigma}_j = \frac{1}{2n} \sum_{i=1}^{2n} \left(\hat{\beta}_j x_i - h_{i,j}\right)\left(\hat{\beta}_j x_i - h_{i,j}\right)^T \tag{3}$$

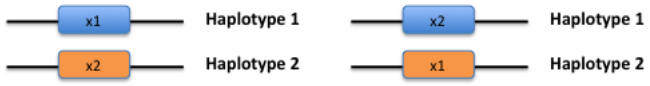

Fig. 1. The two possible phasing arrangements per window. Given the two phased haplotypes of a 1-generation admixed individual confined to a single genomic window and the individual's two ancestral locations x_1, x_2, for most windows it holds that either h^1 originated from x_1 and h^2 originated from x_2, or the other way around.

We follow the standard procedure of adding a small positive constant ($\epsilon = 0.01$) to the diagonal of $\hat{\Sigma}_j$ in order to ensure it is full rank and eliminate potential overfitting of the covariance parameters due to the limited sample size.

2.2 Localization of individuals of 1-generation admixture

We are given the two phased haplotypes h^1, h^2 of an individual whose parents are homogenous, and would like to determine the two corresponding locations of origin x_1, x_2. If phasing was error-free, the problem would be reduced to separately localizing each haplotype; to do that, we can use known methods that infer the location of haplotypes of homogeneous ancestry [26,3]. Unfortunately, perfect phasing is typically not feasible, and as a result, each haplotype is mosaic of parental and maternal segments.

Our model for 1-generation admixed individuals accounts for phasing errors by allowing for the two locations of origin to change along the haplotypes. In addition, the model assumes that there are no phasing errors within a window, and therefore h_j^1 and h_j^2 originate from a single location for every window j. As a result, we allow for one of two *phasing arrangements* per windows: Either h_j^1 originated from x_1 and h_j^2 from x_2, or the other way around (see Figure 1 for illustration). Although the assumption of perfect phasing within windows may not hold for all windows, if the window size is set appropriately, it should hold for most of them.

The exact structure of the model is as follows: We combine the window-specific parameters estimated from the reference panel as in 2.1 into an HMM specified by the triplet (Q, ϵ, δ): Q is the set of states, δ are the transition probabilities, and ϵ are the emission probability functions. The set Q contains 2 x L states: For each window $j \in \{1 \dots L\}$ there are two states, denoted $s_j = \{s_j^1, s_j^2\}$, for the two possible phasing arrangements. Given x_1 and x_2, the two states of window j emit the haplotypes in the window in probabilities that are determined by that MVN densities estimated in equation (2):

$$\begin{aligned}\epsilon_{s_j^1}(h_j^1, h_j^2; x_1, x_2) &= MVN(h_j^1; \beta_j x_1, \Sigma_j) \cdot MVN(h_j^2; \beta_j x_2, \Sigma_j) \\ \epsilon_{s_j^2}(h_j^1, h_j^2; x_1, x_2) &= MVN(h_j^2; \beta_j x_1, \Sigma_j) \cdot MVN(h_j^1; \beta_j x_2, \Sigma_j) \end{aligned} \quad (4)$$

The between-states transition rate is constant across windows and is denoted p_s. This constant is determined by the phasing error rate and the window size,

and its exact choice is discussed in section 2.5. We therefore have

$$\delta(s^{k1}_{(j-1)} \to s^{k2}_j) = \begin{cases} 1 - p_s & \text{if } k1{=}k2 \\ p_s & \text{otherwise.} \end{cases} \tag{5}$$

When performing localization of a 1-generation admixed individual, the HMM we have just described is nearly fully parameterized; the only missing parameters are the two location vectors (x_1, x_2). We use the Baum-Welch algorithm to estimate these parameters, thereby localizing the individual. Specifically, denote by (z^1_j, z^2_j) the indicator variables for the states (s^1_j, s^2_j):

$$z^1_j = I_{s_j = s^1_j} = \begin{cases} 1 & \text{if } s_j = s^1_j \\ 0 & \text{if } s_j = s^2_j \end{cases} \tag{6}$$

and similarly for z^2_j, so that for each window $z^1_j + z^2_j = 1$. In iteration t of the algorithm we use the Forward-Backward algorithm to compute $(z^{1(t)}_j, z^{2(t)}_j)$, the posterior probabilities of the indicator variables, for every window j. We then search for the location parameters that maximize the expected log likelihood:

$$\begin{aligned}(x^{(t)}_1, x^{(t)}_2) = \underset{x_1,x_2}{\operatorname{argmax}} \sum_{j=1}^{L} \Big[& (z^{1(t)}_j \log(MVN(h^1_j; \beta_j x_1, \Sigma_j) \cdot MVN(h^2_j; \beta_j x_2, \Sigma_j)) \\ + \quad & z^{2(t)}_j \log(MVN(h^1_j; \beta_j x_2, \Sigma_j) \cdot MVN(h^2_j; \beta_j x_1, \Sigma_j)) \Big]\end{aligned} \tag{7}$$

The values of x_i, $i = 1, 2$ that maximize the above expression have the following closed form solutions:

$$x^{(t)T}_i = \left(\sum_{j=1}^{L} \left(z^{i(t)}_j h^1_j + (1 - z^{i(t)}_j) h^2_j \right)^T \Sigma^{-1}_j \beta_j \right) \left(\sum_{j=1}^{L} (\beta^T_j \Sigma^{-1}_j \beta_j) \right)^{-1} \tag{8}$$

We repeat the expectation-maximization iterations until convergence of the log likelihood (change smaller than 10^{-8}), allowing for up to 100 iterations.

2.3 A spatial model for individuals of *g*-generations admixture

We now extend the model to describe individuals of g-generation admixture. For such individuals each of the chromosomes (paternal and maternal) has originated from a different (but potentially overlapping) set of up to 2^{g-1} locations. Because the admixture is recent, the pair of locations in window j are highly correlated with the pair of locations of window $j + 1$. In order to capture these correlations, we group each K consecutive windows into a *block*, and assume no recent recombinations within blocks. As a result, all windows within the block contain the same two locations of origin, though perhaps with different phasing

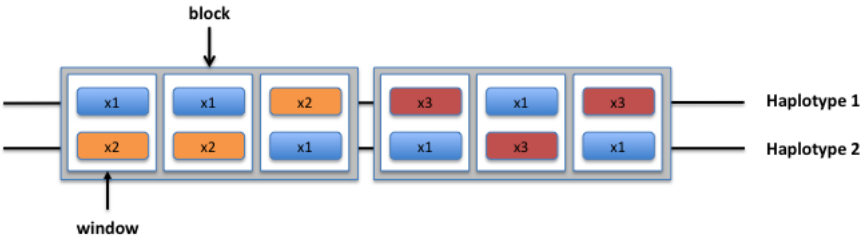

Fig. 2. Model Assumptions for g-generation admixed individuals. Two assumptions are made: (1) The two ancestral locations are constant within block (colored gray), (2) no phasing errors within windows (colored white). In the figure, all windows in the first block contain the locations x_1 and x_2, but different windows in the block have different phasing arrangements. In the second block x_3 replaces x_2 due to a recombination that occurred in the parental DNA.

arrangements. These assumptions are illustrated in Figure 2. We note that in our model windows and blocks have fixed sizes across the genome, but a more accurate approach would be to set their sizes according to the region-specific recombination rates.

We model these assumptions using a two-level HMM defined as follows (and illustrated in Figure 3):

1. Per each block we construct $2^{2(g-1)}$ bottom-level HMMs, one for each unordered pair of parental locations - a paternal one and a maternal one. Each of these HMMs contains $K \times 2$ states, thereby capturing all possible phasing arrangements in the block's windows. These HMMs are identical to the HMM defined in section 2.2, but are confined to a single block.
2. We combine the bottom-level HMMs into a single top-level HMM. Denote by B the number of blocks in the genome, then the top-level HMM has $B \times 2^{2(g-1)}$ states, corresponding to all possible assignments of paternal and maternal locations along the genome.

We estimate the 2^g location vectors using the Baum-Welch algorithm. The Expectation step in iteration t consists of the following two sub-steps:

1. For each block b and for each pair of locations r, we run the Forward-Backward algorithm on the bottom-level HMM. This computation yields the emission probability of block b and pair r in the top-level HMM. The computation also gives us $(z_{br1}^{1(t)}, z_{br1}^{2(t)}) \dots (z_{brK}^{1}, z_{brK}^{2})$, the posterior probabilities of the per-window state variables in this block for this pair.
2. We run the Forward-Backward algorithm on the top level HMM. The transition probabilities between blocks are assumed to be fixed along the genome and are determined by the average genome-wide recombination rate and by g. For block b this computation gives us $(w_{b1}^{(t)} \dots w_{b2^{2(g-1)}}^{(t)})$, the conditional probabilities of the per-block indicator variables corresponding to each of the $2^{2(g-1)}$ states.

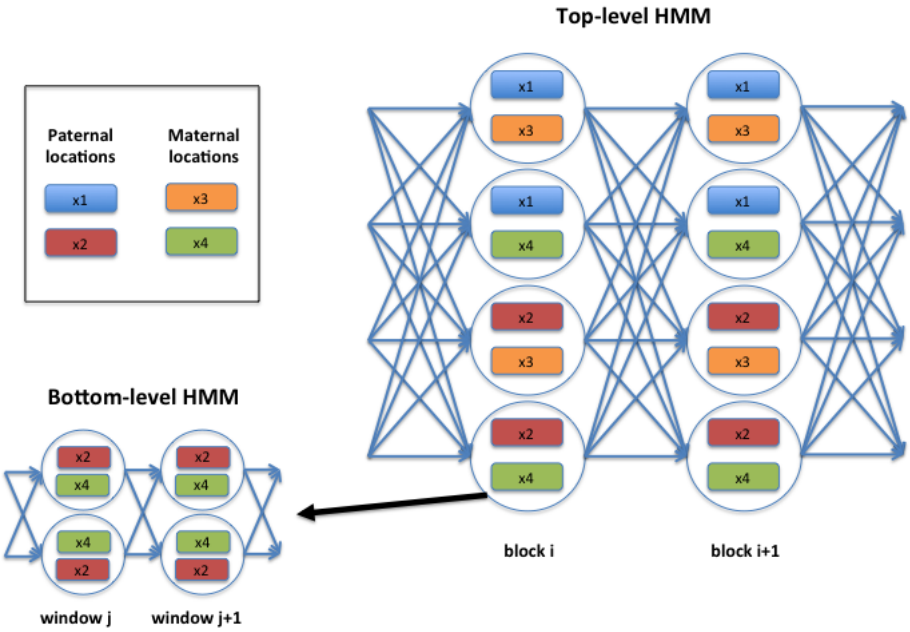

Fig. 3. Two-level HMM for modeling 2-generation admixed individuals. The top-level HMM includes four states per block - one for every unordered combination of paternal and maternal states. In addition, for every block we define a bottom-level HMM over the block's windows: Given the block's top-level state, each window has two states, one for every possible phasing arrangement.

In the maximization step we derive the locations that maximize the expected log likelihood. The optimal x_i has the following closed form solution:

$$x_i^T = \left(\sum_{b=1}^{B}\sum_{r\in R_i}\sum_{j=1}^{K} w_{br}^{(t)}(z_{brj}^{1(t)}h_j^1 + z_{brj}^{2(t)}h_j^2)^T \Sigma_j^{-1}\beta_j\right)\left(\sum_{b=1}^{B}\sum_{r\in R_i}\sum_{j=1}^{K} w_{br}^{(t)}\beta_j^T\Sigma_j^{-1}\beta_j\right)^{-1} \tag{9}$$

where R_i is the set of location pairs that include x_i as one of the two locations.

2.4 Simulation setup

We tested our methods on simulated data generated from the POPRES European samples [16]. Our dataset consists of 364,373 SNPs with minor allele frequency > 0.01 and no-call rate $< 10\%$. We used BEAGLE [4] for phasing and imputation. This data contains 1385 individuals whose four grandparents were reported to originate from the same country. Following phasing we obtained a set of 2×1385 *homogenous haplotypes*, one part of which was used for training the different methods, the other for simulating admixed individuals.

We simulated European admixed haplotypes as mosaics of homogenous haplotypes. Specifically, for a g-generation admixed individual we first drew the number of recombinations from a Poisson distribution with a parameter determined by g, and then uniformly sampled the paternal and maternal locations, separately for the paternal and maternal haplotypes. Finally, we used the homogenous haplotypes to fill out the genotype values according to the determined recombination events. For all simulated individuals we assumed all 2^g ancestral origins to be different from each over. Overall we simulated 1500, 500 and 300 Europeans of 1, 2 and 3-generation admixture, respectively. Finally, we introduced phasing errors into the phased haplotypes at a rate determined through simulations. Specifically, we joined pairs of phased homogenous haplotypes from different individuals to form artificial genotypes, used BEAGLE to phase them, and measured the error rate produced by the phasing procedure.

2.5 Method comparison

We compared LIZARD to two recently published methods for the inference of ancestral locations in recently admixed individuals, SPA [26] and SPAMIX [27]. In a nutshell, SPA's model assumes that a SNP's frequency changes linearly across the geographic space, similarly to LIZARD. However, SPA does not model LD, and only works for 1-generation admixture. As for SPAMIX, it does not model LD either, but does model g-generation admixture as well as admixture LD. In accordance with these differences, SPA has only $\mathcal{O}(m)$ parameters (m is the number of SNPs), all of them spatial, while SPAMIX has $\mathcal{O}(2^g)$ additional parameters that determine the individual's admixture proportions. As for LIZARD, its model contains $\mathcal{O}(ml)$ spatial parameters (l is the window size) in order to capture local LD, and only a fixed number of parameters that determine the HMM transition rates. We note that both SPA and SPAMIX should perform approximately the same on 1-generation admixture due to the lack of admixture-LD in these individuals.

All methods were evaluated on simulated data (see section 2.4) of 1-generation admixed Europeans; in addition, LIZARD and SPAMIX were evaluated on 2,3-generation admixed individuals, and LIZARD on real European samples. All methods were trained on the same set of POPRES homogenous individuals - SPA and SPAMIX on genotypes, LIZARD on haplotypes. LIZARD's window size (l), block size (K) and phasing switch rate (p_p) parameters were tuned on the training data and set to $l = 100$, $K = 20$ and $p_p = 0.1$. These parameters are likely to be optimized by other values in other datasets, but we observe that the method is robust to their setting within a wide range (results not shown). One possible strategy for adjusting these parameters is simulating admixed individuals using each study's specific SNP set and relevant recombination maps, and choosing the optimal values in a cross-validation scheme.

A first measure of performance is the distance in kilometers between the estimated geographical coordinates and the true coordinates; the latter was taken to be the center of the (known) country of origin. The calculation of the distance

Table 1. km error on simulated 1-generation admixed Europeans. For each method we given the 0.5 [0.25, 0.75] quantiles of km error over all simulated individuals.

Method	Error
LIZARD	374.95 [267.76, 519.05]
SPA	1141.4 [706.75, 1674.6]
SPAMIX	1159.6 [723.13, 1704.7]

error involved applying a previously described transformation [17,26,3]. The per-individual error is computed as the average error over all locations. Since for each individual an unordered set of locations is estimated, we choose the permutation that produces the best match between the estimated and the true locations.

A second measure of performance is the fraction of accurate assignments to country of origin. Assignments were obtained by choosing the country whose center is closest to the estimated location. Since multiple countries are assigned per individual, the accuracy we report is the fraction of countries that were correctly detected, averaged over individuals. We emphasize, though, that our method aims at continuous assignment and not at classification, and we report classification here only as a proxy to the former, in the absence of exact location information for the POPRES individuals.

3 Results

3.1 Localization of simulated 1-generation admixed individuals

1-generation admixed individuals were simulated as described in section 2.4. As expected, SPA and SPAMIX achieve approximately the same results, and LIZARD outperforms both of them, presumably due to its improved modeling approach. In terms of km error, LIZARD attains a median error of 374 km compared with SPA's 1141 km and SPAMIX's 1159 km (Table 1). These differences are reflected also in improved accuracy of assignment to country of origin, as shown in Figure 4: LIZARD's average success rate is 57%, while both SPA and SPAMIX attain an average success rate of 42%. The two country pairs for which LIZARD did not attain the highest accuracy both involve Portugal, and we observe that when making these errors LIZARD localized the samples too far to the East, resulting in a mis-classification to Spain. More generally, LIZARD is more likely to make an error when the two countries involved are in geographic proximity; in some of these cases these supposedly wrong assignments may be artifacts of the assignment scheme, which assigns an individual to the country whose center is the closest.

3.2 Localization of real Europeans from the POPRES dataset

We used 254 real 1-generation admixed Europeans from the POPRES dataset to calibrate the performance estimates we obtained in the simulations. LIZARD's

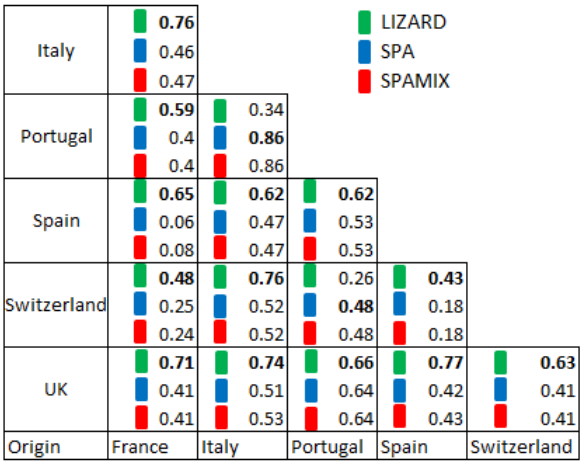

Fig. 4. Accuracy of assignment to country of origin for simulated 1-generation admixed Europeans. For each pair of locations we give the fraction of haplotypes correctly classified to their country of origin, per method. The panel includes only populations that are represented by at least 40 individuals in the training data. In bold is the method which achieved the highest accuracy.

median km error on this data was 368 km, but this number cannot be directly compared to the simulations results due to the difference in ancestral composition between the two test panels. We therefore generated additional simulated datasets so that each real individual is matched with a simulated individual with identical locations of origin. We generated ten such simulated datasets so as to account for the variance resulting from the sampling of the haplotypes. Figure 5 shows that LIZARD's localization error on the real data is similar to the error observed in simulation: For example, LIZARD achieves median errors of 396 and 399 km on the real and simulated datasets, respectively, for individuals originating from Switzerland and France.

3.3 Localization of simulated g-generation admixed individuals

We simulated individuals of g-generation admixture as described in section 2.4. LIZARD localizes individuals of 2-generation admixture who originated from $2^g = 4$ different locations with a median error of 478 km, and individuals of 3-generation admixture and $2^g = 8$ different origins with a median error of 571 km (Figure 6). SPAMIX's error is considerably higher - medians of 1170 and 1332 km for 2 and 3 generations, respectively. The deterioration in the performance of both methods is unsurprising, mostly because as g grows the number of locations increases exponentially with g, while the amount of data to estimate each location decreases exponentially. The problem therefore becomes harder very fast if no additional assumptions about the locations are made, and we discuss possible solutions to this in the Discussion. We also note that

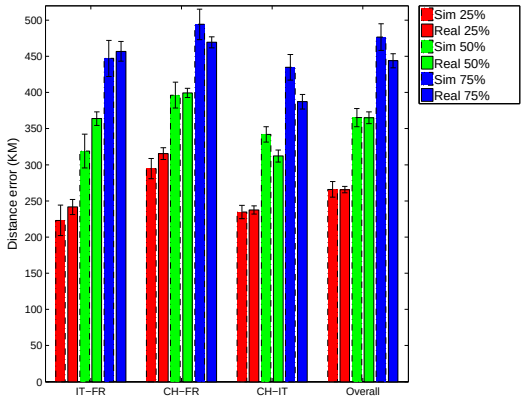

Fig. 5. LIZARD's km error on real data. LIZARD's accuracy was measured on the three main groups of admixed populations in POPRES (IT-Italy; FR-France; CH-Switzerland). The figure gives the 0.25, 0.50 and 0.75 quantiles of km error per group and per method. Error bars for simulated data give the standard error over ten draws of simulated datasets, and for real data the standard error over individuals. The resulting assignment accuracy are 0.68, 0.54 and 0.7 for IT-FR, CH-FR and CH-IT, respectively.

because we choose the best match between true and estimated locations over all possible permutations (see section 2.5), a method that produces random location estimates is expected to score better as g increases. When compared to a random method (Figure 6), LIZARD can be seen to achieve significantly better results for up to 3 generations. Our results therefore suggest that LIZARD is suitable only for recently admixed individuals.

4 Discussion

In this paper we presented LIZARD, a new method for the inference of ancestral coordinates for individuals of recent admixed ancestry. LIZARD is capable of accurately inferring the origins of 1-generation admixtures, and with a lesser success of 2 and 3-generation mixtures. Its improved performance compared with existing approaches is achieved by modeling both long-range genomic correlations due to recent admixture, and short-range correlations due to linkage disequilibrium. As a result of using closed-form optimization formulae, LIZARD runs fast: Its training on a reference set of thousands of haplotypes take a few minutes, and localizing each 1-generation individual takes 45 seconds, on average. We note that LIZARD requires haplotype data, and phasing may take up to a few days, depending on the available computational resources; however, phasing is usually performed in any case as a routine part of the data analysis.

As the number of generations in admixture increase, LIZARD's performance deteriorates rapidly. The main reason is that the length of genomic sequence available for determining each origin decrease exponentially with g. Moreover,

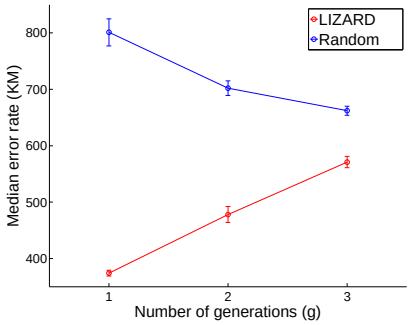

Fig. 6. LIZARD's km error for g-generation admixture. LIZARD is compared with a random assignment. Error bars give the standard error of the median as estimated from a 10-fold cross-validation experiment.

the average length of each single ancestral segment decreases, and hence the long-range correlations decay faster. Finally, in terms of efficiency, the complexity of our algorithm is exponential in g. We close this paper by suggesting a few enhancements that will enable better handling of larger g values.

First, it is often possible to utilize existing information about the ancestral coordinates. In some cases, priors distributions can be formulated, at least for some of the ancestors. In other cases, it is known that all ancestors from one side of the family originate from the same region, and this information can be easily integrated into our optimization. Second, our model assumes that within a parent's haplotype (paternal or maternal), a segment from any location is equally likely to be followed by a segment from any other location. In fact, the transition patterns between locations contain regularities due to the pedigree structure that induced the mixing, and future methods could model these regularities.

More generally, continuous localization as we have attempted here is a qualitatively more difficult task than classification to discrete ancestries. Spatiogenetic modeling of human data is a relatively new research direction, and much work remains to be done in refining the models that underlie current localization methods; we expect that such refinements will yield a significant improvement in localization accuracy of both homogenous and admixed individuals.

Acknowledgements

E.H. is a faculty fellow of the Edmond J. Safra Center at Tel Aviv University. Y.B. was supported in part by a fellowship from the Edmond J. Safra Center for Bioinformatics at Tel-Aviv University. E.H. and Y.B. were also supported in part by the United StatesIsrael Binational Science Foundation (grant 2012304), and by the National Science Foundation (grant III-1217615), and by the Israeli Science Foundation (grant 989/08). E.H, Y.B, and Y.M were partially supported by the German-Israeli Foundation (grant 1094-33.2/ 2010). E.H was also supported by the Israel Science Foundation (grant 1425/13).

References

1. Alexander, D.H., Novembre, J., Lange, K.: Fast model-based estimation of ancestry in unrelated individuals. Genome research 19(9), 1655–1664 (2009)
2. Baran, Y., Pasaniuc, B., Sankararaman, S., Torgerson, D.G., Gignoux, C., Eng, C., Rodriguez-Cintron, W., Chapela, R., Ford, J.G., Avila, P.C., et al.: Fast and accurate inference of local ancestry in latino populations. Bioinformatics 28(10), 1359–1367 (2012)
3. Baran, Y., Quintela, I., Carracedo, Á., Pasaniuc, B., Halperin, E.: Enhanced localization of genetic samples through linkage-disequilibrium correction. The American Journal of Human Genetics 92(6), 882–894 (2013)
4. Browning, S.R., Browning, B.L.: Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. The American Journal of Human Genetics 81(5), 1084–1097 (2007)
5. Bryc, K., Velez, C., Karafet, T., Moreno-Estrada, A., Reynolds, A., Auton, A., Hammer, M., Bustamante, C., Ostrer, H.: Genome-wide patterns of population structure and admixture among hispanic/latino populations. Proceedings of the National Academy of Sciences 107(Supplement 2), 8954–8961 (2010)
6. Churchhouse, C., Marchini, J.: Multiway admixture deconvolution using phased or unphased ancestral panels. Genetic epidemiology 37(1), 1–12 (2013)
7. Gravel, S., Henn, B., Gutenkunst, R., Indap, A., Marth, G., Clark, A., Yu, F., Gibbs, R., Bustamante, C., Altshuler, D., et al.: Demographic history and rare allele sharing among human populations. Proceedings of the National Academy of Sciences 108(29), 11983–11988 (2011)
8. Haiman, C.A., Patterson, N., Freedman, M.L., Myers, S.R., Pike, M.C., Waliszewska, A., Neubauer, J., Tandon, A., Schirmer, C., McDonald, G.J., et al.: Multiple regions within 8q24 independently affect risk for prostate cancer. Nature genetics 39(5), 638–644 (2007)
9. Hinch, A., Tandon, A., Patterson, N., Song, Y., Rohland, N., Palmer, C., Chen, G., Wang, K., Buxbaum, S., Akylbekova, E., et al.: The landscape of recombination in african americans. Nature 476(7359), 170–175 (2011)
10. Jarvis, J., Scheinfeldt, L., Soi, S., Lambert, C., Omberg, L., Ferwerda, B., Froment, A., Bodo, J., Beggs, W., Hoffman, G., et al.: Patterns of ancestry, signatures of natural selection, and genetic association with stature in western african pygmies. PLoS genetics 8(4), e1002641 (2012)
11. Johnson, N.A., Coram, M.A., Shriver, M.D., Romieu, I., Barsh, G.S., London, S.J., Tang, H.: Ancestral components of admixed genomes in a mexican cohort. PLoS genetics 7(12), e1002410 (2011)
12. Kao, W.L., Klag, M.J., Meoni, L.A., Reich, D., Berthier-Schaad, Y., Li, M., Coresh, J., Patterson, N., Tandon, A., Powe, N.R., et al.: Myh9 is associated with nondiabetic end-stage renal disease in african americans. Nature genetics 40(10), 1185–1192 (2008)
13. Maples, B.K., Gravel, S., Kenny, E.E., Bustamante, C.D.: Rfmix: a discriminative modeling approach for rapid and robust local-ancestry inference. The American Journal of Human Genetics 93(2), 278–288 (2013)
14. Menelaou, A., Marchini, J.: Genotype calling and phasing using next-generation sequencing reads and a haplotype scaffold. Bioinformatics 29(1), 84–91 (2013)
15. Moreno-Estrada, A., Gravel, S., Zakharia, F., McCauley, J.L., Byrnes, J.K., Gignoux, C.R., Ortiz-Tello, P.A., Martínez, R.J., Hedges, D.J., Morris, R.W., et al.:

Reconstructing the population genetic history of the caribbean. PLoS genetics 9(11), e1003925 (2013)
16. Nelson, M.R., Bryc, K., King, K.S., Indap, A., Boyko, A.R., Novembre, J., Briley, L.P., Maruyama, Y., Waterworth, D.M., Waeber, G., et al.: The population reference sample, popres: a resource for population, disease, and pharmacological genetics research. The American Journal of Human Genetics 83(3), 347–358 (2008)
17. Novembre, J., Johnson, T., Bryc, K., Kutalik, Z., Boyko, A.R., Auton, A., Indap, A., King, K.S., Bergmann, S., Nelson, M.R., et al.: Genes mirror geography within europe. Nature 456(7218), 98–101 (2008)
18. Price, A.L., Tandon, A., Patterson, N., Barnes, K.C., Rafaels, N., Ruczinski, I., Beaty, T.H., Mathias, R., Reich, D., Myers, S.: Sensitive detection of chromosomal segments of distinct ancestry in admixed populations. PLoS genetics 5(6), e1000519 (2009)
19. Price, A.L., Zaitlen, N.A., Reich, D., Patterson, N.: New approaches to population stratification in genome-wide association studies. Nature Reviews Genetics 11(7), 459–463 (2010)
20. Pritchard, J.K., Stephens, M., Donnelly, P.: Inference of population structure using multilocus genotype data. Genetics 155(2), 945–959 (2000)
21. Seldin, M.F., Pasaniuc, B., Price, A.L.: New approaches to disease mapping in admixed populations. Nature Reviews Genetics 12(8), 523–528 (2011)
22. Smith, M.W., O'Brien, S.J.: Mapping by admixture linkage disequilibrium: advances, limitations and guidelines. Nature Reviews Genetics 6(8), 623–632 (2005)
23. Wang, C., Zöllner, S., Rosenberg, N.A.: A quantitative comparison of the similarity between genes and geography in worldwide human populations. PLoS genetics 8(8), e1002886 (2012)
24. Wegmann, D., Kessner, D., Veeramah, K., Mathias, R., Nicolae, D., Yanek, L., Sun, Y., Torgerson, D., Rafaels, N., Mosley, T., et al.: Recombination rates in admixed individuals identified by ancestry-based inference. Nature genetics 43(9), 847–853 (2011)
25. Wen, X., Stephens, M.: Using linear predictors to impute allele frequencies from summary or pooled genotype data. The annals of applied statistics 4(3), 1158 (2010)
26. Yang, W.Y., Novembre, J., Eskin, E., Halperin, E.: A model-based approach for analysis of spatial structure in genetic data. Nature genetics 44(6), 725–731 (2012)
27. Yang, W.Y., Platt, A., Chiang, C.W.K., Eskin, E., Novembre, J., Pasaniuc, B.: Spatial localization of recent ancestors for admixed individuals. G3: Genes, Genomes, Genetics (2014)
28. Zhu, X., Luke, A., Cooper, R.S., Quertermous, T., Hanis, C., Mosley, T., Gu, C.C., Tang, H., Rao, D.C., Risch, N., et al.: Admixture mapping for hypertension loci with genome-scan markers. Nature genetics 37(2), 177–181 (2005)